University of Kentucky
UKnowledge

Theses and Dissertations--Statistics

Statistics

2019

COMPOSITE NONPARAMETRIC TESTS IN HIGH DIMENSION

Alejandro G. Villasante Tezanos
University of Kentucky, agvi222@uky.edu
Author ORCID Identifier:
https://orcid.org/0000-0001-5108-8637

Digital Object Identifier: https://doi.org/10.13023/etd.2019.339

Recommended Citation

Villasante Tezanos, Alejandro G., "COMPOSITE NONPARAMETRIC TESTS IN HIGH DIMENSION" (2019). *Theses and Dissertations--Statistics*. 42.
https://uknowledge.uky.edu/statistics_etds/42

Alejandro G. Villasante Tezanos, Student

Dr. Solomon W. Harrar, Major Professor

Dr. Katherine L. Thompson, Director of Graduate Studies

COMPOSITE NONPARAMETRIC TESTS IN HIGH DIMENSION

DISSERTATION

A dissertation submitted in partial
fulfillment of the requirements for
the degree of Doctor of Philosophy
in the College of Arts and Sciences
at the University of Kentucky

By

Alejandro G. Villasante Tezanos
Lexington, Kentucky

Director: Dr. Solomon W. Harrar, Professor of Statistics
Lexington, Kentucky
2019

ABSTRACT OF DISSERTATION

COMPOSITE NONPARAMETRIC TESTS IN HIGH DIMENSION

This dissertation focuses on the problem of making high-dimensional inference for two or more groups. High-dimensional means both the sample size (n) and dimension (p) tend to infinity, possibly at different rates. Classical approaches for group comparisons fail in the high-dimensional situation, in the sense that they have incorrect sizes and low powers. Much has been done in recent years to overcome these problems. However, these recent works make restrictive assumptions in terms of the number of treatments to be compared and/or the distribution of the data. This research aims to (1) propose and investigate refined small-sample approaches for high-dimension data in the multi-group setting (2) propose and study a fully-nonparametric approach, and (3) conduct an extensive comparison of the proposed methods with some existing ones in a simulation.

When treatment effects can meaningfully be formulated in terms of means, a semiparametric approach under equal and unequal covariance assumptions is investigated. Composites of F-type statistics are used to construct two tests. One test is a moderate-p version – the test statistic is centered by asymptotic mean – and the other test is a large-p version asymptotic-expansion based finite-sample correction for the mean of the test statistic. These tests do not make any distributional assumptions and, therefore, they are nonparametric in a way. The theory for the tests only requires mild assumptions to regulate the dependence. Simulation results show that, for moderately small samples, the large-p version yields substantial gain in the size with a small power tradeoff.

In some situations mean-based inference is not appropriate, for example, for data that is in ordinal scale or heavy tailed. For these situations, a high-dimensional fully-nonparametric test is proposed. In the two-sample situation, a composite of a Wilcoxon-Mann-Whitney type test is investigated. Assumptions needed are weaker than those in the semiparametric approach. Numerical comparisons with the moderate-p version of the semiparametric approach show that the nonparametric test has very similar size but achieves superior power, especially for skewed data with some amount of dependence between variables.

Finally, we conduct an extensive simulation to compare our proposed methods with other nonparametric test and rank transformation methods. A wide spectrum of simulation settings is considered. These simulation settings include a variety of heavy tailed and skewed data distributions, homoscedastic and heteroscedastic covariance structures, various amounts of dependence and choices of tuning (smoothing window) parameter for the asymptotic variance estimators. The fully-nonparametric and the rank transformation methods behave similarly in terms of type I and type II errors. However, the two approaches fundamentally differ in their hypotheses. Although there are no formal mathematical proofs for the rank transformations, they have a tendency to provide immunity against effects of outliers. From a theoretical standpoint, our nonparametric method essentially uses variable-by-variable ranking which naturally arises from estimating the nonparametric effect of interest. As a result of this, our method is invariant against application of any monotone marginal transformations. For a more practical comparison, real-data from an Encephalogram (EEG) experiment is analyzed.

KEYWORDS: Multivariate Analysis, High Dimension, Statistical Tests.

Author's signature: Alejandro G. Villasante Tezanos

Date: August 1, 2019

COMPOSITE NONPARAMETRIC TESTS IN HIGH DIMENSION

By

Alejandro G. Villasante Tezanos

Director of Dissertation: Solomon W. Harrar

Director of Graduate Studies: Katherine L. Thompson

Date: August 1, 2019

To Samuel, Paula, Santiago and Carolina

ACKNOWLEDGMENTS

I would like to sincerely thank all of the wonderful professors I have had throughout my graduate career at the University of Kentucky. Without your patience and guidance, I would not be where I am today. A special thank goes to Dr. Harrar for his mentorship and friendship that adviced me tirelessly through this process. Thank you, Dr. Harrar, for your patience during this time. I hope to be able to continue working with you in the future and to keep on learning from your expertise. To Dr. Stromberg for his friendship and continuous mentorship, thank you so much for giving me a job that helped me support my family during this time. I will always remember your guidance with so many consultations and projects and for offering great judgement the countless times I came to see you for help. To my friends and classmates Zaid, Sisheng, Liangdong, Lee and many others that spent good and bad times with me during my graduate student years.

Thanks to the many friends here and far that have stepped up to help or to just listen when I needed an ear. Thanks for pushing me when things got hard and for motivating me to stay on the right path. You have all helped me get to where I am today and I will be forever thankful.

I cannot thank my parents enough , Eugenio and Remedios, and my sister Maite, who came to Lexington to help with my kids while I was working on my degree and have been there for me during my entire life. I also want to thank the rest of my family: grandparents, aunts, uncles, cousins... Thank you all for always being a part of my life and supporting me.

And the most important thank of all goes to my kids and my wife. I want to thank my kids for loving me like they do, even without having their dad by their side on many hours, days and weekends. I will never forget this sacrifice and I hope

to be able to make it up to you, I love you. To my wife, for choosing me and for embarking with me in this adventure that has taken so much effort from her. Thank you, my love, for all you have done to make this possible and for all the support that you have given to me through some good and rough times, I love you.

TABLE OF CONTENTS

Acknowledgments iii

List of Tables vii

List of Figures ix

Chapter 1 Introduction 1

Chapter 2 The High Dimensional Problem 4

2.1 Introduction 4

2.2 Classical Approach 4

2.3 High Dimensional Approach 6

2.4 Objective of the Dissertation 14

Chapter 3 Semiparametric High Dimensional Tests 16

3.1 Introduction 16

3.2 Model and Hypothesis 18

3.3 Test Statistic 19

3.4 Main Results 23

3.5 Simulation 25

3.6 Real Data Example 39

3.7 Conclusions 40

3.8 Appendix 41

Chapter 4 Nonparametric Method for Two Samples High Dimensional Tests 60

4.1 Introduction 60

4.2 Model and Hypothesis 62

4.3 Test Statistic 63

4.4 Main Results 64

4.5 Simulation 65

4.6 Conclusions 72

4.7 Appendix 73

Chapter 5 Comparison of Various High Dimensional Tests 76

5.1 Introduction 76

5.2 Compared Methods of Analysis 76

5.3 Simulation 78

5.4 Area Under Reciever Operating Characteristic (ROC) Curve Analysis 98

5.5 Effect of Scaling Transformation 98

5.6 Electroencephalogram Data Analysis 101

5.7 Discussion and Conclusion 104

Chapter 6 Discussion, Conclusion and Future Directions 109

Bibliography 111

Vita 116

LIST OF TABLES

2.1 MANOVA table 6

3.1 Achieved type I error rates for three groups with nominal size $\alpha = 0.05$ for the moderate and large p versions of the equal variance test under Parzen Smoothing Window L. Values L=10, L=20. Sizes $(n_1, n_2, n_3) = (30, 35, 40)$. 27

3.2 Achieved type I error rates for three groups with nominal size $\alpha = 0.05$ for the moderate and large p versions of the equal variance test under Parzen Smoothing Window L. Values L=10, L=20. Sizes $(n_1, n_2, n_3) = (12, 15, 18)$. 28

3.3 Achieved type I error rates for five groups with nominal size $\alpha = 0.05$ for the moderate and large p versions of the equal variance test under Parzen Smoothing Window L. Values L=10, L=20. Sizes $(n_1, n_2, n_3, n_4, n_5) = (30, 35, 40, 25, 28)$. 29

3.4 Achieved type I error rates for five groups with nominal size $\alpha = 0.05$ for the moderate and large p versions of the equal variance test under Parzen Smoothing Window L. Values L=10, L=20. Sizes $(n_1, n_2, n_3, n_4, n_5) = (12, 15, 18, 13, 16)$. 30

3.5 Achieved type I error rates for three groups with nominal size $\alpha = 0.05$ for the moderate and large p versions of the unequal covariance test under Parzen Smoothing Window 3.3. Values L=10, L=20. Sizes $(n_1, n_2, n_3) = (30, 35, 40)$. 31

3.6 Achieved type I error rates for three groups with nominal size $\alpha = 0.05$ for the moderate and large p versions of the unequal covariance test under Parzen Smoothing Window L. Values L=10, L=20. Sizes $(n_1, n_2, n_3) = (12, 15, 18)$. 32

3.7 Achieved type I error rates for five groups with nominal size $\alpha = 0.05$ for the moderate and large p versions of the unequal covariance test under Parzen Smoothing Window L. Values L=10, L=20. Sizes $(n_1, n_2, n_3, n_4, n_5) = (30, 35, 40, 25, 28)$. 33

3.8 Achieved type I error rates for five groups with nominal size $\alpha = 0.05$ for the moderate and large p versions of the unequal covariance test under Parzen Smoothing Window L. Values L=10, L=20. Sizes $(n_1, n_2, n_3, n_4, n_5) = (12, 15, 18, 13, 16)$. 34

3.9 Statistics and pvalues from precipitation example 40

4.1 Achieved type I error rates for two groups. Sample sizes are $n_1 = 40$ and $n_2 = 38$. p stands for dimension and Parzen Smoothing Window is denoted by L. 67

4.2 Achieved type I error rates for two groups. Sample sizes are $n_1 = 20, n_2 = 18$. p stands for dimension and Parzen Smoothing Window is denoted by L. 68

5.1 List of pvalues from all tests by electrode for the EEG dataset. 107

5.2 List of pvalues from all tests by electrode for the EEG dataset cont. . . . 108

LIST OF FIGURES

3.1 Plot of Parzen window function,$-x^2+1$, $-|x|^3+1$ and Trapezoid window from Politis and Romano (1995). 22

3.2 Power plot for equal variance statistic. Errors are generated for three groups from Normal distribution. Sample sizes are $(n_1, n_2, n_3) = (12, 15, 18)$, proportion of shifted means in larger group $\beta = 0.5$, dimension is $p = 1000$. Independence structure is shown on panel (a) and ARMA(2,2) structure is shown on panel (b). 35

3.3 Power plots for equal variance statistic. Errors are generated for three group from Normal distribution. Sample sizes are $(n_1, n_2, n_3) = (12, 15, 18)$, proportion of shifted means in larger group is $\beta = 0.5$ and dimension is $p = 300$. Independence structure is shown on panel (a) and ARMA(2,2) structure on panel (b). 36

3.4 Power plots for equal variance statistic. Errors are generated for three groups from Gamma distribution. Sample sizes are $(n_1, n_2, n_3) = (12, 15, 18)$, proportion of shifted means in larger group is $\beta = 0.5$ and dimension $p = 300$. Independence structure is shown on panel (a) and ARMA(2,2) structure on panel (b). 36

3.5 Power plots for unequal variance statistic. Errors are generated for three groups from Normal distribution. Sample sizes are $(n_1, n_2, n_3) = (12, 15, 18)$, proportion of shifted means in larger groups is $\beta = 0.5$ and dimension $p = 1000$. Independence structure is shown on panel (a) and ARMA(2,2) structure on panel (b). 37

3.6 Power plots for unequal variance statistic. Errors are generated for three groups from Normal distribution. Sample sizes are $(n_1, n_2, n_3) = (12, 15, 18)$, proportion of shifted means in larger group is $\beta = 0.5$ and dimension $p = 300$. Independence structure is shown on panel (a) and ARMA(2,2) structure on panel (b). 38

3.7 Power plots for unequal variance statistic. Errors are generated for three groups from Gamma distribution. Sample sizes are $(n_1, n_2, n_3) = (12, 15, 18)$, proportion of shifted means in larger group is $\beta = 0.5$ and dimension $p = 300$. Independence structure is shown on panel (a) and ARMA(2,2) structure on panel (b). 38

3.8 Time series plot of precipitation daily averages per group 40

4.1 Power plots for VH-np and GCT-mdp. Errors are generated from Cauchy distribution with ARMA structure. Dimension is $p = 1000$. Sample sizes $(n_1, n_2) = (40, 38)$ are represented on panel (a) and $(n_1, n_2) = (20, 18)$ on panel (b). δ is the location shift. 69

4.2 Power plots for VH-np and GCT-mdp. Errors are generated from Cauchy distribution with ARMA structure. Dimension is $p = 300$. Sample sizes $(n_1, n_2) = 40, 38$ are represented on panel (a) and $(n_1, n_2) = (20, 18)$ on panel (b). δ is the location shift. 69

4.3 Power plots for VH-np and GCT-mdp. Errors are generated from Cauchy distribution with independence structure. Dimension is $p = 300$. Sample sizes $(n_1, n_2) = (40, 38)$ are shown on panel (a) and $(n_1, n_2) = (20, 18)$ on panel (b). δ is the location shift. 70

4.4 Power plots for VH-np and GCT-mdp. Errors are generated from GEV distribution with independence structure. Dimension is $p = 300$. Sample sizes $(n_1, n_2) = (40, 38)$ is shown on panel (a) and $(n_1, n_2) = (20, 18)$ on panel (b). λ is the location and shape parameter and is set equal in the second group compared to 0 for both in the first group. 71

4.5 Power plots for VH-np and GCT-mdp. Errors are generated from GEV distribution with ARMA structure. Dimension is $p = 300$. Sample sizes are $(n_1, n_2) = (40, 38)$.$\beta = 0.4$ is shown on panel (a) and $\beta = 0.8$ is shown on panel (b). λ is the location and shape parameter and is set equal in the second group compared to 0 for both in the first group. β is the proportion of variables shifted. 71

4.6 Power plots for VH-np and GCT-mdp. Errors are generated from Normal distribution with independence structure. Sample sizes are $(n_1, n_2) = (20, 18)$. On panel (a), dimension $p = 300$ is represented, $p = 1000$ is on panel (b). δ is the location shift. 72

5.1 Achieved type-I error rates for all tests against sample size and dimension. Errors are generated form Normal distribution with Square-Root-Decay as covariance structure and heteroscedastic. Parzen Smoothing Window used is L=20. Sample sizes are $n_1 = m$ and $n_2 = m$. Panel (a) represents VH-np, panel (b) represents GCT-mdp, panel (c) represents DLRT, panel (d) represents KH and panel (e) represents GCT-lgp. 81

5.2 Achieved type-I error rates for all tests against sample size and dimension. Errors are generated from Cauchy distribution with Square-Root-Decay as covariance structure and heteroscedastic. Parzen Smoothing Window used is L=20. Sample sizes are $n_1 = m$ and $n_2 = m$. Panel (a) represents VH-np, panel (b) represents GCT-mdp, panel (c) represents DLRT, panel (d) represents KH and panel (e) represents GCT-lgp. 82

5.3 Achieved type-I error rates for all tests against sample size and dimension. Errors are generated from Gamma distribution with Square-Root-Decay as covariance structure and heteroscedastic. Parzen Smoothing Window used is L=20. Sample sizes are $n_1 = m$ and $n_2 = m$. Panel (a) represents VH-np, panel (b) represents GCT-mdp, panel (c) represents DLRT, panel (d) represents KH and panel (e) represents GCT-lgp. 83

5.4 Achieved type-I error rates for VH-np against sample size and dimension. Errors are generated from Normal distribution. Parzen Smoothing Window used is L=20. Covariance structures are Independence for panels (a) and (b) and Square-Root-Decay heteroscedastic for panels (c) and (d) . For panels (b) and (d) sample sizes are $n_1 = m$ and $n_2 = m + 5$ and panels (a) and (c) sample sizes are $n_1 = n_2 = m$. 84

5.5 Achieved type-I error rates for all tests against dimension. Errors are generated from Normal distribution, all tests are represented in all panels. Covariance structure is Independence. Parzen Smoothing Window used is L=20. Sample sizes are $n_1 = m$ and $n_2 = m + 5$. $m = 30$ in panel (a), $m = 40$ in panel (b) and $m = 50$ in panel (c). 84

5.6 Achieved type-I error rates for VH-np against sample size and dimension. Errors are generated from Gamma distribution. Parzen Smoothing Window used is L=20. Covariance structures are Independence for panels (a) and (b) and Square-Root-Decay heteroscedastic for panels (c) and (d). Sample sizes are $n_1 = m$ and $n_2 = m + 5$ for panels (b) and (d) and $n_1 = n_2 = m$ for panels (a) and (c). 85

5.7 Achieved type-I error rates for all tests against sample size and correlation parameter (ρ). Errors are generated from Normal distribution. Square-Root-Decay as covariance structure and dimension is $p = 500$. m is the increment in size which is the same on both samples, so sample sizes are $n_1 = m$ and $n_2 = m$. Parzen Smoothing Window used is L=20. Panel (a) is VH test, panel (b) represents GCT-mdp, panel (c) represents DLRT, panel (d) represents KH and panel (e) represents GCT-lgp. 86

5.8 Achieved type-I error rates for all tests against sample size and correlation parameter (ρ). Errors are generated from Cauchy distribution. Square-Root-Decay as covariance structure and dimension is $p = 500$. m is the increment in size which is the same on both samples, so sample sizes are $n_1 = m$ and $n_2 = m$. Parzen Smoothing Window used is L=20. Panel (a) represents VH test, panel (b) represents GCT-mdp, panel (c) represents DLRT, panel (d) represents KH and panel (e) represents GCT-lgp. . . . 87

5.9 Achieved type-I error rates for all tests against sample size and correlation parameter (ρ). Errors are generated from Gamma distribution. Square-Root-Decay as covariance structure and dimension is $p = 500$. Sample sizes are $n_1 = m$ and $n_2 = m$. Parzen Smoothing Window used is L=20. Panel (a) represents VH test, panel (b) represents GCT-mdp, panel (c) represents DLRT, panel (d) represents KH and panel (e) represents GCT-lgp. 88

5.10 Achieved type-I error rates for VH-np against sample size and dimension. Errors generated from Normal distribution, under heteroscedasticity. Parzen Smoothing Window used is L=20. Panel (a) represents independence covariance structure and panel (b) represents Equi-Correlation and panel (c) represents Auto-Regressive and panel (d) represents Square-Root-Decay. Sample sizes are $n_1 = m$ and $n_2 = m$ and p is the dimension. 89

5.11 Achieved type-I error rates for VH-np against sample size and dimension. Errors are generated from Cauchy distribution, under heteroscedasticity. Parzen Smoothing Window used is L=20. Panel (a) represents Equi-Correlation covariance structure, panel (b) represents Auto-Regressive, panel (c) represents Square-Root-Decay. Sample sizes are $n_1 = m$ and $n_2 = m$ and p is the dimension. 90

5.12 Achieved type-I error rates for all tests against dimension. Errors are generated from Cauchy distribution. Covariance structure is Auto-Regressive, sample sizes are $n_1 = m$ and $n_2 = m$. Parzen window parameter used is L=20. $m = 30$ in panel (a), $m = 40$ in panel (b) and $m = 50$ in panel (c). 90

5.13 Achieved type-I error rates for VH-np against sample size and dimension. Errors are generated from Gamma distribution, under heteroscedasticity($\rho = 0.1$). Parzen Smoothing Window used is L=20. Panel (a) represents independence covariance structure, panel (b) represents Equi-Correlation, panel (c) represents Auto-Regressive and panel (d) represents Square-Root-Decay, m is the increment in size which is the same on both samples, so sample sizes are $n_1 = m$ and $n_2 = m$ and p is the dimension. 91

5.14 Achieved type-I error rates for VH-np against sample size and dimension. Errors are generated from Normal distribution, under heteroscedasticity. Panel (a) represents parameter $L = 20$ and panel (b) represents $L = p/2$. Sample sizes are $n_1 = m$ and $n_2 = m$ and p is the dimension. 92

5.15 Power plot for all tests under shifting alternative. Errors are generated from Gamma distribution. Sample sizes are $(n_1, n_2) = (20, 25)$ and dimension $p = 500$. Covariance structure is Auto-Regressive and homoscedastic. Parzen Smoothing Window used is L=20. 93

5.16 Power plot for all tests against proportion of variables shifted alternative. Errors are generated from Gamma distribution. Sample sizes are $(n_1, n_2) = (20, 25)$ and dimension $p = 500$. Parzen Smoothing Window used is L=20. Covariance structure is Auto-Regressive and homoscedastic. 94

5.17 Power plot for all tests under shifting alternative. Errors are generated from Gamma distribution. Covariance structure is homoscedastic Auto-Regressive. Parzen Smoothing Window used is L=20. Proportion of active variables is $\beta = 0.5$. 95

5.18 Power plot for all tests under shifting alternative. Errors are generated from Gamma distribution. Proportion of active variables is $\beta = 0.5$. Size is unbalanced $(n_1, n_2) = (20, 25)$. Dimension is $p = 500$. Parzen Smoothing Window used is L=20. Panel (a) corresponds to Independent covariance matrix, panel (b)corresponds to Auto-Regressive heteroscedastic, panel (c) corresponds to Auto-Regressive homoscedastic, panel (d) corresponds to Square-Root-Decay heteroscedastic and panel (e) corresponds to Square-Root-Decay homoscedastic. 96

5.19 Power plot for all tests under shifting alternative. Errors are generated from Gamma distribution. Sample sizes are $(n_1, n_2) = (20, 25)$, dimension is $p = 500$, covariance structure is Auto-Regressive with proportion of active variables is $\beta = 0.5$. Parzen Smoothing Window used is L=20. . . 97

5.20 ROC curves that plot sensitivity against 1-specificity. Errors are generated from Gamma distribution. Sample sizes are $(n_1, n_2) = (20, 25)$ with $\delta = 4$. Dimension is $p = 500$. Covariance structure is Auto-Regressive and homoscedastic. Parzen Smoothing Window used is L=20. 99

5.21 Power plots for all tests under shifting alternative. Errors are generated from Gamma distribution. Sample sizes are $(n_1, n_2) = (50, 50)$. Dimension is $p = 500$. Covariance structure is $diag(1, ..., p-1)$ and homoscedastic. Parzen Smoothing Window used is L=20. 100

5.22 Power plots for all tests under shifting alternative. Errors are generated from Gamma distribution. Sample sizes are $(n_1, n_2) = (50, 50)$. Dimension is $p = 500$. Covariance structure is $\Sigma_{3i} + diag(1, ..., p-1)$ and homoscedastic. Parzen Smoothing Window used is L=20. 101

5.23 Plots of average ERP's (brain activity) per electrode over time by the Control and Alcoholic groups. 103

5.24 Venn diagram of counts of significant group effect differences for electrodes. 104

5.25 Plot of pvalues of nonparametric tests by electrodes. 104

Chapter 1 Introduction

This dissertation encompasses the comparison of two or more groups of vectors. This comparison can be achieved with a parametric approach by comparing the means of the different groups. When data is not continuous or it is heavily skewed, comparing means may not be appropriate and a nonparametric approach might be more reasonable, i.e. comparing nonparametric quantities such as relative effects. For the two sample comparison of means, the T^2-statistic is defined in Hotelling (1931) (see also Anderson (2003)) as

$$T^2 = [\overline{\boldsymbol{X}}_1 - \overline{\boldsymbol{X}}_2]^\top \left[(\frac{1}{n_1} + \frac{1}{n_2}) S_{\text{pooled}} \right]^{-1} [\overline{\boldsymbol{X}}_1 - \overline{\boldsymbol{X}}_2] \tag{1.1}$$

where S_{pooled} is the pooled sample covariance matrix, $\overline{\boldsymbol{X}}_1$ and $\overline{\boldsymbol{X}}_2$ are the sample mean vectors. This test is invariant under linear transformations. Its exact distribution under the Null hypothesis is known and it is powerful when dimension is small compared to sample size. The test is, however, not well defined when dimension is larger than sample sizes.

For the multiple group comparison of means, one classical approach uses the statistic

$$\Lambda^* = \frac{|W|}{|B + W|} \tag{1.2}$$

where $B = \sum_{i=1}^{a} n_i (\overline{\boldsymbol{X}}_i - \overline{\boldsymbol{X}})(\overline{\boldsymbol{X}}_i - \overline{\boldsymbol{X}})^\top$ and $W = \sum_{i=1}^{a} \sum_{j=1}^{n_i} (\boldsymbol{X}_{ij} - \overline{\boldsymbol{X}}_i)(\boldsymbol{X}_{ij} - \overline{\boldsymbol{X}}_i)^\top$ with subindex i corresponding to group and j the subject within sample group. This test has similar advantages and disadvantages to Hotelling's T^2, in that it requires the sample size to be larger than the dimension. Furthermore, the statistical power would be weak if sample size is not relatively small compared to dimension (Bai and Saranadasa, 1996). When data comes from ordinal variables or data is heavily skewed, comparing means may not be optimal. A more suitable comparison for the two group case would be comparing relative effects. The univariate version of relative effect is defined by Brunner and Munzel (2000) as $\omega = P(X_{11} < X_{21}) + \frac{1}{2} P(X_{11} = X_{21})$. The interpretation for this univariate version is that a random variable in the first group is said “to tend to have smaller values” than a random variable in the second group if $\omega > 1/2$. This quantity can be naturally extended to the multivariate context as the vector $\boldsymbol{\omega} = (\omega_1, \omega_2, ..., \omega_p)^\top$. A test for this extension was proposed by Brunner, Munzel, and Puri (2002) as

$$T_R^2 = n(\widehat{\boldsymbol{\omega}} - \frac{1}{2}\mathbf{1}_p)^\top V_n^{-1}(\widehat{\boldsymbol{\omega}} - \frac{1}{2}\mathbf{1}_p)$$

where $\widehat{\boldsymbol{\omega}}$ is an estimate of $\boldsymbol{\omega}$ based on ranks, $\mathbf{1}_p = (1, 1, ..., 1)^\top$ and V_n^{-1} is also a sample covariance matrix based on ranks. This test is appropriate for skewed and ordinal data, but it is also underpowered or even not defined when sample size is not relatively small to dimension.

The high availability of large datasets has forced science and specifically Statistics to develop new methods. Classical methods solve the comparisons satisfactorily when the sample size is large compared to the dimension. However, in contemporary data analysis, cases in which the dimension far exceeds the sample size are frequently encountered. In recent years, not only has availability to store data increased exponentially but also smartphones and other electronic devices have made it significantly easier to gather information of every activity registered in them. An example of this is geospatial data, time series, and many others. Another example of a larger dimension is genetic data, more specifically microarray gene expressions, where there are thousands of observations per subject and a handful of subjects per group.

To address this need, many different methods have been proposed in the last two decades. Bai and Saranadasa (1996) devised a test to compare two groups that relaxes the restrictions on dimension and size. However, this test is still not powerful when dimension is much larger than sample size. It also assumes a fast decay of covariance structure and higher order dependence. Cai, Liu, and Xia (2013) proposed a test that solves the problem of high p and small n. It is particularly powerful against sparse alternative since its statistic is supremum-based. It also assumes equal covariance matrix for the two groups which is restrictive. Chen and Qin (2010) innovated by not assuming equal covariance in their test, relaxed the higher order dependence and the relationship between n and p. Srivastava, Katayama, and Kano (2013) proposed an invariant test under units of measurement in which it still made some restrictive assumptions in terms of covariance sparsity, but made no assumptions in higher order dependence. All these tests make assumptions on the covariance structures that make them somehow restrictive. Gregory et al. (2015) proposed a test that has milder assumptions than the previous tests in the two sample case setting. The multiple group case has been treated by recent papers as well. Yamada and Srivastava (2012) tackle the multiple group comparison assuming equal covariance matrix. They also make restrictive assumptions in terms of covariance and higher order dependence. Hu, Bai, et al. (2015) and Zhang, Guo, and Zhou (2017) have not assumed equal covariance matrix but still make restrictive assumptions in terms of the dependence structure.

We propose two composite tests that are powerful against weak and dense signal and have milder assumptions and restrictions than the previous methods. One of the tests undertakes the parametric approach and the other test focuses on the nonparametric one. Both make very few distributional assumptions, which makes them very competitive and versatile for various data types. These tests are backed with theoretical results along with extensive simulations.

This dissertation contains six chapters. In Chapter 1, an introduction to the high-dimensional inference is provided. Chapter 2 reviews recent researches done to overcome the issues that arise from high dimensionality. In Chapters 3 and 4, we propose and study a number of new semi-parametric and fully nonparametric tests for high-dimensional group comparison. We conduct an extensive simulation study in Chapter 5. The conclusion of the researches of the dissertation are summarized in Chapter 6 Also, in Chapter 6, future research directions are pointed out.

Chapter 2 The High Dimensional Problem

2.1 Introduction

To contextualize the tests presented in Chapters 3 and 4, influential papers in the topic will be reviewed. In particular, their scope of applications and shortcomings will be discussed. We will divide the review into four sections including this one. Section 2.2 will introduce the classical approach to the problem, followed by the high dimensional approach in Section 2.3. Finally, Section 2.4 will summarize the information and the gaps that we will fill with our proposed tests and simulations.

Let us first set up the model and notations used in the sequel. Assume $\boldsymbol{X}_{i1}, ..., \boldsymbol{X}_{in_i}$ be independent samples, where $\boldsymbol{X}_{ij} = (x_{ij1}, ..., x_{ijp})^\top$, with mean $\boldsymbol{\mu}_i$ and covariance Σ_i, for $i = 1, ..., a$. Here, a is the number of groups or populations to be compared. Also, let n_i, $\overline{\boldsymbol{X}}_i$ and S_i be the sample size, mean and covariance matrix respectively for the i^{th} sample, with $n = \sum_{i=1}^{a} n_i$.

2.2 Classical Approach

The classical approach for this problem extends what is known for unidimensional outcomes to multidimensional outcomes. First, if interest lies in comparing the mean vectors of two populations, i.e. testing the hypothesis of $H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$. A T^2-statistic to test for the equality of mean vectors from two multivariate populations can be developed by analogy to the univariate square of the t-statistic. Similarly to the univariate case, depending on the sample sizes more assumptions may be needed. If sample size is not large enough both populations may need to be assumed normally distributed or even both covariance structures may need to be assumed equal $\Sigma_1 = \Sigma_2$. Given the dimensionality of the problem this assumption is much stronger than the univariate counterpart. For this situation the T^2-statistic is defined in Hotelling (1931) as in 1.1.

Then, the T^2-statistic is distributed as

$$T^2 \sim \frac{n_1 + n_2 - 2}{n_1 + n_2 - p - 1} F_{p, n_1 + n_2 - p - 1}$$

where p is the dimension and F_{p, n_1+n_2-p-1} refers to the Snedecor's F distribution with p and $n_1 + n_2 - p - 1$ degrees of freedom.

If equal covariance cannot be assumed for both populations, then we cannot find an easy statistic whose distribution doesn't depend on the covariance structures. If sample sizes are large enough, p fixed, such that even $n_1 - p$ and $n_2 - p$ are large, then a test similar to T^2 with some modifications is adequate. The statistic

$$T^{*^2} = [\overline{\boldsymbol{X}}_1 - \overline{\boldsymbol{X}}_2 - (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)]^\top \left[(\frac{1}{n_1} S_1 + \frac{1}{n_2} S_2) \right]^{-1} [\overline{\boldsymbol{X}}_1 - \overline{\boldsymbol{X}}_2 - (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)]$$

approximately follows a Chi-square distribution with p degrees of freedom. A finite sample approximation is also available in Krishnamoorthy and Yu (2004).

If we want to extend to $a > 2$ populations or groups, then the univariate test would be ANOVA. The multidimensional approach for this one is called Multivariate Analisys of Variance (MANOVA) as it can be seen in Anderson (2003).

MANOVA, analogically to ANOVA, has a summary table:

Table 2.1: MANOVA table

Source of Variation	Matrix of Sum of Squares and cross products	Degrees of freedom
Treatment	$B = \sum_{i=1}^{a} n_i(\overline{\boldsymbol{X}}_i - \overline{\boldsymbol{X}})(\overline{\boldsymbol{X}}_i - \overline{\boldsymbol{X}})^\top$	$a - 1$
Residual(Error)	$W = \sum_{i=1}^{a} \sum_{j=1}^{n_i} (\boldsymbol{X}_{ij} - \overline{\boldsymbol{X}}_i)(\boldsymbol{X}_{ij} - \overline{\boldsymbol{X}}_i)^\top$	$\sum_{i=1}^{a} n_i - a$
Total(corrected for the mean)	$B + W = \sum_{i=1}^{a} \sum_{j=1}^{n_i} (\boldsymbol{X}_{ij} - \overline{\boldsymbol{X}})(\boldsymbol{X}_{ij} - \overline{\boldsymbol{X}})^\top$	$\sum_{i=1}^{a} n_i - 1$

Table 2.1 has the same form as the ANOVA table only involving sums of squares and cross-product matrices instead of just scalar numbers. A statistic proposed by Wilks 1.2 is a likelihood ratio test that will reject the Null hypothesis when

$$-(n - 1 - \frac{(p-a)}{2}) \ln \Lambda^* = -(n - 1 - \frac{(p-a)}{2}) \ln \frac{|W|}{|B+W|} > \chi^2_{p(a-1)}(\alpha)$$

where $\chi^2_{p(a-1)}(\alpha)$ is the Chi-square distribution with $p(a-1)$ degrees of freedom for test size α.

2.3 High Dimensional Approach

The classical approach in Section 2.2 to the test that we are interested in relies heavily in having a large sample size relative to the dimension of the problem.

Given the actual interest for high dimensional data, it is often questioned whether the vector means of multiple populations are the same or different. It is usually the case that the number of dimensions of such vectors exceeds by far the sample sizes. This is a situation where conventional test statistics such as the previously discussed are not feasible or well defined. When dimension is much larger than the sample size, estimating mean and covariance structure of the vectors is impossible through regular methods such as maximum likelihood. The main difficulty to find tests that are viable for the task is to estimate the dependence relationship between the different observations within subjects. Given the high dimensionality of the problem $p(p+1)/2$ estimates of the variance-covariance matrix are to be found.

Some assumptions will have to be made so that estimation problem can be simplified. Interest in tests for such situations is steadily growing, specially in biological applications (Gadbury et al. (2004), Liao and Chin (2007), Zhang, Zhang, and Wells (2008)).

In these applications, the classical approaches such as Hotelling's T^2-test for two groups or Lawley-Hotelling trace test, Pillai's trace test or Wilks' lambda for multiple groups are no longer powerful or well defined.

Two Sample Problem $a = 2$

Equal Covariance Matrices

The high-dimensional two sample mean comparison was first was first formally studied by Bai and Saranadasa (1996) in the two sample problem where the asymptotic power of the Hotelling's test and Dempster's non-excact test Dempster (1958) are also discussed and a strong dependence on normality assumption is pointed out. A new asymptotic test is proposed without relying on normality of the data:

$$M_n = (\overline{\boldsymbol{X}}_1 - \overline{\boldsymbol{X}}_2)^\top(\overline{\boldsymbol{X}}_1 - \overline{\boldsymbol{X}}_2) - \frac{1}{n_1 n_2}\mathrm{tr}(S_{\mathrm{pooled}}).$$

Under the Assumption 2.3.1- Assumption 2.3.5 , M_n conveniently scaled has asymptotic Normal distribution, i.e.

$$Z_n = \frac{M_n}{\sqrt{\mathrm{Var} M_n}} \xrightarrow{d} \mathcal{N}(0,1), \text{ as } n, p \to \infty.$$

The variance is consitently estimated by

$$\widehat{\mathrm{var}}(M_n) = \frac{2(n+1)}{n} B_n^2$$

and

$$B_n^2 = \frac{1}{(n+2)(n-1)}(\mathrm{tr} S_{\mathrm{pooled}}{}^2 - \frac{1}{n}(\mathrm{tr} S_{\mathrm{pooled}})^2).$$

The following assumptions were needed for the asymptotic results.

Assumption 2.3.1. $\boldsymbol{X}_{ij} = \Gamma \boldsymbol{Z}_{ij} + \boldsymbol{\mu}_j$ *where* Γ *is a* $p \times m$ *matrix* $(m \leq \infty)$ *such that* $\Gamma\Gamma' = \Sigma$ *,* $\boldsymbol{Z}_{ij} = (z_{ij1}, ..., z_{ijm})^\top$ *are iid with mean* $\mathbf{0}$ *and covariance matrix* $\boldsymbol{I}_m$*,* $E(z_{ijk}^4) = 3 + \Delta \leq \infty$ *and* $E(\Pi_{k=1}^m z_{ijk}^{\nu_k}) = 0$ *when at least one* $\nu_k = 1$ *or* $E(\Pi_{k=1}^m z_{ijk}^{\nu_k}) = 1$ *when there are two* ν_k*'s equal to 2, whenever* $\nu_1 + .. + \nu_m = 4$*.*

Assumption 2.3.2. $(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)^\top \Sigma (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1) = o(\frac{n_1+n_2}{n_1 n_2} \Sigma^2)$*.*

Assumption 2.3.3. $\lambda_{\max}\Sigma = o(\sqrt{tr\Sigma^2})$*, where* $\lambda_{\max}(\Sigma)$ *is the maximum eigenvalue of the covariance matrix.*

Assumption 2.3.4. $p/n \to y > 0$*.*

Assumption 2.3.5. *and* $n_1/(n_1 + n_2) \to \kappa \in (0, 1)$*.*

This test is based on the squared Euclidean norm of the difference between the sample mean vectors. Assumption 2.3.1 defines $\boldsymbol{X}_{ij}$ as linear transformations of uncorrelated variables (z_{ijk}) that are centered at 0, along with these properties, moments and moments of cross-products are meant to guarantee a certain pseudo-independence between the components. Not restricting the value of m to be less than p assures certain flexibility on the dependency structure. Assumption 2.3.2 and Assumption 2.3.3 are related to the covariance structure to restrict it so that none of the eigenvalues is too big with respect to the dimension. Assumption 2.3.4 is restricting p and n to be of the same order of magnitude. Assumption 2.3.5 is restricting n_1 and n_2 so that they grow proportionally, avoiding too unbalanced situations.

This test behaves better than the other two classical approaches under non normality but it assumes equal covariance structure. A criticism for this test comes from Assumption 2.3.4. It is commonly the case where p is large and n is small.

