International Journal of Pharmacy and Pharmaceutical Sciences

ISSN- 0975-1491 Vol 8, Suppl 2, 2016

Original Article

IMPROVED ESTIMATION OF COVARIANCE MATRIX IN HOTELLING'S T^2 FOR MICROARRAY DATA

SURYAEFIZA KARJANTO[1]*, NORAZAN MOHAMED RAMLI[2], NOR AZURA MD GHANI[2]

[1]Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA 77300 Melaka, Malaysia, [2]Center for Statistical and Decision Sciences Studies, Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, 40450 Shah Alam, Selangor, Malaysia
Email: suryaefiza@gmail.com

Received: 26 Apr 2016 Revised and Accepted: 28 May 2016

ABSTRACT

The relationship between genes in gene set analysis in microarray data is analyzed using Hotelling's T^2 but the test cannot be applied when the number of samples is larger than the number of variables which is uncommon in the microarray. Thus, in this study, we proposed shrinkage approaches to estimating the covariance matrix in Hotelling's T^2 particularly to cater high dimensionality problem in microarray data. Three shrinkage covariance methods were proposed in this study and are referred as Shrink A, Shrink B and Shrink C. The analysis of the three proposed shrinkage methods was compared with the Regularized Covariance Matrix Approach and Kong's Principal Component Analysis. The performances of the proposed methods were assessed using several cases of simulated data sets. In many cases, the Shrink A method performed the best, followed by the Shrink C and RCMAT methods. In contrast, both the Shrink B and KPCA methods showed relatively poor results. The study contributes to an establishment of modified multivariate approach to differential gene expression analysis and expected to be applied in other areas with similar data characteristics.

Keywords: Gene set analysis, Hotelling's T^2, Microarray analysis, Shrinkage covariance matrix

DOI: http://dx.doi.org/10.22159/ijpps.2016v8s2.15215.

INTRODUCTION

Generally, a single microarray slide may contain thousands of spots and each spots signifying a single gene and all of them representing the entire set of genes of an organism [1]. The widespread of microarray technology was 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 [2].

Up until now, wide applications of this technology have been implemented in various fields such as:

Gene expression profiling

Gene expression profiling is the measurement of the expression (activity) of thousands of genes to create an overall picture of cellular function. As a result, the information obtained from microarray gene expression profiling will probably improve our basic understanding about certain diseases or how specific drugs work in cells [3].

Toxicological Research: Toxicogenomics aims to find the relationship between toxic responses to toxicants (a poisonous agent that is made by humans) and differences in the genetic profiles (the procedure of analyzing the DNA for the purpose of identification) of the objects exposed to such toxicants. Microarray technology provides a tremendous platform for the study of the effect of toxins on the cells and they are passing on to the offspring [4].

Disease Diagnosis: Microarray technology has been implemented in disease diagnosis particularly in the study of cancer commonly known as a genetic disease. Researchers enable to determine the gene expression level in a particular cancer cell. Besides, this will also help researchers to define specific molecular pathways thus the identification of the molecular mechanisms involved in cancer leading to the development of effective drugs as the treatment will be targeted directly to the specific type of cancer [5].

Our objective is comprehensive to test the proposed new shrinkage covariance matrix from our previous extension works [6] for detecting significant gene sets between different samples. We stated in Section 2 about the impact of high dimensionality problem or when the number of genes is larger than the number of samples in sample covariance matrix. We also described in Section 3 that the developed simulation study is to evaluate the performance of our proposed methods in detecting significant gene sets. Then, Section 4 will describe the results and discussion and finally the Section 5 will summarize the findings.

