Enhancing Hotelling's T^2 Statistic using Shrinkage Covariance Matrix for Identifying Differentially Expressed Gene Sets

Suryaefiza Karjanto [a, b], Rasimah Aripin [c], Norazan Mohamed Ramli [d] and Nor Azura Md Ghani [d]

[a] Department of Computer and Mathematical Sciences, Universiti Teknologi MARA, 13500 Permatang Pauh, Pulau Pinang, Malaysia
[b] Laboratory of Department of Computer and Mathematical Sciences, Universiti Teknologi MARA, 13500 Permatang Pauh, Pulau Pinang , Malaysia
[c] Faculty of Science and Technology, Sunway University, Jalan Universiti, Bandar Sunway, 46150 Petaling Jaya, Selangor, Malaysia
[d] Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, 40450 Shah Alam, Selangor, Malaysia.
suryaefiza@gmail.com

Abstract. The breakthrough of microarray technology is a vital research instrument to measure the quantitative and highly parallel of gene expression. In microarray studies, it is common that the data set typically consists of tens of thousands of genes (variables) from just dozens of samples due to various constraints including the high cost of producing microarray chips. As a result, the combined sample covariance matrix in Hotelling's T^2 statistic is not invertible. Therefore the distribution of the resulting statistic is either unknown or insufficient. In this study, shrinkage covariance matrix is proposed to improve Hotelling's T^2 statistic for identification of differentially expressed gene sets. The use of shrinkage covariance matrix overcomes the non-singularity problem in the estimation of sample covariance matrix when the number of variables is larger than the number of samples. The performance of the proposed method was measured using simulation study. The result implies that this approach outperforms existing techniques in many conditions tested.

Keywords: Hotelling's T^2; gene set analysis; shrinkage covariance matrix

1 Introduction

Microarray technology is one of the significant achievements in biotechnology history, developed during the second half of the 1990s. [1] Many researchers admit the breakthrough of this technology as a vital research instrument. The microarray technology can precisely perform simultaneous analysis of thousands of genes in a massively parallel manner to researchers in one experiment, hence it provides valuable knowledge on gene interaction and function [2, 3, 4, 5]. The challenge of understanding the microarray gene expression leads to the development of new methods in the field of statistics for the analysis of gene expression data such as identification of differential gene expression between distinct experimental conditions [6, 7]. The purpose of differential gene expression studies is discovering those genes that produce different expression levels (the rate at which a gene produces the corresponding protein) between samples [4].

In this research, the Hotelling's T^2 statistic is combined with the shrinkage approach as an estimation alternative to estimate the covariance matrix to perform gene set analysis. Gene set analysis [6, 7, 8, 9] directly finds the group of significant functionally related genes in the list of genes defined from GO or some pathway databases. Moreover, prior analysis and results in this area could be matched up and studied as a result of the achievement in gene set analysis method. The main point of gene set analysis is to show that even small expression changes in individual gene of a functionally related genes group can show a significant pattern [8, 9].

For this reason, a method is defined in this study to detect the differential gene sets that produces different expression levels between samples. The method is introduced in Section 3 after a description on the properties of Hotelling's T^2 statistic 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's T^2 is a natural generalization of *t*-statistic. The information for gene interactions is utilized to allow for finding genes whose differential expressions which are not detectable by univariate methods [10, 11] and widely used in the identification of differential gene expression [11, 12, 13]. Let *n* represent the number of slides/samples, and *p* is 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 $X_i = (X_{k1}$, . . . , $X_{ki})^{\mathrm{T}}$ and $X_j = (X_{k1}, . . . , X_{kj})^{\mathrm{T}}$, respectively. The unknown population covariance matrix, $\sum$, is typically estimated by the sample covariance matrix, S_{ij}, for many situations. The sample covariance matrix, S_{ij} is defined as:

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

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

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

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's T^2 is defined as:

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

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

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

The sub-sample covariance matrix, $S^{(1)}$ and $S^{(2)}$ are defined as in equation (1). The maximum likelihood estimator is employed to obtain the sample covariance matrix. This estimator is 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 curse of high dimensionality data. It is common for multivariate test statistics to involve inversion of covariance matrix, including the Hotelling's T^2 statistic. When *p* is near to *n*, 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: ShrinkA

Another alternative to estimate covariance matrix is shrinkage estimation. This technique improves an estimator by reducing the effect of sampling variation and generally this involves converting an unbiased to an improved biased estimator. The shrinkage estimation is introduced by James and Stein [14] and called as "Stein phenomenon". The new estimator solves high dimensional data problem by minimizing the total mean squared error. This estimator outperforms the maximum likelihood estimator especially when the number of variables is greater than three. In general, a biased estimator is added toward unbiased estimator in the form:

$$(1 - f(x)) * x \tag{5}$$

which the amount of shrinkage $f(x)$ need to be specified [15].

