

## DOCTORAL THESIS

### New developments in multiple testing and multivariate testing for high-dimensional data

Hu, Zongliang

*Date of Award:*  
2018

[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

This thesis aims to develop some new and novel methods in advancing multivariate testing and multiple testing for high-dimensional small sample size data.

In Chapter 2, we propose a likelihood ratio test framework for testing normal mean vectors in high-dimensional data under two common scenarios: the one-sample test and the two-sample test with equal covariance matrices. We derive the test statistics under the assumption that the covariance matrices follow a diagonal matrix structure. In comparison with the diagonal Hotelling's tests, our proposed test statistics display some interesting characteristics. In particular, they are a summation of the log-transformed squared  $t$ -statistics rather than a direct summation of those components. More importantly, to derive the asymptotic normality of our test statistics under the null and local alternative hypotheses, we do not need the requirement that the covariance matrices follow a diagonal matrix structure. As a consequence, our proposed test methods are very flexible and readily applicable in practice. Monte Carlo simulations and a real data analysis are also carried out to demonstrate the advantages of the proposed methods.

In Chapter 3, we propose a pairwise Hotelling's method for testing high-dimensional mean vectors. The new test statistics make a compromise on whether using all the correlations or completely abandoning them. To achieve the goal, we perform a screening procedure, pick up the paired covariates with strong correlations, and construct a classical Hotelling's statistic for each pair. While for the individual covariates without strong correlations with others, we apply squared  $t$  statistics to account for their respective contributions to the multivariate testing problem. As a consequence, our proposed test statistics involve a combination of the collected pairwise Hotelling's test statistics and squared  $t$  statistics. The asymptotic normality of our test statistics under the null and local alternative hypotheses are also derived under some regularity conditions. Numerical studies and two real data examples demonstrate the efficacy of our pairwise Hotelling's test.

In Chapter 4, we propose a regularized  $t$  distribution and also explore its applications in multiple testing. The motivation of this topic dates back to microarray

studies, where the expression levels of thousands of genes are measured simultaneously by the microarray technology. To identify genes that are differentially expressed between two or more groups, one needs to conduct hypothesis test for each gene. However, as microarray experiments are often with a small number of replicates, Student's  $t$ -tests using the sample means and standard deviations may suffer a low power for detecting differentially expressed genes. To overcome this problem, we first propose a regularized  $t$  distribution and derive its statistical properties including the probability density function and the moments. The noncentral regularized  $t$  distribution is also introduced for the power analysis. To demonstrate the usefulness of the proposed test, we apply the regularized  $t$  distribution to the gene expression detection problem. Simulation studies and two real data examples show that the regularized  $t$ -test outperforms the existing tests including Student's  $t$ -test and the Bayesian  $t$ -test in a wide range of settings, in particular when the sample size is small.

**Keywords:** High-dimensional data, Hotelling's test, Likelihood ratio test, Log-transformed squared  $t$ -statistic, Noncentral regularized  $t$  distribution, Paired correlations, Power, Regularized  $t$  distribution, Screening, Type I error

# Table of Contents

Declaration	i
Abstract	ii
Acknowledgements	iv
Table of Contents	v
List of Figures	viii
List of Tables	x
Chapter 1 Introduction	1
1.1 Hypothesis Testing in High-Dimensional Data . . . . .	1
1.2 Student's $t$ -Test and Variants . . . . .	4
1.3 Hotelling's $T^2$ Test and Variants . . . . .	6
1.4 High-Dimensional Tests . . . . .	8
1.5 Outline of the Thesis . . . . .	11
Chapter 2 Diagonal Likelihood Ratio Test for Equality of Mean Vectors in High-Dimensional Data	15
2.1 Introduction . . . . .	15
2.2 One-Sample Test . . . . .	18
2.2.1 Diagonal LRT Statistic . . . . .	18
2.2.2 Null Distribution . . . . .	19
2.2.3 Statistical Power . . . . .	22
2.3 Two-Sample Test . . . . .	23

2.4	Monte Carlo Simulation Studies . . . . .	26
2.4.1	Normal Data . . . . .	26
2.4.2	Heavy-Tailed Data . . . . .	30
2.5	Brain Cancer Data . . . . .	32
2.6	Conclusion . . . . .	34
2.7	Technical Details . . . . .	37
2.7.1	Derivation of the One-Sample DLRT Statistic . . . . .	37
2.7.2	Proof of Lemma 2.2.1 . . . . .	38
2.7.3	Proof of Theorem 2.2.1 . . . . .	39
2.7.4	Proof of Corollary 2.2.1 . . . . .	39
2.7.5	Proof of Theorem 2.2.2 . . . . .	40
2.7.6	Derivation of the Two-Sample DLRT Statistic . . . . .	41
2.7.7	Proof of Theorem 2.3.1 . . . . .	42
2.7.8	Proof of Corollary 2.3.1 . . . . .	43
2.7.9	Proof of Theorem 2.3.2 . . . . .	44
2.8	Additional Simulations and Properties . . . . .	45
Chapter 3	Pairwise Hotelling’s Test for Equality of Means in High-Dimensional Data	54
3.1	Introduction . . . . .	54
3.2	One-Sample Test . . . . .	57
3.2.1	Pairwise Hotelling’s Statistic . . . . .	57
3.2.2	Asymptotic Results . . . . .	60
3.2.3	Local Power Analysis . . . . .	62
3.3	Two-Sample Test . . . . .	64
3.3.1	Notation . . . . .	64
3.3.2	Asymptotic Results . . . . .	65
3.4	Monte Carlo Simulation Studies . . . . .	68
3.5	Applications . . . . .	69
3.5.1	Small Round Blue Cell Tumors Data . . . . .	69
3.5.2	Leukemia Data . . . . .	69
3.6	Conclusion . . . . .	72

3.7	Technical Details . . . . .	73
3.7.1	Proof of Theorem 3.2.1 . . . . .	77
3.7.2	Proof of Theorem 3.2.2 . . . . .	78
3.7.3	Proof of Lemma 3.2.1 . . . . .	102
3.7.4	Proof of Theorem 3.3.1 . . . . .	108
3.7.5	Proof of Theorem 3.3.2 . . . . .	109
3.7.6	Proof of Lemma 3.3.1 . . . . .	112
3.8	Additional Simulations . . . . .	117
Chapter 4 Regularized $t$ Distribution: Properties and Applications		118
4.1	Introduction . . . . .	118
4.2	Regularized $t$ Distribution . . . . .	122
4.2.1	Definition . . . . .	124
4.2.2	Probability Density Function . . . . .	126
4.2.3	Moments and Moment Generating Function . . . . .	127
4.3	Noncentral Regularized $t$ Distribution . . . . .	132
4.4	Applications of Regularized $t$ Distribution . . . . .	137
4.5	Two Real Data Examples . . . . .	144
4.6	Conclusion . . . . .	146
4.7	Technical Details . . . . .	147
4.7.1	Proofs of Lemmas . . . . .	147
4.7.2	Derivation of Probability Density Function . . . . .	149
4.7.3	Symmetries and Derivatives of Probability Density Function . . . . .	150
4.7.4	Moments of Regularized $t$ Distribution . . . . .	154
4.7.5	Additional Simulations for Selecting Parameters . . . . .	155
Chapter 5 Future Work		158
Curriculum Vitae		171