”NON-PARAMETRIC STATISTICAL MODELS USING WAVELETS: THEORY AND METHODS”

A Dissertation
Presented to
The Academic Faculty

By

German A. Schnaidt Grez

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Industrial and Systems Engineering

Georgia Institute of Technology

May 2019

”NON-PARAMETRIC STATISTICAL MODELS USING WAVELETS: THEORY AND METHODS”

Approved by:

Dr. Brani Vidakovic, Advisor
School of Industrial and Systems Engineering
Georgia Institute of Technology

Dr. Dave Goldsman
School of Industrial and Systems Engineering
Georgia Institute of Technology

Dr. Justin Romberg
School of Electrical Engineering
Georgia Institute of Technology

Dr. Yao Xie
School of Industrial and Systems Engineering
Georgia Institute of Technology

Dr. Kamran Paynabar
School of Industrial and Systems Engineering
Georgia Institute of Technology

Date Approved: January 18, 2019

Anyone who has never made a mistake has never tried anything new.

Albert Einstein

To Valentina, Santiago, Olivia, Dominga and Sake; my never-ending sources of joy, gratefulness and strength. You are the light that is always there to show me the path.

ACKNOWLEDGEMENTS

To begin, I would like to express my deepest appreciation to my advisor, Dr. Brani Vidakovic for his constant support, kindness, and brilliance. He always went beyond his duties to make sure that me and my family were supported and fine. That, in addition to his brilliant academic guidance was definitely a crucial contribution for the completion of my research. I feel extremely fortunate to work under his advisement, and I am grateful for having the possibility of learning so much from him, not only as an advisor, but also as a mentor and a friend.

I also would like to express my deepest gratitude to my thesis committee: Prof. David Goldsman, Prof. Justin Romberg , Prof. Yao Xie and Prof. Kamran Paynabar; their insightful comments, questions and support were a significant contribution for the completion and quality improvement of my research project.

In addition, for the last four years, Georgia Institute of Technology and ISyE provided me with a comfortable and a nurturing environment to develop my curiosity, allowing me to grow intellectually, academically and professionally. My deepest regards and appreciation to Dr. Alan Erera, Ms. Amanda Ford, Dr. Dawn Strickland, Dr. Joel Sokol and all the faculty and staff members that make ISyE the outstanding school it is. I feel extremely proud of being part of this community of incredible people, and I cannot tell how much I value all the experiences and knowledge that I received during my stay here.

Also, my sincere appreciation to all my family members who gave me their unconditional support all along the way, and most of all, believed in me. You were the fuel that boosted my motivation and determination to make this happen. I really appreciate you being there along the way.

Last but not least, I would like to thank Valentina, Santiago, Olivia and Dominga (my wife and three amazing children) for bearing with me these past few years, and taking this leap of faith of moving away from our country and family to be with me and allow me to pursue my dreams. Without you, this achievement would have been impossible. Thank for your patience, your kindness and unconditional support. Thank you for your generosity and empathy, and thank you for believing in me. Certainly, you were the fuel that boosted my motivation and drive to work hard every day and keep the pace, keeping sight of the things that really matter in life.

Finally, my deepest thanks to CONICYT and the ”BECAS CHILE” program for providing me with financial support during part of my Ph.D. studies (CONICYT PFCHA/DOCTORADO BECAS CHILE/2017-72180070).

TABLE OF CONTENTS

Acknowledgments v

List of Tables xiv

List of Figures xviii

Chapter 1: Introduction and Motivation 1

1.1 Why the use of Wavelets? 1

1.1.1 Some Wavelet Preliminaries 2

1.1.2 Multiresolution Analysis 7

1.1.3 Generation of Local Bases via Wavelets 14

1.1.4 Regularity of Wavelets 18

1.1.5 Approximations and Characterizations of Functional Spaces using Wavelets 22

1.1.6 Daubechies-Lagarias Algorithm 25

1.1.7 Wavelets "Disbalance" Energy in Data 28

1.1.8 Discrete Wavelet Transformations 29

Chapter 2: An Empirical Approach to Survival Density Estimation for Randomly-Censored Data using Wavelets 34

2.1 Introduction 35

2.1.1 Overview of previous and current work in the area 37

2.1.2 About Periodic Wavelets 39

2.2 Survival Density Estimation for right-censored data using Periodized Wavelets . . . 42

2.2.1 Problem statement, assumptions and derivation of the estimator for a density $f(x)$. 42

2.2.2 Estimating $\hat{f}_{J^*}(x)$ in the case of partially observed data. 45

2.2.3 Statistical properties of the Partial Data Estimator assuming $G(y)$ is known . 51

2.2.4 Statistical properties for Partial Data Estimator assuming $G(y)$ unknown. . 53

2.3 Simulation Study 59

2.3.1 Simulation Results. 61

2.3.2 Remarks and comments. 61

2.4 Real Data application and comparison with other Estimators. 73

2.5 Conclusions and Discussion. 76

Chapter 3: Empirical Wavelet-based Estimation for Non-linear Additive Regression Models. 78

3.1 Introduction 79

3.2 Wavelet-based Estimation in Additive Regression Models 83

3.2.1 Problem statement and derivation of the Estimator 84

3.2.2 Asymptotic Properties of the Estimator 88

3.2.3 $\mathbb{L}_2$ Risk Analysis of the Estimator $\hat{f}_J(\mathbf{x})$ 97

3.2.4 Implementation illustration and considerations and comparison to other estimators. 101

3.2.5 Conclusions and Discussion 107

Chapter 4: Least Squares Wavelet-based Estimation for Additive Regression Models using Non Equally-Spaced Designs 109

4.1 Introduction 110

4.2 Wavelet-based Estimation in Additive Regression Models 112

4.3 A Least Squares approach for non-linear Additive model estimation using orthogonal wavelet basis 114

4.3.1 Least Squares problem formulation. 115

4.3.2 Strong consistency of the Linear Least Squares Estimator. 117

4.3.3 Convergence rate of the Wavelet-based Least Squares Estimator. 120

4.3.4 Optimal choice of Estimator parameters $J(n)$ and β_n. 122

4.3.5 Simulation Study 124

4.4 Practical Application of Wavelet based Least Squares Method 155

4.5 Conclusions and Discussion 162

Chapter 5: Bayesian Approach for Non-Linear Additive Regression Models using Conjugate $\mathcal{NIG}$ Structures 165

5.1 Introduction 166

5.2 Bayesian Extension of the Non-linear Additive Regression Problem Using Gaussian Conditional $\mathcal{NIG}$ Model. 170

5.2.1 Obtention of the posterior distribution $\pi(\mathbf{c}_J|\mathbf{y})$ 175

5.2.2 Connection between posterior distribution $\pi(\mathbf{c}_J|\mathbf{y})$ and the Multivariate $t-$ distribution 178

5.2.3 Obtention of the Bayes Estimator $\hat{\mathbf{c}}_J$ 179

5.2.4 Prior Parameter Selection 180

5.2.5 Bayesian Model Implementation 181

5.2.6 Iterative Solution of the Model via Backfitting 183

5.2.7 Simulation Results 185

5.3 Bayesian Estimation using a Mixture $\mathcal{NIG}$ Model. 199

5.3.1 Derivation of the Bayes Estimator and Shrinkage Rule 200

5.3.2 Selection of Hyper-parameters 203

5.4 Bayesian Estimation Using γ-Contaminated $\mathcal{NIG}$ Structures 206

5.4.1 Derivation of the Estimator and Point-Mass Shrinkage Rule 207

5.4.2 Elicitation of Hyper parameters 210

5.5 Simulation Study 213

5.5.1 Remarks and Comments 215

5.6 Conclusions 235

Chapter 6: Multiscale Correlation Analysis in the Wavelet Domain 237

6.1 Introduction 238

6.2 Scale-wise Representation of Sample Correlation via DWT 240

6.3 Some Interesting Correlation Relationships Between Signals and Properties of Wavelet Coefficients 246

6.3.1 Case 1: Perfect Correlation between $\mathbf{x}$ and $\mathbf{y}$ 246

6.3.2 Case 2: Perfect correlation between $\mathbf{x}$ and $\mathbf{y}$ at a particular multiresolution level j_0. 246

6.3.3 Case 3: Perfect correlation between $\mathbf{x}$ and $\mathbf{y}$ at a all multiresolution levels $0 \leq j \leq J-1$, and its translation into the original signal domain correlation. 256

6.3.4 Some Theoretical Properties of wavelet coefficients for stationary, finite energy processes. 257

6.4 Statistical Tests for Multi-scale Correlation in the Wavelet Domain Based on the Whitening Property of DWT . . . 265

6.4.1 Student Test for Normally Distributed Random Variables . . . 265

6.4.2 A Local Test Statistic Based on the Distributional Structure of Wavelet Coefficients for Stationary Sequences, Assuming Normality . . . 267

6.4.3 A Non-Parametric Significance Test Based on the Geometry of the Wavelet Coefficient Sequences. . . . 272

6.4.4 Simulation Study of the Probability of Type I Error for Uncorrelated Stationary Sequences. . . . 285

6.4.5 Simulation Study of the Probability of Type II Error for Correlated Stationary Sequences. . . . 294

6.5 Application Example: Monthly Temperatures Atlanta-Athens, GA. . . . 300

6.6 Conclusions . . . 307

Appendix A: Appendix Chapter 2 . . . 310

A.1 Derivation of the unbiased partial-data estimator. . . . 310

A.2 Proof of Lemma 2.2.1 . . . 314

A.3 Proof of Lemma 2.2.2 . . . 319

A.4 Proof of Lemma 2.2.3 . . . 320

Appendix B: Appendix Chapter 3 . . . 326

B.1 Proof of $\int_0^1 \phi_{jk}^{per}(x)dx = 2^{-\frac{j}{2}}$. . . . 326

B.2 Important results from Multivariate Taylor Series expansion. . . . 326

B.3 Consistency of the Kernel density estimator. . . . 329

B.4 Derivation of an upper bound for $\mathbb{E}\left[\left(\frac{Y\phi_{Jk}^{per}(X_l)}{h(\mathbf{X})}\right)^2\right]$. . . . 333

B.5 Asymptotic correlation between $\frac{Y_i \phi_{Jk}^{per}(X_{il})}{\hat{h}_n(\mathbf{X}_i)}$ and $\hat{\beta}_0$. 335

B.6 Asymptotic convergence of $Cov\left(\hat{c}_{Jk}^{(l)}, \hat{c}_{Js}^{(l)}\right)$. 338

B.7 Proof of Lemma 3.2.4. 340

B.8 Proof of Lemma 3.2.5. 348

B.9 Proof of Lemma 3.2.6. 353

Appendix C: Appendix Chapter 4 355

C.1 Previous Theorems and definitions 355

C.1.1 Theorem P1 (Pollard 1984) 355

C.1.2 Lemma G1 (Györfi et al. 2002) 356

C.1.3 Theorem G2 (Györfi et al. 2002) 356

C.1.4 Theorem G3 (Györfi et al. 2002) 357

C.1.5 Theorem G4 (Györfi et al. 2002) 358

C.1.6 Theorem P2 (Pollard 1984) 359

C.2 Proof of Theorem 4.3.1. 359

C.3 Proof of Lemma 4.3.2. 367

C.4 Proof of Lemma 4.3.3. 370

C.5 Proof of Theorem 4.3.2. 371

C.6 Proof of Lemma 4.3.4. 375

Appendix D: Appendix Chapter 6 377

D.1 Additional Results For Type I and II Error Simulation-Based Performance Studies . 377

D.1.1 Box Plots for Type I Error Simulation Study 377

D.1.2 Box Plots for Type II Error Simulation Study 385

References 399

Vita 400

LIST OF TABLES

1.1 Sobolev α_N^* and Hölder α_N regularity exponents of Daubechies' scaling functions. 22

2.1 AMSE results for Delta distribution with Partial data estimator. 62

2.2 AMSE results for Normal distribution with Partial data estimator. 62

2.3 AMSE results for Bimodal distribution with Partial data estimator. 62

2.4 AMSE results for Strata distribution with Partial data estimator. 62

2.5 AMSE results for Multimodal distribution with Partial data estimator. 62

3.1 Average computational times 107

4.1 RMSE results for Uniform distribution with $\sigma^2 = 0.25$ using Daubechies 4 wavelet filter. 129

4.2 RMSE results for Uniform distribution with $\sigma^2 = 0.75$ using Daubechies 4 wavelet filter. 129

4.3 RMSE results for Uniform distribution with $\sigma^2 = 0.25$ using Coiflets 24 wavelet filter. 130

4.4 RMSE results for Uniform distribution with $\sigma^2 = 0.75$ using Coiflets 24 wavelet filter. 130

4.5 RMSE results for $Beta(\frac{3}{2}, \frac{3}{2})$ distribution with $\sigma^2 = 0.25$ using Daubechies 4 wavelet filter. 143

4.6 RMSE results for $Beta(\frac{3}{2}, \frac{3}{2})$ distribution with $\sigma^2 = 0.75$ using Daubechies 4 wavelet filter. 143

4.7 RMSE results for $Beta(\frac{3}{2}, \frac{3}{2})$ distribution with $\sigma^2 = 0.25$ using Coiflets 24 wavelet filter. 143

4.8 RMSE results for $Beta(\frac{3}{2}, \frac{3}{2})$ distribution with $\sigma^2 = 0.75$ using Coiflets 24 wavelet filter. 143

4.9 Application Data Set characteristics, obtained from [58]. 156

4.10 Comparison results for RMSE for best methods in [58] and the Wavelet-based LS using 4 features. 158

4.11 Comparison results for RMSE for best methods in [58] and the Wavelet-based LS using 1 feature (AT). 159

4.12 Comparison results for RMSE for best methods in [58] and the Wavelet-based LS using 2 features (AT-V). 160

4.13 Comparison results for RMSE for best methods in [58] and the Wavelet-based LS using 3 features (AT-V-RH). 161

5.1 ARMSE comparison between Bayes estimator and Least Squares, for $B = 50$ replications, Daubechies 6 filter and $\sigma = 0.39$. 185

5.2 ARMSE comparison between Bayes estimator and Least Squares, for $B = 50$ replications, Daubechies 6 filter and $\sigma = 0.39$. 186

5.3 ARMSE comparison between Bayes estimator and Least Squares, for $B = 50$ replications, Daubechies 6 filter and $\sigma = 0.39$. 186

5.4 ARMSE comparison between Bayes estimator and Least Squares, for $B = 50$ replications, Daubechies 6 filter and $\sigma = 0.39$. 187

5.5 AMSE(standard deviation) results for Functions in the model, for $N = 1024$. In blue the minimum average MSE, in magenta, the corresponding minimum standard deviation of MSE. 215

5.6 AMSE(standard deviation) results for Functions in the model, for $N = 2048$. In blue the minimum average MSE, in magenta, the corresponding minimum standard deviation of MSE. 216

5.7 AMSE(standard deviation) results for Functions in the model, for $N = 4096$. In blue the minimum average MSE, in magenta, the corresponding minimum standard deviation of MSE. 217

6.1 Agreement table for local significance test 270

6.2 Proposed critical values for the count test statistic. 272

6.3 Proposed critical values for the condition number test statistic. 280

6.4 Obtained results for average $\hat{p}_{Test_m}$ for AR(1) with parameter $\phi = -0.9$. 287

6.5 Obtained results for average $\hat{p}_{Test_m}$ for AR(1) with parameter $\phi = -0.7$. 287

6.6 Obtained results for average $\hat{p}_{Test_m}$ for AR(1) with parameter $\phi = 0.9$. 287

6.7 Obtained results for average $\hat{p}_{Test_m}$ for MA(1) with parameter $\theta = 0.9$. 287

6.8 Obtained results for average $\hat{p}_{Test_m}$ for MA(1) with parameter $\theta = 0.7$. 288

6.9 Obtained results for average $\hat{p}_{Test_m}$ for MA(1) with parameter $\theta = -0.9$. 288

6.10 Obtained results for average $\hat{p}_{Test_m}$ for ARMA(1,1) with parameters $\phi = -0.8$, $\theta = 0.1$. 288

6.11 Obtained results for average $\hat{p}_{Test_m}$ for ARMA(1,1) with parameters $\phi = -0.9$, $\theta = 0.9$. 288

6.12 Obtained results for average $\hat{p}_{Test_m}$ for AR(1) with parameter $\phi = -0.9$, $\beta = 0.25$. 297

6.13 Obtained results for average $\hat{p}_{Test_m}$ for AR(1) with parameter $\phi = -0.7$, $\beta = 0.25$. 297

6.14 Obtained results for average $\hat{p}_{Test_m}$ for AR(1) with parameter $\phi = 0.9$, $\beta = 0.25$. . 297

6.15 Obtained results for average $\hat{p}_{Test_m}$ for MA(1) with parameter $\theta = 0.9$, $\beta = 0.25$. . 297

6.16 Obtained results for average $\hat{p}_{Test_m}$ for MA(1) with parameter $\theta = 0.7$, $\beta = 0.25$. . 298

6.17 Obtained results for average $\hat{p}_{Test_m}$ for MA(1) with parameter $\theta = -0.9$, $\beta = 0.25$. 298

6.18 Obtained results for average $\hat{p}_{Test_m}$ for ARMA(1,1) with parameters $\phi = -0.9$ and $\theta = 0.9$, $\beta = 0.25$. 298

6.19 Obtained results for average $\hat{p}_{Test_m}$ for ARMA(1,1) with parameters $\phi = -0.8$ and $\theta = 0.1$, $\beta = 0.25$. 298

6.20 Estimation results for standardized monthly averages temperatures Athens and Atlanta, GA. This results were obtained using the wavelet filter Symmlet 10. 303

LIST OF FIGURES

1.1 Critical Sampling in $\mathbb{R} \times \mathbb{R}^+$ half-plane ($a = 2^{-j}$ and $b = k\,2^{-j}$). 7

1.2 (a) ϕ and (b) ψ for a given filter h. 11

2.1 Estimate results for Delta distribution, $N = 100, 200, 500, 1000$ using Symmlet5. . 63

2.2 Estimate results for Normal distribution, $N = 100, 200, 500, 1000$ using Symmlet5. 64

2.3 Estimate results for Bimodal distribution, $N = 100, 200, 500, 1000$ using Symmlet5. 65

2.4 Estimate results for Strata distribution, $N = 100, 200, 500, 1000$ using Symmlet5. . 66

2.5 Estimate results for Multimodal distribution, $N = 100, 200, 500, 1000$ using Symmlet5. 67

2.6 Results for 95% empirical quantiles and average estimate for Delta distribution using Symmlet5.(a)-(d) correspond to the partial data approach (for $N = 100, 200, 500, 1000$, respectively). 68

2.7 Results for 95% empirical quantiles and average estimate for Normal distribution using Symmlet5.(a)-(d) correspond to the partial data approach (for $N = 100, 200, 500, 1000$, respectively). 69

2.8 Results for 95% empirical quantiles and average estimate for Bimodal distribution using Symmlet5.(a)-(d) correspond to the partial data approach (for $N = 100, 200, 500, 1000$, respectively). 70

2.9 Results for 95% empirical quantiles and average estimate for Strata distribution using Symmlet5.(a)-(d) correspond to the partial data approach (for $N = 100, 200, 500, 1000$, respectively). 71

2.10 Results for 95% empirical quantiles and average estimate for Multimodal distribution using Symmlet5.(a)-(d) correspond to the partial data approach (for $N = 100, 200, 500, 1000$, respectively). 72

2.11 (a) AMSE for baseline distributions. (b) Q-Q Plot for the density estimates for Bimodal Distribution, $N = 1000$, $x = 0.7$. 72

2.12 Results for the application of the data driven estimators in real datasets. (a) corresponds to Liver metastases data and (b) to marriage duration in the U.S. 75

3.1 Graphic representation of the testing functions for the Additive Model. 102

3.2 Functions estimation for $\mathcal{U}(0, 1)$ designs, for $n = 2048$ samples. In red, the estimated function values at each sample point using AMlet; In black-dashed lines, the actual function shape; In blue lines, the smoothed version of the function values using lowess smoother and the wavelet-based method. In green, the estimated functions via backfitting. 103

3.3 Functions estimation for $\mathcal{U}(0, 1)$ designs, for $n = 4096$ samples. In red, the estimated function values at each sample point using AMlet; In black-dashed lines, the actual function shape; In blue lines, the smoothed version of the function values using lowess smoother and the wavelet-based estimator. In green, the estimated functions via backfitting. 104

3.4 Functions estimation for $Beta(3, 3)$ design, $n = 4096$ samples. In red, the estimated function values at each sample point via AMlet; In black-dashed lines, the actual function shape; In blue lines, the smoothed version of the function values using lowess smoother. In green, the estimated functions via back-fitting. 105

3.5 Box plots for function recovery RMSE for (a) $n = 2048$ samples, and (b) $n = 4096$ samples using $\mathcal{U}(0, 1)$ design. 100 replications were used in the experiment. 106

4.1 Graphic representation of the testing functions for the Additive Model. 126

4.2 RMSE results for Uniform Design using Daubechies and Coiflets filter, for values of $\sigma^2 = 0.25 - 0.75$. 131

4.3 RMSE results for Uniform Design using Daubechies and Coiflets filter, for values of $\sigma^2 = 0.25 - 0.75$. 132

4.4 RMSE results for each function using Uniform Design and Coiflets 24 filter, for values of $\sigma^2 = 0.25\,,\ 0.75$. 133

4.5 Estimated $f_1(x)$ using Uniform Design and Coiflets filter. 134

4.6 Estimated $f_2(x)$ using Uniform Design and Coiflets filter. 135

4.7 Estimated $f_3(x)$ using Uniform Design and Coiflets filter. 136

4.8 Estimated $f_4(x)$ using Uniform Design and Coiflets filter. 137

4.9 Estimated $f_5(x)$ using Uniform Design and Coiflets filter. 138

4.10 Estimated $f_6(x)$ using Uniform Design and Coiflets filter. 139

4.11 Estimated $f_7(x)$ using Uniform Design and Coiflets filter. 140

4.12 Estimated $f_8(x)$ using Uniform Design and Coiflets filter. 141

4.13 Estimated $f_9(x)$ using Uniform Design and Coiflets filter. 142

4.14 RMSE results for Beta Design using Daubechies filters, for values of $\sigma^2 = 0.25\,,\ 0.75$..144

4.15 RMSE results for each function using Coiflets 24 filter, for values of $\sigma^2 = 0.25\,,\ 0.75$..145

4.16 Estimated $f_1(x)$ using Beta Design and Coiflets filter. 146

4.17 Estimated $f_2(x)$ using Beta Design and Coiflets filter. 147

4.18 Estimated $f_3(x)$ using Beta Design and Coiflets filter. 148

4.19 Estimated $f_4(x)$ using Beta Design and Coiflets filter. 149

4.20 Estimated $f_5(x)$ using Beta Design and Coiflets filter. 150

4.21 Estimated $f_6(x)$ using Beta Design and Coiflets filter. 151

4.22 Estimated $f_7(x)$ using Beta Design and Coiflets filter. 152

4.23 Estimated $f_8(x)$ using Beta Design and Coiflets filter. 153

4.24 Estimated $f_9(x)$ using Beta Design and Coiflets filter. 154

4.25 Estimaion result plots over the 95% empirical quantiles region and RMSE (computed using the standardized predictions) obtained over 100 replications. 158

4.26 Estimated $f_1(x)$ and $f_2(x)$ over the 95% empirical quantiles region. The bottom panel illustrates the sample histograms for each considered feature, within the 95% empirical quantiles region. 159

4.27 Estimated $f_3(x)$ and $f_4(x)$ over the 95% empirical quantiles region. The bottom panel illustrates the sample histograms for each considered feature, within the 95% empirical quantiles region. 160

5.1 Estimation result box-plots for the $\log_{10}(ARMSE)$ computed for both Bayes and least squares procedures, using $B = 50$ replications, for each of the testing functions $f_1(x), ..., f_9(x)$. 187

5.2 Estimation result box-plots for the $\log_{10}(ARMSE)$ computed for both Bayes and least squares procedures, using $B = 50$ replications, for each of the testing functions $f_1(x), ..., f_9(x)$. 189

5.3 Estimated function $f_1(x)$ for $N = 1024, 4096$ samples. 190

5.4 Estimated function $f_2(x)$ for $N = 1024, 4096$ samples. 191

5.5 Estimated function $f_3(x)$ for $N = 1024, 4096$ samples. 192

5.6 Estimated function $f_4(x)$ for $N = 1024, 4096$ samples. 193

5.7 Estimated function $f_5(x)$ for $N = 1024, 4096$ samples. 194

5.8 Estimated function $f_6(x)$ for $N = 1024, 4096$ samples. 195

5.9 Estimated function $f_7(x)$ for $N = 1024, 4096$ samples. 196

5.10 Estimated function $f_8(x)$ for $N = 1024, 4096$ samples. 197

5.11 Estimated function $f_9(x)$ for $N = 1024, 4096$ samples. 198

5.12 Functions used in the simulated additive model. 214

5.13 Typical estimated function Blocks for $N = 512,\ 1024$ samples. 219

5.14 Typical estimated function Blocks for $N = 2048,\ 4096,\ 8192$ samples. 220

5.15 Typical estimated function Bumps for $N = 512,\ 1024$ samples. 221

5.16 Typical estimated function Bumps for $N = 2048,\ 4096,\ 8192$ samples. 222

5.17 Typical estimated function Heavisine for $N = 512,\ 1024$ samples. 223

5.18 Typical estimated function Heavisine for $N = 2048,\ 4096,\ 8192$ samples. 224

5.19 Typical estimated function Zero for $N = 512,\ 1024$ samples. 225

5.20 Typical estimated function Zero for $N = 2048,\ 4096,\ 8192$ samples. 226

5.21 Empirical MSE for Blocks on a $\log_{10}$ scale. In each panel from left to right: Bayes Total, Bayes Mixture, Bayes Point Mass, Least Squares, Amlet. 227

5.22 Empirical MSE for Blocks on a $\log_{10}$ scale. In each panel from left to right: Bayes Total, Bayes Mixture, Bayes Point Mass, Least Squares, Amlet. 228

5.23 Empirical MSE for Bumps on a $\log_{10}$ scale. In each panel from left to right: Bayes Total, Bayes Mixture, Bayes Point Mass, Least Squares, Amlet. 229

5.24 Empirical MSE for Bumps on a $\log_{10}$ scale. In each panel from left to right: Bayes Total, Bayes Mixture, Bayes Point Mass, Least Squares, Amlet. 230

5.25 Empirical MSE for Heavisine on a $\log_{10}$ scale. In each panel from left to right: Bayes Total, Bayes Mixture, Bayes Point Mass, Least Squares, Amlet. 231

5.26 Empirical MSE for Heavisine on a $\log_{10}$ scale. In each panel from left to right: Bayes Total, Bayes Mixture, Bayes Point Mass, Least Squares, Amlet. 232

5.27 Empirical MSE for Zero on a $\log_{10}$ scale. In each panel from left to right: Bayes Total, Bayes Mixture, Bayes Point Mass, Least Squares, Amlet. 233

5.28 Empirical MSE for Zero on a $\log_{10}$ scale. In each panel from left to right: Bayes Total, Bayes Mixture, Bayes Point Mass, Least Squares, Amlet. 234

6.1 Plot of the correlation weights w_j for different values of $C_j > 0$. Each of the colored lines represent the shape of $w_j(\beta, C)$ for a fixed value of C. In the plot, C ranges from 0 to 100. The larger C, the smaller the slope of the curve around zero, and the slower it reaches the asymptotic value of 1. 247

6.2 Comparative Plots of uncorrelated (a) vs. correlated (b) signals in the wavelet domain. 250

6.3 Comparative Plots of uncorrelated (a) vs. correlated (b) signals in the time domain, with perfect correlation at scale-level 7. 251

6.4 Comparative boxplots of the typical effects on overall correlation at the original domain, given perfect correlation at the wavelet domain. The experiments were replicated 1000 times. 252

6.5 Comparative boxplots of the typical effects on overall correlation at the original domain, given perfect correlation at the wavelet domain. The experiments were replicated 1000 times. 253

6.6 Comparative boxplots of the typical effects on scale correlations in the wavelet domain, given perfect correlation at each scale level. The experiments were replicated 1000 times. 254

6.7 Comparative boxplots of the typical effects on scale correlations in the wavelet domain, given perfect correlation at each scale level. The experiments were replicated 1000 times. 255

6.8 Comparative box-plots of the typical effects on scale correlations in the wavelet domain and time domain, given perfect correlation at each scale level. (a) illustrates the usual sample correlation in the time domain, (b) shows the corresponding p-values for the test-statistic. The experiments were replicated 1000 times. Values for β_j were chose at random. 264

6.9 Plot of $t-$distribution for 2^j degrees of freedom, $j = 1, ..., J-1$. Note that $\nu > 30$ the distribution closely approximates to a $\mathcal{N}(0, 1)$ distribution. 266

6.10 Typical Histograms of the entries of Table 6.3 for an MA(1) process with parameter $\theta = 0.9$. The experiments were replicated 50000 times. Similar behavior were observed for for the rest of detail levels fir AR(1), WN and ARMA(1,1) model, with no significant differences for the empirical quantiles. 281

6.11 Typical Histograms of the entries of Table 6.3 for an MA(1) process with parameter $\theta = 0.9$. The experiments were replicated 50000 times. Similar behavior were observed for for the rest of detail levels fir AR(1), WN and ARMA(1,1) model, with no significant differences for the empirical quantiles. 282

6.12 Plot of the condition number $\kappa(\widehat{\mathbf{\Sigma}}_j)$ as a function of $\hat{\rho}_{XY}^{(j)}$. 282

6.13 Typical Histograms of the level-wise correlations for uncorrelated sequences of an AR(1) process with $\theta < 0$. The experiments were replicated 50000 times. Similar behavior were observed for MA(1) and WN processes, with no significant differences for the type of decay and empirical quantiles. 283

6.14 Typical Histograms of the level-wise correlations for uncorrelated sequences of an AR(1) process with $\theta < 0$. The experiments were replicated 50000 times. Similar behavior were observed for MA(1) and WN processes, with no significant differences for the type of decay and empirical quantiles. 284