Another approach for the two sample problem under equal covariance assumption was introduced by Cai, Liu, and Xia (2013). They proposed the test statistic:

$$M_{\widehat{\Omega}} = \frac{n_1 n_2}{n_1 + n_2} \max_{1 \leq i \leq p} \frac{\widehat{\boldsymbol{Z}}_i^2}{\widehat{\omega}_{ii}^{(0)}}, \tag{2.1}$$

where $\widehat{\boldsymbol{Z}}_i = (\widehat{Z}_1, \widehat{Z}_2, ..., \widehat{Z}_p)^\top = \widehat{\Omega}(\overline{\mathbf{X}}_1 - \overline{\mathbf{X}}_2)$, $\quad \widehat{\omega}_{ii}^{(0)} = \frac{n_1}{n_1+n_2}\widehat{\omega}_{ii}^{(1)} + \frac{n_2}{n_1+n_2}\widehat{\omega}_{ii}^{(2)}$

$$(\widehat{\omega}_{ij}^{(l)}) := \frac{1}{n_1}\sum_{k=1}^{n_1}(\widehat{\Omega}\boldsymbol{X}_{lk} - \overline{\boldsymbol{X}}_{l\widehat{\Omega}})(\widehat{\Omega}\boldsymbol{X}_{lk} - \overline{\boldsymbol{X}}_{l\widehat{\Omega}})^\top, \, \overline{\boldsymbol{X}}_{l\widehat{\Omega}} = \frac{1}{n_l}\sum_{k=1}^{n_l}\widehat{\Omega}\boldsymbol{X}_{lk},$$

and $\widehat{\Omega}$ estimate of the precision matrix $\Omega = \Sigma^{-1}$.

Let $\Lambda = (\lambda_{ij})$ be the correlation matrix for $\boldsymbol{X}_{1j}$ and $\boldsymbol{X}_{2j}$ and $\Lambda^{(t)} = (\lambda_{ij}^{(t)})$ be the correlation matrix for $\Omega\boldsymbol{X}_{1j}$ and $\Omega\boldsymbol{X}_{2j}$. Cai, Liu, and Xia (2013) make the assumptions:

Assumption 2.3.6. *There exist $C_0 < \infty$ such that $C_0^{-1} \leq \lambda_{\min}(\Sigma) \leq C_0$ where $\lambda_{\min}(\Sigma)$ is the smallest eigenvalue of Σ.*

Assumption 2.3.7. $\max_{1\leq i<j\leq p}\left|\lambda_{ij}\right| \leq r_1 < 1$ *for some constant* $0 < r_1 < 1$.

Assumption 2.3.8. $\max_{1\leq i<j\leq p}\left|\lambda_{ij}^{(t)}\right| \leq r_2 < 1$ *for some constant* $0 < r_2 < 1$.

The assumptions made in this test are basically more specific restrictions in the eigenvalues and entries of both correlation matrices involved. The minimum eigenvalue is bounded so that matrix has full rank and correlations are bounded above by a number smaller than one, which guarantees no perfect correlation between variables.

This test statistic under the null hypothesis follows asymptotically an extreme value type I distribution and hence a test can be performed. This test which is based on a linear transformation of the data by the precision matrix (Ω) is especially advised in the case of sparse alternative (i.e. the mean difference happens only in a small proportion of the variables).

In the case of sparse alternative simulation shows that even though size is similar to the other tests, power is higher.

Tests such as the ones proposed by Bai and Saranadasa (1996), Srivastava and Du (2008), Srivastava (2009) and Chen and Qin (2010) are based on sum of squares statistics which are known to have good power against dense alternatives. However, for a number of applications especially in biology, e.g. imaging anomaly detection and genomics, the means for both groups are the same or almost the same in the sense that the may only differ in a small number of variables.

The main criticism in these tests is the assumption of equal covariance matrix which, as mentioned earlier, is rather strong.

Unequal Covariance Matrices

A step further in the development of these types of tests was introduced by Chen and Qin (2010). This paper is also dealing with two samples but Assumption 2.3.4 is no longer assumed. The test statistic is

$$T_n := \frac{\sum_{i \neq j}^{n_1} \boldsymbol{X}_{1i}^\top \boldsymbol{X}_{1j}}{n_1(n_1-1)} + \frac{\sum_{i \neq j}^{n_2} \boldsymbol{X}_{2i}^\top \boldsymbol{X}_{2j}}{n_2(n_2-1)} - 2\frac{\sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \boldsymbol{X}_{1i}^\top \boldsymbol{X}_{2j}}{n_1 n_2} \tag{2.2}$$

Here also, Assumption 2.3.5 is required. Further, the model is similar to that of Bai and Saranadasa (1996) given in Assumption 2.3.1 with a few changes. Now m is restricted to be greater than p. In addition, the following assumptions are made.

Assumption 2.3.9. $E(z_{ijl_1}^{\alpha_1} z_{ijl_2}^{\alpha_2} ... z_{ijl_q}^{\alpha_q}) = E(z_{ijl_1}^{\alpha_1}) E(z_{ijl_2}^{\alpha_2}) ... E(z_{ijl_q}^{\alpha_q})$ *for a positive integer* q *such that* $\sum_{l=1}^{q} \alpha_l \leq 8$ *and* $l_1 \neq l_2 \neq ... \neq l_q$.

Assumption 2.3.10. $(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \Sigma_i (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) = o[n^{-1} tr(\Sigma_1 + \Sigma_2)^2]$ *as* $n, p \to \infty$.

Assumption 2.3.11. $tr(\Sigma_i \Sigma_j \Sigma_l \Sigma_h) = o(tr^2(\Sigma_1 + \Sigma_2) \quad$ *for* $\quad i, j, l, h = 1$ *or* 2, *as* $p \to \infty$.

The idea for this test originated from the test from Bai and Saranadasa (1996), by eliminating terms in the test statistic that impose size and dimension restrictions but are not useful.

It uses a relaxed version of the model from Bai and Saranadasa (1996) expressed in Assumption 2.3.1 since pseudo-independence property of the components is extended to cross products of up to 8 variables. The test statistic normalized asymptotically follows a standard Normal distribution.

Both Chen and Qin (2010) and Bai and Saranadasa (1996) are invariant under the group of orthogonal transformations but they are not invariant under changes of scale, e.g. changes of units of measurement. Orthogonal transformations preserve angles and distance between points. Examples of such transformations are rotations, symmetries, translations. Units may vary when applying a scale or projection transformations. To overcome the limitations with scale invariance Srivastava, Katayama, and Kano (2013) introduced the test

$$T = \frac{\hat{q}_n}{\sqrt{\widehat{\text{Var}}(\hat{q}_n)c_{p,n}}} = \frac{(\overline{\boldsymbol{X}}_1 - \overline{\boldsymbol{X}}_2)^\top \widehat{D}^{-1}(\overline{\boldsymbol{X}}_1 - \overline{\boldsymbol{X}}_2) - p}{\sqrt{p\widehat{\text{Var}}(\hat{q}_n)c_{p,n}}} \tag{2.3}$$

where

$$\widehat{D} = \frac{\widehat{D}_1}{n_1} + \frac{\widehat{D_2}}{n_2}, \quad \widehat{D}_i = \text{diag}(s_{i11}, ..., s_{ipp}) \text{ and } c_{p,n} = 1 + \frac{\text{tr}R}{p^{3/2}}.$$

The quantity $c_{p,n}$ is a correction term needed for a faster convergence and was given in Srivastava and Du (2008) in connection with a test when the covariance matrices of the two groups are equal. This test makes the following assumptions in addition to Assumption 2.3.5.

Assumption 2.3.12. $0 < c_1 < \min_{1 \leq k \leq p} \sigma_{ikk} \leq \max_{1 \leq k \leq p} \sigma_{ikk} < c_2 < \infty$ *uniformly in* p *where* σ_{ikk} *is the* k^{th} *diagonal entry of* Σ_i.

Assumption 2.3.13. $\lim_{p \to \infty} tr{\Lambda_p}^4/(tr{\Lambda_p}^2)^2 = 0$, *where* $\Lambda_p = D^{-1/2}(\frac{\Sigma_1}{n_1} + \frac{\Sigma_2}{n_2})D^{-1/2}$.

Assumption 2.3.14. $n_m = O(p^\delta), \delta > 1/2, n_m = \min(n_1, n_2)$.

In this test, Assumption 2.3.13 is weaker than Assumption 2.3.11. However, Assumption 2.3.12 is made in exchange, but this assumption is weaker and more reasonable. Recall that, the main advantage of this test with respect to Chen and Qin (2010) and Bai and Saranadasa (1996) is that the test is scale invariant and it should not be affected by the choice of units. Likewise, Feng et al. (2015) proposed a variation of Bai and Saranadasa (1996) and Chen and Qin (2010) that it is also scale invariant using very similar assumptions to those described in Chen and Qin (2010) (Assumptions 2.3.9, 2.3.10 and 2.3.11). For that reason details of this test are omitted.

Another approach was Gregory et al. (2015) based on an average of the t^2 statistic for each variable. Conditions and assumptions for this test are described in depth in Sections 3.1 and 3.2.

More Than Two Samples $a > 2$

Equal Covariance Matrices

An step towards extending the problem from the previous papers to more than two groups was considered among others, by Srivastava and Kubokawa (2013) and Hu, Bai, et al. (2015).

The model studied by Srivastava and Kubokawa (2013) is very similar but extending to multiple groups a regression notation is being used.

They proposed the statistic

$$T_1 = \frac{\text{tr}(\ BD_s^{-1})\ - Np(a-1)(N-2)^{-1}}{2c_{p,N}(a-1)(\text{tr}(\ R^2)\ - N^{-1}p^2)]^{\ 1/2}} \tag{2.4}$$

where $c_{pN} = 1 + (\text{tr}[\ R^2]\ /p^{3/2})$ and $D_s = \text{diag}(S)$.

As in Srivastava, Katayama, and Kano (2013), this test is also invariant under scale transformations. Assumption 2.3.1 extended to multiple groups is assumed here. In addition, the following assumptions are required.

Assumption 2.3.15. $\lim_{p\to\infty}(tr[\ \Lambda^2]\ /p) < \infty$.

Assumption 2.3.16. $\lim_{p\to\infty}(tr[\ \Lambda^4]\ /p^2) = 0$.

Assumption 2.3.17. $N = O(p^\delta)$,$\delta > 1/2$,$a < \infty$.

Assumption 2.3.18. $\lim_{(n,p)\to\infty}\{(p(a-1))^{-1}tr[\ \Lambda M]\ \} = 0$
where

$$M = (\boldsymbol{\mu}_1, ..., \boldsymbol{\mu}_a)\begin{pmatrix} I_{a-1} \\ -\mathbf{1}_{k-1}^\top \end{pmatrix} B \begin{pmatrix} I_{a-1} & -\mathbf{1}_{a-1} \end{pmatrix} \begin{pmatrix} \boldsymbol{\mu}_1^\top \\ \vdots \\ \boldsymbol{\mu}_a^\top \end{pmatrix}$$

and matrix B is defined in Chapter 3 as (3.6).

These assumptions are adaptations from Srivastava, Katayama, and Kano (2013) to an environment where equal covariance matrix is not assumed for all groups. The main criticism to this extension is that equal covariance matrix is assumed.

Cai and Xia (2014) extended the test Cai, Liu, and Xia (2013) to multiple groups with similar adapted conditions which led to the same characteristics.

Under the assumptions of multivariate normality, Schott (2007) , Srivastava (2007), Srivastava and Fujikoshi (2006) and Yamada and Himeno (2015) developed test for the multigroup mean comparison hypothesis. These tests are not invariant under change of units of measurement which we will not consider further.

Unequal Covariance Matrices

Hu, Bai, et al. (2015) is a multiple group test that in the particular case of $a = 2$ coincides with the one proposed by Chen and Qin (2010). The statistic they studied is

$$T_n^{(a)} = \sum_{i<j}^{a} (\overline{\boldsymbol{X}}_i - \overline{\boldsymbol{X}}_j)^\top (\overline{\boldsymbol{X}}_i - \overline{\boldsymbol{X}}_j) - (a-1) \sum_{i=1}^{a} n_i^{-1} \mathrm{tr} S_i. \tag{2.5}$$

They assume multiple group generalizations of Assumption 2.3.2, Assumption 2.3.9, Assumption 2.3.10 and Assumption 2.3.5. Further, they make the following assumptions.

Assumption 2.3.19. *$tr(\Sigma_l \Sigma_d \Sigma_l \Sigma h) = o[tr(\Sigma_l \Sigma_d) tr(\Sigma_l \Sigma_h)]$, $l, d, h \in \{1, 2, ..., a\}$.*

Assumption 2.3.20. *$(\boldsymbol{\mu}_d - \boldsymbol{\mu}_l)^\top \Sigma_d (\boldsymbol{\mu}_d - \boldsymbol{\mu}_h) = o[n^{-1} tr\{(\sum_{i=1}^{a} \Sigma_i)^2\}]$ for $l, d, h \in \{1, 2, ..., a\}$.*

Assumption 2.3.21. *$n_i/n \to \kappa_i \in (0,1)$ for $i = 1, ..., a, \quad as \quad n \to \infty$.*

Under these assumptions they conclude

$$\frac{T_n^{(a)} - \sum_{i<j}^{a} \left\| \boldsymbol{\mu}_i - \boldsymbol{\mu}_j \right\|^2}{\sqrt{\mathrm{Var}(T_n^{(a)})}} \xrightarrow{d} \mathcal{N}(0,1) \tag{2.6}$$

as $n, p \to \infty$.

In numerical studies this test has shown better performance than the test proposed by Srivastava and Kubokawa (2013).

A multiple group test is prosposed by Aoshima and Yata (2013) in which they use some alternative assumptions to normality that don't differ considerably from those used by Chen and Qin (2010) and for that reason details are omitted. It is worth noting however that in this paper they also proposed a confidence region and sample size formula.

Zhang, Guo, and Zhou (2017) implemented a general linear Hypothesis test of mean vectors that has Hu, Bai, et al. (2015) and Yamada and Himeno (2015) as particular cases. It is based on a linear combination of U-statistics and it is applicable to non-normal data without assuming common covariance matrix. This paper innovates in the hypotheses to be tested but assumptions include those in Hu, Bai, et al. (2015) except Assumption 2.3.20 and hence the result does not vary much from the previous tests in terms of the strength of the assumptions.

2.4 Objective of the Dissertation

Most of the tests discussed in Section 2.3 impose restrictions on the covariance structure that are somehow strong. Essentially, they assume factoring of expectations for mixed moments of up to the eight order, they basically assume a certain correlation structure in which variables are linear combinations of pseudo-independent variables such that there is no dependency for higher moments. Gregory et al. (2015) also use the moment of an asymptotic expansion of the statistic to increase rate of convergence. Our proposed test in Chapter 3 is based on this approach but extended to multiple groups.

The same problem has been treated much less extensively in the nonparametric framework for two samples. Wang, Peng, and Li (2015) proposed a test based on mean differences with restrictions such as equal covariance structure and populations coming form certain generalized elliptical distributions. Ghosh and Biswas (2015) studied a distribution free statistic but restricting to elliptical distibutions as well. Wang, Peng, and Li (2015) and Ghosh and Biswas (2015) have other approaches, but these and other different approaches have made similar assumptions for the covariance structure as in the papers discussed in this chapter and are based on mean differences. Our test proposed in Chapter 4 is built as a composite test like in Gregory et al. (2015) using weaker dependency assumptions than the other nonparametric approaches and with difference based on the nonparametric concept of relative group effect defined in 4.2 as opposed to mean differences. The test is distribution free as well but no assumptions on elliptical populations are made. Further, the test admits populations with distributions that could be anything but degenerate.

In order to complement the results discussed in this chapter we propose the following items that will be considered in the remaining part of this dissertation:

- Extend the moment based finite sample correction as it is shown in Section 3.1 to the multigroup.
- Propose a new rank based approach for comparing high-dimensional groups.
- Compare recent High-dimensional tests and their rank-based analogous in an extensive simulation study.

Chapter 3 Semiparametric High Dimensional Tests

3.1 Introduction

Consider the multiple sample problem in which the dimension far exceeds the sample sizes. We are especially interested in the situation where data can be considered to be ordered in space, time or some other index in such a way that the dependence between two components depends on their displacement. This has applications in biology, as it occurs in chromosomal datasets, and many time series datasets.

In the search for statistics that fit our data assumptions, two statistics are investigated. Each of them will have two versions, One designed for groups with common second and higher moments, and another one for groups with different second and higher moments. As it is briefly stated in Chapter 2, the statistics proposed in this chapter can be seen as a multiple group extensions of two group test proposed by Gregory et al. (2015). More specifically, Gregory et al. (2015) proposed a test based on the average of each variable t^2 test called "Generalized Component Test".

The test statistic is based on T_n which is defined as:

$$T_n = \frac{1}{p}\sum_{k=1}^{p} t_{nk}^2$$

where t_{nk}^2 is the square of the t-statistic for the k^{th} variable. That is,

$$t_{nk}^2 = \frac{(\overline{X}_{1i} - \overline{X}_{2i})^2}{\frac{s_{1i}^2}{n_1} + \frac{s_{2i}^2}{n_2}}$$

where $\overline{X}_{1k}$ and $\overline{X}_{2k}$ are the sample means and s_{1k}^2 and s_{2k}^2 are the sample variances , respectively, of the k^{th} variable.

Then, the test statistic is defined by:

$$G_n^{(L)} \equiv p^{1/2}(T_n - (1 + n^{-1}\widehat{a}_n + n^{-2}\widehat{b}_n))/\widehat{\zeta}_n \tag{3.1}$$

where $\widehat{a}_n = (\widehat{c}_{n1} + ... + \widehat{c}_{np})/p$, $\widehat{b}_n = (\widehat{d}_{n1} + ... + \widehat{d}_{np})/p$, and $\widehat{c}_{nk}$ and $\widehat{d}_{nk}$ for $k = 1, ..., p$ are functions of the sample moments described in Gregory et al. (2015).

While (3.1) is meant for large-p version, there is another statistic defined as

$$G_n^{(M)} \equiv p^{1/2}(T_n - 1)/\widehat{\zeta}_n \tag{3.2}$$

which is designed for a moderate p. The large-p statistic is based on finite sample approximation for the center via asymptotic expansion of the first moment of T_n rather than using the mean of the limiting distribution. More precisely, the main difference between (3.2) and (3.1) is that (3.1) centers T_n with its mean correct up to order $O(n^{-2})$ which is achieved by asymptotically expanding $E(T_n)$ whereas (3.2) centers by the asymptotic mean.

Dependency between the variables keeps the Central Limit Theorem from guaranteeing asymptotic normality of the test statistics. It was shown that the statistics $G_n^{(L)}$ and $G_n^{(M)}$ each will converge to a Normal distribution if α-mixing dependence structure holds among the $t_{nk}^2, k = 1, 2, ..., p$.

Let G_n be either one of the statistics (3.2) or (3.1). Then

$$G_n \equiv p^{1/2}(T_n - \widehat{\xi}_n)/\widehat{\zeta}_n \xrightarrow{d} N(0, 1) \text{ as } n \to \infty$$

as $n \to \infty$ where $\widehat{\zeta}_n$ is a consistent estimator of the asymptotic variance τ_∞ and $\widehat{\xi}_n$ is the centering term from (3.2) or (3.1).

This test is a sum-of-squares based test and can be sensitive to dense but otherwise weak alternatives. Gregory et al. (2015) mentions that it would be better to find a sumpremum-based alternative test in situations when the signal is strong but sparse.

We aim to propose and prove similar results for the general multiple sample situation based on an F-type statistic. To that end, the chapter will be structured in eight sections including this introduction, Section 3.2 sets notation for the model and the hypothesis and assumptions. In Section 3.3, the test statistics are defined. Then, Section 3.4 will contain the main results. Simulations will be presented in Section 3.5 using a variety of sample sizes, dimensions and distributions to describe the situations where the test is more useful. This type of problem will be illustrated with a real data example in the subsequent Section 3.6. Conclusions will be presented in Section 3.7. Proofs of the main results and some relevant intermediate results and lemmas are given in the Appendix (Section 3.8).

3.2 Model and Hypothesis

For each $i = 1, 2, ..., a$, let $\boldsymbol{X}_{ij} = (X_{ij1}, ..., X_{ijp})^\top$ be iid for $j = 1, 2, ..., n_i$ with mean $\boldsymbol{\mu}_i$ and covariance matrix Σ_i. Denote by $n = \sum_{i=1}^{a} n_i$ the total sample size and assume the a samples are independent.

The hypothesis of interest is $H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = ... = \boldsymbol{\mu}_a$ versus H_1 :at least $\exists i, j \in \{1, .., a\} : \boldsymbol{\mu}_i \neq \boldsymbol{\mu}_j$. For testing this hypothesis, let

$$\boldsymbol{F}_n = (F_{n1}, F_{n2}, ..., F_{np})^\top \tag{3.3}$$

where $F_{nk} = \frac{MST_k}{MSE_k}$ is the F statistic for an ANOVA test on the k^{th} variable. We will use two different versions of MSE_k depending on the comparison of the second and further moments for the different groups.

When equal second and further moments are assumed we will refer to (3.3). Alternatively, when second and further moments are not assumed to be equal we will refer to vector

$$\boldsymbol{F}'_n = (F'_{n1}, F'_{n2}, ..., F'_{np})^\top. \tag{3.4}$$

For the development of the theory, we will assume a notion of sparsity for the dependence between the variables. Let

$$\alpha_{ij}(s) = \sup_{k \geq 1} \{|P(A \cap B) - P(A)P(B)| : A \in \mathcal{F}_1^k(i, j) \text{ and } B \in \mathcal{F}_{k+s}^{\infty}(i, j), \}$$

for $i = 1, ..., a$ and $j = 1, ..., n_i$ where $\mathcal{F}_a^b(i, j) \equiv \sigma(\{X_{ijk} : a \leq k \leq b\})$. Here $\alpha_{ij}(s)$ is a dependence coefficient that measures the strength of dependence between variables that are at least s indices apart. It provides a measure of the strenght of dependence between variables that are at least s time points (space units) apart. For notation purposes, the subindices in the coefficients $\alpha_{ij}(s)$ are added for the different sequences of variables, but it will be dropped when the sequence is clear by the context.

The following conditions are needed for later use:

Assumption 3.2.1. *For some* $\delta > 0$ $\sum_{s=1}^{\infty} \alpha_{ij}(s)^{\delta/(2+\delta)} < \infty$.

Assumption 3.2.2. *For some* $\delta > 0$, $E|F_{nk}|^{2s+\delta} < b < \infty$ *for all* $k = 1, ..., p$ *for some integer* $s \geq 1$.

Assumption 3.2.3. $\lim_{n\to\infty} \frac{1}{p-s} \sum_{k=1}^{p-s} Cov(F_{nk}, F_{n(k+s)}) = \gamma(s)$ *exists* $\forall s > 0$.

Assumption 3.2.4. $\max\{E|X_{11k}|^{16}, E|X_{21k}|^{16}, ..., E|X_{a1k}|^{16}, k = 1, ..., p\} = O(1)$.

Assumption 3.2.5. $\min\{Var(X_{11k}), Var(X_{21k}), ..., Var(X_{a1k}), k = 1, ..., p\} > b > 0$.

Assumption 3.2.1 appeals to the dependency structure between variables, assuming dependency fades away as variables are further away from each other, at a rate that is not exponential but rather polynomial . Assumption 3.2.2 refers to the fact that the F-type statistic has a finite second or higher moment. Assumption 3.2.3 is needed to control the sum of covariances of the F_{nk}'s and, ultimately, along with Assumption 3.2.2 assure the finiteness of the variance of the statistic F_n. Assumption 3.2.4 implies that the $16^t h$ moment of each variable is finite and bounded. In Assumption 3.2.5 all variable variances are bounded below by a number greater than 0. Assumption 3.2.4 is used in the proof of Theorem 3.4.1 since some expected values of this power are used and need to be finite. Assumption 3.2.5 is needed since the F-type statistic has the second moment of the variables in the denominator. Very small values of this variances will promote very large values of the F-type statistic. Since the scaling $\widehat{\zeta}_n$ in $F_n^{(M)}$ and $F_n^{(L)}$ (defined in the next section) is a function of autocovariance of the F_{nk}'s, large values of F_{nk} will shrink $F_n^{(M)}$ or $F_n^{(L)}$ toward 0 when they should be producing the opposite effect.

3.3 Test Statistic

We will define the statistic proposed for the multiple group problem under equal and unequal covariance. We consider these two different cases separately since the denominators of the F_{nk} statistics would vary slightly in order to enhance the power of the test statistics under their respective assumptions.

Test Statistic Under Equal Covariance

Under the assumption that second and higher moments are common between the different groups, the statistic in this subsection is based on an average of the usual ANOVA F statistic for each variable considered separately. Let

$$F_n = \frac{1}{p}(F_{n1} + F_{n2} + ... + F_{np}) \tag{3.5}$$

where F_{nk} for $k = 1, ..., p$ is the F statistic for the k^{th} variable and defined by $F_{nk} = MST_k/MSE_k$. Here, $MST_k = \overline{\boldsymbol{X}_k}' B \overline{\boldsymbol{X}_k}/(a-1)$, $\overline{\boldsymbol{X}_k} = (\overline{X}_{1k}, \overline{X}_{2k}, ..., \overline{X}_{ak})^\top$,

$$B = \begin{pmatrix} n_1 - \frac{n_1^2}{n} & -\frac{n_1 n_2}{n} & .. & .. & -\frac{n_1 n_a}{n} \\ -\frac{n_1 n_2}{n} & n_2 - \frac{n_2^2}{n} & .. & .. & -\frac{n_2 n_a}{n} \\ .. & .. & .. & .. & .. \\ .. & .. & .. & .. & .. \\ -\frac{n_1 n_a}{n} & -\frac{n_2 n_a}{n} & .. & .. & n_a - \frac{n_a^2}{n} \end{pmatrix} \tag{3.6}$$

and $MSE_k = \sum_{i=1}^{a} (n-a)^{-1}(n_i - 1)s_{ik}^2$, where s_{ik}^2 the unbiased sample variance for the k^{th} variable in the i^{th} sample and $\overline{X}_{ik}$ the sample mean for the k^{th} variable in the i^{th} sample.

Scaling and centering F_n in a manner analogous to (3.1) and (3.2) we propose two test statistics $F_n^{(M)}$ and $F_n^{(L)}$.

The moderate-p version of the statistic is

$$F_n^{(M)} = \frac{F_n - 1}{\widehat{\zeta}_n}$$

and the large-p version of the statistic is

$$F_n^{(L)} = \frac{F_n - (\widehat{a}_n + \frac{\widehat{b}_n}{n})}{\widehat{\zeta}_n}$$

where, $\widehat{a}_n = p^{-1}(\widehat{c}_{n1} + \widehat{c}_{n2} + ... + \widehat{c}_{np})$ and $\widehat{b}_n = p^{-1}(\widehat{d}_{n1} + \widehat{d}_{n2} + ... + \widehat{d}_{np})$.

The sample quantities $\widehat{c}_{nk}$ and $\widehat{d}_{nk}$ are estimates from c_{nk} and d_{nk} defined in (3.19) and (3.20). They are estimated using sample moments described in Subsection 3.8. The scaling factor will consider the dependence structure between elements of the vector, it is defined as

$$\widehat{\zeta}_n^2 \equiv \sum_{|s|<L} w(s/L)\widehat{\gamma}(s),$$

where $\widehat{\gamma}(s)$ is the sample autocovariance defined by

$$\widehat{\gamma}(s) = \frac{1}{p-s}\sum_{k=1}^{p-s}(F_{nk} - F_n)(F_{n(k+s)} - F_n)$$

and the covariance between variables are weighted according to the lag separation between them. A possible choice for the weight function is $w(s/L)$,

$$w(x) = \begin{cases} 1 - 6|x|^2 + 6|x|^3 & \text{if } |x| < 1/2 \\ 2(1-|x|)^3 & \text{if } 1/2 \leq x \leq 1 \\ 0 & \text{if } |x| > 1 \end{cases}$$

which is the Parzen window discussed in Brockwell and Davis (2013), s is the distance away from the diagonal in the covariance matrix and L is the distance from the diagonal where covariance becomes negligible.

The function $w(x)$ is graphed in Figure 3.1 along with $-x^2+1$, $-|x|^3+1$ and Trapezoid windows from Politis and Romano (1995) to emphasize the pace at which the weight decreases as you move away from 0.

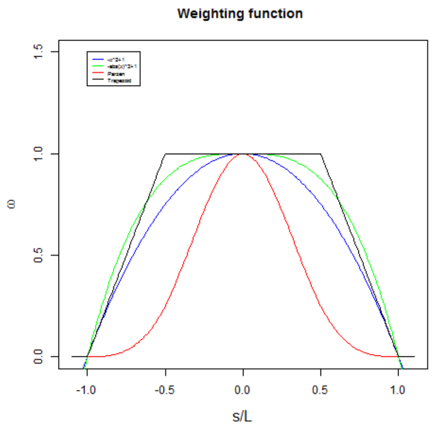


Figure 3.1: Plot of Parzen window function,$-x^2+1$, $-|x|^3+1$ and Trapezoid window from Politis and Romano (1995).

As a function of the ratio between s and L, $w(x)$ gives a reasonable weight for the covariance estimates. Recall that assumptions of α-mixing guarantee that as we move away from any given element in the vector, the dependence and , hence, the correlation fades away. Introducing this weight will lead to a consistent estimator of the asymptotic variance by taking advantage of the assumed α-mixing structure. No further window functions have been investigated.

Test Statistic Under Unequal Covariance

When the group covariance matrices are unequal, we modify MSE_k so that its expectation equals that of MST_k under the null hypothesis. Then, similarly to the statistic defined in the previous section, we define

$$F_n^{'} = p^{-1}(F_{n1}^{'} + F_{n2}^{'} + ... + F_{np}^{'})$$

where

$$F_{nk}^{'} = \frac{MST_k^{'}}{MSE_k^{'}} \text{ for } k = 1, ..., p,$$

with

$$MST_k^{'} = \frac{1}{a-1}\sum_{i=1}^{a}(\overline{X}_{ik} - \overline{X}_{.k})^2 \text{ , } MSE_k^{'} = \frac{1}{a}\sum_{i=1}^{a}\frac{1}{n_i}s_{ik}^2 \text{ and } \overline{X}_{.k} = \frac{1}{a}\sum_{i=1}^{a}\overline{X}_{ik}.$$

The test statistics $F_n^{'(M)}$ and $F_n^{'(L)}$ are defined similarly to $F_n^{(M)}$ and $F_n^{(L)}$. Two different centering options are used.

The moderate-p version of the statistic is

$$F_n^{'(M)} = \frac{F_n' - 1}{\widehat{\zeta_n}}$$

and the large-p version of the statistic is

$$F_n^{'(L)} = \frac{F_n' - (1 + \frac{\widehat{b'_n}}{n})}{\widehat{\zeta_n}}$$

where $\widehat{b'_n} = \frac{1}{p}(\widehat{d'_{n1}} + \widehat{d'_{n2}} + ... + \widehat{d'_{np}})$.

Population quantity d'_{nk} defined in Subsection 3.8 is estimated as $\widehat{d'_{nk}}$. Estimates are calculated by substituting sample moments (Subsection 3.8) in population parameters. The estimator $\widehat{\zeta_n}$ is defined in exactly the same way as in Subsection 3.3 except using the quantities F_n' and F_{nk}'.

3.4 Main Results

We will establish the asymptotic normality of the centered and scaled test statistic F_n. A similar statement is also true for F_n'. The following will be proved:

Theorem 3.4.1. *: Let us assume that $p \equiv p_n = o(n^4)$ and Assumptions 3.2.1,3.2.2 3.2.3, 3.2.4 and 3.2.5 hold with $s = 1$. Then,*

$$sup_{x\in\mathbb{R}}|P(F_n - a_n < x) - \Phi\{\sqrt{p}(x - n^{-1}b_n)/\tau_\infty\}| = o(1)$$

where

$$\tau_\infty^2 = \gamma(0) + 2\sum_{s=1}^{\infty}\gamma(s) < \infty$$

$$a_n = \frac{c_{n1} + c_{n2} + ... + c_{np}}{p} \text{ and } b_n = \frac{d_{n1} + d_{n2} + ... + d_{np}}{p}$$

where c_{nk} and d_{nk} are defined in (3.19 and 3.20) and satisfy

$$a_n \to 1 \text{ and } b_n = O(1) \text{ as } n \to \infty.$$

The proof for this theorem is given in the Appendix, Subsection 3.8. The statement and proof for the statistic under the assumption of non equal variance and higher moments remains the same and it is omitted.

Expansion for the First Moment of F_n and F_n'

Proposition 3.4.1. *: Assuming* $\{X_{1jk}, j = 1, ..., n_1\}, \{X_{2jk}, j = 1, ..., n_2\}, ...,$ $\{X_{ajk}, j = 1, ..., n_a\}$ *are independent and identically distributed random samples for all* $k = 1, .., p$, $E[X_{1jk}] = E[X_{2jk}] = ... = E[X_{ajk}] = \mu_k$, $Var[X_{1jk}] = \sigma^2_{1k}$, $Var[X_{2jk}] = \sigma^2_{2k}, ...,$ $Var[X_{ajk}] = \sigma^2_{ak}$ *and Assumptions3.2.4 and 3.2.5.*

Let

$$F_{nk} = \frac{\overline{\boldsymbol{X}_k}' B \overline{\boldsymbol{X}_k}/(a-1)}{\sum_{i=1}^{a} \frac{(n_i - 1)s^2_{ik}}{n-a}}$$

where s^2_{ik} *is the sample variance of the* k^{th} *variable in the* i^{th} *group and* $n/n_i = O(1) \quad \forall i = 1, ..., a$ *as* $n \to \infty$.

Then

$$E[F_{nk}] = c_{nk} + n^{-1} d_{nk} + o(n^{-\frac{3}{2}}).$$

The proof for this proposition is presented in the Appendix, in Subsection 3.8. The rationale for choosing the centering values is based on the finite sample approximation for the center via asymptotic expansion of the first moment. Depending on how many terms are included in the expansion, the rate of convergence will vary.

$$\mathrm{E}[F_n] = a_n + \frac{1}{n} b_n + o(n^{-\frac{3}{2}}) \text{ and } \mathrm{E}[F_n'] = 1 + \frac{1}{n} b_n' + o(n^{-\frac{3}{2}}).$$

This implies,

$$\mathrm{E}[\sqrt{p}(F_n - (a_n + \frac{1}{n} b_n))] = \sqrt{p} o(n^{-\frac{3}{2}}) \text{ and } \mathrm{E}[\sqrt{p}(F_n' - (1 + \frac{1}{n} b_n'))] = \sqrt{p} o(n^{-\frac{3}{2}}).$$

When only the first term is kept, p needs to grow at the rate $p = o(n^2)$. On the other hand, if the first two terms are included, p needs to grow at the rate $p = o(n^4)$. Therefore, when only a_n is included, F_n is expected to have lower rate of convergence than when both a_n and b_n are included. This shows that as more terms are included in the expansion, the higher theoretical precision as n and p increase. This means that more terms in the expansion lead to a potentially better approximation for the null distribution in smaller sample size situations.

Large-p version of the tests allow for $p = o(n^4)$, but since large-p versions of the tests include higher-order sample moments, they are more sensitive to outliers and their performance under heavy tailed distributed data could be worse than the moderate-p version.

All statements in this section can be applied to F_n' and b_n' as well.

3.5 Simulation

We aim to show the performance of the proposed statistics in terms of size and power under various settings. More precisely, we investigate how the large-p versions of the tests (from now on also called "VH-lgp") compare to the moderate-p versions of the tests (from now on also called "VH-mdp"), in particular in the case of small sample sizes. Since we derived an approximation for the expected value of the statistic, that has a potential to improve the rate of convergence than the method without expansion, we would expect the approximation to perform better in the small sample size environment.

In order to make the simulation as thorough as possible, we have investigated multiple combinations of parameter values. Specifically, effects in the number of groups a, sample sizes n_i and dimension p are investigated. Parzen Smoothing Window parameter (L) is used and needs to be specified before hand to estimate the variance of the test statistic, it dictates the extent to which the dependency is estimated in the variance. In the power simulation, there are other parameters used: δ which expresses the shift of the mean and β, which controls the proportion of the means shifted for the alternative hypothesis.

The results of the simulations are given in Tables 3.1-3.4 for the statistics with equal covariance matrices and in Tables 3.5-3.8 for the statistics with unequal covariance matrices. Power is also compared and displayed in Figures 3.2 to 3.7. The settings for these simulations are more restricted and specified in each figure.

Simulation Design

Sizes were compared for VH-mdp and VH-lgp under the following settings:

- Sample sizes of $(n_1, n_2, n_3) = \{(12, 15, 18), (30, 35, 40)\}$ and $(n_1, n_2, n_3, n_4, n_5) = \{(12, 15, 18, 13, 16)$,$(30, 35, 40, 25, 28)\}$.
- Dimensions: $p = \{300, 1000\}$.
- Two values for the parameter L are also used. $L = \{10, 20\}$.
- Dependence model: Independence and ARMA structure for the errors for the p dimensions.
- Error distribution: N(0,1), centered Gamma(4,2), Uniform(-5,5) and Double exponential(0,1).

For the dependency structure, we used ARMA(q_1,q_2) errors in which each element in the vector depends on the closest elements following the formula:

$$X_t = \varepsilon_t + \sum_{k=1}^{q_1} \varphi_k X_{t-k} + \sum_{k=1}^{q_2} \theta_k \varepsilon_{t-k}$$

where ε_t is a white noise error term and φ_k and θ_k are the coefficients that define the structure. The simulation used here is an ARMA(2,2) model with coefficients $\varphi_1 = 0.8897$, $\varphi_2 = -0.4858$, $\theta_1 = -0.2279$ and $\theta_2 = 0.2488$.

In order to compare power, samples under the alternative hypothesis were generated by shifting half of the means ($\beta = 0.5$) of the largest group by a δ amount. Power was simulated under the same settings as size with the exception of distributions and L. Distribution of errors for power were simulated from Normal and Gamma distributions, Parzen Smoothing Window is set to $L = 20$.

Simulation Results

Size Simulation

The actual size is set to $\alpha = 0.05$ in all the simulations. All tables represent result for 7000 simulations. In all tables there is a mix result for the parameter L. We could not find any pattern or setting for which any of the two values of L picked was more advantageous than the other. We speculate that it might be due to the fact that the dependency structure of the setting is not long range, the pick of L might be more important in that case.

Simulations for size show the following results:

Table 3.1: Achieved type I error rates for three groups with nominal size $\alpha = 0.05$ for the moderate and large p versions of the equal variance test under Parzen Smoothing Window L. Values L=10, L=20. Sizes $(n_1, n_2, n_3) = (30, 35, 40)$.

