FLORIDA STATE UNIVERSITY
COLLEGE OF ARTS AND SCIENCES

NONLINEAR MULTIVARIATE TESTS FOR HIGH-DIMENSIONAL DATA USING
WAVELETS WITH APPLICATIONS IN GENOMICS AND ENGINEERING

By

SENTHIL BALAJI GIRIMURUGAN

A Dissertation submitted to the
Department of Statistics
in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Degree Awarded:
Spring Semester, 2014
Senthil Balaji Girimurugan defended this dissertation on February 05, 2014.
The members of the supervisory committee were:

Eric Chicken
Professor Co-Directing Dissertation

Jinfeng Zhang
Professor Co-Directing Dissertation

Jon Ahlquist
University Representative

Minjing Tao
Committee Member

The Graduate School has verified and approved the above-named committee members, and certifies that the dissertation has been approved in accordance with university requirements.
To my wife Lela.
ACKNOWLEDGMENTS

As I embark on the final stages of my journey as a student of the sciences, I am thankful for spending them with the Department of Statistics at Florida State University.

I have come to know more than a handful of dedicated and highly intelligent professors. I express my sincere gratitude to Dr. Chicken for being the best mentor, academically and personally. His sense of humor and laid back attitude has given me solace during stressful times. In addition, I thank him for various discussions in statistical theory and the theory of wavelets that encouraged me to advance my skills further. I also thank Dr. Jinfeng Zhang for his research support and his advice on several research projects. His guidance has made me a better programmer aside from honing my skills in Biostatistics. I am also thankful to Dr. Jonathan Dennis and his entire lab for their contribution to my research through diligence, intellectually stimulating meetings, and a significant amount of data. Also, I express my utmost gratitude to my committee members, Dr. Minjing Tao, Dr. Debajyothi Sinha and Dr. Jon Ahlquist for being a significant part of this dissertation. Again, I am indebted to the chair and all faculty and staff of the Department of Statistics who have played a very vital role in this achievement.

Last, but not least, I thank Lela for supporting me through this process and sharing the inevitable stress accompanying dissertation and valuable research.
# TABLE OF CONTENTS

List of Tables ....................................................................................... viii
List of Figures ....................................................................................... xi
List of Abbreviations ............................................................................... xvi
List of Symbols ....................................................................................... xvii
Abstract ................................................................................................. xviii

1 Differences in Functional Data ............................................................ 1
   1.1 Examples of Functional Data ....................................................... 2
   1.2 Types of Differences .................................................................. 3

2 Testing Differences in Functional Data .................................................. 10
   2.1 Traditional Testing Methods ..................................................... 10
   2.2 Testing Methods for Functional Data ......................................... 11
      2.2.1 Definition of Statistical Tests for Functional Data ............... 11
      2.2.2 Criteria for Efficient Statistical Testing ............................... 12
      2.2.3 Pointwise Testing in Functional Data ................................. 13
      2.2.4 Measure Based Testing in Functional Data ....................... 13
   2.3 Transform Based Statistical Tests .............................................. 19

3 Wavelet Transforms in Data Analysis .................................................... 21
   3.1 Fourier Transform ................................................................... 21
   3.2 Wavelet Transforms .................................................................. 25
      3.2.1 Wavelets ........................................................................ 26
      3.2.2 Continuous Wavelet Transform (CWT) ............................ 27
      3.2.3 Discrete Wavelet Transform (DWT) .................................. 28
      3.2.4 Formal Definition of DWT .............................................. 28
      3.2.5 Wavelet Filters and Their Frequency Gain ....................... 31
   3.3 Matrix Notation of DWT .......................................................... 32
3.4 Inverse Wavelet Transform .................................................. 34
3.5 Wavelet Thresholding ....................................................... 36

4 A Non-parametric Wavelet Based Test 41
4.1 Traditional Multivariate Test ............................................. 41
   4.1.1 $T^2$ Test Statistic ......................................................... 42
   4.1.2 Wavelet Based Multivariate Statistic ............................... 44
   4.1.3 Choice of Noise Estimator, $\hat{\sigma}$ ................................ 45
   4.1.4 Distribution of $\kappa_\eta$ ................................................. 48
4.2 Thresholding Model using a Binomial Distribution ................. 52

5 Performance of a $\kappa_\eta$ Based Test and Practical Considerations 55
5.1 Performance Simulations .................................................. 55
   5.1.1 Measure of Difference .................................................. 55
   5.1.2 Mean Response ............................................................. 56
5.2 Simulation Setup and Results ............................................. 57
   5.2.1 Two-treatment Setup ..................................................... 57
   5.2.2 Multi-treatment Setup .................................................. 62
   5.2.3 Type of Noise Estimator .............................................. 65
   5.2.4 Choice of Wavelet Filter .............................................. 69
   5.2.5 Level of Threshold ($l_t$) ............................................... 71
   5.2.6 Zero Padding ............................................................... 72
5.3 Approximations for $\kappa_\eta$ ............................................. 72
   5.3.1 Normal Approximation .................................................. 74
   5.3.2 Chi-squared Approximation ......................................... 74
   5.3.3 Binomial-Normal Mixture Approximation ......................... 75

6 Applications of $\kappa_\eta$ Based Tests 77
6.1 Applications in Engineering .............................................. 77
   6.1.1 Data Description ......................................................... 77
   6.1.2 Methods ................................................................. 78
6.2 Applications in Genomics ................................................ 79
6.2.1 Data Types .................................................. 80
6.2.2 Effect of HDAC Inhibitors on HIV Cell Lines ..................... 82
6.2.3 Kaposi’s Sarcoma Associated Herpes Virus (KSHV) Nucleosome Distribution 83
6.2.4 Quality Control in Genomic Profiles ............................. 88
6.2.5 Cancer Data Analysis ........................................ 89
6.3 Applications of \( \kappa_\eta \) in Quantitative Finance .................... 92

7 Conclusions and Future Work .............................. 100

Appendices

A Proofs .................................................. 102
A.1 \( T^2 \) Unbalanced Design ................................... 102
A.1.1 Notation .................................................. 102
A.1.2 Two-treatment Design .................................... 102
A.1.3 Multiple Treatment Design ................................. 103
A.1.4 Wavelet Statistic, \( \kappa_\eta \) ................................. 104
A.1.5 Density and MGF of a \( \lambda^2 \) Distribution .................. 107
A.1.6 Density and MGF of a \( \lambda_M^2 \) Distribution ................. 108
A.1.7 Density of \( \kappa_\eta \) ........................................ 111
A.1.8 Mean and Variance of \( \kappa_\eta \) ............................. 112

B Tables .................................................. 115
B.1 Critical Value Tables ......................................... 116
B.2 iSeg: A Short Summary ....................................... 119

C Supplementary Results .................................. 120
C.1 KSHV-NGS Results ........................................ 121

Bibliography ............................................... 123
Biographical Sketch ........................................ 139
# LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.1</td>
<td>Summary of Type-I error for the simulations and the minimum detectable difference for the two treatment setup, balanced and unbalanced cases.</td>
<td>58</td>
</tr>
<tr>
<td>5.2</td>
<td>Summary of Type-I errors achieved by $\kappa_\eta$ and $\vartheta$ under various experimental settings. The 'la8' filter is employed with a full DWT decomposition and $l_t = 4$. For the balanced case, $r_i = 5, \forall i$ and for the unbalanced case, $r_1 = 10, r_2 = 8, r_3 = 5$ and $r_4 = 4$.</td>
<td>69</td>
</tr>
<tr>
<td>5.3</td>
<td>Summary of $\alpha$ values attained by a test involving $\kappa_\eta$. Here $r_i = 5, \forall i$ for the balanced design and for the unbalanced design $r_1 = 10, r_2 = 5, r_3 = 8$ and $r_4 = 4$, and $\sigma_i^2 = 1 \forall i$. The 'la8' filter is used with a full DWT decomposition and $l_t = 4$. The level of confidence set for the test in the simulations is 0.05 and 200 repetitive simulations are performed to obtain the estimated errors above.</td>
<td>70</td>
</tr>
<tr>
<td>5.4</td>
<td>Summary of $\alpha$ values attained by a test involving $\kappa_\eta$. Here $r_i = 5, \forall i$ for the balanced design and for the unbalanced design $r_1 = 10, r_2 = 5, r_3 = 8$ and $r_4 = 4$, and $\sigma_i^2 = 1 \forall i$. The 'la8' filter is used with a full DWT decomposition and $l_t = 4$. The level of confidence set for the test in the simulations is 0.05 and 200 repetitive simulations are performed to obtain the estimated errors above.</td>
<td>70</td>
</tr>
<tr>
<td>5.5</td>
<td>Summary of $\alpha$ values attained by a test involving $\kappa_\eta$ with varying filter choices. Here $r_i = 5, \forall i$ for the balanced design and for the unbalanced design $r_1 = 10, r_2 = 5, r_3 = 8$ and $r_4 = 4$, and $\sigma_i^2 = 1 \forall i$. The level of confidence set for the test in the simulations is 0.05 and 200 repetitive simulations are performed to obtain the estimated errors above.</td>
<td>71</td>
</tr>
<tr>
<td>5.6</td>
<td>Summary of $\alpha$ values attained by a test involving $\kappa_\eta$ with varying filter choices. Here $r_i = 5, \forall i$ for the balanced design and for the unbalanced design $r_1 = 10, r_2 = 5, r_3 = 8$ and $r_4 = 4$, and $\sigma_i^2 = 1, \sigma_2^2 = 2$ and $\sigma_4^2 = 0.5$. The level of confidence set for the test in the simulations is 0.05 and 200 repetitive simulations are performed to obtain the estimated errors above.</td>
<td>71</td>
</tr>
</tbody>
</table>
Summary of $\alpha$ values attained by a test involving $\kappa_\eta$ with varying filter choices. Here $r_i = 5, \forall i$ for the balanced design and for the unbalanced design, $r_1 = 10, r_2 = 5, r_3 = 8$ and $r_4 = 4$, and $\sigma^2_i = 1 \forall i$. The level of confidence set for the test in the simulations is 0.05 and 200 repetitive simulations are performed to obtain the estimated errors above. The level of threshold $l_t$ is varied from 4 to 10 and the la8 filer is employed in a full DWT decomposition with respect to the dimension of the functional response.

Summary of $\alpha$ values attained by a test involving $\kappa_\eta$ with varying filter choices. Here $r_i = 5, \forall i$ for the balanced design and for the unbalanced design, $r_1 = 10, r_2 = 5, r_3 = 8$ and $r_4 = 4$, and $\sigma^2_{1,3} = 1, \sigma^2_2 = 2$ and $\sigma^2_4 = 0.5$. The level of confidence set for the test in the simulations is 0.05 and 200 repetitive simulations are performed to obtain the estimated errors above. The level of threshold $l_t$ is varied from 4 to 10 and the la8 filer is employed in a full DWT decomposition with respect to the dimension of the functional response.

Classification results for the ISOLET database. Results summarized under 1, 2 and 3 can be found in [124], [42] and [55]. A rigorous summary of unpublished results is also provided in the ISOLET database.

Summary of p-values for each difference map defined in Equation (6.2) and the drugs that resemble HIV responses ($U_2$).

Number of genes with a significant change relative to 0 hour time point using a two-sample $\kappa_\eta$ based test.

Summary of top 10 genes demonstrating significant changes at the 24 hour time point. The genes are sorted according to their p-values obtained using a $\kappa_\eta$ based test.

Number of genes with a significant change relative to 0 hour time point using a two-sample $\kappa_\eta$ based test using NGS data for the KSHV experiment.

Summary of gene profiles that differ among biological replicates. The cell lines GM and IMR are included in the analysis for all degrees of digestion.

Number of genes with a significant change in Grade I (Patients 1357 and 4137) and
Grade III (Patients 386 and 873), using a two-sample \( \kappa \eta \) based test with \( \alpha = 0.05 \).

6.8 \( \kappa \eta \), p-value and the threshold (\( \lambda \)) used in the raw and normalization approaches of \( \kappa \eta \) with stock price data. The p-values are obtained using the normal approximation to \( \kappa \eta \) and the exact p-values are shown within parentheses.

6.9 Summary of p-values for the pairwise comparison procedure with \( \kappa \eta \) using the direct approach.

B.1 Critical values obtained using the exact distribution of \( \kappa \eta \) and its approximations with \( \alpha = 0.1 \).

B.2 Critical values obtained using the exact distribution of \( \kappa \eta \) and its approximations with \( \alpha = 0.05 \).

B.3 Critical values obtained using the exact distribution of \( \kappa \eta \) and its approximations with \( \alpha = 0.01 \).

C.1 Common list of genes that demonstrate a significant change between Normal and Tumour profiles among Grade I cancer patients 1357 and 4137.

C.2 Common list of genes that demonstrate a significant change between Normal and Tumour profiles among Grade III cancer patients 386 and 873.

C.3 Genes that show similarity in tumour behaviour among the genes that change in Grade I cancer patients, 1357 and 4137.
### LIST OF FIGURES

1.1 Daily precipitation amounts in millimeters recorded daily over an entire year from four different weather stations in Canada averaged over 1960 to 1994. ........................................ 3

1.2 A Gaussian process with \( n = 1024 \) and an identity kernel. That is, \( \kappa = I \) and the underlying mean function is \( f_0 = 6 \sin(6x), x \in (0, 1) \). Here, \( \sigma^2 = 1 \). The figure shows two replicates simulated using the aforementioned kernel and mean response. ........ 4

1.3 Global and local differences among signals. .......................................................... 5

1.4 Two signals where the underlying functional responses are different and are given by \( f_1 = 6 \sin(6x) \) and \( f_2 = 6 \cos(6x) \) respectively. Here, \( \sigma^2 = 1 \). ......................... 6

1.5 Differences in the functional response and covariance structures. .............................. 7

1.6 Two copy number variation (CNV) profiles from the cell lines GM05296 and GM13330.[150]. The base signal is assumed to have a mean zero and under normality assumptions, the jumps indicate insertions or deletions of copy number. ......................... 8

2.1 Plot of Type-I error (\( \alpha \)) against the dimension of the mean response. Three treatments are used with replicate sizes as \( r_1 = 2, r_2 = 3, r_3 = 2 \) respectively. The response vectors are in \( \mathbb{R}^3 \). ................................................................. 17

2.2 Effect of replicate sizes, homogeneity on MANOVA Type-I error. The level of the test is \( \alpha = 0.05 \) and an F-approximation for \( \Lambda_W \) is employed in the test. .............. 18

3.1 A noiseless functional response with several frequency components and its corresponding frequency spectrum obtained using Fourier Transforms. ................................. 23

3.2 A noisy functional response with several frequency components and its corresponding frequency spectrum obtained using Fourier Transforms corrupted by Gaussian noise. 25

3.3 Density estimators for the real (Re), imaginary (Im) [Left] and magnitude [Right] of the coefficients of the Fourier transform of the signal shown in Figure 3.2. ...................... 26
3.4 A Haar mother wavelet [top] and scaling function [bottom] .................................. 27

3.5 Commonly used Wavelet Filters and their frequency response ............................. 33

3.6 Proportion of non-zero (NZ) coefficients corresponding to a signal of length $\log_2(N)$, where ‘N’ is the sample size. The underlying “noiseless” function is $\sin(2x)$ and the “haar” wavelet filter is employed. ......................................................... 37

3.7 The underlying “noiseless” function (solid gray line) is $\sin(2x)$ and the “haar” wavelet filter is employed. The piecewise approximation of the function using the reduced set of wavelet coefficients with an IDWT is drawn using a solid black line. .................. 38

3.8 Density of a noisy response, $f(x) = \sin(2x)$ (solid gray line) and the density of the wavelet coefficients (dotted black line). The wavelet coefficients are obtained used a “haar” wavelet with $J = 15$. The sample size is $N = 2^{15}$. ................................. 39

4.1 A truncated normal distribution and a $\chi^2$ distribution using $n = 2^{15}$ with a hard threshold. The histogram of the observations and a smooth density estimator are shown. ................................................................. 46

4.2 Expected number of wavelet coefficients after a hard, universal threshold under $H_0$. Here, the sample size is $n = 2^J$, $J = 5 : 20$ and $l_t = 4$. A Haar wavelet is used with a full DWT decomposition. .......................................................... 52

4.3 Density of $\kappa_\eta$ using analytic expression in Lemma 4.1.7 (solid line) and the histogram of empirical $\kappa_\eta$ values obtained using 5000 simulations. A Haar wavelet is used in DWT and $l_t = 4$ with full decomposition corresponding to $l = \log_2(2^J)$. A hard, universal threshold is applied. Thus, $\lambda = 1.18\sqrt{J}$, $p = 1$ and $q = n/2^{lt}$ .... .............................. 54

5.1 Performance comparison in detecting global differences with $\kappa_\eta$ and $\vartheta$ in a two-treatment, balanced setup where $\sigma_1^2 = \sigma_2^2 = 1$ and $r_1 = r_2 = 2$ .................................................. 58

5.2 Performance comparison in detecting local differences with $\kappa_\eta$ and $\vartheta$ in a two-treatment, balanced setup where $\sigma_1^2 = \sigma_2^2 = 1$ and $r_1 = r_2 = 2$ .................................................. 59

5.3 Performance comparison in detecting global differences with $\kappa_\eta$ and $\vartheta$ in a two-
treatment, unbalanced setup where $\sigma_1^2 = \sigma_2^2 = 1$ and $r_1 = 3$ and $r_2 = 5$ .......................... 60

5.4 Performance comparison in detecting local differences with $\kappa_\eta$ and $\vartheta$ in a two-treatment, unbalanced setup where $\sigma_1^2 = \sigma_2^2 = 1$ and $r_1 = 3$ and $r_2 = 5$ .......................... 60

5.5 Performance comparison for $\kappa_\eta$ and $\vartheta$ in a two-treatment, balanced setup where $\sigma_1^2 = \sigma_2^2 = 1$ and $r_1 = 3$ and $r_2 = 5$ .......................... 61

5.6 Performance comparison for $\kappa_\eta$ and $\vartheta$ in a two-treatment, unbalanced setup where $\sigma_1^2 = \sigma_2^2 = 1$ and $r_1$ and $r_2 = 3$ .......................... 62

5.7 Performance comparison for $\kappa_\eta$ and $\vartheta$ in a two-treatment, balanced setup where $\sigma_1^2 = \sigma_2^2 = 1$ and $r_1 = r_2 = 10$ .......................... 63

5.8 Performance comparison for $\kappa_\eta$ and $\vartheta$ in a two-treatment, unbalanced setup where $\sigma_1^2 = 1$ and $\sigma_2^2 = 2$ and $r_1 = 10$ and $r_2 = 20$ for tests involving $\vartheta$ and $r_1 = 3$ and $r_2 = 5$ for tests involving $\kappa_\eta$. .......................... 64

5.9 Performance comparison for $\kappa_\eta$ and $\vartheta$ in a four-treatment, balanced setup where $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma_4^2 = 1$ and $r_1 = r_2 = r_3 = r_4 = 5$ .......................... 65

5.10 Performance comparison for $\kappa_\eta$ and $\vartheta$ in a four-treatment, balanced setup where $\sigma_1^2 = 1, \sigma_2^2 = 2, \sigma_3^2 = 0.5, \sigma_4^2 = 1$ and $r_1 = r_2 = r_3 = r_4 = 5$ .......................... 66

5.11 Performance comparison for $\kappa_\eta$ and $\vartheta$ in a four-treatment, unbalanced setup where $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma_4^2 = 1$ and $r_1 = 10, r_2 = 5, r_3 = 8, r_4 = 4$ .......................... 67

5.12 Performance comparison for $\kappa_\eta$ and $\vartheta$ in a four-treatment, unbalanced setup where $\sigma_1^2 = 1, \sigma_2^2 = 2, \sigma_3^2 = 0.5, \sigma_4^2 = 1$. Here $r_1 = 10, r_2 = 5, r_3 = 8, r_4 = 4$ for $\kappa_\eta$ based testing and $r_1 = 10, r_2 = 15, r_3 = 18, r_4 = 14$ for $\vartheta$ based testing. .......................... 68

5.13 Histogram of simulated $\kappa_\eta$ values with $p = 1, q = n/2^\nu$. Here, $n = 128$ (top) and $n = 1024$ (bottom). The Haar filter is used with a full DWT decomposition. .......................... 74

5.14 Histogram of simulated $\kappa_\eta$ values with $p = 1, q = n/2^\nu$. Here, $n = 4096$ (top) and $n = 8192$ (bottom). The Haar filter is used with a full DWT decomposition. .......................... 75
6.1 Functional responses for letters B and J from the ISOLET database training set.

6.2 Densities for letters B (black) and J (blue) from the ISOLET database training set.

6.3 Group-I GIV treated response corresponding to Chromosome 1. The figure shows both the actual and smoothed signal (spline-based).

6.4 Chromosomes 1 and 6 showing smoothed, drug treated responses from Group-I compared with untreated response from Group-II. The dotted line corresponds to Group-II response and the coloured lines correspond to the five drug treated difference maps.

6.5 Chromosomes 10 and 17 showing smoothed, drug treated responses from Group-I compared with untreated response from Group-II. The dotted line corresponds to Group-II response and the coloured lines correspond to the five drug treated difference maps.

6.6 Chromosomes 20 and X showing smoothed, drug treated responses from Group-I compared with untreated response from Group-II. The dotted line corresponds to Group-II response and the coloured lines correspond to the five drug treated difference maps.

6.7 Chromosomes 1 and 6 showing smoothed, drug treated responses from Group-I compared with untreated response from Group-II. The dotted line corresponds to Group-II response and the coloured lines correspond to the five drug treated difference maps. Chromosome 2 indicates the case demonstrating similarity between drug treated responses from Group-I and untreated response from Group-II.

6.8 The responses of four genes (two from Chromosome-2 and two from Chromosome-16) from all time points; 0 (dotted, black), 6 (solid, red), 12 (solid, green), 24 (solid, blue) and 48 (solid, cyan) are shown.

6.9 Responses from the KSHV experiment at the time points 0 (dotted, black), 6 (solid, red), 12 (solid, green), 24 (solid, blue) and 48 (solid, cyan) hours. The responses here are obtained from a next generation sequencing platform and the amplitude corresponds to counts (as represented in NGS type data).
6.10 Comparison of biological replicates from GM and IMR cell lines in the DoD experiment with NGS type data for genes TESK1 and SAMD11. ....................................................... 90

6.11 Comparison of biological replicates from GM and IMR cell lines in the DoD experiment with NGS type data for genes CPT1C and FNDC4. ....................................................... 91

6.12 Comparison of Normal (N) and Tumour (T) profiles in Grade I (1357, 4137) and Grade III (386, 873) patients for genes EID2 and USP51. ....................................................... 93

6.13 Comparison of Normal (N) and Tumour (T) profiles in Grade I (1357, 4137) and Grade III (386, 873) patients for genes LOC643837 and TMEM254-AS1. ....................................................... 94

6.14 Three sample paths of an Ornstein-Uhlenbeck process generated with an underlying mean response, $f_0$ as defined in equation 5. ....................................................... 95

6.15 The average of the observed sample plots (with replicates $r=10$ and $r=1000$) plotted alongside the actual functional response. It can be noted that the average becomes an “exact approximation” of $f_0$, the underlying function, as the replicate size increases. 96

6.16 Adjusted Closing prices for the year 2000. ....................................................... 97

6.17 Average Adjusted Closing prices obtained using daily average stock price sample paths from 1983 to 2013 (August 21, 2013) ....................................................... 98

C.1 The top four genes called by $\kappa_\eta$ as significantly changed in the KSHV-NGS analysis. . 122
## LIST OF ABBREVIATIONS

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ANOVA</td>
<td>ANalysis Of VAriance.</td>
</tr>
<tr>
<td>bp</td>
<td>Base Pair</td>
</tr>
<tr>
<td>CWT</td>
<td>Continuous Wavelet Transform.</td>
</tr>
<tr>
<td>DFT</td>
<td>Discrete Fourier Transform.</td>
</tr>
<tr>
<td>DoD</td>
<td>Degree of Digestion</td>
</tr>
<tr>
<td>DNA</td>
<td>Deoxyribonucleic acid</td>
</tr>
<tr>
<td>FFT</td>
<td>Fast Fourier Transform.</td>
</tr>
<tr>
<td>GP</td>
<td>Gaussian Process</td>
</tr>
<tr>
<td>ID</td>
<td>Identification</td>
</tr>
<tr>
<td>IDWT</td>
<td>Inverse Discrete Wavelet Transform.</td>
</tr>
<tr>
<td>ISOLET</td>
<td>ISOlated LETter recognition database.</td>
</tr>
<tr>
<td>KSHV</td>
<td>Kapsoi’s Sarcoma associated Herpes Virus.</td>
</tr>
<tr>
<td>MAD</td>
<td>Median Absolute Deviation.</td>
</tr>
<tr>
<td>MANOVA</td>
<td>Multivariate ANalysis Of VAriance.</td>
</tr>
<tr>
<td>UCSC</td>
<td>University of California, Santa Cruz Genome Browser.</td>
</tr>
</tbody>
</table>
LIST OF SYMBOLS

Symbols and notation commonly used in this document are summarized below.

- $M$: The median operator.
- $\mathcal{W}$: The DWT operator.
- $\lambda$: The DWT threshold parameter.
- $\sigma$: The noise estimator. It corresponds to either the standard deviation or any other estimator of noise.
- $\vartheta$: The Hotelling $T^2$ statistic.
- $\kappa$: The proposed wavelet statistic.
- $\kappa_{\eta}^{(p,q,\lambda)}$: The proposed wavelet statistic with its parameters: p, q and $\lambda$.
- $E[X]$: The expectation of a random variable.
- $Var[X]$: The variance of a random variable.
ABSTRACT

Gaussian processes are not uncommon in various fields of science such as engineering, genomics, quantitative finance and astronomy, to name a few. In fact, such processes are special cases in a broader class of data known as functional data.

When the underlying mean response of a process is a deterministic function, the resulting data from these processes are functional responses and specialized statistical tools are required in their analysis. The methodology discussed in this work offers non-parametric tests that can detect differences in such data with greater power and good control of Type-I error over existing methods. The tests are non-parametric as they do not make assumptions about the underlying model of the data or the functional structure. In other words, the underlying functions need not be defined by any mathematical model.

The incorporation of wavelet transforms makes the test an efficient approach due to its de-correlation properties. These tests are designed primarily to handle functional responses from multiple treatments simultaneously and generally are extensible to high dimensional data. The sparseness introduced by wavelet transforms is another advantage of this test when compared to traditional tests. In addition to offering a theoretical framework, several applications of such tests in the fields of engineering, genomics and quantitative finance are also discussed.

A new test statistic \((\kappa_\eta)\) is proposed to test differences in functional responses obtained from multiple experiments. Several \(\kappa_\eta\)-based power simulations are carried out using a variety of experimental settings which demonstrate good power and control of Type-I error over traditional methods (such as a Hotelling \(T^2\) test and Multivariate ANOVA). Chapter 5 discusses these simulations and their results in detail. The desirable properties of the test statistic are due to the mathematical features of wavelet transforms. In Chapter 3, a gentle, yet elaborate discussion of wavelet transform is provided and Chapter 4 establishes a statistical framework for \(\kappa_\eta\).
In Chapter 6, \( \kappa_\eta \) is employed in testing differences in speech and genomic profiles. The statistic is used to classify English letters using observed speech profiles. The profiles are from the ISOLET database obtained from the UCI Machine Learning repository. In genomics, the statistic is used to detect differences in microarray and Next Generation Sequencing (NGS) based functional responses obtained from experiments conducted by Dr. Jonathan Dennis et al. at Florida State University. These experiments are primarily associated with the detection of biological significance in nucleosome occupancy profiles obtained from tissue samples related to cancer and HIV research.
CHAPTER 1

DIFFERENCES IN FUNCTIONAL DATA

Functional data came into prevalence in the early 1980s when data (without temporal ordering) was viewed as curves [94]. Time series analysis could be perceived as a precursor to the idea of functional data because its methods estimated an underlying smooth function from data with a natural, temporal ordering. Since then data has been perceived as functional in various areas of research such as weather pattern analysis and handwriting recognition [134]. In addition, most responses in processes observed in engineering and other scientific areas such as genomics and quantitative finance are functions.

A n-dimensional functional response from the $i^{th}$ treatment (any response from an experiment) is denoted as,

$$Y_i = f_i + \epsilon_i$$  \hspace{1cm} (1.1)

where $i=1, 2, \ldots t$, where ‘t’ is the number of treatments and $Y_i, f_i, \epsilon_i \in \mathbb{R}^n$. $f_i$ is a deterministic function (the underlying mean response) and $\epsilon_i$ is the noise component. Here, $f_i$ is the underlying functional response corresponding to the $i^{th}$ treatment. It should be noted that the “ideal” definition of functional data perceives data as smooth curves with very little or no noise. However, “noisy” functional data is the primary focus of analysis in this work. That is, an underlying smooth curve with additive noise is the response. Of course, the methods discussed in this work can be applied to “noiseless” functional data where data is interpreted as curves.

In a multivariate case, each variable results in a functional response and the notation for multivariate, functional data can be deduced from equation (1.1). Consider an n-dimensional function response from ‘t’ different treatments. Let ‘j’ denote the number of replicates from the $i^{th}$ treatment.
The functional response can be denoted as,

\[ Y_{ijk} = f_k + \mu_{ik} + \epsilon_{jk} \] (1.2)

\[ = F_{ik} + \epsilon_{jk} \] (1.3)

where, \( \mu_{ik} \) is the effect of the \( i^{th} \) treatment. This effect could either be a constant or a unique functional response specific to the \( i^{th} \) treatment. The terminology from design of experiments can be extended to functional data also. Here, we can define balanced and unbalanced designs [111] for functional data. Therefore, in a balanced design the replicate sizes are identical \( r_1 = r_2 = r_3 = \ldots = r_t = r \) and otherwise for an unbalanced design.

1.1 Examples of Functional Data

Functional data appear in many fields of science and engineering. However, statistical tools to analyze such data have become popular in recent decades due to the increase in data dimensions and its complexity. For instance, the precipitation data collected from four different weather stations are shown in Figure 1.1 [132]. The entire dataset consisting of daily temperature and precipitation measurements from 35 Canadian weather stations can be accessed from the R package, \texttt{fda} [133]. It can be noted from Figure 1.1 that the underlying response is a smooth function. Another commonly known process that can be identified as functional data is a Gaussian process [136]. A Gaussian process (GP) is a stochastic process where every observation with respect to time or space is a Gaussian random variable. A continuous time Gaussian process, \( \{X_t\} \) is defined as,

\[ \{X_t\} = \text{GP}(f_0, \kappa) \] (1.4)

where \( f_0 \) is the underlying mean response (a smooth function) and \( \kappa \) describes the covariance structure of the process (commonly known as the kernel function in Machine Learning applications [81, 24]). In statistics, such processes are discrete realizations of \( \{X_t\} \) and such a process sampled at ‘n’ equal intervals is denoted by \( X_n \). Thus, \( X_n \in \mathbb{R}^n \) and \( \kappa \in \mathbb{R}^{n \times n} \). The simplest covariance structure would assume identical variances, hence, \( \kappa = \sigma^2 I \) and \( I \in \mathbb{R}^n \) is an identity matrix. \( \sigma^2 \) is
Figure 1.1: Daily precipitation amounts in millimeters recorded daily over an entire year from four different weather stations in Canada averaged over 1960 to 1994.

The variance of the process. Figure 1.2 shows a Gaussian process with an identity kernel where the underlying mean response is \( f_0 = 6 \sin(6x) \). Here ‘x’ corresponds to a unit of time or space.

1.2 Types of Differences

In engineering and science, processes are observed and monitored in order to maintain quality and aid in decision-making that benefits a specific cost function. For example, in profile monitoring, a process is said to be in-control (IC) when the process is functioning normally and within agreeable control limits that guarantee the expected quality from the process. A process that is out-of-control (OOC) indicates a deviation from the IC profile and will have a significant effect on the quality of the process. Most statisticians are interested in detecting such differences between IC and OOC processes. This problem of profile monitoring can be further extended to a classification problem \([108]\) or a change point detection problem \([101, 21]\).
Figure 1.2: A Gaussian process with $n = 1024$ and an identity kernel. That is, $\kappa = I$ and the underlying mean function is $f_0 = 6 \sin(6x), x \in (0, 1)$. Here, $\sigma^2 = 1$. The figure shows two replicates simulated using the aforementioned kernel and mean response.

