

Doctoral Dissertation

**Clustering of strata means based on  
pairwise L1 regularized empirical  
likelihood**

Department of Statistics

Graduate School of Chonnam National University

Nong Quynh Van

August 2019

Doctoral Dissertation

# **Clustering of strata means based on pairwise L1 regularized empirical likelihood**

Department of Statistics

Graduate School of Chonnam National University

Nong Quynh Van

Directed by Professor Chi Tim Ng

Department of Statistics, Chonnam National University

August 2019

# Clustering of strata means based on pairwise L1 regularized empirical likelihood

Department of Statistics

Graduate School of Chonnam National University

Nong Quynh Van

The dissertation entitled above, by the graduate student Nong Quynh Van, in partial fulfillment of the requirements for the **Doctor of Philosophy in Statistics** has been deemed by the Professors below.

Jang Sun Baek, Ph.D .....

Jae Sik Jeong, Ph.D .....

Myung Hwan Na, Ph.D .....

Chi Tim Ng, Ph.D .....

Woo Joo Lee, Ph.D .....

August 2019

Dedicated to

My Parents, My Younger Brother and My Family

# Acknowledgements

First and foremost, I am cordially thankful for my supervisor, Dr. Chi Tim Ng who offered the opportunity to study PhD degree for me. Without his support, guidance, knowledge and patience, this research would not have been possible. I dare to say that he is a good PhD advisor.

I wish to express my gratitude to the Vietnamese Ministry of Education and Training.

I would like to thank all students and staff of the Department of Statistics of Chonnam National University for their friendly and good social environment. I also would like to thank members of the Mathematical Statistics Research Laboratory, especially Nguyen Van Cuong, Zhang Lili, Zhang Kaimeng, 박진경, 김태경 for their great help and friendship.

I would like to dedicate this thesis to my parents and my brother who have given me the eternal love and have encouraged me to pursue the long academic journey. Without their sacrifices, I would not have finished my studies and my thesis.

Finally, I would like to take this opportunity to thank my husband for his love and unconditional support. The last word goes to my lovely children, Ngo Bao Chau and Ngo Duc Thanh who have brought me the true meaning of love and have given me the strength to get all things done.

# Contents

|   |            |
|---|------------|
| <b>List of Tables</b>   | <b>v</b>   |
| <b>List of Figures</b>  | <b>vi</b>  |
| <b>Abstract</b>   | <b>vii</b> |
| <b>1 Introduction and Previous Work</b>   | <b>1</b>   |
| 1.1 Introduction . . . . .  | 1          |
| 1.2 Outline of the dissertation . . . . .   | 3          |
| <b>2 Strata Mean Clustering via Regularized Empirical Likelihood</b>                                | <b>5</b>   |
| 2.1 Introduction . . . . .  | 5          |
| 2.2 L1 Regularized Empirical Likelihood Estimation . . . . .  | 6          |
| 2.3 Familywise Error Rate and Bayesian Information Criterion . . . . .                              | 7          |
| 2.4 Algorithm . . . . .   | 9          |
| 2.4.1 One-population $m$ -strata Case . . . . .   | 9          |
| 2.4.2 Two-population $m$ -strata Case . . . . .   | 11         |
| 2.5 Consistency Theory . . . . .  | 12         |
| 2.5.1 Main Theorems . . . . .   | 12         |
| 2.5.2 Proofs of Main Theorems . . . . .   | 14         |
| 2.5.3 Technical Lemmas . . . . .  | 19         |
| 2.6 Simulation studies . . . . .  | 22         |
| 2.7 Real Data examples . . . . .  | 32         |
| 2.7.1 Example 1: Chronic Myelogenous Leukemia Survival Data . . . . .                               | 32         |
| 2.7.2 Example 2: Investigating Structural Change and Monday Effect<br>in the Stock Market . . . . . | 33         |