6.15 Summary plot of average type I error probability for each test statistic 290

6.16 Scatter plots of typical AR(1) with high oscillatory behavior and their respective wavelet coefficients generated from orthogonal DWT using Symmlet 10. The red lines indicate the separation between consecutive scale levels, arranged in an increasing order. 291

6.17 Scatter plots of typical MA(1) with high oscillatory behavior and their respective wavelet coefficients generated from orthogonal DWT using Symmlet 10. The red lines indicate the separation between consecutive scale levels, arranged in an increasing order. 292

6.18 293

6.19 Summary plot of average type II error probability for each test statistic 299

6.21 Monthly averages and standard deviations of temperatures for both Atlanta (solid blue) and Athens (dashed blue), computed across the samples shown in Fig. 6.20a. 301

6.23 Scatter plots of the wavelet coefficients for each scale, corresponding to the DWT of the Athens (x-axis) and Atlanta (y-axis) daily average temperatures. This results were obtained using the wavelet filter Symmlet 10. 305

6.24 Scatter plots of the wavelet coefficients for each scale, corresponding to the DWT of the Athens (x-axis) and Atlanta (y-axis) daily average temperatures. This results were obtained using the wavelet filter Symmlet 10. 306

D.1 Box plots of $\hat{p}_{Test_m}$ for AR(1) with parameter $\phi = -0.9$. For scale levels $J = 4$ to $J = 7$, most of tests remains within the expected 5% type I error; however, for $J = 8$ the tests $C2$, Kendall, Spearman's and T-test exhibit a significantly large deviation from the expected error, with an average of approximately 38%. This implies that on average, for this kind of stochastic processes, wavelet coefficient corresponding to short time scales depart from normality and exhibit heavier tails, which causes an artificial inflation of the likelihood of a false positive classification. 377

D.2 Box plots of $\hat{p}_{Test_m}$ for AR(1) with parameter $\phi = -0.7$. For scale levels $J = 4$ to $J = 7$, most of tests remains within the expected 5% type I error; however, for $J = 8$ the tests $C2$, Kendall, Spearman's and T-test exhibit a significantly large deviation from the expected error, with an average of approximately 14%. This implies that on average, for this kind of stochastic processes, wavelet coefficient corresponding to short time scales depart from normality and exhibit heavier tails, which causes an artificial inflation of the likelihood of a false positive classification. 378

D.3 Box plots of $\hat{p}_{Test_m}$ for AR(1) with parameter $\phi = 0.9$. For all scale levels most of tests remains within the expected 5% type I error. 379

D.4 Box plots of $\hat{p}_{Test_m}$ for MA(1) with parameter $\theta = 0.9$. For scale levels $J = 4$ to $J = 7$, most of tests remains within the expected 5% type I error. However, for $J = 8$, the tests $C2$, Kendall, Spearman's and T-test exhibit a significantly large deviation from the expected error, with an average of approximately 13%. 380

D.5 Box plots of $\hat{p}_{Test_m}$ for MA(1) with parameter $\theta = 0.7$. For scale levels $J = 4$ to $J = 7$, most of tests remains within the expected 5% type I error. However, for $J = 8$, the tests $C2$, Kendall, Spearman's and T-test exhibit a significantly large deviation from the expected error, with an average of approximately 11%. 381

D.6 Box plots of $\hat{p}_{Test_m}$ for MA(1) with parameter $\theta = -0.9$. For all scale levels, most of tests remains within the expected 5% type I error. In particular, the tests $C2$, Kendall, Spearman's and T-test exhibit a slight deviation from the expected error, with an average of approximately 6.5%. 382

D.7 Box plots of $\hat{p}_{Test_m}$ for ARMA(1,1) with parameters $\phi = -0.8$, $\theta = 0.1$. For scale levels $J = 4$ to $J = 7$, most of tests remains within the expected 5% type I error; however, for $J = 8$ the tests $C2$, Kendall, Spearman's and T-test exhibit a significantly large deviation from the expected error. This implies that on average, for this kind of stochastic processes, wavelet coefficient corresponding to short time scales depart from normality, which causes an artificial inflation of the likelihood of a false positive classification. 383

D.8 Box plots of $\hat{p}_{Test_m}$ for ARMA(1,1) with parameters $\phi = -0.9$, $\theta = 0.9$. For scale levels $J = 4$ to $J = 7$, most of tests remains within the expected 5% type I error. . . 384

D.9 Box plots of $\hat{p}_{Test_m}$ (average probability of type II error) for AR(1) with parameter $\phi = -0.9$. For scale levels $J = 5$ to $J = 8$, most of tests remains within the 5% average type II error; however, for $J = 4$ the tests Kendall and Spearman's exhibit a significantly large deviation from the expected error, with an average of approximately 24%. Also, the observed performance of the Condition number test can be considered as good as the statistical tests used as benchmark. 385

D.10 Box plots of $\hat{p}_{Test_m}$ (average probability of type II error) for AR(1) with parameter $\phi = -0.7$. For scale levels $J = 5$ to $J = 8$, most of tests remains within the 5% average type II error; however, for $J = 4$ the tests Kendall and Spearman's exhibit a significantly large deviation from the expected error, with an average of approximately 24%. Also, the observed performance of the Condition number test can be considered as good as the statistical tests used as benchmark. 386

D.11 Box plots of $\hat{p}_{Test_m}$ (average probability of type II error) for AR(1) with parameter $\phi = 0.9$. For scale levels $J = 5$ to $J = 8$, most of tests remains within the 5% average type II error; however, for $J = 4$ the tests Kendall and Spearman's exhibit a significantly large deviation from the expected error, with an average of approximately 24%. Also, the observed performance of the Condition number test can be considered as good as the statistical tests used as benchmark. 387

D.12 Box plots of $\hat{p}_{Test_m}$ (average probability of type II error) for MA(1) with parameter $\theta = -0.9$. For scale levels $J = 5$ to $J = 8$, most of tests remains within the 5% average type II error; however, for $J = 4$ the tests Kendall and Spearman's exhibit a significantly large deviation from the expected error, with an average of approximately 24%. Also, the observed performance of the Condition number test can be considered as good as the statistical tests used as benchmark. 388

D.13 Box plots of $\hat{p}_{Test_m}$ (average probability of type II error) for MA(1) with parameter $\theta = 0.9$. For scale levels $J = 5$ to $J = 8$, most of tests remains within the 5% average type II error; however, for $J = 4$ the tests Kendall and Spearman's exhibit a significantly large deviation from the expected error, with an average of approximately 24%. Also, the observed performance of the Condition number test can be considered as good as the statistical tests used as benchmark. 389

D.14 Box plots of $\hat{p}_{Test_m}$ (average probability of type II error) for MA(1) with parameter $\theta = 0.7$. For scale levels $J = 5$ to $J = 8$, most of tests remains within the 5% average type II error; however, for $J = 4$ the tests Kendall and Spearman's exhibit a significantly large deviation from the expected error, with an average of approximately 24%. Also, the observed performance of the Condition number test can be considered as good as the statistical tests used as benchmark. 390

D.15 Box plots of $\hat{p}_{Test_m}$ (average probability of type II error) for ARMA(1) with parameters $\phi = -0.8$ $\theta = 0.1$. Induced linear relationship given by $\beta = 0.25$. 391

D.16 Box plots of $\hat{p}_{Test_m}$ (average probability of type II error) for ARMA(1) with parameters $\phi = -0.9$ $\theta = 0.9$. Induced linear relationship given by $\beta = 0.25$. 391

SUMMARY

Machine Learning and Data Analytics have become key tools in the advancement of modern society, with a vast variety of applications exhibiting exponential growth in breadth and depth during the past few years. Moreover, the advancement of data-gathering technologies enables the availability of massive amounts of data, which fuels the opportunity for the application and development of new analytics tools to obtain insights and discover patterns, relationships or structures that otherwise wouldn't be possible to identify, thus making an effective and efficient use of it.

This dissertation aims to contribute to the scientific existing methodologies in this context, with focus in the non-parametric statistical domain due to its robustness to prior modeling assumptions and flexibility of application in many different contexts.

In light of this objective, four non-parametric techniques based on wavelets are introduced and analyzed. Applications such as survival density estimation, non-linear additive regression and multiscale correlation analysis are covered, and each topic is studied from both a theoretical and pragmatic perspectives. In fact, a theoretical foundation for each proposed method is developed, and then its applicability is illustrated using simulations studies and real data sets.

This Thesis is structured in six Chapters, each containing the following topics:

In Chapter 1, the motivation for the use of wavelets is provided, and general definitions and results involving their use in statistics are introduced. This aims to the construction of a brief theoretical foundation over which the methodologies introduced in the subsequent Chapters are built upon.

In Chapter 2, the density estimation problem is studied. A non-parametric estimator for probability densities in the presence of randomly censored data is introduced, and their statistical properties are analyzed. In particular, a linear density estimator using empirical wavelet coefficients that are fully data-driven is proposed. This estimator is shown to be asymptotically unbiased, with global mean square consistency. In addition, it performance is evaluated using different exemplary

distributions, with different sample sizes and censoring schemes. On top of that, some implementation recommendations and remarks are provided, providing guidance to future practitioners interested in applying the proposed technique.

In Chapter 3 the problem of non-parametric regression for additive models is investigated, introducing a novel approach using orthogonal projections onto linear functional subspaces. These regression models are useful in the analysis of responses determined by non-linear relationships with multivariate predictors, which provides more flexibility and generality that the traditional multi-dimensional linear regression model. For this purpose, a mean-square consistent estimator based on an orthogonal projection onto a multiresolution space using empirical wavelet coefficients is proposed, and its convergence rates are analyzed when the set of unknown functions can be characterized by a known smoothness index. These results are obtained without the assumption of an equispaced design, a condition that is typically assumed in most wavelet-based procedures. In addition, some qualitative comparison with existing methodologies is provided, illustrating the potential estimation capabilities of the proposed methodology.

In Chapter 4, a the additive regression problem is analyzed from a different viewpoint: the classic least squares solutions using an orthogonal wavelet basis is proposed and its theoretical properties are analyzed. This estimation methodology is based on periodic orthogonal wavelets on the interval [0,1]. A strongly consistent estimator (with respect to the $\mathbb{L}_2$ norm) is introduced, leading to optimal convergence rates up to a logarithmic factor, independent of the dimensionality of the problem. Similarly as in the previous Chapter, these results are obtained without the assumption of an equispaced design for the predictors, which shows the flexibility of wavelets for statistical applications and the power of the least squares approach. This theoretical study is further complemented with a simulation experiment and the application of the method on a real-life data set, enabling the comparison of the proposed methodology with several machine learning algorithms in a real-life scenario.

In Chapter 5, an alternative approach for the non-linear additive regression problem using

Bayesian hierarchical Normal-Inverse-Gamma ($\mathcal{NIG}$) structures is introduced. First, a robust and simple approach that reduces to an l_2-regularized regression model is proposed and implemented. The theoretical derivations of the estimator and predictive distribution are provided, and the hyper-parameter selection is discussed. Furthermore, an implementation algorithm based on a backfitting approach is proposed and its performance is studied via simulation. Secondly, this model is extended to a mixture of $\mathcal{NIG}$ in the expansion coefficients, improving the capacity of the model to adapt to different degrees of smoothness of the unknown functions. Closed form solutions for the Bayes estimator are derived and its structure is discussed. Next, a special case of the previous model is analyzed: a point-mass contaminated $\mathcal{NIG}$ model. This modeling structure aims to enforce a more sparse estimation of the functions in the model, thus providing a more adaptive methodology for irregular functions. Finally, the applicability of these methods is illustrated via a simulation study, and its performance is compared to the least squares approach, the simple $\mathcal{NIG}$ model initially introduced and a method denominated *AMlet*, introduced by Sardy and Tseng (2004)[1]. The obtained results suggest that the Bayesian approach based on $\mathcal{NIG}$ models tends to outperform most of previously existing methods, and is very flexible to implement.

In Chapter 6, the correlation analysis problem is studied from a multiscale perspective via the application of Discrete Wavelet transformations (DWT). A systematic methodology that uses the linearity and orthogonality of the DWT is used to decompose a sample correlation into a weighted sum of scale-wise correlations that have a special additive structure and enable the extraction of information about possible linear relationships at different scale resolutions that are otherwise hidden. In addition, some of its theoretical properties are analyzed, and a non-parametric test is proposed for the assessment of the statistical significance of the scale-wise correlations. This is further complemented by simulation-based performance study and an application use case that analyzes monthly average temperatures between the cities of Atlanta and Athens, GA.

CHAPTER 1
INTRODUCTION AND MOTIVATION

1.1 Why the use of Wavelets?

Wavelets are mathematical tools that have interpretation and application in many scientific fields, such as non-parametric function estimation, signal processing and data compression. In the early 1990s, a seminal research by Donoho and Johnstone[2] and their coauthors made the connection between Wavelets and statistical models, showing that wavelets are appropriate tools to tackle problems such as denoising, regression, and density estimation.

Due to its mathematical properties, Wavelets provide a rich source of useful tools for applications in "time-scale" types of problems. In particular, the wavelet representations enable to represent a time-domain evolution in terms of scale components. In this context, it is possible to find many similarities between wavelet transformations and Fourier transformations. Fourier transformation extract details from the signal frequency, however, the location of a particular frequency within the signal is not captured. This can be obtained by windowing the signal, and then by taking its Fourier transform. The problem with this approach is that the portions of the processed signal are of a fixed length (determined by the window size), which may lead to a local under- or over-fitting. This results from the fact that windows of the same length are used to resolve both high and low frequency components, which in the case of non-stationary signals is particularly inadequate.

For these reasons, statistical multiscale modeling based on wavelets has become a popular area in both theoretical and applied statistics. Wavelet based methods have been under development in areas such as regression, density and function estimation, factor analysis, modeling and forecasting

in time series analysis, and spatial statistics.

In addition, wavelets provide alternative orthonormal bases in a variety statistical problems. Even in the cases in which the traditional orthogonal series are simply replaced by Wavelet bases, the literature has shown that Wavelets often offer better localization and parsimony, leading to better estimation results.

In the next sections, some important definitions and results generated by wavelets are provided with the aim of providing a foundation for the wavelet-based tools that are applied in the subsequent chapters of this Thesis. For a more complete and detailed treatment of the wavelets in statistics, the reader could refer to the book by Vidakovic (1999) [3], the monograph by Antoniadis (1997) [4], and the work by Daubechies (1992)[5]. Most of the material presented in the following sections was obtained from [3], and was used with permission from the authors.

1.1.1 Some Wavelet Preliminaries

The first theoretical results in wavelets are connected with continuous wavelet decompositions of $\mathbb{L}_2$ functions and go back to the early 1980s. Papers of Morlet *et al.* (1982)[6] and Grossmann and Morlet (1985)[7] were among the first on this subject.

Let $\psi_{a,b}(x),\ a \in \mathbb{R}\backslash\{0\}, b \in \mathbb{R}$ be a family of functions defined as translations and re-scales of a single function $\psi(x) \in \mathbb{L}_2(\mathbb{R})$, as follows:

$$\psi_{a,b}(x) = \frac{1}{\sqrt{|a|}}\psi\left(\frac{x-b}{a}\right). \tag{1.1}$$

The normalization by $\frac{1}{\sqrt{|a|}}$ ensures that $||\psi_{a,b}(x)||_{\mathbb{L}_2}$ is independent of a and b. The function ψ

(called *the wavelet function* or *the mother wavelet*) is assumed to satisfy the *admissibility condition,*

$$C_\psi = \int_{\mathbb{R}} \frac{|\Psi(\omega)|^2}{|\omega|} d\omega < \infty, \tag{1.2}$$

where $\Psi(\omega) = \int_R \psi(x) e^{-ix\omega} dx$ is the Fourier transformation of $\psi(x)$. The admissibility condition (1.2) implies:

$$0 = \Psi(0) = \int \psi(x) dx.$$

Also, if $\int \psi(x)dx = 0$ and $\int (1 + |x|^\alpha)|\psi(x)|dx < \infty$ for some $\alpha > 0$, then $C_\psi < \infty$.

Wavelet functions are usually normalized to "have unit energy", i.e., $||\psi_{a,b}(x)|| = 1$.

For any $\mathbb{L}_2$ function $f(x)$, the continuous wavelet transformation is defined as a function of two variables:

$$\mathcal{CWT}_f(a, b) = \langle f, \psi_{a,b} \rangle = \int f(x) \overline{\psi_{a,b}(x)} dx. \tag{1.3}$$

Here, the dilation and translation parameters, a and b, respectively, vary continuously over $\mathbb{R}\backslash\{0\} \times \mathbb{R}$.

Resolution of Identity. When the admissibility condition is satisfied, i.e., $C_\psi < \infty$, it is possible to find the inverse continuous transformation via the relation known as *resolution of identity* or *Calderón's reproducing identity,* which is given by:

$$f(x) = \frac{1}{C_\psi} \int_{\mathbb{R}^2} \mathcal{CWT}_f(a, b) \psi_{a,b}(x) \frac{da\ db}{a^2}.$$

If a is restricted to $\mathbb{R}^+$, which is natural since a can be interpreted as a reciprocal of frequency,

(1.2) becomes:

$$C_\psi = \int_0^\infty \frac{|\Psi(\omega)|^2}{\omega} d\omega < \infty, \tag{1.4}$$

and the *resolution of identity* relation takes the form

$$f(x) = \frac{1}{C_\psi} \int_{-\infty}^{\infty} \int_0^\infty \mathcal{CWT}_f(a,b)\psi_{a,b}(x) \frac{1}{a^2} da\, db. \tag{1.5}$$

Next, we list a few important properties of continuous wavelet transformations.

Shifting Property. If $f(x)$ has a continuous wavelet transformation $\mathcal{CWT}_f(a,b)$, then $g(x) = f(x-\beta)$ has the continuous wavelet transformation $\mathcal{CWT}_g(a,b) = \mathcal{CWT}_f(a, b-\beta)$.

Scaling Property. If $f(x)$ has a continuous wavelet transformation $\mathcal{CWT}_f(a,b)$, then $g(x) = \frac{1}{\sqrt{s}}\, f\left(\frac{x}{s}\right)$ has the continuous wavelet transformation $\mathcal{CWT}_g(a,b) = \mathcal{CWT}_f\left(\frac{a}{s}, \frac{b}{s}\right)$.

Note that both the shifting property and the scaling property result from changing variables under the integral sign in Eq. (1.3).

Energy Conservation. From (1.5),

$$\int_{-\infty}^{\infty} |f(x)|^2 dx = \frac{1}{C_\psi} \int_{-\infty}^{\infty} \int_0^\infty |\mathcal{CWT}_f(a,b)|^2 \frac{1}{a^2} da\, db.$$

Localization. Let $f(x) = \delta(x - x_0)$ be the Dirac pulse at the point x_0. Then,

$$\mathcal{CWT}_f(a, b) = \frac{1}{\sqrt{a}}\psi\left(\frac{x_0 - b}{a}\right).$$

Reproducing Kernel Property. Define $\mathbb{K}(u, v; a, b) = \langle\psi_{u,v}, \psi_{a,b}\rangle_{\mathbb{L}_2}$. Then, if $F(u, v)$ is a continuous wavelet transformation of $f(x)$,

$$F(u, v) = \frac{1}{C_\psi}\int_{-\infty}^{\infty}\int_{0}^{\infty}\mathbb{K}(u, v; a, b)F(a, b)\frac{1}{a^2}da\ db,$$

i.e., $\mathbb{K}$ is a reproducing kernel. The corresponding reproducing kernel Hilbert space (RKHS) is defined as a $\mathcal{CWT}$ image of $\mathbb{L}_2(\mathbb{R})$ – the space of all complex-valued functions F on $\mathbb{R}^2$ for which $\frac{1}{C_\psi}\int_{-\infty}^{\infty}\int_{0}^{\infty}|F(a, b)|^2\frac{da\ db}{a^2}$ is finite.

Characterization of Regularity. Let $\int(1 + |x|)\,|\psi(x)|\,dx < \infty$ and let $\Psi(0) = 0$.
If $f \in \mathbb{C}^\alpha$ (Hölder space with exponent α), then, it follows:

$$|\mathcal{CWT}_f(a, b)| \leq C|a|^{\alpha+1/2}. \tag{1.6}$$

Conversely, if a continuous and bounded function f satisfies (1.6), then $f \in \mathbb{C}^{\alpha\cdot}$ Examples of wavelets having a compact support and an arbitrarily great regularity r have been constructed by Daubechies (See Daubechies 1988[8])..

Discretization of the Wavelet Transformation. The continuous wavelet transformation of a function of one variable results in a function of two variables a, b. Since the transformation is redundant, it is possible to select discrete values of a and b and still have a transformation that is invertible. However, sampling that preserves all information about the decomposed function cannot be coarser than the *critical sampling.*

The critical sampling (see Fig. 1.1) is defined by:

$$a = 2^{-j},\ b = k2^{-j},\ j, k \in \mathbb{Z}. \tag{1.7}$$

This choice of parameters, will result in a minimal basis. Any coarser sampling will not give a unique inverse transformation, meaning that the original function will not be uniquely recoverable. Moreover, under mild conditions on the wavelet function ψ, these sampling scheme generates an orthogonal basis:

$$\{\psi_{jk}(x) = 2^{j/2}\psi(2^j x - k),\ j, k \in \mathbb{Z}\}.$$

Nonetheless, it is possible to select different discretization alternatives. For example, selecting $a = 2^{-j},\ b = k$ will lead to non-decimated (or stationary) wavelets. A more general sampling, is given by:

$$a = a_0^{-j},\ b = k\, b_0\, a_0^{-j},\ j, k \in \mathbb{Z},\ a_0 > 1, b_0 > 0. \tag{1.8}$$

Here, numerically stable reconstructions are possible if the system $\{\psi_{jk},\ j, k \in \mathbb{Z}\}$ constitutes a frame. Here:

$$\psi_{jk}(x) = a_0^{j/2}\,\psi\left(\frac{x - k\, b_0\, a_0^{-j}}{a_0^{-j}}\right) = a_0^{j/2}\psi(a_0^j x - k\, b_0),$$

corresponds to (1.1) evaluated at (1.8).

In the next section, we consider wavelet transformations (wavelet series expansions) for values of a and b given by (1.7). An elegant theoretical framework for critically sampled wavelet transformation is *Mallat's Multiresolution Analysis* (Mallat, 87; 89a, 89b, 98).

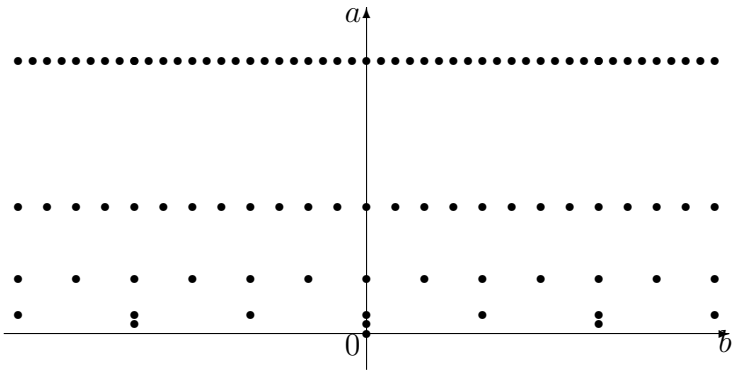


Figure 1.1: Critical Sampling in $\mathbb{R} \times \mathbb{R}^+$ half-plane ($a = 2^{-j}$ and $b = k\ 2^{-j}$).

1.1.2 Multiresolution Analysis

A multiresolution analysis (MRA) is a sequence of closed subspaces $V_n, n \in \mathbb{Z}$ in $\mathbb{L}_2(\mathbb{R})$ such that they lie in a containment hierarchy, as follows:

$$\cdots \subset V_{-2} \subset V_{-1} \subset V_0 \subset V_1 \subset V_2 \subset \cdots . \tag{1.9}$$

The nested spaces have an intersection that contains the zero function only and a union that is dense in $\mathbb{L}(\mathbb{R})$, therefore:

$$\cap_n V_j = \{\mathbf{0}\}, \quad \overline{\cup_j V_j} = \mathbb{L}_2(\mathbb{R}),$$

where $\overline{A}$ denoting the closure of a set A. The hierarchy (1.9) is constructed such that:

(i) The V-spaces are self-similar, meaning:

$$f(2^j x) \in V_j \text{ iff } f(x) \in V_0. \tag{1.10}$$

(ii) There exists *a scaling function* $\phi \in V_0$ whose integer-translates span the space V_0, as follows:

$$V_0 = \left\{ f \in \mathbb{L}_2(\mathbb{R}) |\ \ f(x) = \sum_k c_k \phi(x-k) \right\},$$

and for which the set $\{\phi(\bullet - k),\ k \in \mathbb{Z}\}$ is an orthonormal basis.[1]

Note that mild conditions on ϕ are necessary to move forward in the developments. In this context, it can be assumed that $\int \phi(x)dx \geq 0$. Later, we will show that this integral is in fact equal to 1.

Since $V_0 \subset V_1$, the function $\phi(x) \in V_0$ can be represented as a linear combination of functions from V_1, which implies:

$$\phi(x) = \sum_{k \in \mathbb{Z}} h_k\ \sqrt{2}\phi(2x-k), \tag{1.11}$$

for some coefficients $h_k,\ k \in \mathbb{Z}$. This equation is denominated as the *scaling equation* and it is fundamental in constructing, exploring, and utilizing wavelets.

As an important remark, the coefficients h_n in (1.11) are important in connecting the MRA to the theory of signal processing. The (possibly infinite) vector $\mathbf{h} = \{h_n,\ n \in \mathbb{Z}\}$ will be called a *wavelet filter.* It is a low-pass (averaging) filter as will become clear later by considerations in the Fourier domain.

Theorem 1.1.1. *For the scaling function it holds:*

$$\int_{\mathbb{R}} \phi(x)dx = 1,$$

[1] It is possible to relax the orthogonality requirement. It is sufficient to assume that the system of functions $\{\phi(\bullet - k),\ k \in \mathbb{Z}\}$ constitutes a Riesz basis for V_0.

or, equivalently,

$$\Phi(0) = 1,$$

where $\Phi(\omega)$ *is the Fourier transformation of* ϕ*, given by* $\Phi(\omega) = \int_{\mathbb{R}} \phi(x) e^{-i\omega x} dx$.

To further explore the properties of multiresolution analysis subspaces and their bases, it is useful to work in the Fourier domain. Define the function m_0 as follows:

$$m_0(\omega) = \frac{1}{\sqrt{2}} \sum_{k \in \mathbb{Z}} h_k e^{-ik\omega} = \frac{1}{\sqrt{2}} H(\omega). \tag{1.12}$$

This function (1.12) describes the behavior of the associated filter $\mathbf{h}$ in the Fourier domain, and is sometimes called the *transfer function*. Moreover, note m_0 is periodic with the period 2π and that the filter taps $\{h_n,\ n \in \mathbb{Z}\}$ are the Fourier coefficients of the function $H(\omega) = \sqrt{2}\, m_0(\omega)$.

Now, taking the Fourier transformation of (1.11), it follows:

$$\begin{aligned}
\Phi(\omega) &= \int_{-\infty}^{\infty} \phi(x) e^{-i\omega x} dx \\
&= \sum_k \sqrt{2}\, h_k \int_{-\infty}^{\infty} \phi(2x-k) e^{-i\omega x} dx \\
&= \sum_k \frac{h_k}{\sqrt{2}} e^{-ik\omega/2} \int_{-\infty}^{\infty} \phi(2x-k) e^{-i(2x-k)\omega/2} d(2x-k) \\
&= \sum_k \frac{h_k}{\sqrt{2}} e^{-ik\omega/2}\ \Phi\left(\frac{\omega}{2}\right) \\
&= m_0\left(\frac{\omega}{2}\right) \Phi\left(\frac{\omega}{2}\right).
\end{aligned}$$

Therefore, the relation becomes:

$$\Phi(\omega) = m_0\left(\frac{\omega}{2}\right) \Phi\left(\frac{\omega}{2}\right), \tag{1.13}$$

where $\Phi(\omega)$ is the Fourier transformation of $\phi(x)$.

By iterating (1.13), it is possible to obtain:

$$\Phi(\omega) = \prod_{n=1}^{\infty} m_0\left(\frac{\omega}{2^n}\right), \tag{1.14}$$

which is convergent under very mild conditions on rates of decay of the scaling function ϕ. There are several sufficient conditions for convergence of the product in (1.14). For instance, the uniform convergence on compact sets is assured if the following holds:

(i) $m_0(\omega) = 1$

(ii) $|m_0(\omega) - 1| < C|\omega|^{\epsilon}$, for some positive C and ϵ.

Next, we show two important properties of wavelet filters in the context of an orthogonal multiresolution analysis, *normalization* and *orthogonality.*

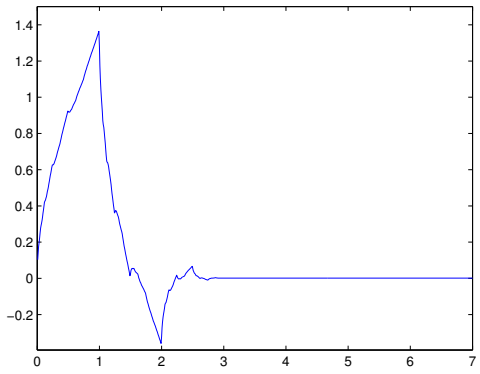
Normalization. For a wavelet filter it follows:

$$\sum_{k\in\mathbb{Z}} h_k = \sqrt{2}. \tag{1.15}$$

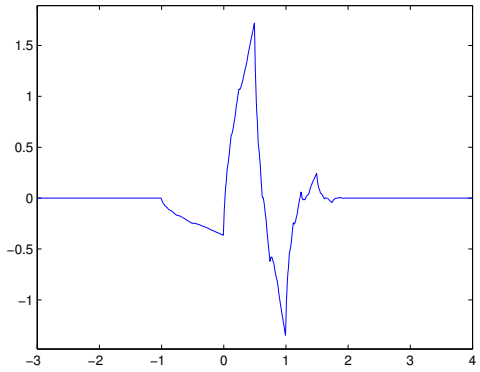
Proof:

$$\begin{aligned}
\int \phi(x)dx &= \sqrt{2}\sum_k h_k \int \phi(2x-k)dx \\
&= \sqrt{2}\sum_k h_k \,\frac{1}{2}\int \phi(2x-k)d(2x-k) \\
&= \frac{\sqrt{2}}{2}\sum_k h_k \int \phi(x)dx.
\end{aligned}$$

Since $\int \phi(x)dx \neq 0$ by assumption, (1.15) holds. This result can also be obtained from



(a)



(b)

Figure 1.2: (a) ϕ and (b) ψ for a given filter h.