Type-I error rates× 100

Error distr.	Dependence Structure	$\widehat{\xi_n}$	$p = 300$		$p = 1000$	
			L=10	L=20	L=10	L=20
Normal	indep	VH-mdp	6.26	7.53	8.30	8.90
		VH-lgp	5.89	6.63	4.83	5.94
	ARMA	VH-mdp	7.27	6.56	7.74	6.76
		VH-lgp	7.03	7.80	6.11	5.77
Gamma	indep	VH-mdp	6.10	7.16	8.51	6.70
		VH-lgp	5.86	6.90	5.30	5.94
	ARMA	VH-mdp	6.17	12.21	7.91	7.47
		VH-lgp	7.11	7.03	6.09	5.93
Uniform	indep	VH-mdp	6.54	7.37	8.13	8.86
		VH-lgp	6.13	7.39	5.63	5.70
	ARMA	VH-mdp	6.74	7.06	7.81	7.39
		VH-lgp	7.69	7.53	5.90	5.54
Double exp	indep	VH-mdp	6.01	6.91	8.96	8.46
		VH-lgp	5.61	6.69	5.26	5.89
	ARMA	VH-mdp	6.93	6.83	7.71	6.90
		VH-lgp	7.19	7.16	6.19	5.81

Pooled variance. As it can be seen on Table 3.1, the proposed test for VH-lgp is closer to nominal α than the test for VH-mdp in most settings. It is especially more accurate for $p = 1000$ since it performs better under all other settings.

Table 3.2: Achieved type I error rates for three groups with nominal size $\alpha = 0.05$ for the moderate and large p versions of the equal variance test under Parzen Smoothing Window L. Values L=10, L=20. Sizes $(n_1, n_2, n_3) = (12, 15, 18)$.

Type-I error rates× 100

Error distr.	Dependence Structure	$\widehat{\xi_n}$	$p = 300$		$p = 1000$	
			L=10	L=20	L=10	L=20
Normal	indep	VH-mdp	10.49	12.60	28.83	29.09
		VH-lgp	6.07	6.99	5.54	5.71
	ARMA	VH-mdp	9.17	9.01	17.59	17.81
		VH-lgp	7.56	7.13	6.50	6.61
Gamma	indep	VH-mdp	10.56	7.00	28.60	28.51
		VH-lgp	6.01	7.11	5.67	6.66
	ARMA	VH-mdp	8.70	8.79	17.97	17.30
		VH-lgp	7.29	6.46	6.16	5.70
Uniform	indep	VH-mdp	11.41	12.14	29.49	30.04
		VH-lgp	6.23	6.50	5.07	5.21
	ARMA	VH-mdp	9.61	9.20	19.07	18.26
		VH-lgp	6.73	7.36	5.69	5.81
Double exp	indep	VH-mdp	11.50	12.21	28.53	27.56
		VH-lgp	6.16	6.76	6.90	7.89
	ARMA	VH-mdp	9.00	8.87	17.57	16.19
		VH-lgp	6.49	6.89	5.56	6.00

In Table 3.2 the setting changes in sample sizes. As can it be seen, in this setting the proposed VH-lgp version of the test performs much better than the VH-mdp versions. In all instances, except under Gamma independent setting with $p = 300$, the VH-lgp version is closer to nominal size. Especially when $p = 1000$, the sizes are in the double digits for the test VH-mdp version and are very close to the nominal value in the VH-lgp version of the test statistic.

Table 3.3: Achieved type I error rates for five groups with nominal size $\alpha = 0.05$ for the moderate and large p versions of the equal variance test under Parzen Smoothing Window L. Values L=10, L=20. Sizes $(n_1, n_2, n_3, n_4, n_5) = (30, 35, 40, 25, 28)$.

Type-I error rates× 100

Error distr.	Dependence Structure	$\widehat{\xi_n}$	$p = 300$		$p = 1000$	
			L=10	L=20	L=10	L=20
Normal	indep	VH-mdp	6.51	6.77	8.79	8.34
		VH-lgp	5.73	6.90	5.11	5.40
	ARMA	VH-mdp	6.91	7.10	7.94	7.47
		VH-lgp	7.07	7.37	5.97	5.90
Gamma	indep	VH-mdp	6.16	7.07	8.80	7.91
		VH-lgp	5.77	6.47	5.79	5.81
	ARMA	VH-mdp	6.70	7.00	7.17	6.94
		VH-lgp	6.73	7.16	5.86	5.89
Uniform	indep	VH-mdp	5.94	7.34	8.57	8.79
		VH-lgp	6.29	6.47	5.39	5.80
	ARMA	VH-mdp	7.11	7.73	8.09	6.56
		VH-lgp	6.71	7.07	6.24	5.96
Double exp	indep	VH-mdp	6.86	7.33	8.64	8.84
		VH-lgp	5.96	7.01	5.49	5.77
	ARMA	VH-mdp	7.03	7.67	7.21	6.67
		VH-lgp	7.11	7.41	6.10	5.86

Table 3.3 shows results as the number of groups changes with relatively larger sample sizes. The difference in this case is not as large as in Tables 3.1 and 3.2 but VH-lgp is still getting closer to the nominal α than VH-mdp under settings when $p = 1000$. When $p = 300$ there are more instances where VH-lgp is closer to nominal α but the results are more mixed even though both tests have values that differ only by 0.04 from nominal α.

Table 3.4: Achieved type I error rates for five groups with nominal size $\alpha = 0.05$ for the moderate and large p versions of the equal variance test under Parzen Smoothing Window L. Values L=10, L=20. Sizes $(n_1, n_2, n_3, n_4, n_5) = (12, 15, 18, 13, 16)$.

Type-I error rates× 100

Error distr.	Dependence Structure	$\widehat{\xi_n}$	$p = 300$		$p = 1000$	
			L=10	L=20	L=10	L=20
Normal	indep	VH-mdp	10.00	10.97	23.01	23.23
		VH-lgp	5.84	6.64	5.66	5.04
	ARMA	VH-mdp	8.80	8.46	15.50	14.20
		VH-lgp	6.87	7.30	5.96	5.89
Gamma	indep	VH-mdp	9.81	10.17	22.44	22.53
		VH-lgp	6.10	6.19	5.13	5.51
	ARMA	VH-mdp	8.77	8.71	15.23	14.33
		VH-lgp	8.77	7.03	6.31	5.93
Uniform	indep	VH-mdp	10.44	10.93	22.20	24.24
		VH-lgp	5.76	6.79	5.41	5.49
	ARMA	VH-mdp	8.80	8.41	16.10	15.63
		VH-lgp	7.19	7.33	6.03	6.44
Double exp	indep	VH-mdp	10.19	11.51	22.49	22.21
		VH-lgp	5.61	6.94	5.46	6.17
	ARMA	VH-mdp	7.86	8.06	15.70	14.03
		VH-lgp	7.06	6.97	7.19	5.60

When the sample size is reduced in the five group case, the result is very similar to the three group case. In all instances, VH-lgp is closer to nominal α. Especially in the $p = 1000$ case, the performance of the VH-lgp is clearly superior with achieved sizes that are very close to α compared to double digit sizes in VH-mdp.

Table 3.5: Achieved type I error rates for three groups with nominal size $\alpha = 0.05$ for the moderate and large p versions of the unequal covariance test under Parzen Smoothing Window 3.3. Values L=10, L=20. Sizes $(n_1, n_2, n_3) = (30, 35, 40)$.

Type-I error rates× 100

Error distr.	Dependence Structure	$\widehat{\xi_n}$	$p = 300$		$p = 1000$	
			L=10	L=20	L=10	L=20
Normal	indep	VH-mdp	6.34	7.36	8.94	9.37
		VH-lgp	5.66	7.11	5.90	5.37
	ARMA	VH-mdp	6.61	7.16	7.47	7.73
		VH-lgp	6.70	7.50	6.69	6.10
Gamma	indep	VH-mdp	5.97	7.39	9.59	9.44
		VH-lgp	6.93	7.50	5.83	6.19
	ARMA	VH-mdp	7.14	6.63	7.63	6.70
		VH-lgp	7.20	7.50	6.46	5.91
Uniform	indep	VH-mdp	6.53	7.14	9.34	9.50
		VH-lgp	5.79	7.64	5.49	5.34
	ARMA	VH-mdp	6.47	6.63	7.59	6.90
		VH-lgp	7.36	7.50	6.39	6.07
Double exp	indep	VH-mdp	6.13	6.90	8.97	8.89
		VH-lgp	6.96	7.49	7.34	6.36
	ARMA	VH-mdp	6.76	7.06	7.89	7.51
		VH-lgp	7.94	8.56	6.56	6.69

Unpooled variance. Table 3.5 shows simulation results for settings with the statistics for unequal covariance matrix. In this table we can see that, for $p = 300$, all values are similar even though the VH-mdp is consistently better. In the case of $p = 1000$, the VH-lgp is consistently better.

Table 3.6: Achieved type I error rates for three groups with nominal size $\alpha = 0.05$ for the moderate and large p versions of the unequal covariance test under Parzen Smoothing Window L. Values L=10, L=20. Sizes $(n_1, n_2, n_3) = (12, 15, 18)$.

Type-I error rates× 100

Error distr.	Dependence Structure	$\widehat{\xi_n}$	$p = 300$		$p = 1000$	
			L=10	L=20	L=10	L=20
Normal	indep	VH-mdp	11.76	13.27	33.26	34.17
		VH-lgp	7.76	8.70	7.81	8.54
	ARMA	VH-mdp	9.27	9.41	21.87	20.07
		VH-lgp	8.71	9.11	8.96	8.64
Gamma	indep	VH-mdp	12.17	15.62	35.39	35.71
		VH-lgp	9.29	10.52	13.36	13.84
	ARMA	VH-mdp	8.59	8.87	22.49	21.64
		VH-lgp	9.50	9.37	9.54	10.00
Uniform	indep	VH-mdp	12.39	13.66	35.94	35.04
		VH-lgp	6.74	7.37	6.33	6.70
	ARMA	VH-mdp	10.96	9.63	21.33	21.04
		VH-lgp	8.47	8.59	8.14	7.24
Double exp	indep	VH-mdp	11.67	13.86	31.63	31.99
		VH-lgp	13.21	14.54	23.24	24.30
	ARMA	VH-mdp	9.04	9.39	20.20	19.54
		VH-lgp	10.09	10.61	12.03	11.51

Table 3.6 provides another example where for smaller samples sizes, VH-lgp has a clear advantage, especially in large p settings. Even though the errors for the two tests were further from α, the comparison in almost every situation is double digits (VH-mdp) vs single digits (VH-lgp). In the case of moderate p the results are again mixed.

Table 3.7: Achieved type I error rates for five groups with nominal size $\alpha = 0.05$ for the moderate and large p versions of the unequal covariance test under Parzen Smoothing Window L. Values L=10, L=20. Sizes $(n_1, n_2, n_3, n_4, n_5) = (30, 35, 40, 25, 28)$.

Type-I error rates× 100

Error distr.	Dependence Structure	$\widehat{\xi_n}$	$p = 300$		$p = 1000$	
			L=10	L=20	L=10	L=20
Normal	indep	VH-mdp	6.86	7.59	9.23	9.40
		VH-lgp	6.34	6.61	5.36	5.73
	ARMA	VH-mdp	7.06	7.26	7.61	7.26
		VH-lgp	7.24	7.45	6.54	6.01
Gamma	indep	VH-mdp	6.94	7.43	9.07	9.89
		VH-lgp	7.63	8.99	10.43	10.61
	ARMA	VH-mdp	7.23	7.30	7.43	7.96
		VH-lgp	7.77	8.06	6.97	6.71
Uniform	indep	VH-mdp	6.61	7.64	8.71	9.31
		VH-lgp	5.29	6.69	5.64	5.71
	ARMA	VH-mdp	6.64	7.07	8.93	7.76
		VH-lgp	6.79	7.61	6.30	5.80
Double exp	indep	VH-mdp	6.66	7.16	9.37	9.30
		VH-lgp	6.84	7.11	6.16	6.59
	ARMA	VH-mdp	7.09	6.83	8.26	7.43
		VH-lgp	7.06	7.87	6.33	6.44

Table 3.7 contains results comparing the performance of the tests for the unequal covariance in the large sample size and five group ($a = 5$) case. As we can see, performance of both tests is similar with VH-lgp being consistently closer to α when $p = 1000$ and being consistently conservative (below α) when $p = 300$.

Table 3.8: Achieved type I error rates for five groups with nominal size $\alpha = 0.05$ for the moderate and large p versions of the unequal covariance test under Parzen Smoothing Window L. Values L=10, L=20. Sizes $(n_1, n_2, n_3, n_4, n_5) = (12, 15, 18, 13, 16)$.

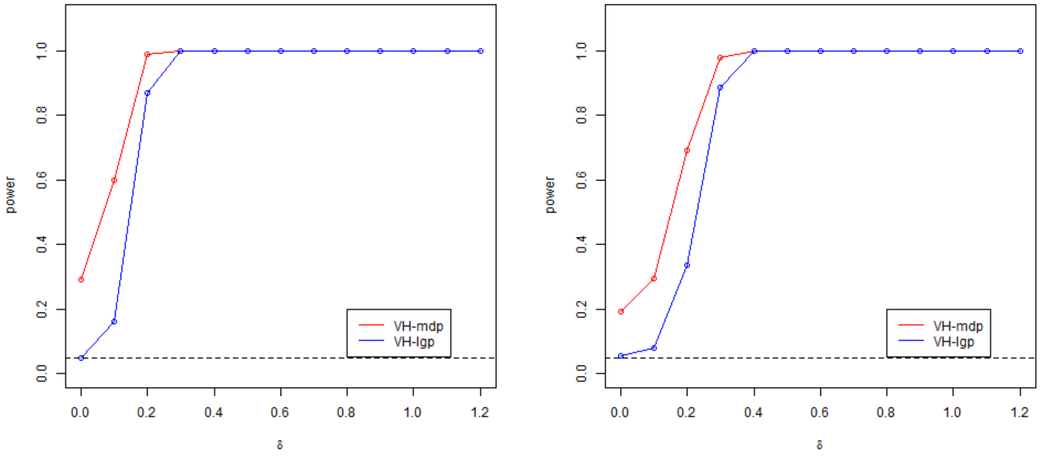
Type-I error rates× 100

Error distr.	Dependence Structure	$\widehat{\xi_n}$	$p = 300$		$p = 1000$	
			L=10	L=20	L=10	L=20
Normal	indep	VH-mdp	11.00	11.69	26.44	26.06
		VH-lgp	6.63	7.80	7.10	7.14
	ARMA	VH-mdp	8.91	9.10	17.56	16.67
		VH-lgp	7.24	7.94	7.44	6.91
Gamma	indep	VH-mdp	11.16	11.61	25.74	26.16
		VH-lgp	13.79	15.44	28.13	28.03
	ARMA	VH-mdp	8.73	9.09	16.79	16.29
		VH-lgp	9.91	9.37	10.19	11.01
Uniform	indep	VH-mdp	11.27	11.84	26.00	26.64
		VH-lgp	6.34	7.04	5.90	6.50
	ARMA	VH-mdp	9.11	8.74	17.34	16.61
		VH-lgp	7.49	7.83	7.13	6.27
Double exp	indep	VH-mdp	10.84	11.23	24.54	24.44
		VH-lgp	10.49	11.44	16.24	16.87
	ARMA	VH-mdp	8.30	8.73	17.26	15.66
		VH-lgp	8.57	9.04	8.96	8.21

In Table 3.8 we can see, similar to the version of the test with equal covariance, that the proposed version of the test is better especially when $p = 1000$. We thought it was worth noting that VH-mdp has sizes closer to α in all settings for centered Gamma except when $p = 1000$ and the dependence has ARMA structure.

Power Simulation

Figures comparing power for the two tests will be shown in some of the settings that were represented in the size tables. Figures are based on 3200 simulations and Parzen Smoothing Window value $L = 20$. The simulation for the alternative was picked by shifting half ($\beta = 0.5$) of the errors in the largest group by δ. For a better organization, we present the results for pooled and unpooled variances separately.



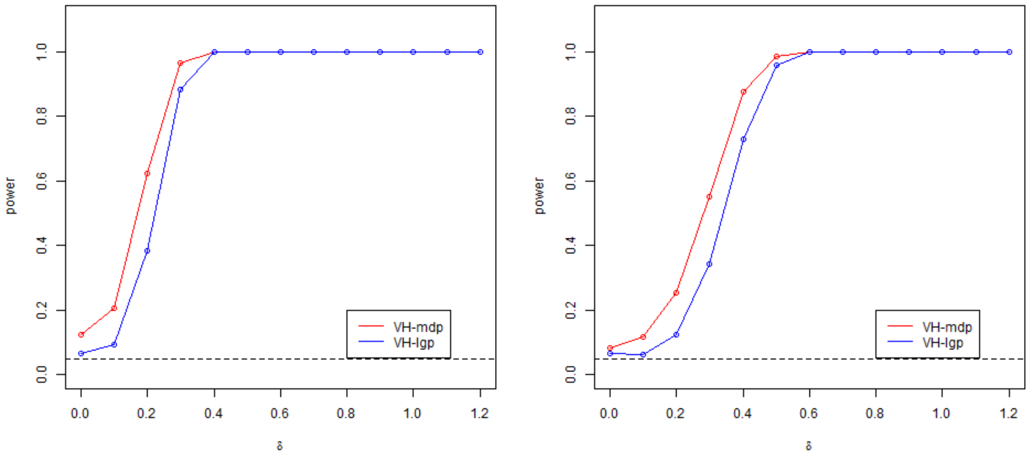
(a)

(b)

Figure 3.2: Power plot for equal variance statistic. Errors are generated for three groups from Normal distribution. Sample sizes are $(n_1, n_2, n_3) = (12, 15, 18)$, proportion of shifted means in larger group $\beta = 0.5$, dimension is $p = 1000$. Independence structure is shown on panel (a) and ARMA(2,2) structure is shown on panel (b).

Pooled variance Figure 3.2 shows that for both, independent and ARMA errors for standard Normal distribution, the tests get to a high degree of power when the largest sample shifted by 0.2 units. Independent errors increase power a little faster than ARMA. We observed that for $p = 1000$ both tests perform similar with VH-lgp version trading off some loss of power in compensation for a much better size.

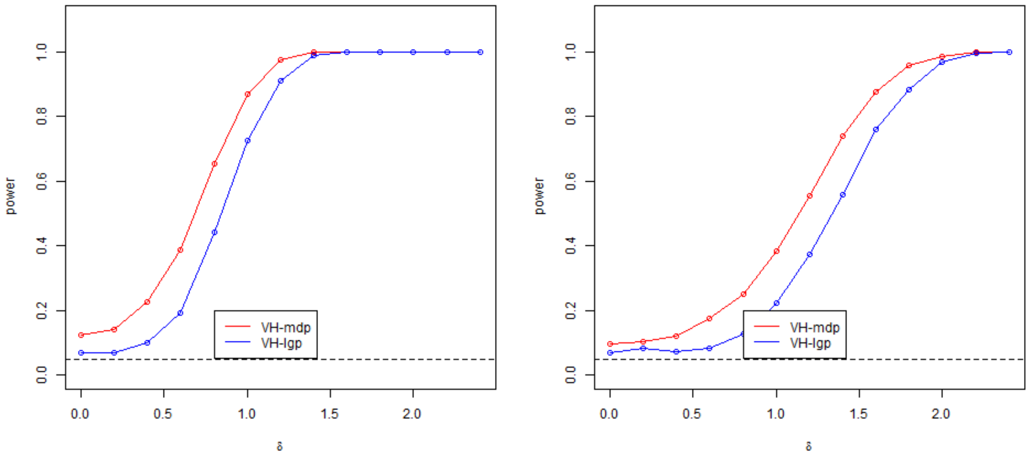
The power plots in Figure 3.3 for Normal distribution where $p = 300$ behave similar to those in Figure 3.2. There is the same trade off in the VH-lgp version of the test with even some initial drop of power in the ARMA simulation.



(a)

(b)

Figure 3.3: Power plots for equal variance statistic. Errors are generated for three group from Normal distribution. Sample sizes are $(n_1, n_2, n_3) = (12, 15, 18)$, proportion of shifted means in larger group is $\beta = 0.5$ and dimension is $p = 300$. Independence structure is shown on panel (a) and ARMA(2,2) structure on panel (b).

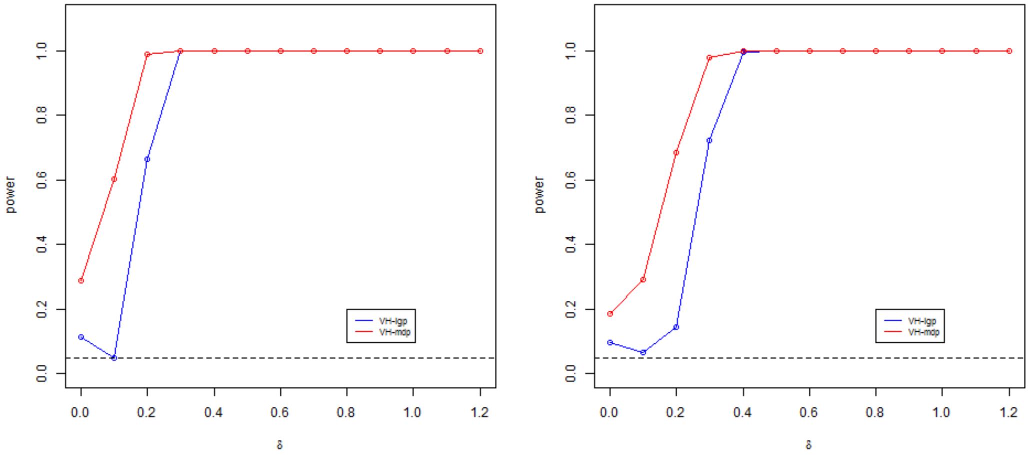


(a)

(b)

Figure 3.4: Power plots for equal variance statistic. Errors are generated for three groups from Gamma distribution. Sample sizes are $(n_1, n_2, n_3) = (12, 15, 18)$, proportion of shifted means in larger group is $\beta = 0.5$ and dimension $p = 300$. Independence structure is shown on panel (a) and ARMA(2,2) structure on panel (b).

Figure 3.4 illustrates the behavior of the statistics under skewed conditions. We can observe the same trade off as in Figures 3.3 and 3.2. Also, power grows at a slower pace for both tests. That slower pace is much more pronounced under dependence structure.



(a)

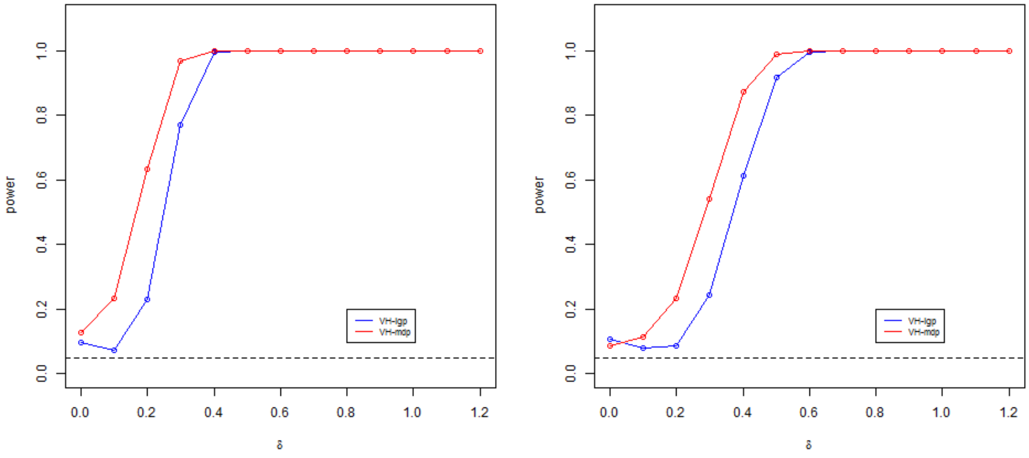
(b)

Figure 3.5: Power plots for unequal variance statistic. Errors are generated for three groups from Normal distribution. Sample sizes are $(n_1, n_2, n_3) = (12, 15, 18)$, proportion of shifted means in larger groups is $\beta = 0.5$ and dimension $p = 1000$. Independence structure is shown on panel (a) and ARMA(2,2) structure on panel (b).

Unpooled variance Figure 3.5 illustrates the behavior of VH-lgp and VH-mdp when variance is not assumed equal under Normal distribution and $p = 1000$. The size of VH-lgp is considerably better, once again, trading it by initial loss of power even more pronounced than in the pooled variance version of the test.

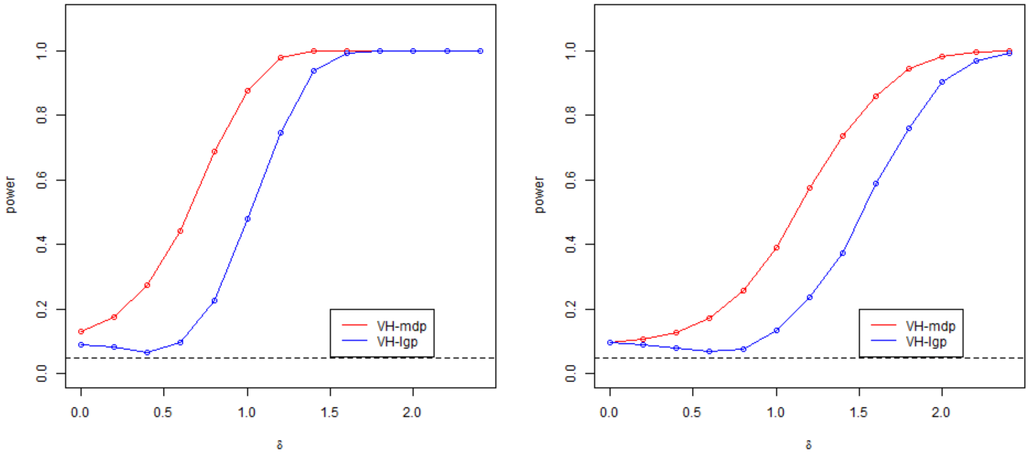
Figure 3.6 illustrates the behavior of the tests when variance is not assumed equal under Normal distribution. Similar pattern is observed as in Figures 3.2, 3.3, and 3.4. In the case of ARMA structure, both size a power are worse for the VH-lgp version.

Figure 3.7 is illustrating the behavior of the statistics for unequal variance under skewed conditions. We can observe the same trade off as in 3.4 . Also power grows



(a) (b)

Figure 3.6: Power plots for unequal variance statistic. Errors are generated for three groups from Normal distribution. Sample sizes are $(n_1, n_2, n_3) = (12, 15, 18)$, proportion of shifted means in larger group is $\beta = 0.5$ and dimension $p = 300$. Independence structure is shown on panel (a) and ARMA(2,2) structure on panel (b).



(a) (b)

Figure 3.7: Power plots for unequal variance statistic. Errors are generated for three groups from Gamma distribution. Sample sizes are $(n_1, n_2, n_3) = (12, 15, 18)$, proportion of shifted means in larger group is $\beta = 0.5$ and dimension $p = 300$. Independence structure is shown on panel (a) and ARMA(2,2) structure on panel (b).

at a slower pace for both tests. That slower pace is much more pronounce under dependence structure. The plot on the right illustrates that VH-mdp version of the test is better in both size and power.

Summary

The sizes under normality are considerably better for VH-lgp in the smaller sample setting. In the larger sample setting, both versions are very similar. Power is generally better in VH-mdp, which is the trade off of VH-lgp for better Type I error.

Under skewness, size is considerably higher than the nominal value in smaller sample size simulation for VH-mdp but VH-lgp it is not too affected. The power is diminished considerably, especially in VH-lgp.

The utility of VH-lgp is especially noticeable in Type I error when sample sizes are small and dimension is large with a trade for initial loss in power. Otherwise VH-mdp seems to have comparable sizes and slightly better power to detect differences.

3.6 Real Data Example

Considering the proposed test is targeting at dense but weak differences as opposed to sparse but strong, a real data example of precipitation at a single station over years comes handy to illustrate. Data was obtained from a database from the National Centers for Environmental Information [1].

Data for 30 years (1986-2016) of precipitation in a single weather station in Miami was obtained. The precipitation records were split into three groups of 10 years each. Dependency is allowed for nearby days but not across years. Therefore, each group is formed by ten years of daily precipitation values. Since comparison is made for each single day of the year, if there are differences over the years they would be weak but dense to describe climate change. For this data set, number of groups is $a = 3$, sample sizes are equal to $n_i = 10$ and dimension is $p = 365$. Time series plot of the group average is shown in Figure 3.8.

[1] https://www.ncdc.noaa.gov/ accessed on October 2018

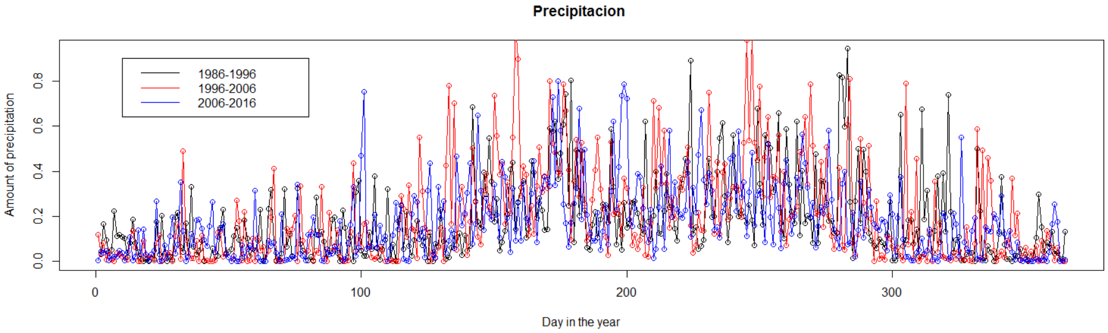


Figure 3.8: Time series plot of precipitation daily averages per group

As it can be observed, there seems to be more spikes that are black or red which come from the earlier groups of years and more blue drops that correspond to the most recent group. The proposed test has a pvalue< 0.001. This would suggest that precipitation is decreasing over the last thirty years.

Table 3.9: Statistics and pvalues from precipitation example

Tests in the example

Test	Test statistic	pvalue
VH-mdp	1.721664	0.08513045
VH-lgp	4.800884	1.57967e-06

As we can see in Table 3.9, VH-mdp does not achieve significance but VH-lgp is highly significant.

3.7 Conclusions

We proposed two statistics to test multiple group differences. The proposed tests are VH-mdp and VH-lgp. Versions under two different assumptions are developed, assuming equal covariance matrix in all groups or assuming unequal covariance matrix in all groups.

Proposed tests in this chapter are shown to asymptotically follow a Normal distribution. The assumption of α-mixing is made in the sample as opposed to the $F_{nk}, i = k, 2, ..., p$. Gregory et al. (2015) assumed α-mixing between the $t^2_{nk}, k = 1, 2, ..., p$ which is much less logical, since α-mixing in the sample could be natural and in the statistics would be more artificial. This makes the result from Theorem 3.4.1 slightly stronger to that of Gregory et al. (2015).

We also showed that the rate of convergence for the statistics from the asymptotic expansion is higher as we develop the expansion further. The drawback of further expansions is having to estimate further moments with the corresponding sensitivity to outliers.

The proposed tests are competitive in the large-p small-n environment when p admits an ordering. Sizes and power were investigated using simulations under various settings. In the simulations, for large p settings, we generally observed better performance of the statistic coming from asymptotic expansion and mixed results when p is moderate. VH-lgp is more robust under skewness, especially when p is larger.

In the example of precipitation over 30 years, we can see that the test picks the difference between the time groups with high significance. This illustrates the statement that this type of test is specially fit when signal is weak and dense.

3.8 Appendix

In this section, we include the complete proof of Proposition 3.4.1 showing the complete expression for $\widehat{a}_n$ and $\widehat{b}_n$, and the rate of convergence of the asymptotic expansion. We also show expression for the value $\widehat{b}_n'$ for the statistic under the assumption of not equal covariance matrix.

Also included in this appendix are Theorem 3.8.1 from Bradley (2005) and Lemmas 3.8.1 and 3.8.2, along with the Lindeberg's CLT theorem for triangular arrays (Theorem 3.8.2 from Billingsley (1995)), which are all used in the proof of our main result, Theorem 3.4.1.

Calculations for the unbiased sample moments up to the fourth order can be found in Subsection 3.8 which are used to estimate the parameters from Proposition 3.4.1. Estimations for the parameters are also included in (3.19) and (3.20) in Subsection 3.8.

Preliminary Lemmas and Theorem

Theorem 3.8.1. *Suppose that for each $n = 1, 2, 3, ...,X^{(n)} := (X_k^{(n)}, k \in \mathbb{Z})$ is a sequence of random variables. Suppose these sequences $X^{(n)}, \quad n = 1, 2, 3, ...$ are independent of each other. Suppose that for each $k \in \mathbb{Z}, \quad h_k; \mathbb{R} \times \mathbb{R} \times \mathbb{R} \times ... \to \mathbb{R}$ is a Borel function. Define the sequence $X := (X_k, k \in \mathbb{Z})$ of random variables by $X_k := h_k(X_k^{(1)}, X_k^{(2)}, X_k^{(3)}, ...), k \in \mathbb{Z}$.*

Then for each $m \geq 1$, $\alpha(m) \leq \sum_{n=1}^{\infty} \alpha_{(n)}(m)$.

The following lemma is useful to take advantage of the α-mixing properties to make sure that variables for which their indices are apart from each other are relatively uncorrelated.

Lemma 3.8.1. *If $Y \in \sigma(X_1, ..., X_i)$ and bounded by B_1, and if $Z \in \sigma(X_{i+n}, X_{i+n+1}...)$ and bounded by B_2, then*

$$\big|E[YZ] - E[Y]\,E[Z]\,\big| \leq 4B_1B_2\alpha(n). \tag{3.7}$$

Lemma 3.8.1 is used to prove the following lemma which we will need in our proof. The following lemma is going a little further than the previous one. If fourth moment of any random variable from the σ-algebra generated by the sequences is bounded, then the correlation also decays according to how far the indexes are from each other.

Lemma 3.8.2. *If $Y \in \sigma(X_1, ..., X_i)$ and $E[Y^4] \leq B_1$, and if $Z \in \sigma(X_{i+n}, X_{i+n+1}...)$ and $E[\,Z^4] \leq B_2$, then*

$$\big|E[YZ] - E[Y]\,E[Z]\,\big| \leq 8(1 + B_1 + B_2)\alpha(n)^{1/2}. \tag{3.8}$$

The following theorem is needed to show convergence of a triangular array in which size (n) and dimension (p) are simultaneously increasing.

Theorem 3.8.2. *Suppose that for each n the sequence $X_{n1}, ..., X_{nr_n}$, where $r_n \to \infty$, is independent and satisfies*

$$E[X_{nk}] = 0,\ \sigma_{nk}^2 = E[X_{nk}^2],\ s_n^2 = \sum_{k=1}^{r_n} \sigma_{nk}^2$$

and

$$\lim_{n\to\infty} \sum_{k=1}^{r_n} \frac{1}{s_n^2} \int_{|X_{nk}|\geq \epsilon s_n} X_{nk}^2 dP = 0 \textit{ for every } \epsilon > 0.$$

Then

$$\frac{S_n}{s_n} \xrightarrow{d} N$$

where $S_n = \Sigma_{k=1}^{r_n} X_{nk}$.

Proof of Theorem 3.4.1

In this section, we show details of the proof for the main result Theorem 3.4.1.

Proof. The proof is divided into four main steps. Step 1 appeals to Bradley (2005) to show that α-mixing in the sample transfers to α-mixing in the F_{ij}'s. Then, Step 2 shows the finiteness of the quantity τ_∞ that corresponds to the variance of the statistic. Step 3, from a modification of the big block little block argument found in Billingsley (1995), will show the asymptotic equivalence of the series of F_{ij}'s to a series of variables that are independent. Step 4, applies Lindeberg's Theorem for triangular arrays and concludes asymptotic normal convergence of the statistic,i.e.

$$p^{-1/2} \sum_{k=1}^{p} (F_{nk} - \mathrm{E}[F_{nk}]) \xrightarrow{d} N(0, \tau_\infty^2)$$

where

$$\begin{aligned} \tau_\infty^2 = \lim_{n\to\infty} \mathrm{Var}(p^{-1/2} \sum_{k=1}^{p} F_{nk}^2) &= \lim_{n\to\infty} p^{-1} \sum_{s=0}^{p-1} \sum_{|k_1-k_2|=s} \mathrm{Cov}(F_{nk_1}, F_{nk_2}) \\ &= \gamma(0) + 2\sum_{s=1}^{\infty} \gamma(s) \end{aligned} \tag{3.9}$$

and

$$\gamma(s) = \lim_{n\to\infty} (p-s)^{-1} \sum_{k=1}^{p-s} \mathrm{Cov}(F_{nk}, F_{n(k+s)}) \ , \ s > 0.$$

Step 1 The assumption of α-mixing in the sample implies that the resulting F_{ij}'s are also α-mixing by Bradley (2005). Bradley's result can be seen in Theorem 3.8.1. Bradley's theorem is defined in an infinite sample but we can take $X^{(n)}$ to be degenerate 0 for $n > \sum n_i$ and the Borel function to be the function that defines the statistic F_{ij}. Hence, Assumption 3.2.1 that was tied to the samples can now be tied to the test statistics.

Step 2 The proof for (3.9) uses the conditions for the moment and α-mixing to show that for any $M \geq 1$

$$p^{-1} \sum_{s=M+1}^{p-1} \sum_{|k_1-k_2|=s} \left|\mathrm{Cov}(F_{nk_1}, F_{nk_2})\right| \leq 2 \sum_{s>M} p^{-1}(p-s)\{\alpha(s)^{\delta/(2+\delta)} \bigvee_{k=1}^{p} (E|F_{nk}|^{2+\delta})^{\frac{2}{2+\delta}}\}$$

$$\leq \sum_{s=M+1}^{\infty} \alpha(s)^{\delta/(2+\delta)} \to 0 \text{ as } M \to \infty.$$

This implies the finiteness of τ_∞.

Step 3 Thus, applying the arguments from Billingsley (1995), we split the sum of $F_{n1} + ... + F_{np}$ into alternate blocks of length b_p and l_p.

We will call

$$U_{npi} = F_{n(i-1)(b_p+l_p)+1} + ... + F_{n(i-1)(b_p+l_p)+b_p} \ , \ 1 \leq i \leq r_p, \tag{3.10}$$

where $r_p = \max\{i : (i-1)(b_p + l_p) + b_p < p\}$

and let

$$V_{npi} = F_{n(i-1)(b_p+l_p)+b_p+1} + ... + F_{ni(b_p+l_p)} \ , \ 1 \leq i < r_p \tag{3.11}$$

$$V_{npr_p} = F_{n(r_p-1)(b_p+l_p)+b_p+1} + ... + F_{np}. \tag{3.12}$$

Then

$$S_{np} = \sum_{i=1}^{r_p} U_{npi} + \sum_{i=1}^{r_p} V_{npi}$$

and we will choose l_p small enough so that the second term in the RHS is small in comparison with the first but large enough so that the variables U_{npi} are nearly independent to be able to use an adaptation of Lyapunov's theorem to show asymptotic normality.