MATERIALS AND METHODS

Proposed Shrinkage Covariance Matrix in Hotelling's T2

The relationship between genes in gene set analysis in microarray data is analyzed using Hotelling's T2 as a multivariate test statistic. However, the test cannot be applied when the number of samples is larger than the number of variables which is uncommon in the microarray. Since the microarray dataset typically consists of tens of thousands of genes from just dozens of samples due to various constraints, the sample covariance matrix is not positive definite and singular, thus it cannot be inverted. Thus, in this study, we proposed shrinkage approaches to estimating the covariance matrix in Hotelling's T2 particularly to cater high dimensionality problem in microarray data. The Hotelling's T2 statistic was combined with the shrinkage approach as an alternative estimation to estimate the covariance matrix in detect significant gene sets.

The proposed shrinkage estimation approach is about taking a weighted average of the sample covariance matrix and a structured matrix or shrinkage target as shrinkage of the sample covariance matrix towards a target matrix of the same dimensions while the shrinkage intensity is the weight that the shrinkage target receives. Three shrinkage covariance methods were proposed in this study and are referred as Shrink A, Shrink B and Shrink C. The following notations are used to describe experimental data generated in the form of two-colour spotted microarrays. Let n represent the number of slides/samples, and p is the total number of genes in a gene set.

The proposed methods provide an alternative to estimate covariance matrix using shrinkage method based on the definition of Ledoit and Wolf [7-9] and Schafer and Strimmer [10]. The approach is adapted to Hotelling's T^2 and is extended to gene set analysis in the microarray study. Throughout this study, three different methods are proposed and they will be termed as Shrink A, Shrink B and Shrink C for the rest of this thesis. Generally, the algorithm for the three proposed methods is outlined below:

Step 1: Prepare the data sets with the pre-processing procedure using suitable and transformation method and normalization method (if necessary). The most common transformation in microarray data analysis is using logarithmic base two for all expression of genes:

$$X_{ki}^{*} = \log_2(X_{ki}) \quad \text{..... (1)}$$

Each of the expression level of the gene for each group is normalized which every extreme value are replaced by the winsorize median absolute deviation. The upper limit of extreme value is replaced by:

$$X_{ki}^{+} = \begin{cases} median\ (X_{ki}) + a \cdot MAD & , X_{ki} > median + a \cdot MAD \\ X_{ki} & , otherwise \end{cases} \quad \text{..... (2)}$$

While the lower limit of extreme value is replaced by:

$$X_{ki}^{-} = \begin{cases} median\ (X_{ki}) - a \cdot MAD & , X_{ki} < median - a \cdot MAD \\ X_{ki} & , otherwise \end{cases} \quad \text{..... (3)}$$

Where three is used in this study as the chosen multiplier. The *MAD* is median absolute deviation which is formulated as below:

$$MAD = \underset{i}{median}\left\{ \left| l_i - \underset{j}{median}\{l_j\} \right| \right\} \quad \text{..... (4)}$$

for a univariate data set $l_1, l_2, , l_n$.

Step 2: Compute the shrinkage target according to the proposed approach.

Step 3: Search the optimal shrinkage intensity using the related definition of the proposed method.

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

Step 5: Compute Hotelling's $T^2$for each of all the gene sets that are measured in data sets as explained in:

$$T^2 = \frac{n_1 n_2}{n}(\bar{X}_i - \bar{X}_j)'\left(S_{shrink}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right)^{-1}(\bar{X}_i - \bar{X}_j) \quad \text{..... (5)}$$

Where the mean, $\bar{X}_i$ was defined as:

$$\bar{X}_i = \frac{1}{n}\sum_{k=1}^{n} X_{ki} \quad \text{..... (6)}$$

and the $\bar{X}_j$ is the mean for X_{kj} and S_{shrink} is shrinkage estimator as modelled in:

$$S_{shrink} = \alpha T_{ij} + (1-\alpha) S_{ij} \quad \text{..... (7)}$$

The sample covariance matrix, S_{ij} was defined as:

$$S_{ij} = \frac{1}{n-1}\sum_{k=1}^{n}(X_{ki} - \bar{X}_i)(X_{kj} - \bar{X}_j) \quad \text{..... (8)}$$

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

$$\alpha = \max\left\{0, \min\left\{\frac{\kappa}{n}, 1\right\}\right\} \quad \text{..... (9)}$$

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

$$\kappa = \frac{\pi - \rho}{\gamma} \quad \text{..... (10)}$$

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 Shrink A, Shrink B and Shrink C.