$m_0(0) = 1$.

Orthogonality. For any $l \in \mathbb{Z}$,

$$\sum_k h_k h_{k-2l} = \delta_l. \tag{1.16}$$

Proof: First observe that from the scaling equation (1.11), it follows that:

$$\begin{aligned} \phi(x)\phi(x-l) &= \sqrt{2} \sum_k h_k \phi(2x-k)\phi(x-l) \\ &= \sqrt{2} \sum_k h_k \phi(2x-k) \sqrt{2} \sum_m h_m \phi(2(x-l)-m). \end{aligned} \tag{1.17}$$

Now, integrating the both sides in (1.17) leads to:

$$\begin{aligned}\delta_l &= 2\sum_k h_k \left[\sum_m h_m \frac{1}{2}\int \phi(2x-k)\phi(2x-2l-m)\,d(2x)\right] \\ &= \sum_k \sum_m h_k h_m \delta_{k,2l+m} \\ &= \sum_k h_k h_{k-2l},\end{aligned}$$

where the last follows from taking $k = 2l + m$.

Note that in the case where $l = 0$, (1.16) becomes:

$$\sum_k h_k^2 = 1. \tag{1.18}$$

Another important result from the orthogonality condition (1.16) is that the convolution of the filter $\mathbf{h}$ with itself, i.e. $\mathbf{f} = \mathbf{h} \star \mathbf{h}$, is an *à trous*.[2]

In addition, the fact that $\{\phi(\bullet - k),\ k \in \mathbb{Z}\}$ constitutes an orthonormal basis for V_0 can be expressed in the Fourier domain in terms of either $\Phi(\omega)$ or $m_0(\omega)$, as follows:

(**a**) In terms of $\Phi(\omega)$:

$$\sum_{l=-\infty}^{\infty} |\Phi(\omega + 2\pi l)|^2 = 1. \tag{1.19}$$

Indeed, from the periodicity of the Fourier transformation and the 2π-periodicity of $e^{i\omega k}$, it is

[2] The attribute *à trous (Fr.)* ($\equiv$ with holes) comes from the property $f_{2n} = \delta_n$, i.e., each tap on even position in $\mathbf{f}$ is 0, except the tap f_0. Such filters are also called half-band filters.

possible to obtain:

$$\begin{aligned}\delta_k &= \int_{\mathbb{R}} \phi(x)\overline{\phi(x-k)}dx \\ &= \frac{1}{2\pi}\int_{\mathbb{R}} \Phi(\omega)\overline{\Phi(\omega)}e^{i\omega k}d\omega \\ &= \frac{1}{2\pi}\int_0^{2\pi}\sum_{l=-\infty}^{\infty}|\Phi(\omega+2\pi l)|^2 e^{i\omega k}d\omega. \end{aligned} \tag{1.20}$$

Here, the last line in (1.20) corresponds to the Fourier coefficient a_k in the Fourier series decomposition of the function $f(\omega)$ defined as:

$$f(\omega) = \sum_{l=-\infty}^{\infty}|\Phi(\omega+2\pi l)|^2.$$

Since the Fourier representation is unique, it follows that $f(\omega) = 1$. As additional results, it is possible to observe that $\Phi(2\pi n) = 0, n \neq 0$, and $\sum_n \phi(x-n) = 1$. The last result follows from inspection of coefficients c_k in the Fourier decomposition of $\sum_n \phi(x-n)$, the series $\sum_k c_k e^{2\pi ikx}$. Since this function is 1-periodic, it implies:

$$c_k = \int_0^1 \left(\sum_n \phi(x-n)\right) e^{-2\pi ikx}dx = \int_{-\infty}^{\infty}\phi(x)e^{-2\pi ikx}dx = \Phi(2\pi k) = \delta_{0,k}.$$

Remark 1.1.1. The identity (1.19) shows that, any set of linearly independent functions spanning V_0, $\{\phi(x-k), k \in \mathbb{Z}\}$, can be orthogonalized in the Fourier domain. Thus, it is possible to obtain an orthonormal basis, which is generated by integer-shifts of the function:

$$\mathcal{F}^{-1}\left[\frac{\Phi(\omega)}{\sqrt{\sum_{l=-\infty}^{\infty}|\Phi(\omega+2\pi l)|^2}}\right]. \tag{1.21}$$

This normalization in the Fourier domain is used in constructing of some wavelet bases.

(**b**) In terms of m_0 :

$$|m_0(\omega)|^2 + |m_0(\omega+\pi)|^2 = 1. \tag{1.22}$$

Since $\sum_{l=-\infty}^{\infty} |\Phi(2\omega + 2l\pi)|^2 = 1$, then using (1.13) implies:

$$\sum_{l=-\infty}^{\infty} |m_0(\omega + l\pi)|^2 |\Phi(\omega + l\pi)|^2 = 1. \tag{1.23}$$

It is possible to split the sum in (1.23) into two sums – one with odd and the other with even indices. Therefore:

$$\begin{aligned} 1 &= \sum_{k=-\infty}^{\infty} |m_0(\omega + 2k\pi)|^2 |\Phi(\omega + 2k\pi)|^2 + \\ &\quad \sum_{k=-\infty}^{\infty} |m_0(\omega + (2k+1)\pi)|^2 |\Phi(\omega + (2k+1)\pi)|^2. \end{aligned}$$

Using relation (1.19) and the 2π-periodicity of $m_0(\omega)$, it follows that:

$$\begin{aligned} 1 &= |m_0(\omega)|^2 \sum_{k=-\infty}^{\infty} |\Phi(\omega + 2k\pi)|^2 + |m_0(\omega+\pi)|^2 \sum_{k=-\infty}^{\infty} |\Phi((\omega+\pi) + 2k\pi)|^2 \\ &= |m_0(\omega)|^2 + |m_0(\omega+\pi)|^2. \end{aligned}$$

1.1.3 Generation of Local Bases via Wavelets

Classical orthonormal bases (Fourier, Hermite, Legendre, etc.) have been widely used in applied mathematics. However, there is a significant limitation shared by many of these clas-

sical bases, which is non-locality. We can say that a basis is non-local when many basis functions are significantly contributing at any value of a decomposition. Moreover, local bases are desirable since they are more adaptive and parsimonious, which leads to in general, better convergence properties and a better flexibility to achieve good approximations for rapidly varying functions with a reasonably small number of expansion coefficients.

When a sequence of subspaces satisfies MRA properties, there exists (though not unique) an orthonormal basis for $\mathbb{L}_2(\mathbb{R})$ given by:

$$\{\psi_{jk}(x) = 2^{j/2}\psi(2^j x - k),\ j, k \in \mathbb{Z}\} \tag{1.24}$$

such that $\{\psi_{jk}(x),\ j\text{-fixed},\ k \in \mathbb{Z}\}$ is an orthonormal basis of the "difference space":

$$W_j = V_{j+1} \ominus V_j.$$

Here, the function $\psi(x) = \psi_{00}(x)$ is called a *wavelet function* or informally, *the mother wavelet*.

Now, it is possible to obtain the wavelet function from the scaling function $\phi(x)$. Since $\psi(x) \in V_1$ (due to the containment $W_0 \subset V_1$), it can be represented as:

$$\psi(x) = \sum_{k\in\mathbb{Z}} g_k\ \sqrt{2}\phi(2x - k), \tag{1.25}$$

for some coefficients $g_k,\ k \in \mathbb{Z}$. Define:

$$m_1(\omega) = \frac{1}{\sqrt{2}}\ \sum_k g_k e^{-ik\omega}. \tag{1.26}$$

By repeating what was done with m_0, it is possible to obtain the Fourier equivalent of (1.25). Indeed:

$$\Psi(\omega) = m_1(\frac{\omega}{2})\Phi(\frac{\omega}{2}). \tag{1.27}$$

Note that the spaces W_0 and V_0 are orthogonal by construction. Therefore, it follows that:

$$\begin{aligned} 0 = \int \psi(x)\phi(x-k)dx &= \frac{1}{2\pi}\int \Psi(\omega)\overline{\Phi(\omega)}e^{i\omega k}d\omega \\ &= \frac{1}{2\pi}\int_0^{2\pi}\sum_{l=-\infty}^{\infty}\Psi(\omega+2l\pi)\overline{\Phi(\omega+2l\pi)}e^{i\omega k}d\omega. \end{aligned}$$

Using the Fourier series argument, as in (1.19), it is possible to conclude:

$$\sum_{l=-\infty}^{\infty}\Psi(\omega+2l\pi)\overline{\Phi(\omega+2l\pi)} = 0.$$

Now, taking into account the definitions of m_0 and m_1, and by observing the derivation process of (1.22), we arrive at:

$$m_1(\omega)\overline{m_0(\omega)} + m_1(\omega+\pi)\overline{m_0(\omega+\pi)} = 0. \tag{1.28}$$

From (1.28), we can argue that there exists a function $\lambda(\omega)$ such that:

$$\begin{bmatrix} m_1(\omega) \\ m_1(\omega+\pi) \end{bmatrix} = \lambda(\omega)\begin{bmatrix} \overline{m_0(\omega+\pi)} \\ -\overline{m_0(\omega)} \end{bmatrix}. \tag{1.29}$$

By substituting $\xi = \omega + \pi$ and by using the 2π-periodicity of m_0 and m_1, it follows from

(1.29):

$$\begin{aligned} \lambda(\omega) &= -\lambda(\omega+\pi), \text{ and} \\ \lambda(\omega) &\quad \text{is } 2\pi\text{-periodic.} \end{aligned} \tag{1.30}$$

Therefore, any function $\lambda(\omega)$ of the form $e^{\pm i\omega}S(2\omega)$, where S is an $\mathbb{L}_2([0,2\pi])$, 2π-periodic function, will satisfy (1.28); however, only the functions for which $|\lambda(\omega)| = 1$ will define an orthogonal basis ψ_{jk} of $\mathbb{L}_2(\mathbb{R})$.

To summarize the construction of orthonormal systems from a scaling function $\phi(x)$, we need to choose $\lambda(\omega)$ such that is satisfies:

(i) $\lambda(\omega)$ is 2π-periodic,

(ii) $\lambda(\omega) = -\lambda(\omega+\pi)$, and

(iii) $|\lambda(\omega)|^2 = 1$.

Standard choices for $\lambda(\omega)$ are $-e^{-i\omega}$, $e^{-i\omega}$, and $e^{i\omega}$; however, any other function satisfying (i)-(iii) will generate a valid m_1. Note that defining $m_1(\omega)$ as:

$$m_1(\omega) = -e^{-i\omega}\overline{m_0(\omega+\pi)}. \tag{1.31}$$

will generate a convenient and standard connection between the filters $\mathbf{h}$ and $\mathbf{g}$ that will be presented next. In fact, this form of m_1 and the equation (1.19) imply that $\{\psi(\bullet - k),\ k \in \mathbb{Z}\}$ is an orthonormal basis for W_0.

Since $|m_1(\omega)| = |m_0(\omega+\pi)|$ (from (1.29), since $|\lambda(\omega)| = 1$), the orthogonality condition (1.22) can be expressed as follows:

$$|m_0(\omega)|^2 + |m_1(\omega)|^2 = 1. \tag{1.32}$$

Note that we can relate g_n and h_n by comparing the definition of m_1 in (1.26) with the following:

$$\begin{aligned} m_1(\omega) &= -e^{-i\omega}\frac{1}{\sqrt{2}}\sum_k h_k e^{i(\omega+\pi)k} \\ &= \frac{1}{\sqrt{2}}\sum_k (-1)^{1-k} h_k e^{-i\omega(1-k)} \\ &= \frac{1}{\sqrt{2}}\sum_n (-1)^n h_{1-n} e^{-i\omega n}, \end{aligned}$$

thus, the relation between g_n and h_n is given by:

$$g_n = (-1)^n\, h_{1-n}. \tag{1.33}$$

In the signal processing literature, the relation defined by (1.33) is known as the *quadrature mirror relation* and the filters $\mathbf{h}$ and $\mathbf{g}$ are referred to as *quadrature mirror filters.*

As was seen from the previous derivations, locality of wavelet bases comes from their construction. In general, most of the wavelets that are used in statistics now are either compactly supported or decay exponentially. An exception are Meyer-type wavelets (with a polynomial decay) used in deconvolution problems.

1.1.4 Regularity of Wavelets

There are many different wavelet bases. An interesting and powerful feature of wavelets is diversity in their properties, since it is possible to construct wavelets with different smoothness, symmetry, oscillatory, support, etc. properties. However, sometimes the desired requirements can be conflicting since some of the properties are exclusive. For example, there is no symmetric real-valued wavelet with a compact support. Similarly, there is no $\mathbb{C}^\infty$-wavelet

function with an exponential decay, among others.

Scaling functions and wavelets can be constructed with a desired degree of smoothness. This regularity (smoothness) of wavelets is connected with the rate of decay of scaling functions, and ultimately with the number of vanishing moments of scaling and wavelet functions. For example, the Haar wavelet has only the "zeroth" vanishing moment (as a consequence of the admissibility condition) resulting in a discontinuous wavelet function.

Theorem 1.1.2 (presented next) shows the important connection between the regularity of wavelets, the number of vanishing moments, and the form of the transfer function $m_0(\omega)$. Its proof follows from the Taylor series argument and the scaling properties of wavelet functions. For details, see Daubechies (1992)[5], pp 153–155.

Now, before introducing Theorem 1.1.2, define:

$$\mathcal{M}_k = \int x^k \phi(x) dx \text{ and } \mathcal{N}_k = \int x^k \psi(x) dx,$$

be the kth moments of the scaling and wavelet functions, respectively.

Theorem 1.1.2. *Let $\psi_{jk}(x) = 2^{j/2}\psi(2^j x - k),\ j,k \in \mathbb{Z}$ be an orthonormal system of functions in $\mathbb{L}_2(\mathbb{R})$, such that for and arbitrary $N \in \mathbb{N}$, $\psi(x)$ satisfies:*

(i) For some constant $C_1 > 0$, and $\alpha > N$:

$$|\psi(x)| \leq \frac{C_1}{(1+|x|)^\alpha}.$$

(ii) $\psi \in \mathbb{C}^{N-1}(\mathbb{R})$.

(iii) The derivatives $\psi^{(k)}(x)$ are bounded for $k \leq N-1$.

Then, ψ has N vanishing moments, $\mathcal{N}_k = 0,\ \ 0 \leq k \leq N-1$.

If, in addition, for some constant $C_2 > 0$ and $\alpha > N$, the scaling function $\phi(x)$ satisfies:

$$|\phi(x)| \leq \frac{C_2}{(1+|x|)^\alpha},\ \alpha > N$$

then, the associated function $m_0(\omega)$ is necessarily of the form

$$m_0(\omega) = \left(\frac{1+e^{-i\omega}}{2}\right)^N \cdot \mathcal{L}(\omega), \tag{1.34}$$

where $\mathcal{L}$ is a 2π-periodic, $\mathbb{C}^{N-1}$-function.

The following definition of regularity is often used in the literature:

Definition 1.1.1. *The multiresolution analysis (or, the scaling function) is said to be r-regular if, for any $\alpha \in \mathbb{Z}$, and a some positive constant C, the scaling function has $r+1$ bounded derivatives in the form:*

$$|\phi^{(k)}(x)| \leq \frac{C}{(1+|x|)^\alpha},$$

for $k = 0, 1, \ldots, r$.

Similarly, it is possible to express the requirement that ψ possesses N vanishing moments in terms of Ψ, m_0, or equivalently, in terms of the filter $\mathbf{h}$.

Assume that a wavelet function $\psi(x)$ has N vanishing moments, i.e.,

$$\mathcal{N}_k = 0,\ k = 0, 1, \ldots, N-1. \tag{1.35}$$

This condition (1.35) can be translated into the Fourier domain as follows:

$$\left.\frac{d^k\Psi(\omega)}{d\omega^k}\right|_{\omega=0} = 0,\ k = 0, 1, \ldots, N-1,$$

which implies:

$$m_1^{(k)}(\omega)\,|_{\omega=0} = m_1^{(k)}(0) = 0,\ k = 0, 1, \ldots, N-1. \tag{1.36}$$

It is possible to verify that in terms of m_0, the relation (1.36) takes the form:

$$m_0^{(k)}(\omega)\,|_{\omega=\pi} = m_0^{(k)}(\pi) = 0,\ k = 0, 1, \ldots, N-1. \tag{1.37}$$

This follows from an inductive argument. In fact, the case $k = 0$ follows from $\Psi(0) = m_1(0)\Phi(0)$ [(1.27) evaluated at $\omega = 0$] and the fact that $\Phi(0) = 1$. Since $\Psi'(0) = \frac{1}{2}m_1'(0)\Psi(0) + \frac{1}{2}m_1(0)\Psi'(0)$ it follows that $m_1'(0) = 0$, as well. Then, $m_1^{(N-1)}(0) = 0$ follows by induction.

Note that the condition $m_1^{(k)}(0) = 0,\ k = 0, 1, \ldots, N-1$ imposes the following constraint on the wavelet-filter coefficients:

$$\sum_{n\in\mathbb{Z}} n^k g_n = \sum_{n\in\mathbb{Z}} (-1)^n n^k h_n = 0,\ \ k = 0, 1, \ldots, N-1. \tag{1.38}$$

Now we can ask the question, how smooth are the wavelets from a given family? For example, as shown by Daubechies, there is an apparent trade-off between the length of support and the regularity index of scaling functions.

In the case of Daubechies family of wavelets, let ϕ be the DAUBN scaling function. The regularity of ϕ can be measured by two popular ways: Sobolev and Hölder regularity exponents.

Let α_N^* be defined as:

$$\begin{aligned} \alpha_N^* &= arg\sup_{\beta} \int (1+|\omega|)^{\beta} |\Phi(\omega)| d\omega \\ &\text{s.t.} \quad \int (1+|\omega|)^{\beta} |\Phi(\omega)| d\omega < \infty, \end{aligned}$$

and let α_N be the exponent of the Hölder space $\mathbb{C}^{\alpha_N}$ to which the scaling function ϕ belongs.

Daubechies (1988) and Daubechies and Lagarias (1991, 1992), obtained regularity exponents for wavelets in the Daubechies family, which are summarized in Table 1.1.

Table 1.1: Sobolev α_N^* and Hölder α_N regularity exponents of Daubechies' scaling functions.

N	1	2	3	4	5	6	7	8	9	10
α_N^*	0.5	1	1.415	1.775	2.096	2.388	2.658	2.914	3.161	3.402
α_N		0.550	0.915	1.275	1.596	1.888	2.158	2.415	2.661	2.902

From Table 1.1, we see that DAUB4 is the first differentiable wavelet, since $\alpha > 1$. More precise bounds on α_N yield that ϕ from the DAUB3 family is, in fact, the first differentiable scaling function ($\alpha_3 = 1.0878$), even though it seams to have a peak at 1. See also Daubechies (1992), page 239, for the discussion.

1.1.5 Approximations and Characterizations of Functional Spaces using Wavelets

Note that any function $f \in \mathbb{L}_2(\mathbb{R})$ can be represented by the orthonormal expansion:

$$f(x) = \sum_{j,k} d_{jk} \psi_{jk}(x).$$

From the MRA, this unique representation corresponds to a multiresolution decomposition fo the form:

$$\mathbb{L}_2(\mathbb{R}) = \bigoplus_{j=-\infty}^{\infty} W_j.$$

Also, for any fixed j_0 the decomposition $\mathbb{L}_2(\mathbb{R}) = V_{j_0} \oplus \bigoplus_{j=j_0}^{\infty} W_j$ corresponds to the expansion:

$$f(x) = \sum_k c_{j_0,k}\phi_{j_0,k}(x) + \sum_{j \geq j_0} \sum_k d_{jk}\psi_{j,k}(x). \tag{1.39}$$

Note that the first sum in (1.39) corresponds to the orthogonal projection $\mathbb{P}_{j_0}$ of f onto V_{j_0}, denoted as $f_{j_0}(x) = \mathbb{P}_{j_0} f(x)$. In fact, by the orthogonality principle $c_{j_0,k} = \langle f(x), \phi_{j_0,k}(x) \rangle_{\mathbb{L}_2}$.

In general, if the regularities of functions f and ϕ are known, then it is possible to bound $||\mathbb{P}_{j_0} f - f|| = ||(\mathbb{I} - \mathbb{P}_{j_0})f||$. In fact, when f has N-continuous derivatives, and ϕ is such that satisfies:

(i) The reproducing kernel generated by $\phi(x)$, for a fixed j_0, i.e. $\mathbb{K}_{j_0}(x,u) = \sum_k \phi_{j_0,k}(x)\phi_{j_0,k}(u)$ is absolutely bounded by a function $F(x) \in \mathbb{L}_2$, that satisfies $\int |x|^N F(x)dx < \infty$, and

(ii) $\int \mathbb{K}_{j_0}(x,u)(u-x)^l = \delta_{0,l}$ for $l = 0, \ldots, N$.

The, there exists a constant $C > 0$ such that:

$$||f(x) - \mathbb{P}_{j_0} f(x)||_{\mathbb{L}_2} \leq \frac{2^{-Nj_0}}{(N-1)!}\, o(2^{-j_0}) \text{ as } j_0 \to \infty.$$

This result follows from the application of a Taylor expansion, and it can be generalized to other functional spaces, as discussed in [5]. The complete proof can be found in Lemma 8.3 [9].

As it was mentioned before, due to it characteristics Wavelets allow for the characterizations

of different functional spaces. For example, a function f belongs to the Hölder space $\mathbb{C}^s$ if and only if there is a constant $C > 0$ such that in an r-regular MRA ($r > s$) the wavelet coefficients satisfy the following two conditions:

$$\begin{aligned} &(i) \qquad |c_{j_0,k}| \leq C, \\ &(ii) \qquad |d_{j,k}| \leq C \cdot 2^{-j(s+\frac{1}{2})},\ j \geq j_0, k \in \mathbb{Z}\,. \end{aligned} \tag{1.40}$$

Similarly, a function f belongs to the Sobolev $\mathbb{W}_2^s(\mathbb{R})$ space if and only if the wavelet coefficients satisfy:

$$\sum_{j,k} |d_{jk}|^2 \cdot (1 + 2^{2js}) < \infty.$$

Even the general (non-homogeneous) Besov spaces, can be characterized by moduli of the wavelet coefficients of its elements. For a given r-regular MRA with $r > \max\{\sigma, 1\}$, the following result holds: (see Meyer 1992, page 200)

Theorem 1.1.3. *Let I_j be a set of indices so that $\{\psi_i,\ i \in I_j\}$ constitutes an o.n. basis of the detail space W_j. There exist two constants $C' \geq C > 0$ such that, for every exponent $p \in [1, \infty]$, for each $j \in \mathbb{Z}$ and for every element $f(x) = \sum_{i \in I_j} d_i \psi_i(x)$ in W_j,*

$$C||f||_p \leq 2^{j/2} 2^{-j/p} \left(\sum_{i \in I_j} |d_i|^p \right)^{1/p} \leq C'||f||_p.$$

This result enables the following characterization of Besov $\mathbb{B}_{p,q}^{\sigma}$ spaces. If the MRA has regularity $r > s$, then wavelet bases are Riesz bases for all $1 \leq p, q \leq \infty,\ 0 < \sigma < r$. The function $f = \sum_k c_{j_0 k} \phi_{j_0 k}(x) + \sum_{j \geq j_0} \sum_k d_{jk} \psi_{jk}(x)$ belongs to $\mathbb{B}_{p,q}^{\sigma}$ space if its wavelet coefficients satisfy the following conditions:

(i) The l_p norm of the scaling coefficients $\{c_{j_0,k},\ k \in \mathbb{Z}\}$ is bounded. Thus:

$$\left(\sum_k |c_{j_0,k}|^p \right)^{1/p} < \infty, \text{ and}$$

(ii) The sequence of detail coefficients given by:

$$\left\{ \left(\sum_{i \in I_j} 2^{j(\sigma+1/2-1/p)} |d_i|^p \right)^{1/p}, \ j \geq j_0 \right\}$$

is an ℓ_q sequence, i.e, $\left[\sum_{j \geq j_0} \left(2^{j(\sigma+1/2-1/p)} (\sum_k |d_{j,k}|^p)^{1/p} \right)^q \right]^{1/q} < \infty$.

The aforementioned results concern with global regularity of functions via Wavelets. On the other hand, it is possible to study the the local regularity of functions by inspecting the magnitudes of their wavelet coefficients. For more details, the the work of Jaffard (1991)[10] and Jaffard and Laurencot (1992)[11] are useful references.

1.1.6 Daubechies-Lagarias Algorithm

It this Thesis, Chapters 2, 3 and 4 propose statistical modeling methodologies that require the evaluation of wavelet and scaling functions at arbitrary points. For this purpose, we describe an algorithm for fast numerical calculation of these quantities, based on the Daubechies-Lagarias (Daubechies and Lagarias, 1991, 1992)[12] *local pyramidal algorithm*.

For example, in Daubechies' families the scaling and wavelet function have no explicit representations (except for the Haar wavelet). As was mentioned before, for applications such as density estimation, and non-linear regression, etc., it necessary to find values of DAUB functions at arbitrary points.

The Daubechies-Lagarias algorithm enables these evaluations of ϕ and ψ at a any point in the

support with arbitrary precision. We illustrate the algorithm on wavelets from the Daubechies family; however, the algorithm works for all orthogonal wavelet filters.

Let ϕ be the scaling function of the DAUBN wavelet (i.e. a wavelet from Daubechies family that has N vanishing moments). In this case, the support of ϕ is $[0, 2N-1]$. Let $x \in (0,1)$, and let $dyad(x) = \{d_1, d_2, \dots d_n, \dots\}$ be the set of 0-1 digits in the dyadic representation of x $(x = \sum_{j=1}^{\infty} d_j 2^{-j})$. By $dyad(x,n)$, we denote the subset of the first n digits from $dyad(x)$, i.e., $dyad(x,n) = \{d_1, d_2, \dots d_n\}$.

Let $\mathbf{h} = (h_0, h_1, \dots, h_{2N-1})$ be the wavelet filter coefficients. We build two $(2N-1) \times (2N-1)$ matrices as:

$$T_0 = (\sqrt{2} \cdot h_{2i-j-1})_{1 \le i,j \le 2N-1} \text{ and } \quad T_1 = (\sqrt{2} \cdot h_{2i-j})_{1 \le i,j \le 2N-1}. \tag{1.41}$$

Then the local pyramidal algorithm can be constructed based on Theorems 1.1.4 and 1.1.5, taken from [3].

Theorem 1.1.4. *Daubechies-Lagarias, [3]*

$$\lim_{n \to \infty} \quad T_{d_1} \cdot T_{d_2} \cdot \dots \cdot T_{d_n} \tag{1.42}$$

$$= \begin{bmatrix} \phi(x) & \phi(x) & \dots & \phi(x) \\ \phi(x+1) & \phi(x+1) & \dots & \phi(x+1) \\ \vdots & & & \\ \phi(x+2N-2) & \phi(x+2N-2) & \dots & \phi(x+2N-2) \end{bmatrix}.$$

As mentioned in [3], the convergence of $||T_{d_1} \cdot T_{d_2} \cdot \dots \cdot T_{d_n} - T_{d_1} \cdot T_{d_2} \cdot \dots \cdot T_{d_{n+m}}||$ to zero, for fixed m, is exponential and constructive, therefore, effective decreasing bounds on the error can be established.

Example 1.1.1. Consider the DAUB2 scaling function ($N = 2$). The corresponding filter is given by $\mathbf{h} = \left(\frac{1+\sqrt{3}}{4\sqrt{2}}, \frac{3+\sqrt{3}}{4\sqrt{2}}, \frac{3-\sqrt{3}}{4\sqrt{2}}, \frac{1-\sqrt{3}}{4\sqrt{2}}\right)$. Using (1.41), the matrices T_0 and T_1 are structured as follows:

$$T_0 = \begin{bmatrix} \frac{1+\sqrt{3}}{4} & 0 & 0 \\ \frac{3-\sqrt{3}}{4} & \frac{3+\sqrt{3}}{4} & \frac{1+\sqrt{3}}{4} \\ 0 & \frac{1-\sqrt{3}}{4} & \frac{3-\sqrt{3}}{4} \end{bmatrix} \quad \text{and} \quad T_1 = \begin{bmatrix} \frac{3+\sqrt{3}}{4} & \frac{1+\sqrt{3}}{4} & 0 \\ \frac{1-\sqrt{3}}{4} & \frac{3-\sqrt{3}}{4} & \frac{3+\sqrt{3}}{4} \\ 0 & 0 & \frac{1-\sqrt{3}}{4} \end{bmatrix}.$$

Now, we will evaluate the scaling function at an arbitrary point, say $x = 0.45$. Twenty "decimals" in the dyadic representation of 0.45 are $dyad(0.45, 20) = \{$ 0, 1, 1, 1, 0, 0, 1, 1, 0, 0, 1, 1, 0, 0, 1, 1, 0, 0, 1, 1 $\}$. In addition to the value at 0.45, the algorithm delivers the values at 1.45 and 2.45 (the values 0.45, 1.45, and 2.45 are contained in the support of ϕ, the interval [0,3]). The values $\phi(0.45), \phi(1.45),$ and $\phi(2.45)$ may be approximated as averages of the first, second, and third row, respectively in the matrix

$$\prod_{i \in dyad(0.45,20)} T_i = \begin{bmatrix} 0.86480582 & 0.86480459 & 0.86480336 \\ 0.08641418 & 0.08641568 & 0.08641719 \\ 0.04878000 & 0.04877973 & 0.04877945 \end{bmatrix}.$$

Using Daubechies-Lagarias algorithm, it is only possible to obtain the values of the scaling function. In many applications, most of the evaluation require the wavelet function as well. For this purpose, it turns out that it is possible to use the same algorithm, due to the following result, taken from [3].