WLOG we will assume $\mathrm{E}[F_{ni}] = 0$.

$$\mathrm{E}[\,S_{np}^4] \,\leq 4! \sum_{i,j,k,l\geq 0, i+j+k+l<p} \big|\mathrm{E}[\,F_{n(1+i)}F_{n(1+i+j)}F_{n(1+i+j+k)}F_{n(1+i+j+k+l)}]\,\big|$$

then, grouping the last three elements in each term and, by Lemma 3.8.2, the RHS is less than or equal to

$$8(1+\mathrm{E}[\,F_{n(1+i)}^4]\,+\mathrm{E}[\,F_{n(1+i+j)}^4F_{n(1+i+j+k)}^4F_{n(1+i+j+k+l)}^4]\,)\alpha(j)^{1/2}$$

and applying Holder's inequality twice in the last expected value and taking

$$\mathrm{E}[\,F_{n*}^4]\,=\max\{\mathrm{E}[\,F_{n(1+i)}^4],\mathrm{E}[\,F_{n(1+i+j)}^4],\mathrm{E}[\,F_{n(1+i+j+k)}^4],\mathrm{E}[\,F_{n(1+i+j+k+l)}^4]\}$$

then, the above expression is at most

$$8(1+\mathrm{E}[\,F_{n*}^4]\,+\mathrm{E}[\,F_{n*}^{12}]\,)\alpha(j)^{1/2} = K_1\alpha(j)^{1/2}.$$

Similarly, grouping the first three elements of each term, $K_1\alpha(l)^{1/2}$ is a bound. The quantity K_1 is also bounded by Assumption 3.2.4.

So,

$$\begin{aligned}\mathrm{E}[\,S_{np}^4]\,&\leq 4!p^2\sum_{j,l\geq 0,j+l<p} K_1\min\{\alpha(j)^{1/2},\alpha(l)^{1/2}\}\\ &\leq K_1p^2\sum_{0\leq j\leq l}\alpha(l)^{1/2} = K_1p^2\sum_{l=0}^{\infty}(l+1)\alpha(l)^{1/2}\end{aligned}$$

by the convergence of the series of $\alpha(k)$ the above series converges and

$$\mathrm{E}[\,S_{np}^4]\,\leq K_2p^2 \tag{3.13}$$

and K_2 independent of p.

Now, let call $b_p = \lfloor p^{4/5}\rfloor$ and $l_p = \lfloor p^{1/5}\rfloor$.

From the definition of r_p

$$b_p \approx p^{4/5}\ ,\ l_p \approx p^{1/5}\ ,\ r_p \approx p^{1/5}. \tag{3.14}$$

Then using Markov's inequality and triangular inequality twice

$$\begin{aligned}
P[\left|\frac{1}{\tau_\infty\sqrt{p}}\sum_{i=1}^{r_p-1}V_{npi}\right| \geq \epsilon] &\leq \frac{E[\left|\frac{1}{\tau_\infty\sqrt{p}}\sum_{i=1}^{r_p-1}V_{npi}\right|]}{\epsilon} \\
&\leq \frac{\sum_{i=1}^{r_p-1}E[\left|\frac{1}{\tau_\infty\sqrt{p}}V_{npi}\right|]}{\epsilon} \\
&\leq \frac{\sum_{i=1}^{r_p-1}\sum_{j=(i-1)(b_p+l_p)+b_p+1}^{i(b_p+l_p)}E[\left|\frac{1}{\tau_\infty\sqrt{p}}I_iF_{nj}\right|]}{\epsilon} \\
&\leq \frac{l_pr_pK'}{\tau_\infty\sqrt{p}\epsilon} \qquad (3.15)
\end{aligned}$$

and this last sequence converges to 0 as p converges to infinity.

Similarly occurs with the last term V_{npr_p}. Therefore,

$$\sum_{i=1}^{r_p}V_{npi}/\tau_\infty\sqrt{p}\xrightarrow{p}0. \qquad (3.16)$$

Let's show now that

$$\sum_{i=1}^{r_p}U_{npi}/\tau_\infty\sqrt{p}\xrightarrow{d}N(0,1). \qquad (3.17)$$

There is a set of independent random variables U'_{npi} that have common distributions with U_{npi}.

Let us apply Lemma 3.8.1 iteratively to the ratio of the characteristic functions of both sets of variables minus 1.

$$\frac{\varphi_{U_1\tau_\infty\sqrt{p}}(t)}{\varphi_{U'_1\tau_\infty\sqrt{p}}(t)}-1=0.$$

$$\begin{aligned}
\frac{\varphi_{(U_1+U_2)/\tau_\infty\sqrt{p}}(t)}{\varphi_{(U'_1+U'_2)/\tau_\infty\sqrt{p}}(t)}-1 &= \mathrm{E}[\,\frac{e^{it(U_1+U_2)/\tau_\infty\sqrt{p}}}{e^{it(U'_1+U'_2)/\tau_\infty\sqrt{p}}}-1] \\
&= \mathrm{E}[\,(e^{it(U_1-U'_1)/\tau_\infty\sqrt{p}}e^{it(U_2-U'_2)/\tau_\infty\sqrt{p}}-1)] \\
&\leq \mathrm{E}[\,e^{it(U_1-U'_1)/\tau_\infty\sqrt{p}}e^{it(U_2-U'_2)/\tau_\infty\sqrt{p}}-1] \\
&\leq 4\alpha_{l_n}.
\end{aligned}$$

In this last step Lemma 3.8.1 is used, 1 is the bound of the absolute value of the characteristic function and the fact that the characteristic function of a degenerate variable in 0 is 1.

Iteratively, if we add another variable

$$\begin{aligned}\frac{\varphi_{(U_1+U_2+U_3)/\tau_\infty\sqrt{p}}(t)}{\varphi_{(U'_1+U'_2+U'_3)/\tau_\infty\sqrt{p}}(t)}-1 &= \mathrm{E}[\,\frac{e^{it(U_1+U_2+U_3)/\tau_\infty\sqrt{p}}}{e^{it(U'_1+U'_2+U'_3)/\tau_\infty\sqrt{p}}}-1] \\ &= \mathrm{E}[\,e^{it((U_1+U_2)-(U'_1+U'_2))/\tau_\infty\sqrt{p}}e^{it(U_3-U'_3)/\tau_\infty\sqrt{p}}-1]\,.\end{aligned}$$

$$\begin{aligned}&\mathrm{E}[\,e^{it((U_1+U_2)/\tau_\infty\sqrt{p}-(U'_1+U'_2)/\tau_\infty\sqrt{p}}e^{it(U_3-U'_3)/\tau_\infty\sqrt{p}}-1] \\ &\qquad\le \mathrm{E}[\,e^{it((U_1+U_2)/\tau_\infty\sqrt{p}-(U'_1+U'_2)/\tau_\infty\sqrt{p}}e^{it(U_3-U'_3)/\tau_\infty\sqrt{p}} \\ &\qquad\quad-\varphi_{(U_1+U_2)/\tau_\infty\sqrt{p}-(U'_1+U'_2)/\tau_\infty\sqrt{p}}(t) \\ &\qquad\quad+\varphi_{(U_1+U_2)/\tau_\infty\sqrt{p}-(U'_1+U'_2)/\tau_\infty\sqrt{p}}(t)-1] \\ &\qquad\le 4\alpha(l_n)+4\alpha(l_n)=4\times 2\times\alpha(l_n).\end{aligned}$$

By induction, the ratio of

$$\sum_{i=1}^{r_p}U_{npi}/\tau_\infty\sqrt{p}\ \text{and}\ \sum_{i=1}^{r_p}U'_{npi}/\tau_\infty\sqrt{p}$$

differ from 1 at most by $4(r_p-1)\alpha(l_p)$.

Since $\alpha(p)=O(p^{-5})$, this difference is $O(n^{-1})$ and the ratio converges uniformly to one. Both sums converge in distribution to a common distribution. So, if we show that the sum of the independent sample converges, the other sum does too.

We know $E[\left|U'_{npi}\right|^2]\approx b_p\tau_\infty$ and $E[\left|U'_{npi}\right|^4]\le Kb_p^2$ by (3.13).

Step 4 From the previous two expressions, we see that Lyapunov's conditions are met for $\delta=2$. Therefore, Lindeberg condition is met and we can appeal to Theorem 3.8.2(proof can be seen in Theorem 27.2, Billingsley (1995)) for the sequence U'_{np1},...,U'_{npr_p} which show the normal convergence of a triangular array and we have the needed result. Finally applying Polya's theorem

$$\sup_{x\in\Re}\left|P(\sqrt{p}[F_n - p^{-1}\sum_{k=1}^{p}\mathrm{E}(F_{nk})] \leq x) - \Phi(x/\tau_\infty)\right| = o(1)$$

$$\Rightarrow \sup_{x\in\Re}\left|P(F_n - p^{-1}\sum_{k=1}^{p}\mathrm{E}(F_{nk}) \leq x) - \Phi(\sqrt{p}x/\tau_\infty)\right| = o(1)$$

$$\Rightarrow \sup_{x\in\Re}\left|P(F_n - a_n \leq x) - \Phi(\sqrt{p}[x - n^{-1}b_n]/\tau_\infty)\right| = o(1)$$

where a_n and b_n are asymptotically bounded sequences and

$$p^{-1}\sum_{k=1}^{p}\mathrm{E}(F_{nk}) = a_n + n^{-1}b_n + o(n^{-3/2}).$$

□

Proof of Proposition 3.4.1

Proof. Let's prove it for a fixed j and this is valid for any j.

We will assume WLOG that $\mathrm{E}[X_{1jk}] = \mathrm{E}[X_{2jk}] = ... = \mathrm{E}[X_{ajk}] = 0$.

Let

$$\Delta_{nk} = \sum_{i=1}^{a}\frac{(n_i-1)}{(n-a)}(s_{ik}^2 - \sigma_{ik}^2)$$

and

$$\tau_k^{-2} = (\sum_{i=1}^{a}\frac{(n_i-1)}{(n-a)}\sigma_{ik}^2)^{-1}$$

then F_{nk} can be approximated using Taylor's expansion for the denominator by

$$\tilde{F}_{nk} = (\overline{\boldsymbol{X}_k}' B\overline{\boldsymbol{X}}_k/(a-1))(\tau_k^{-2} - \tau_k^{-4}\Delta_{nk} + \tau_k^{-6}\Delta_{nk}^2)$$

so

$$F_{nk} - \tilde{F}_{nk} = o(n^{-\frac{3}{2}}). \tag{3.18}$$

To approximate the expected value otherwise complicated to find, $E[\tilde{F}_{nk}]$ is calculated in the following Subsection 3.8. Proof for (3.18) is shown next.

Let

$$F_{nk} = \frac{MST_k}{MSE_k} = (\overline{\boldsymbol{X}}'_k B \overline{\boldsymbol{X}}_k/(a-1))(\sum_{i=1}^{a} \frac{(n_i-1)s_{ik}^2}{n-a})^{-1}.$$

Asymptotic expansion will be performed by developing a Taylor's series up to the third term. We will call $f(x) = x^{-1}$ and we have

$$\sum_{i=1}^{a} \frac{(n_i-1)s_{ik}^2}{n-a} = x$$

from sample that we want to evaluate at

$$\sum_{i=1}^{a} \frac{(n_i-1)\sigma_{ik}^2}{n-a} = x_0$$

from population.

Then, developing Taylor's expansion, we will initially look at the first three elements of this expansion, hence

$$\begin{aligned} f(\sum_{i=1}^{a} \frac{(n_i-1)s_{ik}^2}{n-a}) =& (\sum_{i=1}^{a} \frac{(n_i-1)\sigma_{ik}^2}{n-a})^{-1} \\ & - (\sum_{i=1}^{a} \frac{(n_i-1)\sigma_{ik}^2}{n-a})^{-2} (\sum_{i=1}^{a} \frac{(n_i-1)(s_{ik}^2-\sigma_{ik}^2)}{n-a}) \\ & + (\sum_{i=1}^{a} \frac{(n_i-1)\sigma_{ik}^2}{n-a})^{-3} (\sum_{i=1}^{a} \frac{(n_i-1)(s_{ik}^2-\sigma_{ik}^2)}{n-a})^2 + R_2(\sum_{i=1}^{a} \frac{(n_i-1)s_{ik}^2}{n-a}) \end{aligned}$$

where

$$R_2(\sum_{i=1}^{a} \frac{(n_i-1)s_{ik}^2}{n-a}) = \frac{f^3(\xi_L)}{3!}(\sum_{i=1}^{a} \frac{(n_i-1)}{(n-a)}(s_{ik}^2-\sigma_{ik}^2))^3$$

is the Lagrange remainder and

$$\xi_L \in (\sigma_{ik}^2, s_{ik}^2).$$

Then,

$$\begin{aligned} R_2(\sum_{i=1}^{a} \frac{(n_i-1)s_{ik}^2}{n-a}) =& \frac{f^3(\xi_L)}{3!}(\sum_{i=1}^{a} \frac{(n_i-1)}{(n-a)}(s_{ik}^2-\sigma_{ij}^2))^3 \\ =& \frac{f^3(\xi_L)}{3!}(\sum_{i=1}^{a} O(n^{-\frac{1}{2}}))(\sum_{i=1}^{a} O(n^{-\frac{1}{2}}))(\sum_{i=1}^{a} O(n^{-\frac{1}{2}})) \\ =& \frac{f^3(\xi_L)}{3!}(O(n^{-\frac{1}{2}}))(O(n^{-\frac{1}{2}}))(O(n^{-\frac{1}{2}})) \\ =& O(1)O(n^{-\frac{1}{2}})O(n^{-\frac{1}{2}})O(n^{-\frac{1}{2}}) = O(n^{-\frac{3}{2}}). \end{aligned}$$

$\square$

Calculations of c_{nk}'s and d_{nk}'s

We will find the values of c_{nk} and d_{nk} that are used in Proposition 3.4.1. We will look at one variable of the vector so we will ignore the corresponding index since result is the same for all vector variables.

Gregory et al. (2015) approached the calculation of this moments using cumulant properties as in Leonov and Shiryaev (1959). Calculations will be made from definitions and other properties rather than cumulants in this case. Please note that an extra index for the variables and moments should be added to agree with previous notation.

$$\begin{aligned}
\mathrm{E}[F_n] =& \mathrm{E}[\frac{MST}{MSE}] \\
=& \mathrm{E}[(\overline{\boldsymbol{X}}'B\overline{\boldsymbol{X}}/(a-1))(\sum_{i=1}^{a}\frac{(n_i-1)S_i^2}{n-a})^{-1}] \\
\approx& \mathrm{E}[(\overline{\boldsymbol{X}}'B\overline{\boldsymbol{X}}/(a-1))\tau^{-2}] + \mathrm{E}[(\overline{\boldsymbol{X}}'B\overline{\boldsymbol{X}}/(a-1))\tau^{-4}\Delta_n] \\
&+ \mathrm{E}[(\overline{\boldsymbol{X}}'B\overline{\boldsymbol{X}}/(a-1))\tau^{-6}\Delta_n^2] \\
=& T_1 + T_2 + T_3.
\end{aligned}$$

The expected value of each one of the three terms is calculated separately.

$$T_1 = \mathrm{E}[(\overline{\boldsymbol{X}}'B\overline{\boldsymbol{X}}/(a-1))\tau^{-2}] = \frac{\tau^{-2}}{a-1}\mathrm{E}[\sum_{i=1}^{a}\sum_{j=1}^{a} b_{ij}\overline{X_i X_j}] = \frac{\tau^{-2}}{a-1}\mathrm{E}[\sum_{i=1}^{a} b_{ii}\overline{X_i X_i}]$$

b_{ij} are elements from (3.6). By independence between groups and within groups,the expression simplifies to

$$\begin{aligned}
T_1 =& \frac{\tau^{-2}}{a-1}\mathrm{E}[\sum_{i=1}^{a} b_{ii}\overline{X_i X_i}] = \frac{\tau^{-2}}{a-1}\sum_{i=1}^{a} b_{ii}\mathrm{E}[\overline{X_i X_i}] = \frac{\tau^{-2}}{a-1}\sum_{i=1}^{a}(n_i - \frac{n_i^2}{n})\frac{\sigma_i^2}{n_i} \\
=& \frac{\tau^{-2}}{a-1}\sum_{i=1}^{a}(1-\frac{n_i}{n})\sigma_i^2
\end{aligned}$$

equality follows from definition of b_{ij} and expected value.

Once T_1 is calculated,

$$\begin{aligned}T_2 =& \tau^{-4}\mathrm{E}[(\overline{\boldsymbol{X}}'B\overline{\boldsymbol{X}}/(a-1))\Delta_n] = \tau^{-4}\mathrm{E}[\overline{\boldsymbol{X}}'B\overline{\boldsymbol{X}}/(a-1))\sum_{i=1}^{a}\frac{(n_i-1)}{(n-a)}(S_i^2-\sigma_i^2)]\\ =& \frac{\tau^{-4}}{(a-1)(n-a)}\mathrm{E}[\overline{\boldsymbol{X}}'B\overline{\boldsymbol{X}})\sum_{i=1}^{a}(n_i-1)(S_i^2-\sigma_i^2)]\\ =& \frac{\tau^{-4}}{(a-1)(n-a)}(\mathrm{E}[(\overline{\boldsymbol{X}}'B\overline{\boldsymbol{X}})\sum_{i=1}^{a}(n_i-1)S_i^2] - \mathrm{E}[\overline{\boldsymbol{X}}'B\overline{\boldsymbol{X}})\sum_{i=1}^{a}(n_i-1)\sigma_i^2])\end{aligned}$$

developing matrix from the first element

$$T_2 = \frac{\tau^{-4}}{(a-1)(n-a)}\mathrm{E}[\sum_{j=1}^{a}\sum_{k=1}^{a}\sum_{i=1}^{a} b_{jk}\overline{X}_j\overline{X}_k(n_i-1)S_i^2] - \mathrm{E}[\overline{\boldsymbol{X}}'B\overline{\boldsymbol{X}})\sum_{i=1}^{a}(n_i-1)\sigma_i^2])$$

distinguishing between cross product of same index and different index

$$\begin{aligned}T_2 =& \frac{\tau^{-4}}{(a-1)(n-a)}(\sum_{i\neq j}(n_j-\frac{n_j^2}{n})(n_i-1)\mathrm{E}[\overline{X}_j^2S_i^2] + \sum_{i=1}^{a}(n_i-\frac{n_i^2}{n})(n_i-1)\mathrm{E}[\overline{X}_i^2S_i^2]\\ &-\mathrm{E}[(\overline{\boldsymbol{X}}'B\overline{\boldsymbol{X}})\sum_{i=1}^{a}(n_i-1)\sigma_i^2])\end{aligned}$$

substituting by different index cross product extected values

$$\begin{aligned}T_2 =& \frac{\tau^{-4}}{(a-1)(n-a)}(\sum_{i\neq j}(n_j-\frac{n_j^2}{n})(n_i-1)\frac{1}{n_j}\sigma_j^2\sigma_i^2\\ &+\sum_{i=1}^{a}(n_i-\frac{n_i^2}{n})(n_i-1)\mathrm{E}[\overline{X}_i^2S_i^2] - \sum_{j=1}^{a}(1-\frac{n_j}{n})\sigma_j^2\sum_{i=1}^{a}(n_i-1)\sigma_i^2)\end{aligned}$$

developing $\mathrm{E}[\overline{X}_i^2S_i^2]$

$$\begin{aligned}T_2 =& \frac{\tau^{-4}}{(a-1)(n-a)}(\sum_{i\neq j}(n_j-\frac{n_j^2}{n})(n_i-1)\frac{1}{n_j}\sigma_j^2\sigma_i^2\\ &+\sum_{i=1}^{a}(n_i-\frac{n_i^2}{n})(n_i-1)(\frac{1}{n_i^2(n_i-1)}(\sum_{j=1}^{n_i}\sum_{k=1}^{n_i}\sum_{l=1}^{n_i}\mathrm{E}[X_{ij}X_{ik}X_{il}^2]\\ &-\frac{1}{n_i}\sum_{j=1}^{n_i}\sum_{k=1}^{n_i}\sum_{l=1}^{n_i}\sum_{m=1}^{n_i}\mathrm{E}[X_{ij}X_{ik}X_{il}X_{im}])) - \sum_{j=1}^{a}(1-\frac{n_j}{n})\sigma_j^2\sum_{i=1}^{a}(n_i-1)\sigma_i^2)\end{aligned}$$

using again independence between groups and within groups

$$\begin{aligned}T_2 =&\frac{\tau^{-4}}{(a-1)(n-a)}(\sum_{i\neq j}(n_j - \frac{n_j^2}{n})(n_i-1)\frac{1}{n_j}\sigma_j^2\sigma_i^2\\ &+\sum_{i=1}^{a}(n_i-\frac{n_i^2}{n})(n_i-1)(\frac{1}{n_i^2(n_i-1)}(n_i\mathrm{E}[X_{i1}^4]+n_i(n_i-1)\mathrm{E}[X_{i1}^2]^2-\mathrm{E}[X_{i1}^4]\\ &-3(n_i-1)\mathrm{E}[X_{i1}^2]^2))-\sum_{j=1}^{a}(1-\frac{n_j}{n})\sigma_j^2\sum_{i=1}^{a}(n_i-1)\sigma_i^2)\end{aligned}$$

from definition of moments, results

$$\begin{aligned}T_2 =&\frac{\tau^{-4}}{(a-1)(n-a)}(\sum_{i\neq j}(n_j - \frac{n_j^2}{n})(n_i-1)\frac{1}{n_j}\sigma_j^2\sigma_i^2\\ &+\sum_{i=1}^{a}(n_i-\frac{n_i^2}{n})(n_i-1)(\mu_i^{(4)}/n_i^2+\frac{n_i-3}{n_i^2}\sigma_i^4)-\sum_{j=1}^{a}(1-\frac{n_j}{n})\sigma_j^2\sum_{i=1}^{a}(n_i-1)\sigma_i^2)\\ =&\frac{\tau^{-4}}{(a-1)(n-a)}(\sum_{i\neq j}(1-\frac{n_j}{n})(n_i-1)\sigma_j^2\sigma_i^2\\ &+\sum_{i=1}^{a}(1-\frac{n_i}{n})(n_i-1)(\mu_i^{(4)}/n_i+\frac{n_i-3}{n_i}\sigma_i^4)-\sum_{j=1}^{a}(1-\frac{n_j}{n})\sigma_j^2\sum_{i=1}^{a}(n_i-1)\sigma_i^2).\end{aligned}$$

With T_1 and T_2 calculated, now T_3 will be calculated. In this term the same arguments than in previous terms are used. The only difference is the number of cross product sums is increased.

$$\begin{aligned}T_3 =&\mathrm{E}[(\overline{\boldsymbol{X}}'B\overline{\boldsymbol{X}}/(a-1))\tau^{-6}\Delta_n^2]\\ =&\frac{\tau^{-6}}{(a-1)(n-a)^2}\mathrm{E}[(\overline{\boldsymbol{X}}'B\overline{\boldsymbol{X}})(\sum_{i=1}^{a}(n_i-1)(S_i^2-\sigma_i^2))^2]\end{aligned}$$

developing the square

$$T_3 = \frac{\tau^{-6}}{(a-1)(n-a)^2}\mathrm{E}[(\overline{\boldsymbol{X}}'B\overline{\boldsymbol{X}})(\sum_{i=1}^{a}(n_i-1)(S_i^2-\sigma_i^2))(\sum_{j=1}^{a}(n_j-1)(S_j^2-\sigma_j^2))]$$

factoring

$$T_3 =\frac{\tau^{-6}}{(a-1)(n-a)^2}\mathrm{E}[(\overline{\boldsymbol{X}}'B\overline{\boldsymbol{X}})(\sum_{i=1}^{a}\sum_{j=1}^{a}(n_i-1)(S_i^2-\sigma_i^2)(n_j-1)(S_j^2-\sigma_j^2))]$$

developing the product

$$T_3 =\frac{\tau^{-6}}{(a-1)(n-a)^2}\mathrm{E}[(\overline{\boldsymbol{X}}'B\overline{\boldsymbol{X}})(\sum_{i=1}^{a}\sum_{j=1}^{a}(n_i-1)(n_j-1)(S_i^2S_j^2-S_i^2\sigma_j^2-\sigma_i^2S_j^2+\sigma_i^2\sigma_j^2)]$$

separating independent cross products from dependent cross products

$$\begin{aligned}T_3 =&\frac{\tau^{-6}}{(a-1)(n-a)^2}\mathrm{E}[(\overline{\boldsymbol{X}}'B\overline{\boldsymbol{X}})(\sum_{i=1}^{a}(n_i-1)^2(S_i^4-2S_i^2\sigma_i^2+\sigma_i^4)\\ &+\sum_{j\neq i}^{a}(n_i-1)(n_j-1)(S_i^2S_j^2-S_i^2\sigma_j^2-\sigma_i^2S_j^2+\sigma_i^2\sigma_j^2))]\end{aligned}$$

developing the first quadratic form

$$\begin{aligned}T_3 =&\frac{\tau^{-6}}{(a-1)(n-a)^2}(\mathrm{E}[(\overline{\boldsymbol{X}}'B\overline{\boldsymbol{X}})(\sum_{i=1}^{a}(n_i-1)^2(S_i^4-2S_i^2\sigma_i^2+\sigma_i^4))]\\ &+\mathrm{E}[(\overline{\boldsymbol{X}}'B\overline{\boldsymbol{X}})(\sum_{j\neq i}^{a}(n_i-1)(n_j-1)(S_i^2S_j^2-S_i^2\sigma_j^2-\sigma_i^2S_j^2+\sigma_i^2\sigma_j^2))])\end{aligned}$$

factoring out sums

$$\begin{aligned}T_3 =&\frac{\tau^{-6}}{(a-1)(n-a)^2}(\mathrm{E}[\sum_{j=1}^{a}\sum_{k=1}^{a}\sum_{i=1}^{a}(n_i-1)^2b_{jk}\overline{X_jX_k}(S_i^4-2S_i^2\sigma_i^2+\sigma_i^4)]\\ &+\mathrm{E}[(\sum_{l=1}^{a}\sum_{k=1}^{a}\sum_{j\neq i}^{a}(n_i-1)(n_j-1)b_{lk}\overline{X_lX_k}(S_i^2S_j^2-S_i^2\sigma_j^2-\sigma_i^2S_j^2+\sigma_i^2\sigma_j^2))])\end{aligned}$$

separating independent cross products from dependent cross products in the second term

$$\begin{aligned}T_3 =&\frac{\tau^{-6}}{(a-1)(n-a)^2}(\mathrm{E}[\sum_{j=1}^{a}\sum_{k=1}^{a}\sum_{i=1}^{a}(n_i-1)^2b_{jk}\overline{X_jX_k}(S_i^4-2S_i^2\sigma_i^2+\sigma_i^4)]\\ &+\mathrm{E}[(\sum_{l=1}^{a}\sum_{j\neq i}^{a}(n_i-1)(n_j-1)b_{ll}\overline{X_l^2}(S_i^2S_j^2-S_i^2\sigma_j^2-\sigma_i^2S_j^2+\sigma_i^2\sigma_j^2))]\\ &+2E[\sum_{j\neq i}^{a}(n_i-1)(n_j-1)b_{ij}\overline{X_iX_j}(S_i^2S_j^2))])\end{aligned}$$

doing the same for the first term

$$\begin{aligned}T_3 =&\frac{\tau^{-6}}{(a-1)(n-a)^2}(\mathrm{E}[\sum_{j=1}^{a}\sum_{k=1}^{a}\sum_{i=1}^{a}(n_i-1)^2b_{jk}\overline{X_jX_k}(S_i^4-2S_i^2\sigma_i^2+\sigma_i^4)]\\ &+\mathrm{E}[(\sum_{l\neq j\neq i}^{a}(n_i-1)(n_j-1)b_{ll}\overline{X_l^2}(S_i^2S_j^2-S_i^2\sigma_j^2-\sigma_i^2S_j^2+\sigma_i^2\sigma_j^2))]\\ &+2E[(\sum_{j\neq i}^{a}(n_i-1)(n_j-1)b_{ii}\overline{X_i^2}(S_i^2S_j^2-S_i^2\sigma_j^2-\sigma_i^2S_j^2+\sigma_i^2\sigma_j^2))]\\ &+2E[\sum_{j\neq i}^{a}(n_i-1)(n_j-1)b_{ij}\overline{X_iX_j}(S_i^2S_j^2))])\end{aligned}$$

substituting b_{ij}

$$\begin{aligned} T_3 =& \frac{\tau^{-6}}{(a-1)(n-a)^2}(\mathrm{E}[\sum_{j=1}^{a}\sum_{i=1}^{a}(n_i-1)^2(n_j-\frac{n_j^2}{n})\overline{X_j^2}(S_i^4-2S_i^2\sigma_i^2+\sigma_i^4)] \\ &+\mathrm{E}[(\sum_{l\neq j\neq i}^{a}(n_i-1)(n_j-1)(n_l-\frac{n_l^2}{n})\overline{X_l^2}(S_i^2S_j^2-S_i^2\sigma_j^2-\sigma_i^2S_j^2+\sigma_i^2\sigma_j^2))] \\ &+2\mathrm{E}[(\sum_{j\neq i}^{a}(n_i-1)(n_j-1)(n_i-\frac{n_i^2}{n})\overline{X_i^2}(S_i^2S_j^2-S_i^2\sigma_j^2-\sigma_i^2S_j^2+\sigma_i^2\sigma_j^2))] \\ &+2\mathrm{E}[\sum_{j\neq i}^{a}(n_i-1)(n_j-1)(-\frac{n_in_j}{n}\overline{X_iX_j}(S_i^2S_j^2))]]) \end{aligned}$$

calculating expected values of independent cross products

$$\begin{aligned} T_3 =& \frac{\tau^{-6}}{(a-1)(n-a)^2}(\mathrm{E}[\sum_{j=1}^{a}\sum_{i=1}^{a}(n_i-1)^2(n_j-\frac{n_j^2}{n})\overline{X_j^2}(S_i^4-2S_i^2\sigma_i^2+\sigma_i^4)] \\ &+(\sum_{l\neq j\neq i}^{a}(n_i-1)(n_j-1)(n_l-\frac{n_l^2}{n})\frac{\sigma_l^2}{n_l}(\sigma_i^2\sigma_j^2-\sigma_i^2\sigma_j^2-\sigma_i^2\sigma_j^2+\sigma_i^2\sigma_j^2)) \\ &+2\mathrm{E}[(\sum_{j\neq i}^{a}(n_i-1)(n_j-1)(n_i-\frac{n_i^2}{n})\overline{X_i^2}(S_i^2S_j^2-S_i^2\sigma_j^2-\sigma_i^2S_j^2+\sigma_i^2\sigma_j^2)] \\ &+2\mathrm{E}[\sum_{j\neq i}^{a}(n_i-1)(n_j-1)(-\frac{n_in_j}{n}\overline{X_iX_j}(S_i^2S_j^2))]]) \end{aligned}$$

$$\begin{aligned}
T_3 =&\frac{\tau^{-6}}{(a-1)(n-a)^2}(\mathrm{E}[\sum_{j=1}^{a}\sum_{i=1}^{a}(n_i-1)^2(n_j-\frac{n_j^2}{n})\overline{X_j^2}(S_i^4-2S_i^2\sigma_i^2+\sigma_i^4)] \\
&+2\mathrm{E}[(\sum_{j\neq i}^{a}(n_i-1)(n_j-1)(n_i-\frac{n_i^2}{n})\overline{X_i^2}(S_i^2\sigma_j^2-S_i^2\sigma_j^2-\sigma_i^2\sigma_j^2+\sigma_i^2\sigma_j^2)] \\
&+2\mathrm{E}[\sum_{j\neq i}^{a}(n_i-1)(n_j-1)(-\frac{n_in_j}{n}\overline{X_iX_j}(S_i^2S_j^2))]]) \\
=&\frac{\tau^{-6}}{(a-1)(n-a)^2}(\mathrm{E}[\sum_{j=1}^{a}\sum_{i=1}^{a}(n_i-1)^2(n_j-\frac{n_j^2}{n})\overline{X_j^2}(S_i^4-2S_i^2\sigma_i^2+\sigma_i^4))] \\
&+2\sum_{j\neq i}^{a}(n_i-1)(n_j-1)(-\frac{n_in_j}{n}\mathrm{E}[\overline{X_i}S_i^2]\mathrm{E}[\overline{X_j}S_j^2]) \\
=&\frac{\tau^{-6}}{(a-1)(n-a)^2}(\mathrm{E}[\sum_{j=1}^{a}\sum_{i=1}^{a}(n_i-1)^2(n_j-\frac{n_j^2}{n})\overline{X_j^2}(S_i^4-2S_i^2\sigma_i^2+\sigma_i^4)] \\
&+2\sum_{j\neq i}^{a}(n_i-1)(n_j-1)(-\frac{n_in_j}{n}(\frac{1}{n_i}\mu_i^{(3)}\frac{1}{n_j}\mu_j^{(3)})) \\
=&\frac{\tau^{-6}}{(a-1)(n-a)^2}(\mathrm{E}[\sum_{i\neq j}(n_i-1)^2(n_j-\frac{n_j^2}{n})\overline{X_j^2}(S_i^4-2S_i^2\sigma_i^2+\sigma_i^4)] \\
&+\mathrm{E}[\sum_{i=1}^{a}(n_i-1)^2(n_i-\frac{n_i^2}{n})\overline{X_i^2}(S_i^4-2S_i^2\sigma_i^2+\sigma_i^4))] \\
&+2\sum_{j\neq i}^{a}(n_i-1)(n_j-1)(-\frac{n_in_j}{n}(\frac{1}{n_i}\mu_i^{(3)}\frac{1}{n_j}\mu_j^{(3)})) \\
=&\frac{\tau^{-6}}{(a-1)(n-a)^2}(\mathrm{E}[\sum_{i\neq j}(n_i-1)^2(n_j-\frac{n_j^2}{n})\frac{\sigma_j^2}{n_j}(S_i^4-2S_i^2\sigma_i^2+\sigma_i^4)] \\
&+\mathrm{E}[\sum_{i=1}^{a}(n_i-1)^2(n_i-\frac{n_i^2}{n})\overline{X_i^2}(S_i^4-2S_i^2\sigma_i^2+\sigma_i^4))] \\
&+2\sum_{j\neq i}^{a}(n_i-1)(n_j-1)(-\frac{n_in_j}{n}(\frac{1}{n_i}\mu_i^{(3)}\frac{1}{n_j}\mu_j^{(3)})) \\
=&\frac{\tau^{-6}}{(a-1)(n-a)^2}(\mathrm{E}[\sum_{i\neq j}(n_i-1)^2(n_j-\frac{n_j^2}{n})\frac{\sigma_j^2}{n_j}(S_i^4-2\sigma_i^4+\sigma_i^4)] \\
&+\mathrm{E}[\sum_{i=1}^{a}(n_i-1)^2(n_i-\frac{n_i^2}{n})\overline{X_i^2}(S_i^4-2S_i^2\sigma_i^2+\sigma_i^4))] \\
&+2\sum_{j\neq i}^{a}(n_i-1)(n_j-1)(-\frac{n_in_j}{n}(\frac{1}{n_i}\mu_i^{3}\frac{1}{n_j}\mu_j^{3}))
\end{aligned}$$

$$
\begin{aligned}
T_3 =&\frac{\tau^{-6}}{(a-1)(n-a)^2}(\sum_{i\neq j}(n_i-1)^2(n_j-\frac{n_j^2}{n})\frac{\sigma_j^2}{n_j}((\frac{n_i^2-2n_i+6}{(n_i-1)^2n_i})\mu_i^{(4)} \\
&+(\frac{n_i^2-2n_i+6}{n_i(n_i-1)})\sigma_i^4)-\sigma_i^4)+\mathrm{E}[\sum_{i=1}^{a}(n_i-1)^2(n_i-\frac{n_i^2}{n})\overline{X_i^2}(S_i^4-2S_i^2\sigma_i^2+\sigma_i^4))] \\
&+2\sum_{j\neq i}^{a}(n_i-1)(n_j-1)(-\frac{n_in_j}{n}(\frac{1}{n_i}\mu_i^{(3)}\frac{1}{n_j}\mu_j^{(3)})) \\
=&\frac{\tau^{-6}}{(a-1)(n-a)^2}(\sum_{i\neq j}(n_i-1)^2(1-\frac{n_j}{n})\sigma_j^2((\frac{n_i^2-2n_i+6}{(n_i-1)^2n_i})\mu_i^{(4)} \\
&+(\frac{n_i^2-2n_i+6}{n_i(n_i-1)})\sigma_i^4-\sigma_i^4)+\mathrm{E}[\sum_{i=1}^{a}(n_i-1)^2(n_i-\frac{n_i^2}{n})\overline{X_i^2}(S_i^4-2S_i^2\sigma_i^2+\sigma_i^4))] \\
&+2\sum_{j\neq i}^{a}(n_i-1)(n_j-1)(-\frac{1}{n}(\mu_i^{(3)}\mu_j^{(3)})) \\
=&\frac{\tau^{-6}}{(a-1)(n-a)^2}(\sum_{i\neq j}(n_i-1)^2(1-\frac{n_j}{n})\sigma_j^2((\frac{n_i^2-2n_i+6}{(n_i-1)^2n_i})\mu_i^{(4)}+(\frac{-n_i+6}{n_i(n_i-1)})\sigma_i^4) \\
&+\sum_{i=1}^{a}(n_i-1)^2(1-\frac{n_i}{n})(\frac{1}{n_i^2}\mu_i^{(6)}+\frac{n_i^2(n_i-2)-6n_i(n_i-2)+15(n_i-2)}{n_i^2(n_i-1)}(\sigma_i^2)^3 \\
&+\frac{2n_i^2-8n_i+10}{n_i^2(n_i-1)}(\mu_i^{(3)})^2+\frac{3n_i^2-14n_i+15}{n_i^2(n_i-1)}\mu_i^{(4)}\sigma_i^2) \\
&-2(\sum_{i=1}^{a}(n_i-1)^2(1-\frac{n_i}{n})((\mu_i^{(4)}/n_i+\frac{n_i-3}{n_i}\sigma_i^4)\sigma_i^2) \\
&+\sum_{i=1}^{a}(n_i-1)^2(1-\frac{n_i}{n})\sigma_i^6))] \\
&+2\sum_{j\neq i}^{a}(n_i-1)(n_j-1)(-\frac{1}{n}(\mu_i^{(3)}\mu_j^{(3)})).
\end{aligned}
$$

So, then putting all terms together:

$$
\begin{aligned}
\mathrm{E}[F] =&\mathrm{E}[\frac{MST}{MSE}] \\
=&\mathrm{E}[(\overline{\boldsymbol{X}}'B\overline{\boldsymbol{X}}/(a-1))(\sum_{i=1}^{a}\frac{(n_i-1)S_i^2}{n-a})^{-1}] \\
\approx&\mathrm{E}[(\overline{\boldsymbol{X}}'B\overline{\boldsymbol{X}}/(a-1))\tau^{-2}]+\mathrm{E}[(\overline{\boldsymbol{X}}'B\overline{\boldsymbol{X}}/(a-1))\tau^{-4}\Delta_n]+ \\
&\mathrm{E}[(\overline{\boldsymbol{X}}'B\overline{\boldsymbol{X}}/(a-1))\tau^{-6}\Delta_n^2] \\
=&T_1+T_2+T_3
\end{aligned}
$$

$$\begin{aligned}
\mathrm{E}[F] =&\frac{\tau^{-2}}{a-1}\sum_{i=1}^{a}(1-\frac{n_i}{n})\sigma_i^2 \\
&-\frac{\tau^{-4}}{(a-1)(n-a)}(\sum_{i\neq j}(1-\frac{n_j}{n})(n_i-1)\sigma_j^2\sigma_i^2+\sum_{i=1}^{a}(1-\frac{n_i}{n})(n_i-1)(\mu_i^{(4)}/n_i+\frac{n_i-3}{n_i}\sigma_i^4) \\
&-\sum_{j=1}^{a}(1-\frac{n_j}{n})\sigma_j^2\sum_{i=1}^{a}(n_i-1)\sigma_i^2) \\
&+\frac{\tau^{-6}}{(a-1)(n-a)^2}(\sum_{i\neq j}(n_i-1)^2(1-\frac{n_j}{n})\sigma_j^2((\frac{n_i^2-2n_i+6}{(n_i-1)^2n_i})\mu_i^{(4)}+(\frac{-n_i+6}{n_i(n_i-1)})\sigma_i^4) \\
&+\sum_{i=1}^{a}(n_i-1)^2(1-\frac{n_i}{n})(\frac{1}{n_i^2}\mu_i^{(6)}+\frac{n_i^2(n_i-2)-6n_i(n_i-2)+15(n_i-2)}{n_i^2(n_i-1)}(\sigma_i^2)^3 \\
&+\frac{2n_i^2-8n_i+10}{n_i^2(n_i-1)}(\mu_i^{(3)})^2+\frac{3n_i^2-14n_i+15}{n_i^2(n_i-1)}\mu_i^{(4)}\sigma_i^2) \\
&-2(\sum_{i=1}^{a}(n_i-1)^2(1-\frac{n_i}{n})((\mu_i^{(4)}/n_i+\frac{n_i-3}{n_i}\sigma_i^4)\sigma_i^2)+\sum_{i=1}^{a}(n_i-1)^2(1-\frac{n_i}{n})\sigma_i^6))] \\
&+2\sum_{j\neq i}^{a}(n_i-1)(n_j-1)(-\frac{1}{n}(\mu_i^{(3)}\mu_j^{(3)})).
\end{aligned}$$

Given the expression we are going to call

$$c_n =\frac{\tau^{-2}}{a-1}\sum_{i=1}^{a}(1-\frac{n_i}{n})\sigma_i^2 \tag{3.19}$$

$$\begin{aligned}
d_n =&-\frac{\tau^{-4}}{(a-1)(n-a)}(\sum_{i\neq j}(1-\frac{n_j}{n})(n_i-1)\sigma_j^2\sigma_i^2 \\
&+\sum_{i=1}^{a}(1-\frac{n_i}{n})(n_i-1)(\mu_i^{(4)}/n_i+\frac{n_i-3}{n_i}\sigma_i^4)-\sum_{j=1}^{a}(1-\frac{n_j}{n})\sigma_j^2\sum_{i=1}^{a}(n_i-1)\sigma_i^2) \\
&+\frac{\tau^{-6}}{(a-1)(n-a)^2}(\sum_{i\neq j}(n_i-1)^2(1-\frac{n_j}{n})\sigma_j^2((\frac{n_i^2-2n_i+6}{(n_i-1)^2n_i})\mu_i^{(4)}+(\frac{-n_i+6}{n_i(n_i-1)})\sigma_i^4) \\
&+\sum_{i=1}^{a}(n_i-1)^2(1-\frac{n_i}{n})(\frac{1}{n_i^2}\mu_i^{(6)}+\frac{n_i^2(n_i-2)-6n_i(n_i-2)+15(n_i-2)}{n_i^2(n_i-1)}(\sigma_i^2)^3 \\
&+\frac{2n_i^2-8n_i+10}{n_i^2(n_i-1)}(\mu_i^{(3)})^2+\frac{3n_i^2-14n_i+15}{n_i^2(n_i-1)}\mu_i^{(4)}\sigma_i^2) \\
&-2(\sum_{i=1}^{a}(n_i-1)^2(1-\frac{n_i}{n})((\mu_i^{(4)}/n_i+\frac{n_i-3}{n_i}\sigma_i^4)\sigma_i^2)+\sum_{i=1}^{a}(n_i-1)^2(1-\frac{n_i}{n})\sigma_i^6))] \\
&+2\sum_{j\neq i}^{a}(n_i-1)(n_j-1)(-\frac{1}{n}(\mu_i^{(3)}\mu_j^{(3)})).
\end{aligned} \tag{3.20}$$

Values for d'_{nk}

The statistic for the unpooled variance is calculated similarly to the previous. The calculation steps are omitted since the steps are very similar to the above section.