Suppose the in-control signal (profile) is $f_0$. An out-of-control signal, $f_1$, can exhibit several types of differences. Some of the differences are classified in a generic manner as,

- **Local difference**, a difference that happens on a small interval and it is usually hard to detect. That is,

$$\exists i \in \{1, 2, \ldots, n\} : f_1(x_i) \neq f_0(x_i)$$ (1.5)

An example of such a difference is a spike in a signal. Such a difference could also represent a segment of the signal with noticeable differences as shown in Figure 1.3 (a).
(a) Two signals where the underlying functional response is given by $f = 6 \sin(6x)$ and $\sigma^2 = 1$. The OOC signal, however, has a shift in mean locally between 100 and 250.

(b) Two signals where the underlying functional response is $f = 6 \sin(6x)$ and $\sigma^2 = 1$. The OOC signal, however, has a shift in the mean globally.

Figure 1.3: Global and local differences among signals.

- **Global difference**, a difference that happens on the entire domain of the signal. Such a difference is noticeable and most test statistics can detect such a difference easily. That is,

  $$ f_1(x_i) \neq f_0(x_i), \forall i \in \{1, 2, \ldots, n\} $$

  (1.6)

  A mean shifted version of a signal is a good example of such a difference as shown in Figure 1.3 (b).

- **Functional difference**, a difference where the underlying functional response has changed from the in-control process. That is,

  $$ f_1 = f + \epsilon $$

  (1.7)

  $$ f_0 = g + \epsilon $$

  (1.8)

where $f$ and $g$ are the two underlying functional responses. Such a difference is depicted in Figure 1.4.
Figure 1.4: Two signals where the underlying functional responses are different and are given by \( f_1 = 6 \sin(6x) \) and \( f_2 = 6 \cos(6x) \) respectively. Here, \( \sigma^2 = 1 \).

- **Co-variance Structure difference**, a difference where the underlying covariance structure differs from the in-control process. That is,

\[
  f_1 = f + \epsilon \\
  f_0 = f + \delta
\]  

(1.9) (1.10)

where \( \epsilon \) and \( \delta \) are the noise structures determined by some covariance matrices \( \Sigma_1 \) and \( \Sigma_2 \).

- **Functional and Co-variance Structure difference**, a difference where the underlying covariance structure as well as the functional response differ from the in-control process. That is,

\[
  f_1 = f + \epsilon \\
  f_0 = g + \delta
\]  

(1.11) (1.12)
(a) Two signals where the underlying noise structure is different. The mean response is assumed to be the same as \( f = 6 \sin(6x) \). However, the IC signal has a covariance structure defined by \( \Sigma_1 = \sigma^2 I, \sigma^2 = 1 \) and \( \Sigma_2 = e^{-|t_i - t_j|/l} \), an Ornstein-Uhlenbeck kernel function.

(b) Two signals where the underlying noise structure and the underlying functional response are different. The mean responses are assumed to be different and are given by \( f_1 = 6 \sin(6x) \) and \( f_2 = 6 \cos(6x) \). Also, the IC signal has a covariance structure defined by \( \Sigma_1 = \sigma^2 I, \sigma^2 = 1 \) and the OOC signal has a covariance structure defined by \( \Sigma_2 = e^{-|t_i - t_j|/l} \), an Ornstein-Uhlenbeck kernel function.

Figure 1.5: Differences in the functional response and covariance structures.

where \( f \) and \( g \) are the two underlying functional responses. \( \epsilon \) and \( \delta \) are the noise structures determined by covariance matrices \( \Sigma_1 \) and \( \Sigma_2 \) respectively. An example for such a difference is shown in Figure 1.5 (b).

Intuitively, it can be understood that differences are easy to detect when the underlying functional responses differ entirely. However, a difference in the covariance structure or a local difference is usually difficult to detect and requires thorough statistical analysis to determine the significance of the underlying effects.

The differences discussed above arise commonly in many situations and a similar problem arises in genomics. Gene expression profiles obtained through DNA microarray analysis [112] present information about the DNA in regulating various functions of organisms. Some of the most important features in gene expression analyses include Copy Number Variation (CNV) [73], DNA Methylation [86] and Single Nucleotide Polymorphisms (SNP, pronounced as ‘snip’) [114, 157]. Copy Number Variations (CNV) are particularly significant as they have demonstrated effects in several diseases including cancer [30]. Common forms of copy number variations include deletion of a segment of the DNA or duplication of a segment of the DNA.

From a statistical perspective, a deletion is a decrease in the mean of segment and an insertion
would correspond to an increase in the mean of the segment. Also, such shifts in the mean are local and therefore fall under the category of local differences. Most copy number analyses (CNA) involve the detection of such segments [52, 127, 128]. The method proposed in this work could serve as a precursor to such analysis. That is, the testing method would identify the need for localized detection of copy number variations in long profiles from multiple treatments. This is one of many applications of this testing method and it can be applied to various fields of science involving similar scenarios. For instance, the proposed test can also determine if a gene is significant among profiles from multiple treatments.

In Figure 1.6, each profile from cell lines GM05296 and GM13330 are functional data. That is, the underlying response is a function, although in this experiment [150] the response is assumed to be a flat line with mean zero and no noise. An apposite question in this example is to identify if the differences are significant per profile. Also, if two or more such profiles are compared simultaneously, how would one test for significance among those? This problem is addressed by the proposed test and discussed in this work.

Figure 1.6: Two copy number variation (CNV) profiles from the cell lines GM05296 and GM13330. [150]. The base signal is assumed to have a mean zero and under normality assumptions, the jumps indicate insertions or deletions of copy number.
In the following chapters, existing methods and the proposed testing method are discussed. A mathematical framework is also offered along several simulations and examples motivated by genomics, quantitative finance and other fields of science.
CHAPTER 2

TESTING DIFFERENCES IN FUNCTIONAL DATA

In this chapter we discuss testing methods that can be used to determine differences in functional data. In statistics, significance testing is carried out in two ways:

- **Parametric testing**- The data is assumed to follow a distribution. Therefore, the underlying test statistic follows a known distribution. The test is dependent on the characteristics of the data such as the parameters that model the data.

- **Non-parametric testing**- The data does not follow a known distribution. Therefore, the test is free of distributional assumptions. Such tests are also called “distribution-free” tests. Another case of non-parametric testing allows distribution-based assumptions for the variables but no restrictions are placed on the assumption of the underlying model describing the data [71]. A simple example for the latter non-parametric approach is non-parametric regression. In this work, the proposed test does not make any underlying assumptions about the structure of the underlying function and falls under the category of non-parametric testing.

### 2.1 Traditional Testing Methods

Most statistical testing methods test for a parameter $\Theta \in \mathbb{R}^m$ for the population. As mentioned above, when the parameter follows a known distribution, the resulting test involving such a parameter is a parametric test. For instance, consider a random variable $X \in \mathbb{R}^n$

$$X \sim \Phi(\Theta)$$  \hspace{1cm} (2.1)

where $\Phi$ is a probability distribution [69]. A test involving the random variable $X$ would be based on the sufficient statistics for $\Theta$. The mean is a common parameter tested in a population. Under the assumptions of normality (that is, $\Phi$ is a normal distribution), a t-test [98] is the most preferred statistical test. Although the t-test is effective under such circumstances, such a
traditional test cannot be applied for functional data. Since functional data preserve a structure, averaging functional data would lose structural information and any test involving such a statistic would fail in determining statistical significance. Therefore, traditional testing methods cannot be used directly to test differences in functional data.

2.2 Testing Methods for Functional Data

Even though traditional methods do not serve as the optimal choice for testing differences in functional data, they offer a sound mathematical framework for effective testing methods involving functional data. Consider a functional response, $Y \in \mathbb{R}^n$

$$Y = f(\xi, x) + \epsilon \quad (2.2)$$

where $\epsilon \in \mathbb{R}^n$ is the random noise component and $\xi \in \mathbb{R}^d$ are the parameters that describe the structure of the underlying functional response. It should be noted that ‘d’ doesn’t always equal to ‘n’ in practice. A test involving such a functional response, by definition, is a parametric test.

If we assume that the underlying structure described by $\xi$ is arbitrary, that is,

$$Y = f(x) + \epsilon \quad (2.3)$$

a test involving $Y$ is non-parametric even under the assumption of normality for $\epsilon$. In this chapter, parametric methods and non-parametric methods used to analyze data from situations 2.2 and 2.3 are discussed in detail.

2.2.1 Definition of Statistical Tests for Functional Data

A test for simultaneous comparison of multivariate profiles from multiple treatments can be defined as follows.

Let us consider a set of $t$ n-dimensional responses from $t$ treatments,

$$H_0 : f_1 = f_2 = f_3 = \ldots = f_t \quad (2.4)$$

$$H_1 : f_k \text{ is different for some } k, k \in \{1, 2, \ldots, t\} \quad (2.5)$$
The profiles are observed by samples of size \( n \) as defined in Section 1.2 where the expected value of the mean response over all replicates will correspond to \( f \). The above definition generalizes a test for detecting global differences as well as local differences. Here, \( f_1, f_2, f_3, \ldots f_t \) denote the responses from \( t \) treatments; \( H_0 \) denotes the null hypothesis and \( H_1 \) denotes the alternative hypothesis.

### 2.2.2 Criteria for Efficient Statistical Testing

Several measures are available to determine the efficiency of a statistical test. The most common criteria to measure the efficiency of a statistical test is the Type-I error (\( \alpha \)) and Type-II error (\( \beta \)) [119, 31]. However, \( \beta \) is measured in terms of the power of a statistical test. The Type-I error is defined as the probability of rejecting the null hypothesis when \( H_0 \) is true. That is,

\[
\alpha = \Pr_{\Theta_0}(\Lambda \in \Omega) \tag{2.6}
\]

where, \( \Lambda \) is the test statistic, \( \Theta_0 \) refers to the parameter space defined under the null hypothesis (\( H_0 \)) and \( \Omega \) is the rejection region. Similarly, the power of a statistical test is defined as the probability of rejecting the null when the null hypothesis is false. That is,

\[
\beta^* = \Pr_{\Theta_1}(\Lambda \in \Omega) \tag{2.7}
\]

where, \( \Theta_1 \) refers to the parameter space defined under the alternative hypothesis. Thus, the Type-II error is defined as,

\[
\beta = 1 - \Pr_{\Theta_1}(\Lambda \in \Omega) \tag{2.8}
\]

Since tests involving functional data express \( \Lambda \) as a function of the multivariate response, measures that are commonly used in testing univariate tests can be readily extended to testing with functional data. In addition, other measures such as \( F_1 \) scores [129] and area under the ROC curve (AUC) [97, 174] can also employed to measure the performance of a statistical test.
2.2.3 Pointwise Testing in Functional Data

Many tests can be derived by extending tests applicable for univariate data to functional data. However, the effectiveness of such tests vary dramatically and tests involving functional data involve several considerations such as dimensionality and Familywise Error Rate (FWER)[80, 155]. Therefore, tests tailored specifically for functional data have been studied in the past few decades.

Consider the analysis of variance for functional data as defined in (1.3). In this case, a simple approach is a point-wise testing method. Here, a test statistic is computed for every point in an n-dimensional functional response, across all treatments. An F test statistic was proposed by [135] with the pointwise approach. The statistic was defined as,

\[
F(x_n) = \frac{\text{SSTr}(x_n)/(t - 1)}{\text{SSE}(x_n)/(n - t)}
\]

(2.9)

However, such methods generally lose Type-I error control and would require familywise error rate (FWER) controlling procedures. Although these controlling procedures are easy to implement, each of these methods pose their own demerits in certain situations. For instance, the test statistic would have an F-distribution with numerator and denominator degrees of freedoms as \((t - 1)\) and \((n - t)\) only when the data is assumed to follow a normal distribution.

In addition, as \(n \rightarrow \infty\), the computational complexity of such a test increases tremendously [48]. Some testing strategies have employed the Welch-Satterthwaite approximation [144, 166] to obtain an approximate distribution for the test statistic in pointwise testing procedures [171].

2.2.4 Measure Based Testing in Functional Data

In many statistical tests, the test statistic is expressed as a function of the data. That is, the test statistic is usually a function of the sufficient statistic \(\Psi(T(X))\), where \(T(X)\) is the sufficient statistic [149]. Since \(T(X)\) is mostly a function of the data in \(l_2\), most hypothesis tests in univariate cases are measure-based. The t-test and F-test are the most common among the class of measure based univariate tests. One can use a similar approach with functional data. That is, in order to test differences in functional data one can use a reduced form of the original data rather than the entire data. If data is n-dimensional, a pointwise testing approach would require computations in
the order of $O(n)$. However, the computational time of a measure based test is $O(1)$. It should be noted that the computational complexity in obtaining the measure is assumed to be negligibly small. Although measure based testing is computationally efficient, one of the main cons of such an approach is identifying the measure that would result in a powerful test. The generalized Likelihood Ratio Test (LRT) represents many of the univariate measure based tests that can be extended to multivariate analysis. Suppose $X$ is a random variable distributed with some parameter $\theta$ in a parametric family $f$. A realization of such an $X$ is $(X_1, X_2, \ldots, X_n)$. A hypothesis test to decide on the value of $\theta$ is defined as,

\begin{align}
H_0 & : \theta = \theta_0 \\
H_1 & : \theta = \theta_1
\end{align}

(2.10) (2.11)

A uniformly most powerful (UMP) test [139] as defined by Neyman-Pearson Lemma [118, 120] is defined by a likelihood ratio. That is,

$$
\Lambda = \frac{L(X_1, X_2, \ldots, X_n | \theta = \theta_0)}{\sup_{\theta \in \{\theta_0, \theta_1\}} L(X_1, X_2, \ldots, X_n | \theta)}
$$

(2.12)

$H_0$ is rejected when, $\Lambda(X) < c$, where $0 < c < 1$ and it is chosen based on the level $\alpha$,

$$
\theta = \begin{cases} 
\theta_0, & \Lambda(X) > c \\
\theta_1, & \Lambda(X) \leq c
\end{cases}
$$

(2.13)

For data from the exponential family of distributions [99] where $\eta(\theta)$, $\omega(x)$ and $\gamma(\theta)$ are known functions, the density is given by,

$$
f(x|\theta) = h(x) \exp(\eta(\theta)\omega(x) - \gamma(\theta))
$$

(2.14)
Thus, the likelihood described in (2.12) becomes,

\[ L((X_1, X_2, \ldots, X_n)|\theta) \propto \prod_{i=1}^{n} h(X_i) \exp(\eta(\theta)\omega(X_i) - \gamma(\theta)) \] (2.15)

\[ = \left( \prod_{i=1}^{n} h(X_i) \right) \exp(\sum_{i=1}^{n} \omega(X_i)) \] (2.16)

\[ \propto \exp(\sum_{i=1}^{n} \omega(X_i)) \] (2.17)

\[ \propto \exp(T(X_1, X_2, \ldots, X_n)) \] (2.18)

where \( T(X_1, X_2, \ldots, X_n) \) is the sufficient statistic [45]. Therefore, for data from an exponential family the test statistic is easier to derive and it usually is a function of the sufficient statistic. For test statistics that lack an analytic distribution, one can rely upon computational methods or use approximations for the distribution of the test statistic.

A commonly used approximation for \(- \log(\Lambda(x)^2)\) is a \( \chi^2 \) distribution with \( k \) degrees of freedom where \( k = \dim(\Theta) - \dim(\Theta_0) \). This result holds true for \( n \to \infty \) and is a consequence of Wilks’s Theorem [168, 167]. Therefore, it is natural to assume the existence of an approximating distribution for any test statistic in the parametric, exponential family with large sample sizes. Since functional data involve large samples, it is clear that an approximate distribution is available for the test statistic as long as the test statistic involves a sufficient statistic.

The Multivariate ANalysis Of VAriance (MANOVA) [11, 84, 91] is a test used to measure differences in multivariate data. It is a very common measure based test used for multivariate analysis and a two-sample version of a MANOVA involves the well known Hotelling \( T^2 \) statistic [83, 82].

The measures involved in a MANOVA are mostly based on covariance matrices that are estimated within a treatment \( S_W \) and between treatments \( S_B \).

\[ S_B = \sum_{i=1}^{t} r_i (\overline{Y}_i - \overline{Y})(\overline{Y}_i - \overline{Y})' \] (2.19)

\[ S_W = \sum_{i=1}^{t} (Y_i - \overline{Y}_i)(Y_i - \overline{Y}_i)' \] (2.20)

where \( Y_i \in \mathbb{R}^{n \times r_i}, \overline{Y}_i, \overline{Y} \in \mathbb{R}^n \). For a \( n \)-dimensional mean response, the estimated covariance matrices are in \( \mathbb{R}^{n \times n} \).
The commonly used measures in a MANOVA are formulated using these matrices as,

- **Wilk’s Lambda**: The statistic expressed in terms of the eigen values \( \lambda_i, i = 1,2, \ldots, n \), corresponding to the eigenvalue decomposition of \( S_W^{-1}S_B \).

\[
\Lambda_W = \prod_{i=1}^{n} \frac{\lambda_i}{1 + \lambda_i} \quad (2.21)
\]
\[
= \frac{|S_W|}{|S_B + S_W|} \quad (2.22)
\]

- **Lawley-Hotelling trace**: The statistic is expressed as a measure formulated using the trace of the matrix \( S_B S_W^{-1} \).

\[
\Lambda_{LH} = \sum_{i=1}^{n} \lambda_i \quad (2.23)
\]
\[
\equiv \text{tr}(S_B S_W^{-1}) \quad (2.24)
\]

- **Pillai’s trace**: The statistic is expressed as a measure formulated using the trace of the matrix \( S_B(S_W + S_B)^{-1} \).

\[
\Lambda_P = \sum_{i=1}^{n} \frac{\lambda_i}{1 + \lambda_i} \quad (2.25)
\]
\[
\equiv \text{tr}(S_B(S_W + S_B)^{-1}) \quad (2.26)
\]

- **Roy’s largest root**: The statistic is expressed as a measure formulated using maximum eigen value \( \lambda_i \) resulting from the eigenvalue decomposition of \( S_W(S_B + S_W)^{-1} \)

\[
\Lambda_R = \max_i \lambda_i \quad (2.27)
\]

These measures are quite robust in MANOVA tests although many of them lack an analytic distribution. They usually have an F-approximation with small sample sizes and a \( \chi^2 \)-approximation with large sample sizes. With large samples, all statistics tend to perform in a similar manner yielding good power [11]. The Wilks Lambda statistic (\( \Lambda_W \)) is the statistic corresponding to a likelihood-ratio test (LRT) defined for a MANOVA.
Although MANOVA is robust in multivariate analysis, it poses a great disadvantage with functional data. As with many processes involving functional responses, the dimension (n) of the profile is significantly greater than the number of replicates \((r_i)\) available per treatment. Mathematically, this would result in singular matrices for \(S_B\) and \(S_W\), preventing one from obtaining the measures discussed in (2.22)-(2.27).

In addition, heteroskedasticity and insufficient number of replicates are known to increase Type-I error in a MANOVA. In fact, the assumptions of MANOVA include homogeneity of covariance matrices, independence of the observations, normality, and the existence of an underlying mean response from each treatment. Although some of these assumptions hold true with functional data in most experiments, dependence of observations is very common and violation of this assumption makes MANOVA a less likely candidate for testing functional responses. This behaviour is demon-
(a) Plot of Type-I error ($\alpha$) against the dimension of the mean response. Three treatments are used with replicate sizes as $r_1 = 12$, $r_2 = 15$, $r_3 = 20$ respectively. 500 repetitive tests are carried out to calculate $\alpha$. A zero mean, multivariate normal vector is assumed as the response in $H_0$

(b) Plot of Type-I error ($\alpha$) against the dimension of the mean response. Three treatments are used with replicate sizes as $r_1 = 6$, $r_2 = 6$, $r_3 = 8$ respectively.

Figure 2.2: Effect of replicate sizes, homogeneity on MANOVA Type-I error. The level of the test is $\alpha = 0.05$ and an F-approximation for $\Lambda_W$ is employed in the test.
strated in Figure 2.2. The results are obtained using the Wilk’s Lambda as the test statistic and an appropriate F-approximation is employed [11]. For instance, in Figure 2.2 (a), an increase in \( \alpha \) is observed as the dimension of the mean vector increases. MANOVA seems to offer good control of Type-I error (See Figure 2.2 (b)) when the replicate sizes are reasonably sufficient to estimate the covariance matrices of size \( n \times n \). However, with relatively smaller sizes of replicates, MANOVA seems to lose control of \( \alpha \) and seems to increase dramatically. A similar pattern (See Figure 2.1) is noted when the homogeneity assumption of covariances is violated. This can be rectified with an increase in the number of replicates that will effectively estimate the covariance matrices. However, in practice, getting several replicates is usually cumbersome. Also, as the length of the response is very large, it becomes difficult to meet this requirement. For example, with a response vector of length 4096, one will require a large number of replicates in order to guarantee a reasonable control of Type-I error. Collection of large replicate sizes in certain experiments are not feasible and cost effective. Therefore, a better testing strategy is required. Hence, the application of transform based tests have become an appropriate tool under such circumstances.

### 2.3 Transform Based Statistical Tests

Based on the previous discussion, it can be understood that the dimension of the response has a strong effect on Type-I error control and power of a test. Therefore, an important aspect of testing functional responses is based on efficient dimension reduction without losing information due to the introduction of sparsity. Transform based testing is developed on the premise of the existence of an orthonormal basis and the representation of the original functional response using such a basis. Since statistical testing is based on obtaining measures of variance (as in MANOVA, ANOVA etc.), most transform based methods are effective when such an orthonormal basis exists. The existence of an orthonormal basis guarantees an isometric, linear map between the function and a set of coefficients as defined by Bessel’s inequality [142, 169].

That is, for any function \( f \in \mathbb{H} \) and an orthonormal basis \( \phi_k \),

\[
\sum_{k \in K} | \langle f, \phi_k \rangle |^2 \leq \| f \|^2
\]  

(2.28)
where $\mathcal{K}$ is an index set corresponding to number of basis functions and $\mathbb{H}$ represents a Hilbert space [165]. In the past, transform based statistical tests involved trigonometric basis as defined in Fourier Transforms. For example, the HANOVA test [67, 66] involves Fourier transform coefficients formulated using a statistic using an adaptive Neyman test [118, 120]. According to Neyman, a transform based test is problematic when tested on individual coefficients and a composite statistic using these coefficients is necessary [66].

Intuitively, one can understand Neyman’s argument by associating it with equation (2.28). Other tests have been developed using a wide variety of transforms and methods involved with dimension reduction. In [78], smoothing splines are employed to obtain a test statistic when the response is a curve (functional response). Some authors have used principal component analysis [25, 43, 102] to obtain a minimal set of coefficients which would result in an efficient test statistic. This work focuses primarily on using wavelet transforms to obtain a transform based test statistic for a non-parametric statistical test. Wavelet transforms have been employed in various fields of science and engineering [90, 88, 89, 32]. Over the past few decades, these transforms have become very popular in data analysis and their definition, properties and extensions to non-parametric data analysis are discussed in the chapters ahead.
CHAPTER 3

WAVELET TRANSFORMS IN DATA ANALYSIS

Wavelet transforms are orthogonal, energy preserving transformations with superior time-frequency localization. The most common applications of wavelets have been in the areas of signal and image processing [104, 53] and compression methods [131, 151]. In this chapter, a general introduction of wavelet transforms and their mathematical framework are discussed. In addition, their properties and significance in data analysis are also discussed. Later, such properties are exploited in offering a testing framework for data involving functional responses.

3.1 Fourier Transform

Fourier analyses have played a significant role in signal processing applications and their omnipresence is seen in several fields of science and engineering.

A Fourier analysis consists of extracting frequency components involved in a signal. In data analysis, Fourier analysis is commonly found in time-series models, non-parametric regression, and smoothing techniques. In a Fourier analysis, the original function \( f(x) \), with support \( x \in (-T,T) \), is decomposed into a series of sines and cosines which in effect would measure the frequency components persistent in the signal.

In other words, the function is expressed in terms of a set of trigonometric, orthogonal basis functions. That is,

\[
f(x) = \frac{a_0}{2} + \sum_{k=1}^{\infty} a_k \cos \left( \frac{k\pi x}{T} \right) + b_k \sin \left( \frac{k\pi x}{T} \right)
\]

(3.1)

where \( a_0, a_k \) and \( b_k \) are the Fourier coefficients.
These coefficients can be computed using the following integrals,

\[ a_0 = \frac{1}{T} \int_{-T}^{T} f(x) dx \] (3.2)
\[ a_n = \frac{1}{T} \int_{-T}^{T} f(x) \cos \left( \frac{k\pi x}{T} \right) dx \] (3.3)
\[ b_n = \frac{1}{T} \int_{-T}^{T} f(x) \sin \left( \frac{k\pi x}{T} \right) dx \] (3.4)

In general, the series expansion of the function in (3.1) can be represented as,

\[ f(x) = \sum_{k=-\infty}^{\infty} c_k e^{ikx} \] (3.5)

Clearly, this generalization extends the support of the function to \((-\infty, \infty)\) and the coefficients \(c_k\) are complex. In practice, a discretized version of the Fourier transform is implemented and the Fast Fourier Transform (FFT) \([44, 163]\) is computationally the most efficient approach to obtain these coefficients, where

\[ c_k = \int_{k=-\infty}^{\infty} f(x) e^{-ikx} dx \] (3.6)

These coefficients result in the Fourier frequency spectrum \(F(k)\) of a functional response \(f(x)\) in the original time domain. From a data analysis perspective, the spectrum \(F(k)\) has the same information as the original function. This is shown by a special case of Bessel’s inequality (see equation (2.28)) known as the Parseval’s identity \([152]\),

\[ \sum_{k=-\infty}^{\infty} |c_k|^2 = \frac{1}{2T} \int_{k=-\infty}^{\infty} |f(x)|^2 dx \] (3.7)

In fact, this result is very close to Plancherel’s theorem \([64]\) defined for \(l_2\)-integrable functions commonly discussed in harmonic analysis and the analysis of differential equations. These properties and methods related to Fourier transforms made statistical testing of functional responses better. Some approaches based on the Fourier transform in testing differences in functional responses can be found in \([66]\) and \([67]\). However, amid all of the advantages of Fourier transforms, they have a few drawbacks that would be less efficient statistical testing. For instance, Fourier analysis loses all information associated with sharp, local differences (such as sudden spikes and small segments)
(a) The mean response consisting of four different frequency components $f_1 = 12, f_2 = 24, f_3 = 48$ and $f_4 = 1$. That is, $f(x) = \sin(2\pi(12)x)I_{x \in (0.01,2.56)} + 0.5 \cos(2\pi(24)x)I_{x \in (2.56,5.12)} - \sin(2\pi(48)x)I_{x \in (5.12,7.68)} + \cos(2\pi x)I_{x \in (7.68,10.24)}$

(b) The Fourier spectrum of the signal identifying the frequency components in the signal

Figure 3.1: A noiseless functional response with several frequency components and its corresponding frequency spectrum obtained using Fourier Transforms.

when transformed to the frequency domain. In most signal processing applications where Fourier analysis is employed frequently, a researcher is primarily interested in observing the notable frequency content in the signal.
This information can be obtained efficiently through a Fourier analysis, but local differences and their importance as recognized by statistical testing may not be noticeable in a Fourier analysis. Consider a signal consisting of several frequency components (see Figure 3.1 (a)). It can be noted that the Fourier spectrum in Figure 3.1 (b) identifies all of the frequency components in the signal. However, the locations where frequencies transition from high to low or vice versa is lost entirely in the spectrum.

In the past, the Short-Term Fourier Transform [5, 6] (also known as the windowed Fourier Transform) has been shown to offer time-frequency localization. Although computationally efficient, the method is dependent on the choice of the window length which makes the method not adaptive to every possible response. This disadvantage was one of the motivations for wavelet transforms in addition to computational efficiency, flexibility and adaptability in achieving time-frequency localization. Another noticeable property of Fourier Transforms is the effect of noise in the spectrum of the signal. It can be seen from Figure 3.2 (a) and (b) that the noise of the response has a significant effect on the frequency spectrum obtained using Fourier transformations. Hence, a statistical test based on the coefficients corresponding to the spectrum should discriminate between noise and signal. A proper distribution of the coefficients or the magnitude of these coefficients is sufficient in most cases to obtain a proper test statistic. In [66] and [67], asymptotic distributions were used in obtaining powerful tests. The real and imaginary coefficients from the Fourier transform of the response would have a normal distribution when the functional response is corrupted with Gaussian noise as shown in Figure 3.3. However, the magnitude of these coefficients are not normally distributed. In most cases, the distribution of the magnitude of these coefficients have a Chi-squared distribution and (see Figure 3.3) these have been analyzed in previous work [66]. Although the statistical properties of these coefficients are remarkable, the time-frequency localization and the inability to preserve information pertaining to sharp changes make wavelet transforms a better option than Fourier Transforms.

Wavelet transforms also offer sparsity, computational efficiency and decorrelation which make them relatively superior to Fourier transformation. In data analysis, wavelet transforms are employed in profile monitoring applications, change-point analysis and noise and dimension reduction with applications in engineering and many areas of science.
(a) The mean response consisting of four different frequency components \( f_1 = 12, f_2 = 24, f_3 = 48 \) and \( f_4 = 1 \). That is, 
\[
  f(x) = \sin(2\pi(12)x)\mathbb{1}_{x \in (0.01, 2.56)} + 0.5 \cos(2\pi(24)x)\mathbb{1}_{x \in (2.56, 5.12)} - \sin(2\pi(48)x)\mathbb{1}_{x \in (5.12, 7.68)} + \cos(2\pi x)\mathbb{1}_{x \in (7.68, 10.24)}
\]

(b) The Fourier spectrum of the signal identifying the frequency components in the signal

Figure 3.2: A noisy functional response with several frequency components and its corresponding frequency spectrum obtained using Fourier Transforms corrupted by Gaussian noise.

### 3.2 Wavelet Transforms

As mentioned previously, wavelet transforms were developed due to better time-frequency localization than Fourier transforms. Although different from Fourier transforms, wavelet transforms share a mathematical framework akin to Fourier transforms with minor differences. For example, a trigonometric, orthogonal basis consisting of sines and cosines is used in Fourier transforms whereas wavelet transforms involve a unique, orthogonal basis satisfying some interesting properties [122].
3.2.1 Wavelets

A wavelet $\psi(t)$ is a function that has a narrow, compact support $\Lambda \subset \mathbb{R}$ with some desirable requirements.

- $\int_{-\infty}^{\infty} \psi(t) dt = 0$, commonly referred to as the “admissibility condition”.

- $\int_{-\infty}^{\infty} \psi^2(t) dt = 1$, the “unit energy” condition.

- Therefore, wavelets form a complete orthonormal system (CONS).

The Haar wavelet [27] is the simplest wavelet basis and one of the earliest works related to present day wavelet transforms. In fact, Haar wavelets are closely related to Littlewood-Paley projections [72] in functional analysis. It can be seen in Figure 3.4, that the Haar wavelet is a function with support on a one dimensional ball with support $(0, 1)$. A wide variety of wavelet families are available and each wavelet family offers a unique set of basis functions. The choice on the type
wavelet basis is usually dependent on the problem [122]. One of the most common wavelet functions is the Daubechies filter names after Ingrid Daubechies [51].

### 3.2.2 Continuous Wavelet Transform (CWT)

The continuous wavelet transform of a function \( X(t) \) is defined as,

\[
\xi_{\tau,t} = \frac{1}{\sqrt{\tau}} \int X(s) \psi \left( \frac{s - t}{\tau} \right) ds
\]

(3.8)

where, \( \xi_{\tau,t} \) corresponds to the scale parameter ‘\( \tau \)’, with location parameter ‘\( t \)’. It is important to mention that ‘\( \tau \)’ is a dyadic (power of 2) scale. The CWT obtains a wavelet coefficient for every location of the signal at any scale. Clearly, the transform may offer insight into the behavior of the signal at various resolutions (scales) even though many of these coefficients reveal redundant information in certain applications. Also, evaluating such a transform may require an increased computational cost. Hence, a discrete version of the CWT, known as the discrete wavelet transform (DWT) is adapted in most cases.
3.2.3 Discrete Wavelet Transform (DWT)

The redundancy in a CWT is eliminated by the discrete wavelet transform which chooses the scales in powers of two. Define, \( \tau_k = 2^{k-1} \), the dyadic scales at which the signal is observed. In the ensuing discussion, these scales correspond to the resolution levels of the discrete wavelet transform (DWT). A change in ‘k’ will help the analysis of the function at various time-frequency resolutions. As the scale parameter ‘k’ increases, the function is viewed at short time intervals (high frequencies in the frequency domain) and for smaller values of ‘k’, the function is viewed at larger time intervals (equivalently, low frequency responses).

An appropriate scale may provide the benefit of observing the behaviour of a signal at a global level (smaller scale) or at a local level (larger scale). Hence, a wavelet analysis helps data analysis by “zooming in” and “zooming out” of the function space thereby picking fine and coarse details of the function simultaneously. The DWT of a signal is defined in terms of a discrete wavelet function, commonly known as the “scaling filter” or “wavelet filter”.

