

## DOCTORAL THESIS

### High-dimensional covariance matrix estimation with application to Hotelling's tests

Dong, Kai

*Date of Award:*  
2015

[Link to publication](#)

#### General rights

Copyright and intellectual property rights for the publications made accessible in HKBU Scholars are retained by the authors and/or other copyright owners. In addition to the restrictions prescribed by the Copyright Ordinance of Hong Kong, all users and readers must also observe the following terms of use:

- Users may download and print one copy of any publication from HKBU Scholars for the purpose of private study or research
- Users cannot further distribute the material or use it for any profit-making activity or commercial gain
- To share publications in HKBU Scholars with others, users are welcome to freely distribute the permanent URL assigned to the publication

# ABSTRACT

In recent years, high-dimensional data sets are widely available in many scientific areas, such as gene expression study, finance and others. Estimating the covariance matrix is a significant issue in such high-dimensional data analysis. This thesis focuses on high-dimensional covariance matrix estimation and its application.

First, this thesis focuses on the covariance matrix estimation. In Chapter 2, a new optimal shrinkage estimation of the covariance matrices is proposed. This method is motivated by the quadratic discriminant analysis where many covariance matrices need to be estimated simultaneously. We shrink the sample covariance matrix towards the pooled sample covariance matrix through a shrinkage parameter. Some properties of the optimal shrinkage parameter are investigated and we also provide how to estimate the optimal shrinkage parameter. Simulation studies and real data analysis are also conducted. In Chapter 4, we estimate the determinant of the covariance matrix using some recent proposals for estimating high-dimensional covariance matrix. Specifically, a total of nine covariance matrix estimation methods will be considered for comparison. Through extensive simulation studies, we explore and summarize some interesting comparison results among all compared methods. A few practical guidelines are also made on the sample size, the dimension, and the correlation of the data set for estimating the determinant of high-dimensional covariance matrix. Finally, from a perspective of the loss function, the comparison study in this chapter also serves as a proxy to assess the performance of the covariance matrix estimation.

Second, this thesis focuses on the application of high-dimensional covariance matrix estimation. In Chapter 3, we consider to estimate the high-dimensional covariance matrix based on the diagonal matrix of the sample covariance matrix and apply it to the Hotelling's tests. In this chapter, we propose a shrinkage-based diagonal Hotelling's test for both one-sample and two-sample cases. We also propose several different ways to derive the approximate null distribution under different scenarios of  $p$  and  $n$  for our proposed shrinkage-based test. Simulation studies show that the

proposed method performs comparably to existing competitors when  $n$  is moderate or large, and it is better when  $n$  is small. In addition, we analyze four gene expression data sets and they demonstrate the advantage of our proposed shrinkage-based diagonal Hotelling's test.

Apart from the covariance matrix estimation, we also develop a new classification method for a specific type of high-dimensional data, RNA-sequencing data. In Chapter 5, we propose a negative binomial linear discriminant analysis for RNA-Seq data. By Bayes' rule, we construct the classifier by fitting a negative binomial model, and propose some plug-in rules to estimate the unknown parameters in the classifier. The relationship between the negative binomial classifier and the Poisson classifier is explored, with a numerical investigation of the impact of dispersion on the discriminant score. Simulation results show the superiority of our proposed method. We also analyze four real RNA-Seq data sets to demonstrate the advantage of our method in real-world applications.

**Keywords:** Covariance matrix, Discriminant analysis, High-dimensional data, Hotelling's test, Log determinant, RNA-sequencing data

# Contents

<b>Declaration</b>	<b>i</b>
<b>ABSTRACT</b>	<b>ii</b>
<b>ACKNOWLEDGEMENTS</b>	<b>iv</b>
<b>Table of Contents</b>	<b>v</b>
<b>List of Tables</b>	<b>ix</b>
<b>List of Figures</b>	<b>x</b>
<b>Chapter 1 Introduction</b>	<b>1</b>
1.1 High-dimensional Covariance Matrix Estimation . . . . .	1
1.2 Hotelling's Tests . . . . .	2
1.3 Discriminant Analysis . . . . .	4
1.4 Overall Structure . . . . .	4
<b>Chapter 2 Optimal Shrinkage Estimation of the Covariance Matrices</b>	<b>6</b>
2.1 Introduction . . . . .	6
2.2 Shrinkage Covariance Matrix Estimators . . . . .	8
2.3 Optimal Shrinkage Parameters Estimation . . . . .	8
2.3.1 Optimal Estimator Under the Loss Function $L_1$ . . . . .	9
2.3.2 Optimal Estimator Under the Loss Function $L_2$ . . . . .	12
2.4 Simulation Studies . . . . .	14
2.4.1 Simulation Design . . . . .	14