3.1 ShrinkA Algorithm

The proposed method provides an alternative to estimate covariance matrix by using shrinkage method based on the definition of [5] and [16]. The approach is adapted to Hotelling's T^2 and is extended to gene set analysis in microarray study. In this paper, this method would be termed as ShrinkA. Generally, the algorithm for the proposed method is outlined below:

Step 1: Prepare the data sets with the pre-processing procedure using suitable normalization and transformation method (if needed).

Step 2: Compute the shrinkage target.

Step 3: Find the optimal shrinkage intensity using related definition.

Step 4: Replace the sample covariance matrix in Hotelling's T^2 using the results in Step 2 and Step 3.

Step 5: Calculate Hotelling's T^2 for each of all the gene sets that are measured in data sets.

Step 6: Permute samples for each gene set and declare as significant gene sets according to permutation testing.

Basically, the shrinkage covariance matrix is a linear combination of sample covariance matrix with shrinkage target as a biased estimator and shrinkage intensity α as a weight that the shrinkage target receives [16]. The proportion of each component in shrinkage estimation is determined by:

$$S_{shrink} = \alpha T_{Aij} + (1-\alpha) S_{ij} \tag{6}$$

where shrinkage target, T_{Aij} and shrinkage intensity, α is defined as:

$$\alpha = \max\left\{0, \min\left\{\frac{\kappa_A}{n}, 1\right\}\right\} \tag{7}$$

The shrinkage target for ShrinkA, T_{Aij} [5] is as follows:

$$T_{Aij} = \begin{cases} s_{ii} & if \;\; i = j \\ 0 & if \;\; i \neq j \end{cases} \tag{8}$$

where S_{ii} is sample variance of X_i. The shrinkage target and shrinkage intensity of ShrinkA is applied to equations (6) - (8) and this method ensures that the covariance matrix is always a positive definite and well-defined. The method of [13] differs from ShrinkA with respect to the shrinkage target by using average sample correlation as non-diagonal in shrinkage covariance matrix

Under the assumption that n is fixed while p tends to infinity, [10] proved that κA is as follows:

$$\kappa_A = \frac{\pi}{\gamma_A} \tag{9}$$

where

$$\pi = \sum_{i=1}^{n} \sum_{j=1}^{n} AsyVar\left[\sqrt{n} S_{ij}\right] \tag{10}$$

π is the sum of asymptotic variances of the entries of the sample covariance matrix scaled by $\sqrt{n}$.

$$\gamma_A = \sum_{i=1}^{n} \sum_{j=1}^{n} \left(\sigma_{ij}\right)^2 \tag{11}$$

γ_A is the measurement of the misspecification of the (population) shrinkage target for ShrinkA. Because we do not know the true of κ_A , hence we need to estimate the values of π and γ_A (the details can be found in [16]) in order to produce consistent estimator of κ_A :

$$\hat{\kappa}_A = \frac{\hat{\pi}}{\hat{\gamma}_A} \tag{12}$$

The estimator is proven by [16] to be reliable to estimate the real κ_A . Hence, the consistent estimator for π :

$$\hat{\pi} = \sum_{i=1}^{p} \sum_{j=1}^{p} \hat{\pi}_{ij} \tag{13}$$

$$\hat{\pi}_{ij} = \frac{1}{n} \sum_{k=1}^{n} \left\{ \left(X_{ki} - \overline{X}_i\right)\left(X_{kj} - \overline{X}_j\right) - S_{ij} \right\}^2 \tag{14}$$

Next, σ_{ij} are consistently estimated by S_{ij} respectively therefore the consistent estimator for γ_A :

$$\hat{\gamma}_A = \sum_{i=1}^{n} \sum_{j=1}^{n} \left(S_{ij}\right)^2 \tag{15}$$