3.2.4 Formal Definition of DWT

A discrete wavelet transform (DWT) transforms a function \( f(x) \) such that,

\[
f(x) = \sum_{k=1}^{M_0} V_{0k} \phi_{0k}(x) + \sum_{j=0}^{\infty} \sum_{k=1}^{M_j} W_{jk} \psi_{jk}(x)
\]

That is, the function is represented as a sum of orthogonal wavelet functions. It can be noticed that the decomposition has a construct similar to a Fourier decomposition in equation (3.1). The difference in equation (3.9) is due to the wavelet functions in place of the trigonometric basis. Despite a minor, subtle change in the definition, the DWT offers better time-frequency localization as a primary consequence of the wavelet basis.
The coefficients are obtained using a familiar approach, akin to (3.4). Here,

\[
\psi_{jk} = \int_{-\infty}^{\infty} f(x)2^{j/2}\psi \left(2^j x - k\right) \, dx \tag{3.10}
\]

\[
\phi_{jk} = \int_{-\infty}^{\infty} f(x)2^{j/2}\phi \left(2^j x - k\right) \, dx \tag{3.11}
\]

Since the realization of a functional response in data analysis is discrete, a handy representation of the DWT is derived from signal processing applications. A functional response in data analysis and statistics is a continuous signal sampled at discrete time indices.

For a signal of length ‘N’, that is, \( N \triangleq 2^J \). The original signal can be decomposed into a set of coefficients of length \( \frac{N}{2^p}, p = 1, 2 \ldots P = \log_2(N) - 1 \equiv J - 1 \) where ‘P’ is the highest possible level of DWT.

Intuitively, it can be understood that the level ‘P’ corresponds to the original signal without any transformation. Assume that the topmost decomposition level as level \( J - 1 \) and a top-down approach is carried out to reach level \( J_0 \). Define the scale at level ‘p’ as \( \tau_p \).

The DWT at level ‘p’ will be given as

\[
W'_{pk} = \sum_{l=0}^{N-1} X(l)\delta'_{(l-1)|N} \tag{3.12}
\]

\[
W_{pk} = W'_{pk} \downarrow_2^p \quad k = 0, 1, \ldots \frac{N}{2^p} - 1 \tag{3.13}
\]

In the above equations, \( \downarrow^p \) denotes ‘p’ down-sampling operations and | indicates a modulo operator. \( W_{p,k} \) are the \( p^{th} \) level wavelet coefficients of interest. \( \delta' \) is the periodized version of the wavelet filter, \( \delta_i \). Here, \( W_{pk} \) are coefficients of the discrete wavelet transform. \( \delta_i, i = 1, 2, \ldots L \) denote a discrete wavelet filter of length ‘L’. The filter coefficients are commonly known as “taps” in signal processing applications [39]. The length, L, is chosen to be less than N that is periodized to a length N. This filter is known as the “wavelet” filter.

**Properties of \( \delta_i \).** The first two properties are the previously attributed “unit energy” and “admissibility” conditions of the DWT. The last property ensures orthogonality for every even shift.
The first and the last property collectively form the “orthonormal” property of the wavelet filter. The above filtering operation resulted in \( \frac{N}{2} \) coefficients, the upper half of the wavelet transform. In order to obtain the lower half of the wavelet transform at any given level a different filter has to be employed. This filter is known as “scaling filter”. Let us denote the scale associated with the \( p^{th} \) level scaling filter as \( \lambda_p \). During the inception of wavelet theory, these filters were called the “mother” and “father” wavelet functions. However, with the varied range of signal and image processing applications to which wavelets are applicable, the perception of the wavelet functions as “filters” have become quite common.

**Averaging filter.** The wavelet function can generally be defined in terms of the scaling function [122]. As their names imply, a wavelet filter extracts differences at various scales whereas an averaging filter obtains averages at various scales. In the frequency domain, the wavelet filter has a nominal pass band \( \in (\frac{\omega}{4}, \frac{\omega}{2}) \), corresponding to a “high-pass” filter response, whereas the scaling filter has a nominal pass band \( \in (0, \frac{\omega}{4}) \) corresponding to a “low-pass” filter response.

It can be understood that the “low-pass” filter mirrors the response of a “high-pass” filter below \( \frac{\omega}{4} \) and so the scaling filter is sometimes referred to as a “quadrature mirror filter” corresponding to the wavelet filter. Both filters (as a set) usually form a filter bank at level ‘p’ and generally speaking, the DWT can be perceived as a filter bank with several stages of a filtering.

Formally, the scaling filter can be defined as follows,

\[
\xi_l = (-1)^{l+1} \delta_{L-(t+1)}
\]  

(3.14)

**Properties of \( \xi_l \).** The scaling filter satisfies the following properties,

\[
\sum_{t=0}^{L-1} \xi_t = 0
\]  

\[
\sum_{t=0}^{L-1} \xi_t^2 = 1
\]  

\[
\sum_{t=0}^{L-1} \xi_{t+2k} = 0, k = 0, 1 \ldots n
\]
The lower half of the DWT will be comprised of scaling coefficients of the signal obtained using the scaling filter.

The scaling coefficients at level ‘p’ will be given as

\[ V_{pk}' = \sum_{l=0}^{N-1} X(l)\xi'_{l(l-1)|N} \]   \hspace{1cm} (3.15)

\[ V_{pk} = V_{pk}'|_{p}^{2} \quad k = 0, 1, \ldots N \frac{2^{p}}{2^{p}} - 1 \] \hspace{1cm} (3.16)

\( \xi' \) is the periodized version of the scaling filter. \( W_{pk} \) and \( V_{pk} \) are discrete coefficients [50] offered by a DWT and are equivalent to \( \psi_{pk} \) and \( \phi_{pk} \) given by equation (3.10) and (3.11).

### 3.2.5 Wavelet Filters and Their Frequency Gain

The frequency response is obtained by plotting the squared gain of the frequency spectrum of the filter obtained using a Fourier Transform. That is,

\[ \varphi(\omega) = \int_{-\infty}^{\infty} f(t)e^{-\omega t} dt \] \hspace{1cm} (3.17)

is the continuous Fourier Transform of the function f(t). The equivalent discrete version is given by,

\[ \varphi(\omega_k) = \sum_{n=-\infty}^{\infty} f(n)e^{-\omega n} \] \hspace{1cm} (3.18)

where, f(n) is the discretized version of the function f(t).

In the above setup, \( \omega \) is the angular frequency and it can be represented by a more common notation \( 2\pi f \).

In the frequency domain, the response is given by a complex valued spectral function,

\[ \Omega(\omega) \equiv \Omega(f) = |G(f)|e^{i\phi(f)} \] \hspace{1cm} (3.19)

where \( |G(f)| \) is the gain function and \( \phi(f) \) is the phase-function.
The squared gain function measures the proportional strength of the filtered signal in the frequency domain and the squared gain is given by,

\[ \Omega^2(f) = |G(f)|^2 \]  

(3.20)

In Figure 3.5, this squared gain is plotted against various frequencies. An increased gain at a frequency, say \( f_0 \), would mean that any portion of the signal with a frequency content equal to \( f_0 \) will not be suppressed by the filter with a response function whose squared gain function is \( \Omega^2(f) \). The frequency response is defined by \( \varphi_{\omega_k} \). The choice of filter effects the nature of the filtering (see Figure 3.5). For example, the Haar wavelet filter forms a 1\textsuperscript{st} difference filter whereas the Daubechies-4 (denoted commonly as d4 or d(4)) is a 2\textsuperscript{nd} difference filter. Likewise, the frequency response of a Haar filter is slightly ‘steep’ whereas the d4 filter is rather gradual. Hence, it is advisable to determine a filter that is best applicable to a chosen application in order to reap all the benefits of the Discrete Wavelet Transform.

### 3.3 Matrix Notation of DWT

Matrices define the fundamental building blocks of any computational method involved in data analysis, especially in higher dimensional problems. Hence, it is good practice to implement or express a mathematical problem in terms of matrices whenever possible. Wavelet transforms are computationally efficient, requiring \( O(\log_2(n)) \) operations, and most common wavelet packages implement them with matrix operations. In this section, the discrete wavelet transform is expressed using matrices so that the implementation of DWT for real-time applications become straightforward.

Consider a signal \( X(t) \), discretized to ‘N’ sample points. The discretized (sampled) signal as a vector would be \( X \in \mathbb{R}^n \). The \( j \)\textsuperscript{th} level DWT is constructed by filtering \( X \) with a set of wavelet filter coefficients via circular convolution.

The \( j \)\textsuperscript{th} level wavelet filter coefficients (periodized to length ‘N’) are \([\delta_0, \delta_1, \ldots, \delta_{N-1}] \in \mathbb{R}^n \). This set of filter coefficients are circularly shifted \( N/2^j \) times, resulting in a wavelet filter coefficient matrix.
Figure 3.5: Commonly used Wavelet Filters and their frequency response
comprising \(N/2^{j}\) rows, say, \(\Delta_j \in \mathbb{R}^{(N/2^j) \times N}\). If the maximum level of decomposition is \(J_0 = \log_2(N)\), then \(J_0\) such matrices can be obtained. \(\Delta_j\) would correspond to the upper half of the DWT matrix corresponding to the wavelet filter. In a similar manner, we can obtain another matrix \(\Xi_j\) consisting of scaling filter coefficients (periodized to length ‘N’) given by \([\xi_0, \xi_1, \ldots, \xi_{N-1}] \in \mathbb{R}^n\).

Thus, the entire DWT matrix is given by,

\[
\Theta = \begin{bmatrix} \Delta_j \\ \Xi_j \end{bmatrix} \in \mathbb{R}^{N \times N} \tag{3.21}
\]

The \(j^{th}\) level wavelet and scaling coefficients can be obtained by performing a matrix multiplication given by,

\[
C_j = \Theta X', C_j \in \mathbb{R}^{N/2^{(j-1)}} \tag{3.22}
\]

\(C_j\) is a vector of coefficients (consisting of both wavelet and scaling coefficients) at level \(j\). That is,

\[
C_j = \begin{bmatrix} W_j \\ V_j \end{bmatrix} \in \mathbb{R}^{N/2^{(j-1)}} \tag{3.23}
\]

where \(W_j \in \mathbb{R}^{N/2^j}\) and \(V_j \in \mathbb{R}^{N/2^j}\). At this point, it is imperative to address that the DWT matrix is an orthonormal matrix. As mentioned previously, orthogonality and orthonormality are highly desired properties of a wavelet transform. These requirements will be satisfied by defining the filter coefficients in an appropriate manner so that the DWT matrix remains orthonormal. The choice of wavelet coefficients are based on these conditions and a detailed discussion on this can be found in [153].

### 3.4 Inverse Wavelet Transform

The goal of any mathematical transform is to express the signal in a different space (domain). In most cases, the transformation is well defined so that an inverse transform exists. By definition, the wavelet transform can be inverted to obtain the original signal without any loss of information.
The topic of inverse wavelet transform usually leads into another methodology unique to wavelet transforms known as “multi-resolution analysis (MRA)”. As the name implies, an MRA aids the analysis of a signal at varying resolutions. That is, the same signal can be observed at varying resolutions. This concept was expressed in quite an informal way. A precise mathematical definition of an MRA is offered in the Appendix. This digression to MRA is significant to inverting a wavelet transform because the end result of an MRA with some mathematical additions will result in the original signal.

From the previous section, the coefficients at level ‘j’ were denoted as $C_j$. The level ‘j’ decomposition of an MRA is obtained as follows,

$$X_j = \Theta C_j \quad (3.24)$$

$$= \Psi_j C_j + \Xi_j C_j = D_j + S_j \quad (3.25)$$

Here, $D_j$ is known as the $j^{th}$ level detail and $S_j$ is known as the $j^{th}$ level smooth.

From the definition of an MRA, it can be understood that the $j^{th}$ level smooth is unnecessary in a reconstruction algorithm. This is due to the fact that the $j^{th}$ level smooth can be obtained $(j - 1)^{st}$ level detail and smooth functions.

Thus, if a signal $X$ is decomposed into $J = \log_2(N)$ levels, then it can be shown that

$$X = \sum_{j=1}^{J_0} D_j + S_j \quad (3.26)$$

results in the original signal. A detailed discussion on multi-resolution analysis can be found in [103] and a less intense yet mathematically rigorous analysis is available in [122].
3.5 Wavelet Thresholding

After a discussion on wavelet transforms followed by MRA and reconstruction, one may begin to wonder about the exact purpose of such a transform. Intuitively, we can be convinced of the benefits of analyzing a signal at various resolutions by observing differences and averages at varying scales. However, the true potential of wavelet transforms is revealed in dimension reduction and non-parametric kernel density estimation applications. The dimension reducing capability of wavelet transforms is derived using thresholding methods. If a signal $X(t)$ consists of two components $S(t)$ and $\epsilon(t)$, where $S(t)$ is the true signal part of $X(t)$ and $\epsilon(t)$ is the additive noise component.

A wavelet transform of $X(t)$, using $X \in \mathbb{R}^N$ will yield a set of coefficients at several levels. The overall dimension of the coefficient vector over all levels is still ‘$N$’. That is,

$$C_j \in \mathbb{R}^{N/2^j} \Rightarrow [C_1, C_2, \ldots, C_J] \equiv C \in \mathbb{R}^N \quad (3.27)$$

Clearly, there is no obvious dimension reduction, at least not yet. Since the signal is comprised of both a true signal component and a noise component, the resulting wavelet coefficients will contain information corresponding to the true signal as well as the noise component. However, the magnitude of the signal component is significantly larger than the noise component. Thus, one can eliminate the spurious noise coefficients by setting a threshold, say $\lambda$. Any coefficient greater than $\lambda$ will be retained and the others discarded. Since thresholding is a coefficient selection scheme resulting in significant dimension reduction of coefficients, it is sometimes referred to as a “wavelet shrinkage” method when the coefficients are mapped to $\mathbb{R}$ instead of $\mathbb{Z}_{0,1}$, the binary field.

Thus, $X \in \mathbb{R}^N$ can now be expressed using $C \in \mathbb{R}^M$, where $M \ll N$ coefficients. The dimension reduced coefficients will result in a de-noised version of the signal upon reconstruction. The choice of threshold is not unique and depends on the application. The determination of a threshold has been addressed quite rigorously and still remains to be a topic of interest among researchers. It is true, however, that the actual set of wavelet coefficients of an $l_2$-integrable functional response is “sparse”. That is, a function with no noise will always result in a sparse set of coefficients. Under this condition, it is reasonable to impose a threshold that can ignore some of the spurious
empirical coefficients which may be due to the noise component rather than the underlying functional response. In Figure 3.6, it can be seen that the number of non-zero coefficients required for perfect reconstruction (that is, representation of the response) decreases rapidly as the length of the function increases. For smaller sample sizes, the proportion of non-zero coefficients is 18.75% and rapidly approaching zero. For the largest sample size of $2^{15}$, the proportion of non-zero coefficients is $\approx 0.0793\%$. This would correspond to roughly 26 coefficients instead of the original sample of size $2^{15}$. In Figure 3.7, by discarding extremely small coefficients, the inverse DWT is able to offer a reasonable piecewise approximation to the original response with a mean squared error of $\approx 0.00021$.

Thus, with an appropriate method to discriminate a noise-related coefficient from a response-related coefficient, a reasonable approximation can be obtained. Also, this sparsity property makes DWT attractive in very high dimensional problems. Some of the thresholding schemes are,

- Stein’s Unbiased Risk Estimate (SURE) threshold, a level dependent thresholding method.
Figure 3.7: The underlying “noiseless” function (solid gray line) is $\sin(2x)$ and the “haar” wavelet filter is employed. The piecewise approximation of the function using the reduced set of wavelet coefficients with an IDWT is drawn using a solid black line.

- Universal threshold, $\lambda = \sigma \sqrt{2 \log(N)}$.

- BayesShrink method, that uses Bayesian inference in determining a threshold.

Since the choice of threshold determines the effectiveness of a DWT, several methods are available [61, 62, 60, 59] and threshold selection continues to be a favorable topic of research. Such schemes have incorporated a variety of statistical methodologies ranging from Bayesian methods [38], non-linear Bayesian analysis [160], cross-validation [117], Breiman’s non-negative garrote [75], false discovery rate (FDR) based methods [1, 2] and others [28, 35, 40, 89]. Regardless of the choice of thresholding scheme, the manner in which a threshold is applied is limited to two types,

- Hard thresholding. A coefficient $\theta$ is retained using the following condition,

$$
\theta = \begin{cases} 
\theta, & |\theta| > \lambda \\
0, & \text{o.w}
\end{cases}
$$
Figure 3.8: Density of a noisy response, \( f(x) = \sin(2x) \) (solid gray line) and the density of the wavelet coefficients (dotted black line). The wavelet coefficients are obtained using a “haar” wavelet with \( J = 15 \). The sample size is \( N = 2^{15} \).

- Soft thresholding [58]. This scheme yields better signal isolation. The condition to retain a coefficient is given by,

\[
\theta = \begin{cases} 
\theta - \lambda, & \theta > \lambda \\
0, & |\theta| \leq \lambda \\
\theta + \lambda, & \theta < -\lambda 
\end{cases}
\]

Most of the analyses in this work use a hard thresholding approach which is carried out with \( \lambda \) (the universal threshold) obtained using the simplest thresholding scheme: VisuShrink [62].

In addition, a DWT offers “whitening” or de-correlation [17, 46] which can aid in the analysis of non-stationary processes. Under the assumptions of normality of a functional response, the wavelet coefficients are normally distributed (shown in Figure 3.8). Since the distribution is preserved after transformation, statistical analyses can be carried out easily on the wavelet coefficients.
In this chapter, the Fourier transform is identified as a viable tool for statistical analysis. Although very effective, it had a few disadvantages that are overcome by wavelet transforms. The key aspects of discrete wavelet transforms that make them handy tools in statistical analyses are,

- Sparsity, a very nice feature extensible to high dimensional problems.
- De-correlation, a property that helps in the analysis of non-stationary and heavily correlated processes.
- Time-frequency localization, an effective way to preserve localized differences in statistical testing.
- Computational complexity. Although Fourier transforms are efficiently employed via a Fast Fourier Transform (FFT) offering a computational cost of $O(N \log(N))$, a DWT has a computational cost of $O(N)$ via the pyramid algorithm [51].

With these properties and simplicity, an efficient non-parametric test can be employed for testing in high dimensional problems with functional responses. This will be discussed in the chapter ahead.
CHAPTER 4

A NON-PARAMETRIC WAVELET BASED TEST

Wavelet transforms have been recognized as an effective tool in statistical data analysis. The properties of discrete wavelet transforms discussed in Chapter 3 form the basis of their effectiveness. Several authors have employed wavelet transforms in the past for analyses in problems from a variety of fields. Some of the best applications of wavelets are in image and signal processing applications [131, 151, 164]. In statistics, wavelet transforms are used in density estimation [36], non-parametric regression [116, 26, 29, 92], experimental design [14], quality control and profile monitoring [37, 32], variable selection [10, 33] and time series analysis [70, 125, 126] to name a few. In this chapter, a non-parametric statistical test is proposed for testing differences among functional responses from multiple treatments.

4.1 Traditional Multivariate Test

The proposed non-parametric test for functional data derives some of its properties from multivariate analysis. The multivariate test is modified with the introduction of wavelet transforms that aims to improve power and achieve good control of Type-I error. An initial approach on this area can be found in [170]. A Hotelling $T^2$ test is a commonly used multivariate test analogous to a $t$-test in a univariate test.

The Hotelling $T^2$ statistic proposed by Harold Hotelling is a multivariate generalization of a $T^2$ statistic. The power of this statistic in detecting local and global differences defines a pivotal role of this thesis.

**Corollary 4.1.1** A random variable, $X$, has a Hotelling $T^2$ distribution, defined as, $T^2_{p,m}$, if and only if,

$$\frac{(m - p + 1)X}{mp} \sim F_{m-p+1}$$

(4.1)
The distribution in Corollary 4.1.1 can now be extended to a higher dimensional random variable as follows. Suppose X has a multivariate Normal distribution, that is,

\[ X \sim N(\mu, \Sigma) \]  

(4.2)

where \(X, \mu, \Sigma \in \mathbb{R}^p\). It can be shown that the following statistic based on the random variable X will have a Hotelling \(T^2\) distribution defined in Corollary 4.1.1.

\[ S = n (\bar{X} - \mu)' W^{-1} (\bar{X} - \mu) \sim T^2_{p,n-1} \]  

(4.3)

In equation 4.3, \(W\) corresponds to the sample variance defined as,

\[ W = (X - \bar{X})' (X - \bar{X}) / (n - 1) \]  

(4.4)

### 4.1.1 \(T^2\) Test Statistic

As mentioned previously, the Hotelling statistic can be used in detecting differences between profiles. Consider multivariate responses obtained from two different treatments, namely \(Y_1\) and \(Y_2\). Assume that the responses are defined in a manner similar to 1.1 and each treatment has \(r_1\) and \(r_2\) replicates. Here, \(Y_1, Y_2 \in \mathbb{R}^n\) and the variances are identical among all treatments. That is,

\[ Y_{1j_1} \sim N(\mu_1, \Sigma) \]  

(4.5)

\[ Y_{2j_2} \sim N(\mu_2, \Sigma) \]  

(4.6)

\[ \Sigma = \sigma^2 I, I \in \mathbb{R}^{n \times n} \]  

(4.7)

where ‘I’ is the identity matrix and \(j_k = 1, 2, \ldots r_k; k = 1, 2\). The mean response of each treatment over all replicates is given by \(\bar{Y}_1\) and \(\bar{Y}_2\) and it can be shown under the null hypothesis, \(H_0\), \(\bar{Y}_1 - \bar{Y}_2 \sim N\left(0, \sigma^2 \left(\frac{1}{r_1} + \frac{1}{r_2}\right)^{-1} I\right)\).

The mean squared error (MSE) is a well known estimate for the variance \((\sigma^2)\) and for the n-
dimensional case the MSE can be given as,

$$W = \frac{2}{n(r_1 + r_2 - 2)} \sum_{i=1}^{2}(Y_i - \bar{Y}_i)'(Y_i - \bar{Y}_i)$$

(4.8)

The test statistic can now be formulated, under $H_0$, as

$$T^2 = \frac{(\bar{Y}_1 - \bar{Y}_2)'W^{-1}(\bar{Y}_1 - \bar{Y}_2)}{n\left(\frac{1}{r_1} + \frac{1}{r_2}\right)}$$

(4.9)

The distribution of this statistic, under $H_0$, is $F(n,n(r_1+r_2-2))$ (See Section A.1.2). This can be extended to a design with $t$ treatments, each with a functional response of length $n$. This design is unbalanced and the $n$-dimensional multivariate responses over $t$ treatment variables are defined as,

$$Y_{ijk} = \mu + \alpha_{ijk} + \epsilon_{ijk}$$

(4.10)

$$= \mu_{ijk} + \epsilon_{ijk}$$

(4.11)

with,

$$i = 1,2,\ldots,t$$

$$j = 1,2,\ldots,r_i$$

$$k = 1,2,\ldots,n$$

$$\epsilon_{ij} \sim N(0,I)$$

Lemma 4.1.2 For the design in equation (4.11), the statistic ($\vartheta$) to test differences in a functional response is,

$$\vartheta = \frac{(\varphi - t) \sum_{k=1}^{n}(\bar{Y}_{i,k} - \bar{Y}_{..})^T(\bar{Y}_{i,k} - \bar{Y}_{..})}{(t - 1) \sum_{i=1}^{t} \sum_{j=1}^{r_i}(Y_{ij} - \bar{Y}_{i.})^T(Y_{ij} - \bar{Y}_{i.})}$$

(4.12)

Using the definition of an F-distribution, $\vartheta$ has an F-distribution with numerator and denominator
degrees of freedom as \( n(t - 1) \) and \( n(\varphi - t) \) respectively. Here, \( \varphi = \sum_{i=1}^{t} r_i \).

The proof of Lemma 4.1.2 can be found in section A.1.3.

Tests which have used DWT in the past have done so with a focus on the coefficients [159]. Such tests, because of their similarity to an ANalysis Of VAriance (ANOVA), are called Wavelet ANOVA (WANOVA) due to the incorporation of DWT. Some authors have attributed such testing and analysis without the use of DWT to a functional ANOVA (FANOVA) [3], which makes the name WANOVA [138, 141, 76, 77] quite apparent. Clearly, \( \vartheta \) is a measure-based test statistic and it has similarities to a Hotelling [82, 83] statistic and other measure based statistics (see Section 2.2.4) defined for multivariate data. The proposed non-parametric aims to modify this statistic with the aid of wavelet transformations.

### 4.1.2 Wavelet Based Multivariate Statistic

In this section, the Hotelling \( T^2 \) statistic is modified to offer better power and Type-I control using wavelet transforms. Since wavelet transforms are orthogonal, energy preserving transforms [162], it can be shown, with Bessel's Inequality and Pareval's Identity (see Chapter 3),

\[
\| Y_{ijk} - Y_{...} \|_2^2 \geq W[\| Y_{ijk} - Y_{...} \|_2^2] \tag{4.13}
\]

where \( W \) is an operator denoting a discrete wavelet transform (DWT).

Thus,

\[
\vartheta \leq W[\vartheta] \tag{4.14}
\]

**Lemma 4.1.3** The wavelet based statistic, \( \kappa_{\vartheta} \), under the assumptions of normality, is

\[
\kappa_{\vartheta} = \sum_{i=1}^{t} \sum_{k=1}^{T_i} \tilde{\vartheta}_{ik}^2 \tag{4.15}
\]
where $\tilde{\theta}_{ik}$ is the $k^{th}$ wavelet coefficient (after shrinkage) for the $i^{th}$ treatment obtained using

$$W \left[ \gamma^{-1}(\bar{Y}_i - \bar{Y}) \right]' W \left[ \gamma^{-1}(\bar{Y}_i - \bar{Y}) \right]$$

(4.16)

Lemma 4.1.3 is proven in section A.1.4 and $\gamma$ is also defined in A.1.4. Thus, with an appropriate choice of a threshold, a wavelet filter and decomposition level, $\kappa_\eta$ can be obtained.

4.1.3 Choice of Noise Estimator, $\hat{\sigma}$

The estimator for noise, $\hat{\sigma}$, can be obtained in many ways. Also, this estimator can be evaluated per treatment or as an overall pooled estimator. Since the standard deviation is an unbiased estimator, it can be used directly. The median absolute deviation (MAD) estimator is quite robust in wavelet applications [122] and it can also be used. The estimate $\hat{\sigma}$ is,

$$\hat{\sigma} = M[|W_{(J-1)k} - M|W_{(J-1)k}|]$$

(4.17)

where $M$ is the ‘median’ operator and only the coefficients of the top-most level are considered as they capture most of the noisy components of the signal. In the proposed test, the noise estimator is evaluated per treatment and either the standard deviation or MAD is used as the estimator based on the application.

Corollary 4.1.4 Suppose the $i^{th}$ wavelet coefficient is given by,

$$\theta_i \sim N(\mu, \tau^2)$$

(4.18)

where $i = 1, 2, \ldots m$, $\mu$ is the mean of a wavelet coefficient and $\tau^2$ is the variance of the wavelet coefficient. The sum of squared wavelet coefficients after a threshold of $\lambda$ will have a truncated $\chi^2$ tail distribution with $m$ degrees of freedom $\chi_w^2 [\lambda]$.

The truncated chi-squared distribution in Corollary (4.24) is different from the conventional truncated distributions. The commonly used truncated normal random variable $X$ has support $X \in (\lambda_1, \lambda_2)$, $-\infty < \lambda_1 < \lambda_2 < \infty$ and the random variable $Y = X^2$ will have a truncated chi-squared distribution with support $(0, \lambda^2]$ where $\lambda = \max(\lambda_1, \lambda_2)$. 

45
(a) A truncated normal distribution with $\lambda_1 = -0.3$ and $\lambda_2 = 0.3$ where $\mu = 0$ and $\sigma = 1$

(b) A truncated $\chi^2$ distribution with $\lambda = 0.3$ and one degree of freedom

Figure 4.1: A truncated normal distribution and a $\chi^2$ distribution using $n = 2^{15}$ with a hard threshold. The histogram of the observations and a smooth density estimator are shown.
Within the context of wavelet shrinkage, a truncated normal random variable (see Figure 4.1 (a)) has support \((-\infty, -\lambda] \cup [\lambda, \infty)\), where \(\lambda\) is the threshold used with a hard thresholding scheme (see Chapter 3). Thus, the support of the corresponding truncated Chi-squared distribution (see Figure 4.1 (a)) is \(\chi^2_1[\lambda] \) is \([\lambda^2, \infty)\).

From Figures 4.1 (a) and 4.1 (b), it can be seen that a point mass exists at ‘0’ as a result of the threshold. However, this point mass will be taken care of, or more precisely, removed, due to the definition of \(\kappa_\eta\), the wavelet transform based test statistic. \(\chi^2_1[\lambda]\) plays a crucial role in \(\kappa_\eta\). In fact, this distribution distinguishes a wavelet based statistical analysis from conventional statistical approaches. There are several important properties of \(\chi^2_1[\lambda]\) and these are vital in obtaining an exact distribution of the test statistic. The following summarizes some key properties of \(\chi^2_1[\lambda]\).

**Lemma 4.1.5** Suppose \(X \sim \chi^2_1[\lambda], x \in [\lambda^2, \infty)\), where \(\lambda\) is the value of the threshold (the truncation parameter) under hard thresholding scheme, then,

\[
\begin{align*}
\text{i. The density of } X & \text{ is,} \\
 f_X(x) & = \frac{x^{-1/2}e^{-x/2}}{2\sqrt{2\pi}(1 - \Phi(\lambda))} \\
& \quad (4.19)
\end{align*}
\]

\[
\text{ii. The moment generating function (MGF) of } X \text{ is,}
\]

\[
M_X(t) = \frac{1}{\sqrt{1 - 2t}} \quad t \leq 0 \quad (4.20)
\]

It can be seen that the truncated chi-squared distribution is somewhat similar to a regular Chi-squared. This fact should not be surprising given Figures 4.1 (a) and (b). Lemma 4.1.5 is proved in section A.1.5. Using this result, an analytic expression for the density of \(X \sim \chi^2_1[\lambda], x \in [k\lambda^2, \infty)\) can be obtained. Although the proposed test statistic \(\kappa_\eta\) involves \(\chi^2_1[\lambda]\), it is useful to determine the density of \(\chi^2_1[\lambda]\) because of its significance in obtaining approximations to \(\kappa_\eta\). From equation (4.22), it can be noted that several wavelet coefficients after a threshold form \(\kappa_\eta\). It can be inferred under the assumption of independence and normality that the distribution of a truncated Chi-
squared distribution with ‘M’ degrees of freedom is \( X \sim \chi^2_M[\lambda], x \in [M\lambda^2, \infty) \) where \( X = \sum_{i=1}^{M} X_i \) and \( X_i \sim \chi^2[\lambda], x \in [\lambda^2, \infty) \).

**Lemma 4.1.6** Suppose \( X \sim \chi^2_M[\lambda], x \in [M\lambda^2, \infty) \), where \( \lambda \) is the value of the threshold (the truncation parameter) under hard thresholding scheme. Then,

i. The density of \( X \) is,

\[
f_X(x) = \frac{x^{-M/2}e^{-x/2}}{2^{M/2}\Gamma(M/2)\left(1 - \Phi_G\left(\frac{M\lambda^2}{2}\right)\right)}
\]  

(4.21)

ii. The moment generating function (MGF) of \( X \) is,

\[
M_X(t) = \frac{1}{\sqrt{(1-2t)^M}}, \quad t \leq 0
\]  

(4.22)

where \( M \geq 1 \) and \( \Phi_G(\bullet) \) is the cumulative distribution function of a \( \text{Gamma}(M/2,1) \) random variable.

This proof of Lemma (4.1.7) can be easily derived using Lemma (4.1.5). A formal derivation is demonstrated in A.1.6.

As mentioned earlier in this chapter, Lemmas (4.1.5) and (4.1.7) form the mathematical basis for the distribution of \( \kappa_\eta \). The distribution and its closed form expression are obtained in the following section.