Theorem 1.1.5. *Vidakovic (1999)[3] Let x be an arbitrary real number, let the wavelet be*

given by its filter coefficients, and let $\mathbf{u}$ *with* $2N-1$ *be a vector defined as:*

$$\mathbf{u}(x) = \{(-1)^{1-\lfloor 2x \rfloor} h_{i+1-\lfloor 2x \rfloor}, i = 0, \ldots, 2N-2\}.$$

If for some i *the index* $i+1-\lfloor 2x \rfloor$ *is negative or larger than* $2N-1$*, then the corresponding component of* $\mathbf{u}$ *is equal to 0.*

Let the vector $\mathbf{v}$ *be defined as follows:*

$$\mathbf{v}(x,n) = \frac{1}{2N-1}\mathbf{1}' \prod_{i \in dyad(\{2x\},n)} T_i,$$

where $\mathbf{1}' = (1, 1, \ldots, 1)$ *is the row-vector of ones. Then,*

$$\psi(x) = \lim_{n \to \infty} \mathbf{u}(x)' \mathbf{v}(x,n),$$

and the limit is constructive.

As noted in [3], the proof of this theorem is a straightforward but somewhat tedious re-expression of (1.25).

1.1.7 Wavelets "Disbalance" Energy in Data

By the use of orthogonal wavelets transformations[3], it is possible to detect the uneven distribution of energy within a signal. This feature of signals is very useful in applications such as data compression, since a signal can be well described by only a few energetic components. Similarly, since Wavelet transformations map a signal into a two dimensional space (i.e. scale and location), this energetic disbalance can be translated into scale-wise energy contributions. In particular, this application will be studied in Chapter 6, where we propose

[3]Here, we emphasize the orthogonal nature of the transformation, since it is crucial for the energy conservation after the mapping into the wavelet domain. This follows from Parseval's theorem.

a systematic way to express usual time series correlation into a weighted sum of scale-level correlations between wavelet expansion coefficient.

1.1.8 Discrete Wavelet Transformations

Discrete wavelet transformations (DWT) enable the mapping of data from the time domain (the original or input data, signal vector's original domain) to the wavelet domain. These transformations are linear and they can be defined by matrices of dimension $n \times n$ if they are applied to inputs of size n. Indeed, when the decimated type transformation is used, the resulting vector has the same size of the original signal. Depending on the boundary conditions, the transformation matrices can be either orthogonal or "close" to orthogonal. In the former case, when the utilized matrix is orthogonal, the transformation corresponds simply to a rotation in $\mathbb{R}^n$, where the signal vectors can be interpreted as coordinates of a single point. The coordinates of the point in the new, rotated space correspond to the discrete wavelet transformation of the original coordinates.

In 1989, Mallat (1989a,b)[13] formally defined the link between wavelets, multiresolution analyses and cascade algorithms, producing a constructive and efficient procedure for implementing the discrete wavelet transformation. His results relate the expansion wavelet coefficients from different multiresolution levels in the transformation by filtering the signal with two filters $\mathbf{h}$ and $\mathbf{g}$.

This direct relation between the original signal and the expansion coefficients from the space V_J, for some multiresolution index J is very convenient. Indeed, it is exact for wavelets such as Haar, Shannon, some biorthogonal and halfband-filter wavelets (interpolating wavelets) and close to exact for other kinds of wavelets, for example coiflets. Then, coarser smooth and complementing detail spaces are (V_{J-1}, W_{J-1}), (V_{J-2}, W_{J-2}), etc. Note that decreasing the index in V-spaces is equivalent to coarsening or smoothing the approximation to the

original data.

Along this line, by a simple substitution of indices in the scaling equations (1.11) and (1.25), it is possible to obtain:

$$\phi_{j-1,l}(x) = \sum_{k\in\mathbb{Z}} h_{k-2l}\phi_{jk}(x) \text{ and } \psi_{j-1,l}(x) = \sum_{k\in\mathbb{Z}} g_{k-2l}\phi_{jk}(x). \tag{1.43}$$

These relations in (1.43) are fundamental in developing the cascade algorithm, as it will be shown next.

Suppose a multiresolution analysis $\cdots \subset V_{j-1} \subset V_j \subset V_{j+1} \subset \ldots$. Since $V_j = V_{j-1} \oplus W_{j-1}$, any function $v_j \in V_j$ can be uniquely represented as $v_j(x) = v_{j-1}(x) + w_{j-1}(x)$, where $v_{j-1} \in V_{j-1}$ and $w_{j-1} \in W_{j-1}$. It is a common practice in the literature to denote the coefficients associated with $\phi_{jk}(x)$ and $\psi_{jk}(x)$ by c_{jk} and d_{jk}, respectively.

Under these definitions it follows,

$$\begin{aligned} v_j(x) &= \sum_k c_{j,k}\phi_{j,k}(x) \\ &= \sum_l c_{j-1,l}\phi_{j-1,l}(x) + \sum_l d_{j-1,l}\psi_{j-1,l}(x) \\ &= v_{j-1}(x) + w_{j-1}(x). \end{aligned} \tag{1.44}$$

By using the general scaling equations (1.43), orthogonality of $w_{j-1}(x)$ and $\phi_{j-1,l}(x)$ for any

j and l, and additivity of inner products, it is possible to obtain:

$$\begin{aligned} c_{j-1,l} &= \langle v_j, \phi_{j-1,l} \rangle \\ &= \langle v_j, \sum_k h_{k-2l} \phi_{j,k} \rangle \\ &= \sum_k h_{k-2l} \langle v_j, \phi_{j,k} \rangle \\ &= \sum_k h_{k-2l} c_{j,k}. \end{aligned} \tag{1.45}$$

Using the same argument, it follows that $d_{j-1,l} = \sum_k g_{k-2l} c_{j,k}$.

Note that the cascade algorithm also works in reverse direction. In fact, expansion coefficients in the next finer scale corresponding to V_j can be obtained from the coefficients corresponding to V_{j-1} and W_{j-1}. The relation given by:

$$\begin{aligned} c_{j,k} &= \langle v_j, \phi_{j,k} \rangle \\ &= \sum_l c_{j-1,l} \langle \phi_{j-1,l}, \phi_{j,k} \rangle + \sum_l d_{j-1,l} \langle \psi_{j-1,l}, \phi_{j,k} \rangle \\ &= \sum_l c_{j-1,l} h_{k-2l} + \sum_l d_{j-1,l} g_{k-2l}, \end{aligned} \tag{1.46}$$

describes a single step in the reconstruction algorithm.

Note that from Eq.(1.44) each each function $v_j(x)$ can be expressed via a change of basis. For example, the change of basis in V_1 from $\mathcal{B}_1 = \{\phi_{1k}(x), k \in Z\}$ to $\mathcal{B}_2 = \{\phi_{0k}, k \in Z\} \cup \{\psi_{0k}, k \in Z\}$ can be obtained through a matrix multiplication. Since this can be applied to any arbitrary multiresolution index j, then it is possible to define the DWT via matrix multiplication.

Discrete Wavelet Transformations as Matrix Transformations

Suppose the length of the input signal is 2^J, and let $\mathbf{h} = \{h_s,\ s \in \mathbb{Z}\}$ be the wavelet filter and $N > 0$ to be an appropriately chosen constant.

Denote by H_k a matrix of size $(2^{J-k} \times 2^{J-k+1}),\ k = 1, \ldots$ with entries given by:

$$h_s, \quad s = (N-1) + (i-1) - 2(j-1) \text{ modulo } 2^{J-k+1}, \tag{1.47}$$

at the position (i, j).

Observe that H_k is a circulant matrix, its ith row is 1st row circularly shifted to the right by $2(i-1)$ units. This circularity results from using the *modulo* operator in (1.47).

By analogy, it is possible to define a matrix G_k by using the filter $\mathbf{g}$. A version of G_k corresponding to the already defined H_k can be obtained by changing h_i by $(-1)^i h_{N+1-i}$. The constant N is a shift parameter and affects the position of the wavelet on the time scale. For filters from the Daubechies family, a standard choice for N is the number of vanishing moments.

Note that the matrix $\begin{bmatrix} H_k \\ G_k \end{bmatrix}$ is a basis-change matrix in 2^{J-k+1} dimensional space; consequently, it is unitary.

Therefore,

$$I_{2^{J-k}} = [H_k'\ G_k'] \begin{bmatrix} H_k \\ G_k \end{bmatrix} = H_k' \cdot H_k + G_k' \cdot G_k.$$

and

$$I = \begin{bmatrix} H_k \\ G_k \end{bmatrix} \cdot [H_k'\ G_k'] = \begin{bmatrix} H_k \cdot H_k' & H_k \cdot G_k' \\ G_k \cdot H_k' & G_k \cdot G_k' \end{bmatrix}.$$

That implies,

$$H_k \cdot H_k' = I,\ G_k \cdot G_k' = I,\ G_k \cdot H_k' = H_k \cdot G_k' = 0,\ \text{and}\ H_k' \cdot H_k + G_k' \cdot G_k = I.$$

Now, for a given sequence y, the J-step wavelet transformation denoted as $\mathbf{d}$ is given by $\mathbf{d} = W_J \cdot \mathbf{y}$, where

$$W_1 = \begin{bmatrix} H_1 \\ G_1 \end{bmatrix},\ W_2 = \begin{bmatrix} \begin{bmatrix} H_2 \\ G_2 \end{bmatrix} \cdot H_1 \\ G_1 \end{bmatrix},$$

$$W_3 = \begin{bmatrix} \begin{bmatrix} \begin{bmatrix} H_3 \\ G_3 \end{bmatrix} \cdot H_2 \\ G_2 \end{bmatrix} \cdot H_1 \\ G_1 \end{bmatrix}, \ldots$$

Note that the obtained vector $\mathbf{d} = W_k \cdot \mathbf{y}_i$ has the following structure:

$$\mathbf{d} = [\mathbf{c}^{J-k}; \mathbf{d}^{J-k}; \mathbf{d}^{J-k+1}; ...; \mathbf{d}^{J-2}; \mathbf{d}^{J-1}] \tag{1.48}$$

In the last expression k corresponds to the number of steps in the DWT (usually, $k = J$). Also, it is important to mention that due to the decimated nature of the chosen DWT (in this case), the size of the vector $\mathbf{d}$ is also N (as in the original data vector $\mathbf{y}_i$). In (1.48), $\mathbf{c}^{J-k}$ corresponds to the smooth coefficients at scale level $J - k$; similarly, $\mathbf{d}^{J-k}$ corresponds to the set of detail coefficients at the scale level $J - k$.

In the next Chapter, a methodology for the robust estimation of survival probability densities in the presence of randomly censored data based on the use of wavelet approximations is introduced and analyzed.

CHAPTER 2

AN EMPIRICAL APPROACH TO SURVIVAL DENSITY ESTIMATION FOR RANDOMLY-CENSORED DATA USING WAVELETS

Density estimation is a classical problem in statistics and has received considerable attention when both the data has been fully observed and in the case of partially observed (censored) samples. In survival analysis or clinical trials, a typical problem encountered in the data collection stage is that the samples may be censored from the right. The variable of interest could be observed partially due to the presence of a set of events that occur at random and potentially censor the data. Consequently, developing a methodology that enables robust estimation of the lifetimes in such setting is of high interest for researchers.

In this Chapter, we propose a non-parametric linear density estimator using empirical wavelet coefficients that are fully data driven. We derive an asymptotically unbiased estimator constructed from the complete sample based on an inductive bias correction procedure. Also, we provide upper bounds for the bias and analyze the large sample behavior of the expected $\mathbb{L}_2$ estimation error based on the approach used by Stute (1995)[14], showing that the estimates are asymptotically normal and possess global mean square consistency.

In addition, we evaluate the proposed approach via a theoretical simulation study using different exemplary baseline distributions with different sample sizes. In this study, we choose a censoring scheme that produces a censoring proportion of 40% on average. Finally, we apply the proposed estimator to real data-sets previously published, showing that the proposed wavelet estimator provides a robust and useful tool for the non-parametric estimation of the survival time density function.

2.1 Introduction

Density estimation is a classical problem in statistics and has received considerable attention when both the data has been fully observed and also in the case of partially observed (censored) samples. See [15, 16, 17] for thorough discussions about this topic. In areas such as survival analysis, the estimate of the lifetime density function is of a major importance. In fact, the knowledge of how the lifetimes behave in medical follow-up research or reliability analysis is paramount to get insights, draw conclusions, derive results, make comparisons and/or characterize the underlying death/failure process.

In general, the density estimation problem can be approached from either a parametric or non-parametric perspective. In the first case, an assumption is made about the particular distribution or family of distributions to which the density of interest belongs. As can immediately be observed, that approach causes the estimated function to be completely dependant on the such assumption which may prove of high benefit in the case when it is correct or close-to correct. However, if the elicited family for the target density is not correct, the parametric approach may lead to unsatisfactory results.

Because of the uncertainty about parametric family, the non-parametric approach for density estimation has become a popular topic of research in statistics. In particular, popular methods for density estimation include kernel and nearest neighbors methods [18]. Another approach for the aforementioned problem consists of the use of orthogonal series (see [19, 20]). In this approach, wavelets can be utilized since they can generate orthonormal bases for functions belonging to different functional spaces such as $\mathbb{L}_2(\mathbb{R})$, Sobolev, Besov, etc.

One of the first uses of wavelets in density estimation could be traced back to papers by Doukhan and Leon (1990)[21], Antoniadis and Carmona (1991)[22], Kerkyacharian and Picard (1992)[23] and Walter (1992)[24]. Moreover, due to their locality in both time and fre-

quency and their exceptional approximation properties, wavelets provide a good choice for density estimation. See e.g. Meyer (1992)[25], Daubechies (1992)[5], Donoho and Johnstone (1994, 1995, 1998)[26, 27, 28] for detailed discussions about the properties of wavelets in this context. Also, in Vidakovic (1999)[3] an extensive and thorough discussion of wavelets and their application in statistical modeling can be found.

Even though wavelets offer major advantages for curve estimation, there is a potential problem associated with their use in density estimation: there is no guarantee that the estimates are positive or integrate to 1 when using general scaling functions ϕ. As described in [18], the negative values may appear often in the tails of the target distribution. Nonetheless, that can be addressed; a possible remedial approach is the estimation of the square root of the density which allows then to square back to get a non-negative estimate integrating to 1 (as can be see in Pinheiro and Vidakovic (1997) [29]).

In survival analysis or clinical trials, a typical problem encountered in the data collection stage is that the samples may be censored from the right. The variable of interest may be prevented to be fully observed due to the presence of random events (typically assumed to be independent of the variable of interest) and potentially censor the data. A common example of right censoring in clinical trials is the situation in which a patient leaves the study before its termination or was still alive by the end of the observation period. In these cases, only a subset of the observations are fully observed lifetimes; the others are partially observed and it is only known that the actual lifetime was greater than equal to the time at which the subject ceased to be observed (i.e. the censored time).

Let $X_1, ..., X_N$ be i.i.d. survival times with a common unknown density function f. Also, let $T_1, ..., T_N$ be i.i.d. censoring times with a common unknown density g. Typically (and in the sequel) it is assumed that for $i = 1, ..., N$ $X_i \perp T_i$ (here, $\perp$ stands for statistical independence). In the context of partially observed data, instead of fully observing $X_1, ..., X_N$,

we observed an i.i.d. sequence $\{Y_i\,,\delta_i\}_{i=1}^{N}$, where $Y_i = \min\left(X_i\,,T_i\right)$ and $\delta_i = \mathbf{1}_{(X_i \leq T_i)}$. The function $\mathbb{1}_{(\cdot)}$ stands for the indicator function.

In this Chapter, we propose a linear estimator based on an orthogonal projection onto a defined multiresolution space V_J using empirical wavelet coefficients that are fully data driven. We derive an asymptotically unbiased estimator constructed from the complete sample based on a an inductive bias correction. Also, we provide estimates for the bias and large sample behavior of the expected $\mathbb{L}_2$ error based on the approach used by Stute (1995)[14]. In addition, we evaluate the performance of the proposed estimator via a simulation study using different exemplary unimodal and multimodal baseline distributions under different sample sizes. For this purpose, we chose an exponential censoring scheme that produces a censoring proportion of 40% on average. Finally, we apply the proposed estimator to real data-sets previously used in other published results in the field of non-parametric density estimation.

Our results are based on wavelets periodic on the interval $[0,1]$ and are derived under the assumption that both densities f and g are continuous and the survival function of the censoring random variable T is bounded from below by an exponentially decaying function. Also, we assume that the scaling function ϕ is absolutely integrable and the multiresolution space index J used for the projection is chosen as a function of the sample size N as $J = \lfloor \log_2(N) - \log_2(\log(N)) \rfloor$. The only assumption that we impose on the target density f is that it belongs to the s-sobolev space H^s.

2.1.1 Overview of previous and current work in the area

In the context of wavelets applied to density estimation with complete data, Donoho, et al. (1992) [28] proposed a wavelet estimator based on thresholded empirical wavelet coefficients and investigate the minimax rates of convergence over a wide range of Besov function classes $B_{\sigma pq}$. They choose the resolution of projection spaces such that the estimator achieves the

proper convergence rates. As it can be seen in recent literature, their work is fundamental for subsequent research in the field.

A work by Vanucci (1998) [30] provides overview of different wavelet-based density estimators, emphasizing their properties and comparison with classical estimators. In her paper, the author provides a general description of an orthonormal wavelet basis, focusing on the properties that are essential for the construction of wavelet density estimators. Also, a description of linear and thresholded density estimators is provided. This works constitutes a comprehensive reference for density estimation in the context of complete data.

Following the available results in the context of complete-data density estimation (i.e. no censoring), Pinheiro and Vidakovic (1997) [29] propose estimators of the square root of a density based on compactly supported wavelets. Their estimator is a bona-fide density with $\mathbb{L}_1$ norm equal to 1, taking care of possible negative values resulting from the usual estimation of the density f.

Now in the context of density estimation with censored data, Antoniadis et al. (1999) [20] proposed a wavelet method based on dividing the time axis into a dyadic number of intervals and counting the number of occurrences within each one. Then, they use wavelets smoothers based on wavelets on the interval (see [5]) to get the survival function of the observations. Also, they obtain the best possible asymptotic mean integrated square error (MISE) convergence rate under the assumption that the target density f is $r-$times continuously differentiable and the censoring density g is continuous.

Later on, Li (2003)[31] provides a non-linear wavelet-based density estimator under random censorship that uses a thresholded series expansion of the sub-density $f_1(x) = f(x)\mathbb{1}_{\{x \leq T\}}$ where $T < \tau_H$ and $\tau_H = \inf\{x : F_Y(x) = 1\}$. This approach is based on compactly supported ϕ and ψ (father and mother wavelet, respectively) and detail coefficients d_{jk} are thresh-

olded according to $\tilde{d}_{jk} = \hat{d}_{jk} \mathbb{1}_{\left\{|\hat{d}_{jk}|>\delta\right\}}$ for a suitable defined threshold δ and parameter $j = q$ for the wavelet expansion. In his work, Li provides and asymptotic expansion for the MISE and calculate the convergence rates under smoothness and regularity assumptions on the target density f. This work is then further extended in Li (2007) [32], where the minimax optimality of the thresholded wavelet-based estimator is investigated over a large range of Besov function classes.

One of the most recent works in the context of censored data was developed by Zou and Liang (2017) [33]. They define a non-linear wavelet estimator for the right censoring model in the case when the censoring indicator δ is missing at random. They develop an asymptotic expression for the MISE which is robust under the presence of discontinuities in f. Their estimator reduces to the one proposed by Li (2003) when the censoring indicator missing at random does not happen and a bandwidth in non-parametric estimation is close to zero.

2.1.2 About Periodic Wavelets

For the implementation of the functional estimator, we choose periodic wavelets as an orthonormal basis. Even though this kind of wavelets exhibit poor behaviour near the boundaries (when the analyzed function is not periodic, high amplitude wavelet coefficients are generated in the neighborhood of the boundaries) they are typically used due to the relatively simple numerical implementation and compact support. Also, as was suggested by Donoho and Johnstone (1994)[34], this simplification affects only a small number of wavelet coefficients at each resolution level.

Periodic wavelets in $[0, 1]$ are defined by a modification of the standard scaling and wavelet

functions:

$$\phi_{j,k}^{per}(x) = \sum_{l\in\mathbb{Z}} \phi_{j,k}(x-l)\,, \tag{2.1}$$

$$\psi_{j,k}^{per}(x) = \sum_{l\in\mathbb{Z}} \psi_{j,k}(x-l)\,. \tag{2.2}$$

It is possible to show, as in [35], that $\left\{\phi_{0,0}^{per}(x),\ \psi_{j,k}^{per}(x), 0 \le k \le 2^j-1, j \ge 0\right\}$ constitutes an orthonormal basis for $\mathbb{L}_2[0,1]$. Consequently, due to the hierarchical containment of the spaces, it follows that $\cup_{j=0}^{\infty} V_j^{per} = \mathbb{L}_2[0,1]$, where V_j^{per} is the space spanned by $\left\{\phi_{j,k}^{per}(x), 0 \le k \le 2^j-1\right\}$. This allows to represent a function f with support in $[0,1]$ as:

$$f(x) = \langle f(x), \phi_{0,0}^{per}(x)\rangle \phi_{0,0}^{per}(x) + \sum_{j\ge 0}\sum_{k=0}^{2^j-1} \langle f(x), \psi_{j,k}^{per}(x)\rangle \psi_{j,k}^{per}(x)\,. \tag{2.3}$$

Also, for a fixed $j = J$, we can obtain an orthogonal projection of $f(x)$ onto V_J denoted as $\mathbf{P}_J(f(x))$ given by:

$$\mathbf{P}_J(f(x)) = \sum_{k=0}^{2^J-1} \langle f(x), \phi_{J,k}^{per}(x)\rangle \phi_{J,k}^{per}(x) \tag{2.4}$$

Since periodized wavelets provide a basis for $\mathbb{L}_2([0,1])$, we have that $\parallel f(x) - \mathbf{P}_J(f(x)) \parallel_2 \to 0$ as $J \to \infty$. Also, it can be shown that $\parallel f(x) - \mathbf{P}_J(f(x)) \parallel_\infty \to 0$ as $J \to \infty$. Therefore, we can see that $\mathbf{P}_J(f(x))$ uniformly converges to f as $J \to \infty$. Similarly, as discussed in [5] it is possible to assess the approximation error for a certain density of interest f using a truncated projection (i.e. for a certain chosen detail space J). For example, using the s-th Sobolev norm of a function defined as:

$$\parallel f(x) \parallel_{H^s} = \sqrt{\int (1+|x|^2)^s |f(x)|^2 dx}\,, \tag{2.5}$$

one defines the H^s sobolev space, as the space that consists of all functions f whose s-

Sobolev norm exists and is finite. As it is shown in [5]:

$$\| f(x) - \mathbf{P}_J(f(x)) \|_2 \leq 2^{-J \cdot s} \cdot \| f \|_{H^s[0,1]} \ . \tag{2.6}$$

From (2.6), for a pre-specified $\epsilon > 0$ one can choose J such that $\| f(x) - \mathbf{P}_J(f(x)) \|_2 \leq \epsilon$. In fact, a possible choice of J could be:

$$J \geq -\lceil \frac{1}{s} \log_2 (\frac{\epsilon}{\| f \|_{H^s[0,1]}}) \rceil \, . \tag{2.7}$$

Therefore, it is possible to approximate a desired function to arbitrary precision using the MRA generated by a wavelet basis.

As a final comment to this brief introductory section about periodic wavelets, it is important to point out the relation between discrete wavelets coefficients (i.e. those obtained through DWT[1]) and the continuous wavelet coefficients (i.e. those obtained through the CWT[2]): As shown by Antoniadis and Bigot (2001)[36], we have that because of the difference in orthonormality conditions between the continuous and discrete case, we have that:

$$\begin{aligned} c_{j_0 k} &\approx \sqrt{n} \alpha_{j_0 k} \\ d_{jk} &\approx \sqrt{n} \beta_{jk} \end{aligned}$$

where $c_{j_0 k}$ and d_{jk} correspond to the discrete wavelet coefficients, n is the sample size and $\alpha_{j_0 k}$, β_{jk} are the coefficients corresponding to the CWT.

[1]Discrete wavelet transformation
[2]Continuous wavelet transformation

2.2 Survival Density Estimation for right-censored data using Periodized Wavelets

2.2.1 Problem statement, assumptions and derivation of the estimator for a density $f(x)$.

Consider a sample of iid lifetimes (non-negative) of the form $\tilde{X}_1, ..., \tilde{X}_N$ drawn from a random variable $\tilde{X} \sim \tilde{f}(\cdot)$, with unknown density $\tilde{f} \in \mathbb{L}_2(\mathbb{R})$. Furthermore, let $\tau_{\tilde{X}} = \inf\left\{\tilde{x} : \tilde{F}_{\tilde{X}}(\tilde{x}) = 1\right\}$, where $\tilde{F}_{\tilde{X}}(\tilde{x})$ corresponds to the cumulative density function (cdf) of the random variable $\tilde{X}$.

Define the target density (i.e. the density to be estimated) as $\tilde{f}_c(\tilde{x}) = \tilde{f}(\tilde{x})\mathbb{1}_{\left\{\tilde{x} \leq \tau_{\tilde{X}}\right\}}$, which corresponds to $\tilde{f}(\cdot)$ constrained to the interval $[0, \tau_{\tilde{X}}]$. This definition implies that $\tilde{f}_c(\tilde{x}) = \tilde{f}(\tilde{x})$, for $\tilde{x} \leq \tau_{\tilde{X}}$.

From the observed sample $\tilde{X}_1, ..., \tilde{X}_N$, and a pre-specified $\tau > 0$, define the normalized random variable $X = \frac{1}{\tau}\tilde{X}$. Then, it follows:

$$f_X(x) = \tau\, f_{\tilde{X}}(\tau x)\mathbb{1}_{\left\{x \leq \frac{\tau_{\tilde{X}}}{\tau}\right\}}, \tag{2.8}$$

for the domain-restricted density $\tilde{f}_c(\tilde{x})$.

Remarks

(i) If $\tau = \tau_{\tilde{X}}$ the normalized random variable X has support in [0,1] with density given by $f(x) = f_X(x)$.

(ii) In practice, since $\tilde{f}$ is not known, it is possible to select $\tau = \max\left\{\tilde{X}_1, ..., \tilde{X}_N\right\}$; this, since in general $\tilde{X}_{(N)} \xrightarrow{\mathbb{P}} \tau_{\tilde{X}}$ where the operator $\xrightarrow{\mathbb{P}}$ denotes convergence in probability.

(iii) Note that the definition $\tilde{f}_c(\tilde{x}) = \tilde{f}(\tilde{x})\mathbb{1}_{\left\{\tilde{x} \leq \tau_{\tilde{X}}\right\}}$ corresponds exactly to the conditional density $\tilde{f}_{\tilde{X}|\tilde{X} \leq \tau_{\tilde{X}}}(\tilde{x})$.

In the sequel, it will be assumed that the random variable X was obtained presented above, with a probability density of the form (2.8).

Representing $f(x)$ using Wavelets

Using a multiresolution analysis (MRA) based on periodized wavelets in $[0, 1]$, the density $f(\cdot)$ can be expressed as:

$$f(x) = \sum_{j \in \mathbb{Z}} \sum_{k \geq 0} d_{jk} \cdot \psi_{jk}^{per}(x) \,. \tag{2.9}$$

Using the hierarchical structure of the MRA, for a pre-specified multiresolution scale $J = J_0$, (2.9) can be expressed as:

$$f(x) = \sum_{k \in \mathbb{Z}} c_{J_0,k} \cdot \phi_{J_0,k}^{per}(x) + \sum_{j \geq J_0} \sum_{k \in \mathbb{Z}} d_{jk} \cdot \psi_{jk}^{per}(x) \,, \tag{2.10}$$

for $\phi_{jk}^{per}(x) = 2^{\frac{j}{2}} \phi^{per}(2^j x - k)$, and $\psi_{jk}^{per}(x) = 2^{\frac{j}{2}} \psi^{per}(2^j x - k)$ for $j, k \in \mathbb{Z}$.

Because periodic extensions of wavelets in $[0, 1]$ are used, the support of the scaling function $\phi_{jk}^{per}(x)$ and the wavelet function $\psi_{jk}^{per}(x)$ is $[k \cdot 2^{-j}, (k+1) \cdot 2^{-j}]$ where $k = 0, ..., 2^{j-1}$, and by the Strang-fix condition $j \geq 0$.

From (2.10), the summation over the MRA scale index j goes from J_0 to ∞. This implies that it is possible to approximate $f(\cdot)$ by truncating the summation up to scale index J^*. Therefore, it follows:

$$\hat{f}_{J^*}(x) = \sum_{k \in \mathbf{K}(J_0)} c_{J_0,k} \cdot \phi_{J_0,k}^{per}(x) + \sum_{j \geq J_0}^{J^*} \sum_{k \in \mathbf{K}(j)} d_{jk} \cdot \psi_{jk}^{per}(x) \,, \tag{2.11}$$

where $\mathbf{K}(J_0) = \left\{k \in \mathbb{N} \,|\, 0 \leq k \leq 2^{J_0 - 1}\right\}$ and $\mathbf{K}(j) = \left\{k \in \mathbb{N} \,|\, 0 \leq k \leq 2^{j-1}\right\}$. In the sequel, the value of J^* will be assumed to be selected as a function of the sample size N.