When we put everything together, $\hat{\kappa}_A$ becomes equation (13) and finally, consistent estimator of α in equation (8) for ShrinkA is defined as:

$$\hat{\alpha} = \max\left\{0, \min\left\{\frac{\hat{\kappa}_A}{n}, 1\right\}\right\} \tag{16}$$

The performance of our approach is evaluated by comparing the results with those obtained from two other methods: by using principal component analysis which is proposed by Kong's principal component analysis (KPCA) [11] to solve the high dimensionality problem; and (2) the Regularized Covariance Matrix Approach (RCMAT), introduced by Yates and Reimers [15]. The RCMAT is quite similar with

ShrinkA but the covariance matrix in Hotelling's T^2 is regularized by using the following identity matrix to replace the shrinkage target in equation (8):

$$T_{ij} = \begin{cases} 1 & if \ i = j \\ 0 & if \ i \neq j \end{cases} \tag{17}$$

Since the shrinkage target is penalized to zero and the diagonal to one, information from the covariance matrix would not be fully utilized [15]. The shrinkage intensity, α in equation (8) is reduced from 1 towards 0 by increments of 0.01 and the optimum shrinkage intensity will be achieved when the smallest positive eigenvalue is bigger than the reciprocal of the number of genes in the gene set. The optimum intensity will ensure the covariance matrix is a positive definite and invertible. RCMAT and KPCA comparable with our approach since they were also using Hotelling's T^2 for testing differentially expressed gene sets.

3.2 A Simulation Study

Multivariate normal distribution data sets were generated using *mvrnorm* function in the *MASS* package. We assumed the generated data as correlation matrix by using *rcorrmatrix* function in the *clusterGeneration* package in *R* Programming Language. In a series of extensive computer simulations, the proposed shrinkage covariance matrix framework was investigated in terms of power of performance. The proposed method was written by using the freely available language *R*. This language can be found at http://cran.r-project.org/. We generated similar simulated data by following Yates and Reimers [15] to make comparisons with both methods.

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

$$fx\left(x_{1,.......,}x_p\right) = \frac{1}{\left(2\pi\right)^{p/2}\left|\Sigma\right|^{1/2}} e^{-\left(x-\mu\right)'\Sigma^{-1}\left(x-\mu\right)/2} \tag{18}$$

The gene set variances were set at one and assumed that the number of samples for both groups is equal. The simulated data sets were set to explore the performance of proposed method for following hypothesis/conditions: 1) There is difference (separation) exists between groups (alternative hypothesis) and; 2) Paired comparison (proposed method *p*-value/ compared method *p*-value). Each condition was permutated 10000 times [18] and 100 data sets were generated. The simulated data sets were constructed by changing the four parameters: 1) Number of variables; 2) Number of sample sizes; 3) Axis of variation (a major axis of variation (first eigenvector, e_1) or a minor axis of variation (p/3 eigenvector, $e_{p/3}$)) and; 4) Amount of separation between groups (de_i).

4 Results and Discussion

We investigated the power of performance between methods when high dimensionality problem occurred. This problem generated the simulated data sets with number of variables smaller than number of samples ($n<p$) and number of variables larger than number of samples ($n>p$) cases. As expected, we observed that the mean of p-value increased as number of variables increased with fixed number of samples, but the ShrinkA was easily detected (with lower mean of p-value) the true difference between two groups rather than RCMAT and KCPA at most of the conditions (Table1).