The moment calculation in this case is:

$$\begin{aligned}
\mathrm{E}[F'] =& \mathrm{E}[\frac{MST'}{MSE'}] = \mathrm{E}[(\overline{\boldsymbol{X}}' B \overline{\boldsymbol{X}}/(a-1))(\sum_{i=1}^{a} \frac{(n_i-1)S_i^2}{n-a})^{-1}] \\
\approx& \mathrm{E}[(\overline{\boldsymbol{X}}' B \overline{\boldsymbol{X}}/(a-1))\tau^{-2}] + \mathrm{E}[(\overline{\boldsymbol{X}}' B \overline{\boldsymbol{X}}/(a-1))\tau^{-4}\Delta_n] \\
&+ \mathrm{E}[(\overline{\boldsymbol{X}}' B \overline{\boldsymbol{X}}/(a-1))\tau^{-6}\Delta_n^2] \\
=& 1 - ((\sum_{i=1}^{a} \frac{\sigma_i^2}{n_i})^{-2})(\sum_{i\neq j}^{a} \frac{\sigma_i^2}{n_i}\frac{\sigma_j^2}{n_j} + \sum_{i=1}^{a}(\mu_i^{(4)} + (n_i-3)\frac{(\sigma_i^2)^2}{n_i^3}) \\
&+ 1 + ((\sum_{i=1}^{a} \frac{\sigma_i^2}{n_i})^{-3})(\sum_{i=1}^{a}((n_i(n_i-1)(n_i-2)(\sigma_i^2)^3) + 2n_i(n_i-1)(\mu_i^{(3)})^2 \\
&+ 4n_i(n_i-1)\mu_i^{(4)}\sigma_i^2 + n_i\mu_i^{(6)}) + \frac{1}{n_i^2}(15n_i(n_i-1)(n_i-2)(\sigma_i^2)^3 \\
&+ 15n_i(n_i-1)(\sigma_i^2)^3 + 15n_i(n_i-1)\sigma_i^2\mu_i^{(4)} + 20n_i(n_i-1)(\mu_i^{(3)})^2 \\
&+ n_i\mu_i^6)\frac{1}{(n_i-1)^2 n_i^4}) - \frac{2}{a-1}(\sum_{i\neq j}^{a} \frac{\mu_i^{(3)}}{n_i^2}\frac{\mu_j^{(3)}}{n_j^2}) + 2(\sum_{i\neq j}^{a} \frac{(n_i-3)(\sigma_i^2)^2 + \mu_i^{(4)}}{n_i^3}\frac{\sigma_j^2}{n_j}) \\
&+ \sum_{i\neq j}^{a}((\frac{\sigma_i^2}{n_i})\frac{1}{(n_j-1)n_j^3}((n_j^2 - 2n_j + 3)(\sigma_j^2)^2 + (n_j-1)\mu_j^{(4)})) \\
&+ \sum_{i\neq j\neq k}^{a} \frac{\sigma_i^2}{n_i}\frac{\sigma_j^2}{n_j}\frac{\sigma_k^2}{n_k}) - 2(\sum_{i=1}^{a} \frac{\sigma_i^2}{n_i})^{-2}(\sum_{i\neq j}^{a} \frac{\sigma_i^2}{n_i}\frac{\sigma_j^2}{n_j}) + \frac{1}{n_i^3}\sum_{i=1}^{a}(\mu_i^{(4)} + (n_i-3)(\sigma_i^2)^3) + 1).
\end{aligned}$$

Estimates for unequal covariance in this case $c'_n = 1$ and d'_n is the remaining expression in the above equation.

Unbiased Sample Moments

In order to calculate the statistic the values of a_n, b_n and b'_n need to be estimated from the expressions with population moments. Calculations found the following estimated moments:

$$\widehat{\mu}_i = \overline{X}_i. \tag{3.21}$$

$$\widehat{\sigma}_i^2 = \frac{n_i}{n_i - 1} \sum_{j=1}^{n_i} (x_{ij} - \overline{X}_i)^2 / n_i. \tag{3.22}$$

$$\widehat{\mu}_i^{(3)} = \frac{n_i^2}{(n_i - 1)(n_i - 2)} \sum_{j=1}^{n_i} (x_{ij} - \overline{X}_i)^3 / n_i. \tag{3.23}$$

$$\widehat{\mu}_i^{(4)} = \frac{n_i^3}{(n_i^2 - 3n_i + 3)(n_i - 1)} \sum_{j=1}^{n_i} \frac{1}{n_i} (x_{ij} - \overline{X}_i)^4 - \frac{6n_i - 9}{(n_i^2 - 3n_i + 3)n_i} \sum_{j \neq k}^{n_i} x_{ij}^2 x_{ik}^2. \tag{3.24}$$

These moments are substituted in the expressions with population moments resulting in the final statistics.

Chapter 4 Nonparametric Method for Two Samples High Dimensional Tests

4.1 Introduction

Nonparametric methods are well known for being more robust against nonnormality and other general conditions than their parametric counterparts. In our interest to be as general as possible, we will introduce a nonparametric test statistic in this chapter for quantifying group or treatment differences. The core test statistic in Chapter 3 is a composite version of the square of student's t statistic. For a nonparametric test, a composite Wilcoxon-Mann-Whitney test (Brunner and Munzel, 2000) will be used.

Classical nonparametric tests formulate tests in terms of distribution functions rather than parameters. The challenges with this formulation are (a) alternative hypothesis is difficult to interpret (b) tests can not easily be inverter to construct confidence intervals. To overcome these challenges some characteristic of the distribution functions is often investigated to compare treatments. In this respect, we will use the concept of nonparametric relative group effect that we define in (4.1) to motivate the use of the univariate test from Brunner and Munzel (2000). The variable-by-variable univariate tests are combined to propose a multivariate composite test in the same way as in Chapter 3.

Let us first introduce the concept of relative group effect as it applies to the marginal distributions of the k^{th} variable in the i^{th} group. The random variable X_{ijk} where $i = 1, 2$, $j = 1, ..., n_i$ and $k = 1, ..., p$ is the k^{th} variable for the j^{th} subject from the i^{th} sample group. Suppose

$$X_{ijk} \sim F_{ik}$$

for $j = 1, ..., n_i$. The distribution functions $F_{ik}(x)$ are arbitrary non-degenerate distributions. In our investigation to compare group effects, we study the so-called nonparametric relative treatment effect. In the nonparametric literature, the notation tipically used for this quantity is p but to avoid confusion with the notation for the dimension, we denote this relative treatment effect by ω. In order to accomodate binary, ordered categorical, discrete and continuous data types in a unified manner we will use the normalized version of the distribution function, defined as

$$F_{ik}(x) = \frac{1}{2}\{F_{ik}^{+}(x) + F_{ik}^{-}(x)\} = P(X_{i1k} < x) + \frac{1}{2}P(X_{i1k} = x),$$

where $F_{ik}^{-}(x) = P(X_{i1k} < x)$ and $F_{ik}^{+}(x) = P(X_{i1k} \leq x)$ are the left and right continuous versions of the distibution function. The relative effect for the j^{th} variable is defined by

$$\omega_k = P(X_{11k} < X_{21k}) + \frac{1}{2}P(X_{11k} = X_{21k}). \tag{4.1}$$

The relative effect has the interpretation that if ω_k is greater than 1/2, observations on the k^{th} variable in the first sample tend to have smaller values than observations on the k^{th} variable in the second sample and viceversa if ω_k is smaller than 1/2. If $\omega_k = 1/2$ the two variables are tendentiously equal. For example, for Normal distribution functions, where F_{ik} has expectations μ_{ik} and variances σ_{ik}^2, it can be shown that $\mu_{1k} = \mu_{2k}$ if and only if $\omega_k = \frac{1}{2}$. Therefore, $\omega_k = 1/2$ does not necessarily imply that $F_{1k} = F_{2k}$. In some cases, it could contain a parametric hypothesis as a special case.

For convenience we express ω_k in terms of distribution functions. It can be easily shown that:

$$\omega_k = \int F_{1k} dF_{2k}.$$

In this chapter we consider a hypothesis testing about ω_k's in the high-dimensional asymptotic framework. To that end, the chapter will be organized in seven sections including this Introduction Section 4.1. Section 4.2 introduces the model and hypothesis of interest. We propose the test statistic in Section 4.3. In Section 4.4 asymptotic results for the test statistic are stated and the results are used to construct asymptotic tests. The finite sample performance of the tests is investigated via simulation study in Section 4.5. Finally, we will end the chapter with some conclusions in Section 4.6. All technical details are shifted to the Appendix 4.7.

4.2 Model and Hypothesis

Model can also be formulated in terms of independent random vectors. Let

$$\boldsymbol{X}_{ij} = (X_{ij1}, X_{ij2}, ..., X_{ijp})^\top$$

be independent random vectors for $i = 1, 2$ and $j = 1, ..., n_i$ with $n = n_1 + n_2$.

From the definition of the marginal relative group effects (see Section 4.1), the multivariate nonparametric effect of interest is:

$$\boldsymbol{\omega} = (\omega_1, ..., \omega_p)^\top.$$

The global hypothesis of interest is

$$H_0 : \boldsymbol{\omega} = \frac{1}{2}\mathbf{1}.$$

The statistic used to test this hypothesis is defined in Section 4.3 but to state assumptions, the univariate version of the statistic will be presented here. Let

$$W_{nk} = \frac{\sqrt{n}(\widehat{\omega}_k - \frac{1}{2})}{\widehat{\sigma}_{nk}}$$

where

$$\widehat{\omega}_k = \frac{1}{n_1}(\overline{R}_{2.k} - \frac{n_2+1}{2}), \quad \widehat{\sigma}^2_{nk} = n \cdot \left[\widehat{\sigma}^2_{1k}/n_1 + \widehat{\sigma}^2_{2k}/n_2\right], \tag{4.2}$$

$$\widehat{\sigma}^2_{ik} = S^2_{ik}/(n - n_i)^2 \text{ and } S^2_{ik} = \frac{1}{n_i - 1}\sum_{j=1}^{n_i}(R_{ijk} - R^{(i)}_{jk} - \overline{R}_{i.k} + \frac{n_1+1}{2})^2.$$

Here, R_{ijk} refers to the mid-rank of X_{ijk} among all values on the k^{th} variable in the two samples and $R^{(i)}_{jk}$ refers to the rank of X_{ijk} among the n_i observations of the k^{th} variable in the i^{th} sample. Further, $\overline{R}_{i.k}$ is the mean of the ranks in the i^{th} sample for the k^{th} variable (see also Section 4.7 for formal definitions).

The assumptions for this new statistic are stated analogously to those in Chapter 3. Recall,

$$\alpha_{ij}(s) = \sup_{k \geq 1}\{|P(A \cap B) - P(A)P(B)| : A \in \mathcal{F}^k_1(i,j) \text{ and } B \in \mathcal{F}^\infty_{k+s}(i,j)\},$$

where $\mathcal{F}^b_a(i,j) \equiv \sigma(\{X_{ijk} : a \leq k \leq b\})$. Here $\alpha_{ij}(s)$ is a dependence coefficient that measures the strength of dependence between two groups of variables that are at least s indices apart.

Assumption 4.2.1. *For some* $\delta \in (0, \infty), \sum_{s=1}^{\infty} \alpha_{ij}(s)^{\delta/(2+\delta)} < \infty$ *for* $i = 1, 2$ *and* $j = 1, ..., n_i$.

Assumption 4.2.2. *For some* $\delta \in (0, \infty), E|W_{nk}^2|^{2l+\delta} < b < \infty$ *for all* $k = 1, ..., p$ *for some integer* $l \geq 1$.

Assumption 4.2.3. $\lim_{n\to\infty} \frac{1}{p-s} \sum_{i=1}^{p-s} Cov(W_{nk}^2, W_{n(k+s)}^2) = \gamma(s)$ *exists* $\forall s > 0$.

Assumption 4.2.4. $\sup_{n\geq 1}\{E|W_{nk}|^{16}, k = 1, ..., p\} = O(1)$.

Assumption 4.2.5. $\inf_{n\geq 1}\{Var(W_{nk}), k = 1, ..., p\} > b > 0$.

The assumptions are also made in Chapter 3 and they are needed here for the same reasons.

4.3 Test Statistic

The test statistic will be built from the square of the statistic described in Section 4.2. More precisely,

$$W = \sqrt{p}\frac{W_n - 1}{\widehat{\zeta_n}} \tag{4.3}$$

with

$$W_n = \sum_{k=1}^{p} \frac{W_{nk}^2}{p} \tag{4.4}$$

where $\widehat{\zeta_n}$ is defined similarly to the definition given in Chapter 3 as

$$\widehat{\zeta_n}^2 \equiv \sum_{|s|<L} w(s/L)\widehat{\gamma(s)},$$

where

$$\widehat{\gamma}(s) = \frac{1}{p-s}\sum_{k=1}^{p-s}(W_{nk}^2 - W_n)(W_{n(k+s)}^2 - W_n)$$

and

$$w(x) = \begin{cases} 1 - 6|x|^2 + 6|x|^3 & \text{if } |x| < 1/2 \\ 2(1 - |x|)^3 & \text{if } 1/2 \leq x \leq 1 \\ 0 & \text{if } |x| > 1 \end{cases}$$

is the Parzen Smoothing Window (see Brockwell and Davis (2013)). As mentioned in Chapter 3, introducing this weight (i.e.$w(x)$) leads to a consistent estimator of the asymptotic variance under the assumed α-mixing structure.

4.4 Main Results

The theoretical results for this chapter can derived from Chapter 3. It needs to be proved that the univariate statistic preserves (inherits) the α-mixing condition from the samples. From Section 4.3 it can be seen (Subsection 4.7) that the test statistic W_{nk} is a Lebesgue function of the data values fro the k^{th} variable. By the same argument as in Chapter 3 (see argument in Step 1 from Theorem 3.8, also Bradley (2005)) α-mixing property transfers from the sequence $\{X_{ijk} : k = 1, 2, ...\}$ to the statistics W_{nk} for $k = 1, 2, ...$ To see this, note the statistics are in terms of mid-ranks. Ranks are derived from empirical CDF's, which are Lebesgue functions of independent random variables. Therefore $\{(W_{nk}^2, k = 1, 2, ...\}$ has the same α-mixing property as $\{X_{ijk}; k = 1, 2, ...\}$. By adjusting the Assumptions 4.2.1, 4.2.2, 4.2.3, 4.2.4 and 4.2.5 appropriately, results from Chapter 3 are applicable W_n.

Proposition 4.4.1. *Let us assume that $p \equiv p_n = o(n^2)$ and Assumptions 4.2.1, 4.2.2, 4.2.3, 4.2.4 and 4.2.5 hold with $s = 1$.*

Then,

$$\sup_{x\in\mathbb{R}} |P(W_n - 1 < x) - \Phi\{\sqrt{p}(x)/\tau_\infty\}| = o(1)$$

where $\tau_\infty^2 = \gamma(0) + 2\sum_{k=1}^{\infty} \gamma(k)$.

Proof. Conditions and assumptions are analogous to those in Theorem 3.4.1 except the difference in the definitions of how W_{nk}, W_n and W. By the arguments discussed in the beginning of this section, $\{(W_{nk}^2, k = 1, 2, ...\}$ has the same α-mixing property as $\{X_{ijk}; k = 1, 2, ...\}$. Therefore, the result is a consequence of Theorem 3.4.1. □

4.5 Simulation

The aim of this section is to show the performance, in terms of size and power, of the moderate-p version of the statistic from Gregory et al. (2015) (from now on also referred to as "GCT-mdp") and the nonparametric test proposed in this chapter (from now on also referred to as "VH-np"), in particular in the case of small sample sizes and large dimension. Since inference for two groups is of interest, we slightly modify the settings from those used in the simulations in Chapter 3.

In order to make the simulation as thorough as possible, we have investigated effects of diference in sample sizes n_i and dimension p under multiple scenarios for dependence, error distribution and hte parameter L. L is the size of the Parzen Smoothing Window needed for the estimation of the asymptotic variance. In the power simulation, δ quantifies the departure from null hypothesis. More precisely, the groups compared have means $\boldsymbol{\mu}_1 = \mathbf{0}_p$ and $\boldsymbol{\mu}_2 = \delta\mathbf{1}_p$.

Simulation Design

We compare the sizes of GCT-mdp and VH-np under the following settings:

- Sample sizes $(n_1, n_2) = \{(20, 18), (40, 38)\}$.
- Dimension $p \in \{300, 1000\}$.
- Size of the Parzen Smoothing Window $L \in \{10, 20\}$.
- Dependence model: Independence and ARMA models.
- Error distribution: GEV(0,1,0), Cauchy(0,3), N(0,1), centered Gamma(4,2), Uniform(-5,5) and Double exponential(0,1).

The ARMA model is as defined in Section 3.5. We used ARMA(2,2) model with coefficients $\varphi_1 = 0.8897$, $\varphi_2 = -0.4858$, $\theta_1 = -0.2279$ and $\theta_2 = 0.2488$. Given the nature of the statistic, we added two distributions to the list of distributions used in the simulation in Chapter 3. We included Generalized Extreme Value distribution ($GEV(, \mu, \sigma, \xi)$ with parameters $\mu = 0$, $\sigma = 1$ and $\xi = 0$, and Cauchy distribution

($Cauchy(\mu, \gamma)$) with parameters $\mu = 0$ and $\gamma = 3$ to compare the sizes under skewed and heavy tailed distributions. Another type of alternative points considered in the power simulation is generated from GEV(λ,1,λ) in one group and from GEV(0,1,0) in the other group. As λ gets large ($\lambda > 1$), none of the moments of GEV(λ,1,λ) exist (Hosking, Wallis, and Wood, 1985). In this case, we anticipate that GCT-mdp may behave poorly or worse than VH-np considering the assumptions that were made for the two tests. Results are shown in Tables 4.1 and 4.2. Power is also investigated for three types of distributions that we anticipated the nonparametric statistic could perform better. Details on the settings for these power plots can be seen in captions on Figures 4.1, 4.2, 4.3, 4.4 and 4.6. The nominal size is set to $\alpha = 0.05$ in all simulations.

Simulation Results

The number of simulations is set to 7000.

Size Simulation

When sample sizes are moderately small, Table 4.1, the sizes are very similar for the two statistics except that VH-np shows a slightly better performance in the moderate-p case and vice-versa in the large-p case.

The choice of the Parzen Smoothing Window size doesn't appear to have a considerable effect on the sizes, but the smaller value (L=10) consistently reduces the size to make it closer to the nominal level, especially when the errors are independent (white noise). The opposite effect is observed when errors have ARMA(2,2) structure.

When sample sizes are small, Table 4.2, the test sizes are further away from the nominal value and they are fairly comparable in the moderate p case. The sizes for both statistics are inflated in the large p case, while VH-np is more inflated.

The choice of the Parzen Smoothing Window size doesn't modify the tests sizes considerably, but the comparison is almost identical to that when sample sizes are moderately small. Both Tables 4.1 and 4.2 suggest that the choice of L may play a role in the size moderated by sample sizes and it is definitely a factor to consider when some degree of dependence exists.

Table 4.1: Achieved type I error rates for two groups. Sample sizes are $n_1 = 40$ and $n_2 = 38$. p stands for dimension and Parzen Smoothing Window is denoted by L.

Type-I error rates× 100

Error distr.	Dependence Structure	Statistic	$p = 300$		$p = 1000$	
			L=10	L=20	L=10	L=20
GEV	indep	VH-np	6.13	7.08	7.59	8.67
		GCT-mdp	5.67	6.75	7.61	8.50
	ARMA	VH-np	6.58	6.78	7.08	7.14
		GCT-mdp	5.91	7.16	6.66	6.47
Normal	indep	VH-np	6.36	7.41	8.28	9.16
		GCT-mdp	6.16	6.44	7.34	8.58
	ARMA	VH-np	6.66	6.50	7.28	7.16
		GCT-mdp	6.92	6.92	6.64	8.05
Gamma	indep	VH-np	5.69	6.44	8.16	8.59
		GCT-mdp	5.97	6.75	7.67	7.89
	ARMA	VH-np	6.91	7.09	7.69	6.84
		GCT-mdp	6.72	6.92	6.67	6.44
Uniform	indep	VH-np	6.03	7.20	8.61	8.16
		GCT-mdp	5.89	6.78	7.66	8.25
	ARMA	VH-np	7.00	7.14	7.08	6.63
		GCT-mdp	6.55	7.17	6.56	6.86
Double exp	indep	VH-np	5.25	7.19	9.19	8.27
		GCT-mdp	5.66	6.41	7.72	7.92
	ARMA	VH-np	7.09	7.30	6.92	6.44
		GCT-mdp	6.63	7.06	6.94	6.69
Cauchy	indep	VH-np	6.08	7.27	8.45	8.88
		GCT-mdp	5.56	6.17	6.00	6.17
	ARMA	VH-np	8.77	8.00	8.78	7.08
		GCT-mdp	7.38	6.73	7.38	6.52

Table 4.2: Achieved type I error rates for two groups. Sample sizes are $n_1 = 20, n_2 = 18$. p stands for dimension and Parzen Smoothing Window is denoted by L.

Type-I error rates× 100

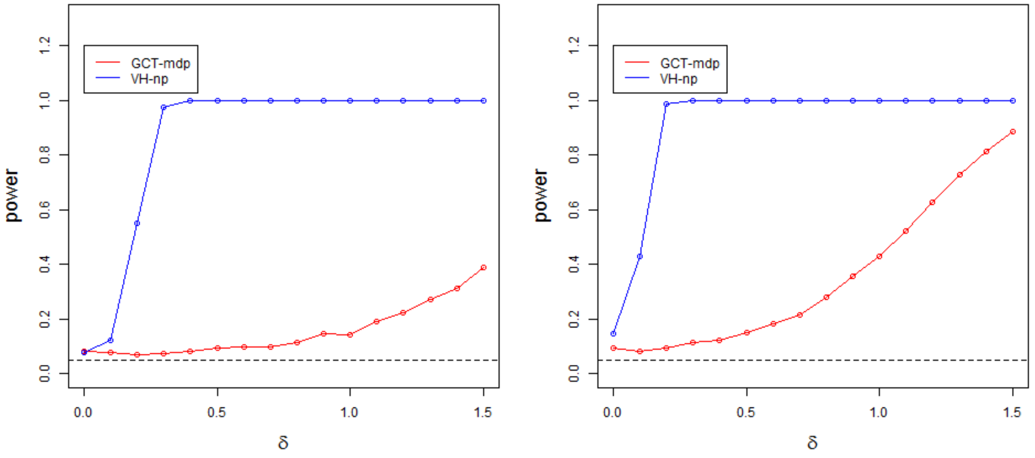
Error distr.	Dependence Structure	Statistic	$p = 300$		$p = 1000$	
			L=10	L=20	L=10	L=20
GEV	indep	VH-np	8.92	9.70	22.59	24.50
		GCT-mdp	7.95	9.69	20.17	19.64
	ARMA	VH-np	7.48	8.02	14.94	14.64
		GCT-mdp	8.03	8.25	13.81	12.56
Normal	indep	VH-np	8.84	9.34	23.81	24.95
		GCT-mdp	9.13	9.73	20.95	20.89
	ARMA	VH-np	7.22	7.31	14.94	14.67
		GCT-mdp	7.80	7.75	13.97	12.66
Gamma	indep	VH-np	8.94	9.48	22.30	24.33
		GCT-mdp	9.55	9.52	20.28	21.13
	ARMA	VH-np	7.80	7.11	15.20	14.22
		GCT-mdp	6.97	7.50	13.30	12.69
Uniform	indep	VH-np	8.00	9.38	23.81	24.09
		GCT-mdp	8.36	9.72	20.20	21.61
	ARMA	VH-np	7.44	7.41	15.31	14.14
		GCT-mdp	6.56	7.69	13.42	13.81
Double exp	indep	VH-np	8.72	9.47	23.47	23.63
		GCT-mdp	8.25	9.97	18.55	20.14
	ARMA	VH-np	8.11	7.56	14.80	14.11
		GCT-mdp	7.44	7.50	13.03	12.78
Cauchy	indep	VH-np	8.56	9.80	23.02	23.86
		GCT-mdp	6.98	8.09	12.58	13.25
	ARMA	VH-np	8.78	7.86	15.63	12.55
		GCT-mdp	8.30	7.98	10.39	9.22

Power Simulation

Power plots were generated to compare the performance of both statistics under some of the settings that were represented in the size tables. Simulation size is 3200.

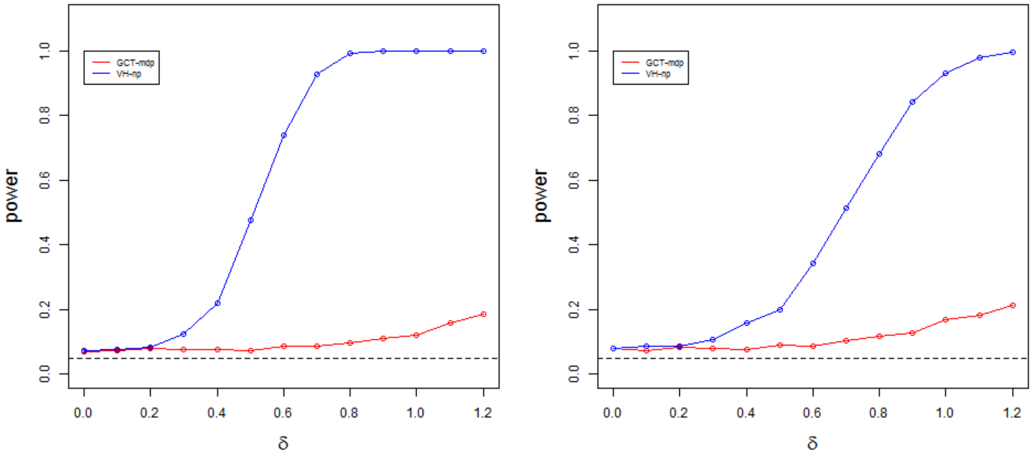
In Figure 4.1, dimension is $p = 1000$ and set equal for both plots. In these plots, we can see a clear advantage of VH-np since it shows a lot more robustness against dependence structure. It is also worth mentioning that size is inflated for both tests if dimension is large and sample size is moderately small.

The setting in Figure 4.2, is very similar to that in Figure 4.3. Indeed, the only difference being that errors follow an ARMA model in the former and are independent



(a) (b)

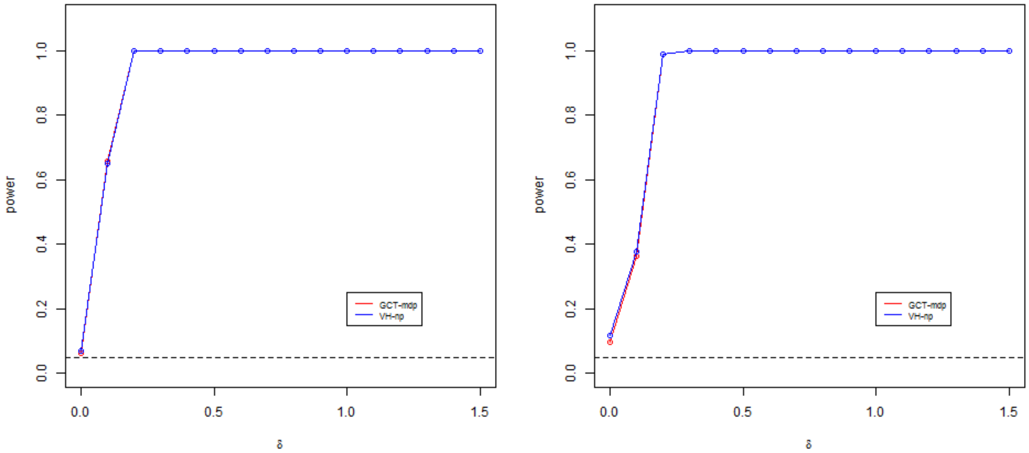
Figure 4.1: Power plots for VH-np and GCT-mdp. Errors are generated from Cauchy distribution with ARMA structure. Dimension is $p = 1000$. Sample sizes $(n_1, n_2) = (40, 38)$ are represented on panel (a) and $(n_1, n_2) = (20, 18)$ on panel (b). δ is the location shift.



(a) (b)

Figure 4.2: Power plots for VH-np and GCT-mdp. Errors are generated from Cauchy distribution with ARMA structure. Dimension is $p = 300$. Sample sizes $(n_1, n_2) = 40, 38$ are represented on panel (a) and $(n_1, n_2) = (20, 18)$ on panel (b). δ is the location shift.

in the latter. As we can see in these graphs, VH-np has a much better performance compared to GCT-mdp. It looks like VH-np is more robust to dependence structure.



(a)

(b)

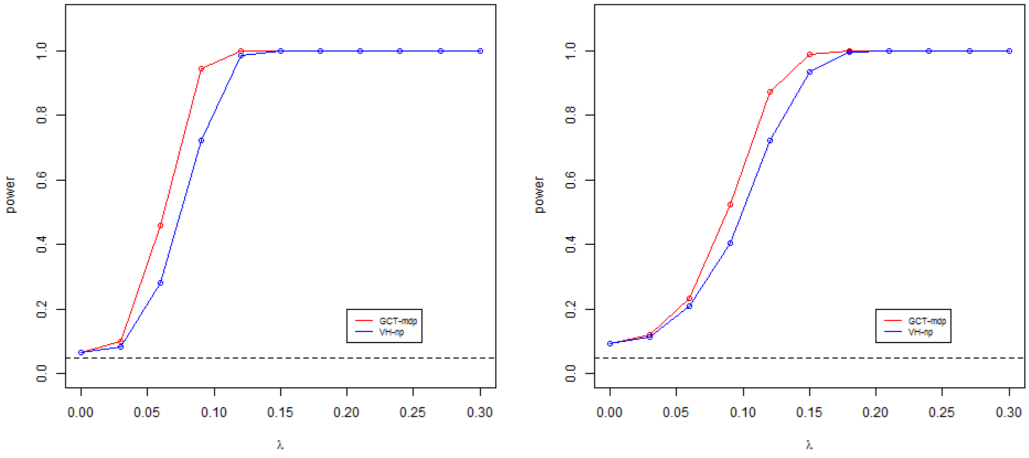
Figure 4.3: Power plots for VH-np and GCT-mdp. Errors are generated from Cauchy distribution with independence structure. Dimension is $p = 300$. Sample sizes $(n_1, n_2) = (40, 38)$ are shown on panel (a) and $(n_1, n_2) = (20, 18)$ on panel (b). δ is the location shift.

In Figure 4.3, we can see that with sample size around 40 per group, the two tests pick up power faster and have better sizes compared to smaller sample size cases, otherwise the two tests compare very similarly.

The settings in Figure 4.4 are the same as in Figure 4.3 except the error distribution is GEV. The results are also similar but it looks like GCT-mdp shows improvement compared to the settings in Figure 4.3. More specifically, it gains power a little faster than VH-np. It is worth noting that the parameter λ corresponds to location and shape i.e. $\mu = \lambda$ and $\xi = \lambda$ and that makes alternative in one of the samples to change location and shape at the same time.

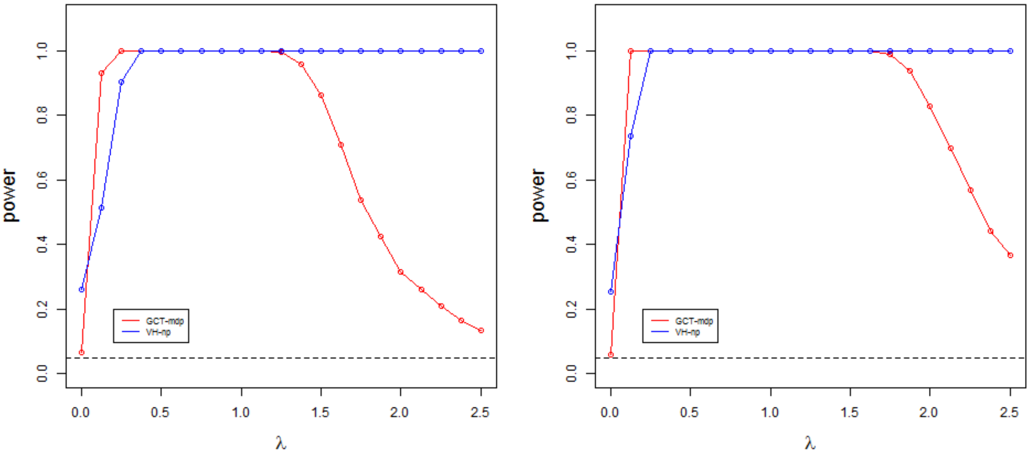
In Figure 4.5, for ARMA setting, we observe that as λ gets larger than 1 the power of GCT-mdp drops. On the contrary, power for VH-np stays up when moments don't exist any more. Size for VH-np is however inflated.

Finally, for small and equal sample sizes, Figure 4.6, the two tests perform well,



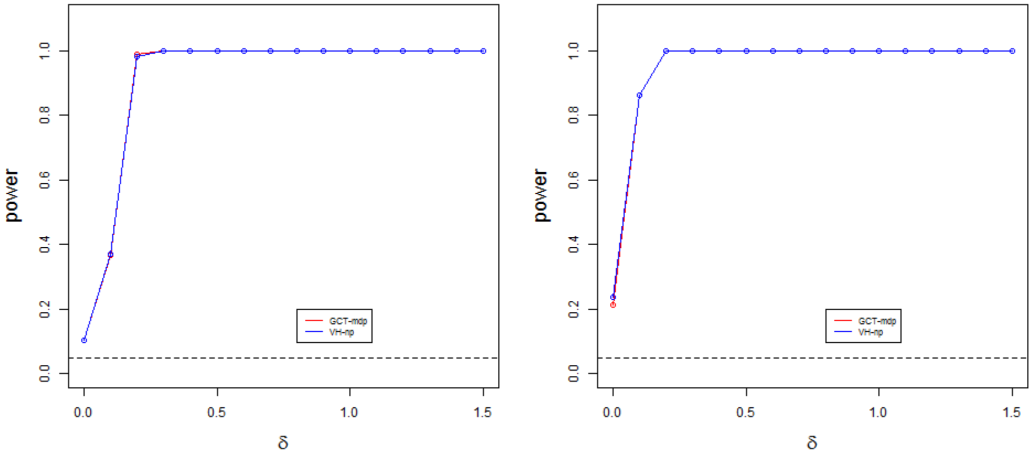
(a) (b)

Figure 4.4: Power plots for VH-np and GCT-mdp. Errors are generated from GEV distribution with independence structure. Dimension is $p = 300$. Sample sizes $(n_1, n_2) = (40, 38)$ is shown on panel (a) and $(n_1, n_2) = (20, 18)$ on panel (b). λ is the location and shape parameter and is set equal in the second group compared to 0 for both in the first group.