### 4.1.4 Distribution of \( \kappa_\eta \)

The proposed wavelet based test statistic, \( \kappa_\eta \) is dependent on a few quantities,

- Dimension of the functional response, \((n)\)
- Maximum level of wavelet decomposition, \((l_j)\)
- Maximum level of thresholding, \((l_t)\)
- The amount of threshold, \(\lambda\).
By definition of a DWT, \( l_t \leq l_j \). Thus, \( \kappa_\eta \) is a shorter, convenient notation for \( \kappa\{n,l_j,l_t,\lambda\} \equiv \kappa\{\eta\} \) where \( \eta = \{n,l_j,l_t,\lambda\} \). From Lemma (4.1.3), \( \kappa_\eta \) is the sum of wavelet coefficients after a threshold. However, \( l_t \) determines the number of coefficients that are actually thresholded. For an \( n \)-dimensional functional response from the \( i \)th treatment, where \( n = 2^J \), the test statistic can be expressed as,

\[
\kappa_\eta = \sum_{k=1}^{n_t} \hat{\theta}_{ik}^2 + \sum_{k=1}^{n-n_t} \theta_{ik}^2
\]  
(4.23)

where,

\[
n_t = \frac{2^J}{2^t}
\]  
(4.24)

In equation (4.23), \( \hat{\theta} \) correspond to the thresholded wavelet coefficients whereas \( \theta \) corresponds to the wavelet coefficients without a threshold. Under the assumptions of normality and recognizing that \( \hat{\theta}_{ik} \sim \chi^2_1[\lambda] \) and \( \theta_{ik} \sim \chi^2_1 \),

\[
\sum_{k=1}^{n-n_t} \hat{\theta}_{ik}^2 \sim \chi^2_{(n-n_t)}[\lambda] \quad (4.25)
\]

\[
\sum_{k=1}^{n_t} \theta_{ik}^2 \sim \chi^2_{(n_t)} \quad (4.26)
\]

Thus, the density of \( \kappa_\eta \) is,

\[
f_{\kappa_\eta}(x) = f_{\chi^2_{(n-n_t)}}(x) \otimes f_{\chi^2_{(n-n_t)}[\lambda]}(x)
\]  
(4.27)

where \( \otimes \) represents a convolution operator. With the densities \( f_{\chi^2_{(n-n_t)}}(x) \) and \( f_{\chi^2_{(n-n_t)}[\lambda]}(x) \), the density of \( f_{\kappa_\eta}(x) \) can be evaluated numerically. Thus, the exact distribution of \( f_{\kappa_\eta}(x) \) has only an analytic expression in terms of known distributions.

In addition, approximations to the density in (4.27) are possible using common distributions with analytic expressions. Such approximations are discussed in Chapter 5. In addition, with statistical testing, \( \kappa_\eta \) has certain properties that make computation of the density in equation (4.29) straightforward. For statistical testing, \( \kappa_\eta \) is defined as having two parameters, \( p \) and \( q \); where \( p = (n-n_t) \) and \( q = n_t \) and can be denoted as \( \kappa^{p,q}_\eta \).
Lemma 4.1.7 The exact distribution of $\kappa_{p,q}^{\eta,\lambda}$ with a threshold of $\lambda$ is obtained using the density function,

$$f_{\kappa_{\eta}}(x) = f_C(x) \frac{1 - \Phi_B(p\lambda^2, q/2, p/2)}{1 - \Phi_G(p\lambda^2/2)}$$

(4.28)

where $C \sim \chi_{p+q}^2$, $\Phi_B(\bullet)$ and $\Phi_G(\bullet)$ are the cumulative distribution functions of a Beta and Gamma random variables respectively.

A detailed derivation of the density for $\kappa_\eta$ is given in A.1.7. The result of Lemma (4.1.7) offers a way to obtain critical values directly from the distribution of $\kappa_\eta$ instead of a computational approach.

Proposition 4.1.8 Under the assumptions of normality and a hard, universal threshold, for a statistical test involving functional responses with a hypothesis as defined in (2.5) the statistic $\kappa_\eta$ under $H_0$ has $p = 1$ and $q = n_t$.

Proof Consider a functional response, $f \in \mathbb{R}^n$. Under $H_0$, $f \sim N(0, 1)$ and $\lambda = \sqrt{2 \ln(n)}$. Define,

$$\Theta = \mathbb{W}[f]$$

(4.29)

where $\Theta \in \mathbb{R}^n$. Let ‘X’ denote the number of coefficients that remain after hard thresholding. For convenience, assume a full DWT decomposition of $f$ and $n = 2^J$. Then,

$$\lambda = \sqrt{2 \times \ln(2) \log_2(2^J)}$$

(4.30)

$$\approx \sqrt{1.386J}$$

(4.31)

$$\approx 1.18\sqrt{J}$$

(4.32)

The probability of a non-zero wavelet coefficient ($\pi$) after a threshold is,

$$\Pr(|\theta_i| \geq \lambda) = \Pr(\theta_i > \lambda) + \Pr(-\theta_i > \lambda)$$

(4.33)

$$= \Pr(\theta_i > \lambda) + \Pr(\theta_i < -\lambda)$$

(4.34)

$$= 2\Pr(\theta_i > \lambda)$$

(4.35)

$$= 2(1 - \Pr(\theta_i \leq \lambda))$$

(4.36)

$$= 2(1 - \Phi(1.18\sqrt{J}))$$

(4.37)
where $\Phi(\bullet)$ is the CDF of a standard normal distribution. Equation (4.37) holds true due to the assumption of normality. Since functional responses are high-dimensional, an assumption of $J \geq 5$ is reasonable. Thus, for $J \geq 5$,

$$\pi = \Pr(\theta \geq \lambda) \leq 5e^{-2} \quad (4.38)$$

Assuming $l_t$ as the level up to which a threshold is applied, the number of coefficients for which a threshold is applied is $n_t$. The number of non-zero coefficients ($N_{\theta}$) among these $n_t$ coefficients can be modelled as a binomial distribution with parameters $n_t$ and probability of success as $\pi$. That is,

$$N_{\theta} \sim \text{Bin}(n_t, \pi) \quad (4.39)$$

Therefore, the expected number of non-zero wavelet coefficients is,

$$E[N_{\theta}] = n_t \pi \quad (4.40)$$

with $J \geq 5$, $E[N_{\theta}] \leq \frac{1}{4}$. Thus, a conservative estimator for the expected value is ‘1’ and the expected value corresponds to $p$ in the proposition.

The bounds established in equation (4.40) can be easily verified via simulation. For instance, let $l_t = 4$. Under the assumptions of normality, the expected number of coefficients that might exceed the threshold defined in (4.37) under $H_0$ is shown in Figure 4.2. Note that even for smaller sample sizes the expected number of coefficients is well within bounds and setting $p$ as stated in Proposition (4.1.8) is reasonable.

The results in Figure 4.2 remain almost the same when $l_t$ is changed and a different wavelet filter is employed. Hence, the result in Proposition 4.1.8 is very generalized and can be extended to a variety of wavelet based analyses. The probability, $\pi$, holds true for functional response of any known distribution provided the distribution of wavelet coefficients are known. For an arbitrary
threshold, $\delta$, the probability is,

$$\pi = 2(1 - F(\delta))$$  \hspace{1cm} (4.41) $$

where $F$ is the cumulative distribution function of the wavelet coefficients.

### 4.2 Thresholding Model using a Binomial Distribution

The proposition (4.1.8) is proved with a binomial distribution assumption for the wavelet coefficients after a hard threshold. This can be demonstrated more elegantly as shown in this section. The definition of $\pi$ in (4.41) is used here. Each wavelet coefficient can be defined as,

$$\tilde{\theta}_i = \theta_i B_i$$  \hspace{1cm} (4.42) $$

where $B_i$ is a Bernoulli random variable with probability of success, $\pi$.

The parameter, $p$, in $\kappa_{q}^{p,q}$, under the assumptions of normality, independence of wavelet coeffi-
cients, and hard thresholding is,

\[
p = E\left[\sum_{i=1}^{(n-n_t)} \tilde{\theta}_i^2\right] \tag{4.43}
\]

\[
= E\left[\sum_{i=1}^{(n-n_t)} \theta_i^2 \B_i^2\right] \tag{4.44}
\]

\[
= \sum_{i=1}^{(n-n_t)} E[B_i^2] E[\theta_i^2] \tag{4.45}
\]

\[
= \sum_{i=1}^{(n-n_t)} \pi E[\theta_i^2] \tag{4.46}
\]

\[
= (n - n_t)\pi \tag{4.47}
\]

Thus, \([p] = 1\) as shown in Figure 4.2.

**Corollary 4.2.1** For a level \(\alpha\) test involving ‘t’ treatments where the response is \(Y_i \in \mathbb{R}^n\) such that,

\[
Y = f_i + \epsilon \sim N(f, \Sigma) \tag{4.48}
\]

\[
i = 1, 2, \ldots t \tag{4.49}
\]

a measure based test statistic for such a test is,

\[
\kappa_\eta = \sum_{i=1}^{t} \sum_{k=1}^{(n-m_i)} \tilde{\theta}_{ik}^2 \tag{4.50}
\]

Under \(H_0\), \(\kappa_\eta \sim \kappa_\eta^{1,m}\) where \(\sum_{i=t}^{t} m_i\), and \(m_i = \frac{n}{2l_i}\) for the \(i^{th}\) treatment with \(l_i\) as the level up to which a threshold is applied to the wavelet coefficients of \(i^{th}\) response. Then, the rejection region is defined as,

\[
\Omega = \kappa_\eta \geq \Pr_{H_0}(\kappa_\eta \geq \kappa_\eta^{1,m}) \leq \alpha \tag{4.51}
\]
Figure 4.3: Density of $\kappa_\eta$ using analytic expression in Lemma 4.1.7 (solid line) and the histogram of empirical $\kappa_\eta$ values obtained using 5000 simulations. A Haar wavelet is used in DWT and $l_t = 4$ with full decomposition corresponding to $l = \log_2(2^l)$. A hard, universal threshold is applied. Thus, $\lambda = 1.18\sqrt{J}$, $p = 1$ and $q = n/2^l$.

Under $H_0$, the simulated and exact density of $\kappa_\eta$ are obtained and shown in Figure 4.3. It can be observed that the density specified in (4.27) is the exact distribution for $\kappa_\eta$ although the tail probability is slightly underestimated by the exact distribution for smaller sample sizes. This is expected behaviour for wavelet transformations as they are well suited for high dimensional responses and it is attributed to the sparsity property.

In the preceding chapter, the density of $\kappa_\eta$ is used to assess the performance of a non-parametric test using $\kappa_\eta$. Also, several approximations and other considerations pertaining to $\kappa_\eta$ and its use with practical data analysis are discussed.
CHAPTER 5

PERFORMANCE OF A $\kappa_\eta$ BASED TEST AND PRACTICAL CONSIDERATIONS

A test for functional responses as defined in (2.2) can be tested using $\kappa_\eta$ (see Chapter 4) and the Hotelling $T^2$ statistic, $\vartheta$ (see (A.1.2)). In chapter 2, it was demonstrated that a MANOVA is not the best option for testing functional responses with very high dimensions.

The Hotelling $T^2$ statistic can also be used to test functional responses. However, like MANOVA based testing, $\vartheta$ has performance issues due to heteroskedasticity and requirement of a minimum replicate size. In this chapter, the performance of a non-parametric test using $\kappa_\eta$ as the test statistic is compared with tests involving $\vartheta$ under various experimental parameters. The performance is measured in terms of Type-I and Type-II errors (see (2.2.2)). Most simulations in this chapter and all statistical analyses are carried out using the R programming environment [130].

5.1 Performance Simulations

Simulations are carried out with $\kappa_\eta$ and $\vartheta$ as test statistics in an $\alpha$-level test and the results are summarized. In the ensuing simulations and results, two types are differences are considered, local and global (See Chapter 1), with profiles involving varying dimensions, replicate sizes, and balanced and unbalanced experimental designs. The underlying functional response of at least one treatment is defined to be different from the other treatments. A difference ($\delta$) is introduced in the simulation and the power of the statistical test in detecting this difference with reasonable type-I error control is studied.

5.1.1 Measure of Difference

The difference measure used in the simulation is the well known Integrated Mean Squared Error (IMSE) of two functions $f_1(x)$ and $f_2(x)$. Here, $f_1$ and $f_2$ are the underlying mean responses of
the actual signal. Thus, the difference measure can be formally defined as,

\[ \delta = \int [f_1(x) - f_2(x)]^2 dx \]  \hspace{1cm} (5.1)

However, in a computational sense, we can define the measure as,

\[ \delta = \frac{\sum_{i=1}^{n} [f_{1i} - f_{2i}]^2}{n\sigma^2} \]  \hspace{1cm} (5.2)

Since \( \sigma \) is known, a priori, in the simulation. It can be used directly to define the difference measure. \( \sigma \) is used merely as a normalization factor.

For \( T \) treatments, \( \delta_d \) can be approximated as,

\[ \delta = \frac{\sum_{i=1}^{n} [f_{Ti} - \bar{f}_i]^2}{n\sigma^2} \]  \hspace{1cm} (5.3)

That is, a difference is introduced in one of the \( T \) treatments. Here, the last treatment is chosen to keep the simulation simple. However, one can randomly select the treatment (or several treatments) so that the total difference is \( \delta \). Also, one can simulate more than one local difference among profiles as long as the total measure of difference among all treatments is \( \delta \).

### 5.1.2 Mean Response

In general, the underlying mean response could be any measurable function. Keeping simplicity in mind one may resign to the class of polynomial functions defined as,

\[ f_k(x) = \sum_{j=0}^{N} a_j^{(k)} x^j \]  \hspace{1cm} (5.4)

The mean responses in the simulation are a transcendental function of degree 1 (linear function) where,

\[ f = 7x^2 + \sin(x) + 6\cos(2x) + 3x^2 \]  \hspace{1cm} (5.5)
For a global difference, a simple mean shift is considered. That is, \( f_2(x) = f_1(x) + \hat{\delta}_d \). A local difference can be simulated by introducing a mean shift on an interval \( \mathcal{A} \subset \mathcal{D} \), where \( \mathcal{D} \) is the domain of the functions.

### 5.2 Simulation Setup and Results

In the simulation, the following approach is commonly employed across every variant.

- The local difference is simulated on the interval (0.5,0.6) where the domain of the function is \([0,1]\).
- The sample size is denoted by \( n \), usually a power to 2 (a wavelet requirement).
- The Haar wavelet is used for tests that involve wavelet based statistics. The choice of the wavelet is “haar” as it represents the “simplest” wavelet basis.
- The noise level is defined by \( \sigma^2 \)
- The power curves are smoothed using a smoothed spline estimator using the \texttt{smooth.spline()} function in \texttt{R}. The actual power estimations are displayed along with the smoothed curve estimates.
- The median absolute deviation (MAD) is used as the estimator for \( \sigma \). However, the conventional standard deviation could also be used as the estimator.

#### 5.2.1 Two-treatment Setup

In this section, the results of a two treatment setup are obtained. For the balanced case, the replicate size is 2 for both treatments and for the unbalanced case, the replicate sizes are 3 and 5 for treatment 1 and 2 respectively. The level of significance is set to 0.05 and the exact distribution of \( \kappa_\eta \) is used to estimate the critical values.

It is possible to obtain the critical values by sampling from the distribution of \( \kappa_\eta \) by employing Monte-Carlo methods such as the Metropolis-Hastings algorithm [34]. However, the existence of an exact distribution evades the necessity for such computational analysis. It can be noted from Figure 5.1 that \( \kappa_\eta \) offers better power and Type-I error control under this setup. With local differ-
(a) The power curves for the two-treatment, balanced setup with \( n = 1024 \). A global difference is expected to be detected by the tests. The solid blue line corresponds to the power curve for the Hotelling \( T^2 \) based test and the solid red line corresponds to the power curve for the \( \kappa_\eta \) test.

(b) Minimum detectable global difference for \( \kappa_\eta \) (solid red) and \( \vartheta \) (solid blue). The solid black line corresponds to the underlying mean response (see equation 5.6) with no difference.

Figure 5.1: Performance comparison in detecting global differences with \( \kappa_\eta \) and \( \vartheta \) in a two-treatment, balanced setup where \( \sigma_1^2 = \sigma_2^2 = 1 \) and \( r_1 = r_2 = 2 \).

Table 5.1: Summary of Type-I error for the simulations and the minimum detectable difference for the two treatment setup, balanced and unbalanced cases.

<table>
<thead>
<tr>
<th></th>
<th>Balanced</th>
<th>Unbalanced</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( \vartheta )</td>
<td>( \kappa_\eta )</td>
</tr>
<tr>
<td>( \alpha )</td>
<td>Global 0.08 Local 0.05</td>
<td>Global 0.03 Local 0.06</td>
</tr>
<tr>
<td>( \delta_{\min} )</td>
<td>0.2745 0.3019</td>
<td>0.0980 0.1255</td>
</tr>
</tbody>
</table>

In some situations, it can be noted from Figure 5.2 that both \( \kappa_\eta \) as well as \( \vartheta \) require a relatively larger magnitude of local difference but the wavelet test statistic is better at detection than its counterpart. A similar analysis is carried out with an unbalanced setup. It can be observed from Figures 5.1, 5.2, 5.3 and 5.4 that \( \kappa_\eta \) performs better than \( \vartheta \). The performance of \( \vartheta \) is still reasonable although the minimum \( \delta \) that can be detected is slightly larger than \( \kappa_\eta \). As noted in [109], this performance is affected by heteroskedasticity.

In some situations, heteroskedasticity may not be an issue by itself but when analyses involve heteroskedasticity and unbalanced replicate sizes, it is likely that both Type-I error and Type-II errors are inflated. Table 5.2.2 summarizes the estimated \( \alpha \) and the minimum detectable difference.

It is obvious that \( \kappa_\eta \) seems to perform better than \( \vartheta \). Hence, the succeeding analyses are
(a) The power curves for the two-treatment, balanced setup with \( n = 1024 \). A local difference is expected to be detected by the tests. The solid blue line corresponds to the power curve for the Hotelling \( T^2 \) based test and the solid red line corresponds to the power curve for the \( \kappa_\eta \) test.

(b) Minimum detectable local difference for \( \kappa_\eta \) (solid red) and \( \vartheta \) (solid blue). The solid black line corresponds to the underlying mean response (see equation 5.6) with no difference.

Figure 5.2: Performance comparison in detecting local differences with \( \kappa_\eta \) and \( \vartheta \) in a two-treatment, balanced setup where \( \sigma_1^2 = \sigma_2^2 = 1 \) and \( r_1 = r_2 = 2 \).

generalized to multiple sample sizes under varying experimental setups and the results are plotted. This approach ensures brevity and better visual interpretation. Also, functional responses of lengths greater than 1000 are considered. That is, \( n = 1024, 2048, 4096, \) and \( 8192 \), which would identify the effectiveness of the tests involving \( \kappa_\eta \) in higher dimensions. The results of a balanced setup with \( r_1 = r_2 = 10 \) and \( \sigma_1^2 = \sigma_2^2 = 1 \) are shown in Figure 5.5. It can be noted that \( \kappa_\eta \) is better capable of detecting small differences under noisy conditions than \( \vartheta \). The results of a balanced setup with \( r_1 = 10 \) and \( r_2 = 3 \) and \( \sigma_1^2 = \sigma_2^2 = 1 \) are shown in Figure 5.6. The unbalanced setup doesn’t seem to have an effect on the performance of either test statistic. In Figure 5.7, the results of a balanced but heteroskedastic case is considered. In this case, \( r_1 = 10 \) and \( r_2 = 10 \) and \( \sigma_1^2 = 1 \) and \( \sigma_2^2 = 2 \). Both \( \kappa_\eta \) and \( \vartheta \) perform well with heteroskedasticity with \( \kappa_\eta \) offering a slight advantage over its counterpart. Finally, a heteroskedastic, unbalanced setup is considered. For \( \kappa_\eta \), the replicate sizes are \( r_1 = 10 \) and \( r_2 = 3; \sigma_1^2 = 1 \) and \( \sigma_2^2 = 2 \). From Figure 5.8, the performance of \( \kappa_\eta \) remains quite efficient under heteroskedastic and unbalanced replicate sizes. However, this is not the case with \( \vartheta \). The results of \( \vartheta \) can only be obtained with larger replicate sizes that tend to mitigate, up to a certain degree, the effects of heteroskedasticity. The replicate sizes are \( r_1 = 10 \) and \( r_2 = 20 \). \( \vartheta \) demonstrates poor performance under heteroskedasticity and unbalanced replicate sizes. An increase in replicate sizes did not improve the performance and a significantly large number of
(a) The power curves for the two-treatment, unbalanced setup with $n = 1024$. A global difference is expected to be detected by the tests. The solid blue line corresponds to the power curve for the Hotelling $T^2$ based test and the solid red line corresponds to the power curve for the $\kappa_\eta$ test.

(b) Minimum detectable global difference for $\kappa_\eta$ (solid red) and $\vartheta$ (solid blue). The solid black line corresponds to the underlying mean response (see equation 5.6) with no difference.

Figure 5.3: Performance comparison in detecting global differences with $\kappa_\eta$ and $\vartheta$ in a two-treatment, unbalanced setup where $\sigma_1^2 = \sigma_2^2 = 1$ and $r_1 = 3$ and $r_2 = 5$

(a) The power curves for the two-treatment, unbalanced setup with $n = 1024$. A local difference is expected to be detected by the tests. The solid blue line corresponds to the power curve for the Hotelling $T^2$ based test and the solid red line corresponds to the power curve for the $\kappa_\eta$ test.

(b) Minimum detectable global difference for $\kappa_\eta$ (solid red) and $\vartheta$ (solid blue). The solid black line corresponds to the underlying mean response (see equation 5.6) with no difference.

Figure 5.4: Performance comparison in detecting local differences with $\kappa_\eta$ and $\vartheta$ in a two-treatment, unbalanced setup where $\sigma_1^2 = \sigma_2^2 = 1$ and $r_1 = 3$ and $r_2 = 5$
(a) The power curves for the two-treatment, balanced setup with n=1024 (black), 2048 (red), 4096 (green) and 8192 (blue). A global difference is expected to be detected by the tests. $\kappa_\eta$ is used as the test statistic.

(b) The power curves for the two-treatment, balanced setup with n=1024 (black), 2048 (red), 4096 (green) and 8192 (blue). A global difference is expected to be detected by the tests. $\vartheta$ is used as the test statistic.

(c) The power curves for the two-treatment, balanced setup with n=1024 (black), 2048 (red), 4096 (green) and 8192 (blue). A local difference is expected to be detected by the tests. $\kappa_\eta$ is used as the test statistic.

(d) The power curves for the two-treatment, balanced setup with n=1024 (black), 2048 (red), 4096 (green) and 8192 (blue). A local difference is expected to be detected by the tests. $\vartheta$ is used as the test statistic.

Figure 5.5: Performance comparison for $\kappa_\eta$ and $\vartheta$ in a two-treatment, balanced setup where $\sigma_1^2 = \sigma_2^2 = 1$ and $r_1 = r_2 = 10$
replicates might be necessary to improve it, if possible. Clearly, $\vartheta$ suffers from very high Type-II error under conditions of heteroskedasticity. This setup demonstrates one of the advantages of $\kappa_{\eta}$ over $\vartheta$.

### 5.2.2 Multi-treatment Setup

The analysis in the previous section is extended to a multi-treatment scenario where several treatments are compared simultaneously. As mentioned earlier in Section 5.1.1, one of ‘t’ treatments is simulated to be different from the other treatment profiles. The results of a multi-treatment,
Figure 5.7: Performance comparison for $\kappa_\eta$ and $\vartheta$ in a two-treatment, balanced setup where $\sigma_1^2 = 2$ and $\sigma_2^2 = 2$ and $r_1 = r_2 = 10$.

balanced and homoskedastic simultaneous test for profiles of varying lengths are shown in Figure 5.9. The performance is similar to the two-treatment case, with $\kappa_\eta$ being able to perform better detection than $\vartheta$. A similar conclusion can be derived for a multi-treatment, balanced setup under heteroskedasticity from the results in Figure 5.10.

A similar analysis can be carried out with an unbalanced setup under homoskedasticity (see Figure 5.11) and heteroskedasticity (see Figure 5.12). It can be noted that $\vartheta$ presents problems in controlling Type I and Type II errors similar to the two treatment setup. Of course, a significantly large number replicates (see Figure 5.12) may alleviate this problem, but it may not be optimal in
(a) The power curves for the two-treatment, unbalanced setup with n=1024 (black), 2048 (red), 4096 (green) and 8192 (blue). A global difference is expected to be detected by the tests. $\kappa_\eta$ is used as the test statistic.

(b) The power curves for the two-treatment, unbalanced setup with n=1024 (black), 2048 (red), 4096 (green) and 8192 (blue). A global difference is expected to be detected by the tests. $\vartheta$ is used as the test statistic.

(c) The power curves for the two-treatment, unbalanced setup with n=1024 (black), 2048 (red), 4096 (green) and 8192 (blue). A local difference is expected to be detected by the tests. $\kappa_\eta$ is used as the test statistic.

(d) The power curves for the two-treatment, unbalanced setup with n=1024 (black), 2048 (red), 4096 (green) and 8192 (blue). A local difference is expected to be detected by the tests. $\vartheta$ is used as the test statistic.

Figure 5.8: Performance comparison for $\kappa_\eta$ and $\vartheta$ in a two-treatment, unbalanced setup where $\sigma_1^2 = 1$ and $\sigma_2^2 = 2$ and $r_1 = 10$ and $r_2 = 20$ for tests involving $\vartheta$ and $r_1 = 3$ and $r_2 = 5$ for tests involving $\kappa_\eta$.

practical applications, thus making $\kappa_\eta$ a viable alternative.

In Table 5.2.2, the Type-I error achieved by $\kappa_\eta$ and $\vartheta$ under various experimental settings are summarized. The error for $\vartheta$ is 1 under conditions of heteroskedasticity and unbalanced replicate sizes. As discussed earlier, increasing replicate sizes seems to decrease Type-I error, but a very large number of replicates may be necessary to achieve good control of the error. From the analyses in this section, the effectiveness of $\kappa_\eta$ is realized, and in the following sections several practical considerations in using $\kappa_\eta$ as a test statistic in non-parametric testing are discussed.
The power curves for the four-treatment, balanced setup with $n=1024$ (black), 2048 (red), 4096 (green) and 8192 (blue). A global difference is expected to be detected by the tests. $\kappa_\eta$ is used as the test statistic.

The power curves for the four-treatment, balanced setup with $n=1024$ (black), 2048 (red), 4096 (green) and 8192 (blue). A global difference is expected to be detected by the tests. $\vartheta$ is used as the test statistic.

The power curves for the four-treatment, balanced setup with $n=1024$ (black), 2048 (red), 4096 (green) and 8192 (blue). A local difference is expected to be detected by the tests. $\kappa_\eta$ is used as the test statistic.

The power curves for the four-treatment, balanced setup with $n=1024$ (black), 2048 (red), 4096 (green) and 8192 (blue). A local difference is expected to be detected by the tests. $\vartheta$ is used as the test statistic.

Figure 5.9: Performance comparison for $\kappa_\eta$ and $\vartheta$ in a four-treatment, balanced setup where $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma_4^2 = 1$ and $r_1 = r_2 = r_3 = r_4 = 5$

The primary measure of $\kappa_\eta$ in testing high-dimensional profiles is Type-I error ($\alpha$) control. Since $\kappa_\eta$ achieves good power, it is sufficient to measure the effectiveness of $\kappa_\eta$ using $\alpha$.

5.2.3 Type of Noise Estimator

$\hat{\sigma}$, the estimator for the noise parameter, plays a significant role in the statistic and several estimators are available. In practice, one of the following approaches can be considered,

- **Known**: The estimator is known and supplied in evaluating the statistic. Although very effective, this is not realistic as $\sigma$ is usually unknown.
(a) The power curves for the four-treatment, balanced setup with $n=1024$ (black), 2048 (red), 4096 (green) and 8192 (blue). A global difference is expected to be detected by the tests. $\kappa_\eta$ is used as the test statistic.

(b) The power curves for the four-treatment, balanced setup with $n=1024$ (black), 2048 (red), 4096 (green) and 8192 (blue). A global difference is expected to be detected by the tests. $\vartheta$ is used as the test statistic.

(c) The power curves for the four-treatment, balanced setup with $n=1024$ (black), 2048 (red), 4096 (green) and 8192 (blue). A local difference is expected to be detected by the tests. $\kappa_\eta$ is used as the test statistic.

(d) The power curves for the four-treatment, balanced setup with $n=1024$ (black), 2048 (red), 4096 (green) and 8192 (blue). A local difference is expected to be detected by the tests. $\vartheta$ is used as the test statistic.

Figure 5.10: Performance comparison for $\kappa_\eta$ and $\vartheta$ in a four-treatment, balanced setup where $\sigma_1^2 = 1, \sigma_2^2 = 2, \sigma_3^2 = 0.5, \sigma_4^2 = 1$ and $r_1 = r_2 = r_3 = r_4 = 5$

- **Standard deviation**: The unbiased noise estimator. Under normality assumptions, it can be assumed to be one.

- **MAD**: The median absolute deviation has been suggested by authors [122] engaged in wavelet based data analysis and it offers good mathematical properties in certain applications. More precisely, the estimate $\hat{\sigma}$ is,

$$\hat{\sigma} = \mathbb{M}[|W_{(J-1)k} - \mathbb{M}[W_{(J-1)k}]|]$$

(5.6)
Figure 5.11: Performance comparison for $\kappa_\eta$ and $\vartheta$ in a four-treatment, unbalanced setup where $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma_4^2 = 1$ and $r_1 = 10, r_2 = 5, r_3 = 8, r_4 = 4$
(a) The power curves for the four-treatment, unbalanced setup with n=1024 (black), 2048 (red), 4096 (green) and 8192 (blue). A global difference is expected to be detected by the tests. $\kappa_{\eta}$ is used as the test statistic.

(b) The power curves for the four-treatment, unbalanced setup with n=1024 (black), 2048 (red), 4096 (green) and 8192 (blue). A global difference is expected to be detected by the tests. $\vartheta$ is used as the test statistic.

(c) The power curves for the four-treatment, unbalanced setup with n=1024 (black), 2048 (red), 4096 (green) and 8192 (blue). A local difference is expected to be detected by the tests. $\kappa_{\eta}$ is used as the test statistic.

(d) The power curves for the four-treatment, unbalanced setup with n=1024 (black), 2048 (red), 4096 (green) and 8192 (blue). A local difference is expected to be detected by the tests. $\vartheta$ is used as the test statistic.

Figure 5.12: Performance comparison for $\kappa_{\eta}$ and $\vartheta$ in a four-treatment, unbalanced setup where $\sigma_1^2 = 1, \sigma_2^2 = 2, \sigma_3^2 = 0.5, \sigma_4^2 = 1$. Here $r_1 = 10, r_2 = 5, r_3 = 8, r_4 = 4$ for $\kappa_{\eta}$ based testing and $r_1 = 10, r_2 = 15, r_3 = 18, r_4 = 14$ for $\vartheta$ based testing.
Table 5.2: Summary of Type-I errors achieved by $\kappa_\eta$ and $\vartheta$ under various experimental settings. The 'la8' filter is employed with a full DWT decomposition and $l_t = 4$. For the balanced case, $r_i = 5$, $\forall i$ and for the unbalanced case, $r_1 = 10$, $r_2 = 8$, $r_3 = 5$ and $r_4 = 4$.

<table>
<thead>
<tr>
<th></th>
<th>$\sigma_i^2 = 1 \forall i$</th>
<th>$\sigma_i^2 = \sigma_3^2 = 1$; $\sigma_2^2 = 2 \sigma_4^2 = 0.5$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\kappa_\eta$</td>
<td>$\vartheta$</td>
</tr>
<tr>
<td>Balanced</td>
<td>1024</td>
<td>0.055</td>
</tr>
<tr>
<td></td>
<td>2048</td>
<td>0.075</td>
</tr>
<tr>
<td></td>
<td>4096</td>
<td>0.035</td>
</tr>
<tr>
<td></td>
<td>8192</td>
<td>0.115</td>
</tr>
<tr>
<td>Unbalanced</td>
<td>1024</td>
<td>0.055</td>
</tr>
<tr>
<td></td>
<td>2048</td>
<td>0.06</td>
</tr>
<tr>
<td></td>
<td>4096</td>
<td>0.06</td>
</tr>
<tr>
<td></td>
<td>8192</td>
<td>0.075</td>
</tr>
</tbody>
</table>

where $\bar{\alpha}$ is the ‘median’ operator and only the coefficients of the top-most level are considered as they capture most of the noisy components of the signal. The estimate in (5.6) can be scaled under normality assumptions in order to obtain a ‘consistent’ estimator. Such an estimator resembles the robust estimator proposed by Donoho and Johnstone.

The effect of the noise estimator on Type-I error is summarized in Table 5.3. From Tables 5.3 and 5.4, it is apparent that the choice of the estimator does not affect the Type-I errors attained.