In the wavelet series approximation of $f(\cdot)$ defined by (2.11), the coefficients $c_{J_0,k}$ and d_{jk} are given by the orthogonal projection of $f(\cdot)$ onto each subspace $V_{J_0}^{per}$ and W_j^{per} in the MRA[3]. Here, $V_{J_0}^{per}$ and W_j^{per} correspond to the functional spaces spanned by $\left\{\phi_{J_0,k}^{per}, 0 \leq k \leq 2^{J_0}-1\right\}$, and $\left\{\psi_{j,k}^{per}, 0 \leq k \leq 2^j-1, J_0 \leq j \leq J^*\right\}$ respectively. Using this definitions, it follows:

$$c_{J_0,k} = \int_0^1 f(x) \cdot \phi_{J_0,k}^{per}(x)dx = \langle f(x), \phi_{J_0,k}^{per}(x)\rangle, \tag{2.12}$$

$$d_{jk} = \int_0^1 f(x) \cdot \psi_{j,k}^{per}(x)dx = \langle f(x), \psi_{j,k}^{per}(x)\rangle. \tag{2.13}$$

Clearly, since f is a probability density, (2.12) and (2.13) can be represented as:

$$c_{J_0,k} = \mathbb{E}_f[\phi_{J_0,k}^{per}(X)], \tag{2.14}$$

$$d_{jk} = \mathbb{E}_f[\psi_{j,k}^{per}(X)]. \tag{2.15}$$

Substituting (2.14) and (2.15) in (2.11), $\hat{f}_{J^*}(x)$ takes the form:

$$\hat{f}_{J^*}(x) = \sum_{k\in\mathbf{K}(J_0)} \mathbb{E}_f[\phi_{J_0,k}^{per}(X)] \cdot \phi_{J_0,k}^{per}(x) + \sum_{j\geq J_0}^{J^*} \sum_{k\in\mathbf{K}(j)} \mathbb{E}_f[\psi_{j,k}^{per}(X)] \cdot \psi_{jk}^{per}(x). \tag{2.16}$$

Using (2.16) and assuming $X_1, ..., X_N \sim f(\cdot)$ are iid, for $f(\cdot)$ unknown, it is possible to estimate the coefficients $c_{J_0,k}$ and d_{jk} from the sample as follows:

$$\tilde{c}_{J_0,k} = \frac{1}{N}\sum_{i=1}^{N} \phi_{J_0,k}^{per}(X_i), \tag{2.17}$$

$$\tilde{d}_{j,k} = \frac{1}{N}\sum_{i=1}^{N} \psi_{j,k}^{per}(X_i). \tag{2.18}$$

[3]In fact, from the MRA approach we have that $V_{J^*}^{per} = V_{J_0}^{per} \oplus \cup_{j=J_0}^{J^*} W_j^{per}$.

Therefore, the data-driven estimated density $\hat{f}_{J^*}(x)$ can be expressed as:

$$\hat{f}_{J^*}(x) = \sum_{k \in \mathbf{K}(J_0)} \left(\frac{1}{N} \sum_{i=1}^{N} \phi_{J_0,k}^{per}(X_i) \right) \cdot \phi_{J_0,k}^{per}(x) + \sum_{j \geq J_0}^{J^*} \sum_{k \in \mathbf{K}(j)} \left(\frac{1}{N} \sum_{i=1}^{N} \psi_{j,k}^{per}(X_i) \right) \cdot \psi_{jk}^{per}(x) \,. \tag{2.19}$$

From (2.19), it follows that $\hat{f}_{J^*}(x)$ was constructed based on fully observed realizations of the lifetime random variable X. Therefore, a natural extension is the modification of (2.19) to allow the introduction of partially observed (censored) samples; in particular, we will focus on the case of right-censored data.

2.2.2 Estimating $\hat{f}_{J^*}(x)$ in the case of partially observed data.

Consider a random variable X that is distributed with an unknown density $f(x)$. Furthermore, suppose an observed sample $\{Y_i, \delta_i\}_{i=1}^{N}$ that is composed on both fully, and partially observed realizations of X. In the sample, Y_i is defined as:

$$Y_i = \min(X_i, T_i) \quad i = 1, ..., N \,, \tag{2.20}$$

for $T_1, ..., T_N$ being iid random variables from an unknown distribution $T \sim g(t)$, which is the right-censoring sequence that causes some realizations from X to be partially observed, and is assumed to be independent of X. Also δ_i, representing the censoring indicator, is defined as:

$$\delta_i = \mathbb{1}_{(X_i \leq T_i)} \quad i = 1, ..., N \,, \tag{2.21}$$

where $\mathbb{1}_{(X_i \leq T_i)} = 1$ if and only if $(X_i \leq T_i)$ and 0 otherwise. Therefore, $\delta_i = 0$ represents a life-time X_i that was observed only up to time T_i, for which we can only conclude that $X_i > T_i$.

Since the observed data is $\{Y_i, \delta_i\}_{i=1}^{N}$, from (2.20) and (2.21), the joint distribution of the pair

(Y, δ) can be obtained as follows:

$$\begin{aligned}
\mathbb{P}(Y \leq y, \delta = 1) &= \mathbb{P}\left(\min(X, T) \leq y, X \leq T\right) \\
&= \int_{-\infty}^{y} \mathbb{P}\left(T \geq x\right) f(x) dx \\
&= \int_{-\infty}^{y} \left(1 - G(x)\right) f(x) dx \,, \qquad (2.22)
\end{aligned}$$

where $G(x) = \mathbb{P}\left(T \leq x\right)$. Similarly, for $\mathbb{P}(Y \leq y, \delta = 0)$ and a fixed y, it follows:

$$\begin{aligned}
\mathbb{P}(Y \leq y, \delta = 0) &= \mathbb{P}\left(\min(X, T) \leq y, X > T\right) \\
&= \int_{-\infty}^{+\infty} \mathbb{P}\left(T \leq \min(x, y)\right) f(x) dx \\
&= \int_{-\infty}^{y} \mathbb{P}\left(T \leq x\right) f(x) dx + \int_{y}^{+\infty} \mathbb{P}\left(T \leq y\right) f(x) dx \\
&= \int_{-\infty}^{y} G(x) f(x) dx + G(y) \int_{y}^{+\infty} f(x) dx \\
&= \int_{-\infty}^{y} G(x) f(x) dx + G(y)(1 - F(y)) \,. \qquad (2.23)
\end{aligned}$$

From (2.22) and (2.23) it follows:

$$f_{Y,\delta}(y, \delta) = f(y)^{\delta}(1 - G(y))^{\delta} g(y)^{1-\delta}(1 - F(y))^{1-\delta} \,. \qquad (2.24)$$

Similarly, from (2.24), the marginal density of the complete-data sample Y can be expressed as:

$$f_Y(y) = f_X(y)(1 - G_T(y)) + g_T(y)(1 - F_X(y)) \,, \qquad (2.25)$$

where the subscripts X and T are placed to emphasize the relation between each density function and its corresponding random variable.

Assuming $0 < G_T(y) < 1$, $f(x)$, from (2.25) it follows that $f(x)$ can be expressed as:

$$f_X(y) = \frac{f_Y(y)}{1 - G_T(y)} - \frac{(1 - F_X(y))g_T(y)}{1 - G_T(y)}\,. \tag{2.26}$$

As was mentioned in 2.2.1, the next sections assume that the observed data has been normalized according to $\tau = \max\{Y_1, ..., Y_N\}$, to restrict the support of the random variable X to the interval $[0, 1]$.

Complete Data Estimator

From (2.17) and (2.18), (2.25) and (2.26), the wavelet coefficients $c_{J_0,k}$ in the orthogonal wavelet expansion can be expressed as:

$$\begin{aligned} c_{J_0,k} &= \int_0^1 f(x) \cdot \phi^{per}_{J_0,k}(x)dx \\ &= \int_0^1 \left(\frac{f_Y(y)}{1 - G_T(y)} - \frac{(1 - F_X(y))g_T(y)}{1 - G_T(y)} \right) \cdot \phi^{per}_{J_0,k}(x)dx\,. \end{aligned}$$

Therefore:

$$c_{J_0,k} = \mathbb{E}_Y \left[\frac{\phi^{per}_{J_0,k}(Y)}{(1 - G(Y))} \right] - \mathbb{E}_T \left[\frac{(1 - F(Y))\phi^{per}_{J_0,k}(Y)}{(1 - G(Y))} \right]\,. \tag{2.27}$$

Similarly, for the coefficients $d_{j,k}$, it follows:

$$d_{j,k} = \mathbb{E}_Y \left[\frac{\psi^{per}_{j,k}(Y)}{(1 - G(Y))} \right] - \mathbb{E}_T \left[\frac{(1 - F(Y))\psi^{per}_{j,k}(Y)}{(1 - G(Y))} \right]\,. \tag{2.28}$$

Remarks:

(i) Expressions (2.27) and (2.28) are valid assuming $0 < G(y) < 1$ for $y \in [0, 1]$.

(ii) In the case of non-censored data, $G = \delta_\infty$ (i.e. Dirac at ∞) and, for $i = 1, ..., N$ $\delta_i = 1$. Therefore, $f_{Y,\delta} = f(x)$. Thus, (2.27) and (2.28) collapse into $\frac{1}{N}\sum_{i=1}^{N} \phi_{Jk}^{per}(Y_i)$ and $\frac{1}{N}\sum_{i=1}^{N} \psi_{Jk}^{per}(Y_i)$ respectively, which is the usual orthogonal-series density estimator scheme.

Using an empirical approach as in (2.17) and (2.18), it follows:

$$\tilde{c}_{J_0,k} = \frac{1}{N}\sum_{i=1}^{N} \frac{\phi_{J_0,k}^{per}(Y_i)}{1 - G(Y_i)} - \frac{1}{N}\sum_{i=1}^{N} \frac{\mathbb{1}_{(\delta_i=0)}(1 - F(Y_i))\phi_{J_0,k}^{per}(Y_i)}{(1 - G(Y_i))}\,, \tag{2.29}$$

provided $0 < G(Y_i) < 1$, for $i = 1, ..., N$.

Finally, the data-driven estimated density $\hat{f}_{J^*}(x)$ can be expressed as:

$$\hat{f}_{J^*}(x) = \sum_{k\in\mathbf{K}(J_0)} \left(\frac{1}{N}\sum_{i=1}^{N} \alpha_i^{\phi} \cdot \phi_{J_0,k}^{per}(Y_i)\right) \cdot \phi_{J_0,k}^{per}(x) + \sum_{j\geq J_0}^{J^*} \sum_{k\in\mathbf{K}(j)} \left(\frac{1}{N}\sum_{i=1}^{N} \alpha_i^{\psi} \cdot \psi_{j,k}^{per}(Y_i)\right) \cdot \psi_{jk}^{per}(x)\,, \tag{2.30}$$

where:

$$\alpha_i^{\phi} = \alpha_i^{\psi} = \frac{1}{1 - G(Y_i)} - \frac{\mathbb{1}_{(\delta_i=0)}(1 - F(Y_i))}{1 - G(Y_i)}\,, \tag{2.31}$$

for $i = 1, ..., N$.

As can be seen from (2.30) and (2.31), the computation of (2.30) implies addressing the following issues:

(i) Estimation of $G(Y_i)$ and $F(Y_i)$ for $i = 1, ..., N$.

(ii) Computation of α_i^{ϕ}, for $i = 1, ..., N$.

(iii) Computation of $\phi_{J_0,k}^{per}(Y_i)$ and $\psi_{j,k}^{per}(Y_i)$ for $i = 1, ..., N$, $j = J_0, ..., J^*$ and $0 \leq k \leq 2^{j-1}$.

Naturally, $G(Y_i)$ and $F(Y_i)$ can be obtained using the Kaplan-Meier estimator, which is well known for its robustness in the presence of censored data. Similarly, $\phi^{per}_{J_0,k}(Y_i)$ and $\psi^{per}_{j,k}(Y_i)$ we can computed using Daubechies-Lagarias algorithm.

Denote $\left\{(Y_{(i)}, \tilde{\delta}_{(i)})\right\}_{i=1}^{N}$ as the ranked sample $\{(Y_i, \delta_i)\}_{i=1}^{N}$ with respect to Y_i, where $\tilde{\delta}_{(i)} = 1 - \delta_{(i)}$. Using Kaplan-Meier, it follows:

$$\hat{G}_N(Y_{(i)}) = \hat{G}(Y_{(i)}) = \sum_{k=1}^{i} \left(\frac{\tilde{\delta}_{(k)}}{N-k+1} \prod_{j=1}^{k-1} (1 - \frac{\tilde{\delta}_{(j)}}{N-j+1}) \right), \tag{2.32}$$

$$\hat{F}_N(Y_{(i)}) = \hat{F}(Y_{(i)}) = \sum_{k=1}^{i} \left(\frac{\delta_{(k)}}{N-k+1} \prod_{j=1}^{k-1} (1 - \frac{\delta_{(j)}}{N-j+1}) \right), \tag{2.33}$$

for $i = 1, ..., N$. Thus, the estimated density $\hat{f}_{J^*}(x)$ can be expressed as:

$$\begin{aligned} \hat{f}_{J^*}(x) &= \sum_{k \in \mathbf{K}(J_0)} \left(\frac{1}{N} \sum_{i=1}^{N} \alpha^{\phi}_{(i)} \cdot \phi^{per}_{J_0,k}(Y_{(i)}) \right) \cdot \phi^{per}_{J_0,k}(x) \\ &\quad + \sum_{j \geq J_0}^{J^*} \sum_{k \in \mathbf{K}(j)} \left(\frac{1}{N} \sum_{i=1}^{N} \alpha^{\psi}_{(i)} \cdot \psi^{per}_{j,k}(Y_{(i)}) \right) \cdot \psi^{per}_{jk}(x), \end{aligned} \tag{2.34}$$

where:

$$\alpha^{\phi}_{(i)} = \alpha^{\psi}_{(i)} = \frac{1}{1 - \hat{G}(Y_{(i)})} - \frac{\mathbb{1}_{(\delta_i = 0)}(1 - \hat{F}(Y_{(i)}))}{1 - \hat{G}(Y_{(i)})}, \tag{2.35}$$

for $0 < \hat{G}(Y_{(i)}) < 1$, $i \in \subset \{1, ..., N\}$, $\mathbf{K}(J_0) = \left\{0, 1, ..., 2^{J_0} - 1\right\}$, and $\mathbf{K}(j) = \{0, 1, ..., 2^j - 1;\ j \geq J_0\}$.

From section 2.1.2, for a properly chosen multiresolution index J, the estimated density $\hat{f}_J(x)$ can be approximated by a truncated projection $\mathbf{P}_J(f(x))$ onto a multiresolution space V_J spanned by the functions $\left\{\phi^{per}_{Jk},\ 0 \leq k \leq 2^J - 1\right\}$. Under this setting, $\hat{f}_{J^*}(x)$ takes the

form:

$$\hat{f}_J(x) = \sum_{k=0}^{2^J-1} \tilde{c_{Jk}} \cdot \phi_{J,k}^{per}(x)\,, \tag{2.36}$$

where:

$$\tilde{c_{Jk}} = \frac{1}{N} \sum_{i=1}^{N} \alpha_{(i)}^{\phi} \cdot \phi_{J,k}^{per}(Y_{(i)})\,. \tag{2.37}$$

Note here that in the expansion, we only use the scaling functions $\phi_{J,k}^{per}(\cdot)$ evaluated at the sample points. This representation is equivalent to (2.11), where the detail coefficients $d_{j,k}, k = 0, ..., J-1$ can be obtained from $\left\{c_{J,k}, k = 0, ..., 2^J - 1\right\}$ using Mallat's algorithm.

Partial-Data Estimator assuming $G(y)$ is known.

From definition (2.36), using an iterative bias-correction procedure it is possible to obtain an unbiased estimator for (2.36), which is given by:

$$\hat{f}^{PD}(x) = \sum_{k=0}^{2^J-1} \tilde{c}_{Jk} \cdot \phi_{J,k}^{per}(x)\,, \tag{2.38}$$

where:

$$\tilde{c}_{Jk} = \frac{1}{N} \sum_{i=1}^{N} \frac{\mathbb{1}_{(\delta_i=1)}}{1 - \hat{G}(Y_i)} \phi_{Jk}^{per}(Y_i), \text{ and} \tag{2.39}$$

$$\mathbb{E}\left[\tilde{c}_{Jk}\right] = c_{Jk}\,. \tag{2.40}$$

The corresponding derivation can be found in section A.1 of the appendix.

Remark From (2.39), it is possible to observe that the "partial data" definition comes from the fact that the estimator uses only the samples corresponding to actual observations of the survival time X, as opposed to (2.36) which uses the complete sample $Y_1, ..., Y_N$. A similar estimator is proposed by Efromovich in [19] using a fourier basis instead of wavelets. For

the rest of the sequel, we will focus our theoretical Analysis in this type of estimator.

2.2.3 Statistical properties of the Partial Data Estimator assuming $G(y)$ is known .

Mean Square Consistency.

Now we investigate the mean-square convergence of the estimator $\hat{f}^{PD}(x)$.

Lemma 2.2.1. *Define:*

$$\mu_J(x) = \mathbb{E}\left[\hat{f}^{PD}(x)\right] = f_J(x)\,, \tag{2.41}$$

$$\sigma_J^2(x) = Var\left[\hat{f}^{PD}(x)\right]\,. \tag{2.42}$$

Assume the following conditions are satisfied:

(i) The scaling function ϕ that generates the orthonormal set $\left\{\phi_{Jk}^{per}, 0 \leq k \leq 2^J - 1\right\}$ has compact support and satisfies $||\theta_\phi(x)||_\infty = C < \infty$, for $\theta_\phi(x) := \sum_{r\in\mathbb{Z}} |\phi(x-r)|$.

(ii) $\exists\, F \in \mathbb{L}_2(\mathbb{R})$ such that $|K(x,y)| \leq F(x-y)$, for all $x, y \in \mathbb{R}$, where

$$K(x,y) = \sum_{k\in\mathbb{Z}} \phi(x-k)\phi(y-k).$$

(iii) For $s = m+1$, $m \geq 1$, integer, $\int |x|^s F(x)dx < \infty$.

(iv) $\int (y-x)^l K(x,y)dy = \delta_{0,l}$ for $l = 0, ..., s$.

(v) The density f belongs to the s-sobolev space $W_2^s([0,1], A)$, $A > 0$ defined as:

$$W_2^s([0,1],A) = \left\{f \mid f \in \mathbb{L}_2([0,1]),\, \exists\, f^{(1)}, ..., f^{(s)} \text{ s.t. } f^{(l)} \in \mathbb{L}_2([0,1]),\, l = 1, ..., s,\;\; ||f||_\infty \leq A\right\}.$$

Then, it follows:

$$\sup_{f \in W_2^s([0,1],A)} \mathbb{E}\left[||\hat{f}^{PD}(x) - f(x)||_2^2\right] \leq C_1 \frac{2^J}{N} + C_2 2^{-2sJ}\,, \textit{and} \tag{2.43}$$

for $J = \lfloor \log_2(N) - \log_2(\log(N)) \rfloor$*:*

$$\sigma_J^2(x) = \mathcal{O}(\log(N)^{-1})\,, \tag{2.44}$$

$$\mathbb{E}\left[\| f(x) - \hat{f}^{PD}(x) \|_2^2\right] \leq \mathcal{O}(N^{-s}\log(N)^s) \tag{2.45}$$

for $C_1 > 0\,,\ C_2 > 0$ *independent of* J *and* N*, provided* $\exists\ \alpha_1 \mid 0 < \alpha_1 < \infty$*,* $C_T \in (0,1)$ *such that* $(1 - G(y)) \geq C_T e^{-\alpha_1 y}$ *for* $y \in [0,1)$*, and* $0 \leq F(y) \leq 1\ \forall y \in [0,1]$.

The proof can be found in section A.2 of the appendix.

Based on (2.44), it is possible to observe that $\sigma_J^2(x) \to 0$ as $N \to \infty$, which implies that $\hat{f}^{PD}(x)$ is consistent for $f(x)$, for all $x \in [0,1]$ and $f \in W_2^s([0,1],A)$.

Remarks

Note that from (2.45), it is possible to choose the multiresolution level J such that the upper bound for the $\mathbb{L}_2$ risk is minimized. In this context, it is possible to show that $J^*(N) = \frac{1}{2s+1}\log_2\left(\frac{2sC_2}{C_1}\right) + \frac{1}{2s+1}\log_2(N)$ achieves that result. Moreover, under this choice of J, it follows:

$$\sup_{f \in W_2^s([0,1],A)} \mathbb{E}\left[||\hat{f}^{PD}(x) - f(x)||_2^2\right] \leq \tilde{C} N^{-\frac{2s}{2s+1}}\,,$$

for a constant $\tilde{C} > 0$, independent of N and s.

2.2.4 Statistical properties for Partial Data Estimator assuming $G(y)$ unknown.

In the previous section, we showed that $f^{PD}(x)$ is unbiased for $f_J(x)$ and mean square consistent for $f(x) \in W_2^s([0,1], A)$, assuming G known and the multiresolution index J for the orthogonal projection onto the space V_J was chosen as $J = \lfloor \log_2(N) - \log_2(\log(N)) \rfloor$.

Naturally, assuming G is known may be questionable because of both the nature of the nonparametric density estimation approach, and its practical application. In most of real life cases neither the target density f, nor the censoring density g are known, so making assumptions about them could undermine the robustness and quality of the estimated functions.

In this section we approach the problem of deriving the partial-data estimator using the data driven wavelet coefficients proposed in (2.39). In particular, we investigate the statistical properties of the partial data estimator through the application the methodology proposed by Stute (1995) [14] that approximates Kaplan-Meier integrals by the average of i.i.d. random variables plus a remainder that decays to zero at a certain rate.

Asymptotic Unbiasedness.

As was proposed in (2.39), $\tilde{c}_{Jk} = \frac{1}{N}\sum_{i=1}^{N} \frac{\mathbb{1}_{(\delta_{(i)}=1)}}{1-\hat{G}(Y_i)} \phi_{Jk}^{per}(Y_i)$. Using the methodology and results proposed by Stute in [14], and assumptions defined in B.3.1, it follows:

$$\sum_{i=1}^{N} W_{(i)} \phi_{Jk}^{per}(Y_{(i)}) = \frac{1}{N}\sum_{i=1}^{N} \delta_i \phi_{Jk}^{per}(Y_i)\gamma_0(Y_i) + \frac{1}{N}\sum_{i=1}^{N} U_i + R_N\,, \tag{2.46}$$

where $W_{(i)} = d\hat{F}_N(x)$ is the Kaplan-Meier probability mass function of the random variable X based on the sample, $\gamma_0(Y_i) = \frac{1}{1-G_T(Y_i)}$ and $U_i = (1-\delta_i)\gamma_1(Y_i) - \gamma_2(Y_i)$ for $i = 1, ..., N$.

Similarly, $\gamma_1(x) = \gamma_{1,Jk}(x)$ and $\gamma_2(x) = \gamma_{2,Jk}(x)$ are given by the following expressions:

$$
\begin{aligned}
\gamma_{1,Jk}(x) &= \frac{1}{1-F_Y(x)} \int_x^{\tau_H} \phi_{Jk}^{per}(u) f_X(u) du \,, \\
\gamma_{2,Jk}(x) &= \int_{-\infty}^{\tau_H} C(\min\{x,u\}) \phi_{Jk}^{per}(u) f_X(u) du \,, \text{where} \\
C(x) &= \int_{-\infty}^{x^-} \frac{g_T(u)du}{(1-F_Y(u))(1-G_T(u))} \,.
\end{aligned}
$$

In addition, assume the following conditions are satisfied (from Stute [14]):

$$
\int \phi^2(x) \gamma_0^2(x) f_{Y,\delta=1}(x) dx < \infty \,, \tag{2.47}
$$

$$
\int |\phi(x)| \sqrt{C(x)} f_X(x) dx < \infty \,. \tag{2.48}
$$

Condition (2.47) corresponds to the requirement of finite second moment (modified) on the scaling function $\phi(x)$, while condition (2.48) incorporates a modification on the first moment of $\phi(x)$ with respect to f_X that allows to control de bias in $\int \phi_{Jk}^{per}(u) \hat{f}_N(u) du$. For further details, see [14] and [37].

From the definitions above, it follows:

$$
\mathbb{E}\left[\phi_{Jk}^{per}(Y) \delta \gamma_0(Y)\right] = c_{Jk} \,, \tag{2.49}
$$

assuming $x < \tau_H$ for $\tau_H = \inf\{x : F_Y(x) = 1\}$.

Also, from (2.32) and (2.33), it follows that $d\hat{F}_N(x) = \hat{f}_N(x)$; indeed:

$$
d\hat{F}_N(x) = \begin{cases} 0 & \text{if } x \notin \left\{Y_{(1)}, ..., Y_{(N)}\right\} \\ \frac{\delta_{(i)}}{N-i+1} \prod_{j=1}^{i-1} \left(1 - \frac{\delta_{(j)}}{n-j+1}\right) & \text{if } x = Y_{(i)} \,, i = 1, ..., N \end{cases}
$$

After some algebra, it follows:

$$d\hat{F}_N(x) = \frac{\delta_{(i)}}{N-i+1} \prod_{j=1}^{i-1} \left(\frac{n-j}{n-j+1} \right)^{\delta_{(j)}} . \tag{2.50}$$

Moreover, $\frac{1}{1-\hat{G}_N(Y_{(i)})}$ can be expressed as:

$$\frac{1}{1-\hat{G}_N(Y_{(i)})} = \frac{N}{N-i+1} \prod_{j=1}^{i-1} \left(\frac{n-j}{n-j+1} \right)^{\delta_{(j)}} . \tag{2.51}$$

Therefore, putting together (2.50) and (2.51), it follows:

$$\frac{\delta_{(i)}}{N(1-\hat{G}_N(Y_{(i)}))} = \frac{\delta_{(i)}}{N-i+1} \prod_{j=1}^{i-1} \left(\frac{n-j}{n-j+1} \right)^{\delta_{(j)}} = d\hat{F}_N(x) . \tag{2.52}$$

These results altogether imply:

$$\int \phi_{Jk}^{per}(u) \hat{f}_N(u) du = \tilde{c}_{Jk} . \tag{2.53}$$

From Stute (1995), results (2.47)-(2.53) imply that (2.46) can be expressed as:

$$\int \phi_{Jk}^{per}(u) \hat{f}_N(u) du = \frac{1}{N} \sum_{i=1}^{N} \delta_i \phi_{Jk}^{per}(Y_i) \gamma_0(Y_i) + \frac{1}{N} \sum_{i=1}^{N} U_i + R_N , \tag{2.54}$$

where U_i i.i.d. for $i = 1, ..., N$ with $\mathbb{E}[U_1] = 0$, $\mathbb{E}[U_1^2] = \sigma^2 < \infty$ and $|R_N| = \mathcal{O}(N^{-1} \log(N))$.

Therefore:

$$\begin{aligned} \mathbb{E}\left[\int \phi_{Jk}^{per}(u) \hat{f}_N(u) du \right] &= \mathbb{E}\left[\frac{1}{N} \sum_{i=1}^{N} \delta_i \phi_{Jk}^{per}(Y_i) \gamma_0(Y_i) \right] + \mathbb{E}\left[\frac{1}{N} \sum_{i=1}^{N} U_i \right] + \mathcal{O}(N^{-1} \log(N)) , \\ &= c_{Jk} + \mathcal{O}(N^{-1} \log(N)) . \end{aligned} \tag{2.55}$$

Thus, $bias(\tilde{c}_{Jk}) = \mathcal{O}(N^{-1}\log(N))$, which implies that the partial data approach is asymptotically unbiased. The exact bias can be obtained by following the details presented in [14].

$\mathbb{L}_2$ *Risk Analysis.*

Following the same methodology and assumptions used in the previous section, we investigate the estimation error for the partial data approach, in the case where G is unknown.

Lemma 2.2.2. *Under the assumptions and definitions stated in B.3.1 and 2.2.4, by choosing* $J = \lfloor \log_2(N) - \log_2(\log(N)) \rfloor$*, it follows:*

$$\sup_{f \in W_2^s([0,1],A)} \mathbb{E}\left[\| f(x) - \hat{f}^{PD}(x) \|_2^2 \right] = \mathcal{O}(N^{-s}\log(N)^s)\,. \tag{2.56}$$

$$\tag{2.57}$$

The corresponding proofs can be found in section A.3 of the appendix.

Remarks

(i) Observe that by following the same methodology as in A.2, it is possible to obtain:

$$\sup_{f \in W_2^s([0,1],A)} \mathbb{E}\left[||\hat{f}^{PD}(x) - f(x)||_2^2 \right] \leq C_1 \frac{2^J}{N} + C_2 2^{-2sJ}\,,$$

for $C_1 = \frac{||F||_2^2 e^{2\gamma}}{C^2}$ and $C_2 > 0$, independent of N and J, provided that $\exists\ \gamma > 0$, and $C_T \in (0,1)$ such that $(1 - \hat{G}_N(y)) \geq C_T e^{-\gamma y}$ for $y \in [0,1)$.

(ii) The last result implies that by choosing $J^*(N) = \frac{1}{2s+1}\log_2\left(\frac{2s\,C_2}{C_1}\right) + \frac{1}{2s+1}\log_2(N)$, the $\mathbb{L}_2$ risk of the estimator $\hat{f}^{PD}(x)$ (when G is unknown) is also mean square consistent, and achieves a convergence rate of the order $\sim N^{-\frac{2s}{2s+1}}$. This implies that as long as the empirical survival function of the censoring random variable obtained from the

Kaplan-Meier estimator is bounded from below by an exponentially decaying function, the knowledge of the its cdf does not affects the statistical properties of the estimator.

Limiting Distribution.

In this section, we investigate the limiting distribution of the partial data estimator $\hat{f}^{PD}(x)$. Similarly as in sections 2.2.4 and 2.2.4, we will use results proposed in [14] as framework for our analysis.