Table 1: The shrinkage combinations for shrink A, Shrink B and Shrink C

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

Step 6: Permute samples for each gene set thus claim the significance of gene sets according to permutation testing. The discussion of permutation testing elaborated in:

$$\hat{p} = \frac{\sum_{i=1}^{M} I(t_i \geq t^*)}{M} \quad \ldots\ldots\ldots\ldots\ldots (11)$$

Where M is the permutation test be used, where $t_i, i = 1,\ldots, M$ is Hotelling's T^2 statistic that compute from the permutation.

A simulation study

For a better interpretation of multivariate structure in the gene set, the correlation matrix was used. The multivariate normal distribution data was generated using *mvrnorm* function in the *MASS* package of *R* language (http://cran. r-project. org/). The generated data was assumed as correlation matrix using *rcorrmatrix* function in the *cluster Generation* package. All programming codes and packages are written in R language (http://cran.r-project.org/).

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 joint density function:

$$fx(x_{1,\ldots\ldots,}x_p) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} e^{-(x-\mu)'\Sigma^{-1}(x-\mu)/2} \quad \ldots\ldots (12)$$

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 constructed according to requirement of cases by changing the four parameters:

- Increasing number of variables of 10 and 30;
- Increasing number of sample sizes of 20;
- The different axis of variation of a major axis of variation (first eigenvector, e_1) or a minor axis of variation ($p/3$ eigenvector, $e_{p/3}$). The two levels a relatively high and a low variance;

Increasing the amount of separation between groups (de_i) of 0.25, 0.50 and 1.00. The three levels represent a low, a moderate and a relatively high separation.

The analysis of the three proposed shrinkage methods was compared with the Regularized Covariance Matrix Approach Testing (RCMAT) [11] and Kong's Principal Component Analysis (KPCA) [12]. This condition was compared the order of magnitude of proposed method based the degree of reduction in the proposed method *p*-values relative to RCMAT and KPCA. For presentation, we illustrate only those conditions where the number of samples is 20. The paired comparisons between two methods were set up as the logarithmic base ten of ratio of the proposed *p*-values to the RCMAT *p*-values or KPCA *p*-values for the same data under several different conditions. In addition, a ratio below 0.0001 is replaced by 0.00009 to avoid the invalid value of logarithmic.

This condition compares the order of magnitude of proposed method based the degree of reduction in the proposed method *p*-values relative to RCMAT and KPCA. We used the paired comparison between two methods using the amount of proportion between the proposed method *p*-value/RCMAT *p*-value and proposed *p*-value/KPCA *p*-value to investigate the power of the proposed approach. Depending on the performance of proposed method that being compared, the graph would shift either to the left or right toward zero. A shift to the left toward zero exhibits the *p*-value of proposed method is smaller or reduced than compared method. In contrast, a shift to the right toward zero shows the *p*-value of proposed method is larger than compared method. Hence, we are looking for the lower amount of proportion or reduction degree to consider the proposed method as a good method or vice versa.