5.2.4 Choice of Wavelet Filter

There are several type of wavelet filters and in most situations the choice of filters is dependent on the type of application. Some of the commonly used filters are Haar [161], Daubechies (d) [51] and least asymmetric (la) [63] filters. Also, the filter length (the number of filter taps/coefficients) tends to effect the results in signal processing applications. Although the choice of filter is crucial in signal processing applications, from a data analysis standpoint it may not cause a significant difference in test performance.

Table 5.5 summarizes the value of $\alpha$ attained when different wavelet filters are employed. It can be seen that the choice of wavelet filter (or filter length) doesn’t affect the control of Type-I error under homoskedasticity. In Table 5.6 a similar analysis is carried out for a heteroskedastic setup.
Table 5.3: Summary of $\alpha$ values attained by a test involving $\kappa_\eta$. Here $r_i = 5, \forall i$ for the balanced design and for the unbalanced design $r_1 = 10, r_2 = 5, r_3 = 8$ and $r_4 = 4$, and $\sigma_i^2 = 1 \forall i$. The ‘la8’ filter is used with a full DWT decomposition and $l_t = 4$. The level of confidence set for the test in the simulations is 0.05 and 200 repetitive simulations are performed to obtain the estimated errors above.

<table>
<thead>
<tr>
<th>n</th>
<th>Global MAD</th>
<th>Global SD</th>
<th>Local MAD</th>
<th>Local SD</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Balanced</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1024</td>
<td>0.105</td>
<td>0.015</td>
<td>0.05</td>
<td>0.035</td>
</tr>
<tr>
<td>2048</td>
<td>0.04</td>
<td>0.025</td>
<td>0.05</td>
<td>0.08</td>
</tr>
<tr>
<td>4096</td>
<td>0.055</td>
<td>0.035</td>
<td>0.08</td>
<td>0.045</td>
</tr>
<tr>
<td>8192</td>
<td>0.035</td>
<td>0.06</td>
<td>0.085</td>
<td>0.05</td>
</tr>
<tr>
<td>Unbalanced</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1024</td>
<td>0.05</td>
<td>0.045</td>
<td>0.04</td>
<td>0.055</td>
</tr>
<tr>
<td>2048</td>
<td>0.055</td>
<td>0.06</td>
<td>0.055</td>
<td>0.03</td>
</tr>
<tr>
<td>4096</td>
<td>0.045</td>
<td>0.05</td>
<td>0.03</td>
<td>0.065</td>
</tr>
<tr>
<td>8192</td>
<td>0.085</td>
<td>0.05</td>
<td>0.055</td>
<td>0.045</td>
</tr>
</tbody>
</table>

Table 5.4: Summary of $\alpha$ values attained by a test involving $\kappa_\eta$. Here $r_i = 5, \forall i$ for the balanced design and for the unbalanced design $r_1 = 10, r_2 = 5, r_3 = 8$ and $r_4 = 4$, and $\sigma_{1,3}^2 = 1, \sigma_2^2 = 2$ and $\sigma_4^2 = 0.5$. The ‘la8’ filter is used with a full DWT decomposition and $l_t = 4$. The level of confidence set for the test in the simulations is 0.05 and 200 repetitive simulations are performed to obtain the estimated errors above.

<table>
<thead>
<tr>
<th>n</th>
<th>Global MAD</th>
<th>Global SD</th>
<th>Local MAD</th>
<th>Local SD</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Balanced</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1024</td>
<td>0.065</td>
<td>0.025</td>
<td>0.065</td>
<td>0.04</td>
</tr>
<tr>
<td>2048</td>
<td>0.06</td>
<td>0.04</td>
<td>0.055</td>
<td>0.065</td>
</tr>
<tr>
<td>4096</td>
<td>0.08</td>
<td>0.075</td>
<td>0.085</td>
<td>0.075</td>
</tr>
<tr>
<td>8192</td>
<td>0.065</td>
<td>0.08</td>
<td>0.085</td>
<td>0.03</td>
</tr>
<tr>
<td>Unbalanced</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1024</td>
<td>0.07</td>
<td>0.03</td>
<td>0.04</td>
<td>0.07</td>
</tr>
<tr>
<td>2048</td>
<td>0.08</td>
<td>0.04</td>
<td>0.09</td>
<td>0.045</td>
</tr>
<tr>
<td>4096</td>
<td>0.1</td>
<td>0.06</td>
<td>0.075</td>
<td>0.035</td>
</tr>
<tr>
<td>8192</td>
<td>0.105</td>
<td>0.1</td>
<td>0.08</td>
<td>0.03</td>
</tr>
</tbody>
</table>

70
Table 5.5: Summary of $\alpha$ values attained by a test involving $\kappa_\eta$ with varying filter choices. Here $r_i = 5, \forall i$ for the balanced design and for the unbalanced design $r_1 = 10, r_2 = 5, r_3 = 8$ and $r_4 = 4$, and $\sigma_i^2 = 1, \forall i$. The level of confidence set for the test in the simulations is 0.05 and 200 repetitive simulations are performed to obtain the estimated errors above.

<table>
<thead>
<tr>
<th>n</th>
<th>d2 (Haar)</th>
<th>d4</th>
<th>d8</th>
<th>d16</th>
<th>la8</th>
<th>la16</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Global</td>
<td>Local</td>
<td>Global</td>
<td>Local</td>
<td>Global</td>
<td>Local</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Balanced</td>
<td>1024</td>
<td>0.05</td>
<td>0.025</td>
<td>0.04</td>
<td>0.04</td>
<td>0.055</td>
</tr>
<tr>
<td></td>
<td>2048</td>
<td>0.04</td>
<td>0.02</td>
<td>0.05</td>
<td>0.045</td>
<td>0.03</td>
</tr>
<tr>
<td></td>
<td>4096</td>
<td>0.05</td>
<td>0.045</td>
<td>0.035</td>
<td>0.055</td>
<td>0.25</td>
</tr>
<tr>
<td></td>
<td>8192</td>
<td>0.085</td>
<td>0.045</td>
<td>0.045</td>
<td>0.045</td>
<td>0.02</td>
</tr>
<tr>
<td>Unbalanced</td>
<td>1024</td>
<td>0.065</td>
<td>0.03</td>
<td>0.04</td>
<td>0.055</td>
<td>0.045</td>
</tr>
<tr>
<td></td>
<td>2048</td>
<td>0.03</td>
<td>0.04</td>
<td>0.055</td>
<td>0.065</td>
<td>0.055</td>
</tr>
<tr>
<td></td>
<td>4096</td>
<td>0.035</td>
<td>0.05</td>
<td>0.05</td>
<td>0.04</td>
<td>0.05</td>
</tr>
<tr>
<td></td>
<td>8192</td>
<td>0.04</td>
<td>0.05</td>
<td>0.08</td>
<td>0.045</td>
<td>0.065</td>
</tr>
</tbody>
</table>

It can be observed that longer filters seem to cause a very slight increase in $\alpha$ with longer profiles.

5.2.5 Level of Threshold ($l_t$)

Since the level of DWT has a direct association with the frequency content of the signal, it is natural to assume that the level of thresholding applied in $\kappa_\eta$ will have an effect on its performance. By definition, $\kappa_\eta$ accounts for such effects and the statistical properties of the statistic are flexible enough to handle different levels of threshold. Under the assumptions of normality, $\kappa_\eta$ performs independently from the level of threshold and the results are generally consistent with varying levels of threshold, $l_t$.

In Table 5.7, $\alpha$ attained by $\kappa_\eta$ in homoskedastic functional responses among multiple treatments for varying levels of threshold are summarized. It can be observed that a reasonable value of $l_t$ is
Table 5.7: Summary of $\alpha$ values attained by a test involving $\kappa_\eta$ with varying filter choices. Here $r_i = 5, \forall i$ for the balanced design and for the unbalanced design, $r_1 = 10, r_2 = 5, r_3 = 8$ and $r_4 = 4$, and $\sigma^2_i = 1\forall i$. The level of confidence set for the test in the simulations is 0.05 and 200 repetitive simulations are performed to obtain the estimated errors above. The level of threshold $l_t$ is varied from 4 to 10 and the la8 filter is employed in a full DWT decomposition with respect to the dimension of the functional response.

<table>
<thead>
<tr>
<th>Level of Threshold ($l_t$)</th>
<th>$l_t = 4$</th>
<th>$l_t = 5$</th>
<th>$l_t = 6$</th>
<th>$l_t = 7$</th>
<th>$l_t = 8$</th>
<th>$l_t = 9$</th>
<th>$l_t = 10$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Balanced</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1024</td>
<td>0.025</td>
<td>0.04</td>
<td>0.03</td>
<td>0.045</td>
<td>0.035</td>
<td>0.04</td>
<td>0.035</td>
</tr>
<tr>
<td>2048</td>
<td>0.01</td>
<td>0.02</td>
<td>0.06</td>
<td>0.01</td>
<td>0.075</td>
<td>0.04</td>
<td>0.045</td>
</tr>
<tr>
<td>4096</td>
<td>0.03</td>
<td>0.045</td>
<td>0.095</td>
<td>0.055</td>
<td>0.045</td>
<td>0.045</td>
<td>0.06</td>
</tr>
<tr>
<td>8192</td>
<td>0.045</td>
<td>0.03</td>
<td>0.065</td>
<td>0.06</td>
<td>0.075</td>
<td>0.055</td>
<td>0.075</td>
</tr>
<tr>
<td>Unbalanced</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1024</td>
<td>0.02</td>
<td>0.04</td>
<td>0.034</td>
<td>0.065</td>
<td>0.08</td>
<td>0.065</td>
<td>0.08</td>
</tr>
<tr>
<td>2048</td>
<td>0.04</td>
<td>0.06</td>
<td>0.045</td>
<td>0.03</td>
<td>0.045</td>
<td>0.05</td>
<td>0.075</td>
</tr>
<tr>
<td>4096</td>
<td>0.075</td>
<td>0.055</td>
<td>0.03</td>
<td>0.04</td>
<td>0.06</td>
<td>0.075</td>
<td>0.065</td>
</tr>
<tr>
<td>8192</td>
<td>0.08</td>
<td>0.065</td>
<td>0.045</td>
<td>0.045</td>
<td>0.05</td>
<td>0.055</td>
<td>0.09</td>
</tr>
</tbody>
</table>

at most 8 for most functional responses. The results are obtained by varying $l_t$ between 4 and 10, inclusive. With the la8 filter and longer functional responses the Type-I error increases slightly and it is recommended that $l_t$ is kept below 6 for functional responses with dimension 4096 or higher. The analysis is repeated with heteroskedastic data and the results are summarized in Table 5.8.

5.2.6 Zero Padding

The length of a response is set to be a power of 2 in this work. This is a generic requirement for a DWT. However, in certain applications, it may not be plausible to meet this requirement. Under such circumstances, the response can be padded with zeroes such that the padded length is a power of 2. However, the number of padded zeroes must be excluded from $q$ as the number of zeroes are known a priori.

5.3 Approximations for $\kappa_\eta$

The distribution of $\kappa_\eta$ can be approximated using known distributions under the assumptions of normality. In fact, all analyses in the previous sections are carried out with an approximation to $\kappa_\eta$, the normal approximation to $\kappa_\eta$ to be specific.
Table 5.8: Summary of α values attained by a test involving $\kappa_\eta$ with varying filter choices. Here $r_i = 5, \forall i$ for the balanced design and for the unbalanced design, $r_1 = 10, r_2 = 5, r_3 = 8$ and $r_4 = 4$, and $\sigma^2_{1,3} = 1, \sigma^2_{2} = 2$ and $\sigma^2_{4} = 0.5$. The level of confidence set for the test in the simulations is 0.05 and 200 repetitive simulations are performed to obtain the estimated errors above. The level of threshold $l_t$ is varied from 4 to 10 and the la8 filter is employed in a full DWT decomposition with respect to the dimension of the functional response.

<table>
<thead>
<tr>
<th>Level of Threshold ($l_t$)</th>
<th>$l_t = 4$</th>
<th>$l_t = 5$</th>
<th>$l_t = 6$</th>
<th>$l_t = 7$</th>
<th>$l_t = 8$</th>
<th>$l_t = 9$</th>
<th>$l_t = 10$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Global</td>
<td>Local</td>
<td>Global</td>
<td>Local</td>
<td>Global</td>
<td>Local</td>
<td>Global</td>
</tr>
<tr>
<td>Balanced</td>
<td>1024</td>
<td>0.035</td>
<td>0.055</td>
<td>0.045</td>
<td>0.06</td>
<td>0.055</td>
<td>0.115</td>
</tr>
<tr>
<td>Unbalanced</td>
<td>1024</td>
<td>0.06</td>
<td>0.045</td>
<td>0.06</td>
<td>0.065</td>
<td>0.075</td>
<td>0.065</td>
</tr>
<tr>
<td></td>
<td>2048</td>
<td>0.05</td>
<td>0.06</td>
<td>0.055</td>
<td>0.04</td>
<td>0.055</td>
<td>0.06</td>
</tr>
<tr>
<td></td>
<td>4096</td>
<td>0.05</td>
<td>0.06</td>
<td>0.04</td>
<td>0.055</td>
<td>0.075</td>
<td>0.06</td>
</tr>
<tr>
<td></td>
<td>8192</td>
<td>0.115</td>
<td>0.07</td>
<td>0.065</td>
<td>0.065</td>
<td>0.06</td>
<td>0.05</td>
</tr>
</tbody>
</table>

Proposition 5.3.1 Let $X \sim \kappa(\eta)^{(p,q,\lambda)}$, then the mean and variance of $\kappa(\eta)^{(p,q,\lambda)}$ are given by,

$$
\mu_{\kappa(\eta)^{(p,q,\lambda)}} = E[X] = q + \mu \quad (5.7)
$$

$$
\sigma^2_{\kappa(\eta)^{(p,q,\lambda)}} = Var[X] = 2q + \sigma^2 \quad (5.8)
$$

where,

$$
\mu = \frac{8p\Gamma(3)(1 - \Phi_G(\lambda/2, 3, 1))}{\sqrt{2\pi}} \quad (5.9)
$$

$$
\sigma^2 = \frac{32p\Gamma(5)(1 - \Phi_G(\lambda/2, 5, 1))}{\sqrt{2\pi}} - p\mu^2 \quad (5.10)
$$

and $\Phi_G(\bullet, s_1, s_2)$ is the cumulative distribution of a Gamma distribution with parameters $s_1$ and $s_2$ respectively.

The proof of Proposition 5.3.1 is given in Section A.1.8. The approximations can be obtained using Proposition 5.3.1.
Figure 5.13: Histogram of simulated $\kappa_{\eta}$ values with $p = 1, q = n/2$. Here, $n = 128$ (top) and $n = 1024$ (bottom). The Haar filter is used with a full DWT decomposition.

5.3.1 Normal Approximation

The normal approximation to $\kappa_{\eta}$ is obtained using Central Limit Theorem (CLT) [140]. With high-dimensional profiles, as a consequence of CLT,

$$\frac{(S_n - \mu_{\kappa_{\eta}^{p,q}})\sqrt{n}}{\sigma_{\kappa_{\eta}^{p,q}}} \sim \xi$$

(5.11)

where $\xi \sim N(0,1)$ and $S_n \sim \kappa_{\eta}^{(p,q,\lambda)}$.

5.3.2 Chi-squared Approximation

Since most likelihood ratio tests have a Chi-squared distribution [31], $\kappa_{\eta}$ can be approximated by a Chi-squared distribution as well. The degrees of freedom of such an approximation under the null distribution is $p + q$. A better approximation of the degrees of freedom can be obtained using the normal approximation in 5.3.1. Thus, the Chi-squared approximation is given by,

$$S_n \sim \chi^2_K$$

(5.12)
where $K = \lceil q + \mu \rceil$, where $\mu$ is defined in equation 5.10.

### 5.3.3 Binomial-Normal Mixture Approximation

It was shown in chapter 4 that a reasonable value for ‘$p$’ is 1. The above approximations use $p$ as 1 in most applications. However, an alternative approach is to use probabilities for each value of ‘$p$’. The probability of a non-zero coefficient after DWT is given by $\pi$ (see Chapter 4). Let ‘$P$’ be the number of non-zero coefficients (the degrees of freedom, $p$) for $\kappa_{\eta}^{p,q}$. Then, for a level of threshold $(l_t)$,

$$P \sim \text{Bin}(n/2^t, \pi) \tag{5.13}$$

$$\omega_i = \Pr(P = i) \tag{5.14}$$

Define,

$$Z_{\kappa_{\eta}} = \sum_{i=1}^{N_t} \Pr(P = i)\kappa_{\eta}^{i,q} \tag{5.15}$$
Due to independence and normality assumptions,

\[ \mu_{Z^{(p,q,\lambda)}} = E[Z] = \sum_{i=1}^{N_t} \omega_i \mu_{\kappa^{(i,q,\lambda)}} \quad (5.16) \]

\[ \sigma^2_{Z^{(p,q,\lambda)}} = Var[Z] = \sum_{i=1}^{N_t} \omega_i^2 \sigma^2_{\kappa^{(i,q,\lambda)}} \quad (5.17) \]

Using CLT, an approximation similar to (5.11) can be obtained for \( \kappa_\eta \) expressed in terms of mixture probabilities given by \( \omega_i \). That is,

\[ \frac{(Z_n - \mu_{Z^{(p,q,\lambda)}}) \sqrt{n}}{\sigma_{Z^{(p,q,\lambda)}}} \sim \xi \quad (5.18) \]

where \( \xi \sim N(0,1) \) and \( Z_n = \frac{1}{n} \sum_{i=1}^{n} Z_i \), \( Z_i \sim \kappa^{(i,q,\lambda)} \). It can be seen from the tables in Chapter B and Figures 5.13 and 5.14, the approximations serve as good estimators for the exact distribution of \( \kappa_\eta \) especially for longer profiles. The approximations seem to get better as the signal dimension increases.
CHAPTER 6

APPLICATIONS OF $\kappa_\eta$ BASED TESTS

In Chapter 5, the performance of $\kappa_\eta$ is shown to be better than $\vartheta$, the Hotelling $T^2$ statistic, when differences in high dimensional profiles. In addition, several statistical properties of $\kappa_\eta$ and its exact distribution offer a sound framework for statistical testing.

In this chapter, $\kappa_\eta$ is used in real-life applications in detecting changes in profiles from various fields of science such as engineering, genomics and quantitative finance.

6.1 Applications in Engineering

Discrete Wavelet Transforms (DWT), as mentioned repeatedly in previous chapters, is a fascinating concept employed in electrical and communications engineering. Hence, our first application is employed to one of the simplest classification tasks in electrical engineering, signal processing. The data used in this analyses is the ISOlated LETter Speech Recognition (ISOLET) database which can be downloaded from the UCI Machine Learning Repository [18]. The ISOLET database is a well-known data collection which has been used in various research topics in the fields of machine learning and artificial intelligence. [42, 124, 16, 100, 54, 55, 8, 9]. Therefore, the ISOLET database will serve as a benchmark dataset in explaining the effectiveness $\kappa_\eta$ as a classifier.

6.1.1 Data Description

The data consists of 6 groups: isolet1 through isolet5 are training examples and isolet6 is a test case. Each group consists of 30 speakers and each speaker pronounced all letters in the English Alphabet twice. 2 profiles (Letter ‘F’) are missing in the training example and one profile (Letter ‘M’) is missing in the test example. Thus, in total, there are 6238 training and 1559 test profiles. The dimension (n) of each profile is 617. The 618th observation in the profile represents the letter index. Of course, the grouping can be changed and reassigned for a different analysis. The usual
(a) Mean Response for letter B pronounced by speakers in the training database
(b) Mean Response for letter J pronounced by speakers in the training database

Figure 6.1: Functional responses for letters B and J from the ISOLET database training set.

grouping is employed to ensure easy comparison with previous results.

6.1.2 Methods

In the past, several methods have been employed to classify the letters accurately. Some of the notable methods are Opt Backpropagation neural network with 56 hidden units and 26 output units (one-per-class) [42], a method using the error-correction codes [55], Linear Discriminant Analysis (LDA), Principal Component Analysis (PCA), MRMI [156] and a modification of the MRMI method [124]. From Figure 6.1, it can be seen that the differences between pronunciations of two letters are extremely small. In a statistical sense, the densities of these two profiles are nearly identical (see Figure 6.2). Thus, there exists a lack of separability when a p-value based decision boundary is employed.

For the results summarized in Table 6.1, an ‘la8’ filter is used with $l_t = 4$ and a maximum level DWT decomposition (here, it is 10) is carried out. The noise estimator is set to ‘1’ for both treatments. The actual standard deviation is approximately equal to 0.14 and it tends to cause overdispersion resulting in relatively poor performance. Although zero padding is employed, it is not accounted for in this analysis as critical values are not estimated. The classification task is performed by assigning a label from $\{1, 2, \ldots, 26\}$ to each test case using the following criterion,

$$t_{ij} = \arg\min_k T[Y_{ij1} - \bar{Z}_k]$$  \hspace{1cm} (6.1)
where, $i,k \in \{1,2,\ldots,26\}$, $j_1 = 1,2,\ldots,30$ representing the 30 replicates in each test case except for letter ‘M’. $Y_{ij_1}$ represent the test profiles, $Y_{ij_1} \in \mathbb{R}^{j_1 \times n}$ and $Z_{j_2k} \in \mathbb{R}^{j_2 \times n}$. Here, $j_2 = 1,2,\ldots,240$ across 4 training groups except letter ‘F’ where $j_2 = 238$. In equation (6.1), $T(\bullet)$ refers to the evaluation of $\kappa_\eta$ on $T[Y_{ij_1} - \bar{Z}_k]$ as a two-treatment test. The results are compared to previously published results (see Table 6.1).

The classification task using $\kappa_\eta$ requires proper selection of parameters like other methods such as MRMI. LDA and PCA do not require such tuning [124]. The average classification rate attained by $\kappa_\eta$ is 97.56% and the average error rate is 2.44%.

The tests here involve classification of fundamental speech signals. Based on these results, it can be inferred that $\kappa_\eta$ may be a suitable candidate for advanced testing in speech signals, recognition and classification.

### 6.2 Applications in Genomics

In genomic analysis related to this work, $\kappa_\eta$ is used as an effective tool in analyzing genomic profiles from various experiments. Most of the experimental data in this work is contributed by
Table 6.1: Classification results for the ISOLET database. Results summarized under 1, 2 and 3 can be found in [124], [42] and [55]. A rigorous summary of unpublished results is also provided in the ISOLET database.

<table>
<thead>
<tr>
<th>Method</th>
<th>% Correct</th>
<th>% Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\kappa\eta)</td>
<td>97.56</td>
<td>2.44</td>
</tr>
<tr>
<td>Opt- Backpropagation(^2)</td>
<td>95.9</td>
<td>4.1</td>
</tr>
<tr>
<td>Opt-Backpropagation(^3)</td>
<td>96.73</td>
<td>3.37</td>
</tr>
<tr>
<td>Modified MRMI(^1)</td>
<td>72.26</td>
<td>17.74</td>
</tr>
<tr>
<td>LDA(^1)</td>
<td>71.21</td>
<td>28.79</td>
</tr>
<tr>
<td>MRMI(^2)</td>
<td>70.56</td>
<td>29.44</td>
</tr>
<tr>
<td>PCA(^1)</td>
<td>59.85</td>
<td>40.15</td>
</tr>
</tbody>
</table>

Dennis et al. and his lab personnel at Florida State University. \(\kappa\eta\) is employed to monitor quality control among genomics profile replicates and in identifying differences in responses with varying effects (such as drugs, disease etc.). Some of the results of \(\kappa\eta\) are discussed in this section.

In genomics, some of the important statistical problems involve identifying segments of significance within a genomic profile, variable selection and classification. These problems are studied broadly with respect to experiments from various areas of life sciences. \(\kappa\eta\) can be employed in experiments involving classification and variable selection, whereas other tools are necessary to isolate significant segments from a profile. In situations where segmentation of genomic profiles are sought, a need for segmentation must be established. This initial step can be accomplished by comparing genomic profiles for differences and \(\kappa\eta\) can be readily used. S.B. Girimurugan et al. proposed a new approach to identify segments in genomic profiles [172] which showed better performance over existing methods [128, 127, 52, 19, 96, 123]. Hence, for analysis pertaining to segmentation, the iSeg [172] algorithm is used in this work. Readers interested in these methods are encouraged to follow the references to gain more valuable information behind their methodologies.

6.2.1 Data Types

Data involved in genomic analyses arise from a variety of experiments, each offering a unique insight into the human genome (see Chapter 1). However, data in genomics can be categorized into two types,
Table 6.2: Summary of p-values for each difference map defined in Equation (6.2) and the drugs that resemble HIV responses ($U_2$)

<table>
<thead>
<tr>
<th>Chromosome</th>
<th>Dimension (n)</th>
<th>Significant Drugs</th>
<th>$\kappa_\eta$ p-values</th>
</tr>
</thead>
<tbody>
<tr>
<td>chr1</td>
<td>10720</td>
<td>{φ}</td>
<td>{7.9e-81,6.3e-134,7.3e-62,5.6e-61,4.7e-92}</td>
</tr>
<tr>
<td>chr2</td>
<td>11440</td>
<td>{GIV, VOR, BEL, VPA, PMA}</td>
<td>{1,1,1,1}</td>
</tr>
<tr>
<td>chr3</td>
<td>9144</td>
<td>{GIV, VOR, BEL, VPA, PMA}</td>
<td>{1,1,1,1}</td>
</tr>
<tr>
<td>chr4</td>
<td>8722</td>
<td>{GIV, VOR, BEL, VPA, PMA}</td>
<td>{1,1,1,1}</td>
</tr>
<tr>
<td>chr5</td>
<td>8312</td>
<td>{GIV, VOR, BEL, VPA, PMA}</td>
<td>{1,1,1,1}</td>
</tr>
<tr>
<td>chr6</td>
<td>7910</td>
<td>{φ}</td>
<td>{7.7e-127,7e-201,5.3e-154,2.4e-126,3.3e-128}</td>
</tr>
<tr>
<td>chr7</td>
<td>7400</td>
<td>{GIV, VOR, BEL, VPA, PMA}</td>
<td>{1,1,1,1,1}</td>
</tr>
<tr>
<td>chr8</td>
<td>6776</td>
<td>{GIV, VOR, BEL, VPA, PMA}</td>
<td>{0.97,0.027,0.98,1.0,0.92}</td>
</tr>
<tr>
<td>chr9</td>
<td>5695</td>
<td>{BEL, VPA, PMA}</td>
<td>{1.6e-07,1.9e-10,1.2e-05,4.3e-05,5.4e-06}</td>
</tr>
<tr>
<td>chr10</td>
<td>6372</td>
<td>{φ}</td>
<td>{8e-259,0.63e-252,1.2e-242,2.2e-273}</td>
</tr>
<tr>
<td>chr11</td>
<td>6144</td>
<td>{GIV, VOR, BEL, VPA, PMA}</td>
<td>{1.099,1.1,1}</td>
</tr>
<tr>
<td>chr12</td>
<td>6106</td>
<td>{GIV, VOR, BEL, VPA, PMA}</td>
<td>{0.013,1.7e-06,0.04,0.26,0.013}</td>
</tr>
<tr>
<td>chr13</td>
<td>4638</td>
<td>{GIV, VOR, BEL, VPA, PMA}</td>
<td>{0.013,0.32,0.52,0.48,0.006}</td>
</tr>
<tr>
<td>chr14</td>
<td>4220</td>
<td>{GIV, VOR, BEL, VPA, PMA}</td>
<td>{1,1,1,1}</td>
</tr>
<tr>
<td>chr15</td>
<td>3838</td>
<td>{GIV, VOR, BEL, VPA, PMA}</td>
<td>{0.57,0.066,0.21,0.71,0.74}</td>
</tr>
<tr>
<td>chr16</td>
<td>3719</td>
<td>{GIV, VOR, BEL, VPA, PMA}</td>
<td>{1,1,1,1}</td>
</tr>
<tr>
<td>chr17</td>
<td>3734</td>
<td>{φ}</td>
<td>{1.3e-08,1.9e-18,1.9e-08,4.8e-07,1e-08}</td>
</tr>
<tr>
<td>chr18</td>
<td>3622</td>
<td>{GIV, VOR, BEL, VPA, PMA}</td>
<td>{1,1,1,1}</td>
</tr>
<tr>
<td>chr19</td>
<td>2573</td>
<td>{GIV, BEL, VPA, PMA}</td>
<td>{0.11,5.1e-11,0.15,0.45,0.42}</td>
</tr>
<tr>
<td>chr20</td>
<td>2869</td>
<td>{φ}</td>
<td>{6.5e-103,2.3e-117,4.2e-115,5.7e-94,8.1e-107}</td>
</tr>
<tr>
<td>chr21</td>
<td>1631</td>
<td>{GIV, VOR, BEL, VPA, PMA}</td>
<td>{0.98,0.83,0.99,1.0,0.98}</td>
</tr>
<tr>
<td>chr22</td>
<td>1667</td>
<td>{GIV, VOR, BEL, VPA, PMA}</td>
<td>{0.89,0.98,0.92,0.92,0.85}</td>
</tr>
<tr>
<td>chrX</td>
<td>6647</td>
<td>{φ}</td>
<td>{1.3e-161,5.1e-165,2.4e-112,7.1e-150,6e-127}</td>
</tr>
<tr>
<td>chrY</td>
<td>1038</td>
<td>{φ}</td>
<td>{1.4e-201,4.9e-152,5.2e-132,7.7e-195,0}</td>
</tr>
</tbody>
</table>

- **Array CGH data**: Comparative Genomic Hybridization (CGH) [93] is a technique used to detect copy number changes using DNA microarrays. This technology is one of the commonly used DNA sequencing techniques. The data obtained from such experiments are log$_2$ – ratio of measurement intensities and generally are normally distributed.

- **NGS data**: Next Generation Sequencing [145, 107] is the most recent DNA sequencing technique. NGS data offers high resolution sequencing data requiring high-throughput analysis. Data obtained from NGS experiments are high dimensional and the data are usually counts. $\kappa_\eta$ is used in the analysis of both types of data. However, additional statistical methodologies such as variance-stabilization are required with NGS data.
6.2.2 Effect of HDAC Inhibitors on HIV Cell Lines

Histone deacetylase (HDAC) Inhibitors (HDIs) are a class of drugs used in the treatment of epilepsy and inflammatory diseases. It is hypothesized that nucleosomal changes caused by the human immunodeficiency virus (HIV) is similar to those caused by HDIs. To study this, a microarray experiment is employed to obtain data from cells with and without HIV. Each of these cells are treated with HDIs: GIVinostat (GIV), VORinostat (VOR), BELinostat (BEL), ValProic Acid (VPA) and Phorbol Myristic Acid (PMA). The untreated (UNT) baseline data is also obtained for both with HIV (Group-II) and without HIV (Group-I) cases. To verify the hypothesis, a wavelet based test is carried out across the entire genome (Chromosomes 1 through 22, Chromosome X and Chromosome Y). A difference map is obtained as,

\[ D_j = Y_{1j} - U_1 \]  \hspace{1cm} (6.2)

and employed in a two sample test involving \( \kappa \) against \( U_2 \). A difference map from Group-I is shown in Figure 6.2.2. The wavelet based test is very flexible in this experiment as testing is done
without any replicate data whereas the Hotelling $T^2$ statistic ($\vartheta$) is dependent on replicate size (at least two replicates in one of the treatments). Here, $U_1$ and $U_2$ are baseline, untreated data from groups 1 (no HIV) and 2 (HIV) respectively. $Y_{ij}$ represent the profiles treated with drugs GIV (1), VOR (2), BEL (3), VPA (4) and PMA (5) in Group-I respectively. Using a $\kappa_\eta$ based test on noisy profiles obtained across the genome, it can be seen from Table 6.2 that Chromosomes 1, 6, 10, 17, 20, X and Y indicate a significant difference between drug treated responses from Group-I and the untreated response from Group-II.

The total number of probes covering the entire genome is 134937. Among these probes, 39290 of them correspond to the chromosomes that didn't show a similarity toward the untreated response in Group-II. Thus, it can be concluded that approximately 70.89% of the assessed genome indicate similarities between HDI treated responses from cells without HIV and responses from HIV cells. From Figures 6.4, 6.5, 6.6 and 6.7, it is evident that $\kappa_\eta$ captures the contained differences in noisy, high dimensional profiles effectively.

6.2.3 Kaposi's Sarcoma Associated Herpes Virus (KSHV) Nucleosome Distribution

The objective of the experiment is to identify the time period demonstrating the most changes in nucleosome redistribution upon KSHV reactivation. The time periods observed are 0, 6, 12, 24
Figure 6.5: Chromosomes 10 and 17 showing smoothed, drug treated responses from Group-I compared with untreated response from Group-II. The dotted line corresponds to Group-II response and the coloured lines correspond to the five drug treated difference maps.