As seen in (2.54), (2.55), Theorem 1.1 of [14] and the SLLN (Strong Law of Large Numbers), the following results hold:

$$\frac{1}{N}\sum_{i=1}^{N}\frac{\delta_i \phi_{Jk}^{per}(Y_i)}{1-G(Y_i)} \quad \xrightarrow{\mathbb{P}} \quad c_{Jk}\,, \tag{2.58}$$

$$R_N \quad \xrightarrow{\mathbb{P}} \quad 0\,, \tag{2.59}$$

where (2.58) follows from the SLLN (assuming the expectation is finite), and (2.59) from the fact that $|R_N| = \mathcal{O}_{\mathbb{P}}(\frac{1}{\sqrt{N}})$, as shown in [14]. Using Slutzky's theorem (see [38]), it follows:

$$\tilde{c}_{Jk} - \frac{1}{N}\sum_{i=1}^{N}\frac{\delta_i \phi_{Jk}^{per}(Y_i)}{1-G(Y_i)} - R_N \overset{\mathbb{D}}{=} \frac{1}{N}\sum_{i=1}^{N} U_i\,, \tag{2.60}$$

where $U_i = (1-\delta_i)\gamma_1(Y_i) - \gamma_2(Y_i)$, $i = 1, ..., N$ are i.i.d. zero-mean and finite variance random variables with $\mathbb{E}\left[U_1^2\right] = \sigma^2$. Also, from the definitions of $\gamma_1(x)$ and $\gamma_2(x)$, it follows that $\sigma^2 = \sigma_{Jk}^2$ since it depends on the scaling function $\phi_{Jk}^{per}(x)$. Now, by the CLT (Central Limit Theorem) it follows:

$$\frac{1}{\sqrt{N}}\sum_{i=1}^{N} U_i \xrightarrow{\mathbb{D}} N(0, \sigma_{Jk}^2)\,. \tag{2.61}$$

Combining results (2.60), (2.61), Slutzky's theorem implies:

$$\sqrt{N}(\tilde{c}_{Jk} - c_{Jk}) \xrightarrow{\mathbb{D}} N(0, \sigma^2_{Jk}) \,. \tag{2.62}$$

Similarly, it follows:

$$\sqrt{N}\left(\hat{f}^{PD}(x) - f(x)\right) = \sum_{k=0}^{2^J-1} \sqrt{N}(\tilde{c}_{Jk} - c_{Jk})\phi^{per}_{Jk}(x) \,. \tag{2.63}$$

Lemma 2.2.3. *For $c > 0$, $\beta > 1$ and x in a neighborhood of 1, assume the following conditions hold:*

(i) $(1 - F_X) \sim c(1 - G_T)^\beta$

(ii) $C(x) \leq \frac{1}{(1-F_X(x))(1-G_T(x))}$

Then, it follows:

$$\sqrt{N}\left(\hat{f}^{PD}(x) - f(x)\right) \xrightarrow{\mathbb{D}} N\left(0\,, \sum_{k=0}^{2^J-1} \sigma^2_{Jk}(\phi^{per}_{Jk}(x))^2 + 2\sum_{k<l} \sigma_{J,kl}\phi^{per}_{Jk}(x)\phi^{per}_{Jl}(x)\right), \tag{2.64}$$

for $k, l = 0, ..., 2^J - 1$,

$$\sigma^2_{Jk} \;=\; \mathbb{E}\left[((1-\delta)\gamma_{1,Jk}(Y) - \gamma_{2,Jk}(Y))^2\right],$$

and

$$\sigma_{J,kl} = \mathbb{E}\left[\frac{\delta^2\phi^{per}_{Jk}(Y)\phi^{per}_{Jl}(Y)}{(1-G(Y))^2} - c_{Jk}c_{Jl}\right],$$

provided assumptions detailed in B.3.1, (2.47), (2.48) are satisfied and $J = \lfloor \log_2(N) - \log_2(\log(N)) \rfloor$.

The corresponding proof can be found in section A.4 of the appendix.

Remarks

(a) Note that condition (iiii) indicates that there is enough information about the tails of the target density f; also, the larger the values of β, the heavier the tails of the censoring distribution, compared to the tails of the survival time distribution.

(b) As described in [37] and [14], the condition of $\beta > 1$ is required so that the bias of $\tilde{c}_{Jk} - c_{Jk}$ achieves a convergence rate better that $a\, N^{-\frac{1}{2}}$ for some non-vanishing a which may cause that (2.46) is no longer valid.

(c) As it can be seen in (A.67), the fact that $\hat{f}^{PD}(x)$ presents asymptotic normality brings to discussion the possibility that the estimates may be negative, as was previously mentioned in 2.2.4 and discussed in [18].

2.3 Simulation Study

In this section, we investigate the estimation performance of $\hat{f}^{PD}(x)$ and evaluate it with respect to the AMSE (Average Mean Squared Error) via a simulation study. For this purpose, we choose a set of exemplary baseline functions that resemble important features that continuous survival times that can be encountered in practice could posses. To simplify the simulations, we chose functions that are supported in an interval close to [0,1]. A brief description of each chosen function follows:

(a) **Delta.** This corresponds to a R.V. $X \sim N(0.5\,, 0.02^2)$. The idea is to have an extreme spatially heterogeneous curve that has support over a small region. The goal is to represent situations when a short but abrupt deviation from a process may happen.

(b) **Normal.** This corresponds to the usual Normal distribution with parameters $\mu = 0.5$ and $\sigma = 0.15$.

(c) **Bimodal.** This corresponds to a mixture of 2 Normal distributions and has the form $f(x) = 0.5\,X_1 + 0.5\,X_2$ where $X_1 \sim N(0.4\,, 0.12^2)$ and $X_2 \sim N(0.7\,, 0.08^2)$.

(d) **Strata.** This corresponds to a mixture of 2 Normal distributions and has the form $f(x) = 0.5\,X_1 + 0.5\,X_2$ where $X_1 \sim N(0.2\,, 0.06^2)$ and $X_2 \sim N(0.7\,, 0.08^2)$. The idea is to represent a function that is supported over 2 separate subintervals.

(e) **Multimodal.** This functions corresponds to a mixture of 3 Normal distributions and has the form $f(x) = \frac{1}{3}\,X_1 + \frac{1}{3}\,X_2 + \frac{1}{3}\,X_3$ where $X_1 \sim N(0.2\,, 0.06^2)$, $X_2 \sim N(0.5\,, 0.05^2)$ and $X_3 \sim N(0.7\,, 0.05^2)$. The idea of this function is to represent multimodal survival times which are expected to occur in heterogeneous populations.

An advantage of using simulated data in the case of censored data is that the values for both X and T are known for all samples; also, the controlled-environment approach allows the investigation of the estimator's performance for different sample sizes and censoring schemes. For testing purposes, we choose a censoring random variable $T \sim Exp(\lambda)$ with $\lambda = 0.8$, which produces approximately 45% censored samples at each generated datasets. Also, we use samples sizes $N = 100, 200, 500, 1000$ and measure the global error given by:

$$\hat{MSE} = \frac{1}{B}\sum_{b=1}^{B}\frac{1}{N}\sum_{i=1}^{N}\left(f(x_i) - \hat{f}_{N,b}(x_i)\right)^2\,, \tag{2.65}$$

where B is the number of replications of the experiment and N is the number of samples. For all experiments we choose $B = 1000$ and the wavelet filter Symmlet5. To implement simulations, we generate 2 independent random samples $\{X_i\}_{i=1}^{N}$ and $\{T_i\}_{i=1}^{N}$. X_i random variables were drawn from each one of the aforementioned distributions, while $T_i \overset{i.i.d.}{\sim} Exp(\lambda)$. Also, we included in the simulation study the complete data estimator as we found of interest to

observe its performance and compare it to the partial data approach.

2.3.1 Simulation Results.

In this section, we summarize the results obtained for each baseline distribution. In particular, the following results are provided:

(a) Tables 2.1 to 2.5 present details for AMSE results obtained for each baseline distribution used in the study.

(b) In figures 2.1 - 2.5, dashed lines (red and blue) correspond to the average estimates for $\hat{f}^{PD}(x)$, computed at each data point x from all $B = 1000$ replications. The black line indicates the actual density function and the light blue and blue continuous lines represents the best estimates among all replications (i.e. the one with the smallest AMSE).

(c) In figures 2.6 - 2.10, dashed lines (red and green) correspond to the empirical 95% quantiles computed at each data point x from all $B = 1000$ replications, for $\hat{f}^{b}(x)$ and $\hat{f}^{PD}(x)$ respectively. The blue and magenta lines show the average density estimates for the complete and partial data approach, respectively. The black line indicates the actual density function.

(d) Figure 2.11a shows the AMSE vs. sample size plot.

(e) Figure 2.11b exemplifies the asymptotic normality behavior of the density estimates, as proposed in 2.2.4.

2.3.2 Remarks and comments.

(i) From the resulting figures, it is possible to observe that the proposed estimator is able to accurately estimate the underlying density in the presence of right-censored observa-

Table 2.1: AMSE results for Delta distribution with Partial data estimator.

PD Estimator	$N = 100$	$N = 200$	$N = 500$	$N = 1000$
Mean AMSE	2.5954	0.3674	0.1856	0.2216
St.Dev. AMSE	0.0986	0.1680	0.1301	0.1009
Min AMSE	2.5149	0.2010	0.0112	0.0216
Max AMSE	3.5061	1.3967	0.8243	0.6893

Table 2.2: AMSE results for Normal distribution with Partial data estimator.

PD Estimator	$N = 100$	$N = 200$	$N = 500$	$N = 1000$
Mean AMSE	0.1219	0.0821	0.0385	0.0214
St.Dev. AMSE	0.0858	0.0524	0.0230	0.0129
Min AMSE	0.0036	0.0086	0.0037	0.0031
Max AMSE	0.5426	0.5058	0.1764	0.0872

Table 2.3: AMSE results for Bimodal distribution with Partial data estimator.

PD Estimator	$N = 100$	$N = 200$	$N = 500$	$N = 1000$
Mean AMSE	0.1764	0.1041	0.0494	0.0296
St.Dev. AMSE	0.1110	0.0620	0.0275	0.0175
Min AMSE	0.0175	0.0123	0.0041	0.0030
Max AMSE	0.9177	0.4933	0.1850	0.1323

Table 2.4: AMSE results for Strata distribution with Partial data estimator.

PD Estimator	$N = 100$	$N = 200$	$N = 500$	$N = 1000$
Mean AMSE	0.2468	0.1422	0.0731	0.0491
St.Dev. AMSE	0.1485	0.0854	0.0420	0.0243
Min AMSE	0.0225	0.0130	0.0078	0.0102
Max AMSE	1.0432	0.6857	0.3657	0.1783

Table 2.5: AMSE results for Multimodal distribution with Partial data estimator.

PD Estimator	$N = 100$	$N = 200$	$N = 500$	$N = 1000$
Mean AMSE	0.3838	0.2183	0.1321	0.2216
St.Dev. AMSE	0.1595	0.1108	0.0652	0.2193
Min AMSE	0.0619	0.0289	0.0171	0.2193
Max AMSE	1.0382	0.5863	0.4589	0.2193

tions. Also, the observed values for the estimates (Best and Mean) with respect to the sample size, suggests a bias effect in the vicinity of the underlying distribution modes.

(ii) In terms of the sensibility of the estimator's performance to the kind of scaling functions used to span the projection subspace, we observed during our experiments that results obtained using Symmlets, Coiflets and Daubechies wavelets are similar. The main difference relates to the computational efficiency of the algorithm, which is primarily affected by the length of the corresponding filter.

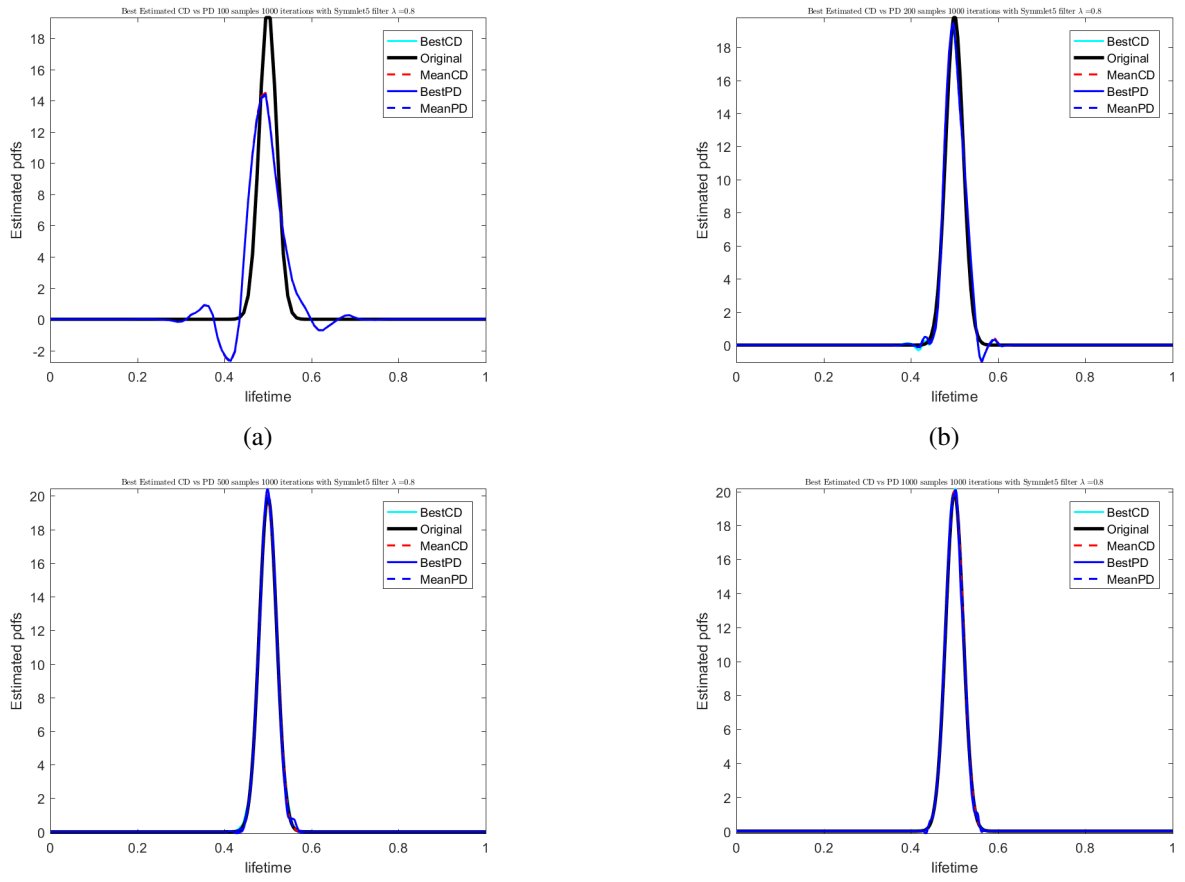


Figure 2.1: Estimate results for Delta distribution, $N = 100, 200, 500, 1000$ using Symmlet5.

(iii) From the quantiles plots, the empirical quantiles of the estimated densities contain the actual values of the target density in most of its support. Moreover, for all baseline distributions except for the Multimodal, this is the case. On the contrary, the regions where the 95% empirical quantiles do not contain the true density value are observed to occur in the vicinities of the distribution modes. This could be caused by the choice of the multiresolution index J, the post-processing smoothing procedure and/or by the censoring effect.

(iv) As the sample size increases, it was observed that the interval $|\hat{f}_{N\,0.975}(x) - \hat{f}_{N\,0.025}|$ monotonically decreases in coherence with the theoretical convergence results shown

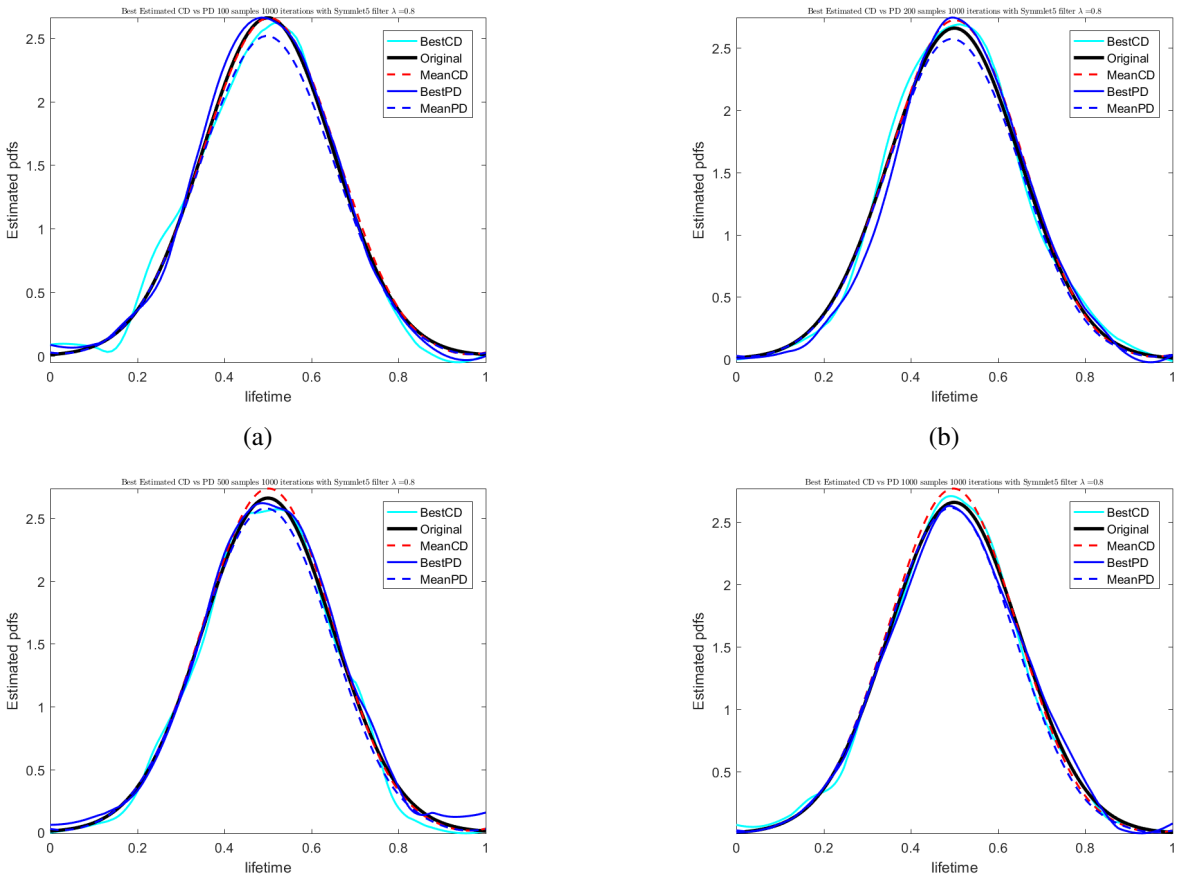


Figure 2.2: Estimate results for Normal distribution, $N = 100, 200, 500, 1000$ using Symmlet5.

in section 2.2.4.

(v) From the AMSE plot (2.11a), it is possible to observe that all baseline distributions present a similar error decay behavior. Moreover, results contained in tables 2.1 to 2.5, imply that as N grows, the standard deviation and range of AMSE decays in accordance with the convergence rates proposed for both estimators.

(vi) Figure 2.11b, suggest normality of the estimated density values, which is coherent with results presented in section 2.2.4. This property of the estimators allows the construction of confidence intervals and the application of standard statistical inference tools that could be useful in practical situations. However, to make this applicable, the Vari-

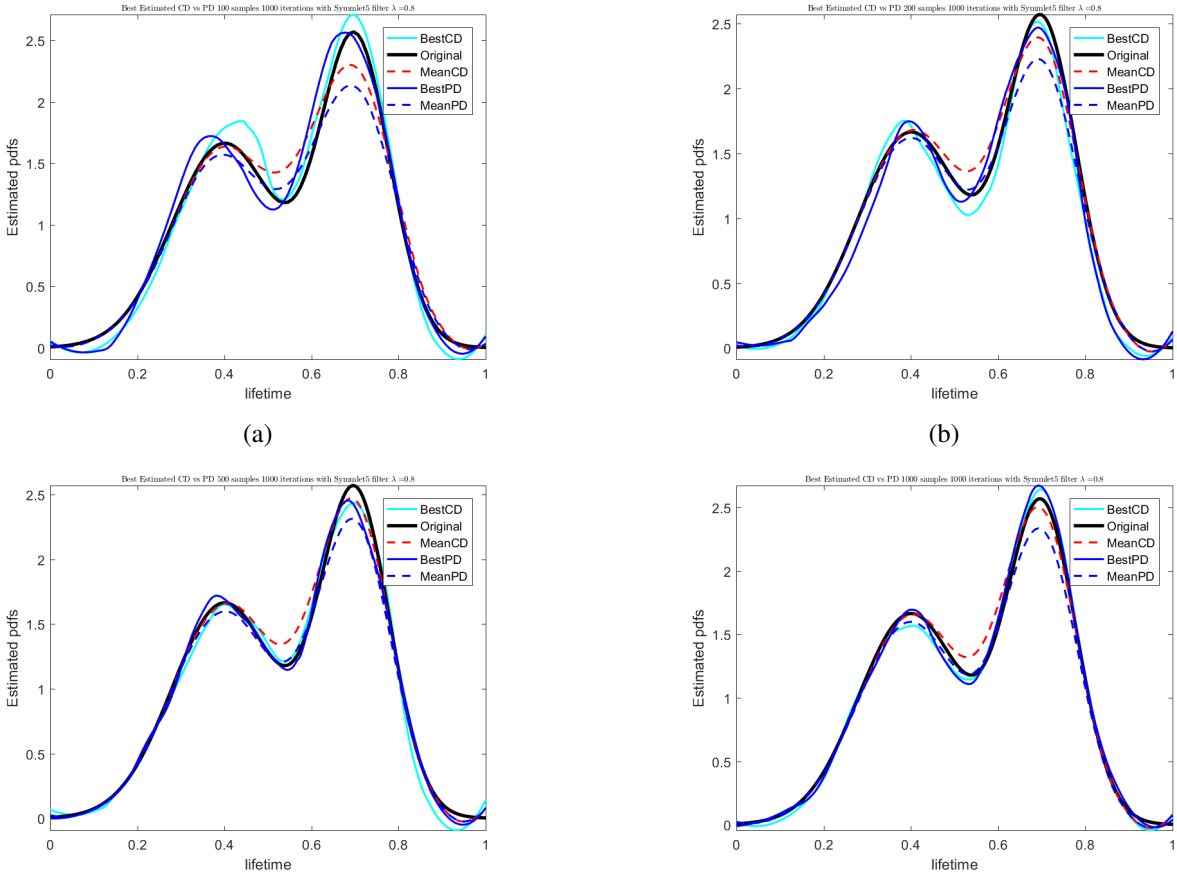


Figure 2.3: Estimate results for Bimodal distribution, $N = 100, 200, 500, 1000$ using Symmlet5.

ance of $\hat{f}^{PD}(x)$ in accordance with (A.67) needs to be estimated.

(vii) In most of presented figures it is possible to observe that at the extremes of the support sometimes the estimated density values are slightly negative. This effect is consistent with the boundary effect noted in [18] by Antoniadis. As was mentioned in the introduction, a possible remedial measure could be application the approach proposed by [29]. Another possibility is using $\hat{f}_+(x) = \max\left\{0, \hat{f}^{PD}(x)\right\}$, as proposed in [20].

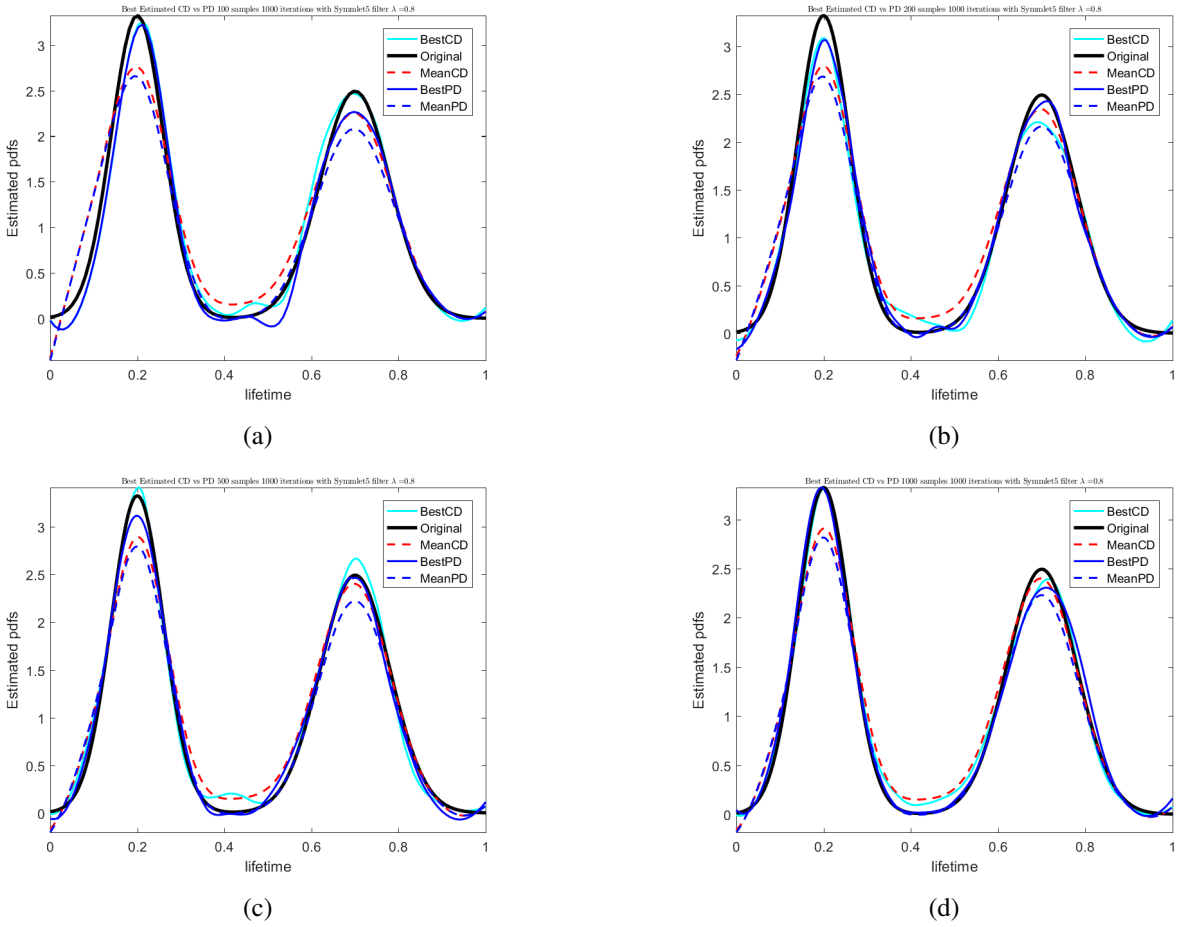


Figure 2.4: Estimate results for Strata distribution, $N = 100, 200, 500, 1000$ using Symmlet5.

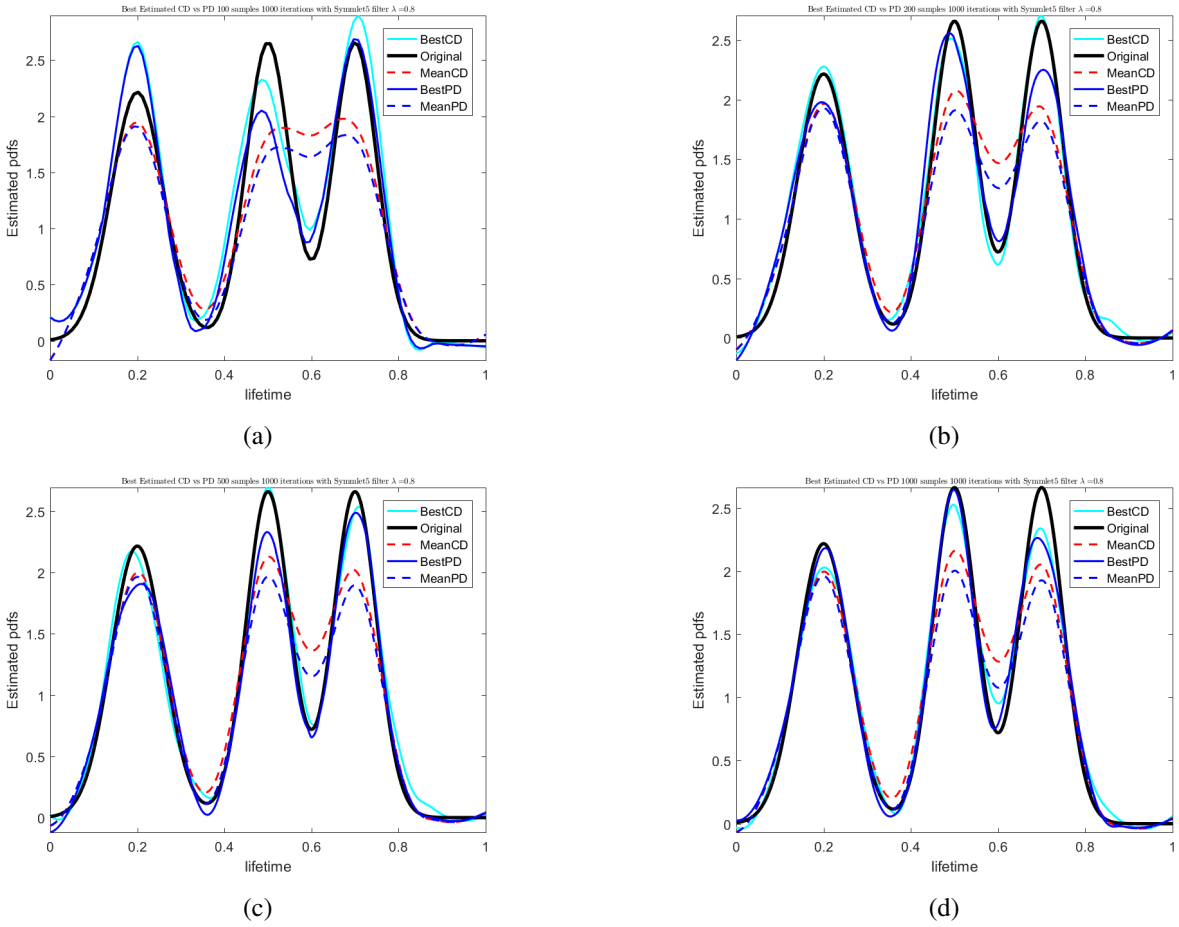


Figure 2.5: Estimate results for Multimodal distribution, $N = 100, 200, 500, 1000$ using Symmlet5.

