UNIVERSITI TEKNOLOGI MARA

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

SURYAEFIZA KARJANTO

Thesis submitted in fulfillment
of the requirements for the degree of
Doctor of Philosophy

Faculty of Computer and Mathematical Sciences

February 2017

ABSTRACT

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.

ACKNOWLEDGEMENT

My utmost gratitude to Allah SWT for His guidance and give me the strength, courage and assistance throughout my life especially during the difficult times.

The printed pages of this thesis hold far more than the conclusion of years of study. These pages also reflect the relationships with many generous and inspiring people I have met since beginning my Ph.D. work.

To my main supervisor Dr Norazan Mohamed Ramli, always a mentor who encourages me by words, thoughtful criticism and attention.

To my ex-main supervisor Dr Rasimah Aripin for showing me by a good example and shares her enthusiasm on my early years of study.

To my co-supervisor Associated Prof. Dr Nor Azura Md Ghani for giving precious comments on my work.

To my parent and for being patient and great support, and my siblings Juffrizal, Nor Suhaida, Nur Azura, Iswandi, Rafidah, Nor Amalina Nabila, Adham, Aliff Irfan and Sara for being best family.

To all my colleagues from Department of Computer and Mathematical Sciences, Universiti Teknologi MARA (Pulau Pinang) and Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA (Melaka) Kampus Jasin for understand me to complete my doctoral programme.

TABLE OF CONTENTS

CHAPTER ONE
INTRODUCTION

1.1 OVERVIEW

The first part of this chapter provides a basic overview of microarray technology. In the Section 1.3 some research motivation to this research is described followed by an explanation of the problem statement in Section 1.4. The objective of this study will be presented on Section 1.5. The detail explanation of the significance of this study is also included in Section 1.6. Then description of scope and limitations of the study described in Section 1.7. Finally, the descriptions how this thesis will be organized are in the last sections.

1.2 RESEARCH BACKGROUND

Some fundamental theories and molecular biology concepts are introduced in this section. In the following parts, microarray technologies that generate the data sets are then described. Some commonly used microarray data analysis also presented.

1.2.1 Microarray Technology

Microarray technology is one of the significant achievements in biotechnology history, 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 (Schena *et al.*, 1996).

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. But, in a precise definition, microarray involves placing an orderly arrangement of thousands of gene sequences in a grid on a suitable surface usually a glass slide. Each gene sequences on the glass represents as a single gene and a sample containing deoxyribose nucleic acid (DNA) or ribonucleic acid (RNA) is placed in contact with the gene chip. Generally, a single microarray slide may contain thousands