2.4.2	Simulation Results . . . . .	14
2.5	Real Data Analysis . . . . .	17
2.6	Proofs . . . . .	20
2.6.1	Proof of Theorem 1 . . . . .	24
2.6.2	Proof of Theorem 2 . . . . .	25
2.6.3	Proof of Theorem 3 . . . . .	26
2.6.4	Proof of Theorem 4 . . . . .	27
2.6.5	Proof of Theorem 5 . . . . .	33
2.6.6	Proof of Theorem 6 . . . . .	34
2.6.7	Proofs of (2.11) and (2.12) . . . . .	35
2.6.8	Proof of (2.14) . . . . .	36
2.6.9	Proof of Theorem 7 . . . . .	37
2.6.10	Proof of Theorem 8 . . . . .	38

**Chapter 3 Shrinkage-Based Diagonal Hotelling’s Tests for High-Dimensional  
Small Sample Size Data 40**

3.1	Introduction . . . . .	40
3.2	Improving the Diagonal Hotelling’s Tests . . . . .	44
3.2.1	The Diagonal Hotelling’s Tests . . . . .	45
3.2.2	Shrinkage-Based Diagonal Hotelling’s Tests . . . . .	46
3.3	Null Distributions of Shrinkage-Based Diagonal Hotelling’s Tests for Small Sample Size . . . . .	48
3.3.1	Chi-squared Approximation . . . . .	49
3.3.2	Normal Approximation . . . . .	51
3.4	Monte Carlo Simulation Studies . . . . .	53
3.4.1	Simulation Design . . . . .	53
3.4.2	Simulation Results . . . . .	54
3.5	Case Studies . . . . .	59
3.6	Discussion . . . . .	63
3.7	Proofs . . . . .	64

3.7.1	Proof of Lemma 6 . . . . .	64
3.7.2	Derivation of formula (3.10) . . . . .	65

**Chapter 4 A Comparison of Methods for Estimating the Determinant  
of High-Dimensional Covariance Matrix 67**

4.1	Introduction . . . . .	67
4.2	Methods for Estimating $\theta$ . . . . .	70
4.2.1	Diagonal Estimation . . . . .	70
4.2.2	Shrinkage Estimation . . . . .	72
4.2.3	Sparse Estimation . . . . .	74
4.2.4	Factor Model Estimation . . . . .	76
4.3	Simulation Studies . . . . .	77
4.3.1	Setup I . . . . .	77
4.3.2	Setup II . . . . .	78
4.3.3	Setup III . . . . .	79
4.4	Conclusion and Discussion . . . . .	80
4.5	Proofs . . . . .	82
4.5.1	Proof of Theorem 13 . . . . .	82

**Chapter 5 NBLDA: Negative Binomial Linear Discriminant Analysis  
for RNA-Seq Data 90**

5.1	Introduction . . . . .	90
5.2	Negative Binomial Linear Discriminant Analysis . . . . .	93
5.2.1	Methodology . . . . .	93
5.2.2	Parameter Estimation . . . . .	97
5.3	Simulation Studies . . . . .	98
5.3.1	Simulation Design . . . . .	99
5.3.2	Simulation Results . . . . .	103
5.4	Real Data Analysis . . . . .	103
5.4.1	Gene Selection . . . . .	104
5.4.2	Results . . . . .	106

5.5 Discussion . . . . .	107
<b>Chapter 6 Summary</b>	<b>109</b>
<b>Bibliography</b>	<b>111</b>
<b>Curriculum Vitae</b>	<b>123</b>