(a) (b)

Figure 4.5: Power plots for VH-np and GCT-mdp. Errors are generated from GEV distribution with ARMA structure. Dimension is $p = 300$. Sample sizes are $(n_1, n_2) = (40, 38)$.$\beta = 0.4$ is shown on panel (a) and $\beta = 0.8$ is shown on panel (b). λ is the location and shape parameter and is set equal in the second group compared to 0 for both in the first group. β is the proportion of variables shifted.



(a) (b)

Figure 4.6: Power plots for VH-np and GCT-mdp. Errors are generated from Normal distribution with independence structure. Sample sizes are $(n_1, n_2) = (20, 18)$. On panel (a), dimension $p = 300$ is represented, $p = 1000$ is on panel (b). δ is the location shift.

but the sizes for the two statistics are considerably inflated when $p = 1000$. We can say that the two statistics behave similarly in a normal independent setting.

4.6 Conclusions

We investigated a nonparametric test statistic for two group comparison in high dimensions. The test statistic was defined in manner similar to the one in Chapter 3, but motivated by the nonparametric Wilcoxon-Mann-Whitney type statistic of Brunner and Munzel (2000).

The proposed test statistic is shown to asymptotically follow a Normal distribution. Mild moment conditions and strong mixing (α-mixing) dependence is required to establish this result. The new nonparametric test is compared with CGT-mdp in a simulation. In finite samples, sizes are very similar for both statistics with a little advantage for GCT-mdp. Both statistics are liberal when small sample size and large dimension were combined. However, if sample size is not large enough, the larger the dimension gets, it will make the size performance worse. As we showed in

Chapter 3, this observation suggests that n and p have to be large enough and the assumed rate $p = o(n^2)$ appears necessary, or else the approximate normality of the test may not hold.

When comparing power, the simulations show a clear advantage for VH-np in heavily tailed distributions such as Cauchy. The power simulation for Generalized Extreme Value distribution is special in that λ varies both location and shape parameters. In this case, under ARMA structure, VH-np shows more robustness when moments no longer exist ($\lambda > 1$). Simulations show a great advantage of VH-np when there exist correlation between the variables.Otherwise, under independence the two tests perform very similarly with CGT-mdp having a slight advantage.

In summary, VH-np has a better performance overall . In almost all cases, for moderate p the sizes for both tests are comparable and the power is either very similar or clearly advantageous for VH-np. This makes one think that the nonparametric version of the statistic is preferable to the moderate p version of Gregory et al. (2015).

4.7 Appendix

Technical details needed for the main result in Section 4.4 are presented in this section.

Some Definitions

To fix notation about mid ranks we define

$$R_{ijk} = n \cdot \widehat{H}_k(x_{ij}) + \frac{1}{2}$$

is the rank among all elements in both samples admitting ties where

$$\widehat{H}_k(x) = \sum_{i=1}^{2} \frac{n_i}{n} \widehat{F}_{ik}(x)$$

and we can write

$$\widehat{F}_{ik}(x) = \frac{1}{n_i} \sum_{j=1}^{n_i} c(x - X_{ijk}), i = 1, 2.$$

where

$$c(x) = \tfrac{1}{2}(c^-(x) + c^+(x)), \quad c^+(x) = \begin{cases} 0 & x < 0 \\ 1 & x \geq 0 \end{cases}, \quad c^-(x) = \begin{cases} 0 & x \leq 0 \\ 1 & x > 0 \end{cases}.$$

Also,

$$\overline{R}_{i.k} = \frac{1}{n_i} \sum_{j=1}^{n_i} R_{ijk}$$

is the mean of the ranks in the i^{th} sample for the k^{th} variable and

$$R_{jk}^{(i)} = n_i \widehat{F}_{ik}(X_{ijk}) + \frac{1}{2}$$

the rank of X_{ijk} among the n_i observations of the ith sample.

Statistic as a Lebesgue Function

The statistic W is an scaled average of the squared W_{ni}'s, and the univariate statistics can be written as

$$W_{nk} = \frac{\sqrt{n}(\widehat{\omega}_k - \frac{1}{2})}{\widehat{\sigma}_{nk}}$$

and $\widehat{\omega}_k$ can be written, using definitions in Subsection 4.7, as

$$\begin{aligned}
\widehat{\omega}_k &= \frac{1}{n_1}(\overline{R}_{2.k} - \frac{n_2+1}{2}) \\
&= \frac{1}{n_1}(\frac{1}{n_2}\sum_{j=1}^{n_2} R_{2jk} - \frac{n_2+1}{2}) \\
&= \frac{1}{n_1}(\frac{1}{n_2}\sum_{j=1}^{n_2}(n \cdot \widehat{H}_k(X_{2jk}) + \frac{1}{2}) - \frac{n_2+1}{2}) \\
&= \frac{1}{n_1}(\frac{1}{n_2}\sum_{j=1}^{n_2} n(\cdot \sum_{i=1}^{2} \frac{n_i}{n} \widehat{F}_{ik}(X_{2jk}) + \frac{1}{2}) - \frac{n_2+1}{2}) \\
&= \frac{1}{n_1}(\frac{1}{n_2}\sum_{j=1}^{n_2}(\sum_{i=1}^{2}\sum_{m=1}^{n_i} c(X_{2jk} - X_{ijm}) + \frac{1}{2}) - \frac{n_2+1}{2}).
\end{aligned}$$

As we can see the estimator for the relative effect is a Lebesgue function of the random sample. Also, $\widehat{\sigma}_{nk}$ can be written as functions of S_{ik}^2 where, using again definitions on Subsection 4.7, S_{ik}^2 can be written as:

$$
\begin{aligned}
S_{ik}^2 =& \frac{1}{n_i - 1}\sum_{j=1}^{n_i}(R_{ijk} - R_{jk}^{(i)} - \overline{R}_{i.k} + \frac{n_1 + 1}{2})^2 \\
=& \frac{1}{n_i - 1}\sum_{j=1}^{n_i}((\sum_{i=1}^{2}\sum_{m=1}^{n_i} c(X_{ijk} - X_{imk}) + \frac{1}{2}) - n_i(\frac{1}{n_i}\sum_{m=1}^{n_i} c(X_{imk} - X_{ijl}))+ \\
& \frac{1}{2} - \frac{1}{n_i}\sum_{j=1}^{n_i}(\sum_{r=1}^{2}\sum_{m=1}^{n_i} c(X_{ijk} - X_{rmk}) + \frac{1}{2}) + \frac{n_1 + 1}{2})^2
\end{aligned}
$$

which is a Lebesgue function of the sample.

Chapter 5 Comparison of Various High Dimensional Tests

5.1 Introduction

With a number of methods proposed in Chapters 3 and 4, we will explore the results and some possible extensions compared to other alternative methods proposed in literature. We will study numerically some parametric tests but using ranks instead of raw values and other nonparametric methods. We will use ranks instead of the raw values to illustrate numerically the effect of introducing dependency between the observed samples. The results may suggest the possibility of a modification of assumptions or even new statistics. We will also illustrate the tests' application in data analysis to real-data from an encephalograph (EEG) experiment.

5.2 Compared Methods of Analysis

Diagonal Likelihood Ratio Test (DLRT)

An interesting approach is taken by Hu, Tong, and Genton (2019). A composite test statistic is derived from the likelihood ratio assuming that covariance matrices follow a common diagonal matrix structure. To derive the asymptotic normality under the null and local alternatives the assumption of diagonal covariance matrix is not needed, α-mixing is then assumed. The following hypotheses are tested,

$$H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 \text{ versus } H_1 : \boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2.$$

For these hypotheses, the likelihood ratio test statistic becomes,

$$T_2 = n \sum_{k=1}^{p} \log(1 + \frac{t_{nk}^2}{\nu_2}) := \sum_{k=1}^{p} V_{nk},$$

where t_{nk}^2 is the regular version of t^2 with pooled variance for the k^{th} variable and $\nu_2 = n - 2$. One difficulty of this approach is to calculate the mean and variance of the log transformed. These quantities are calculated using the digamma function

Ψ. Defining $D(x) = \Psi\{(x+1)/2\} - \Psi(x/2)$, $G_1 = nD(\nu_2)$ and $G_2 = n^2\{D^2(\nu_2) - 2D'(\nu_2)\}$, then

$$E[V_{nj}] = G_1 \text{ and } \mathrm{Var}(V_{nj}) = G_2 - G_1^2.$$

Further, the following is true,

$$\frac{T_2 - pG_1}{\tau_2\sqrt{p}} \xrightarrow{d} N(0,1) \text{ as } p \to \infty,$$

where τ_2^2 is the asymptotic variance of the statistic.

This result has a corollary in which T_2 is corrected by an asymptotic expansion of the moment rather than the asymptotic mean. In this case, a formula is reached for any expansion level. The result is a little restrictive since it assumes the sequence $\{V_{nj}\}$ is stationary, along with α-mixing.

Nonparametric Test for Two samples

Our method proposed in Chapter 4 (called "VH-np") is included for comparison. One of the main advantages of this method is that no distributional assumptions are made except that distributions are non-degenerate.

High-Dimensional Rank-Based Test

A different approach to VH-np was taken by Kong and Harrar (from now on "KH"). The statistic is constructed naturally from the nonparametric concept of relative effect defined in Chapter 4. The relative effect as defined in Chapter 4 is:

$$\omega_k = P(X_{11k} < X_{21k}) + \frac{1}{2}P(X_{11k} = X_{21k}),$$

for $k = 1, ..., p$. They use an average of the univariate distribution functions as a baseline reference. The relative effect in this case is defined as:

$$\omega_{ik} = E[Y_{i1k}] = \int H dF_{ik} \text{ where } H(x) = \frac{1}{2p}\sum_{k=1}^{p}\{F_{1k}(x) + F_{2k}(x)\}.$$

It can be noted that one big difference with VH-np is that, instead of comparing the distribution of one of the samples on the k^{th} variable to the distribution of the second sample on the same variable, it is compared to the average distribution of all variables from both samples. Given these definitions, they estimated the parameters ω_{ik} naturally using the empirical distribution in place of the unknown true distribution, i.e.

$$\widehat{\omega}_{ik} = \frac{1}{2p}[\frac{\overline{R}^{(i)}_{.k} - 1/2}{n_i} + \frac{\overline{R}_{i.k} - \overline{R}^{(i)}_{.k}}{n - n_i}] \tag{5.1}$$

with notation as defined in Chapter 4. The test statistic (T_n) used is that defined in Chen and Qin (2010) with the exception that mid-ranks are used instead of raw values. A consistent estimator for the variance of this statistic is proposed and finally the following result is presented.

$$\widehat{\sigma}_n^{-1} T_n(\widehat{\boldsymbol{Y}}^c) \xrightarrow{d} N(0,1) \text{ as } n, p \to \infty.$$

This test assumes α-mixing as VH-np along with some regularity conditions. There are differences in definitions of the relative effect quantity and that affects the results.

General Component Tests

The methods GCT-mdp and GCT-lgp described in Gregory et al. (2015) are a special two group case of our methods proposed in Chapter 3 for unequal covariance matrices. They are also included for comparison but using mid-ranks instead of raw values.

5.3 Simulation

We will explore the performance of the statistics proposed in Section 5.2 under various settings. In order to make the simulation as thorough as possible, we have investigated multiple combinations of parameter values. Specifically, effects in the number of groups a, sample sizes n_i and dimension p are investigated. In the statistics DLRT, VH and GCT, the Parzen Smoothing Window L needs to be specified beforehand to estimate the asymptotic variance of the test statistic. In the power simulation, δ expresses the shift of the variable means and β is the proportion of means shifted.

Simulation Design

When considering $a = 2$, we compare sizes for methods explained in Section 5.2 under the following settings:

- Sample sizes $(n_1, n_2) = \{(20, 20), (30, 30), (40, 40), (50, 50), (10, 15), (20, 25), (30, 35)(40, 45), (50, 55)\}$.

  These sizes cover balanced and moderately unbalanced situations.

- Dimension $p \in \{100, 150, 200, 250, 300, 350, 400, 450, 500\}$.

  There are various assumptions in the methods discussed in this chapter about the relationship between n and p. We expect this range of values to show the possible interaction effects of sizes and dimension.

- For methods that require L to be defined, we will use $L = 0.5p$ and fixed $L = 20$.

- Covariance structure:

  - Independence: $\Sigma_1 = I_p$ and $\Sigma_2 = I_p$.
  - Equi-Correlation: $\Sigma_1 = 0.5I_p + 0.5J_p$ and $\Sigma_2 = (1 - \rho)I_p + \rho J_p$.
  - Auto-Regressive: $\Sigma_1 = (0.5^{|j-j_1|})$ and $\Sigma_2 = (\rho^{|j-j_1|})$.
  - Square-Root-Decay: $\Sigma_1 = (0.5|j - j_1|^{-1/2})$ and $\Sigma_2 = (\rho|j - j_1|^{-1/2})$.

  These settings cover situations from independence between variables to short and long range dependence. The parameter ρ will take values 0.1, 0.2, 0.3, 0.4 and 0.5 to illustrate the behavior of tests in homoscedasticity and heteroscedasticity.

- Marginal error distribution: Cauchy(0,3), N(0,1), centered Gamma(4,2). These settings include error coming from symmetric distributions to heavily tailed and heavily skewed distributions.

All combinations from these settings were simulated but we only present some plots to avoid redundancy. The specific settings are described in each plot.

Power is investigated for the different tests described in Section 5.2. We will focus our interest in some specific settings listed below:

- sample sizes $(n_1, n_2) = \{(20, 20), (20, 25), (50, 50)\}$.
- Dimension $p = \{100, 500\}$.
- $L = 20$.
- Dependence model: Independence, Auto-Regressive and Square-Root-Decay.
- Correlation parameter $\rho = \{0.1, 0.5\}$.
- Error distribution: Cauchy(0,3), N(0,1), centered Gamma(4,2).

Two types of alternative hypotheses will be examined. We will explore a mean shift in the two sample variables (δ) and we will investigate a change in the proportion of variables shifted for a fixed change (β).

Size Simulation Results

Effect of Sample Size and Dimension

All statistics considered in this subsection are mostly compared in independence and Square-Root-Decay structure to check their behaviour for sample size and dimension. Other covariance structures followed similar patterns. Their differences are described in Subsection 5.3.

Normal: In Figure 5.1, all tests converge to nominal value as size and dimension increase simultaneously. KH and GCT-lgp are particularly accurate in the case where n and p are relatively small.

Cauchy: For Cauchy distribution, Figure 5.2 shows that KH is performing well under all values of sample sizes and dimensions. For fixed sample size, KH gets closer to nominal size as p increases but it is not the case for all other tests. However, as both sample size and dimension increase simultaneously, the trends shown by all tests are similar.

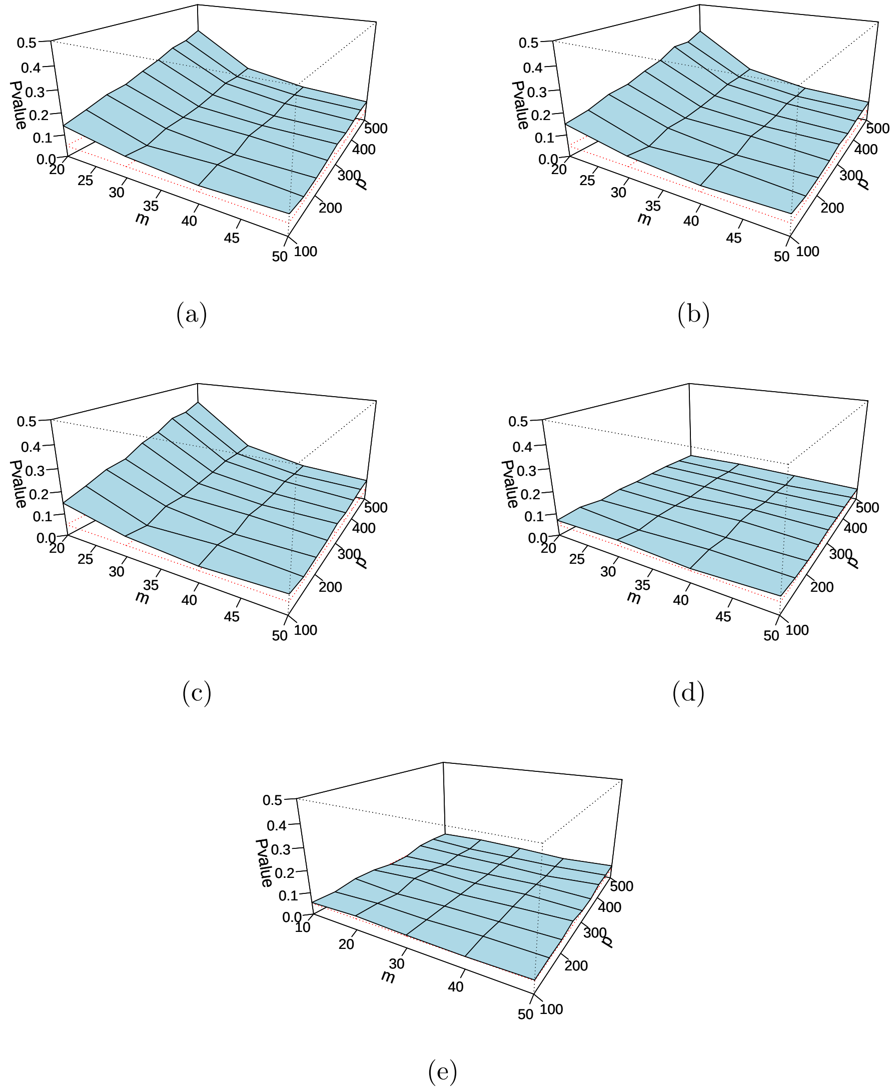
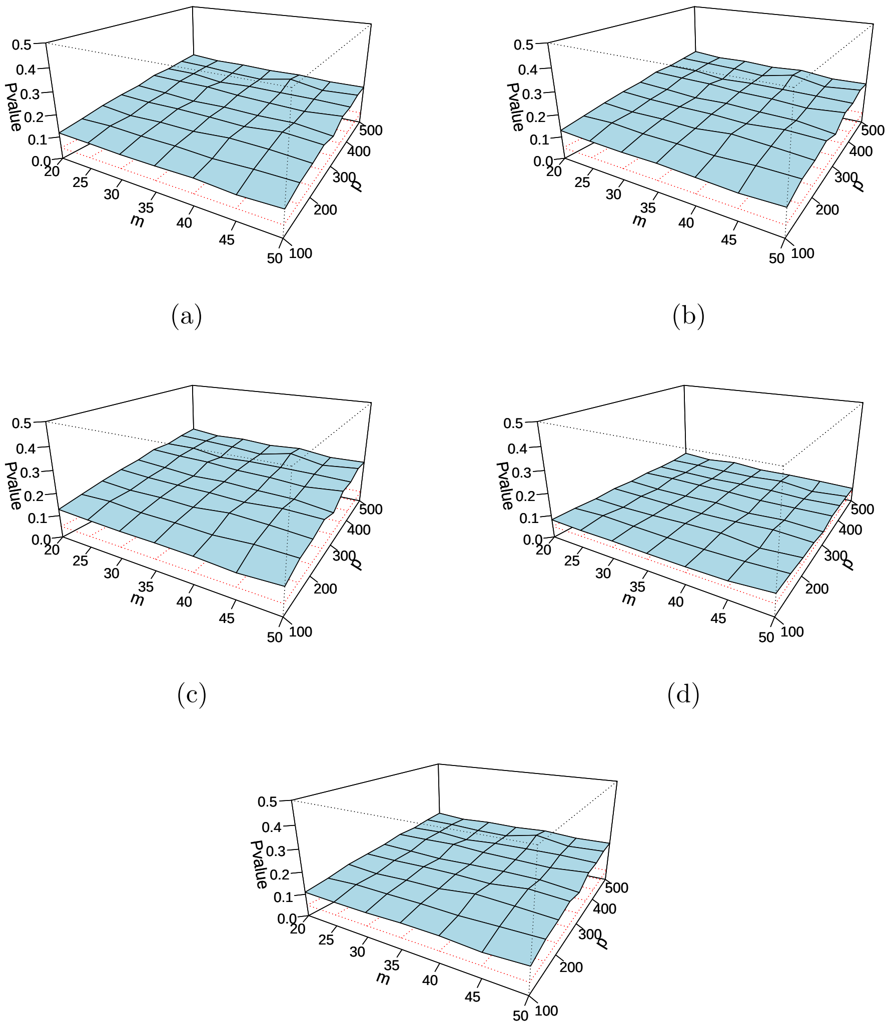


Figure 5.1: Achieved type-I error rates for all tests against sample size and dimension. Errors are generated form Normal distribution with Square-Root-Decay as covariance structure and heteroscedastic. Parzen Smoothing Window used is L=20. Sample sizes are $n_1 = m$ and $n_2 = m$. Panel (a) represents VH-np, panel (b) represents GCT-mdp, panel (c) represents DLRT, panel (d) represents KH and panel (e) represents GCT-lgp.

Gamma: In Figure 5.3, we can observe the behavior of all tests is remarkably good for most sizes and dimensions. Seems specially good GCT-lgp in panel (e).

Effect of Unbalancedness

The behaviour of the statistics is explored under unbalanced samples.



(a)

(b)

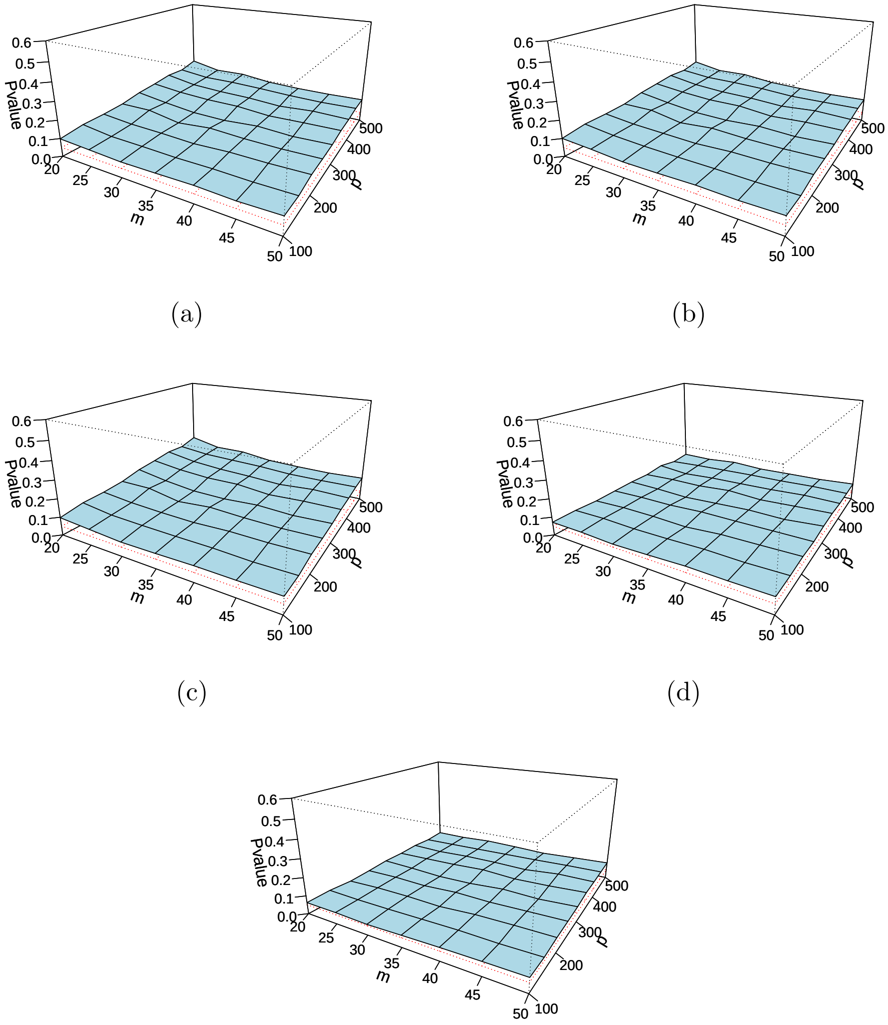
(c)

(d)

(e)

Figure 5.2: Achieved type-I error rates for all tests against sample size and dimension. Errors are generated from Cauchy distribution with Square-Root-Decay as covariance structure and heteroscedastic. Parzen Smoothing Window used is L=20. Sample sizes are $n_1 = m$ and $n_2 = m$. Panel (a) represents VH-np, panel (b) represents GCT-mdp, panel (c) represents DLRT, panel (d) represents KH and panel (e) represents GCT-lgp.

Normal: In Figure 5.4, we observe the behaviour of VH-np comparing unbalanced and balanced samples. It seems that sizes are more affected in the balanced case than in the unbalanced case. Especially when sample sizes are smaller. As we can see in Figure 5.5, for fixed sample sizes, KH and GCT-lgp, maintain stable size as p increases but the others tests increase the size. The difference is much less prominent



(a)

(b)

(c)

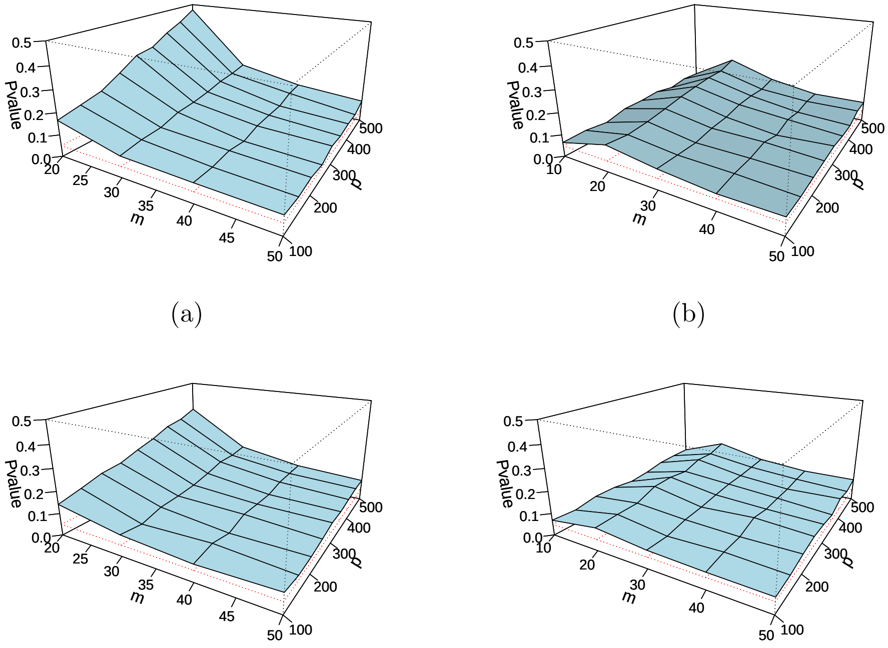
(d)

(e)

Figure 5.3: Achieved type-I error rates for all tests against sample size and dimension. Errors are generated from Gamma distribution with Square-Root-Decay as covariance structure and heteroscedastic. Parzen Smoothing Window used is L=20. Sample sizes are $n_1 = m$ and $n_2 = m$. Panel (a) represents VH-np, panel (b) represents GCT-mdp, panel (c) represents DLRT, panel (d) represents KH and panel (e) represents GCT-lgp.

as sample sizes increase.

Gamma: For Gamma distributed errors, the unbalanced small sample sizes have higher error rates than the balanced (see Figure 5.6) and the difference is increased as dimension increases. When sample sizes are increased, both, balanced and unbal-



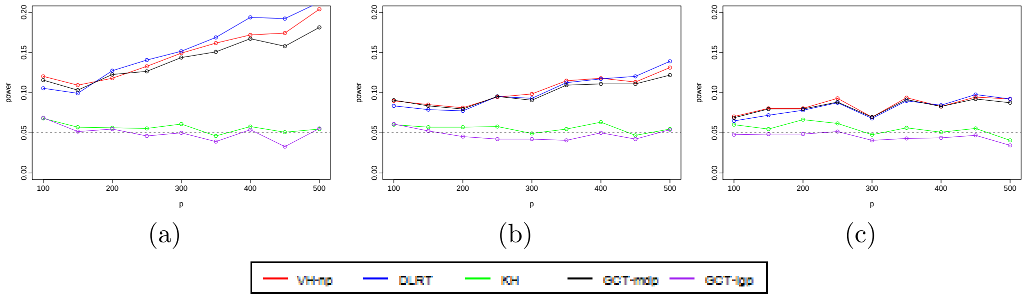
(a)

(b)

(c)

(d)

Figure 5.4: Achieved type-I error rates for VH-np against sample size and dimension. Errors are generated from Normal distribution. Parzen Smoothing Window used is L=20. Covariance structures are Independence for panels (a) and (b) and Square-Root-Decay heteroscedastic for panels (c) and (d) . For panels (b) and (d) sample sizes are $n_1 = m$ and $n_2 = m{+}5$ and panels (a) and (c) sample sizes are $n_1 = n_2 = m$.



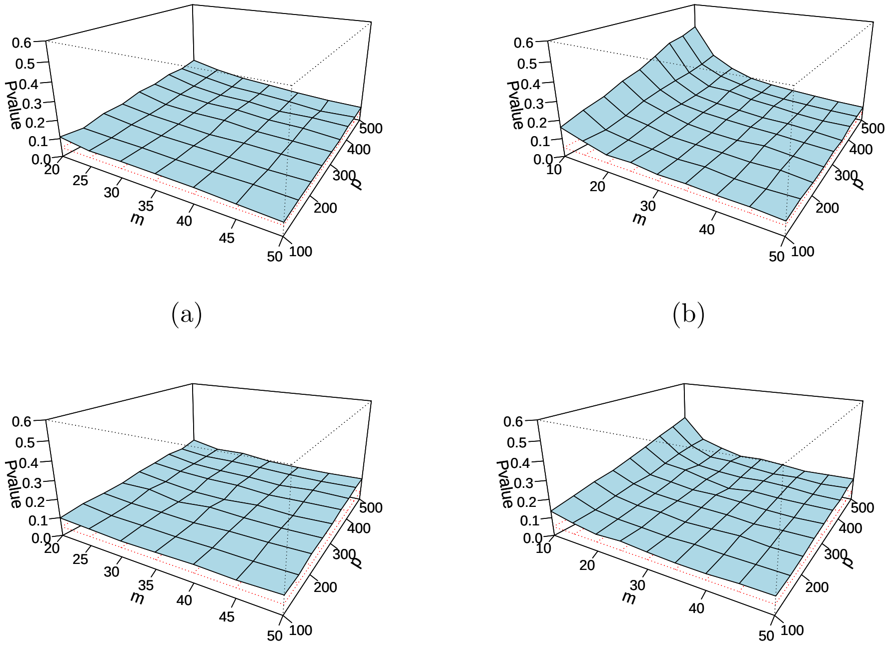
(a)

(b)

(c)

Figure 5.5: Achieved type-I error rates for all tests against dimension. Errors are generated from Normal distribution, all tests are represented in all panels. Covariance structure is Independence. Parzen Smoothing Window used is L=20. Sample sizes are $n_1 = m$ and $n_2 = m + 5$. $m = 30$ in panel (a), $m = 40$ in panel (b) and $m = 50$ in panel (c).

anced error rates are fairly similar.



(a)

(b)

(c)

(d)

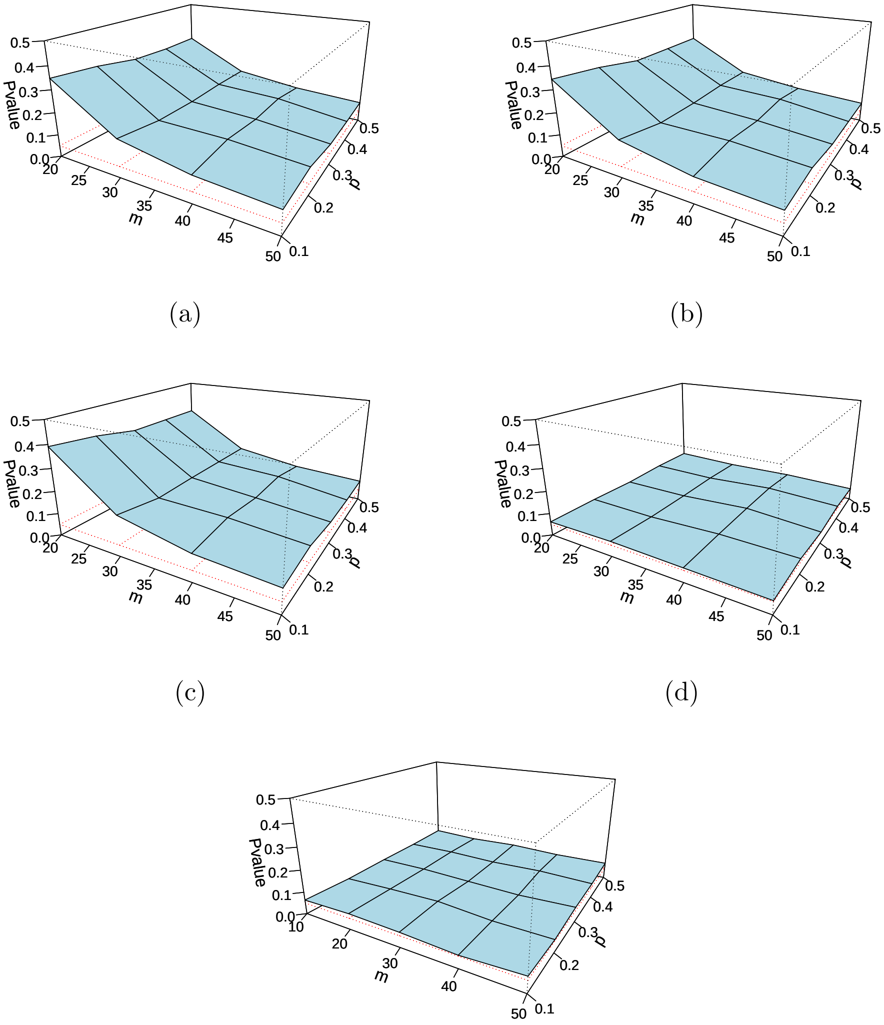
Figure 5.6: Achieved type-I error rates for VH-np against sample size and dimension. Errors are generated from Gamma distribution. Parzen Smoothing Window used is L=20. Covariance structures are Independence for panels (a) and (b) and Square-Root-Decay heteroscedastic for panels (c) and (d). Sample sizes are $n_1 = m$ and $n_2 = m + 5$ for panels (b) and (d) and $n_1 = n_2 = m$ for panels (a) and (c).

Effect of Size and Correlation

All plots shown to compare the effect of size and correlation are chosen to have dimension $p = 500$ and covariance structure a Square-Root-Decay. we chose this setting because it is a more appropriate setting for the real life problem that we are trying to solve.

Normal: As we can see on panels from Figure 5.7, error rates increase as heteroscedasticity increases. KH and GCT-lgp tests seem to be less affected by heteroscedasticity. It is a common trend in the other settings' plots.

Cauchy: The pattern in plots with Cauchy distributed errors (see Figure 5.8) seem different to that of normal. KH behaves very well compared to the other tests. The



(a)

(b)

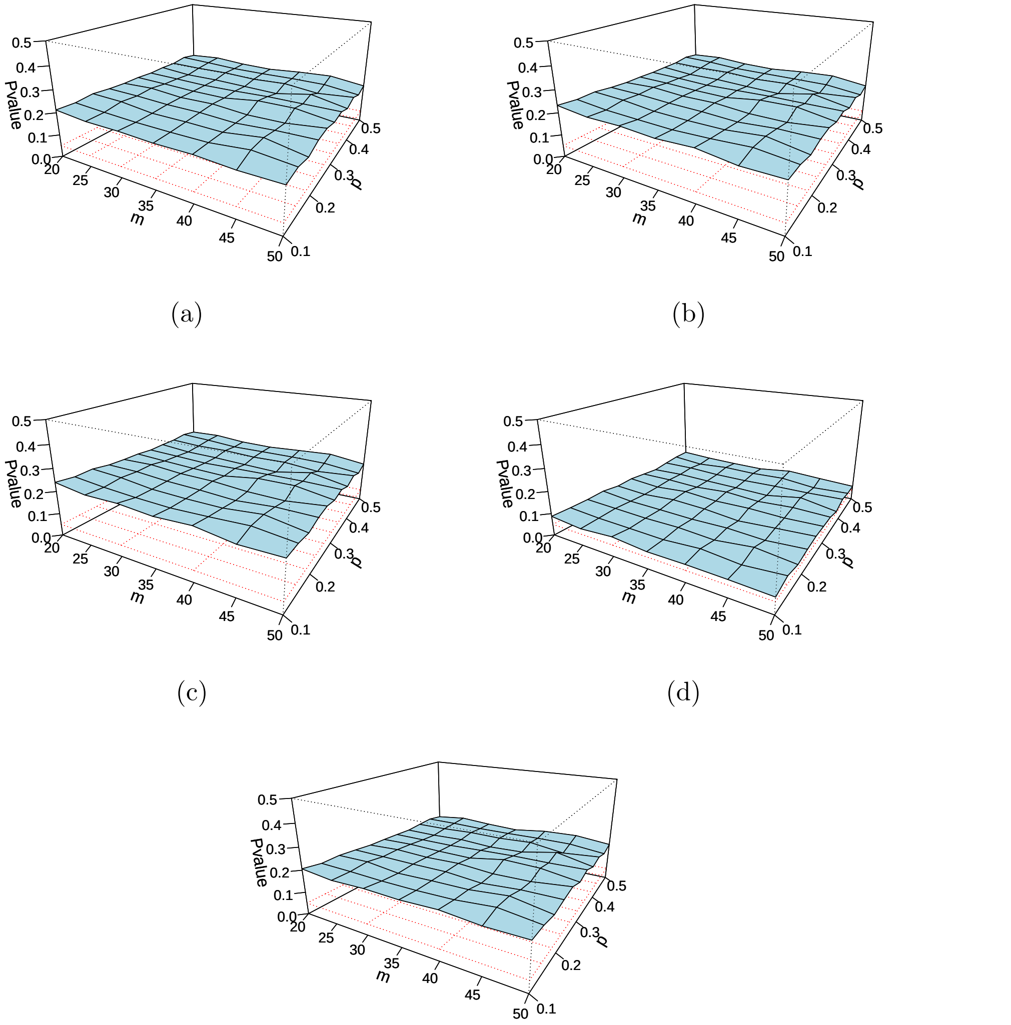
(c)

(d)

(e)

Figure 5.7: Achieved type-I error rates for all tests against sample size and correlation parameter (ρ). Errors are generated from Normal distribution. Square-Root-Decay as covariance structure and dimension is $p = 500$. m is the increment in size which is the same on both samples, so sample sizes are $n_1 = m$ and $n_2 = m$. Parzen Smoothing Window used is L=20. Panel (a) is VH test, panel (b) represents GCT-mdp, panel (c) represents DLRT, panel (d) represents KH and panel (e) represents GCT-lgp.

achieved sizes of all other tests are far from the nominal sizes for small sample sizes and seem to be getting closer to nominal size as sample sizes increase but at a much slower rate than for the normally distibuted errors.