RESULTS AND DISCUSSION

Fig. 1–fig. 6 and fig. 7–fig. 12 display the results of ratios of Shrink A, Shrink B and Shink C respect to RCMAT and KPCA for the cumulative distribution functions with specified amount of separations and axis of variations for 10 variables and 30 variables respectively.

fig. 1 clearly showed that 10 per cent of Shrink C p-values were at least reduced 100 times smaller than the corresponding RCMAT p-value for a separation of 0.25 along major axis compared to other proposed methods. Fig. 2 shows that in relative to KPCA, approximately 20 percent of the Shrink C *p*-values is reduced to 3.16 times of the corresponding KPCA *p*-values for a separation of 0.25 along the minor axis. For 0.5 separations, about 40 percent probability of Shrink B *p*-values being smaller 3.16 times than the corresponding RCMAT *p*-values and being smaller 3.16 times than the corresponding KPCA *p*-values are shown in fig. 3 and fig. 4 respectively. Relative to RCMAT and KPCA, 40 percent of Shrink C *p*-values were reduced by 3.16 and were reduced by 3.16 for 1.0 separations along the major axis and minor as shown in fig. 5 and fig. 6 respectively.

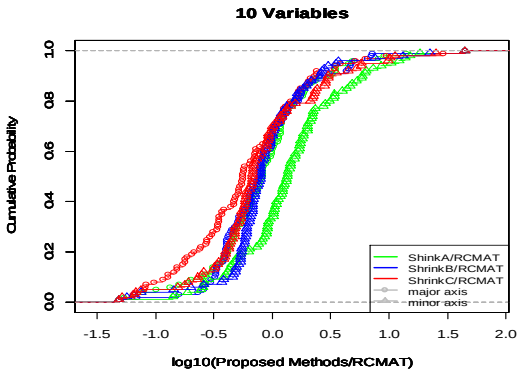

Fig. 1: Cumulative distribution function of ratio of shrink A, shrink B, shrink C and RCMAT p-values with 0.25 separation for 10 variables

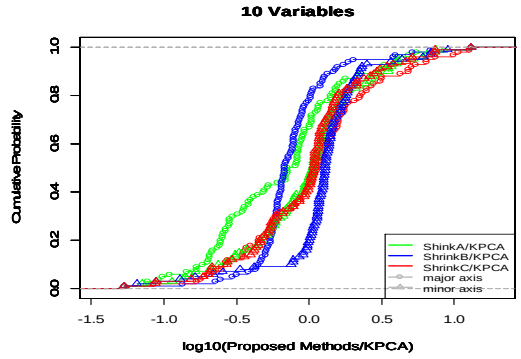

Fig. 2: Cumulative distribution function of ratio of shrink A, shrink B, shrink C and KPCA p-values with 0.25 separation for 10 variables

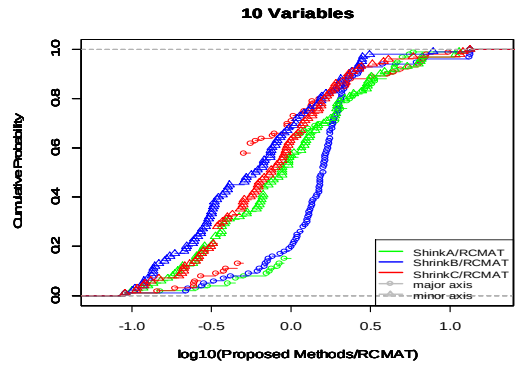

Fig. 3: Cumulative distribution function of ratio of shrink A, shrink B, shrink C and RCMAT KPCA p-values with 0.50 separation for 10 variables

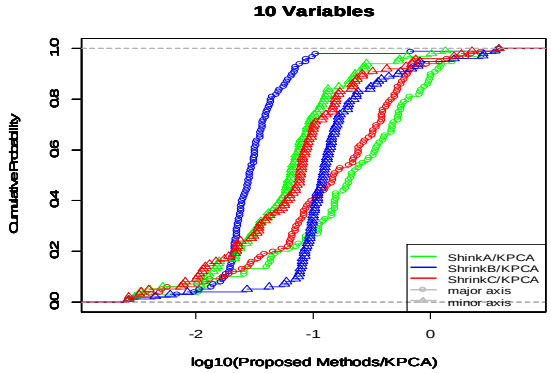

Fig. 4: Cumulative distribution function of ratio of shrink A, shrink B, shrink C and KPCA p-values With 0.50 separation for 10 variables