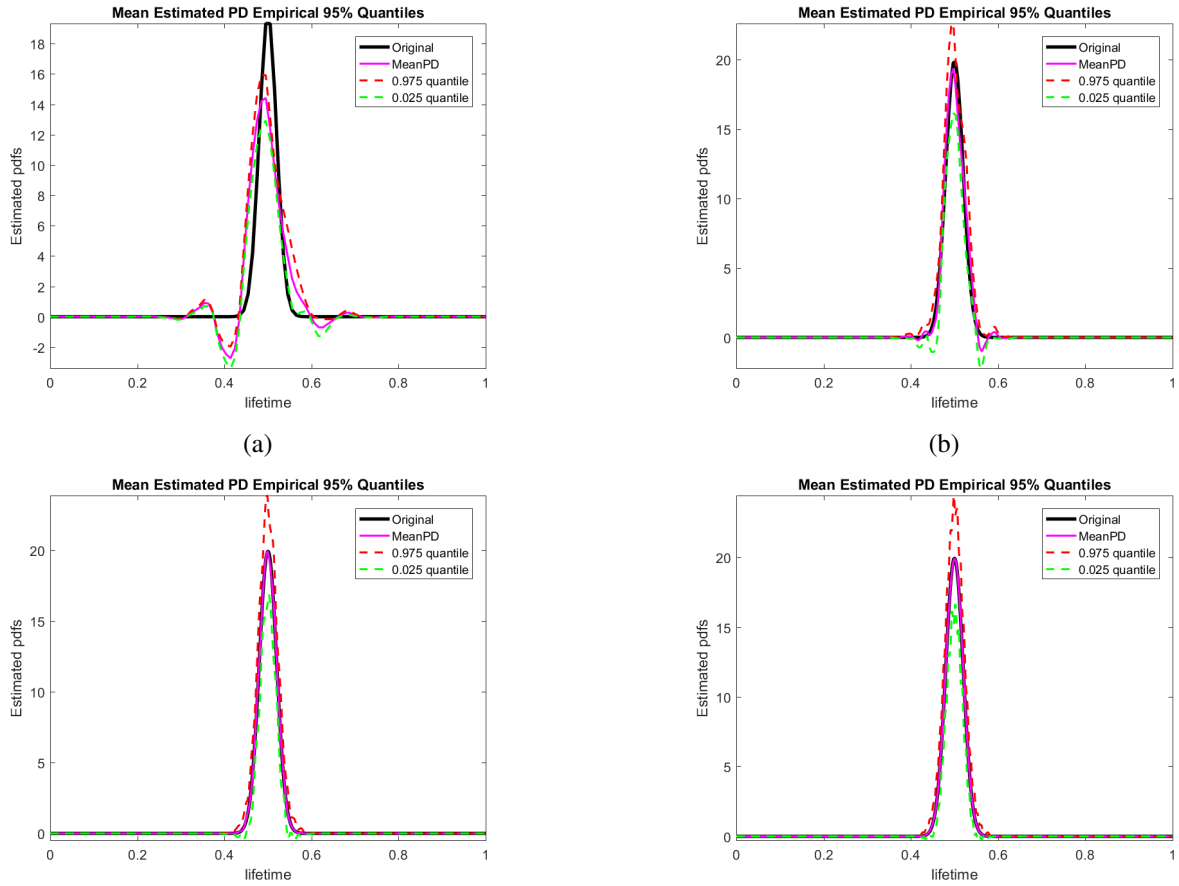
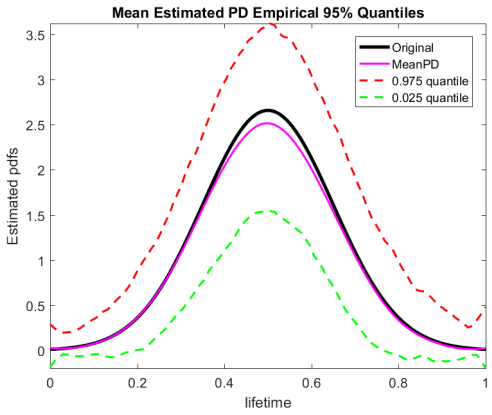
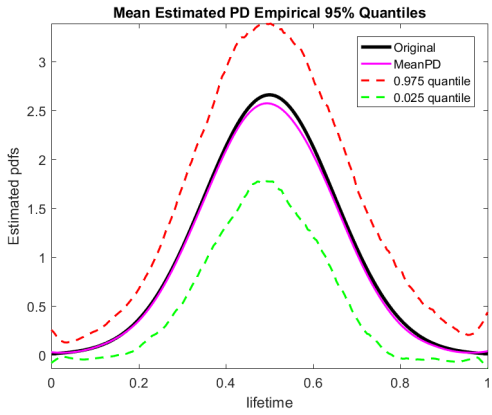


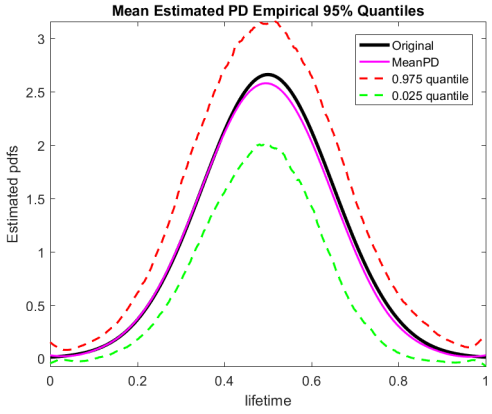
Figure 2.6: Results for 95% empirical quantiles and average estimate for Delta distribution using Symmlet5.(a)-(d) correspond to the partial data approach (for $N = 100, 200, 500, 1000$, respectively).



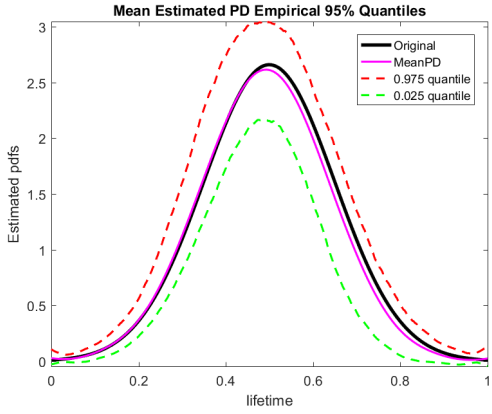
(a)



(b)



(c)



(d)

Figure 2.7: Results for 95% empirical quantiles and average estimate for Normal distribution using Symmlet5.(a)-(d) correspond to the partial data approach (for $N = 100, 200, 500, 1000$, respectively).

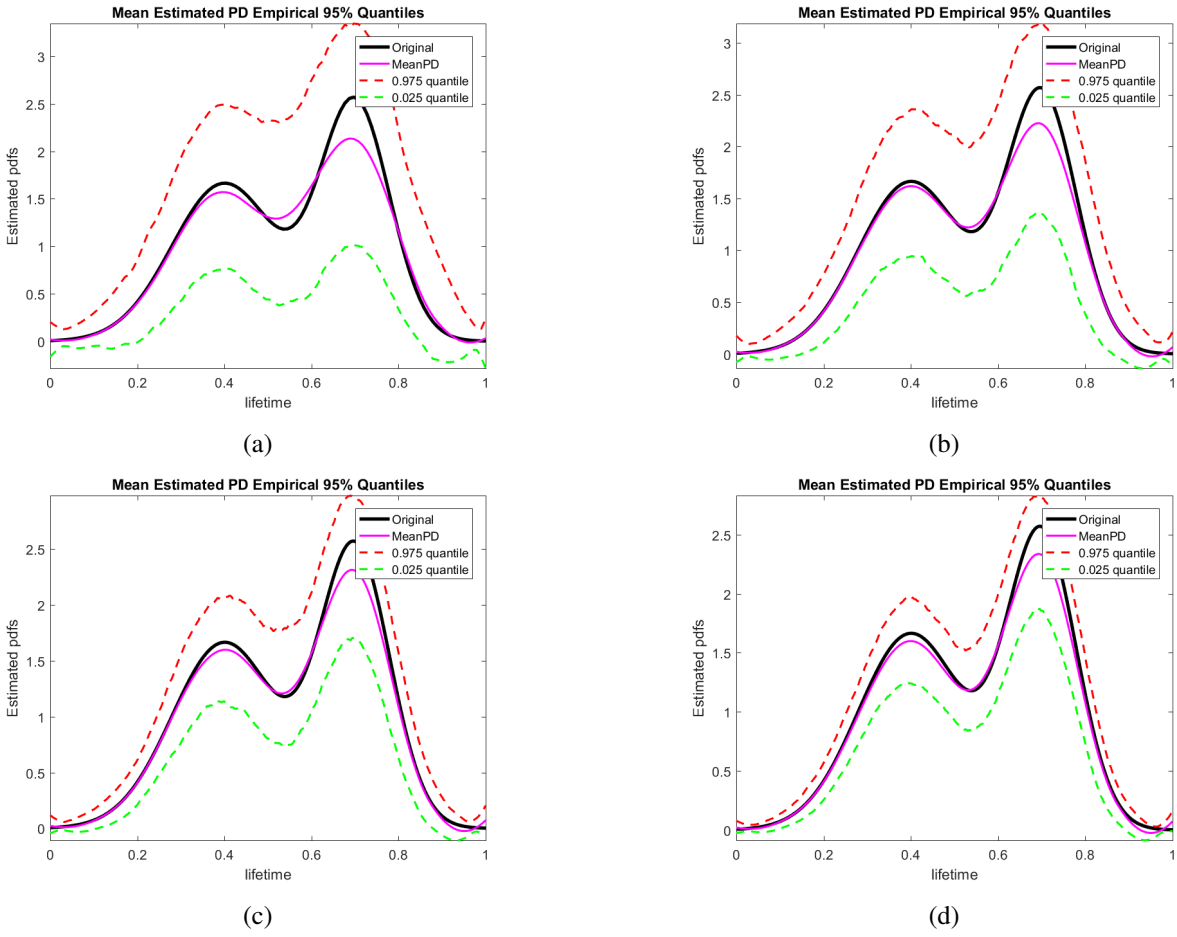


Figure 2.8: Results for 95% empirical quantiles and average estimate for Bimodal distribution using Symmlet5.(a)-(d) correspond to the partial data approach (for $N = 100, 200, 500, 1000$, respectively).

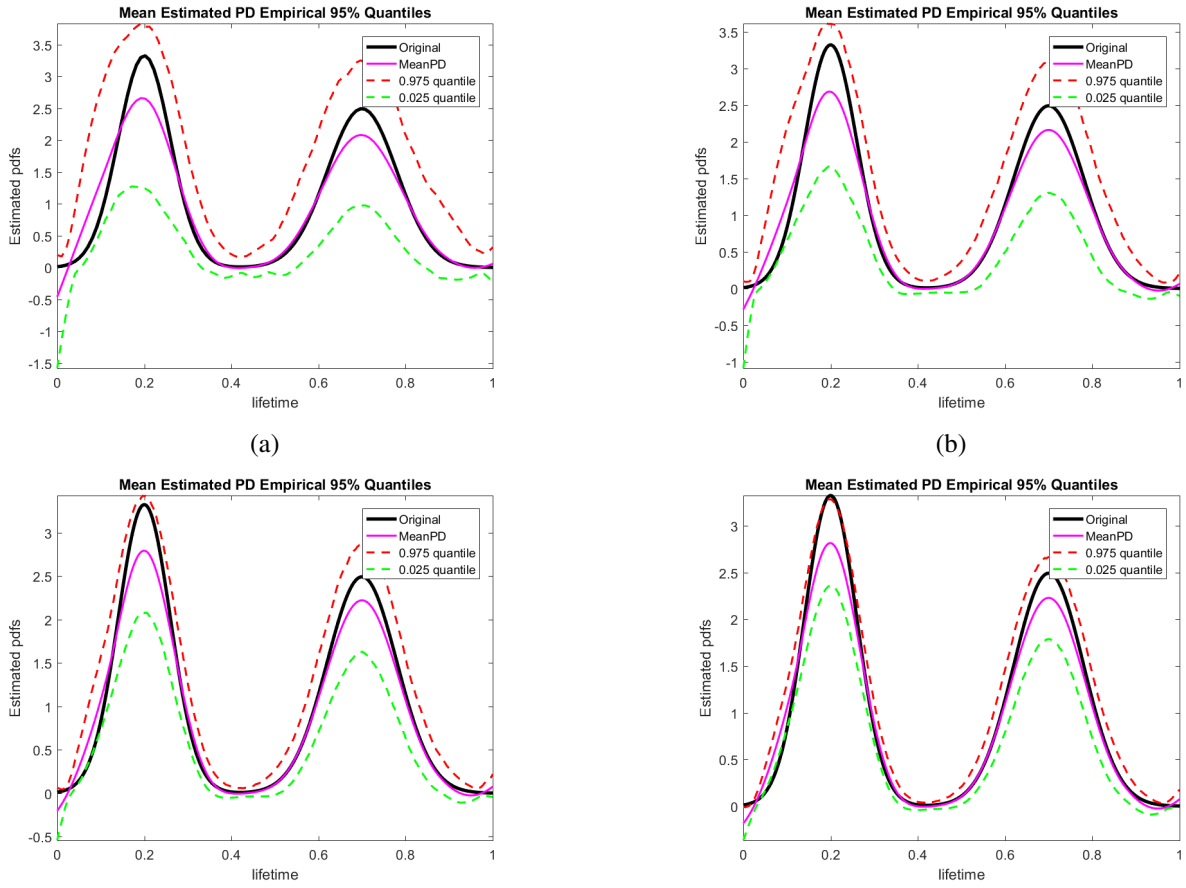
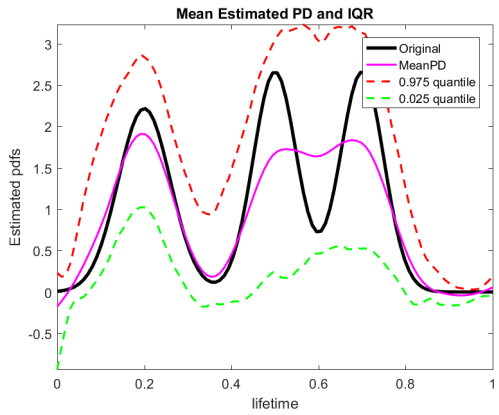
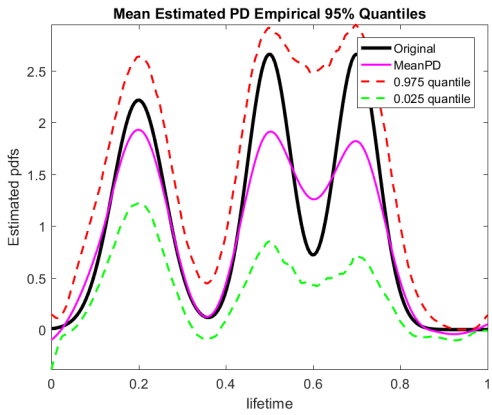


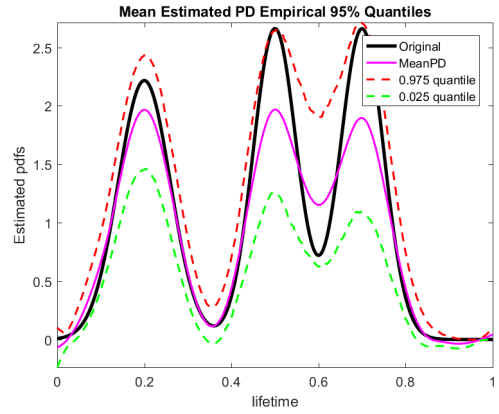
Figure 2.9: Results for 95% empirical quantiles and average estimate for Strata distribution using Symmlet5.(a)-(d) correspond to the partial data approach (for $N = 100, 200, 500, 1000$, respectively).



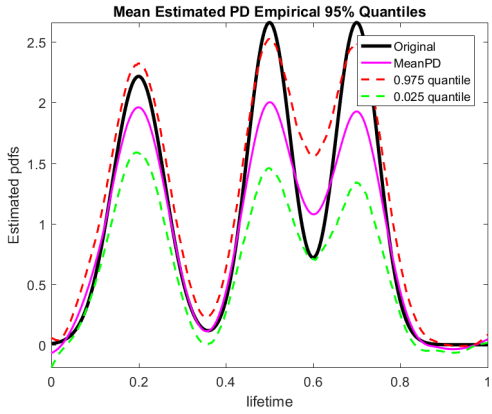
(a)



(b)

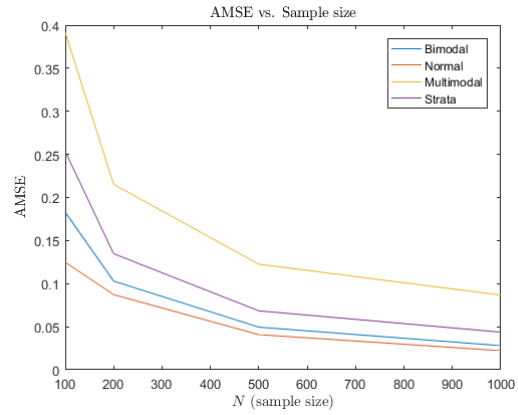


(c)

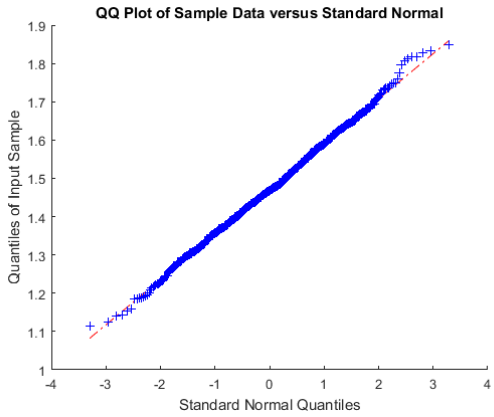


(d)

Figure 2.10: Results for 95% empirical quantiles and average estimate for Multimodal distribution using Symmlet5.(a)-(d) correspond to the partial data approach (for $N = 100, 200, 500, 1000$, respectively).



(a)



(b)

Figure 2.11: (a) AMSE for baseline distributions. (b) Q-Q Plot for the density estimates for Bimodal Distribution, $N = 1000$, $x = 0.7$.

2.4 Real Data application and comparison with other Estimators.

In this section we consider the implementation of the proposed estimator on the datasets utilized by Antoniadis et al. in [20]. To compare our approach with other popular estimators, we will also use the non-parametric Kernel density estimator with optimal bandwidth and the smoothed histogram using local polynomials based on the actual samples.

The first application considers the data studied by Haupt and Mansmann (1995)[4]. In their research, they analized the survival times for patients with liver metastases from a colorectal tumour without other distant metastases. In their data, they have a total of 622 patients from which 43.64% of the samples are censored. The obtained results are given in Fig.2.12 (a).

Our next practical application, considers the study of marriage dissolution based on a longitudinal survey conducted in the U.S.[5] The unit of observation is the couple and the event of interest is the time from marriage to divorce. Interviewed and widowhood are considered as censoring events. Couples with different educational levels and ethnicity were considered. The original data considered 3371 couples with 30.61% of samples being censored. The obtained results are given in Fig.2.12 (b).

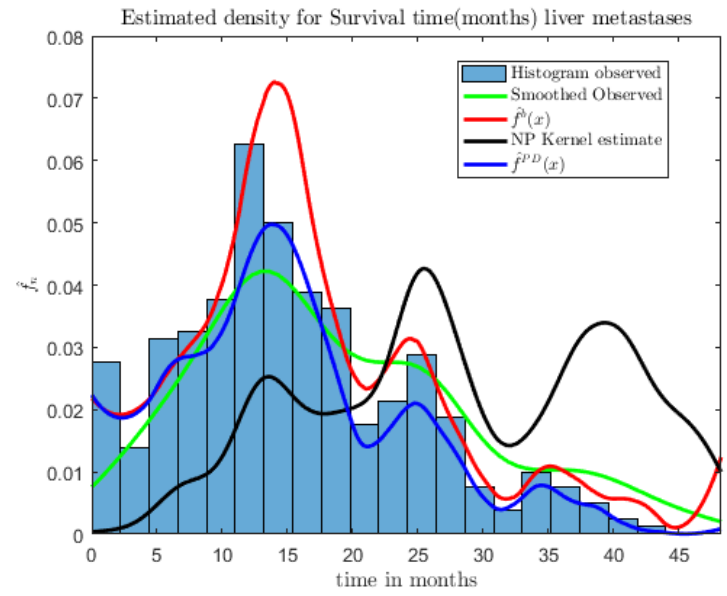
From figure 2.12 (a), it can be observed that the complete data estimator (in red) shows boundary effects, since after 45 months, according to the data there are almost no patients alive. However, both complete data and partial data estimators are able to catch the individual modes shown by the histogram without over smoothing as compared to the smoothed histogram (in green). Also, the estimators are able to keep the proportions between the histogram modes as compared to the Kernel density estimator with universal bandwidth (in

[4]The data set is available at CART for Survival Data. Statlib Archive `http://lib.stat.cmu.edu/S/survcart`.

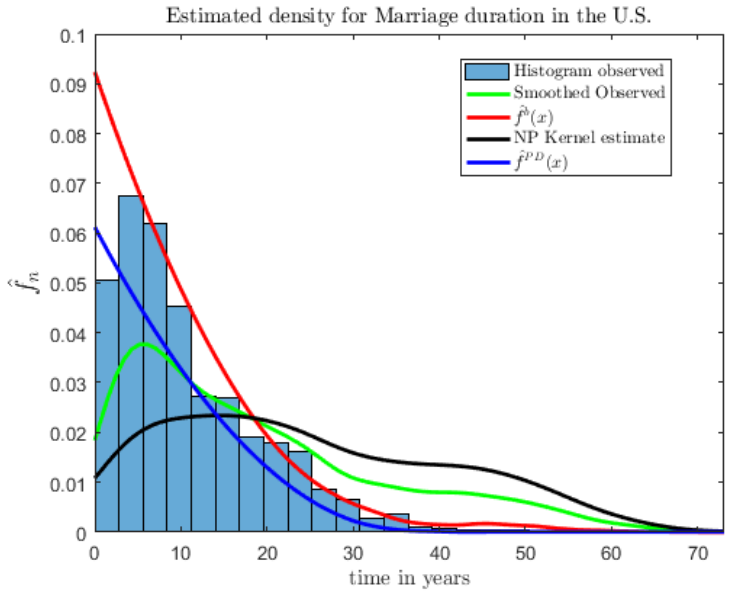
[5]Data set available at `http://data.princeton.edu/wws509/datasets` and was adapted from an example in the software aML (See Lillard and Panis (2000), aML Multilevel Multiprocess Statistical Software, Release 1.0, EconWare, LA, California.)

black).

From figure 2.12 (b), it is possible to observe the fairly exponential behavior of the density estimates. Both the complete data and the partial data are able to follow the rate of decay of the Histogram envelope and do not overestimate the density values in the right tails, which is consistent with the data (from data, it is highly unlikely that a certain couple would last married longer than 45 years); both local polynomial and kernel density estimator fail to account for that fact, while assigning significant density to times above 40 years.



(a)



(b)

Figure 2.12: Results for the application of the data driven estimators in real datasets. (a) corresponds to Liver metastases data and (b) to marriage duration in the U.S.

2.5 Conclusions and Discussion.

This Chapter introduced an empirical wavelet-based method to estimate the density in the case of randomly censored data. We proposed estimators based on use of the partial and complete sample, showing statistical properties of bias, consistency and limiting distribution. Also, we derived convergence rates for the expected $\mathbb{L}_2$ error, proposing the optimal choice for the multiresolution index J that is used for the approximation space.

Both estimators were implemented and tested using different baseline distributions via a theoretical simulation study, showing good performance in the presence of significantly censored data. The obtained results show that in theory, the estimator attains the large sample behavior that was proposed: it is asymptotically unbiased and mean-square consistent, which provides certain performance guarantees that makes the estimator suitable for practical applications.

Regarding the effect of censoring in the estimates, we observed that the introduced method is robust enough to handle censoring proportions of nearly 50% while achieving acceptable estimation results. Moreover, in the case of no censoring, the estimates converge to the usual orthogonal wavelet-series estimator (See remarks in section 2.2.4).

From a real data application viewpoint, the proposed method was capable to detect modes in the data that other smoothing methods typically fail to account, avoiding the problem of modes over-smoothing. Also, the estimators were capable of capturing the exponential rates of decay of the underlying densities, preventing the overestimation of likelihood values in regions of the support with near-zero empirical mass.

Finally, some of the drawbacks that were observed throughout this Chapter were the possibility of obtaining negative values for the density estimates (highly likely at the tails) and also boundary problems resulting from the periodic wavelet extension approach. Also, another

important remark worth noting is the fact that it is possible that the estimated density does no integrate to 1. Nonetheless, for most of these problems there are possible solutions such as the ones proposed in [18] and [29].

CHAPTER 3

EMPIRICAL WAVELET-BASED ESTIMATION FOR NON-LINEAR ADDITIVE REGRESSION MODELS.

Additive regression models are actively researched in the statistical field because of their usefulness in the analysis of responses determined by non-linear relationships with multivariate predictors. In this kind of statistical models, the response depends linearly on unknown functions of predictor variables and typically, the goal of the analysis is to make inference about these functions.

In this Chapter, we consider the problem of Additive Regression with random designs from a novel viewpoint: we propose an estimator based on an orthogonal projection onto a multiresolution space using empirical wavelet coefficients that are fully data driven. In this setting, we derive a mean-square consistent estimator based on periodic wavelets on the interval $[0, 1]$. For construction of the estimator, we assume that the joint distribution of predictors is non-zero and bounded on its support; We also assume that the functions belong to a Sobolev space and integrate to zero over the [0,1] interval, which guarantees model identifiability and convergence of the proposed method. Moreover, we provide the $\mathbb{L}_2$ risk analysis of the estimator and derive its convergence rate.

Theoretically, we show that this approach achieves good convergence rates when the dimensionality of the problem is relatively low and the set of unknown functions is sufficiently smooth. In this approach, the results are obtained without the assumption of an equispaced design, a condition that is typically assumed in most wavelet-based procedures.

Finally, we show practical results obtained from simulated data, demonstrating the poten-

tial applicability of our method in the problem of additive regression models with random designs.

3.1 Introduction

Additive regression models are popular in the statistical field because of their usefulness in the analysis of responses determined by non-linear relationships involving multivariate predictors. In this kind of statistical models, the response depends linearly on unknown functions of the predictors and typically, the goal of the analysis is to make inferences about these functions. This model has been extensively studied through the application of piece-wise polynomial approximations, splines, marginal integration, as well as back-fitting or functional principal components. Chapter 15 of [39], Chapter 22 of [9] and [40], [41] and [42] feature thorough discussions of the issues related to fitting such models and provide a comprehensive overview and analysis of various estimation techniques for this problem.

In general, the additive regression model relates a univariate response Y to predictor variables $\mathbf{X} \in \mathbb{R}^p$, $p \geq 1$, via a set of unknown non-linear functions $\{f_l \,|\, f_l : \mathbb{R} \to R\,,\, l = 1, ..., p\}$. The functions f_l may be assumed to have a specified parametric form (e.g. polynomial) or may be specified non-parametrically, simply as "smooth functions" that satisfy a set of constraints (e.g. belong to a certain functional space such as a Besov or Sobolev, Lipschitz continuity, spaces of functions with bounded derivatives, etc.). Though the parametric estimates may seem more attractive from the modeling perspective, they can have a major drawback: a parametric model automatically restricts the space of functions that is used to approximate the unknown regression function, regardless of the available data. As a result, when the elicited parametric family is not "close" to the assumed functional form the results obtained through the parametric approach can be misleading. For this reason, the non-parametric approach has gained more popularity in statistical research, providing a more general, flexible and robust

approach in tasks of functional inference.

In this Chapter we propose a linear functional estimator based on an orthogonal projection onto a specified multiresolution space V_J using empirical wavelet coefficients that are fully data driven. Here, V_J stands for the space spanned by the set of scaling functions of the form $\left\{\phi_{Jk}^{per},\, 0 \leq k \leq 2^J - 1\right\}$, generated by a specified wavelet filter. Since we assume predictors $\mathbf{X} \in \mathbb{R}^p\,,\, p \geq 1$ are random with an unknown distribution, we introduce a kernel density estimator in the model to estimate its density. In this setting, we propose a mean-square consistent estimator for the constant term and the wavelet coefficients in the orthogonal series representation of the model. Our results are based on wavelets periodic on the interval $[0, 1]$ and are derived under a set of assumptions that guarantee identifiability and convergence of the proposed estimator. Moreover, we derive convergence rates for the $\mathbb{L}_2$ risk and propose a practical choice for the multiresolution index J to be used in the wavelet expansion. In this approach, we obtain stated results without the assumption of an equispaced design, a condition that is typically assumed in most wavelet-based procedures.

Our choice of wavelets as an orthonormal basis is motivated by the fact that wavelets are well localized in both time and scale (frequency), and possess superb approximation properties for signals with rapid local changes such as discontinuities, cusps, sharp spikes, etc.. Moreover, the representation of these signals in the form of wavelet decompositions can be accurately done using only a few wavelet coefficients, enabling sparsity and dimensionality reduction. This adaptivity does not, in general, hold for other standard orthonormal bases (e.g. Fourier basis) which may require many compensating coefficients to describe signal discontinuities or local bursts.

We also illustrate practical results for the proposed estimator using different exemplary functions and random designs, under different sample sizes, demonstrating the suitability of the proposed methodology.