Figure 6.6: Chromosomes 20 and X showing smoothed, drug treated responses from Group-I compared with untreated response from Group-II. The dotted line corresponds to Group-II response and the coloured lines correspond to the five drug treated difference maps.
Figure 6.7: Chromosomes 1 and 6 showing smoothed, drug treated responses from Group-I compared with untreated response from Group-II. The dotted line corresponds to Group-II response and the coloured lines correspond to the five drug treated difference maps. Chromosome 2 indicates the case demonstrating similarity between drug treated responses from Group-I and untreated response from Group-II and 48 hours, where 0 hour corresponds to the initial reactivation time point. This experiment has been studied using data obtained from microarrays [147] and high resolution NGS data. The microarray data studied here is published and publicly available at the Gene Expression Omnibus (GEO) [146].

**Microarray Data.** iSLK.219 cells are used in the analysis and the responses are measured at 0, 6, 12, 24 and 48 hours. The responses are measured for sets of genes on a per chromosome basis over the entire genome. A two-sample $\kappa_\eta$ test is performed using a ‘la8’ filter and a full DWT decomposition proportional to signal length with zero padding. $l_t$ is set at 4 and profiles with a minimum length of 32 are considered. The analysis is carried out as a multiple testing problem on individual genes. A false discovery rate (FDR) controlling procedure [22] is employed to determine the significance cut-off for all multiple comparisons. Figure 6.2.3 shows the individual gene responses at various time-points from two chromosomes (2 and 16). There are 498 genes observed across the entire genome except Chromosome-Y. The two samples in the test are $Y_0$ corresponding to 0 hour and $Y_k, k \in \{6, 12, 24, 48\}$ respectively. The FDR is controlled at 0.05 and the resulting cut-off among all tests is 0.0213. Using this cut-off, the number of genes with a significant change at each time point are summarized in Table 6.3. From Table 6.3, it can be observed that most of the changes (approximately 69.88%) occur during the 24-hour time point.
Figure 6.8: The responses of four genes (two from Chromosome-2 and two from Chromosome-16) from all time points; 0 (dotted, black), 6 (solid, red), 12 (solid, green), 24 (solid, blue) and 48 (solid, cyan) are shown.

Table 6.3: Number of genes with a significant change relative to 0 hour time point using a two-sample $\kappa_\eta$ based test.

<table>
<thead>
<tr>
<th></th>
<th>0-6h</th>
<th>0-12h</th>
<th>0-24h</th>
<th>0-48h</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>95</td>
<td>237</td>
<td>348</td>
<td>182</td>
</tr>
</tbody>
</table>

In this analysis, the profiles are normalized with respect to their standard deviation prior to testing in order to satisfy normality assumptions of $\kappa_\eta$.

**NGS data.** As mentioned earlier, NGS data offers better resolution of the genome and hence a greater insight into genomic behaviour in cells. The KSHV experiment is also carried out on the NGS platform. The genome is measured in terms of fragments (2000 base pairs per fragment) corresponding to regions of interest (usually genes, transcription sites etc.). For this data, the genome was sequenced for 21857 genes. The response is averaged count data corresponding to every sequenced gene per chromosome. Some of the sample responses are shown in Figure 6.10.

The data is comprised of averaged counts and the normality is ensured with the use of variance-stabilization techniques [65, 113] such as the square root transformation [20, 74], the logarithmic...
Figure 6.9: Responses from the KSHV experiment at the time points 0 (dotted, black), 6 (solid, red), 12 (solid, green), 24 (solid, blue) and 48 (solid, cyan) hours. The responses here are obtained from a next generation sequencing platform and the amplitude corresponds to counts (as represented in NGS type data).

Table 6.4: Summary of top 10 genes demonstrating significant changes at the 24 hour time point. The genes are sorted according to their p-values obtained using a $\kappa_\eta$ based test.

<table>
<thead>
<tr>
<th>UCSC-ID</th>
<th>Chromosome</th>
<th>Gene Symbol</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>uc001dut.3</td>
<td>chr1</td>
<td>AMY2A</td>
<td>6.017995e-177</td>
</tr>
<tr>
<td>uc010dfa.1</td>
<td>chr17</td>
<td>ABCA10</td>
<td>5.246727e-125</td>
</tr>
<tr>
<td>uc001phg.2</td>
<td>chr11</td>
<td>MMP10</td>
<td>1.708354e-112</td>
</tr>
<tr>
<td>uc001gtj.4</td>
<td>chr1</td>
<td>CFH</td>
<td>4.746537e-110</td>
</tr>
<tr>
<td>uc022aop.1</td>
<td>chr7</td>
<td>MIR548F4</td>
<td>5.387554e-106</td>
</tr>
<tr>
<td>uc009xgp.3</td>
<td>chr1</td>
<td>KMO</td>
<td>3.322961e-101</td>
</tr>
<tr>
<td>uc001gho.3</td>
<td>chr1</td>
<td>FMO4</td>
<td>9.25444e-98</td>
</tr>
<tr>
<td>uc002jio.4</td>
<td>chr17</td>
<td>KCNJ16</td>
<td>1.22244e-94</td>
</tr>
<tr>
<td>uc002jal.4</td>
<td>chr17</td>
<td>TANC2</td>
<td>1.494158e-86</td>
</tr>
<tr>
<td>uc003yez.3</td>
<td>chr8</td>
<td>SLC26A7</td>
<td>6.679895e-76</td>
</tr>
</tbody>
</table>

transformation or the Anscombe transformation [12]. In this analysis, the square root transformation is employed. However, a suitable alternative is the Anscombe transformation. A two-sample $\kappa_\eta$ based test is carried out with $l_t = 4$ and a la8 filter. A full DWT decomposition is carried out with respect to profile length including zero padding.
Table 6.5: Number of genes with a significant change relative to 0 hour time point using a two-sample $\kappa_\eta$ based test using NGS data for the KSHV experiment.

<table>
<thead>
<tr>
<th></th>
<th>0-6h</th>
<th>0-12h</th>
<th>0-24h</th>
<th>0-48h</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2</td>
<td>37</td>
<td>3731</td>
<td>196</td>
</tr>
</tbody>
</table>

The variance introduced by count type data is handled by a square root transformation. The $p$-values are evaluated using a Chi-squared approximation to $\kappa_\eta$. A significance cut-off is established using the Benjamini-Hochberg procedure for an FDR of 0.05. This cut-off is found to be 0.0503 for this analysis. It can be seen from Table 6.5 that most of the changes occur during the 24 hour time point, a conclusion similar to the microarray based analysis (see Table 6.3). The number of genes changed at the 24 hour time point relative to the total gene count is less than the previous estimate obtained using the microarray analysis. However, from a statistical standpoint, the chances of observing a count in Table 6.5 is more significant the count in Table 6.3.

The topmost four significant genes from the 24 hour time point are plotted in Figures ??, ?? and C.1 along with results from a post hoc segmentation that isolates specific regions of significance. It is apparent that most changes are detected effectively by $\kappa_\eta$ and the segments of significance are obtained with an iSeg [172] implementation. A short summary about iSeg is given in Section B.2. The results of segmentation with the responses are shown in Appendix C. The top 10 genes with significant changes are shown in Table 6.4.

6.2.4 Quality Control in Genomic Profiles

In this section, the use of $\kappa_\eta$ to establish quality control in genomic profiles is demonstrated. In this analysis, genetic responses from cells subjected to low and high degrees of digestion (DoD) are studied.

Two different cell lines are used in the experiment: GM12878 (GM) and IMR. Each cell line has two biological replicates, GM7, GM11, IMR15 and IMR18. It is vital for the biological replicates to be identical so that the conclusions are consistent. In this analysis, $\kappa_\eta$ is used to determine the quality of these two replicates. The analysis is carried out on four different profile types from
Table 6.6: Summary of gene profiles that differ among biological replicates. The cell lines GM and IMR are included in the analysis for all degrees of digestion.

<table>
<thead>
<tr>
<th>Cut-off</th>
<th>H0</th>
<th>L0</th>
<th>H140</th>
<th>L140</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>GM</td>
<td>IMR</td>
<td>GM</td>
<td>IMR</td>
</tr>
<tr>
<td><strong>α</strong></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td><strong>α_{FDR}</strong></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

These cell lines, High (H) and Low (L) degrees of digestions measured with profiles centered along 0-140bp and 0-200bp. The 0-140bp profiles are denoted by H140 and L140 while the 0-200bp profiles are denoted by H0 and L0. The comparison of biological replicates for each cell line is shown in Figures 6.10 and 6.11.

In the analyses, 18458 genes are studied per experiment and each profile is of length 200. The data corresponds to fragmented count data obtained using NGS. In Table 6.6, the number of genes that differ significantly among replicates are summarized. The traditional cut-off of $\alpha = 0.05$ is used in addition to a FDR controlled cut-off $\alpha_{FDR} = 5.93 \times 10^{-7}$ evaluated for an FDR of 0.1. It can be seen at level $\alpha$ that several profiles are found to be significantly different in H140. Due to the large number of genes, an FDR correction is employed and based on a cut-off corrected for a specified FDR (FDR=0.1). The profiles did not differ significantly across biological replicates. Therefore, based on a $\kappa_\eta$ test for this experiment, the biological replicates seem to be in-control and pass quality control for further analysis.

6.2.5 Cancer Data Analysis

In this analysis, data from cancer patients at various grades (Grades I and III) are analyzed. The patient identification numbers are 386, 873, 1357 and 4137. Patients 1357 and 4137 correspond to Grade I cancer and patients 386 and 873 correspond to Grade-III.

The experiment is carried out on an NGS platform and 21857 genes are observed per sample. Each gene profile corresponds to fragmented, count data and has a length of 200. This experiment investigates nucleosomal changes at various grades of cancer. First, a comparison is carried out between Normal (N) and Tumor (T) profiles from Grade I and Grade III cancer cells. A $\kappa_\eta$ based test is employed per gene, for all 21857 genes. A few profiles from this experiment are shown in
(a) Comparison of biological replicates GM7 (top, red) with GM15 (top, black) and IMR15 (bottom, red) with IMR18 (bottom, black) for gene TESK1.

(b) Comparison of biological replicates GM7 (top, red) with GM15 (top, black) and IMR15 (bottom, red) with IMR18 (bottom, black) for gene SAMD11.

Figure 6.10: Comparison of biological replicates from GM and IMR cell lines in the DoD experiment with NGS type data for genes TESK1 and SAMD11.
(a) Comparison of biological replicates GM7 (top, red) with GM15 (top, black) and IMR15 (bottom, red) with IMR18 (bottom, black) for gene CPT1C.

(b) Comparison of biological replicates GM7 (top, red) with GM15 (top, black) and IMR15 (bottom, red) with IMR18 (bottom, black) for gene FNDC4.

Figure 6.11: Comparison of biological replicates from GM and IMR cell lines in the DoD experiment with NGS type data for genes CPT1C and FNDC4.
Table 6.7: Number of genes with a significant change in Grade I (Patients 1357 and 4137) and Grade III (Patients 386 and 873), using a two-sample $\kappa_\eta$ based test with $\alpha = 0.05$.

<table>
<thead>
<tr>
<th></th>
<th>1357</th>
<th>4137</th>
<th>386</th>
<th>873</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>399</td>
<td>5947</td>
<td>75</td>
<td>307</td>
</tr>
</tbody>
</table>

Figures 6.12 and 6.13. Using $\kappa_\eta$ and $\alpha = 0.05$, the number of changed genes are summarized in Table 6.7. Among the genes a common list of genes that demonstrate changes in Grade I and Grade III are obtained (see Chapter C for the lists). It can be noted that 125 genes demonstrate changes in Grade I and only 10 genes demonstrate changes in Grade III. Among the genes that demonstrated change in Grade I, the genes that showed a similar change in cancer profiles are observed. A $\kappa_\eta$ based test is carried out to identify a set of genes showing identical changes in the tumour profiles of Grade I patients, 1357 and 4137. This set of genes are summarized in C.2. The genes that demonstrate similar changes in Grade I (see Table C.2) are analyzed further in Grade III. It is found that none of the genes that demonstrated a change in Grade I showed changes in Grade III.

6.3 Applications of $\kappa_\eta$ in Quantitative Finance

In this section, $\kappa_\eta$ is used to detect differences in an Ornstein-Uhlenbeck process, a mean-reverting process that is a special case of a Gaussian process [136].

Let $\{X_t\}$ denote a Gaussian process with a covariance function $\kappa_{t,s}$. A discrete version of $\{X_t\}$, $X_n$, involves the sampling of $\{X_t\}$ at equidistant intervals, usually, $\delta t = \frac{t}{n}$. Thus, $X_n$ is an infinite dimensional multivariate Gaussian with zero mean and covariance specified by $\kappa_{i,j} \in \mathbb{R}^{n \times n}$. $\{X_t\}$ is an Ornstein-Uhlenbeck process if the covariance function is defined as,

$$\kappa_{t,s} = e^{-\frac{|t-s|}{l}}$$  \hspace{1cm} (6.3)

where $l$ is the length-scale parameter of the process. In this analysis, $l = 1$ in all simulations. The effect of $l$ on power and Type-I error would be analyzed in future work on this topic.
Figure 6.12: Comparison of Normal (N) and Tumour (T) profiles in Grade I (1357, 4137) and Grade III (386, 873) patients for genes EID2 and USP51.
Figure 6.13: Comparison of Normal (N) and Tumour (T) profiles in Grade I (1357, 4137) and Grade III (386, 873) patients for genes LOC643837 and TMEM254-AS1.
Any phenomenon described by an Ornstein-Uhlenbeck process can be modelled by estimating the parameters describing the process. Such methods involve stochastic calculus and several approaches have been developed in the past [158]. Here, assume that a set of sample paths from an experiment is available and the underlying mean response is given by, $f_0$. That is,

$$X_n \sim GP(f_0, \kappa)$$ \hspace{1cm} (6.4)

If the observed number of such sample paths is large, then the mean of sample paths is a good estimator for $f_0$. Otherwise, the mean of a few sample paths may not describe the underlying mean response properly. For instance, in Figure 6.14, three sample paths of an Ornstein-Uhlenbeck process are shown. The underlying functional response that created these sample paths is $f_0$, defined by,

$$f_0 = 0.51 \frac{\sin(12x)}{\cos(x)}$$ \hspace{1cm} (6.5)

From Figure 6.15, it can be observed with very few replicates ($r=10$) that the underlying mean response is estimated, but not accurately. With a larger set of replicates ($r=1000$), the estimation
improves significantly. In general, with a good signal to noise ratio or a larger set of replicates, the mean of observed sample paths approximates the underlying functional response with greater accuracy. $\kappa_\eta$ seeks to identify differences in the underlying function with a small set of sample paths and when the signal strength is relatively weak.

Previous work on statistical tests for Ornstein-Uhlenbeck processes involved testing for the parameters [95] and tests for stability of such processes [110]. The proposed test aims at testing for differences in the underlying mean response of several sample paths generated by an Ornstein-Uhlenbeck process from $K$ groups. Given the effectiveness of $\kappa_\eta$ in detecting differences among high dimensional profiles, an analysis of stock prices that are usually modelled by Ornstein-Uhlenbeck processes; more precisely, geometric brownian motion (GBM) is performed using $\kappa_\eta$. Stock price data obtained through Yahoo Finance [85] is used in this analysis. Historical stock prices from four different firms are analyzed. Daily stock price data for Advanced Micro Devices (AMD), Intel Corporation (INTC), Qualcomm (QCOM) and Texas Instruments (TXN) are obtained from 1983 to 2013 (current). For QCOM, data was available from 1991 onward. The objective is to determine differences in the historical stock prices between these firms. Also, one can determine if the average stock prices differed from one decade to another. For the observed stock prices, $S_t$ is
usually a geometric brownian motion and \( X_t = \log(S_t) \), is an Ornstein-Uhlenbeck process. Thus, there are four groups and several replicates for each group. Since the lengths of stock prices are not dyadic (in powers of 2), zero padding is carried out. It should be noted that the parameters can be adjusted to account for zero padding while obtaining critical values. The adjusted closing prices in 2000 for the four groups are shown in Figure 6.3. The profiles are not very long with these data, but \( \kappa \eta \) tends to perform better with longer profiles as shown in previous analyses [76, 77].

The number of sample paths for each group is \( r_{AMD} = 31, r_{INTC} = 28, r_{QCOM} = 23 \) and \( r_{TXN} = 31 \), thus leading to an unbalanced, multivariate test. The profile lengths are adjusted to 256 to ensure that all profiles across all groups have identical lengths. The mean response for each group is shown in Figure 6.3. The data is used in three different ways with a normality assumption. First, the data is used as is with an assumption of unit variance for the mean response after normalizing each profile by its standard deviation. This is called the raw approach.

In the second scenario, the standard deviation is estimated for the mean response and used for \( \hat{\sigma} \). This latter approach is called the normalization approach. The maximum level of decomposition is set to 8 with \( J_t = 4 \). With the raw approach, due to normality, the universal, hard threshold is employed. A manual, hard threshold is applied where \( \lambda = \hat{\sigma} \sqrt{2 \log(n)} \) and \( \hat{\sigma}^2 = \sum_{i=1}^{4} \hat{\sigma}_i^2/4 \) for
Figure 6.17: Average Adjusted Closing prices obtained using daily average stock price sample paths from 1983 to 2013 (August 21, 2013)

Table 6.8: $\kappa_\eta$, p-value and the threshold ($\lambda$) used in the raw and normalization approaches of $\kappa_\eta$ with stock price data. The p-values are obtained using the normal approximation to $\kappa_\eta$ and the exact p-values are shown within parentheses.

<table>
<thead>
<tr>
<th></th>
<th>Raw</th>
<th>Normalization</th>
<th>Direct</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\kappa_\eta$</td>
<td>781.33</td>
<td>634.64</td>
<td>641.35</td>
</tr>
<tr>
<td>p-value</td>
<td>$1.892017e-144$</td>
<td>$5.801e-93$</td>
<td>$4.557e-94$</td>
</tr>
<tr>
<td>$\lambda$</td>
<td>3.302</td>
<td>0.9107</td>
<td>3.302</td>
</tr>
</tbody>
</table>

the normalization approach. Lastly, a direct approach uses estimation of the noise parameter and $\hat{\gamma}$ (see Chapter 4) is evaluated using which $\kappa_\eta$ is calculated. The normal approximation to $\kappa_\eta$ is used to find the p-values. The la8 filter is used in obtaining the DWT. The results in testing for differences in the mean responses among the four groups is summarized in Table 6.8. It can be seen that all approaches rejected the $H_0$ at a significance level of 0.05. Now that $H_0$ is rejected, a pairwise comparison test would reveal groups that are identical (if any). To perform a pairwise comparison, a two sample $\kappa_\eta$ test is employed. The p-values are obtained using the direct approach and the results is summarized in Table 6.9.

Since $\kappa_\eta$ can be used as a classifier (see ISOLET application), unsupervised learning of the four firms using the p-values can be obtained using the raw and normalization approaches. A significance level of 0.05 for the raw approach and the normalization approach is used. Thus, the firms
Table 6.9: Summary of p-values for the pairwise comparison procedure with $\kappa_\eta$ using the direct approach.

<table>
<thead>
<tr>
<th></th>
<th>TXN</th>
<th>INTC</th>
<th>QCOM</th>
</tr>
</thead>
<tbody>
<tr>
<td>AMD</td>
<td>1.153e-32</td>
<td>4.691e-90</td>
<td>7.690e-09</td>
</tr>
<tr>
<td>TXN</td>
<td></td>
<td>0.4652</td>
<td>7.788e-35</td>
</tr>
<tr>
<td>INTC</td>
<td></td>
<td></td>
<td>3.389e-81</td>
</tr>
</tbody>
</table>

can be readily classified into two groups. It can be deduced that AMD and QCOM form separate classes from TXN and INTC.
CHAPTER 7

CONCLUSIONS AND FUTURE WORK

In this work, a non-parametric test using wavelet transforms is proposed. The sparsity of wavelet transforms combined with its denoising and whitening properties makes the proposed test an effective methodology to analyze high-dimensional profiles. Since an underlying model is not assumed in this work, the results can be extended to a variety of situations where a parametric model is not applicable.

Here, the statistic and its mathematical framework have been developed under the assumption of normality. In many cases, a Generalized Gaussian distribution is used for modeling wavelet coefficients [57]. Some authors have employed mixture models to model such coefficients [7], however such approaches are restrictive when data is not Gaussian. In this work, a well defined statistical model is offered for the modeling of wavelet coefficients, albeit under the assumptions of normality. Moreover, the current work focuses primarily on a statistical framework when hard thresholding is employed. Hence, an extension of the framework would aim at obtaining such a distribution under varying types of threshold.

Many circumstances (including genomics, consider NGS data) do not offer flexibility with assumptions of normality and appropriate measures are required to establish consistent analyses. Thus, it becomes imperative to explore the behaviour of the coefficients of a Discrete Wavelet Transform (DWT) for data from any distribution in the exponential family. Several authors have developed interesting methods to address this issue [87, 15] and the foundations developed in this work will focus on improving and developing solutions to this problem. The effectiveness of the the proposed test is due to the desirable properties of DWT. However, as explained by Coifman et al. [41], a best basis algorithm is capable of representing a signal (response) effectively. Thus, a better basis than a standard DWT basis could result in optimal testing with respect to a given entropy function. Wavelet Packet Transforms (WPT) currently aim at offering best basis selection with respect to a given cost function. Hence, $\kappa_\eta$ based tests can be extended to using WPT instead of DWT to offer tests flexible to a given application with a specified cost function.
The current work considers high-dimensional profiles (or functional responses). However, high-dimensional data occur in more complex cases such as images, shapes and 3D structures. Wavelet transforms have been used in the analysis of shapes [173] and more generally, manifolds [143, 13, 49, 115]; images [105], in network topology and spatial statistics [47]. This work can be extended to testing and could result in efficient data analysis in such areas.

In analyzing stochastic processes, the current work considered the Ornstein-Uhlenbeck process as it is a non-stationary process and the data structure is not heavily correlated. However, several other stochastic processes can be analyzed with \( \kappa_{\eta} \) with necessary modification. Such processes include, but are not limited to, the Wiener process, the fractional Brownian motion and the Brownian bridge. Past research that incorporate wavelet transforms and such processes focused on the interpretation of the underlying correlation structure [154, 106, 68] and further analysis on \( \kappa_{\eta} \) would exploit such ideas to facilitate testing with such processes.

Bayesian statistics is very crucial to analysis with longitudinal data and biostatistics. The current implementation of \( \kappa_{\eta} \) did not employ any Bayesian fundamentals in its framework. However, Bayesian techniques are not new to wavelet transforms. In fact, one of the thresholding schemes used in wavelet based statistical analysis has its roots in Bayesian principles [61, 62]. Other methods in Bayesian analysis involving wavelet transforms deal with image and signal processing [121, 137, 23]. With past research and developments in Bayesian methods involving wavelet transforms, many other advancements are possible in the area of non-parametric testing with \( \kappa_{\eta} \).

A final area of research incorporating \( \kappa_{\eta} \) will be in variable selection. Previous works on this topic involved wavelets and Bayesian analysis [148, 56] in semi-parametric and parametric modeling [10] situations. \( \kappa_{\eta} \) can be used in variable selection, especially with high-dimensional data, primarily in a non-parametric setting. In summary, there are multiple areas of research that can be advanced and improved with \( \kappa_{\eta} \). Some of these areas could be computationally more intensive than others and computational efficiency is up to par in the current implementation of \( \kappa_{\eta} \). However, modifications to \( \kappa_{\eta} \) may be required to accommodate newer topics of research with respect to statistical accuracy and computational efficiency.
A.1 \( T^2 \) Unbalanced Design

In this chapter, a derivation of the unbalanced design of a Hotelling \( T^2 \) test statistic is provided.

A.1.1 Notation

\[ Y_{ijk} = k^{th} \text{ response of the } j^{th} \text{ replicate for the } i^{th} \text{ treatment} \]

\[ Y_{ij} \in \mathbb{R}^n \text{ and } \overline{Y}_{i,k} = \frac{1}{r_j} \sum_{k=1}^{r_j} Y_{ikn} \in \mathbb{R}^n \]

Assume, \( Y_{ij} \sim N_n(f_i, \sigma^2 I) \in \mathbb{R}^n \), where \( f_i \) is the mean response of the \( i^{th} \) treatment. \( \Rightarrow \overline{Y}_{i,k} \sim N_n \left( f_i, \frac{\sigma^2}{r_i} I \right) \), where \( I \) is the identity matrix.

A.1.2 Two-treatment Design

\[ \overline{Y}_{1,k} \sim N_n \left( f_1, \frac{\sigma^2}{r_1} I \right) \quad \text{(A.1)} \]

\[ \overline{Y}_{2,k} \sim N_n \left( f_2, \frac{\sigma^2}{r_2} I \right) \quad \text{(A.2)} \]

\[ \Rightarrow \overline{Y}_{1,k} - \overline{Y}_{2,k} \sim N_n \left( f_1 - f_2, \sigma^2 \left( \frac{1}{r_1} + \frac{1}{r_2} \right) I \right) \quad \text{(A.3)} \]

\[ \frac{\overline{Y}_{1,k} - \overline{Y}_{2,k}}{\sigma \sqrt{\left( \frac{1}{r_1} + \frac{1}{r_2} \right)}} \sim N_n(f_1 - f_2, I) \Rightarrow \frac{(\overline{Y}_{1,k} - \overline{Y}_{2,k})^2}{\sigma^2 \left( \frac{1}{r_1} + \frac{1}{r_2} \right)} \sim \chi^2_1 \quad \text{(A.4)} \]
\[ \sum_{k=1}^{n} \left( Y_{1,k} - Y_{2,k} \right)^2 \over \sigma^2 \left( \frac{1}{r_1} + \frac{1}{r_2} \right) \sim \chi^2_n \tag{A.5} \]

The SSE can be derived as,

\[ SSE = \sum_{i=1}^{2} \sum_{j=1}^{r_i} \sum_{k=1}^{n} \left( Y_{ij}(x_k) - \bar{Y}_{i,i}(x_k) \right)^2 \Rightarrow MSE = \frac{SSE}{n(r_1 + r_2 - 2)} \tag{A.6} \]

For the \( k^{th} \) response, define

\[ (r_1 - 1)s^2_{1k} = \sum_{j=1}^{r_1} \left( Y_{ij}(x_k) - \bar{Y}_{1,i}(x_k) \right)^2 \sim \chi^2_{r_1-1} \tag{A.7} \]

Similarly,

\[ (r_2 - 1)s^2_{2k} = \sum_{j=1}^{r_2} \left( Y_{ij}(x_k) - \bar{Y}_{2,i}(x_k) \right)^2 \sim \chi^2_{r_2-1} \tag{A.8} \]

Hence,

\[ \frac{(r_1 - 1)s^2_{1k}}{\sigma^2} + \frac{(r_2 - 1)s^2_{2k}}{\sigma^2} \sim \chi^2_{(r_1+r_2-2)} \tag{A.9} \]

\[ \Rightarrow \sum_{i=1}^{n} \sum_{j=1}^{r_i} \sum_{k=1}^{n} \left( Y_{ij}(x_k) - \bar{Y}_{i,i}(x_k) \right)^2 \sim \chi^2_{n(r_1+r_2-2)} \tag{A.10} \]

Thus, the test statistic \( T^2 \), under the null hypothesis is,

\[ T^2 = \frac{\sum_{k=1}^{n} \left( Y_{1,k} - Y_{2,k} \right)^2}{MSE} \sim \frac{\frac{1}{n} \chi^2_n}{\chi^2_{n(r_1+r_2-2)}} \Rightarrow T^2 \sim F(n, n(r_1 + r_2 - 2)) \tag{A.11} \]

### A.1.3 Multiple Treatment Design

For the design defined in equation (1.3), a decomposition can be obtained as follows,

\[ Y_{ij}(x_k) - \bar{Y}_{..}(x_k) = Y_{ij}(x_k) - \bar{Y}(x_k) + \bar{Y}_{i..}(x_k) - \bar{Y}_{..}(x_k) \tag{A.12} \]

\[ = Y_{ijk} - \bar{Y}_{1,k} + \bar{Y}_{1,k} - \bar{Y}_{..k} \tag{A.13} \]
The decomposition of variance (within and between) treatments can be obtained as,

\[ \sum_{i=1}^{t} \sum_{j=1}^{r_i} \sum_{k=1}^{n} (Y_{ijk} - Y_{..})^2 = \sum_{i=1}^{t} \sum_{k=1}^{n} (Y_{i,k} - Y_{..})^2 + \sum_{i=1}^{t} \sum_{j=1}^{r_i} \sum_{k=1}^{n} (Y_{ijk} - Y_{i,k})^2 \] (A.14)

\[ SST = SSTr + SSE \] (A.15)

where SST is the total sum of squares, SSTr is the sum of squares corresponding to the treatments and SSE is the sum of squares corresponding to the error. Under the assumptions of normality,

\[ \sum_{i=1}^{t} \sum_{k=1}^{n} (Y_{i,k} - Y_{..})^2 \sim \chi^2_{n(t-1)} \] (A.16)

\[ \sum_{i=1}^{t} \sum_{j=1}^{r_i} \sum_{k=1}^{n} (Y_{ijk} - Y_{i,k})^2 \sim \chi^2_{n(\varphi-t)} \] (A.17)

where \( \varphi = \sum_{i=1}^{t} r_i \).

Thus, the test statistic \( \vartheta \) is given by,

\[ \vartheta = \frac{\sum_{k=1}^{n}(Y_{i,k} - Y_{..})^T (Y_{i,k} - Y_{..})}{n(t-1)} \]

\[ = \frac{(\varphi - t) \sum_{k=1}^{n}(Y_{i,k} - Y_{..})^T (Y_{i,k} - Y_{..})}{(t-1) \sum_{i=1}^{t} \sum_{j=1}^{r_i}(Y_{ij} - Y_{i.})^T (Y_{ij} - Y_{i.})} \] (A.18)

\[ \equiv \frac{(\varphi - t) \sum_{k=1}^{n}(Y_{i,k} - Y_{..})^T (Y_{i,k} - Y_{..})}{(t-1) \sum_{i=1}^{t} \sum_{j=1}^{r_i}(Y_{ij} - Y_{i.})^T (Y_{ij} - Y_{i.})} \] (A.19)

A.1.4 Wavelet Statistic, \( \kappa_\eta \)

Since,

\[ \| Y_{ijk} - Y_{..} \|_2^2 \geq \mathbb{V}[\| Y_{ijk} - Y_{..} \|_2^2] \] (A.20)

\[ \equiv \)
the DWT of \( \vartheta \) is given by,

\[
\mathbb{W}[\vartheta] = \mathbb{W} \left[ \sum_{k=1}^{n} (\bar{Y}_{i,k} - \bar{Y}...)^T (\bar{Y}_{i,k} - \bar{Y}...) \right] \\
\quad \quad = \mathbb{W} \left[ \frac{( \varphi - t ) \sum_{k=1}^{n} (\bar{Y}_{i,k} - \bar{Y}...)^T (\bar{Y}_{i,k} - \bar{Y}...) }{n(\varphi - t)} \right] \\
\quad \quad = \mathbb{W} \left[ \frac{(t - 1)^{-1} \sum_{k=1}^{n} (\bar{Y}_{i,k} - \bar{Y}...)^T (\bar{Y}_{i,k} - \bar{Y}...)}{\hat{\sigma}^2} \right] \\
\quad \quad = \mathbb{W} \left[ (t - 1)^{-1} (\bar{Y}_{i} - \bar{Y})' (\bar{Y}_{i} - \bar{Y}) \right] \\
\quad \quad < \mathbb{W} \left[ \| \bar{Y}_{i} - \bar{Y} \|_2^2 \right] = \| \theta_{ik} \|_2^2 \\
\quad \quad \leq \| \bar{Y}_{i} - \bar{Y} \|_2^2 \\
\quad \quad = \| \tilde{\theta}_{ik} \|_2^2
\]

(A.22)

(A.23)

(A.24)

(A.25)

(A.26)

(A.27)

(A.28)

(A.29)

where \( \bar{Y}_{i} - \bar{Y} = \gamma^{-1}(\bar{Y}_{i} - \bar{Y}) \), \( \theta_{ik} \) form the set of wavelet coefficients \((k = 1, 2, \ldots \leq n)\) corresponding to the \( i \)th treatment. Since, wavelet transforms maintain the statistical properties of signal, an appropriate shrinkage scheme can be used to discriminate coefficients between the true functional response and noise resulting in \( \tilde{\theta}_{ik} \). The denominator of \( \vartheta \) estimates the variance and it can be replaced by \( \hat{\gamma} \), any relevant function of the estimator for noise. Also,

\[
\| \tilde{\theta}_{ik} \|_2^2 \leq \| \theta_{ik} \|_2^2
\]

(A.30)

and so, \( \| \tilde{\theta}_{ik} \|_2^2 \) is a very good approximation for \( \| \bar{Y}_{i} - \bar{Y} \|_2^2 \) under noisy conditions. Thus,

\[
\kappa_{\eta}^* = \| \tilde{\theta}_{ik} \|_2^2 \\
\quad = \sum_{j=1}^{k} \tilde{\theta}_{ik}^2
\]

(A.31)

(A.32)
will be a reasonable statistic to test differences within one treatment and using a similar approach, intuitively, the statistic for multiple treatments is,

\[ \kappa_{ij} = \sum_{i=1}^{t} \sum_{j=1}^{k} \theta_{ik}^2 \]  

(A.33)

where ‘t’ is the number of treatments.

