



UNIVERSITI
TEKNOLOGI
MARA

Institut
Pengajian
Siswazah

THE DOCTORAL RESEARCH ABSTRACTS

TWELFTH
ISSUE

Volume: 12, Issue 12

October 2017

IGS Biannual Publication



Name : SURYAEFIZA KARJANTO

Title : SHRINKAGE ESTIMATION OF COVARIANCE MATRIX IN HOTELLING'S T^2 FOR DIFFERENTIALLY EXPRESSED GENE SETS

Supervisor : DR. NORAZAN MOHAMED RAMLI (MS)
ASSOC. PROF. DR. NOR AZURA MD GHANI (CS)

The microarray technology performs simultaneous analysis of thousands of genes in a massively parallel manner in one experiment, hence providing valuable knowledge on gene interaction and function. The understanding of microarray data has led to the development of new methods in statistics such as detection of differentially expressed genes. The microarray analysis was first employed for individual or single gene, but recently it has been applied to a gene set or a group of the gene. The relationship between genes in gene set is analysed using Hotelling's T^2 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 T^2 particularly to cater high dimensionality problem in microarray data. The Hotelling's T^2 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 ShrinkA, ShrinkB and ShrinkC. The ShrinkA is the simplest approach with both the non-diagonal element of shrinkage target and the sum of asymptotic covariances of the entries of the shrinkage

target are assumed as zero. The shrinkage target of ShrinkB is assumed as the square root of the multiplication of variance of two groups for non-diagonal element and has a same element on diagonal with other approaches. The ShrinkC approach is quite similar to ShrinkB except with the addition of average sample correlation. 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 three conditions of simulated data sets. Firstly, the performance of the proposed methods were assessed when no difference (separation) exists between two groups (null hypotheses), followed by the second one which measures the performance when there is a difference (separation) between groups (alternative hypotheses) and finally, for paired comparison. Method validation was also done using real microarray data sets such as diabetes and leukemia. In many conditions whether simulation or real data studies the ShrinkA method performed slightly better than the ShrinkC and RCMAT methods. In contrast, both the ShrinkB and KPCA performed relatively poorly in this analysis. The robust trimmed mean is integrated into the shrinkage matrix to reduce the influence of outliers and consequently increases its efficiency. The performance of the proposed method is measured using several simulation designs. The results are expected to outperform existing techniques in many tested conditions tested. 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.