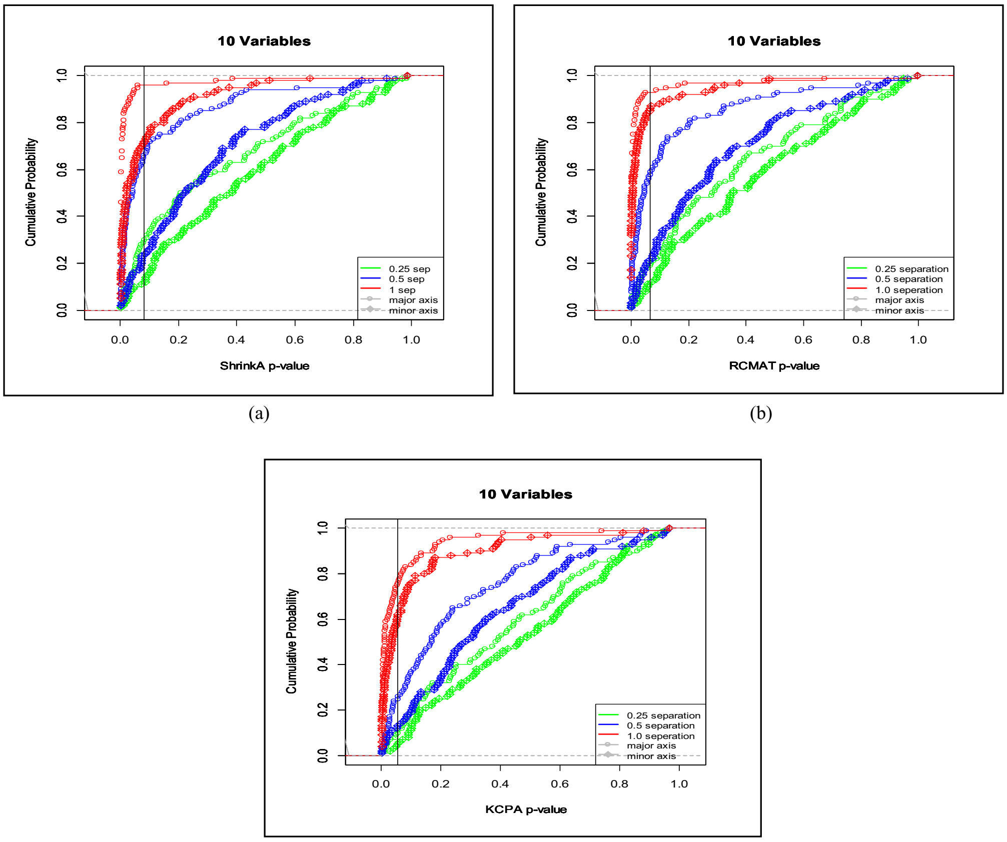
For instance, when number of variables changes from 10 to 30 (with fixed number of samples), the mean of p-value shifted from 0.3261 to 0.3373 for ShrinkA along a major axis of variation and amount of separation is 0.25. For same conditions, mean of p-value shifted from 0.3457 to 0.4213 for RCMAT and from 0.4053 to 0.4430 for KCPA. As the amount of separation increased to 0.5 and 1.0 along a major axis of variation, the mean of p-value is also found to be lower than other methods. In addition, all results of the cumulative distribution function of each methods of nominal p-values is illustrated in Figure 1 (10 variables) and Figure 2 (30 variables).

The relative power between the two methods is shown in Figure 3 for 10 variables and Figure 4 for 30 variables. These figures compare the order of magnitude of ShrinkA based the reduction degree in the ShrinkA p-values relative to the other two methods. Figure 3(a) shows that relative to RCMAT, 10 per cent of the ShrinkA p-values were at least reduced 3.16 times smaller than the corresponding RCMAT p-value for a separation of one along minor axis and a separation of 0.25 along major axis. Relative to KCPA, 20 per cent of the ShrinkA p-values being smaller 3.16 times than the corresponding KPCA p-value for all separations along both axes except separations of 0.25 along minor axis and 0.5 along major axis as shown in Figure 3(b).

Similar improvements were also observed for other combinations of separation and axis for 30 variables (Figure 4(a) and Figure 4(b)). For all separations, about 20 per cent probability of ShrinkA p-values being smaller than 3.16 times than the corresponding RCMAT p-values is shown in Figure 4(a). The 20 per cent of ShrinkA p-values is about 10 times smaller than the corresponding KPCA p-value for all separations except separation of one and 0.25 along a major axis (Figure 4(b)).

Table 1. The mean of the nominal *p*-values of ShrinkA, RCMAT and KCPA.

n, p	**Axis of variation, Amount of separation**	**ShrinkA**	**RCMAT**	**KCPA**
10,20	Major, 0.25	0.3261	0.3457	0.4053
10,20	Minor, 0.25	0.4142	0.4175	0.4764
10,20	Major, 0.5	0.1084	0.1223	0.2395
10,20	Minor, 0.5	0.4153	0.2877	0.3459
10,20	Major,1.0	0.0027	0.0014	0.0192
10,20	Minor, 1.0	0.0055	0.0399	0.0956
30,20	Major, 0.25	0.3373	0.4213	0.4430
30,20	Minor, 0.25	0.4493	0.4735	0.4669
30,20	Major, 0.5	0.1324	0.2092	0.3912
30,20	Minor, 0.5	0.3748	0.3875	0.4395
30,20	Major, 1.0	0.0004	0.0050	0.1560
30,20	Minor, 1.0	0.1594	0.1532	0.3336