Q.E.D

The estimate for \( \hat{\gamma} \) in A.29 can be expressed as follows. Consider,

\[ E[\bar{Y}_i - \bar{Y}_.] = 0 \]  

(A.34)

then,

\[ Var[\bar{Y}_i - \bar{Y}_.] = E[(\bar{Y}_i - \bar{Y}_.)^2] \]

(A.35)

\[ = Var[\bar{Y}_i] + Var[\bar{Y}_.] - 2Cov(\bar{Y}_i, \bar{Y}_.) \]

(A.36)

\[ = \sigma_i^2 + \frac{1}{t^2} \sum_{i=1}^{t} \sigma_i^2 - 2Cov(\bar{Y}_i, \frac{1}{t} \sum_{j=1}^{t} \bar{Y}_j) \]

(A.37)

\[ = \sigma_i^2 + \frac{1}{t^2} \sum_{i=1}^{t} \sigma_i^2 - \frac{2}{t} \left( \sigma_i^2 + \sum_{j=1}^{t} Cov(\bar{Y}_i, \bar{Y}_j) \right), \quad \text{where } i \neq j \]

(A.38)

\[ = \sigma_i^2(1 - \frac{2}{t}) + \frac{1}{t^2} \sum_{i=1}^{t} \sigma_i^2 - \frac{2}{t} \left( \sum_{j=1}^{t} Cov(\bar{Y}_i, \bar{Y}_j) \right), \quad \text{where } i \neq j \]

(A.39)

In general, if the total variance is \( \sigma^2 \), the variance of the \( i^{th} \) treatment with \( r_i \) replicates is \( \frac{1}{r_i} \sigma^2 \), thus,

\[ \hat{\gamma} = \sigma^2 \left[ \frac{1}{r_i} \left( \frac{(t-2)}{t} \right) + \frac{1}{t^2} \sum_{i=1}^{t} \frac{1}{r_i} \right] - \frac{2}{t} \rho_{ij} \]

(A.41)

where, for \( i \neq j \),

\[ \rho_{ij} = Cov(\bar{Y}_i, \bar{Y}_j) \]

(A.42)
and,
\[
\hat{\sigma}^2 = \text{Var}(Y_{ijk} - \bar{Y}_{i,k}) \tag{A.43}
\]

**A.1.5 Density and MGF of a $\chi^2_{\lambda}$ Distribution**

Consider a truncated normal random variable with mean ‘0’ and standard deviation ‘1’, then its density is given by,
\[
f_X(x; 0, 1, \lambda) = \frac{\phi(x)}{2\Phi(-\lambda)} \mathbb{I}_{(-\infty,-\lambda]} + \frac{\phi(x)}{2(1 - \Phi(\lambda))} \mathbb{I}_{[\lambda,\infty)} \tag{A.44}
\]
where $\mathbb{I}(\bullet)$ is the indicator function, $\Phi(\bullet)$ and $\phi(\bullet)$ are the cumulative distribution function and the density function of a standard normal random variable respectively. That is,
\[
\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \tag{A.45}
\]
\[
\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt \tag{A.46}
\]

Define, $C = X^2$ which is a $\chi^2_{\lambda}$ random variable. Then, using random variable transformations [31], the density of ‘C’ can be obtained as,
\[
f_C(c) = \frac{1}{2\sqrt{c}} \left[ f_X(\sqrt{c}) + f_X(-\sqrt{c}) \right] \tag{A.47}
\]
Since, $c \in [\lambda^2, \infty)$,
\[
f_C(c) = \frac{1}{\sqrt{c}} \left[ f_X(\sqrt{c}) \right] \tag{A.48}
\]
\[
= \frac{1}{2\sqrt{c}} \frac{\phi(\sqrt{c})}{(1 - \Phi(\lambda))} \tag{A.49}
\]
Therefore,
\[
f_C(c) = \frac{1}{2\sqrt{2\pi} \sqrt{c} (1 - \Phi(\lambda))} e^{-c/2} \tag{A.50}
\]
\[
= \frac{1}{2\sqrt{2\pi} (1 - \Phi(\lambda))} e^{-c/2} \mathbb{I}_{[\lambda^2,\infty)} \tag{A.51}
\]
The moment generating function (MGF) can be obtained for $\chi^2_1$ using the density obtained in (A.51). Using the definition of a moment generating function, the MGF of $C$ can be obtained as,

$$M_C(t) = E[e^{Ct}]$$ (A.52)

$$= \int_{\chi^2} e^{ct} \frac{1}{2\sqrt{2\pi}} \frac{c^{-1/2}e^{-c/2}}{(1 - \Phi(\lambda))} dc$$ (A.53)

$$= \int_{\chi^2} \frac{1}{2\sqrt{2\pi}} \frac{c^{-1/2}e^{-c(1-2t)/2}}{(1 - \Phi(\lambda))} dc$$ (A.54)

Now, using a change of variable $u = (1 - 2t)c$, equation becomes,

$$M_C(t) = \int_{(1-2t)\lambda^2}^{\infty} \frac{1}{2\sqrt{2\pi}} \frac{u^{-1/2}e^{-u/2}}{(1 - 2t)(1 - \Phi(\lambda))} du$$ (A.56)

and with integration (using (A.51)), $M_C(t)$ can be derived as $\sqrt{1 - 2t}\mathbb{1}_{t \leq 0}$

Q.E.D

A.1.6 Density and MGF of a $\chi^2_M$ Distribution

Let,

$$C_M = \sum_{i=1}^{M} X_i^2$$ (A.58)

$$= \sum_{i=1}^{M} X_i^2$$ (A.59)

where, $supp(C_M) \in [M\lambda^2, \infty)$ and $X_i$ are truncated normal random variable as defined in (A.44), then,

$$F_{C_M}(c) = \Pr(C_M \leq c)$$ (A.60)

$$= \int_{X_1} \int_{X_2} \ldots \int_{X_M} \prod_{i=1}^{M} f_{X_i}(x_i) dx_i$$ (A.61)
Using (A.44),

\[
F_{CM}(c) = \frac{1}{\sqrt{[2(2\pi(1 - \Phi(\lambda)))]^M}} \int_{X_1} \int_{X_2} \ldots \int_{X_M} e^{-\sum_{i=1}^{M} \frac{x_i^2}{2}} \prod_{i=1}^{M} dx_i \quad (A.62)
\]

Geometrically, one can interpret \( C_M \) as \( \sqrt{R} \), where \( R \) is the radius of an \((M - 1)\)-sphere (the surface of the \(M\)-ball)\([79]\) centered at \( \emptyset \in \mathbb{R}^M \), the origin.

Define, \( \xi = (2\sqrt{2\pi(1 - \Phi(\lambda))})^{-1} \), then the integral in (A.62) can be expressed in terms of the surface area of a \((M - 1)\) sphere, \( A_M \),

\[
F_{CM}(c) = \frac{1}{\xi^M} \int_{\sqrt{M} \lambda}^{\sqrt{c}} A e^{-R^2/2} dR \quad (A.63)
\]

where,

\[
R \in [\sqrt{M} \lambda, \sqrt{c}] \quad (A.64)
\]

\[
A_M = \frac{MR^{M-1} \pi^{M/2}}{\Gamma \left( \frac{M}{2} + 1 \right)} \quad (A.65)
\]

Using equations (A.65) and (A.65) in (A.63) with the property of gamma functions \( \Gamma \left( \frac{M}{2} + 1 \right) = \frac{M}{2} \Gamma \left( \frac{M}{2} \right) \),

\[
F_{CM}(c) = \frac{2\pi^{M/2}}{\xi^M \Gamma \left( \frac{M}{2} \right)} \int_{\sqrt{M} \lambda}^{\sqrt{c}} R^{M-1} e^{-R^2/2} dR \quad (A.66)
\]

Using the second fundamental theorem of calculus \([4]\) in equation (A.66), the density of \( C_M \) is derived as,

\[
f_{CM}(c) = \frac{\pi^{M/2} c^{M/2 - 1} e^{c/2}}{\xi^M \Gamma \left( \frac{M}{2} \right)} \quad (A.67)
\]
Replacing $\xi$, the density can be obtained as,

$$f_{CM}(c) \propto \frac{c^{M/2-1}e^{c/2}}{2^{M/2+1}(1 - \Phi(\lambda))^{M}\Gamma\left(\frac{M}{2}\right)}$$  \hspace{1cm} (A.68)

$$= K \frac{c^{M/2-1}e^{c/2}}{2^{M/2+1}(1 - \Phi(\lambda))^{M}\Gamma\left(\frac{M}{2}\right)}$$  \hspace{1cm} (A.69)

Equation (A.69) represents an unnormalized density and by integrating $f_{CM}$ over the support of $C_M$, the actual normalized density, $f(C_M)$, can be obtained. Therefore, we require,

$$K \int_{M^{\lambda^2}}^\infty f_{CM}(c)dc = 1$$  \hspace{1cm} (A.70)

where ‘$K$’ is the normalizing constant. That is,

$$K \int_{M^{\lambda^2}}^\infty \frac{c^{M/2-1}e^{c/2}}{2^{M/2+1}(1 - \Phi(\lambda))^{M}\Gamma\left(\frac{M}{2}\right)}dc = 1$$  \hspace{1cm} (A.71)

Let $b = \frac{c}{2}$, then

$$K \int_{M^{\lambda^2}}^\infty f_{CM}(c)dc = K \int_{M^{\lambda^2/2}}^\infty \frac{1}{2(1 - \Phi(\lambda))^{M}} \int_{M^{\lambda^2/2}}^\infty b^{M/2-1}e^{-b}db = 1$$  \hspace{1cm} (A.72)

The integral in equation (A.73) corresponds to the upper incomplete gamma function, $\Gamma\left(\frac{M}{2}, M^{\lambda^2/2}\right)$. Equivalently, the upper incomplete gamma function can be expressed using the cumulative distribution function of a Gamma($s,1$) random variable, $\Phi_G(\bullet)$. Here, ‘$s$’ is the scale parameter. Using, $\Gamma(s,x) = \left[1 - \Phi_G(x)\right]\Gamma(s)$ in equation (A.73) and setting it equal to 1, the normalizing constant can be found as,

$$K = \frac{2(1 - \Phi(\lambda))^{M}}{1 - \Phi_G(M^{\lambda^2/2})}$$  \hspace{1cm} (A.74)
Using equation (A.74) in (A.69), the density can be deduced as,

\[ f_{CM}(c) = \frac{c^{M/2-1}e^{-c/2}}{2^{M/2}[1 - \Phi_G(M\lambda^2/2)]\Gamma(M/2)} \] (A.75)

Due to independence,

\[ E[e^{CMt}] = E \left[ e^{\sum_{i=1}^{M} X_i t} \right] = \prod_{i=1}^{M} E[e^{X_i t}] \] (A.76) = (A.77)

and since \( X_i \sim \chi^2_1[\lambda] \), the moment generating function can be found as,

\[ M_{CM}(t) = (1 - 2t)^{-M/2}, \quad t \leq 0 \] (A.78)

A.1.7 Density of \( \kappa_\eta \)

By definition,

\[ \kappa_\eta^{(p,q,\lambda)} = \sum_{i=1}^{p} X_i^2 + \sum_{i=1}^{q} Y_i^2 \] (A.79)

where \( X_i \) is a random variable with a normal distribution with mean ‘0’ and standard deviation ‘1’, \( Y_i \) is a truncated normal random variable with truncation parameter \( \lambda \). That is,

\[ Y_i = X_i I_{|X_i|>\lambda} \] (A.80)

Thus,

\[ Z = \sum_{i=1}^{p} X_i^2 \sim \chi^2_p[\lambda] \] (A.81)

\[ W = \sum_{i=1}^{q} Y_i^2 \sim \chi^2_q \] (A.82)

Define,

\[ S = W + Z = \kappa_\eta \Rightarrow f_S(s) = \int_{\lambda^2}^{s} f_W(s-t)f_Z(t)dt \] (A.83)
Since,

\[ f_W(w) = \frac{w^{q/2-1}e^{-w/2}}{2^{q/2}\Gamma(q/2)} \]  \hspace{2cm} (A.84)

\[ f_Z(z) = \frac{z^{q/2-1}e^{-z/2}}{2^{q/2}\Gamma(p/2)(1 - \Phi_G(\nu\lambda^2/2))} \]  \hspace{2cm} (A.85)

where \( G \sim \text{Gamma}(p/2, 1) \). Thus,

\[ f_S(s) = \int_{p\lambda^2}^{s} \frac{(s-t)^{q/2-1}e^{-(s-t)/2}}{2^{q/2}\Gamma(q/2)} \frac{t^{q/2-1}e^{-t/2}}{1 - \Phi_G(\nu\lambda^2/2)} dt \]  \hspace{2cm} (A.86)

\[ = \frac{1}{\Gamma(p/2)\Gamma(q/2)2^{(p+q)/2}} \int_{p\lambda^2}^{s} (s-t)^{q/2-1}e^{-s/2}t^{p/2-1} dt \]  \hspace{2cm} (A.87)

\[ = \frac{1}{\Gamma(p/2)\Gamma(q/2)2^{(p+q)/2}(1 - \Phi_G(\nu\lambda^2/2))} \int_{p\lambda^2}^{s} (s-t)^{q/2-1}e^{-s/2}t^{p/2-1} dt \]  \hspace{2cm} (A.88)

\[ = \frac{e^{-s/2}s^{q/2-1}}{\Gamma(p/2)\Gamma(q/2)2^{(p+q)/2}(1 - \Phi_G(\nu\lambda^2/2))} \int_{p\lambda^2}^{s} (1 - t/s)^{q/2-1}(t/s)^{p/2-1} dt \]  \hspace{2cm} (A.89)

Using a substitution, \( r = t/s \), the above integral becomes,

\[ = \frac{e^{-s/2}s^{q/2-1}}{\Gamma(p/2)\Gamma(q/2)2^{(p+q)/2}} \int_{p\lambda^2}^{1} (1 - r)^{q/2-1}r^{p/2-1} dr \]  \hspace{2cm} (A.91)

\[ = \frac{e^{-s/2}s^{q/2-1}}{\Gamma((p+q)/2)2^{(p+q)/2}(1 - \Phi_B(j\lambda^2/p, q/2))} \]  \hspace{2cm} (A.92)

\[ = f_C(s) \left( 1 - \Phi_B(p\lambda^2/s, p/2, q/2) \right) \]  \hspace{2cm} (A.93)

where \( C \sim \chi^2_{(p+q)} \).

\[ Q.E.D \]

A.1.8 Mean and Variance of \( \kappa_{\eta} \)

Consider,

\[ X_i \sim Z_i \mathbb{I}_{\{|Z_i| > \lambda\}} \]  \hspace{2cm} (A.94)
where \( Z_i \sim N(0, 1) \). Under normality assumptions,

\[
\sum_{i=1}^{p} X_i^2 \sim \chi_p^2
\]  

(A.95)

Thus,

\[
E[X_i^2] = \int_{-\infty}^{\infty} z_i^2 \frac{1}{\sqrt{2\pi}} e^{-z_i^2/2} \{ I_{\{Z_i>|\lambda}\} \} dz_i
\]

(A.96)

Define, \( Y = z_i/2 \), then,

\[
E[X_i^2] = \int_{-\infty}^{-\lambda/2} 4y^2 \frac{1}{\sqrt{2\pi}} e^{-y} (2) dy + \int_{\lambda/2}^{\infty} 4y^2 \frac{1}{\sqrt{2\pi}} e^{-y} (2) dy
\]

(A.98)

\[
= \frac{4\Gamma(3)}{\sqrt{2\pi}} \int_{\lambda/2}^{\infty} y^{3-1} e^{-y} (2) dy \quad \frac{1}{\Gamma(3)} (2) dy
\]

(A.99)

\[
E[X_i^2] = 8\frac{\Gamma(3)}{\sqrt{2\pi}} \{ 1 - \Phi_G(\lambda/2, 3, 1) \} \equiv \mu
\]

(A.100)

where \( \mu \) is defined in (5.10). To find the variance, define, \( W = \sum_{i=1}^{p} X_i^2 \). Then,

\[
Var(W) = pVar(X_i^2)
\]

(A.101)

\[
= p(E[X_i^4] - \mu^2)
\]

(A.102)

\[
E[X_i^4] = \int_{-\infty}^{\infty} z_i^4 \{ I_{\{Z_i>|\lambda}\} \} e^{-z_i^2/2} dz_i
\]

(A.103)

Define, \( Y = z_i/2 \), then,

\[
E[X_i^4] = \int_{-\infty}^{-\lambda/2} 16y^4 \frac{1}{\sqrt{2\pi}} e^{-y} (2) dy + \int_{\lambda/2}^{\infty} 16y^5 \frac{1}{\sqrt{2\pi}} e^{-y} (2) dy
\]

(A.104)

\[
E[X_i^4] = \frac{16\Gamma(5)}{\sqrt{2\pi}} \int_{\lambda/2}^{\infty} y^{5-1} e^{-y} (2) dy \quad \frac{1}{\Gamma(5)} (2) dy
\]

(A.105)

\[
= \frac{32\Gamma(5)}{\sqrt{2\pi}} \{ 1 - \Phi_G(\lambda/2, 5, 1) \}
\]

(A.106)
Thus,

$$\sigma^2 = \frac{32p\Gamma(5)}{\sqrt{2\pi}} [1 - \Phi_G(\lambda/2, 5, 1)] - p\mu^2$$  \hspace{1cm} (A.107)$$

where $\sigma^2$ is as defined in (5.10).
APPENDIX B

TABLES

In this chapter, tables of critical values obtained using various approximations to the distribution of $\kappa_\eta$ are summarized. The notations used for each approximation (see Chapter 5) in the table of critical values are shown below.

- **I**: The Normal approximation
- **II**: The Chi-squared approximation
- **III**: The Binomial-Normal mixture approximation
- **IV**: The Binomial-Chi squared mixture approximation
- **V**: The exact distribution
- **SIM**: A simulated distribution under $H_0$.

The universal threshold ($\lambda$) is employed under a hard thresholding rule to obtain $\kappa_\eta^{(p,q,\lambda)}$. The Haar filter is employed and the critical values are obtained for varying levels of threshold ($l_1$). The distribution of the statistic under $H_0$ is obtained through repetitive simulations (500 simulations) and a critical value is obtained using the respective quantiles of the simulated statistics. These empirical critical values are given by SIM. The critical value tables are obtained for three levels of significance ($\alpha$), 0.01, 0.05 and 0.1.

In addition to varying levels of threshold, the critical values are summarized for varying sample sizes. The sample sizes include 128, 256, 512, 1024, 2048 and 4096. These tables can be used directly in applications involving such sample sizes.
B.1 Critical Value Tables

Table B.1: Critical values obtained using the exact distribution of $\kappa_\eta$ and its approximations with $\alpha = 0.1$.

<table>
<thead>
<tr>
<th>n</th>
<th>Type</th>
<th>$l_t = 2$</th>
<th>$l_t = 3$</th>
<th>$l_t = 4$</th>
<th>$l_t = 5$</th>
<th>$l_t = 6$</th>
</tr>
</thead>
<tbody>
<tr>
<td>128</td>
<td>I</td>
<td>43.75</td>
<td>25.30</td>
<td>18.57</td>
<td>16.62</td>
<td>15.89</td>
</tr>
<tr>
<td></td>
<td>II</td>
<td>60.64</td>
<td>43.49</td>
<td>34.90</td>
<td>30.60</td>
<td>28.44</td>
</tr>
<tr>
<td></td>
<td>III</td>
<td>48.44</td>
<td>29.70</td>
<td>19.90</td>
<td>14.78</td>
<td>12.11</td>
</tr>
<tr>
<td></td>
<td>IV</td>
<td>42.19</td>
<td>24.08</td>
<td>14.83</td>
<td>10.10</td>
<td>7.70</td>
</tr>
<tr>
<td></td>
<td>V</td>
<td>43.62</td>
<td>24.82</td>
<td>14.84</td>
<td>9.46</td>
<td>6.52</td>
</tr>
<tr>
<td></td>
<td>SIM</td>
<td>48.78</td>
<td>32.77</td>
<td>22.70</td>
<td>20.41</td>
<td>15.71</td>
</tr>
<tr>
<td>256</td>
<td>I</td>
<td>79.98</td>
<td>43.76</td>
<td>25.69</td>
<td>19.66</td>
<td>17.94</td>
</tr>
<tr>
<td></td>
<td>II</td>
<td>94.59</td>
<td>60.45</td>
<td>43.31</td>
<td>34.72</td>
<td>30.41</td>
</tr>
<tr>
<td></td>
<td>III</td>
<td>84.30</td>
<td>48.24</td>
<td>29.49</td>
<td>19.68</td>
<td>14.54</td>
</tr>
<tr>
<td></td>
<td>IV</td>
<td>77.62</td>
<td>42.22</td>
<td>24.06</td>
<td>14.72</td>
<td>9.92</td>
</tr>
<tr>
<td></td>
<td>V</td>
<td>79.77</td>
<td>43.69</td>
<td>24.79</td>
<td>14.75</td>
<td>9.34</td>
</tr>
<tr>
<td></td>
<td>SIM</td>
<td>83.46</td>
<td>48.88</td>
<td>34.46</td>
<td>24.22</td>
<td>18.49</td>
</tr>
<tr>
<td>512</td>
<td>I</td>
<td>149.98</td>
<td>79.99</td>
<td>43.75</td>
<td>26.19</td>
<td>20.80</td>
</tr>
<tr>
<td></td>
<td>II</td>
<td>162.22</td>
<td>94.41</td>
<td>60.27</td>
<td>43.13</td>
<td>34.53</td>
</tr>
<tr>
<td></td>
<td>III</td>
<td>153.98</td>
<td>84.11</td>
<td>48.04</td>
<td>29.28</td>
<td>19.46</td>
</tr>
<tr>
<td></td>
<td>IV</td>
<td>146.71</td>
<td>77.57</td>
<td>42.20</td>
<td>24.00</td>
<td>14.59</td>
</tr>
<tr>
<td></td>
<td>V</td>
<td>149.69</td>
<td>79.83</td>
<td>43.67</td>
<td>24.73</td>
<td>14.66</td>
</tr>
<tr>
<td></td>
<td>SIM</td>
<td>152.13</td>
<td>84.95</td>
<td>50.55</td>
<td>32.33</td>
<td>24.39</td>
</tr>
<tr>
<td>1024</td>
<td>I</td>
<td>286.62</td>
<td>149.98</td>
<td>80.00</td>
<td>43.77</td>
<td>26.81</td>
</tr>
<tr>
<td></td>
<td>II</td>
<td>296.47</td>
<td>162.03</td>
<td>94.22</td>
<td>60.09</td>
<td>42.94</td>
</tr>
<tr>
<td></td>
<td>III</td>
<td>290.21</td>
<td>153.81</td>
<td>83.93</td>
<td>47.85</td>
<td>29.08</td>
</tr>
<tr>
<td></td>
<td>IV</td>
<td>282.06</td>
<td>146.55</td>
<td>77.54</td>
<td>42.16</td>
<td>23.91</td>
</tr>
<tr>
<td></td>
<td>V</td>
<td>286.12</td>
<td>149.75</td>
<td>79.82</td>
<td>43.62</td>
<td>24.65</td>
</tr>
<tr>
<td></td>
<td>SIM</td>
<td>289.11</td>
<td>156.95</td>
<td>82.96</td>
<td>50.47</td>
<td>34.02</td>
</tr>
<tr>
<td>2048</td>
<td>I</td>
<td>555.56</td>
<td>286.58</td>
<td>149.99</td>
<td>79.98</td>
<td>43.79</td>
</tr>
<tr>
<td></td>
<td>II</td>
<td>562.54</td>
<td>296.29</td>
<td>161.85</td>
<td>94.04</td>
<td>59.90</td>
</tr>
<tr>
<td></td>
<td>III</td>
<td>557.99</td>
<td>290.04</td>
<td>153.64</td>
<td>83.75</td>
<td>47.67</td>
</tr>
<tr>
<td></td>
<td>IV</td>
<td>548.59</td>
<td>281.77</td>
<td>146.49</td>
<td>77.51</td>
<td>42.11</td>
</tr>
<tr>
<td></td>
<td>V</td>
<td>554.08</td>
<td>286.18</td>
<td>149.75</td>
<td>79.78</td>
<td>43.56</td>
</tr>
<tr>
<td></td>
<td>SIM</td>
<td>559.05</td>
<td>289.82</td>
<td>155.34</td>
<td>88.64</td>
<td>52.02</td>
</tr>
<tr>
<td>4096</td>
<td>I</td>
<td>1087.09</td>
<td>555.54</td>
<td>286.59</td>
<td>149.97</td>
<td>79.97</td>
</tr>
<tr>
<td></td>
<td>II</td>
<td>1089.97</td>
<td>562.37</td>
<td>296.11</td>
<td>161.67</td>
<td>93.85</td>
</tr>
<tr>
<td></td>
<td>III</td>
<td>1086.78</td>
<td>557.84</td>
<td>289.88</td>
<td>153.48</td>
<td>83.59</td>
</tr>
<tr>
<td></td>
<td>IV</td>
<td>1075.61</td>
<td>548.14</td>
<td>281.67</td>
<td>146.47</td>
<td>77.48</td>
</tr>
<tr>
<td></td>
<td>V</td>
<td>1083.02</td>
<td>554.13</td>
<td>286.17</td>
<td>149.71</td>
<td>79.72</td>
</tr>
<tr>
<td></td>
<td>SIM</td>
<td>1085.45</td>
<td>559.08</td>
<td>289.80</td>
<td>153.47</td>
<td>85.55</td>
</tr>
</tbody>
</table>
Table B.2: Critical values obtained using the exact distribution of $\kappa_\eta$ and its approximations with $\alpha = 0.05$.

<table>
<thead>
<tr>
<th>n</th>
<th>Type</th>
<th>$l_t = 2$</th>
<th>$l_t = 3$</th>
<th>$l_t = 4$</th>
<th>$l_t = 5$</th>
<th>$l_t = 6$</th>
</tr>
</thead>
<tbody>
<tr>
<td>128</td>
<td>I</td>
<td>47.41</td>
<td>28.07</td>
<td>20.61</td>
<td>18.29</td>
<td>17.38</td>
</tr>
<tr>
<td></td>
<td>II</td>
<td>67.32</td>
<td>49.85</td>
<td>41.09</td>
<td>36.70</td>
<td>34.50</td>
</tr>
<tr>
<td></td>
<td>III</td>
<td>52.28</td>
<td>32.76</td>
<td>22.45</td>
<td>17.02</td>
<td>14.17</td>
</tr>
<tr>
<td></td>
<td>IV</td>
<td>44.82</td>
<td>26.08</td>
<td>16.45</td>
<td>11.50</td>
<td>8.98</td>
</tr>
<tr>
<td></td>
<td>V</td>
<td>47.27</td>
<td>27.64</td>
<td>17.08</td>
<td>11.31</td>
<td>8.12</td>
</tr>
<tr>
<td></td>
<td>SIM</td>
<td>53.12</td>
<td>36.36</td>
<td>27.80</td>
<td>23.54</td>
<td>22.66</td>
</tr>
<tr>
<td>256</td>
<td>I</td>
<td>84.83</td>
<td>47.41</td>
<td>28.43</td>
<td>21.66</td>
<td>19.62</td>
</tr>
<tr>
<td></td>
<td>II</td>
<td>101.87</td>
<td>67.13</td>
<td>49.67</td>
<td>40.91</td>
<td>36.52</td>
</tr>
<tr>
<td></td>
<td>III</td>
<td>89.27</td>
<td>52.06</td>
<td>32.53</td>
<td>22.22</td>
<td>16.77</td>
</tr>
<tr>
<td></td>
<td>IV</td>
<td>81.26</td>
<td>44.85</td>
<td>26.06</td>
<td>16.33</td>
<td>11.29</td>
</tr>
<tr>
<td></td>
<td>V</td>
<td>84.61</td>
<td>47.34</td>
<td>27.61</td>
<td>16.99</td>
<td>11.18</td>
</tr>
<tr>
<td></td>
<td>SIM</td>
<td>90.76</td>
<td>56.02</td>
<td>37.06</td>
<td>31.94</td>
<td>26.24</td>
</tr>
<tr>
<td>512</td>
<td>I</td>
<td>156.54</td>
<td>84.83</td>
<td>47.41</td>
<td>28.90</td>
<td>22.78</td>
</tr>
<tr>
<td></td>
<td>II</td>
<td>170.58</td>
<td>101.69</td>
<td>66.95</td>
<td>49.48</td>
<td>40.72</td>
</tr>
<tr>
<td></td>
<td>III</td>
<td>160.61</td>
<td>89.07</td>
<td>51.86</td>
<td>32.32</td>
<td>21.99</td>
</tr>
<tr>
<td></td>
<td>IV</td>
<td>151.80</td>
<td>81.17</td>
<td>44.83</td>
<td>25.99</td>
<td>16.18</td>
</tr>
<tr>
<td></td>
<td>V</td>
<td>156.23</td>
<td>84.67</td>
<td>47.32</td>
<td>27.54</td>
<td>16.89</td>
</tr>
<tr>
<td></td>
<td>SIM</td>
<td>160.74</td>
<td>90.86</td>
<td>55.73</td>
<td>37.08</td>
<td>28.02</td>
</tr>
<tr>
<td>1024</td>
<td>I</td>
<td>295.55</td>
<td>156.55</td>
<td>84.82</td>
<td>47.42</td>
<td>29.48</td>
</tr>
<tr>
<td></td>
<td>II</td>
<td>306.64</td>
<td>170.39</td>
<td>101.50</td>
<td>66.76</td>
<td>49.29</td>
</tr>
<tr>
<td></td>
<td>III</td>
<td>299.21</td>
<td>160.43</td>
<td>88.89</td>
<td>51.66</td>
<td>32.11</td>
</tr>
<tr>
<td></td>
<td>IV</td>
<td>289.25</td>
<td>151.58</td>
<td>81.13</td>
<td>44.79</td>
<td>25.90</td>
</tr>
<tr>
<td></td>
<td>V</td>
<td>295.06</td>
<td>156.29</td>
<td>84.66</td>
<td>47.27</td>
<td>27.46</td>
</tr>
<tr>
<td></td>
<td>SIM</td>
<td>300.18</td>
<td>159.15</td>
<td>89.86</td>
<td>57.90</td>
<td>37.75</td>
</tr>
<tr>
<td>2048</td>
<td>I</td>
<td>567.76</td>
<td>295.51</td>
<td>156.52</td>
<td>84.85</td>
<td>47.44</td>
</tr>
<tr>
<td></td>
<td>II</td>
<td>575.62</td>
<td>306.46</td>
<td>170.20</td>
<td>101.31</td>
<td>66.56</td>
</tr>
<tr>
<td></td>
<td>III</td>
<td>570.37</td>
<td>299.04</td>
<td>160.26</td>
<td>88.71</td>
<td>51.48</td>
</tr>
<tr>
<td></td>
<td>IV</td>
<td>558.79</td>
<td>288.87</td>
<td>151.50</td>
<td>81.10</td>
<td>44.74</td>
</tr>
<tr>
<td></td>
<td>V</td>
<td>566.42</td>
<td>295.12</td>
<td>156.28</td>
<td>84.62</td>
<td>47.20</td>
</tr>
<tr>
<td></td>
<td>SIM</td>
<td>570.91</td>
<td>297.57</td>
<td>160.44</td>
<td>92.90</td>
<td>59.53</td>
</tr>
<tr>
<td>4096</td>
<td>I</td>
<td>1103.92</td>
<td>567.68</td>
<td>295.53</td>
<td>156.51</td>
<td>84.85</td>
</tr>
<tr>
<td></td>
<td>II</td>
<td>1107.46</td>
<td>575.44</td>
<td>306.28</td>
<td>170.01</td>
<td>101.11</td>
</tr>
<tr>
<td></td>
<td>III</td>
<td>1103.96</td>
<td>570.22</td>
<td>298.88</td>
<td>160.09</td>
<td>88.54</td>
</tr>
<tr>
<td></td>
<td>IV</td>
<td>1090.07</td>
<td>558.19</td>
<td>288.74</td>
<td>151.49</td>
<td>81.09</td>
</tr>
<tr>
<td></td>
<td>V</td>
<td>1100.17</td>
<td>566.47</td>
<td>295.11</td>
<td>156.25</td>
<td>84.56</td>
</tr>
<tr>
<td></td>
<td>SIM</td>
<td>1104.95</td>
<td>572.54</td>
<td>296.50</td>
<td>160.94</td>
<td>92.16</td>
</tr>
</tbody>
</table>
Table B.3: Critical values obtained using the exact distribution of $\kappa_\eta$ and its approximations with $\alpha = 0.01$.