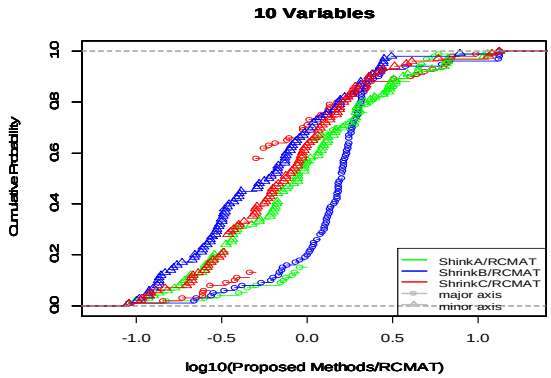

Fig. 5: Cumulative distribution function of ratio of shrink A, shrink B, shrink C and RCMAT p-values with 1.0 separation for 10 variables

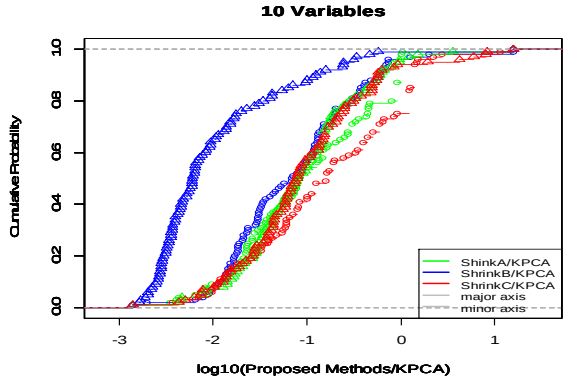

Fig. 6: Cumulative distribution function of ratio of shrink A, shrink B, shrink C and KPCA p-values with 1.0 separation for 10 variables

Fig. 7 and fig. 8 show approximately 20 percent of the Shrink B *p*-values that were at least reduced 3.16 times smaller than the corresponding RCMAT *p*-values and KPCA *p*-values for a separation of 0.25 along major axis respectively.

Fig. 9 showed that approximately 20 percent of the Shrink B *p*-values being smaller 3.16 times than the corresponding RCMAT *p*-values for 0.5 separations along the minor axis. About 20 percent probability of all proposed methods being smaller than 3.16 times than the corresponding KPCA *p*-values for a separation of 0.5 along minor axis was shown in fig. 10. The 40 percent of Shrink A *p*-values and Shrink C *p*-values were about 100 times smaller than the corresponding RCMAT *p*-values and KPCA *p*-values for a separation of one along the minor axis as depicted in fig. 11 and fig. 12 respectively.

Fig. 1–fig. 12 were captured the comparability of the two methods using cumulative distribution function plots. We can see the relative power or reduction degree between all proposed methods and RCMAT or KPCA is higher although when the number of variables is larger than the number of samples thus the results suggest the good performance of all proposed methods. However, the different conditions exhibit different results and most the reduction degree between all proposed methods and KPCA is higher compared to RCMAT.

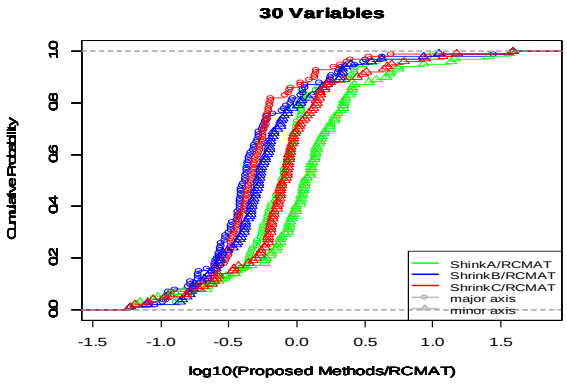

Fig. 7: Cumulative distribution function of ratio of shrink A, Shrink B, Shrink C and RCMAT p-values with 0.25 separation for 30 variables