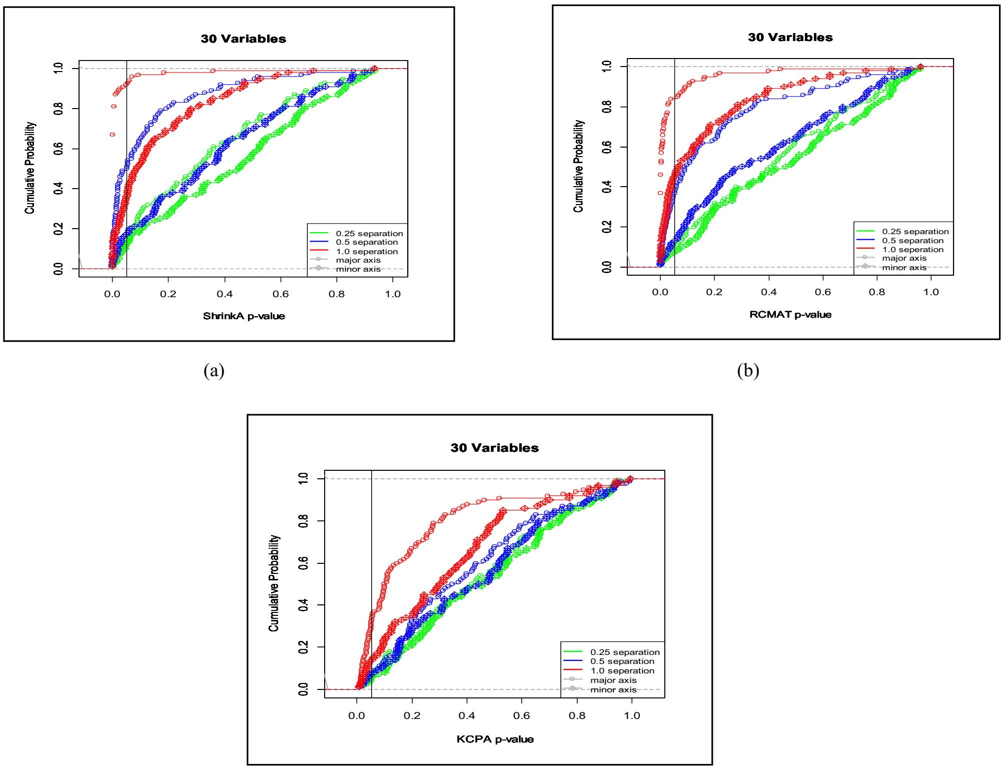


(a)

(b)

(c)

Fig. 1. Cumulative distribution function of ShrinkA, RCMAT and KCPA nominal *p*-values for 10 variables.

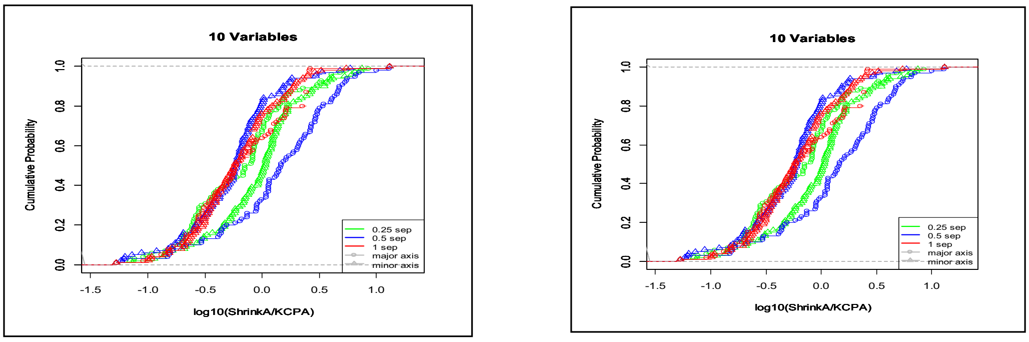


(a)

(b)