(a)

(b)

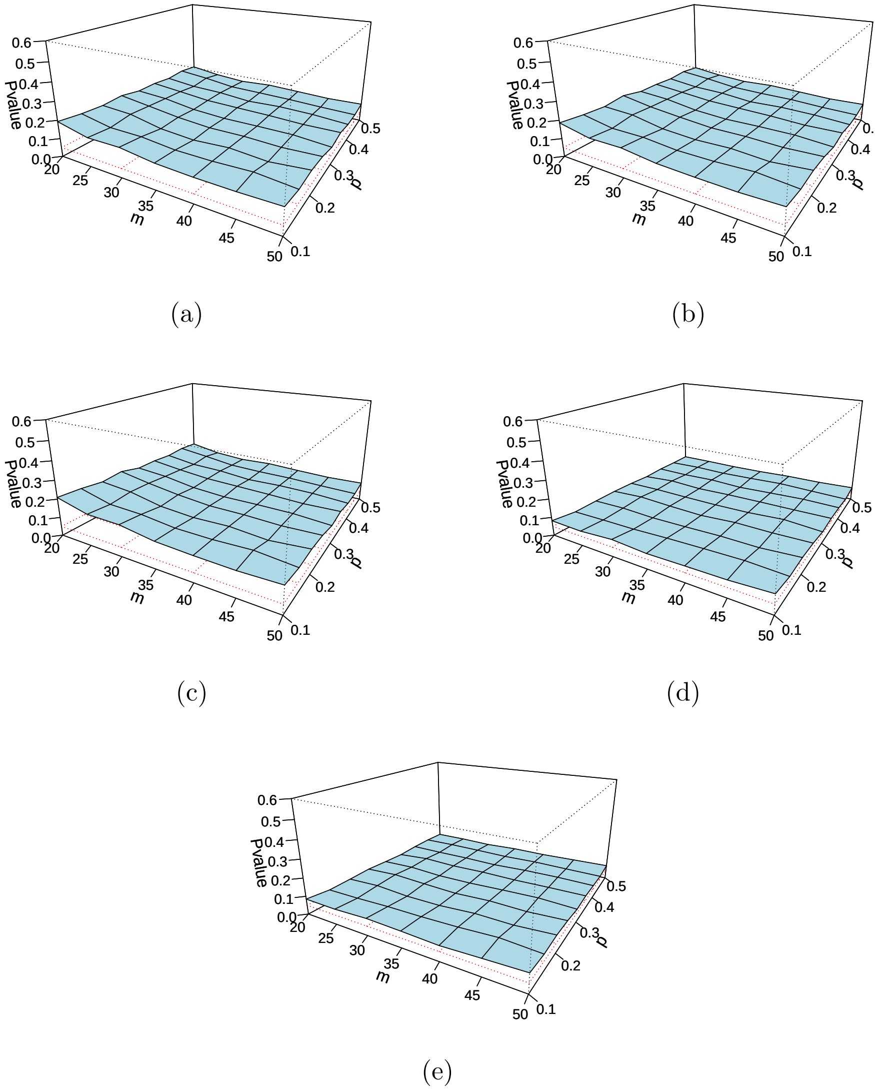
(c)

(d)

(e)

Figure 5.8: Achieved type-I error rates for all tests against sample size and correlation parameter (ρ). Errors are generated from Cauchy distribution. Square-Root-Decay as covariance structure and dimension is $p = 500$. m is the increment in size which is the same on both samples, so sample sizes are $n_1 = m$ and $n_2 = m$. Parzen Smoothing Window used is L=20. Panel (a) represents VH test, panel (b) represents GCT-mdp, panel (c) represents DLRT, panel (d) represents KH and panel (e) represents GCT-lgp.

Gamma: The results for Gamma distributed errors shown in Figure 5.9 are better than normal errors for small sample sizes but get closer to nominal size at a slower pace than the normal case. GCT-lgp and KH behave remarkably well for smaller and larger $n's$ and under homoscedastic and heteroscedastic settings. All other test



(a)

(b)

(c)

(d)

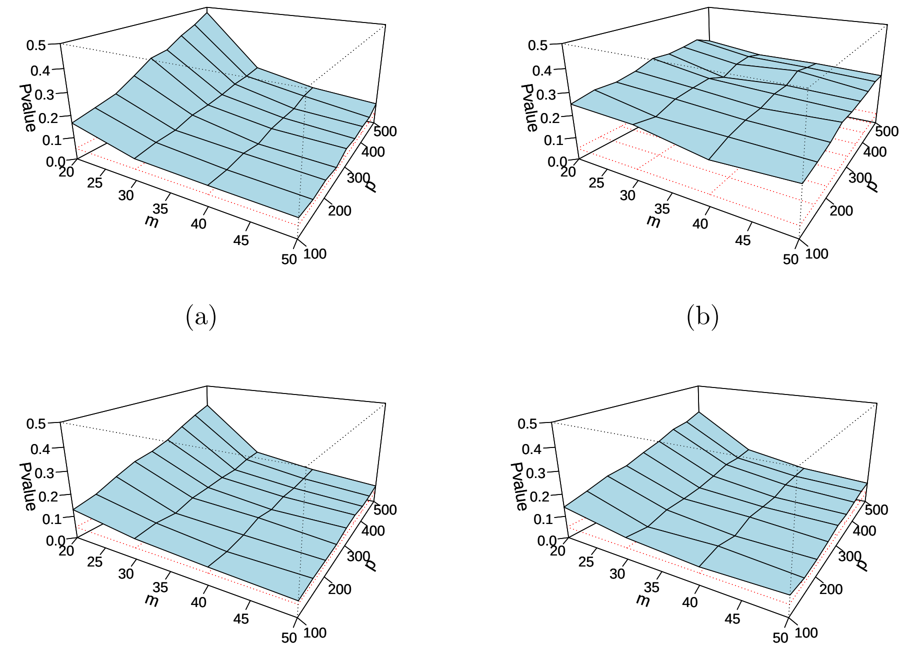
(e)

Figure 5.9: Achieved type-I error rates for all tests against sample size and correlation parameter (ρ). Errors are generated from Gamma distribution. Square-Root-Decay as covariance structure and dimension is $p = 500$. Sample sizes are $n_1 = m$ and $n_2 = m$. Parzen Smoothing Window used is L=20. Panel (a) represents VH test, panel (b) represents GCT-mdp, panel (c) represents DLRT, panel (d) represents KH and panel (e) represents GCT-lgp.

behave better under an homoscedatic setting than heteroscedastic.

Effect of Covariance Structure

The different covariance structures are compared on VH-np.The value of ρ is set to 0.1, which is the most heteroscedastic case. Error rates are plotted against sample size and dimension.



(a)

(b)

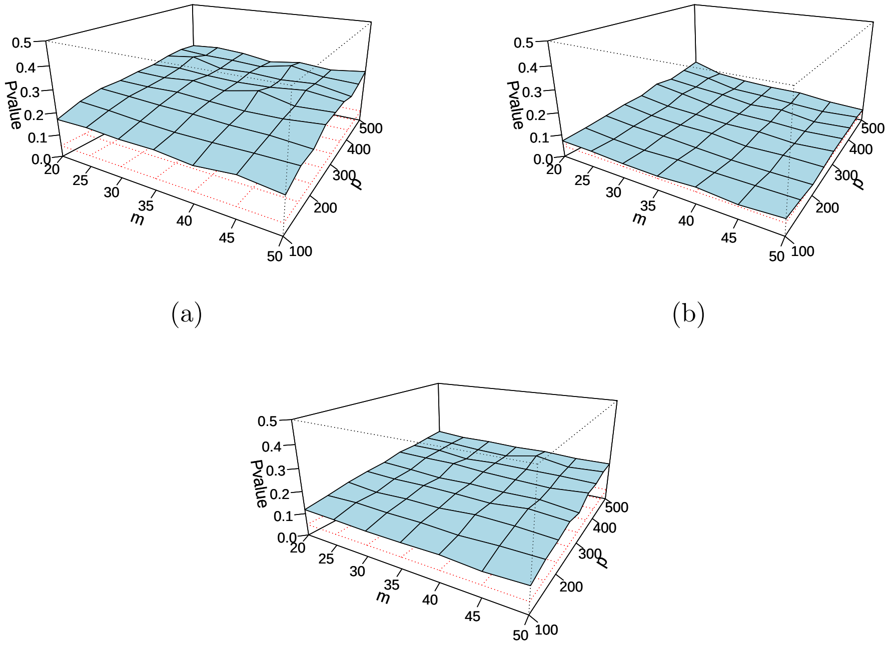
(c)

(d)

Figure 5.10: Achieved type-I error rates for VH-np against sample size and dimension. Errors generated from Normal distribution, under heteroscedasticity. Parzen Smoothing Window used is L=20. Panel (a) represents independence covariance structure and panel (b) represents Equi-Correlation and panel (c) represents Auto-Regressive and panel (d) represents Square-Root-Decay. Sample sizes are $n_1 = m$ and $n_2 = m$ and p is the dimension.

Normal: Comparing different covariance structures in Figure 5.10, we can see the good behavior of VH test for all covariance structures except in the Equi-Correlation structure case where we observe much slower convergence.

Cauchy: As we can see in Figure 5.11, in a similar way to errors generated from other distributions, the proposed test (VH-np) behaves worst under the Equi-Correlation covariance structure. In Figure 5.12 we observe that for smaller sample size, the error rates increase in all tests except on KH and GCT-lgp. As sample size increases, the difference between the tests fades away.

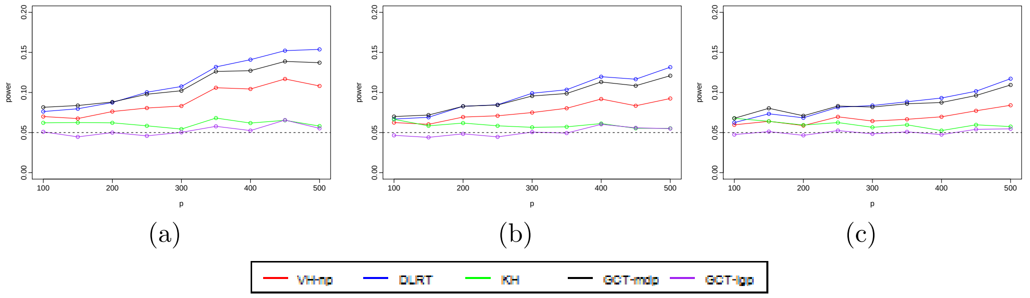


(a)

(b)

(c)

Figure 5.11: Achieved type-I error rates for VH-np against sample size and dimension. Errors are generated from Cauchy distribution, under heteroscedasticity. Parzen Smoothing Window used is L=20. Panel (a) represents Equi-Correlation covariance structure, panel (b) represents Auto-Regressive, panel (c) represents Square-Root-Decay. Sample sizes are $n_1 = m$ and $n_2 = m$ and p is the dimension.



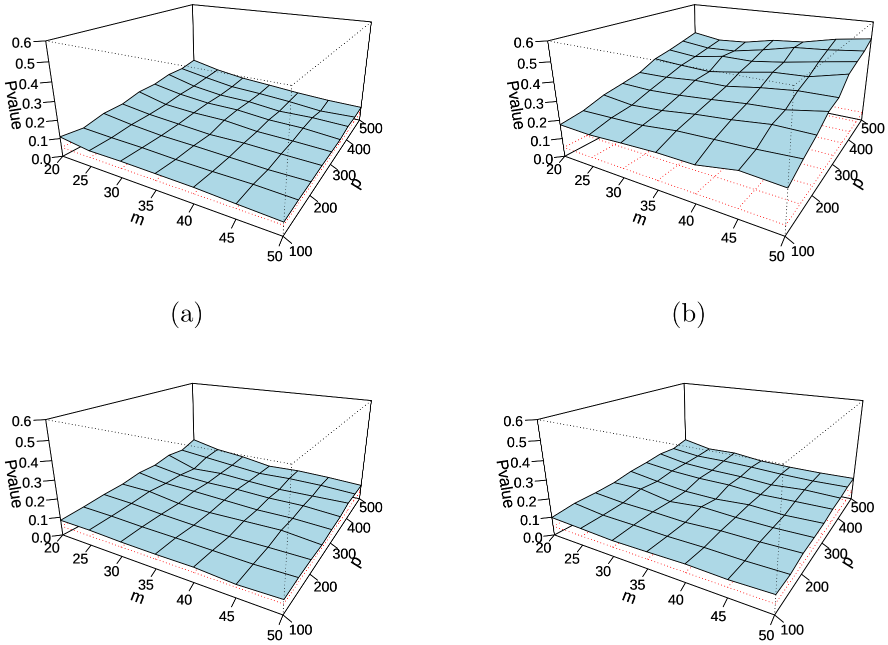
(a)

(b)

(c)

Figure 5.12: Achieved type-I error rates for all tests against dimension. Errors are generated from Cauchy distribution. Covariance structure is Auto-Regressive, sample sizes are $n_1 = m$ and $n_2 = m$. Parzen window parameter used is L=20. $m = 30$ in panel (a), $m = 40$ in panel (b) and $m = 50$ in panel (c).

Gamma: As Figure 5.13 shows, the Gamma distribution has a similar effect on VH-np as Cauchy for the different covariance structures.



(a)

(b)

(c)

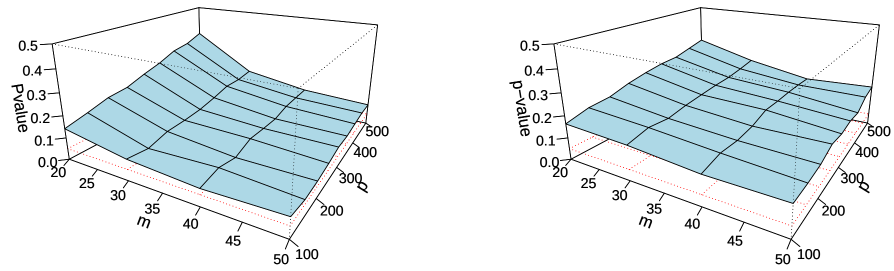
(d)

Figure 5.13: Achieved type-I error rates for VH-np against sample size and dimension. Errors are generated from Gamma distribution, under heteroscedasticity($\rho = 0.1$). Parzen Smoothing Window used is L=20. Panel (a) represents independence covariance structure, panel (b) represents Equi-Correlation, panel (c) represents Auto-Regressive and panel (d) represents Square-Root-Decay, m is the increment in size which is the same on both samples, so sample sizes are $n_1 = m$ and $n_2 = m$ and p is the dimension.

Effect of L (Parzen Smoothing Window)

Heteroscedastic autocorrelated setting. Error rates are plotted with size increment and dimension.

Normal: Looking at Figure 5.14, we can see that increasing the L has a negative impact on the error rates, more so as p and n increase.



(a)

(b)

Figure 5.14: Achieved type-I error rates for VH-np against sample size and dimension. Errors are generated from Normal distribution, under heteroscedasticity. Panel (a) represents parameter $L = 20$ and panel (b) represents $L = p/2$. Sample sizes are $n_1 = m$ and $n_2 = m$ and p is the dimension.

Power simulation Results

Plots were generated for all settings described in the simulation design. We will focus on settings where sample is unbalanced ($n_1 = 20, \quad n_2 = 25$), errors are distributed as Gamma and dimension is larger ($p = 500$). We are interested in data that is homoscedastic and Auto-Regressive. Power plots in this section follow these settings unless changes are noted.

Effect of δ and β

In Figures 5.15 and 5.16, there is a noticeable difference in the sizes of two homogeneous groups of test. One group includes VH-np, GCT-mdp and DLRT and the other group has KH and GCT-lgp. Their power curves behave similarly but comparing the groups, type I error rates are more inflated for the former group. KH is especially interesting since it has a good size and picks power faster than GCT-lgp. We can see that for small proportion of variables shifted (β), the power grows at a much slower pace. That pace increases for all tests as β increases. We also observe in Figure 5.16 that for small signal the proportion of variables shifted doesn't affect much the behaviour of any of the tests. As signal increases we observe a similar behaviour in the comparison of tests as we see in Figure 5.15. When shift is one unit all tests reach full power considerably fast. When $\beta = 0.5$ all tests except GCT-lgp

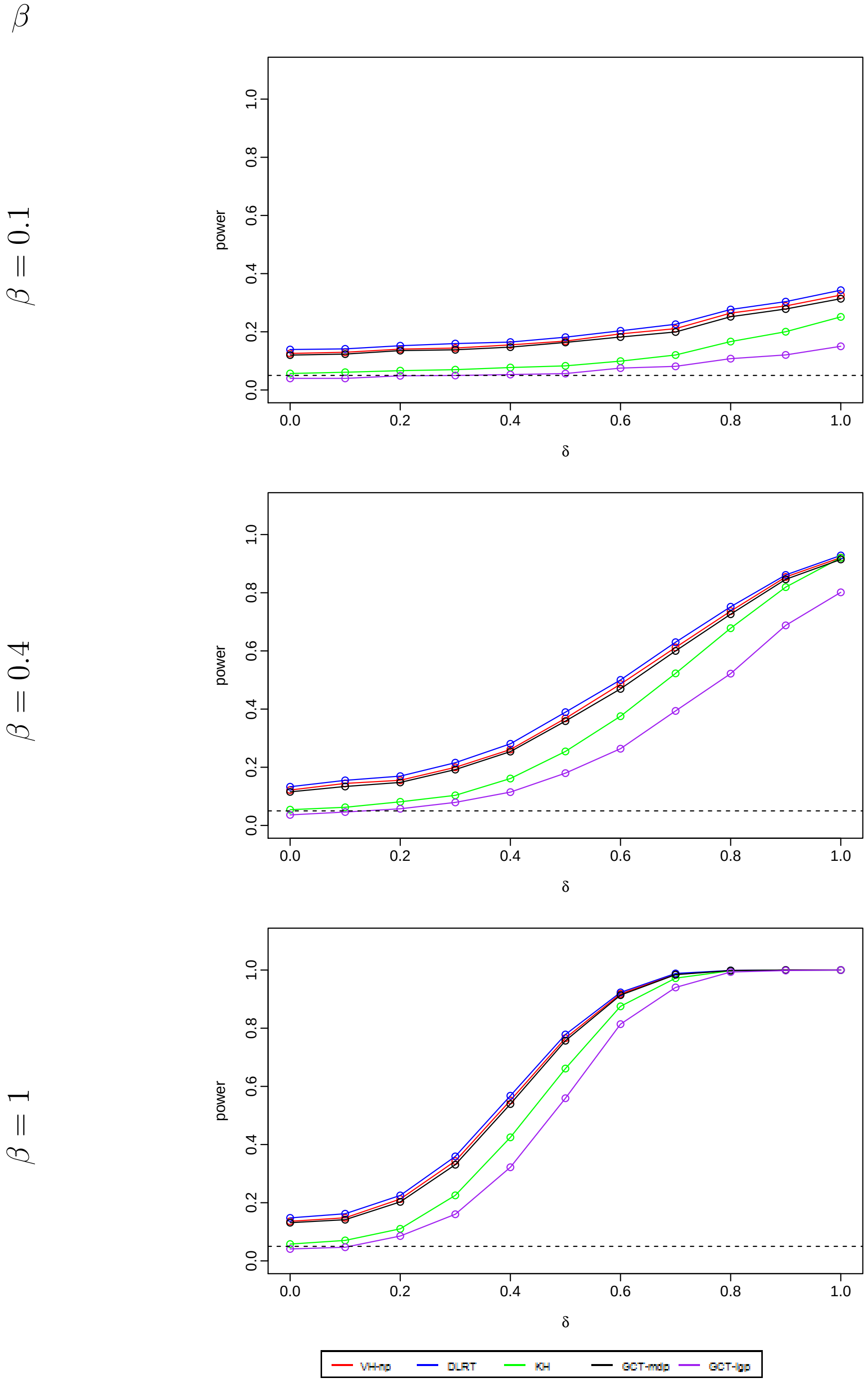


Figure 5.15: Power plot for all tests under shifting alternative. Errors are generated from Gamma distribution. Sample sizes are $(n_1, n_2) = (20, 25)$ and dimension $p = 500$. Covariance structure is Auto-Regressive and homoscedastic. Parzen Smoothing Window used is L=20.

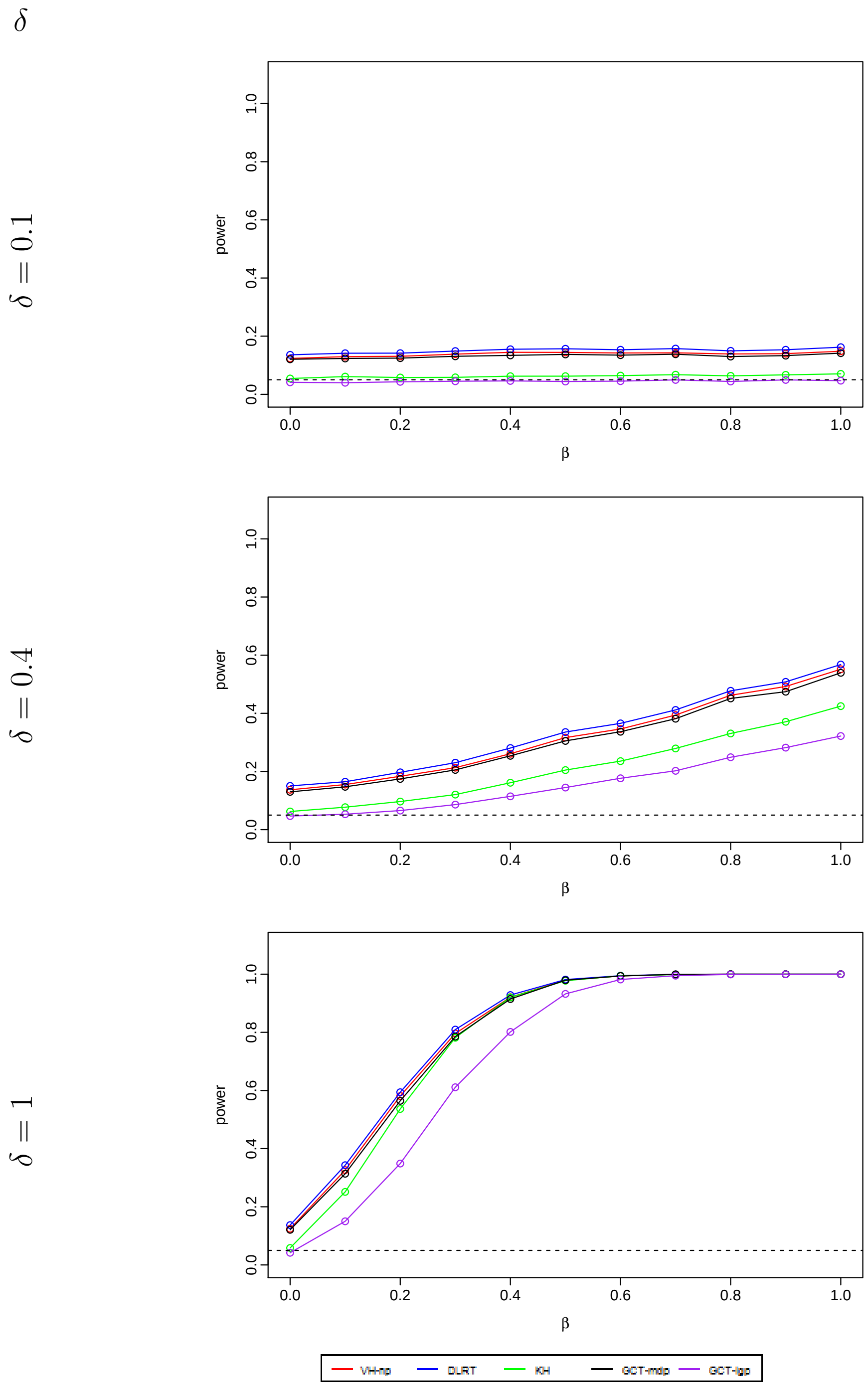


Figure 5.16: Power plot for all tests against proportion of variables shifted alternative. Errors are generated from Gamma distribution. Sample sizes are $(n_1, n_2) = (20, 25)$ and dimension $p = 500$. Parzen Smoothing Window used is L=20. Covariance structure is Auto-Regressive and homoscedastic.

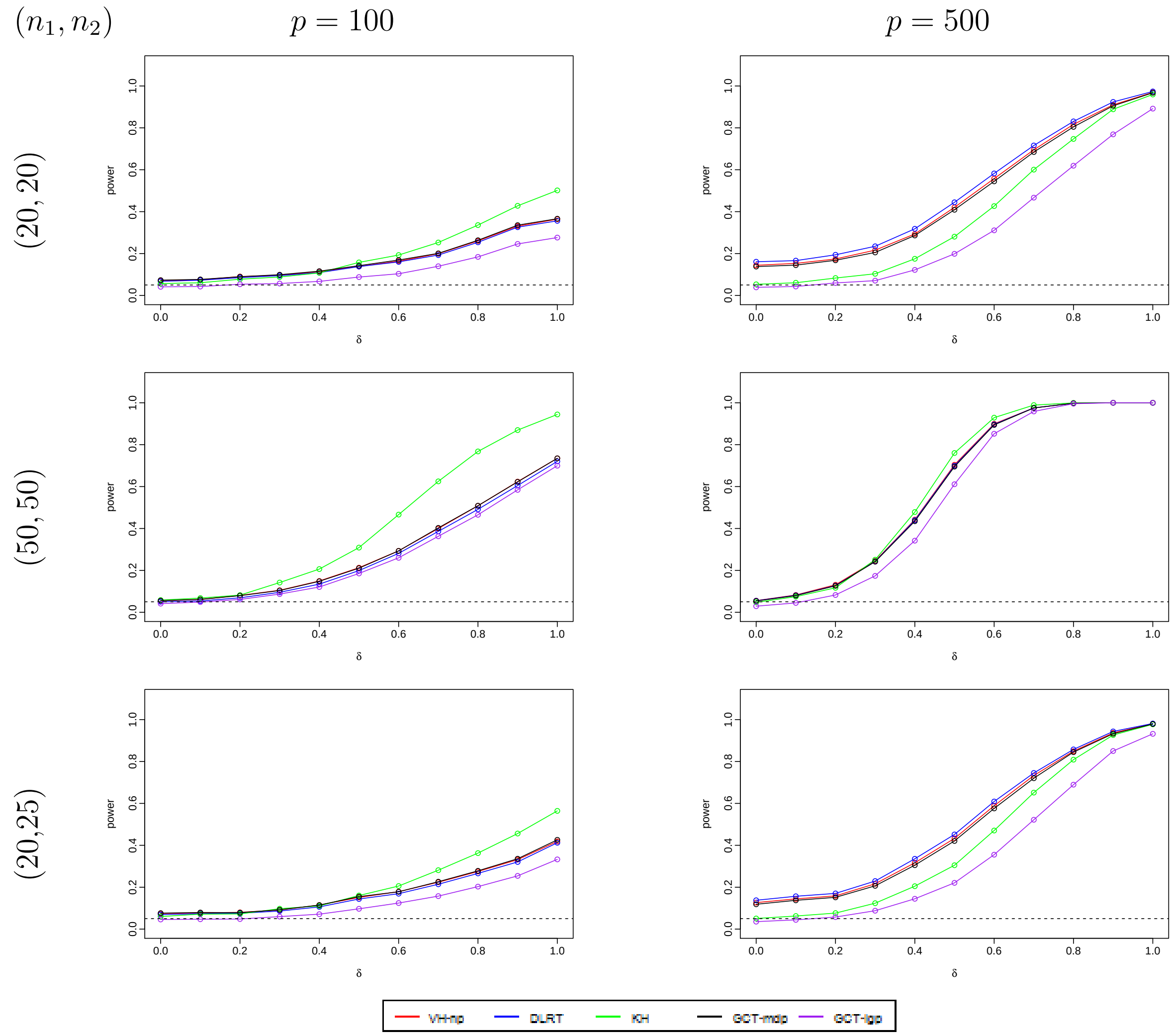


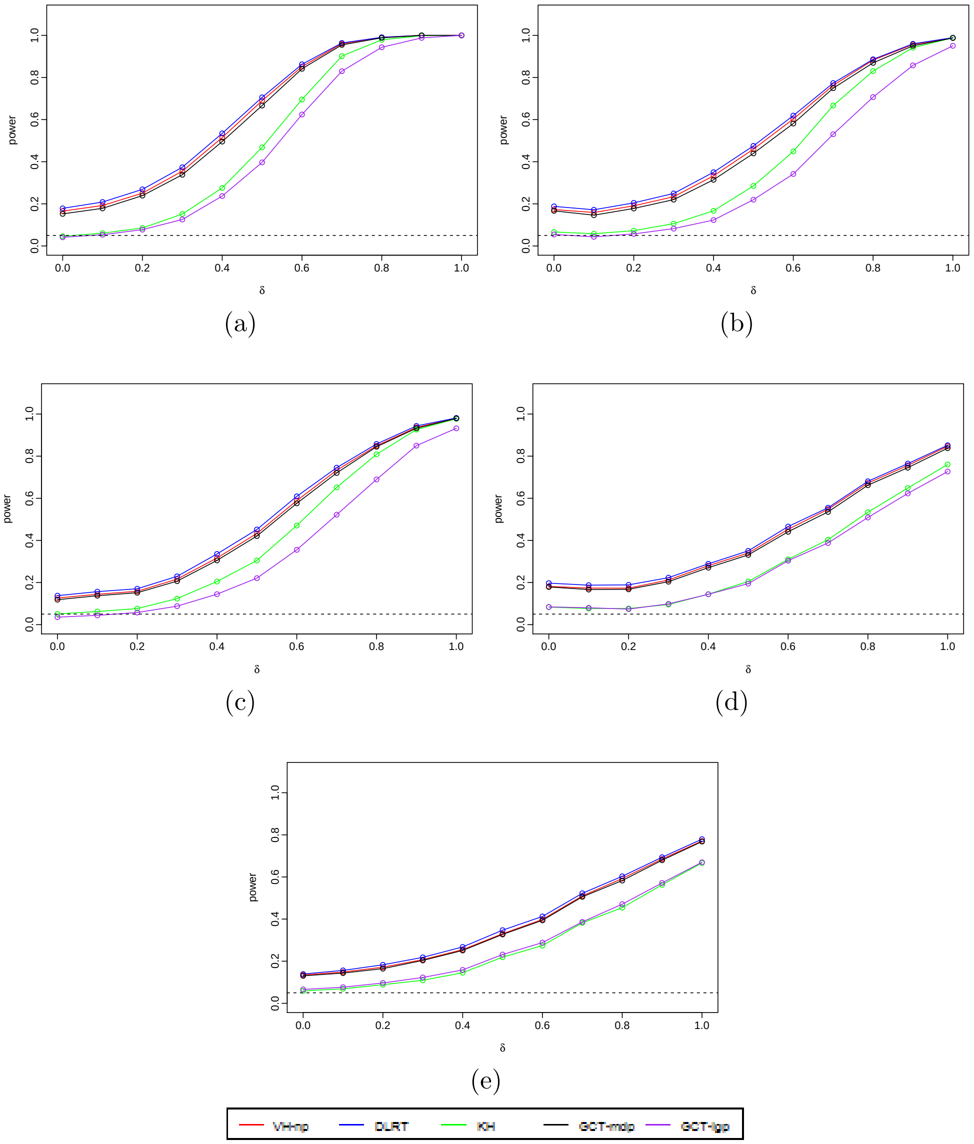
Figure 5.17: Power plot for all tests under shifting alternative. Errors are generated from Gamma distribution. Covariance structure is homoscedastic Auto-Regressive. Parzen Smoothing Window used is L=20. Proportion of active variables is $\beta = 0.5$.

are whithin less than five hundredths from one.

Effect of Dimension and Size

In Figure 5.17, we notice that for smaller p there is a slight advantage of KH compared to all of the others that behave similarly, even in the unbalanced case. When p is larger, we see KH and GCT-lgp have noticeably better size compared to the other methods. If sample sizes are relatively smaller or unbalanced, the size of VH-np, DLRT and GCT-mdp are off and pick up power very similarly.

Effect of Covariance Matrix



(a)

(b)

(c)

(d)

(e)

Figure 5.18: Power plot for all tests under shifting alternative. Errors are generated from Gamma distribution. Proportion of active variables is $\beta = 0.5$. Size is unbalanced $(n_1, n_2) = (20, 25)$. Dimension is $p = 500$. Parzen Smoothing Window used is L=20. Panel (a) corresponds to Independent covariance matrix, panel (b)corresponds to Auto-Regressive heteroscedastic, panel (c) corresponds to Auto-Regressive homoscedastic, panel (d) corresponds to Square-Root-Decay heteroscedastic and panel (e) corresponds to Square-Root-Decay homoscedastic.

In Figure 5.18, the comparison between the tests is almost the same as in other settings but we can see that under independence, power curves are slightly steeper and in Auto-Regressive covariance structure power increases faster than in Square-Root-Decay structure. When comparing homoscedastic to heteroscedastic, panels are very similar.

Effect of Distribution and Homoscedasticity

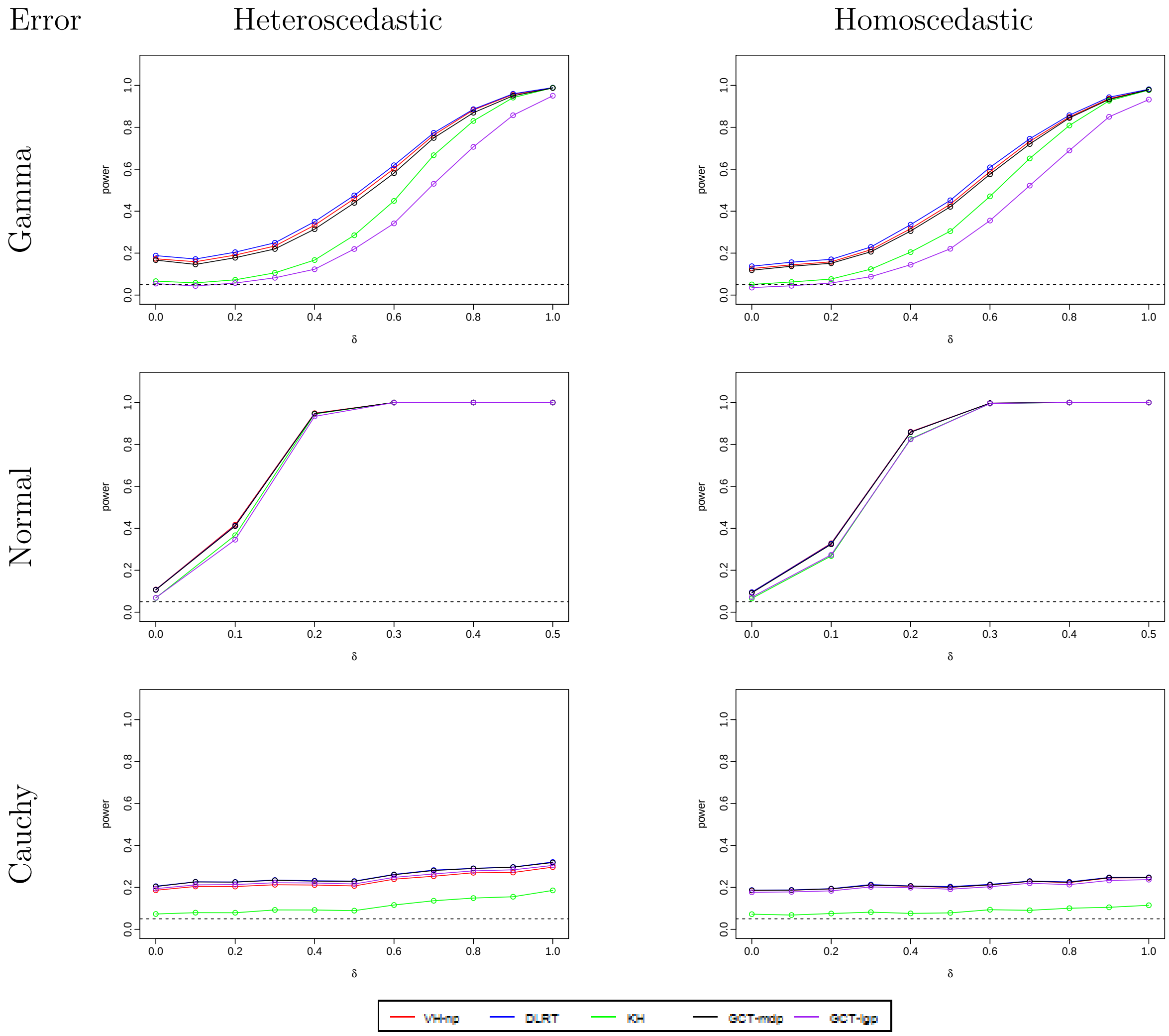


Figure 5.19: Power plot for all tests under shifting alternative. Errors are generated from Gamma distribution. Sample sizes are $(n_1, n_2) = (20, 25)$, dimension is $p = 500$, covariance structure is Auto-Regressive with proportion of active variables is $\beta = 0.5$. Parzen Smoothing Window used is L=20.

In Figure 5.19, we observe big differences in power curve plots for the different distributions. Comparison between the different tests is still very similar to previous figures. For Normal distribution, the power curve is steeper and all tests are very close to each other. For Gamma distribution, power is not so steep and sizes of VH-np, DLRT and GCT-mdp are considerably off. For Cauchy, none of the tests pick power for the differences investigated in this simulation. When comparing homoscedastic to heteroscedastic, slightly better results are observed for homoscedastic settings.

5.4 Area Under Reciever Operating Characteristic (ROC) Curve Analysis

We explored the sensitivity and specificity of the tests with ROC curves. We aimed to see the behavior of the tests when the shift was large ($\delta = 4$) and the proportion of shifted means was small and slightly modified ($\beta = 0, 0.025, 0.05$). The ROC curve plots 1-specificity in the x-axis and sensitivity in the y-axis. The good behavior of a test is measured by the area under the curve. The higher the area, the better.