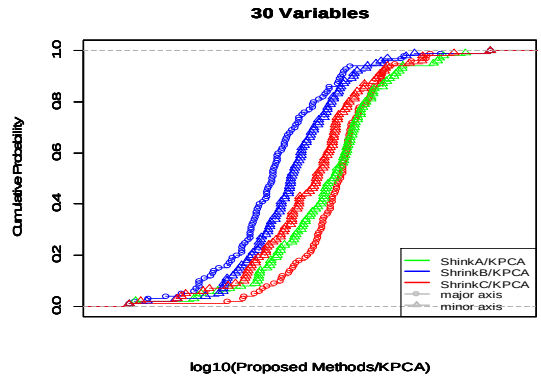

Fig. 8: Cumulative distribution function of ratio of shrink A, shrink B, shrink C and KPCA p-values with 0.25 Separation for 30 variables

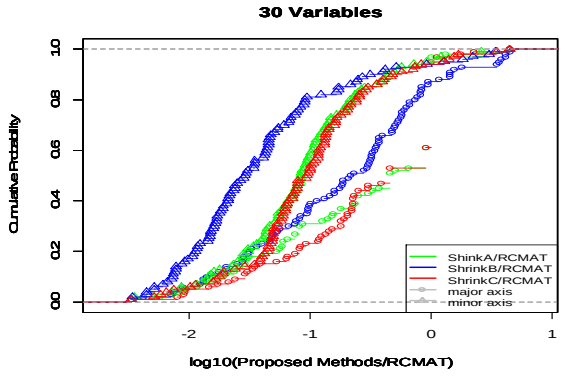

Fig. 9: Cumulative distribution function of ratio of shrink A, shrink B, shrink C and RCMAT p-values With 0.50 separation for 30 variables

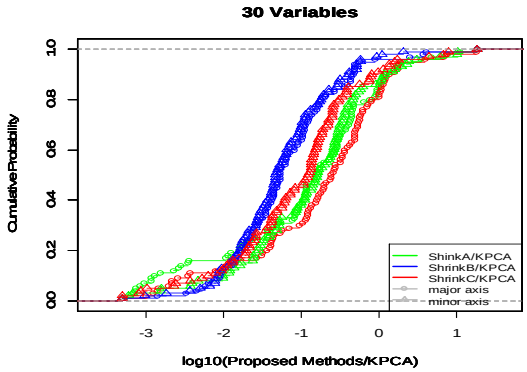

Fig. 10: Cumulative distribution function of ratio of shrink A, shrink B, shrink C and KPCA p-values with 0.50 separation for 30 variables

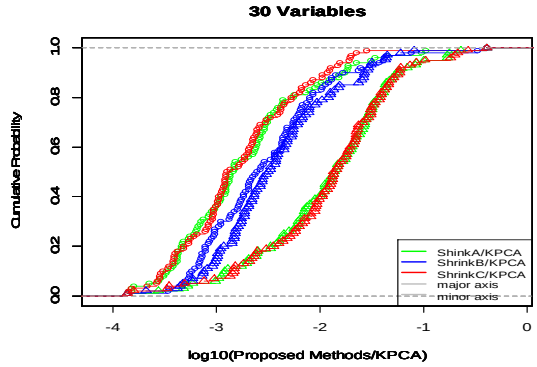

Fig. 11: Cumulative distribution function of ratio of shrink A, shrink B, shrink C and RCMAT p-values with 1.0 separation for 30 variables