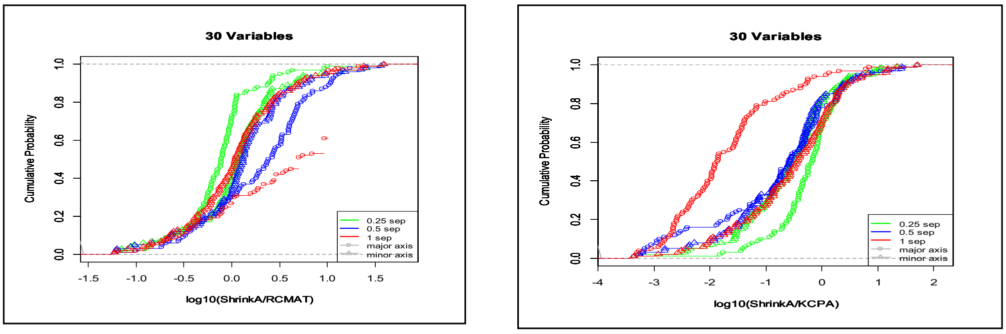
(c)

Fig. 2. Cumulative distribution function of ShrinkA, RCMAT and KCPA nominal *p*-values for 30 variables.



(a) (b)

Fig. 3. Cumulative distribution function of paired comparison between ShrinkA *p*-value/ RCMAT *p*-value and ShrinkA *p*-value/KCPA *p*-value for 10 variables.



(a) (b)

Fig. 4. Cumulative distribution function of paired comparison between ShrinkA *p*- value/ RCMAT *p*-value and ShrinkA *p*-value/KCPA *p*-value for 30 variables.

5 Conclusion

This study discovers the potential of the shrinkage approach to estimate the covariance matrix for microarray data, particularly in differential gene expression area. The use of shrinkage covariance matrix overcomes the non-singularity problem in the estimation of sample covariance matrix when the number of variables is larger than the number of samples. The performance of the proposed method was measured using simulation study and the results show that this approach outperforms existing techniques in many conditions tested. The results are expected to be of interest for further applications in other areas of research with similar data characteristics.

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, R. W., and Brown, P. O., Quantitative monitoring of gene expression patterns with a complementary DNA microarray, *Science*, 1995; **270(5235)**: 467-470.

[2] Babu, M. M., Introduction to microarray data analysis, *Computational Genomics: Theory and Application*, 2004; 225-249.

[3] Szabo A., Boucher K., Jones D., Tsodikov D., Klebanov L.E.V.B and Yakovlev A.Y., 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] Schäfer J. and Strimmer K., A Shrinkage Approach to Large-Scale Covariance Matrix Estimation and Implications for Functional Genomics, *Statistical Applications in Genetics and Molecular Biology,* 2005*;* **4(1)**: 32.

[6] Mootha V.K., Lindgren C.M., Eriksson K.F., Subramanian A., Sihag S., Lehar L., Puigserver P., Carlsson E., Ridderstraale M., Laurila E. *et al*. ,Pgc-1 α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes, *Nature Genetics* **34(3)**, 267–273 (2003).

[7] Song S. and Black M.A., Microarray-based gene set analysis: A comparison of current methods, *BMC Bioinformatics,* 2008; **9**: 502.

[8] Dopazo J., Functional interpretation of microarray experiments, *OMICS,* 2006; **10**, 398–410.

[9] Nam D. and Kim S., Gene-set approach for expression pattern analysis, *Brief Bioinform* 2008; **9**: 189-197.

[10] Lu Y., Liu P.Y., Xiao P. and Deng H.W., Hotelling's T 2 multivariate profiling for detecting differential expression in microarrays, *Bioinformatics* 2005; **21(14)**: 3105–3113.

[11] Kong S.W., Pu W.T. and Park P.J., A multivariate approach for integrating genome-wide expression data and biological knowledge, *Bioinformatics* 2006; **22(19)**: 2373-2380.

[12] Haydenn D., Lazar P., and Schoenfeld D., Assessing statistical significance in microarray experiments using the distance between microarrays, *PLoS One*, 2009; ***4*(6)**: e5838.

[13] Karjanto, Suryaefiza, and Rasimah Aripin., Shrinkage covariance matrix approach for microarray data, In *AIP Conference Proceedings,* 2013; **1522**:1262.

[14] James W. and Stein C., Estimation with quadratic loss, *Proc. 4th Berkeley Sympos. Math. Statist. and Prob.,* 1961; **4(1):** 361-379.

[15] Yates P.D. and Reimers M.A, RCMAT: A regularized covariance matrix approach to testing gene sets, *BMC Bioinformatics,* 2009; **10**: 300.

[16] 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.