|          |   |           |
|----------|---|-----------|
| 2.7.3    | Example 3: Microarray Data of Breast Cancer Patients . . . . .        | 34        |
| 2.8      | Discussion . . . . .  | 39        |
| <b>3</b> | <b>Deriving hypotheses testing via penalized empirical likelihood</b> | <b>40</b> |
| 3.1      | Introduction . . . . .  | 40        |
| 3.2      | One-sample mean test with empirical likelihood . . . . .              | 41        |
| 3.3      | Two samples mean test with empirical likelihood . . . . .             | 43        |
| 3.4      | Simulation studies . . . . .  | 45        |
| 3.4.1    | One-sample Mean Tests based on Empirical Likelihood . . . . .         | 45        |
| 3.4.2    | Two-sample Mean Tests based on Empirical Likelihood . . . . .         | 45        |
| 3.5      | Real data examples . . . . .  | 48        |
| 3.6      | Discussion . . . . .  | 49        |
| <b>4</b> | <b>Discussion and Conclusion</b>                                      | <b>51</b> |
|          | <b>References</b>   | <b>53</b> |
|          | <b>초록</b>   | <b>58</b> |
|          | <b>Appendix</b>   | <b>59</b> |
|          | <b>Appendix A Calculation of Derivatives in section 3.2</b>           | <b>59</b> |

# List of Tables

|      |   |    |
|------|---|----|
| 2.1  | Mis-classification Rate for Chisquare distribution. . . . .   | 25 |
| 2.2  | Mis-classification Rate for Gamma distribution with $v = 1$ . . . . .   | 26 |
| 2.3  | Compare the influence of variance of Gamma distribution to the performance. . . . .   | 27 |
| 2.4  | The influence of variance of Gamma distribution in small distance between cluster's means case. . . . .   | 27 |
| 2.5  | Cumulative propotion of number groups in k-th cluster for m=40. . . . .   | 28 |
| 2.6  | Cumulative propotion of number groups in k-th cluster for m=200. . . . .  | 29 |
| 2.7  | Mis-classification Rate for Chi-square distributed Unbalanced Data. . . . .   | 30 |
| 2.8  | Mis-classification Rate for Example 2-Gamma with fixed variance . . . . .   | 31 |
| 2.9  | Detected cluster treatments . . . . .   | 36 |
| 2.10 | Detected cluster absolute returns of years . . . . .  | 36 |
| 2.11 | Detected cluster genes . . . . .  | 36 |
| 3.1  | Powers of one-sample mean penalized empirical likelihood test and empirical likelihood ratio test. PLT for Penalized Likelihood Test. ELRT for Empirical Likelihood Ratio Test. . . . .               | 46 |
| 3.2  | Powers of two-sample means test via penalized empirical likelihood and empirical likelihood ratio test. PLT for Penalized Likelihood Test. ELRT for Empirical Likelihood Ratio Test. . . . .          | 47 |
| 3.3  | Statistic values and Critical values of penalized empirical likelihood test and empirical likelihood ratio test. PLT for Penalized Likelihood Test. ELRT for Empirical Likelihood Ratio Test. . . . . | 49 |



# List of Figures

|     |   |    |
|-----|---|----|
| 2.1 | Box plot for comparison average of absolute returns between clusters. . . | 36 |
| 2.2 | Data arranged using Heatmap. . . . .                                      | 37 |
| 2.3 | Box plot for comparison average of means between clusters. . . . .        | 38 |

# Abstract

To determine the pairwise equalities-in-mean of a vast amount of subsamples with unknown distributions, a clustering approach is developed based on  $L_1$  regularized empirical likelihood. Under the clustering approach, all possible contradictory conclusions are ruled out automatically. On the contrary, the decision rules based on many existing pairwise comparison procedures can generate contradictory results. Moreover, under certain mild conditions, the proposed clustering method enjoys the consistency property that with probability going to one, the equalities-in-mean of all pairs of subsamples can be determined correctly. An exterior point algorithm is presented for the clustering. The applications of the proposed methods are demonstrated using stock market data and microarray data of breast cancer patients.