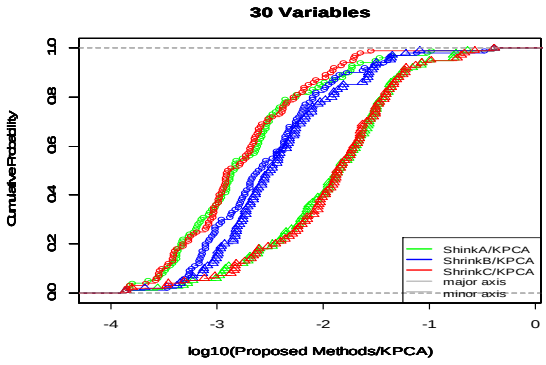

Fig. 12: Cumulative distribution function of ratio of shrink A, shrink B, shrink C and KPCA p-values with 1.0 separation for 30 variables

CONCLUSION

The multivariate test statistic of classical Hotelling's T^2 is used to integrate the correlation when assessing changes in activity level across biological conditions. In the other word, the shrinkage covariance matrix incorporating correlation between genes for detecting significant gene sets. The motivation to this idea is the correlation is an important measure of the relationship among genes and the relatedness is regarded.

We tested the ability of our newly proposed methods on simulated data sets. Based on the given conditions, the examples from simulated data sets showed that the newly proposed methods performed better than other methods. The results from simulation studies indicated that the Shrink A, Shrink B and Shrink C always outperformed the other methods in most conditions. This study also found that Shrink A gave the best results followed by Shrink C and Shrink B methods in most conditions.

According to our methodology, only two groups or experimental conditions (such as normal and treatment groups) with an equally number of group size is the suitable characteristics. However, it is also recommended to integrate our proposed study into more than two groups.

ACKNOWLEDGMENT

Special thanks also go to Universiti Teknologi MARA for supporting this research under the Research Grant d No. 600-RMI/DANA 5/3/CIFI (65/2013).

CONFLICT OF INTERESTS

Declared none

REFERENCES

1. Wang Z, Zineddin B, Liang J, Zeng N, Li Y, Du M, *et al*. cDNA microarray adaptive segmentation. Neurocomputing 2014;142:408-18.
2. Zvara Á, Kitajka K, Faragó N, Puskás LG. Microarray technology. Acta Biological Szegediensis 2015;59:51-67.
3. Cooper-Knock J, Kirby J, Ferraiuolo L, Heath PR, Rattray M, Shaw PJ. Gene expression profiling in human neurodegenerative disease. Nat Rev Neurol 2012;8:518-30.
4. Altenburger R, Scholz S, Schmitt-Jansen M, Busch W, Escher BI. Mixture toxicity revisited from a toxicogenomic perspective. Environ Sci Technol 2012;46:2508-22.
5. Tran B, Dancey JE, Kamel-Reid S, McPherson JD, Bedard PL, Brown AM, *et al*. Cancer genomics: technology, discovery, and translation. J Clin Oncol 2012;30:647-60.
6. Karjanto S, Ramli NM, Aripin R, Ghani NAM. Improved statistical test using shrinkage covariance matrix for identifying differential gene sets. J Appl Environ Biol Sci 2014;1:302-10.
7. Ledoit O, Wolf M. A well-conditioned estimator for large-dimensional covariance matrices. J Multivariate Anal 2001;88:365–411.
8. Ledoit O, Wolf M. Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. J Empirical Finance 2003;10:603-21.
9. Ledoit O, Wolf M. Honey, I shrunk the sample covariance matrix. J Portfolio Management 2004;31:110-9.
10. Schäfer J, Strimmer K. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. Stat Appl Genet Mol Biol 2005;4:32.
11. Yates PD, Reimers MA. RCMAT: a regularized covariance matrix approach to testing gene sets. BMC Bioinf 2009;10:300.
12. Kong SW, Pu WT, Park PJ. A multivariate approach for integrating genome-wide expression data and biological knowledge. Bioinformatics 2006;22:2373-80.

How to cite this article

- Suryaefiza Karjanto, Norazan Mohamed Ramli, Nor Azura MD Ghaninor Azura MD Ghani. Improved estimation of covariance matrix in hotelling's t^2 for microarray data. Int J Pharm Pharm Sci 2016;8 Suppl 2:26-30.