<table>
<thead>
<tr>
<th>n</th>
<th>Type</th>
<th>$l_t = 2$</th>
<th>$l_t = 3$</th>
<th>$l_t = 4$</th>
<th>$l_t = 5$</th>
<th>$l_t = 6$</th>
</tr>
</thead>
<tbody>
<tr>
<td>128</td>
<td>I</td>
<td>54.79</td>
<td>33.83</td>
<td>25.04</td>
<td>21.93</td>
<td>20.43</td>
</tr>
<tr>
<td></td>
<td>II</td>
<td>79.85</td>
<td>61.77</td>
<td>52.69</td>
<td>48.14</td>
<td>45.87</td>
</tr>
<tr>
<td></td>
<td>III</td>
<td>59.98</td>
<td>39.03</td>
<td>27.79</td>
<td>21.77</td>
<td>18.59</td>
</tr>
<tr>
<td></td>
<td>IV</td>
<td>49.76</td>
<td>29.82</td>
<td>19.48</td>
<td>14.13</td>
<td>11.39</td>
</tr>
<tr>
<td></td>
<td>V</td>
<td>54.64</td>
<td>33.47</td>
<td>21.85</td>
<td>15.36</td>
<td>11.69</td>
</tr>
<tr>
<td></td>
<td>SIM</td>
<td>65.48</td>
<td>46.27</td>
<td>35.83</td>
<td>36.41</td>
<td>37.72</td>
</tr>
<tr>
<td>256</td>
<td>I</td>
<td>94.44</td>
<td>54.79</td>
<td>34.15</td>
<td>26.06</td>
<td>23.32</td>
</tr>
<tr>
<td></td>
<td>II</td>
<td>115.54</td>
<td>79.66</td>
<td>61.59</td>
<td>52.51</td>
<td>47.96</td>
</tr>
<tr>
<td></td>
<td>III</td>
<td>99.10</td>
<td>59.76</td>
<td>38.79</td>
<td>27.53</td>
<td>21.50</td>
</tr>
<tr>
<td></td>
<td>IV</td>
<td>88.07</td>
<td>49.78</td>
<td>29.81</td>
<td>19.34</td>
<td>13.87</td>
</tr>
<tr>
<td></td>
<td>V</td>
<td>94.20</td>
<td>54.71</td>
<td>33.44</td>
<td>21.75</td>
<td>15.21</td>
</tr>
<tr>
<td></td>
<td>SIM</td>
<td>103.88</td>
<td>69.40</td>
<td>50.83</td>
<td>39.01</td>
<td>46.78</td>
</tr>
<tr>
<td>512</td>
<td>I</td>
<td>169.31</td>
<td>94.45</td>
<td>54.79</td>
<td>34.56</td>
<td>27.14</td>
</tr>
<tr>
<td></td>
<td>II</td>
<td>186.26</td>
<td>115.35</td>
<td>79.47</td>
<td>61.39</td>
<td>52.32</td>
</tr>
<tr>
<td></td>
<td>III</td>
<td>173.54</td>
<td>98.89</td>
<td>59.54</td>
<td>38.55</td>
<td>27.28</td>
</tr>
<tr>
<td></td>
<td>IV</td>
<td>161.35</td>
<td>87.93</td>
<td>49.76</td>
<td>29.73</td>
<td>19.17</td>
</tr>
<tr>
<td></td>
<td>V</td>
<td>168.99</td>
<td>94.27</td>
<td>54.69</td>
<td>33.36</td>
<td>21.64</td>
</tr>
<tr>
<td></td>
<td>SIM</td>
<td>174.30</td>
<td>106.08</td>
<td>72.50</td>
<td>46.58</td>
<td>41.37</td>
</tr>
<tr>
<td>1024</td>
<td>I</td>
<td>312.75</td>
<td>169.29</td>
<td>94.44</td>
<td>54.80</td>
<td>35.07</td>
</tr>
<tr>
<td></td>
<td>II</td>
<td>325.74</td>
<td>186.07</td>
<td>115.15</td>
<td>79.27</td>
<td>61.19</td>
</tr>
<tr>
<td></td>
<td>III</td>
<td>316.58</td>
<td>173.35</td>
<td>98.70</td>
<td>59.33</td>
<td>38.33</td>
</tr>
<tr>
<td></td>
<td>IV</td>
<td>302.74</td>
<td>161.02</td>
<td>87.87</td>
<td>49.71</td>
<td>29.62</td>
</tr>
<tr>
<td></td>
<td>V</td>
<td>312.32</td>
<td>169.05</td>
<td>94.25</td>
<td>54.63</td>
<td>33.27</td>
</tr>
<tr>
<td></td>
<td>SIM</td>
<td>315.27</td>
<td>182.16</td>
<td>105.03</td>
<td>66.91</td>
<td>45.74</td>
</tr>
<tr>
<td>2048</td>
<td>I</td>
<td>591.24</td>
<td>312.79</td>
<td>169.32</td>
<td>94.43</td>
<td>54.80</td>
</tr>
<tr>
<td></td>
<td>II</td>
<td>600.16</td>
<td>325.55</td>
<td>185.87</td>
<td>114.94</td>
<td>79.05</td>
</tr>
<tr>
<td></td>
<td>III</td>
<td>594.09</td>
<td>316.41</td>
<td>173.17</td>
<td>98.51</td>
<td>59.13</td>
</tr>
<tr>
<td></td>
<td>IV</td>
<td>577.91</td>
<td>302.18</td>
<td>160.91</td>
<td>87.85</td>
<td>49.67</td>
</tr>
<tr>
<td></td>
<td>V</td>
<td>590.05</td>
<td>312.38</td>
<td>169.05</td>
<td>94.21</td>
<td>54.56</td>
</tr>
<tr>
<td></td>
<td>SIM</td>
<td>601.74</td>
<td>316.48</td>
<td>180.63</td>
<td>100.13</td>
<td>69.31</td>
</tr>
<tr>
<td>4096</td>
<td>I</td>
<td>1135.86</td>
<td>591.15</td>
<td>312.76</td>
<td>169.31</td>
<td>94.45</td>
</tr>
<tr>
<td></td>
<td>II</td>
<td>1140.28</td>
<td>599.97</td>
<td>325.35</td>
<td>185.66</td>
<td>114.72</td>
</tr>
<tr>
<td></td>
<td>III</td>
<td>1136.68</td>
<td>593.93</td>
<td>316.25</td>
<td>173.00</td>
<td>98.33</td>
</tr>
<tr>
<td></td>
<td>IV</td>
<td>1117.20</td>
<td>577.04</td>
<td>302.00</td>
<td>160.90</td>
<td>87.84</td>
</tr>
<tr>
<td></td>
<td>V</td>
<td>1132.83</td>
<td>590.11</td>
<td>312.38</td>
<td>169.01</td>
<td>94.15</td>
</tr>
<tr>
<td></td>
<td>SIM</td>
<td>1139.19</td>
<td>582.55</td>
<td>323.58</td>
<td>180.16</td>
<td>102.28</td>
</tr>
</tbody>
</table>
B.2 iSeg: A Short Summary

A brief description of iSeg, developed by S.B. Girimurugan et al., used in segmentation is given below.

- **Step-I:** For a given profile, iSeg scans a large number of segments starting with a minimum window length, $W_{\text{min}}$, and up to a maximum window length, $W_{\text{max}}$, which can be defined by a user. The window length increases by a fixed multiplicative factor, called power factor, $\rho$, which can also be defined by the user. All tuning parameters have default values, they are determined by the user depending on dataset to be analyzed and using benchmark datasets, if they are available.

- **Step-II:** p-values are computed for all the segments and a set of non-overlapping segments which are most significant among all the possible segments are detected. The test statistic is designed based on a normal assumption with mean under $H_0$. The standard deviation is estimated using median absolute deviation (MAD) on the whole profile and assumed to be known. The detection of overlapping segments are done using a data structure based on balanced binary trees.

- **Step-III:** A refinement step consisting of expansion, shrinking, and merging of significant segments is performed to obtain optimal segments with smaller p-values.

- **Step-IV:** The identified regions are sorted and selected based on a desired false discovery rate (FDR) using the Benjamini-Hochberg procedure.

A biological significance cut-off (BC) can be applied in step II to filter out segments with mean values smaller than the cut-off. iSeg can be accessed freely on-line at http://cloud.stat.fsu.edu/iSeg.
### APPENDIX C

**SUPPLEMENTARY RESULTS**

Table C.1: Common list of genes that demonstrate a significant change between Normal and Tumour profiles among Grade I cancer patients 1357 and 4137

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>uc001bls.1</td>
<td>chr1</td>
<td>CEFP5</td>
<td>uc002vrt.4</td>
<td>chr20</td>
<td>DTY1</td>
<td>uc001ef.2</td>
<td>chr19</td>
<td>LOC644554</td>
</tr>
<tr>
<td>uc002eqb.1</td>
<td>chr10</td>
<td>TNKAC1L1</td>
<td>uc002tcc.4</td>
<td>chr22</td>
<td>CBLN4</td>
<td>uc001bhe.3</td>
<td>chr22</td>
<td>SYCE5</td>
</tr>
<tr>
<td>uc001fmi.3</td>
<td>chr1</td>
<td>LMNA</td>
<td>uc002sra.3</td>
<td>chr20</td>
<td>PHACTR3</td>
<td>uc001hka.4</td>
<td>chr3</td>
<td>TREX1</td>
</tr>
<tr>
<td>uc001gji.4</td>
<td>chr1</td>
<td>SELP</td>
<td>uc002ykk.1</td>
<td>chr21</td>
<td>MIR99A</td>
<td>uc001hmb.2</td>
<td>chr3</td>
<td>LINC00969</td>
</tr>
<tr>
<td>uc001itx.4</td>
<td>chr10</td>
<td>MKX</td>
<td>uc003alk.2</td>
<td>chr22</td>
<td>PISD</td>
<td>uc001hba.4</td>
<td>chr4</td>
<td>ZNF876P</td>
</tr>
<tr>
<td>uc001loa.1</td>
<td>chr11</td>
<td>SGC1B1C</td>
<td>uc003aap.1</td>
<td>chr22</td>
<td>TMPRSS6</td>
<td>uc001khx.3</td>
<td>chr7</td>
<td>ABCB5</td>
</tr>
<tr>
<td>uc001lyf.4</td>
<td>chr11</td>
<td>CNN1A10</td>
<td>uc003cg.2</td>
<td>chr3</td>
<td>UBE2E2</td>
<td>uc001peq.3</td>
<td>chr1</td>
<td>ZBTB17B</td>
</tr>
<tr>
<td>uc002mnc.3</td>
<td>chr1</td>
<td>ABC8C</td>
<td>uc003cpx.3</td>
<td>chr3</td>
<td>KIF9</td>
<td>uc001ple.3</td>
<td>chr10</td>
<td>TAF3</td>
</tr>
<tr>
<td>uc002nng.2</td>
<td>chr11</td>
<td>SLC15A3</td>
<td>uc003crd.3</td>
<td>chr3</td>
<td>KLHL13</td>
<td>uc001nak.2</td>
<td>chr15</td>
<td>ANP32A-IT1</td>
</tr>
<tr>
<td>uc001nto.2</td>
<td>chr11</td>
<td>TUT1</td>
<td>uc003cxw.3</td>
<td>chr3</td>
<td>LAMB2</td>
<td>uc001vrt.2</td>
<td>chr17</td>
<td>TRPV1</td>
</tr>
<tr>
<td>uc001qav.4</td>
<td>chr11</td>
<td>ESAM</td>
<td>uc003dir.1</td>
<td>chr3</td>
<td>SEMA3G</td>
<td>uc002chm.2</td>
<td>chr19</td>
<td>C9orf71</td>
</tr>
<tr>
<td>uc001qet.3</td>
<td>chr11</td>
<td>TINAGL1</td>
<td>uc003dxy.3</td>
<td>chr3</td>
<td>UBE2E2</td>
<td>uc010ai.1</td>
<td>chr22</td>
<td>MIR1308</td>
</tr>
<tr>
<td>uc001sam.3</td>
<td>chr12</td>
<td>KRT6A</td>
<td>uc003dxx.3</td>
<td>chr3</td>
<td>LOC33974</td>
<td>uc010bap.2</td>
<td>chr3</td>
<td>PBSS42</td>
</tr>
<tr>
<td>uc001sib.4</td>
<td>chr12</td>
<td>MMP19</td>
<td>uc003dxj.2</td>
<td>chr3</td>
<td>GUCA1C</td>
<td>uc011bem.1</td>
<td>chr3</td>
<td>SNAR-I</td>
</tr>
<tr>
<td>uc001skm.3</td>
<td>chr12</td>
<td>GPR182</td>
<td>uc003fkk.3</td>
<td>chr3</td>
<td>TTC14</td>
<td>uc011kde.1</td>
<td>chr7</td>
<td>DKKFzp434L192</td>
</tr>
<tr>
<td>uc001nlq.3</td>
<td>chr12</td>
<td>ZNF10</td>
<td>uc003hia.3</td>
<td>chr4</td>
<td>FRAS1</td>
<td>uc011nmd.2</td>
<td>chr9</td>
<td>FAM163B</td>
</tr>
<tr>
<td>uc001ntc.2</td>
<td>chr13</td>
<td>USPL1</td>
<td>uc003jgg.3</td>
<td>chr4</td>
<td>SMARCA5</td>
<td>uc021ona.1</td>
<td>chr1</td>
<td>LOC729041</td>
</tr>
<tr>
<td>uc001ruit.3</td>
<td>chr13</td>
<td>RFXP2</td>
<td>uc003lel.3</td>
<td>chr4</td>
<td>RNF175</td>
<td>uc021on.1</td>
<td>chr1</td>
<td>MIR5015</td>
</tr>
<tr>
<td>uc001vkn.3</td>
<td>chr13</td>
<td>RNF13B</td>
<td>uc003mve.1</td>
<td>chr4</td>
<td>STOX2</td>
<td>uc021orf.1</td>
<td>chr12</td>
<td>LOC100507091</td>
</tr>
<tr>
<td>uc001xdj.2</td>
<td>chr12</td>
<td>PSMA3</td>
<td>uc003nja.3</td>
<td>chr4</td>
<td>CARD6</td>
<td>uc021tqz.1</td>
<td>chr17</td>
<td>GNGT2</td>
</tr>
<tr>
<td>uc001xsh.3</td>
<td>chr14</td>
<td>GPATCH2L</td>
<td>uc003oqj.3</td>
<td>chr4</td>
<td>F2RL1</td>
<td>uc021wlm.1</td>
<td>chr22</td>
<td>MIR301B</td>
</tr>
<tr>
<td>uc001zab.1</td>
<td>chr15</td>
<td>SNORD115-33</td>
<td>uc003dzy.4</td>
<td>chr4</td>
<td>AP3B1</td>
<td>uc021wym.1</td>
<td>chr3</td>
<td>HYAL3</td>
</tr>
<tr>
<td>uc001ztr.3</td>
<td>chr15</td>
<td>HIGD2B</td>
<td>uc003pfr.4</td>
<td>chr4</td>
<td>NRG2</td>
<td>uc021zxc.1</td>
<td>chr3</td>
<td>NUDT16</td>
</tr>
<tr>
<td>uc002ava.3</td>
<td>chr15</td>
<td>CBX5</td>
<td>uc003qym.2</td>
<td>chr5</td>
<td>CEP120</td>
<td>uc022sap.1</td>
<td>chr7</td>
<td>MIR548F</td>
</tr>
<tr>
<td>uc002avb.3</td>
<td>chr15</td>
<td>BBS4</td>
<td>uc003rve.2</td>
<td>chr5</td>
<td>NRG2</td>
<td>uc022xdl.1</td>
<td>chr8</td>
<td>LOC100127983</td>
</tr>
<tr>
<td>uc002bgm.3</td>
<td>chr16</td>
<td>TNRC6A</td>
<td>uc003sfo.3</td>
<td>chr6</td>
<td>PGDB1</td>
<td>uc022xkl.1</td>
<td>chr9</td>
<td>MIR4291</td>
</tr>
<tr>
<td>uc002ecs.5</td>
<td>chr16</td>
<td>ZNF267</td>
<td>uc003sng.4</td>
<td>chr6</td>
<td>KIAA1586</td>
<td>uc031als.1</td>
<td>chr5</td>
<td>OR2V1</td>
</tr>
<tr>
<td>uc002ejq.5</td>
<td>chr16</td>
<td>ZNF267</td>
<td>uc003vul.3</td>
<td>chr6</td>
<td>SELINC1</td>
<td>uc031asg.1</td>
<td>chr5</td>
<td>OR2V1</td>
</tr>
<tr>
<td>uc002ejn.3</td>
<td>chr17</td>
<td>MTIM</td>
<td>uc003wko.3</td>
<td>chr6</td>
<td>ULBP3</td>
<td>uc031asg.1</td>
<td>chr5</td>
<td>OR2V1</td>
</tr>
<tr>
<td>uc002ggg.3</td>
<td>chr17</td>
<td>SPA87</td>
<td>uc003xmo.4</td>
<td>chr7</td>
<td>CCDC132</td>
<td>uc031asg.1</td>
<td>chr5</td>
<td>OR2V1</td>
</tr>
<tr>
<td>uc002hwi.3</td>
<td>chr17</td>
<td>KRTAP9-4</td>
<td>uc003yxw.3</td>
<td>chr7</td>
<td>CREB3L2</td>
<td>uc031asg.1</td>
<td>chr5</td>
<td>OR2V1</td>
</tr>
<tr>
<td>uc002hxy.3</td>
<td>chr17</td>
<td>DNAJC7</td>
<td>uc003ykd.3</td>
<td>chr7</td>
<td>LOC100507156</td>
<td>uc031asg.1</td>
<td>chr5</td>
<td>OR2V1</td>
</tr>
<tr>
<td>uc002hqs.3</td>
<td>chr17</td>
<td>GFAP</td>
<td>uc003zio.1</td>
<td>chr7</td>
<td>STC1</td>
<td>uc031asg.1</td>
<td>chr5</td>
<td>OR2V1</td>
</tr>
<tr>
<td>uc002iop.1</td>
<td>chr17</td>
<td>ABH3</td>
<td>uc003zyj.3</td>
<td>chr8</td>
<td>C8orf86</td>
<td>uc031asg.1</td>
<td>chr5</td>
<td>OR2V1</td>
</tr>
<tr>
<td>uc002iop.1</td>
<td>chr17</td>
<td>TAC4</td>
<td>uc003zyj.3</td>
<td>chr8</td>
<td>MAPK15</td>
<td>uc031asg.1</td>
<td>chr5</td>
<td>OR2V1</td>
</tr>
<tr>
<td>uc002kji.4</td>
<td>chr18</td>
<td>THOC1</td>
<td>uc003zio.4</td>
<td>chr9</td>
<td>PPAPDC2</td>
<td>uc031asg.1</td>
<td>chr5</td>
<td>OR2V1</td>
</tr>
<tr>
<td>uc002kqt.3</td>
<td>chr18</td>
<td>CIDEA</td>
<td>uc003zio.3</td>
<td>chr9</td>
<td>CDC37L1</td>
<td>uc031asg.1</td>
<td>chr5</td>
<td>OR2V1</td>
</tr>
<tr>
<td>uc002hfl.1</td>
<td>chr18</td>
<td>RNF152</td>
<td>uc003zio.2</td>
<td>chr9</td>
<td>MLLT3</td>
<td>uc031asg.1</td>
<td>chr5</td>
<td>OR2V1</td>
</tr>
<tr>
<td>uc002mqi.2</td>
<td>chr18</td>
<td>NDUFA11</td>
<td>uc003zio.4</td>
<td>chr9</td>
<td>PRSS3</td>
<td>uc031asg.1</td>
<td>chr5</td>
<td>OR2V1</td>
</tr>
<tr>
<td>uc002ozj.3</td>
<td>chr19</td>
<td>HKR1</td>
<td>uc004ats.2</td>
<td>chr9</td>
<td>FGD3</td>
<td>uc031asg.1</td>
<td>chr5</td>
<td>OR2V1</td>
</tr>
<tr>
<td>uc002pft.4</td>
<td>chr19</td>
<td>CYP2A13</td>
<td>uc004btx.1</td>
<td>chr9</td>
<td>CRB2</td>
<td>uc031asg.1</td>
<td>chr5</td>
<td>OR2V1</td>
</tr>
<tr>
<td>uc002pjk.3</td>
<td>chr19</td>
<td>FAM83E</td>
<td>uc004bto.1</td>
<td>chr9</td>
<td>LCN2</td>
<td>uc031asg.1</td>
<td>chr5</td>
<td>OR2V1</td>
</tr>
<tr>
<td>uc002pwo.3</td>
<td>chr19</td>
<td>SIGLEC10</td>
<td>uc004cdv.4</td>
<td>chr9</td>
<td>ADAMT8</td>
<td>uc031asg.1</td>
<td>chr5</td>
<td>OR2V1</td>
</tr>
<tr>
<td>uc002qgb.4</td>
<td>chr19</td>
<td>TNNT1</td>
<td>uc004dit.3</td>
<td>chr9</td>
<td>ZNF182</td>
<td>uc031asg.1</td>
<td>chr5</td>
<td>OR2V1</td>
</tr>
<tr>
<td>uc002qk.4</td>
<td>chr19</td>
<td>SUV420H2</td>
<td>uc004fsc.3</td>
<td>chr9</td>
<td>SMIC1A</td>
<td>uc031asg.1</td>
<td>chr5</td>
<td>OR2V1</td>
</tr>
<tr>
<td>uc002pqj.3</td>
<td>chr19</td>
<td>NLRP5</td>
<td>uc004fus.3</td>
<td>chr9</td>
<td>SASH3</td>
<td>uc031asg.1</td>
<td>chr5</td>
<td>OR2V1</td>
</tr>
<tr>
<td>uc002qsc.2</td>
<td>chr19</td>
<td>ZSCAN22</td>
<td>uc004fyn.3</td>
<td>chr9</td>
<td>TMLHE</td>
<td>uc031asg.1</td>
<td>chr5</td>
<td>OR2V1</td>
</tr>
<tr>
<td>uc002vlg.3</td>
<td>chr2</td>
<td>TNP1</td>
<td>uc009yhc.1</td>
<td>chr11</td>
<td>KCNC1</td>
<td>uc031asg.1</td>
<td>chr5</td>
<td>OR2V1</td>
</tr>
</tbody>
</table>
Table C.2: Common list of genes that demonstrate a significant change between Normal and Tumour profiles among Grade III cancer patients 386 and 873

<table>
<thead>
<tr>
<th>UCID</th>
<th>Chromosome</th>
<th>Gene.Symbol</th>
</tr>
</thead>
<tbody>
<tr>
<td>uc001exq.3</td>
<td>chr1</td>
<td>ZNF687</td>
</tr>
<tr>
<td>uc001fcb.3</td>
<td>chr1</td>
<td>S100A2</td>
</tr>
<tr>
<td>uc001fgb.3</td>
<td>chr1</td>
<td>CKS1B</td>
</tr>
<tr>
<td>uc001fqg.3</td>
<td>chr1</td>
<td>PRCC</td>
</tr>
<tr>
<td>uc001ftj.1</td>
<td>chr1</td>
<td>AIM2</td>
</tr>
<tr>
<td>uc002qvn.1</td>
<td>chr1</td>
<td>FAM72D</td>
</tr>
<tr>
<td>uc003xwn.3</td>
<td>chr8</td>
<td>VCPIP1</td>
</tr>
<tr>
<td>uc003xwo.2</td>
<td>chr8</td>
<td>C8orf44</td>
</tr>
<tr>
<td>uc021pai.1</td>
<td>chr1</td>
<td>MIR4258</td>
</tr>
<tr>
<td>uc031pqw.1</td>
<td>chr1</td>
<td>MIR5187</td>
</tr>
</tbody>
</table>

Table C.3: Genes that show similarity in tumour behaviour among the genes that change in Grade I cancer patients, 1357 and 4137

<table>
<thead>
<tr>
<th>UCID</th>
<th>Chromosome</th>
<th>Gene.Symbol</th>
</tr>
</thead>
<tbody>
<tr>
<td>uc001bls.1</td>
<td>chr1</td>
<td>CEP85</td>
</tr>
<tr>
<td>uc002iop.1</td>
<td>chr17</td>
<td>ABI3</td>
</tr>
<tr>
<td>uc003lev.2</td>
<td>chr5</td>
<td>NRG2</td>
</tr>
<tr>
<td>uc011bap.2</td>
<td>chr3</td>
<td>PRSS42</td>
</tr>
<tr>
<td>uc001bta.3</td>
<td>chr1</td>
<td>TINAGL1</td>
</tr>
<tr>
<td>uc002ipp.1</td>
<td>chr17</td>
<td>TAC1</td>
</tr>
<tr>
<td>uc003nkz.3</td>
<td>chr6</td>
<td>SERINC1</td>
</tr>
<tr>
<td>uc011mdm.2</td>
<td>chr9</td>
<td>MIR4295</td>
</tr>
<tr>
<td>uc001fni.4</td>
<td>chr1</td>
<td>MIR4677</td>
</tr>
<tr>
<td>uc002kqt.4</td>
<td>chr18</td>
<td>MIR4895</td>
</tr>
<tr>
<td>uc003pdj.3</td>
<td>chr6</td>
<td>MIR4900</td>
</tr>
<tr>
<td>uc011mdm.2</td>
<td>chr9</td>
<td>MIR4901</td>
</tr>
<tr>
<td>uc001ggi.4</td>
<td>chr1</td>
<td>MIR5102</td>
</tr>
<tr>
<td>uc002lih.1</td>
<td>chr18</td>
<td>MIR5103</td>
</tr>
<tr>
<td>uc003pyy.1</td>
<td>chr6</td>
<td>MIR5104</td>
</tr>
<tr>
<td>uc021ona.1</td>
<td>chr1</td>
<td>MIR5105</td>
</tr>
<tr>
<td>uc001loa.1</td>
<td>chr11</td>
<td>MIR5106</td>
</tr>
<tr>
<td>uc002opt.4</td>
<td>chr19</td>
<td>MIR5107</td>
</tr>
<tr>
<td>uc003xdj.3</td>
<td>chr8</td>
<td>MIR5108</td>
</tr>
<tr>
<td>uc001mnc.3</td>
<td>chr11</td>
<td>MIR5109</td>
</tr>
<tr>
<td>uc002pjn.2</td>
<td>chr19</td>
<td>MIR5110</td>
</tr>
<tr>
<td>uc003xdw.1</td>
<td>chr8</td>
<td>MIR5111</td>
</tr>
<tr>
<td>uc001nqn.2</td>
<td>chr11</td>
<td>MIR5112</td>
</tr>
<tr>
<td>uc002pwo.3</td>
<td>chr19</td>
<td>MIR5113</td>
</tr>
<tr>
<td>uc003xdw.1</td>
<td>chr8</td>
<td>MIR5114</td>
</tr>
<tr>
<td>uc001nto.2</td>
<td>chr11</td>
<td>MIR5115</td>
</tr>
<tr>
<td>uc002qjb.4</td>
<td>chr11</td>
<td>MIR5116</td>
</tr>
<tr>
<td>uc001qav.4</td>
<td>chr11</td>
<td>MIR5117</td>
</tr>
<tr>
<td>uc002qmj.3</td>
<td>chr19</td>
<td>MIR5118</td>
</tr>
<tr>
<td>uc001sam.3</td>
<td>chr12</td>
<td>MIR5119</td>
</tr>
<tr>
<td>uc001sib.4</td>
<td>chr12</td>
<td>MIR5120</td>
</tr>
<tr>
<td>uc001smk.3</td>
<td>chr12</td>
<td>MIR5121</td>
</tr>
<tr>
<td>uc001ulq.3</td>
<td>chr12</td>
<td>MIR5122</td>
</tr>
<tr>
<td>uc001utc.2</td>
<td>chr13</td>
<td>MIR5123</td>
</tr>
<tr>
<td>uc001utt.3</td>
<td>chr13</td>
<td>MIR5124</td>
</tr>
<tr>
<td>uc001vnk.3</td>
<td>chr13</td>
<td>MIR5125</td>
</tr>
<tr>
<td>uc001xsh.2</td>
<td>chr13</td>
<td>MIR5126</td>
</tr>
<tr>
<td>uc001zab.1</td>
<td>chr15</td>
<td>MIR5127</td>
</tr>
<tr>
<td>uc002ava.3</td>
<td>chr15</td>
<td>MIR5128</td>
</tr>
<tr>
<td>uc002avb.3</td>
<td>chr15</td>
<td>MIR5129</td>
</tr>
<tr>
<td>uc002bgm.3</td>
<td>chr15</td>
<td>MIR5130</td>
</tr>
<tr>
<td>uc002bnl.3</td>
<td>chr15</td>
<td>MIR5131</td>
</tr>
<tr>
<td>uc002dmm.3</td>
<td>chr16</td>
<td>MIR5132</td>
</tr>
<tr>
<td>uc002ecs.5</td>
<td>chr16</td>
<td>MIR5133</td>
</tr>
<tr>
<td>uc002ejn.3</td>
<td>chr16</td>
<td>MIR5134</td>
</tr>
<tr>
<td>uc002hyo.3</td>
<td>chr17</td>
<td>MIR5135</td>
</tr>
<tr>
<td>uc002ihq.3</td>
<td>chr17</td>
<td>MIR5136</td>
</tr>
<tr>
<td>uc002ihq.3</td>
<td>chr17</td>
<td>MIR5137</td>
</tr>
<tr>
<td>uc002ihq.3</td>
<td>chr17</td>
<td>MIR5138</td>
</tr>
</tbody>
</table>

C.1 KSHV-NGS Results

In this section, some supplementary results from the KSHV-NGS analysis are summarized. Specifically, the topmost four genes that demonstrate changes at the 24 hour time point are shown below.
(a) Responses (actual and variance stabilized) with segmentation results for gene, AMY2A in chromosome 1. The black and red line in the top and middle figures correspond to 0 hour and 24 hour responses respectively. The blue and red segments indicate regions of significance detected by iSeg with no biological cut-off and a biological cut-off of 1.

(b) Responses (actual and variance stabilized) with segmentation results for gene, ABCA10 in chromosome 17. The black and red line in the top and middle figures correspond to 0 hour and 24 hour responses respectively. The blue and red segments indicate regions of significance detected by iSeg with no biological cut-off and a biological cut-off of 1.

(c) Responses (actual and variance stabilized) with segmentation results for gene, MMP10 in chromosome 11. The black and red line in the top and middle figures correspond to 0 hour and 24 hour responses respectively. The blue and red segments indicate regions of significance detected by iSeg with no biological cut-off and a biological cut-off of 1.

(d) Responses (actual and variance stabilized) with segmentation results for gene, CFH in chromosome 1. The black and red line in the top and middle figures correspond to 0 hour and 24 hour responses respectively. The blue and red segments indicate regions of significance detected by iSeg with no biological cut-off and a biological cut-off of 1.

Figure C.1: The top four genes called by $\kappa_\eta$ as significantly changed in the KSHV-NGS analysis.
BIBLIOGRAPHY


[106] Elias Masry. The wavelet transform of stochastic processes with stationary increments and


Brani Vidakovic. Wavelet-based functional data analysis: theory, applications and ramifications.


BIOGRAPHICAL SKETCH

Senthil Balaji Girimurugan, affectionately known to many as BeeJay and Sentil, pursued his degree in Electrical Engineering when he was 17. While studying engineering, he grew fond of computer programming, theoretical mathematics and signal processing and was fascinated by their use in real-life situations. Rather than merely obtaining a practical overview of the discipline, Sentil was keen on understanding the mathematical foundations behind signals and systems. This eventually led him to acquiring a Master’s in Electrical Engineering and a Master’s in Mathematics with a focus on stochastic processes from Clemson University.

As a capstone to his academic curiosities, he intended to pursue a Doctoral degree and chose Statistics as his major at Florida State University. During his well-spent time at Florida State University, he enjoyed being an instructor and a researcher in the field. With the knowledge, experiences, and connections gained from Florida State University, he aspires to pursue a career in scientific research and teaching.

Aside from writing proofs, coding scripts and executing simulations, his interests are in open source projects, biking, fitness, cooking, music and travel.