As we can see in Figure 5.20, the area under the curve is good for all tests but it is extremely good for KH. Even when proportion of means shifted is small, KH sensitivity is really high compared to all other methods. Among the rest of the methods, it seems that GCT-lgp has the lowest area under the curve and the rest have very similar behavior.

5.5 Effect of Scaling Transformation

From the simulations shown previously in this Chapter, it seems that KH is almost unbeatable in all situations. However, knowing that KH uses overall average of CDF's as reference and overall ranks from all variables, we wonder if changes in scale or units of measurement for different variables will have an effect in the test's behavior. To elicit some answers to this question we ran some simulations in which we introduced a scale difference among the variables. This, in real data, would happen if variables have different scales or units of measurement a frequent situation in microarray data. We were interested in the effect of a scale difference among the variables

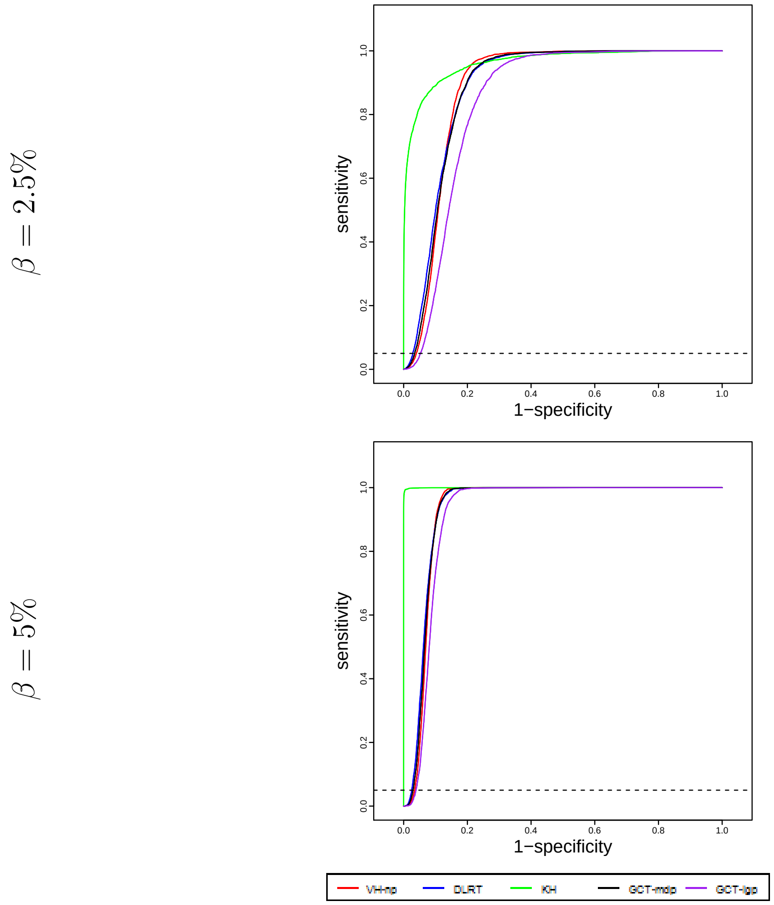


Figure 5.20: ROC curves that plot sensitivity against 1-specificity. Errors are generated from Gamma distribution. Sample sizes are $(n_1, n_2) = (20, 25)$ with $\delta = 4$. Dimension is $p = 500$. Covariance structure is Auto-Regressive and homoscedastic. Parzen Smoothing Window used is L=20.

in skewed data for unbalanced samples with moderate sample sizes. Simulation with settings similar to some of these in Section 5.3. We considered sample sizes $(n_1, n_2) \in \{(30, 35), (50, 50)\}$ since we know from Section 5.3 that for small sample sizes, the Type-I error rates of VH-np, GCT-mdp and DLRT are slightly inflated. We set dimension to $p = 500$. The parameters investigated for the alternative are δ and β as described in Section 5.3. The scale difference among the variables introduced in the covariance structure redefines them as:

- Independence: $\Sigma_1 = \text{diag}(0, ..., p-1)$ and $\Sigma_2 = \text{diag}(0, ..., p-1)$.

- Square-Root-Decay: $\Sigma_1 = (0.5|j - j_1|^{-1/2}) + \text{diag}(0, ..., p-1)$ and $\Sigma_2 = (\rho|j - j_1|^{-1/2}) + \text{diag}(0, ..., p-1)$.

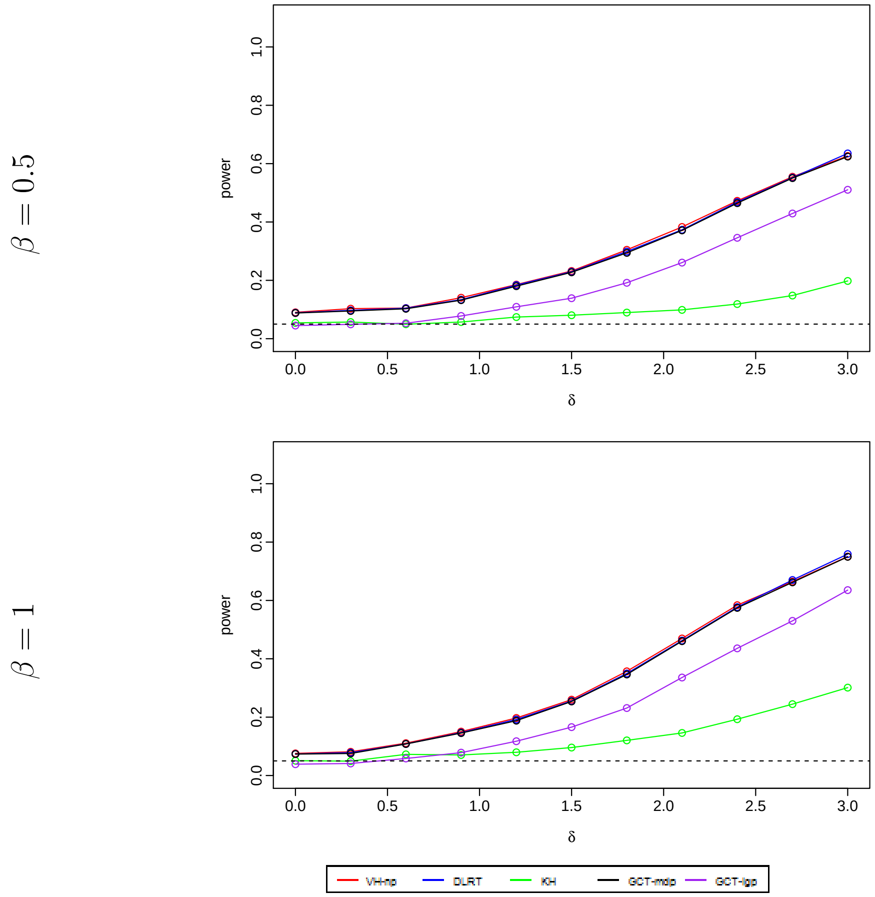


Figure 5.21: Power plots for all tests under shifting alternative. Errors are generated from Gamma distribution. Sample sizes are $(n_1, n_2) = (50, 50)$. Dimension is $p = 500$. Covariance structure is $diag(1, ..., p-1)$ and homoscedastic. Parzen Smoothing Window used is L=20.

From Figure 5.21, VH-np, DLRT and GCT-mdp are not affected by the scale difference. The test GCT-lgp shows better size with some power trade off. KH is affected considerably. Figure 5.22 shows that when dependence and reduction in sample size is introduced, power is reduced but patterns shown by the tests remain the same. We can also see that reducing sample size moderately affects the sizes of VH-np, DLRT and GCT-mdp.

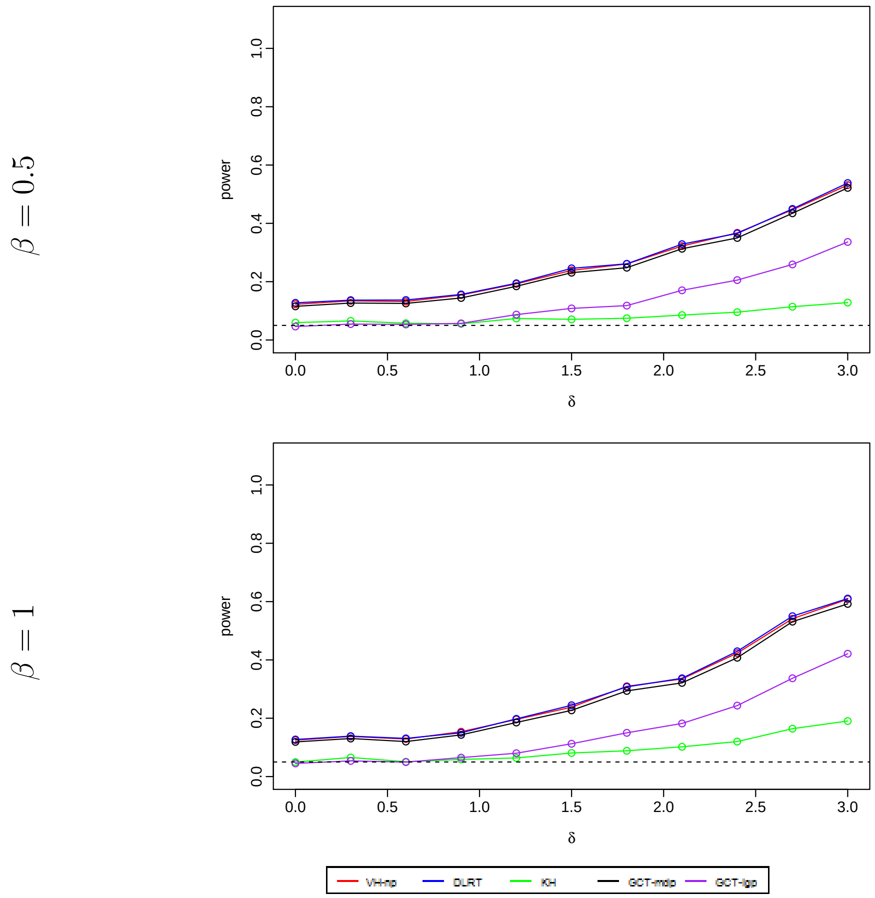


Figure 5.22: Power plots for all tests under shifting alternative. Errors are generated from Gamma distribution. Sample sizes are $(n_1, n_2) = (50, 50)$. Dimension is $p = 500$. Covariance structure is $\Sigma_{3i} + diag(1, ..., p-1)$ and homoscedastic. Parzen Smoothing Window used is L=20.

5.6 Electroencephalogram Data Analysis

The Electroencephalogram (EEG) data[1] we used in this analysis comes from a large study to examine associations of genetic predisposition to alcoholism. This data can be found at the University of California-Irvine Machine Learning Repository. For this study, sixty-four electrodes were placed in the subjects' scalps. Each one of these electrodes or channels are named according to the anatomical location of the placement of the electrode (Fp-Pre frontal, F-frontal lobe, T-temporal lobe, P-parietal lobe, O-occipital lobe and C-central). The name also contains a number,

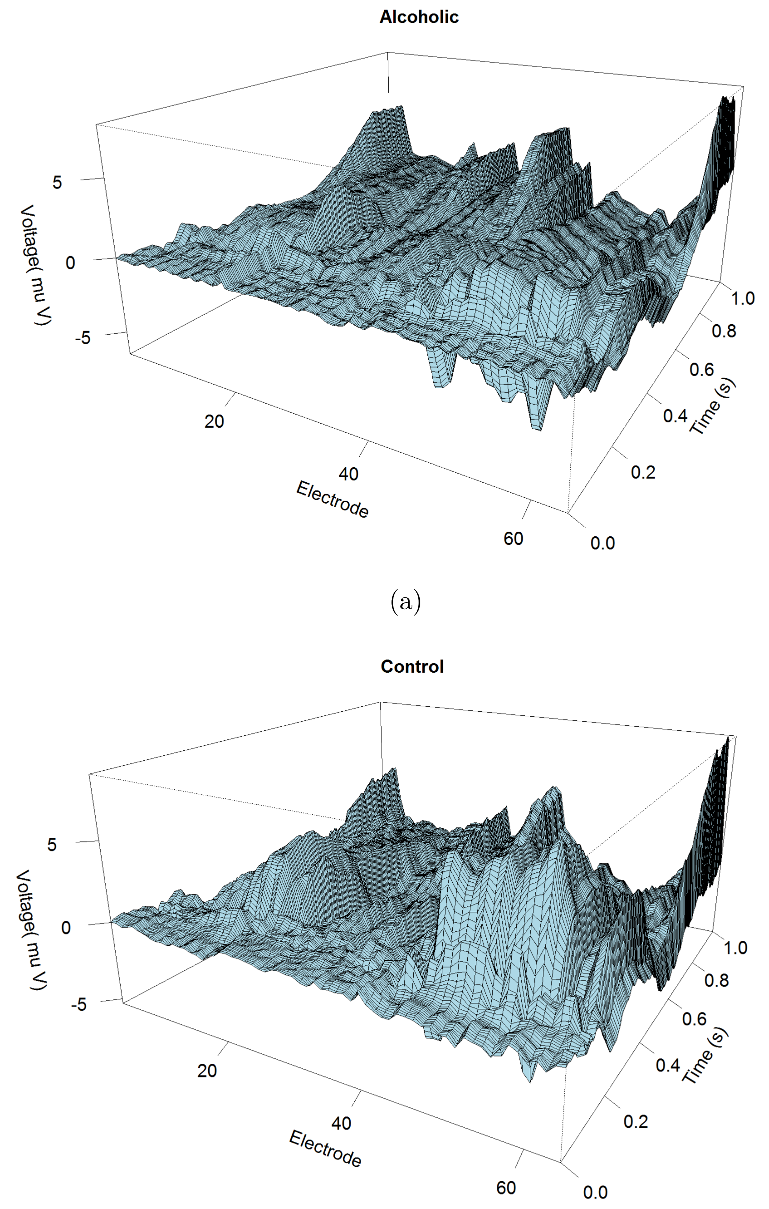
[1] data can be found at https://archive.ics.uci.edu/ml/datasets/eeg+database

which identifies the hemisphere of the brain (odd number for the left hemisphere, even number for the right hemisphere and letter z (zero) for the mid-line). The electrodes were used to measure Event-Related Potentials (ERP), which were recorded 256 times for one second.

In this study, there are two groups of subjects, the alcoholic and the control. Subjects were exposed to pictures of objects selected from a picture set; each subject was presented with either a single stimulus (S1) or to two stimuli (S1 and S2). For a more in-depth account of the EEG data, see Harrar and Kong (2016). ERP reading from an electrode indicates the level of electrical activity (in μvolts) in the region of the brain where the electrode is placed. In this dissertation, we analyze the data only for the single stimulus (S1) exposure using the methods investigated in this chapter. ERP data averages from the different objects for the two groups are plotted in Figure 5.23.

FDR adjusted p-values for channel-by-channel results are displayed in Tables 5.1 and 5.2. A summary of the pvalues is represented in Figure 5.24. Nonparametric and parametric tests are represented in different diagrams since they are testing different hypotheses. Each number inside the circles reprsents the number of significant channels for each test and the number on the lower right corner represents the number of channels that were not significant for any of the tests.

From panel (b) in Figure 5.24, we note that the VH-np declares the activity at 17 more channels to be significantly different compared to KH. Parametric methods coincide in the channels for which they find significance, a total of 47. In Figure 5.25, bar plot of the FDR adjusted p-values are shown for the nonparametric methods (VH-np and KH). The horizontal reference line (black dashed line) marks = 0.05 level of significance. Magnitude of discrepancies can be seen in it. It can be noted that KH declares significant differences in most locations (channels) that are away from the frontal lobe. Our proposed VH-np find significance in all of those locations and some more in the frontal lobe. It does not find significant difference in AF7, F8, FC4 and FP2 which all contain the letter F referring to their location in the frontal lobe.

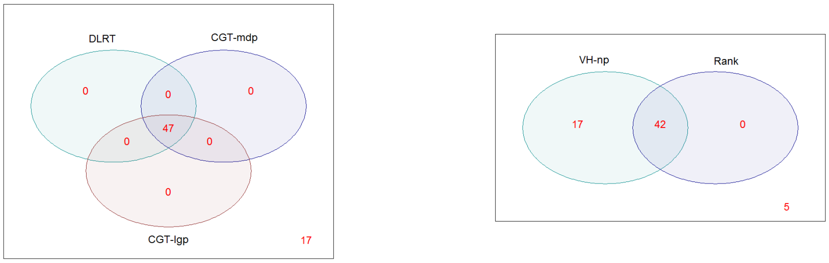


(a)

(b)

Figure 5.23: Plots of average ERP's (brain activity) per electrode over time by the Control and Alcoholic groups.

This experiment was described in Porjesz and Begleiter (2003). They report to expect most differences between both groups to be in between 300 and 700 ms. More recent studies such as Acharya, Sree, et al. (2012), Acharya, S, et al. (2014) and Bae et al. (2017) have investigated the same experiment from different approaches and concluded that groups have significant differences but did not get into details of



(a)

(b)

Figure 5.24: Venn diagram of counts of significant group effect differences for electrodes.

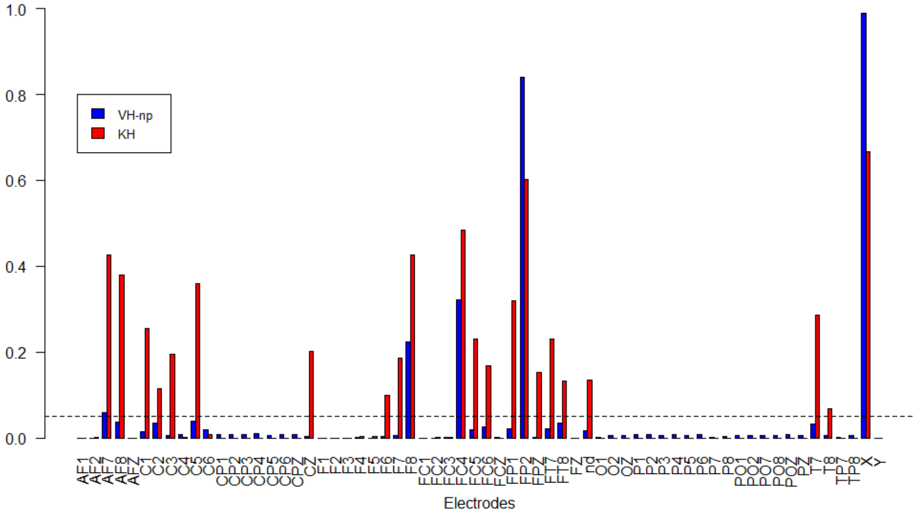


Figure 5.25: Plot of pvalues of nonparametric tests by electrodes.

which channels were different and which were not. In this analysis, we corroborate their results.

5.7 Discussion and Conclusion

The comparisons made in this chapter bring us to a few conclusions that can help in this field. Kong and Harrar have extended a well known parametric result from Chen and Qin (2010) to a nonparametric environment. Our proposed VH-np can also be viewed as a nonparametric two group extension of GCT-mdp. We illustrated how nonparametric extensions of DLRT or GCT compare to KH and VH-np under various settings.

It seems that tests statistics compared in simulation from Section 5.3 group in two sets that have homogeneous group behavior, one contains DLRT, VH-np and GCT-mdp, the other includes KH and GCT-lgp. DLRT, VH-np and GCT-mdp perform very similarly in terms of size and power under all settings investigated. KH and GCT-lgp perform very similarly in terms of size but KH shows a clear advantage in terms of power under the smaller n and smaller p settings. Both groups of tests behave similarly in larger setting for n and p, but sizes are clearly further from nominal values in the group of VH-np when sample size and dimension are smaller. Power is slightly advantageous in KH.

Mid-ranks were used instead of raw observations for the parametric tests in the simulation study and we observed that their behaviour is very similar to the nonparametric tests. This suggests that rank transforms of parametric tests might be studied in the future; we did not investigate this further.

Parametric and nonparametric tests are usually testing different hypotheses. Hypotheses in parametric tests are stated in terms of the mean vectors and in nonparametric tests are stated in terms of nonparametric relative effects. So, it is not fair to compare them in many circumstances. When comparing numerical simulations of the two strictly nonparametric methods, there is an apparent similarity under the settings investigated that does not show a clear analytical difference. As we can see in definitions of relative effect (4.2) and relative effect (5.1), there is one substantial difference. VH uses variable-by-variable ranks and KH uses overall ranks for the comparisons. This is a difference that could affect the behaviour of KH when variables have different scales. If there are no scale differences between variables, KH seems slightly advantageous, specially for size. However, if variables have different scales or units of measurements, it is shown in Section 5.5 that VH-np is not affected when KH is. More extensive numerical simulations may clarify this issue but we did not investigate it further.

Table 5.1: List of pvalues from all tests by electrode for the EEG dataset.

Electrode	VH-np	DLRT	KH	GCT-mdp	GCT-lgp
AF1	<0.001	<0.001	<0.001	<0.001	<0.001
AF2	<0.001	<0.001	0.0016	<0.001	<0.001
AF7	0.0587	1.0000	0.4273	1.0000	1.0000
AF8	0.0379	1.0000	0.3803	1.0000	1.0000
AFZ	<0.001	<0.001	<0.001	<0.001	<0.001
C1	0.0150	0.8320	0.2560	0.8335	0.8775
C2	0.0347	0.0683	0.1142	0.0674	0.0830
C3	0.0058	0.0875	0.1955	0.0873	0.1222
C4	0.0077	0.0244	<0.001	0.0245	0.0262
C5	0.0395	0.1773	0.3591	0.1757	0.1985
C6	0.0188	0.0272	0.0078	0.0270	0.0289
CP1	0.0086	0.0078	<0.001	0.0083	0.0087
CP2	0.0076	0.0061	<0.001	0.0069	0.0071
CP3	0.0077	0.0056	<0.001	0.0061	0.0062
CP4	0.0095	0.0058	<0.001	0.0067	0.0068
CP5	0.0068	0.0061	<0.001	0.0067	0.0070
CP6	0.0077	0.0078	<0.001	0.0088	0.0090
CPZ	0.0071	0.0178	<0.001	0.0179	0.0249
CZ	0.0043	0.0045	0.2014	0.0049	0.0059
F1	<0.001	<0.001	<0.001	<0.001	<0.001
F2	<0.001	<0.001	<0.001	<0.001	<0.001
F3	<0.001	<0.001	<0.001	<0.001	<0.001
F4	<0.001	<0.001	0.0041	<0.001	<0.001
F5	<0.001	<0.001	0.0041	<0.001	<0.001
F6	0.0045	0.0046	0.0991	0.0050	0.0059
F7	0.0067	0.0986	0.1850	0.1002	0.1316
F8	0.2228	0.5982	0.4273	0.5995	0.6719
FC1	<0.001	<0.001	<0.001	<0.001	<0.001
FC2	<0.001	<0.001	0.0014	<0.001	<0.001
FC3	<0.001	<0.001	0.0013	<0.001	<0.001
FC4	0.3217	0.8169	0.4841	0.8192	0.8676
FC5	0.0188	0.2091	0.2303	0.2094	0.2496
FC6	0.0263	0.0199	0.1693	0.0197	0.0232
FCZ	<0.001	<0.001	<0.001	<0.001	<0.001
FP1	0.0224	1.0000	0.3188	1.0000	1.0000
FP2	0.8405	1.0000	0.6014	1.0000	1.0000
FPZ	0.0020	0.6097	0.1515	0.6151	0.8136
FT7	0.0214	0.5667	0.2303	0.5629	0.6068
FT8	0.0347	0.1436	0.1336	0.1423	0.1599
FZ	<0.001	<0.001	<0.001	<0.001	<0.001

Table 5.2: List of pvalues from all tests by electrode for the EEG dataset cont.

Electrode	VH-np	DLRT	KH	GCT-mdp	GCT-lgp
nd	0.0172	0.0178	0.1354	0.0179	0.0208
O1	0.0020	0.0022	<0.001	0.0026	0.0027
O2	0.0050	0.0045	<0.001	0.0049	0.0052
OZ	0.0068	0.0066	<0.001	0.0071	0.0074
P1	0.0077	0.0053	<0.001	0.0057	0.0059
P2	0.0086	0.0061	<0.001	0.0067	0.0068
P3	0.0068	0.0045	<0.001	0.0050	0.0056
P4	0.0088	0.0058	<0.001	0.0067	0.0068
P5	0.0067	0.0045	<0.001	0.0049	0.0052
P6	0.0072	0.0049	<0.001	0.0057	0.0059
P7	0.0019	0.0013	<0.001	0.0017	0.0018
P8	0.0045	0.0030	<0.001	0.0036	0.0038
PO1	0.0068	0.0046	<0.001	0.0050	0.0056
PO2	0.0067	0.0045	<0.001	0.0049	0.0052
PO7	0.0050	0.0025	<0.001	0.0032	0.0033
PO8	0.0058	0.0045	<0.001	0.0049	0.0052
POZ	0.0086	0.0078	<0.001	0.0083	0.0087
PZ	0.0067	0.0045	<0.001	0.0049	0.0052
T7	0.0331	0.1566	0.2853	0.1558	0.1752
T8	0.0058	0.0117	0.0674	0.0116	0.0184
TP7	0.0015	0.0033	<0.001	0.0036	0.0038
TP8	0.0052	0.0046	<0.001	0.0050	0.0056
X	0.9904	0.2091	0.6678	0.2094	0.4580
Y	<0.001	<0.001	<0.001	<0.001	<0.001

Chapter 6 Discussion, Conclusion and Future Directions

In this dissertation, we studied parametric and nonparametric methods for high dimensional inference. We proposed two parametric statistics to test multiple group differences. We also proposed a fully nonparametric statistic to test two group differences.

The parametric tests are composites of variable-by-variable F-type statistics. One of them centers by an asymptotic mean value and is intended for moderate dimension. The other centers by its expanded mean correct up to order $n^{-3/2}$ and is devised for large dimension. The tests do not assume equal covariance matrix for the groups and, under weak dependence, follow asymptotic Normal distributions. We also showed that the rate of convergence for the statistics from the asymptotic expansion is higher as we develop the expansion further. The drawback of further expansions is having to estimate further moments with the corresponding sensitivity to outliers.

We investigated a nonparametric composite test statistic for two-group comparisons based on a variable-by-variable Wilcoxon-Mann-Whitney type statistic. Under mild moment conditions and weak dependence, the proposed test statistic is shown to asymptotically follow a Normal distribution. There is a great advantage of this method compared to the parametric methods when there exists correlation between the variables, especially for heavy tailed distributions.

We illustrated, via extensive simulation, how nonparametric extensions of some recent parametric methods and a nonparametric method compare to our proposed nonparametric test. Mid-ranks were used instead of raw observations for the parametric tests showing that behaviour is very similar to the nonparametric tests. This suggests that rank transforms of parametric tests might be studied successfully in the future. Since parametric and nonparametric tests are devised to test different hypotheses, we focused on nonparametric tests. When comparing numerical simulations of the strictly nonparametric methods, there is one substantial difference: our method uses variable-by-variable ranks and the other nonparametric method uses overall ranks for the comparisons. Therefore, if variables have different scales or units of measurements, it is shown that our method is advantageous.

One possible criticism to these results is that they require data to be ordered or indexed so that dependence decays based on index displacement.

Results presented make one ponder about what could be done to continue this research. Edgeworth expansion can be developed for the large-p versions of the parametric statistics. This expansion could lead to even more precision in the tests. Also, for the nonparametric test, we see a possibility to reduce the assumptions for the same conclusion. Consequently, working in this direction may produce a stronger result. All proposed tests are L_2 norm based which makes them competitive under weak but dense alternatives. An extension to other types of alternatives could be to add a power parameter to the univariate statistics instead of just having squared statistics. We can use this parameter to construct an adaptive test that will be powerful under various situations of sparsity and signal strength.

Bibliography

Hotelling, Harold (1931). “The Generalization of Student’s Ratio”. In: *The Annals of Mathematical Statistics* 2.3, pp. 360–378. DOI: 10.1214/aoms/1177732979.

Dempster, A. P. (1958). “A High Dimensional Two Sample Significance Test”. In: *The Annals of Mathematical Statistics* 29.4, pp. 995–1010. DOI: 10.1214/aoms/1177706437.

Leonov, V. P. and A. N. Shiryaev (1959). “On a Method of Calculation of Semi-Invariants”. In: *Theory of Probability & Its Applications* 4.3, pp. 319–329. DOI: 10.1137/1104031. eprint: https://doi.org/10.1137/1104031. URL: https://doi.org/10.1137/1104031.

Hosking, J. R. M., J. R. Wallis, and E. F. Wood (1985). “Estimation of the Generalized Extreme-Value Distribution by the Method of Probability-Weighted Moments”. In: *Technometrics* 27.3, pp. 251–261. DOI: 10.1080/00401706.1985.10488049. eprint: https://www.tandfonline.com/doi/pdf/10.1080/00401706.1985.10488049. URL: https://www.tandfonline.com/doi/abs/10.1080/00401706.1985.10488049.

Billingsley, P. (1995). *Probability and Measure*. Wiley Series in Probability and Statistics. Wiley. ISBN: 9780471007104. URL: https://books.google.com/books?id=z39jQgAACAAJ.

Politis, Dimitris N. and Joseph P. Romano (1995). “BIAS-CORRECTED NONPARAMETRIC SPECTRAL ESTIMATION”. In: *Journal of Time Series Analysis* 16.1, pp. 67–103. DOI: 10.1111/j.1467-9892.1995.tb00223.x.

Bai, Zhidong and Hewa Saranadasa (1996). “EFFECT OF HIGH DIMENSION: BY AN EXAMPLE OF A TWO SAMPLE PROBLEM”. In: *Statistica Sinica* 6.2, pp. 311–329. ISSN: 10170405, 19968507. URL: http://www.jstor.org/stable/24306018.

Brunner, Edgar and Ullrich Munzel (2000). “The Nonparametric Behrens Fisher Problem: Asymptotic Theory and a Small Sample Approximation”. eng. In: *Biometrical Journal* 42.1, p. 17 25. ISSN: 0323 3847.

Brunner, Edgar, Ullrich Munzel, and Madan L. Puri (2002). “The multivariate nonparametric Behrens-Fisher problem”. In: *Journal of Statistical Planning and Inference* 108.1–2, pp. 37–53. DOI: 10.1016/s0378-3758(02)00269-0.

Anderson, T. W. (Theodore Wilbur) (2003). *An introduction to multivariate statistical analysis*. 3rd ed. Wiley series in probability and statistics. Hoboken, NJ: Wiley-Interscience. ISBN: 0471360910.

Porjesz, Bernice and Henri Begleiter (2003). “Alcoholism and human electrophysiology”. In: *Alcohol research & health : the journal of the National Institute on Alcohol Abuse and Alcoholism* 27, pp. 153–60.

Gadbury, G. L et al. (2004). “Power and sample size estimation in high dimensional biology”. In: *Statistical Methods in Medical Research* 13.4, pp. 325–338. DOI: 10.1191/0962280204sm369ra.

Krishnamoorthy, K. and Jianqi Yu (2004). “Modified Nel and Van der Merwe test for the multivariate Behrens-Fisher problem”. In: *Statistics & Probability Letters* 66.2, pp. 161–169. DOI: 10.1016/j.spl.2003.10.012.

Bradley, Richard C. (2005). “Basic Properties of Strong Mixing Conditions. A Survey and Some Open Questions”. In: *Probability Surveys* 2.0, pp. 107–144. DOI: 10.1214/154957805100000104.

Srivastava, Muni Shanker and Yasunori Fujikoshi (2006). “Multivariate analysis of variance with fewer observations than the dimension”. In: *Journal of Multivariate Analysis* 97.9, pp. 1927–1940. DOI: 10.1016/j.jmva.2005.08.010.

Liao, J.G. and K.-V. Chin (2007). “Logistic regression for disease classification using microarray data: model selection in a large p and small n case”. In: *Bioinformatics* 23.15, pp. 1945–1951. DOI: 10.1093/bioinformatics/btm287.

Schott, James R. (2007). “Some high-dimensional tests for a one-way MANOVA”. In: *Journal of Multivariate Analysis* 98.9, pp. 1825–1839. DOI: 10.1016/j.jmva.2006.11.007.

Srivastava, Muni Shanker (2007). "Multivariate Theory for Analyzing High Dimensional Data". In: *JOURNAL OF THE JAPAN STATISTICAL SOCIETY* 37.1, pp. 53–86. DOI: 10.14490/jjss.37.53.

Srivastava, Muni Shanker and Meng Du (2008). "A test for the mean vector with fewer observations than the dimension". In: *Journal of Multivariate Analysis* 99.3, pp. 386–402. DOI: 10.1016/j.jmva.2006.11.002.

Zhang, Min, Dabao Zhang, and Martin T Wells (2008). "Variable selection for large p small n regression models with incomplete data: Mapping QTL with epistases". In: *BMC Bioinformatics* 9.1. DOI: 10.1186/1471-2105-9-251.

Srivastava, Muni Shanker (2009). "A test for the mean vector with fewer observations than the dimension under non-normality". In: *Journal of Multivariate Analysis* 100.3, pp. 518–532. DOI: 10.1016/j.jmva.2008.06.006.

Chen, Song Xi and Ying-Li Qin (2010). "A two-sample test for high-dimensional data with applications to gene-set testing". In: *The Annals of Statistics* 38.2, pp. 808–835. DOI: 10.1214/09-aos716.

Acharya, U. Rajendra, S. Vinitha Sree, et al. (2012). "AUTOMATED DIAGNOSIS OF NORMAL AND ALCOHOLIC EEG SIGNALS". In: *International Journal of Neural Systems* 22.03, p. 1250011. DOI: 10.1142/s0129065712500116.

Yamada, Takayuki and Muni S. Srivastava (2012). "A Test for Multivariate Analysis of Variance in High Dimension". In: *Communications in Statistics - Theory and Methods* 41.13-14, pp. 2602–2615. DOI: 10.1080/03610926.2011.581786. eprint: https://doi.org/10.1080/03610926.2011.581786. URL: https://doi.org/10.1080/03610926.2011.581786.

Aoshima, Makoto and Kazuyoshi Yata (2013). "Asymptotic Normality for Inference on Multisample, High-Dimensional Mean Vectors Under Mild Conditions". In: *Methodology and Computing in Applied Probability* 17.2, pp. 419–439. DOI: 10.1007/s11009-013-9370-7.

Brockwell, P.J. and R.A. Davis (2013). *Time Series: Theory and Methods*. Springer Series in Statistics. Springer New York. ISBN: 9781489900043. URL: https://books.google.com/books?id=DJ%5C_lBwAAQBAJ.

Cai, T. Tony, Weidong Liu, and Yin Xia (2013). "Two-sample test of high dimensional means under dependence". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 76.2, pp. 349–372. DOI: 10.1111/rssb.12034.

Srivastava, Muni Shanker, Shota Katayama, and Yutaka Kano (2013). "A two sample test in high dimensional data". In: *Journal of Multivariate Analysis* 114, pp. 349–358. DOI: 10.1016/j.jmva.2012.08.014.

Srivastava, Muni Shanker and Tatsuya Kubokawa (2013). "Tests for multivariate analysis of variance in high dimension under non-normality". In: *Journal of Multivariate Analysis* 115, pp. 204–216. DOI: 10.1016/j.jmva.2012.10.011.

Acharya, U. Rajendra, Vidya. S, et al. (2014). "Computer-aided diagnosis of alcoholism-related EEG signals". In: *Epilepsy & Behavior* 41, pp. 257–263. DOI: 10.1016/j.yebeh.2014.10.001.

Cai, T. Tony and Yin Xia (2014). "High-dimensional sparse MANOVA". In: *Journal of Multivariate Analysis* 131, pp. 174–196. DOI: 10.1016/j.jmva.2014.07.002.

Feng, Long et al. (2015). "Two-sample behrens-fisher problem for high-dimensional data". In: *Statistica Sinica*. DOI: 10.5705/ss.2014.048.

Ghosh, Anil K. and Munmun Biswas (2015). "Distribution-free high-dimensional two-sample tests based on discriminating hyperplanes". In: *TEST* 25.3, pp. 525–547. DOI: 10.1007/s11749-015-0467-x.

Gregory, Karl Bruce et al. (2015). "A Two-Sample Test for Equality of Means in High Dimension". In: *Journal of the American Statistical Association* 110.510. PMID: 26279594, pp. 837–849. DOI: 10.1080/01621459.2014.934826. eprint: https://doi.org/10.1080/01621459.2014.934826. URL: https://doi.org/10.1080/01621459.2014.934826.

Hu, Jiang, Zhidong Bai, et al. (2015). "On testing the equality of high dimensional mean vectors with unequal covariance matrices". In: *Annals of the Institute of Statistical Mathematics* 69.2, pp. 365–387. DOI: 10.1007/s10463-015-0543-8.

Wang, Lan, Bo Peng, and Runze Li (2015). "A High-Dimensional Nonparametric Multivariate Test for Mean Vector". In: *Journal of the American Statistical Association* 110.512, pp. 1658–1669. DOI: 10.1080/01621459.2014.988215.

Yamada, Takayuki and Tetsuto Himeno (2015). “Testing homogeneity of mean vectors under heteroscedasticity in high-dimension”. In: *Journal of Multivariate Analysis* 139, pp. 7–27. DOI: 10.1016/j.jmva.2015.02.005.

Harrar, Solomon W. and Xiaoli Kong (2016). “High-dimensional multivariate repeated measures analysis with unequal covariance matrices”. In: *Journal of Multivariate Analysis* 145, pp. 1–21. DOI: 10.1016/j.jmva.2015.11.012.

Bae, Youngoh et al. (2017). “Automated network analysis to measure brain effective connectivity estimated from EEG data of patients with alcoholism”. In: *Physiological Measurement* 38.5, pp. 759–773. DOI: 10.1088/1361-6579/aa6b4c.

Zhang, Jin Ting, Jia Guo, and Bu Zhou (2017). “Linear hypothesis testing in high-dimensional one-way MANOVA”. In: *Journal of Multivariate Analysis* 155, pp. 200–216. DOI: 10.1016/j.jmva.2017.01.002.

Hu, Zongliang, Tiejun Tong, and Marc G. Genton (2019). “Diagonal likelihood ratio test for equality of mean vectors in high-dimensional data”. In: *Biometrics*. DOI: 10.1111/biom.12984.

Vita

- Place of birth: San Javier, Murcia, Spain.
- Education
  - 2014 Master of Science in Statistics, University of Kentucky.
  - 2004 Master of Science in Statistics, Universidad Miguel Hernandez.
  - 2001 Bachelor of Science in Statistics, Universidad Miguel Hernandez.
- Professional Positions
  - 2016-2019 Data Management Specialist Sr., University of Kentucky.
  - 2015-2016 Research Assistant, University of Kentucky.
  - 2012-2015 Teaching Assistant, University of Kentucky.
  - 2004-2012 Secondary Math teacher, Schools in Spain and US.