As it was mentioned, additive regression models have been studied by many authors using a wide variety of approaches. The approaches include marginal integration, back-fitting, least squares (including penalized least squares), orthogonal series approximations, and local polynomials. Short descriptions of the most commonly used techniques are provided next:

(i) **Marginal Integration**. This method was proposed by Tjostheim and Auestad (1994)[43] and Linton and Nielsen (1995)[44] and later generalized by Chen et al. (1996)[45]. The marginal integration idea is based on the estimation of the effects of each function in the model using sample averages of kernel functions by keeping a variable of interest fixed at each observed sample point, while changing the remaining ones. This method has been shown to produce good results in simulation studies (Sperlich et al., 1999)[46]. However, the marginal integration performance over finite samples tends to be inadequate when the dimension of the predictors is large. In particular, the bias-variance trade-off of the estimator in this case is challenging: for a given bandwidth there may be too few data points $\mathbf{x}_i$ for any given $\mathbf{x}$, which inflates the estimator variance and reduces its numerical stability. On the other hand, choosing larger bandwidth may reduce the variability but also enlarge the bias.

(ii) **Back-fitting**. This approach was first introduced by Buja et al. (1989)[47] and further developed by Hastie and Tibshirani (1990)[48]. This technique uses nonparametric regression to estimate each additive component, and then updates the preliminary estimates. This process continues in an iterative fashion until convergence. One of the drawbacks of this method is that it has been proven to be theoretically challenging to analize. In this context, Opsomer and Ruppert (1997)[49] investigated the properties of a version of back-fitting, and found that the estimator was not oracle efficient[1]. Later on, Mammen et al. (1999)[50] and Mammen and Park (2006)[51] proposed ways to

[1] An oracle efficient estimator is such that each component of the model can be estimated with the same convergence rate as if the rest of the model components were known.

modify the backfitting approach to produce estimators with better statistical properties such as oracle efficiency and asymptotic normality, and also free of the curse of dimensionality. Even though this is a popular method, it has been shown that its efficiency decreases when the unknown functions are observed at nonequispaced locations.

(iii) **Series based methods using wavelets**. One important benefit of wavelets is that they are able to adapt to unknown smoothness of functions (Donoho et al. (1995)[27]). Most of the work using wavelets is based on the requirement of equally spaced measurements (e.g. at equal time intervals or a certain response observed on a regularly spaced grid). Antoniadis et al. (1997)[4] propose a method using interpolations and averaging; based on the observed sample, the function is approximated at equally spaced dyadic points. In this context, most of the methods that use this kind of approach lead to wavelet coefficients that can be computed via a matrix transformation of the original data and are formulated in terms of a continuous wavelet transformation applied to a constant piecewise interpolation of the observed samples. Pensky and Vidakovic (2001)[52] propose a method that uses a probabilistic model on the design of the independent variables and can be applied to non-equally spaced designs (NESD). Their approach is based on a linear wavelet-based estimator that is similar to the wavelet modification of the Nadaraja-Watson estimator (Antoniadis et al. (1994)[53]). In the same context, Amato and Antoniadis (2001)[54] propose a wavelet series estimator based on tensor wavelet series and a regularization rule that guarantees an adaptive solution to the estimation problem in the presence of NESD.

(iv) **Other methods based on wavelets**. Different approaches from the previously described that are wavelet-based have been also investigated. Donoho et al. (1992)[55] proposed an estimator that is the solution of a penalized Least squares optimization problem preventing the problem of ill-conditioned design matrices. Zhang and Wong (2003)[56] proposed a two-stage wavelet thresholding procedure using local polyno-

mial fitting and marginal integration for the estimation of the additive components. Their method is adaptive to different degrees of smoothness of the components and has good asymptotic properties. Later on Sardy and Tseng (2004)[1] proposed a non-linear smoother and non-linear back-fitting algorithm that is based on `WaveShrink`, modeling each function in the model as a parsimonious expansion on a wavelet basis that is further subjected to variable selection (i.e. which wavelets to use in the expansion) via non-linear shrinkage.

As was discussed before in the context of the application of wavelets to the problem of additive models in NESD, another possibility is just simply ignore the nonequispaced condition on the predictors and apply the wavelet methods directly to the observed sample. Even though this might seem a somewhat crude approach, we will show that it is possible to implement this procedure via a relatively simple algorithm, obtaining good statistical properties and estimation results.

3.2 Wavelet-based Estimation in Additive Regression Models

Suppose that instead of the typical linear regression model $y = \sum_{j=1}^{p} \beta_j x_j + \beta_0 + \epsilon$ which assumes linearity in the predictors $\mathbf{x} = (x_1, ..., x_p)$, we have the following:

$$\begin{aligned} f(\mathbf{x}) &= \beta_0 + f_A(\mathbf{x}) + \sigma \cdot \epsilon \\ &= \beta_0 + \sum_{j=1}^{p} f_j(x_j) + \sigma \cdot \epsilon \end{aligned} \tag{3.1}$$

where ϵ, independent of $\mathbf{x}$, $\mathbb{E}[\epsilon] = 0$, $\mathbb{E}[\epsilon^2] = 1$, $\sigma > 0$, $\sigma < \infty$. Similarly, $\mathbf{x}_i \overset{\text{iid}}{\sim} h(\mathbf{x})$, an unknown design density of observations and $\{f_1(\cdot), ..., f_p(\cdot)\}$ are unknown real-valued functions $f_l : \mathbb{R} \to \mathbb{R}$ to be estimated.

3.2.1 Problem statement and derivation of the Estimator

Suppose that we are able to observe a sample $\{y_i = f(\mathbf{x}_i), \mathbf{x}_i\}_{i=1}^n$ where $\mathbf{x}_1, ..., \mathbf{x}_n \overset{iid}{\sim} h(\mathbf{x})$. We are interested in estimating β_0 and $\{f_1(\cdot), ..., f_p(\cdot)\}$. For simplicity (without loss of generality) and identifiability, we assume:

(A1) The density $h(\mathbf{x})$ is of the continuous type and has support in $[0,1]^p$. Also, we assume $\exists \epsilon_h > 0$ s.t. $h(\mathbf{x}) \geq \epsilon_h \ \forall \mathbf{x} \in [0,1]^p$.

(A2) For $k = 1, ..., p$, $\int_0^1 f_k(x)dx_k = 0$.

(A3) For $k = 1, ..., p$, $\underset{x\in[0,1]}{sup} |f_k(x)| \leq M_k < \infty$ and $\underset{x\in[0,1]}{inf} \{f_k(x)\} \geq m_k > -\infty$. This implies that for $k = 1, ..., p$, $f_k \in \mathbb{L}_2([0,1])$.

(A4) The design density $h()$ belongs to a generalized Holder class of functions of the form:

$$\mathbb{H}(\beta, L) = \{h : |\partial^{\boldsymbol{\alpha}} h(\mathbf{x}) - \partial^{\boldsymbol{\alpha}} h(\mathbf{y})| \leq L \parallel \mathbf{x} - \mathbf{y} \parallel_1^{\beta - |\boldsymbol{\alpha}|}, \forall \boldsymbol{\alpha} \in \mathbb{N}^p, \ s.t. \ |\boldsymbol{\alpha}| = \lfloor \beta \rfloor, \ \forall \mathbf{x}, \mathbf{y} \in [0,1]^p\} \tag{3.2}$$

where $\partial^{\boldsymbol{\alpha}} f := \partial_1^{\alpha_1} \cdot ... \cdot \partial_p^{\alpha_p} f = \frac{\partial^{|\boldsymbol{\alpha}|} f}{\partial x_1^{\alpha_1} \cdot ... \cdot \partial x_p^{\alpha_p}}$, and $|\boldsymbol{\alpha}| := \sum_{j=1}^p \alpha_j$. Also, suppose that $|\partial^{\boldsymbol{\alpha}} h| \leq M_h$, for all $\mathbf{x} \in [0,1]^p$ and $|\boldsymbol{\alpha}| \leq \lfloor \beta \rfloor$.

(A5) The density $h(\mathbf{x})$ is uniformly bounded in $[0,1]^p$, that is, $\forall \mathbf{x} \in [0,1]^p$, $|h(\mathbf{x})| \leq M$, $M < \infty$.

Furthermore, since $\left\{\phi_{J,k}^{per}(x), 0 \leq k \leq 2^J\right\}$ as $J \to \infty$ spans $\mathbb{L}_2([0,1])$, each of the functions in 4.1 can be represented as:

$$f_l(x) = \lim_{j\to\infty} \sum_{k=0}^{2^j - 1} c_{jk}^{(l)} \cdot \phi_{jk}^{per}(x), \quad l = 1, ..., p\,, \tag{3.3}$$

where $c_{jk}^{(l)}$ denotes the $j,k-$th wavelet coefficient of the $l-$th function in the model. Similarly,

for some fixed J that $f_{l,J}(x)$, $l = 1, ..., p$ is the orthogonal projection of $f_l(x)$, onto the multiresolution space V_J. Therefore, $f_{l,J}(x)$ can be expressed as:

$$f_{l,J}(x) = \sum_{k=0}^{2^J-1} c_{Jk}^{(l)} \cdot \phi_{Jk}^{per}(x), \quad l = 1, ..., p\,, \tag{3.4}$$

where:

$$c_{Jk}^{(l)} = \langle f_l(x), \phi_{Jk}^{per}(x) \rangle = \int_0^1 f_l(x)\phi_{Jk}^{per}(x)dx, \quad l = 1, ..., p\,. \tag{3.5}$$

Based on the model (4.1) and (4.3), it is possible to approximate $f(\mathbf{x})$ by an orthogonal projection $f_J(\mathbf{x})$ onto the multiresolution space spanned by the set of scaling functions:

$$\left\{\phi_{J,k}^{per}(x), 0 \leq k \leq 2^J - 1\right\},$$

by approximating each of the functions $f_l()$ as described above. Therefore, $f_J(\mathbf{x})$ can be expressed as:

$$f_J(\mathbf{x}) = \beta_0 + \sum_{l=1}^{p} \sum_{k=0}^{2^J-1} c_{Jk}^{(l)} \phi_{Jk}^{per}(x) \tag{3.6}$$

Now, the goal is for a pre-specified multiresolution index J, to use the observed samples to estimate the unknown constant β_0 and the orthogonal projections of the functions $f_{l,J}(x)$, $l = 1, ..., p$.

Remarks

(i) Note that the scaling function $\phi(x)$ for the wavelet basis $\left\{\phi_{J,k}^{per}(x), 0 \leq k \leq 2^J - 1]\right\}$ is absolutely integrable in $\mathbb{R}$. Therefore, $\int_{\mathbb{R}} |\phi(x)|dx = C_\phi < \infty$.

(ii) Also, from the above conditions, the variance of the response $y(\mathbf{x})$ is bounded for every $\mathbf{x} \in \mathbb{R}^p$.

(iii) The assumption that the support of the random vector $\mathbf{X}$ is $[0,1]^p$ can be always satisfied by carrying out appropriate monotone increasing transformations of each dimensional component, even in the case when the support before transformation is unbounded. In practice, it would be sufficient to transform the empirical support to $[0,1]^p$.

Derivation of the estimator for β_0

From the model definition presented in (4.1), and assumption (**A2**) we have that:

$$\begin{aligned} \int_{[0,1]^p} (\beta_0 + \sum_{l=1}^{p} f_l(x_l)) d\mathbf{x} &= \beta_0 + \sum_{l=1}^{p} \int_0^1 f_l(x_l) dx_l \\ &= \beta_0 \end{aligned} \tag{3.7}$$

Therefore, under assumptions (**A1**) and the last result, it is possible to obtain β_0 as:

$$\beta_0 = \mathbb{E}_{\mathbf{X},\epsilon}\left[\frac{f(\mathbf{X})}{h(\mathbf{X})}\right] . \tag{3.8}$$

Indeed,

$$\begin{aligned} \mathbb{E}_{\mathbf{X},\epsilon}\left[\frac{f(\mathbf{X})}{h(\mathbf{X})}\right] &= \mathbb{E}_{\mathbf{X},\epsilon}\left[\frac{\beta_0 + \sum_{j=1}^{p} f_j(X_j) + \sigma \cdot \epsilon}{h(\mathbf{X})}\right] \\ &= \beta_0 + \mathbb{E}_{\mathbf{X}}\left[\frac{\sum_{j=1}^{p} f_j(X_j)}{h(\mathbf{X})}\right] \\ &= \beta_0 . \end{aligned}$$

As a result of (3.8), a natural data-driven estimator of β_0 is

$$\hat{\beta}_0 = \frac{1}{n}\sum_{i=1}^{n} \frac{y_i}{\hat{h}_n(\mathbf{x}_i)}\,, \tag{3.9}$$

where $\hat{h}_n(\cdot)$ is a suitable non-parametric density estimator of $h()$, e.g. a kernel density estimator.

Derivation of the estimator for the wavelet coefficients $c_{Jk}^{(l)}$

Based on the multiresolution space spanned by the orthonormal functions $\left\{\phi_{J,k}^{per}(x)\right\}$, (4.4) and assumption (**A2**), the wavelet coefficients for each functional can be represented as:

$$c_{Jk}^{(l)} = \int_0^1 f_l(x_l)\phi_{Jk}^{per}(x_l)dx_l\,. \tag{3.10}$$

Expanding the right-hand-side (rhs) of the last equation, we get:

$$\begin{aligned}
\int_0^1 f_l(x_l)\phi_{Jk}^{per}(x_l)dx_l &= \int_0^1 f_l(x_l)\left(\phi_{Jk}^{per}(x_l) - 2^{-\frac{J}{2}}\right)dx_l \\
&= \int_0^1 \left(\beta_0 + \sum_{j=1}^{p} f_j(x_j)\right)\left(\phi_{Jk}^{per}(x_l) - 2^{-\frac{J}{2}}\right)dx_l \\
&= \int_{[0,1]^{p-1}}\int_0^1 \left(\beta_0 + \sum_{j=1}^{p} f_j(x_j)\right)\left(\phi_{Jk}^{per}(x_l) - 2^{-\frac{J}{2}}\right)dx_l d\mathbf{x}_{(-l)}\,,
\end{aligned}$$

where $\mathbf{x}_{(-l)}$ corresponds to the random vector $\mathbf{x}$ without the $l-$th entry. It is easy to see that (3.10) holds because of assumption (**A2**) and the fact that $\int_0^1 \phi_{Jk}^{per}(x)dx = 2^{-\frac{J}{2}}$. The proof for this last claim can be found in B.1.

Now, if we consider (**A1**), we can see that an alternative way to express (3.10) could be:

$$c_{Jk}^{(l)} = \mathbb{E}_{\mathbf{X},\epsilon}\left[\frac{f(\mathbf{X})(\phi_{Jk}^{per}(x_l) - 2^{-\frac{J}{2}})}{h(\mathbf{X})}\right]. \tag{3.11}$$

Indeed,

$$\begin{aligned}
\mathbb{E}_{\mathbf{X},\epsilon}\left[\frac{f(\mathbf{X})(\phi_{Jk}^{per}(x_l) - 2^{-\frac{J}{2}})}{h(\mathbf{X})}\right] &= \mathbb{E}_{\mathbf{X},\epsilon}\left[\frac{(\beta_0 + \sum_{j=1}^{p} f_j(X_j) + \sigma\cdot\epsilon)(\phi_{Jk}^{per}(x_l) - 2^{-\frac{J}{2}})}{h(\mathbf{X})}\right] \\
&= \int_{[0,1]^p}\left(\beta_0 + \sum_{j=1}^{p} f_j(x_j)\right)\left(\phi_{Jk}^{per}(x_l) - 2^{-\frac{J}{2}}\right)d\mathbf{x} \\
&= c_{Jk}^{(l)}\,.
\end{aligned}$$

From (3.11), similarly as for β_0, we obtain a natural data-driven estimator of $c_{Jk}^{(l)}$ as:

$$\hat{.c}_{Jk}^{(l)} = \frac{1}{n}\sum_{i=1}^{n}\frac{y_i\left(\phi_{Jk}^{per}(x_{il}) - 2^{-\frac{J}{2}}\right)}{\hat{h}_n(\mathbf{x}_i)} \tag{3.12}$$

3.2.2 Asymptotic Properties of the Estimator

In this section, we study the asymptotic properties of the estimates proposed in (3.9) and (3.12) and propose necessary and sufficient conditions for the pointwise mean squared consistency of the estimator, under assumptions (**A1**)-(**A5**).

Unbiasedness and Consistency of $\hat{\beta}_0$

Next, we analyze the asymptotic behavior of the estimator $\hat{\beta}_0$ assuming assumptions (**Ak1**)-(**Ak4**) stated in B.3 hold.

Asymptotic Behavior of $\mathbb{E}(\hat{\beta}_0)$

From (B.10) and the hierarchy of convergence for random variables, it follows that for a fixed $\mathbf{x}$, $\hat{h}_n(\mathbf{x}) \xrightarrow{\mathbb{D}} h(\mathbf{x})$. Let's consider now a function $g : [\epsilon_h, M] \to [0, B_h]$, for $\epsilon_h > 0$, $B_h < \infty$, defined as $g(\hat{h}_n(\mathbf{x})) = \frac{1}{\hat{h}_n(\mathbf{x})}$. Since $\hat{h}_n(\mathbf{x})$ satisfies (**A5**)-(**A6**), $g(h)$ is bounded and continuous, which implies:

$$\mathbb{E}\left[\frac{1}{\hat{h}_n(\mathbf{x})}\right] \underset{n\to\infty}{\to} \frac{1}{h(\mathbf{x})}\,. \tag{3.13}$$

In fact, since $g(\hat{h}_n(\mathbf{x})) = \frac{1}{\hat{h}_n(\mathbf{x})}$ is continuous in $(0, \infty)$ and admits infinitely many derivatives , by using a Taylor series expansion around $h(\mathbf{x})$ and results (B.12) and (B.15), it is possible to obtain:

$$\begin{aligned}\left|\mathbb{E}_{\mathbf{X}_1,\ldots,\mathbf{X}_n}\left[\left(\frac{1}{\hat{h}_n(\mathbf{x})}\right)^k - \left(\frac{1}{h(\mathbf{x})}\right)^k\right]\right| &\leq \frac{1}{\epsilon_h^{k+2}}\left\{|Bias(\hat{h}_n(\mathbf{x}))| + Var(\hat{h}_n(\mathbf{x})) + Bias(\hat{h}_n(\mathbf{x}))^2\right\} \\ &\leq C\left\{\delta^\beta + \frac{1}{n\delta^p} + \delta^{2\beta}\right\}, \end{aligned} \tag{3.14}$$

for $k \geq 1$ and a sufficiently large $C > 0$ (independent of n, δ).

Therefore, under the choice $\delta \sim n^{-\frac{1}{2\beta+p}}$, $\mathbb{E}\left[\left(\frac{1}{\hat{h}_n(\mathbf{x})}\right)^k\right]$ converges to $\left(\frac{1}{h(\mathbf{x})}\right)^k$ at a rate $\sim n^{-\frac{\beta}{2\beta+p}}$ for $k \geq 1$. Here the expectation is taken with respect to the joint density of the iid sample.

Similarly, the last result leads to:

$$\mathbb{E}_{\mathbf{X}_1,\ldots,\mathbf{X}_n}\left[\left(\frac{1}{\hat{h}_n(\mathbf{x})} - \frac{1}{h(\mathbf{x})}\right)^2\right] \longrightarrow 0\,, \tag{3.15}$$

as $n \to \infty$ at a rate $\sim n^{-\frac{2\beta}{2\beta+p}}$.

Now, letting **x** to be random, using conditional expectation it is possible to obtain:

$$\mathbb{E}\left[\frac{1}{\hat{h}_n(\mathbf{X})}\right] = \mathbb{E}_{\mathbf{X}}\left[\mathbb{E}_{\mathbf{X}_1,\ldots,\mathbf{X}_n|\mathbf{X}}\left(\frac{1}{\hat{h}_n(\mathbf{x})}|\mathbf{X}\right)\right]. \tag{3.16}$$

From (3.13) and the last result, the dominated convergence theorem implies:

$$\mathbb{E}\left[\frac{1}{\hat{h}_n(\mathbf{x})}\right] \underset{n\to\infty}{\longrightarrow} 1 \tag{3.17}$$

Using the definition of $\hat{\beta}_0$ and the model (4.1), we obtain:

$$\begin{aligned}
\mathbb{E}\left[\hat{\beta}_0\right] &= \beta_0 + \mathbb{E}\left[\frac{\sum_{l=1}^{p} f_l(X_l)}{\hat{h}_n(\mathbf{X})}\right] \\
&= \beta_0 + \mathbb{E}_{\mathbf{X}}\left[\mathbb{E}_{\mathbf{X}_1,\ldots,\mathbf{X}_n|\mathbf{X}}\left(\frac{\sum_{l=1}^{p} f_l(X_l)}{\hat{h}_n(\mathbf{X})}|\mathbf{X}\right)\right] \\
&= \beta_0 + \mathbb{E}_{\mathbf{X}}\left[\sum_{l=1}^{p} f_l(X_l)\cdot\mathbb{E}_{\mathbf{X}_1,\ldots,\mathbf{X}_n|\mathbf{X}}\left(\frac{1}{\hat{h}_n(\mathbf{X})}|\mathbf{X}\right)\right].
\end{aligned} \tag{3.18}$$

Therefore, from (3.13)-(3.17) and under (**A2**),(**A3**), the dominated convergence leads to:

$$\mathbb{E}\left[\hat{\beta}_0\right] \underset{n\to\infty}{\longrightarrow} \beta_0\,, \tag{3.19}$$

which shows that $\hat{\beta}_0$ is asymptotically unbiased for β_0.

Asymptotic Behavior of Var($\hat{\beta}_0$)

From the definition of $\hat{\beta}_0$ and (4.1), we can see that:

$$
\begin{aligned}
Var(\hat{\beta}_0) &= \frac{1}{n} Var\left(\frac{Y}{\hat{h}_n(\mathbf{X})}\right) \\
&\leq \frac{1}{n}\mathbb{E}\left[\frac{Y^2}{\hat{h}_n(\mathbf{X})^2}\right] \\
&\leq \frac{1}{n}\mathbb{E}_{\mathbf{X}}\left[\mathbb{E}_{\mathbf{X}_1,\ldots,\mathbf{X}_n|\mathbf{X}=\mathbf{x}}\left(\frac{Y^2}{\hat{h}_n(\mathbf{X})^2}|\mathbf{X}=\mathbf{x}\right)\right] . \qquad (3.20)
\end{aligned}
$$

Now, if $n \to \infty$, from conditions (**A2**) and (**A3**), and the dominated convergence theorem, it follows:

$$
\mathbb{E}_{\mathbf{X}}\left[\mathbb{E}_{\mathbf{X}_1,\ldots,\mathbf{X}_n|\mathbf{X}=\mathbf{x}}\left(\frac{Y^2}{\hat{h}_n(\mathbf{X})^2}|\mathbf{X}=\mathbf{x}\right)\right] \underset{n\longrightarrow\infty}{\to} \mathbb{E}\left[\frac{Y^2}{h(\mathbf{X})^2}\right] . \qquad (3.21)
$$

Thus,

$$
Var(\hat{\beta}_0) \underset{n\to\infty}{\longrightarrow} 0\,, \qquad (3.22)
$$

provided $\mathbb{E}\left[\frac{Y^2}{h(\mathbf{X})^2}\right] < \infty$.

Finally, putting together (3.19) and (3.22) we obtain that $\hat{\beta}_0$ is consistent for β_0.

Unbiasedness and Consistency of the $\hat{c}_{Jk}^{(l)}$

In this section, we study the asymptotic behavior of the wavelet coefficient estimators $\hat{c}_{Jk}^{(l)}$ for a fixed J, assuming that conditions (**A1**)-(**A5**) and (**Ak1**)-(**Ak4**) hold.

Asymptotic Behavior of $\mathbb{E}(\hat{c}_{Jk}^{(l)})$

For a fixed J, $l = 1, ..., p$, and $k = 0, ..., 2^J - 1$,we have that $\hat{c}_{Jk}^{(l)} = \frac{1}{n}\sum_{i=1}^{n} \frac{y_i\left(\phi_{Jk}^{per}(x_{il}) - 2^{-\frac{J}{2}}\right)}{\hat{h}_n(\mathbf{x}_i)}$. Therefore,

$$\mathbb{E}\left[\hat{c}_{Jk}^{(l)}\right] = \mathbb{E}\left[\frac{Y\phi_{Jk}^{per}(X_l)}{\hat{h}_n(\mathbf{X})}\right] - 2^{-\frac{J}{2}}\mathbb{E}\left[\hat{\beta}_0\right] . \tag{3.23}$$

Following the same argument as in the case of the asymptotic behavior of $\hat{\beta}_0$, we find that the first term of (3.23) can be represented as:

$$\begin{aligned}
\mathbb{E}\left[\frac{Y\phi_{Jk}^{per}(X_l)}{\hat{h}_n(\mathbf{X})}\right] &= \mathbb{E}_{\mathbf{X}}\left[\mathbb{E}_{\mathbf{X}_1,...,\mathbf{X}_n|\mathbf{X}}\left(\frac{Y\phi_{Jk}^{per}(X_l)}{\hat{h}_n(\mathbf{X})}|\mathbf{X}\right)\right] \\
&= \mathbb{E}_{\mathbf{X}}\left[Y\phi_{Jk}^{per}(X_l)\cdot\mathbb{E}_{\mathbf{X}_1,...,\mathbf{X}_n|\mathbf{X}}\left(\frac{1}{\hat{h}_n(\mathbf{X})}|\mathbf{X}\right)\right] .
\end{aligned}$$

Since J is assumed fixed and (**A3**) holds, by the dominated convergence theorem, it follows that:

$$\mathbb{E}_{\mathbf{X}}\left[Y\phi_{Jk}^{per}(X_l)\cdot\mathbb{E}_{\mathbf{X}_1,...,\mathbf{X}_n|\mathbf{X}}\left(\frac{1}{\hat{h}_n(\mathbf{X})}|\mathbf{X}\right)\right] \underset{n\to\infty}{\longrightarrow} \mathbb{E}\left[\frac{Y\phi_{Jk}^{per}(X_l)}{h(\mathbf{X})}\right] . \tag{3.24}$$

Furthermore, by (**A3**) and (B.1):

$$\begin{aligned}
\mathbb{E}\left[\frac{Y\phi_{Jk}^{per}(X_l)}{h(\mathbf{X})}\right] &= \int_{[0,1]^p}\left(\beta_0 + \sum_{j=1}^{p} f_j(x_j)\right)\phi_{Jk}^{per}(x_l)d\mathbf{x} \\
&= \int_0^1 f_l(x_l)\phi_{Jk}^{per}(x_l)dx_l + 2^{-\frac{J}{2}}\beta_0 \\
&= c_{Jk}^{(l)} + 2^{-\frac{J}{2}}\beta_0 .
\end{aligned} \tag{3.25}$$

Finally, putting together the last result and (3.19), it follows:

$$\mathbb{E}\left[\hat{c}_{Jk}^{(l)}\right] \underset{n\to\infty}{\longrightarrow} c_{Jk}^{(l)}\,, \tag{3.26}$$

which shows that the wavelet coefficient estimators $\hat{c}_{Jk}^{(l)}$ are asymptotically unbiased, for J fixed, $l = 1, ..., p$, and $k = 0, ..., 2^J - 1$.

Asymptotic Behavior of Var($\hat{c}_{Jk}^{(l)}$)

For a fixed J, $l = 1, ..., p$ and $k = 0, ..., 2^J - 1$, $\hat{c}_{Jk}^{(l)} = \frac{1}{n}\sum_{i=1}^{n} \frac{Y_i\left(\phi_{Jk}^{per}(X_{il}) - 2^{-\frac{J}{2}}\right)}{\hat{h}_n(\mathbf{x}_i)}$, the variance of $\hat{c}_{Jk}^{(l)}$ is given by:

$$\begin{aligned} Var\left(\hat{c}_{Jk}^{(l)}\right) &= Var\left(\frac{1}{n}\sum_{i=1}^{n}\frac{Y_i\phi_{Jk}^{per}(X_{il})}{\hat{h}_n(\mathbf{X}_i)} - 2^{-\frac{J}{2}}\hat{\beta}_0\right) \\ &= \frac{1}{n}Var\left(\frac{Y\phi_{Jk}^{per}(X_l)}{\hat{h}_n(\mathbf{X})}\right) + 2^{-J}Var\left(\hat{\beta}_0\right) - 2Cov\left(\frac{1}{n}\sum_{i=1}^{n}\frac{Y_i\phi_{Jk}^{per}(X_{il})}{\hat{h}_n(\mathbf{X}_i)}\,,\, 2^{-\frac{J}{2}}\hat{\beta}_0\right) \\ &= \frac{1}{n}V_{c1} + 2^{-J}V_{c2} + 2V_{c3}\,. \end{aligned}$$

By using the model defined in (4.1) we find that for $V_{c1} = \frac{1}{n}Var\left(\frac{Y\phi_{Jk}^{per}(X_l)}{\hat{h}_n(\mathbf{X})}\right)$ it follows:

$$V_{c1} = Var_{\mathbf{X}}\left(\mathbb{E}_{\mathbf{X}_1,...,\mathbf{X}_n|\mathbf{X}}\left[\frac{Y\phi_{Jk}^{per}(X_l)}{\hat{h}_n(\mathbf{X})}|\mathbf{X}\right]\right) + \mathbb{E}_{\mathbf{X}}\left[Var_{\mathbf{X}_1,...,\mathbf{X}_n|\mathbf{X}}\left(\frac{Y\phi_{Jk}^{per}(X_l)}{\hat{h}_n(\mathbf{X})}|\mathbf{X}\right)\right],$$

$$= Var_{\mathbf{X}}\left(Y\phi_{Jk}^{per}(X_l)\cdot\mathbb{E}_{\mathbf{X}_1,...,\mathbf{X}_n|\mathbf{X}}\left[\tfrac{1}{\hat{h}_n(\mathbf{X})}|\mathbf{X}\right]\right) + \mathbb{E}_{\mathbf{X}}\left[(Y\phi_{Jk}^{per}(X_l))^2\cdot Var_{\mathbf{X}_1,...,\mathbf{X}_n|\mathbf{X}}\left(\tfrac{1}{\hat{h}_n(\mathbf{X})}|\mathbf{X}\right)\right]. \tag{3.27}$$

By the dominated convergence theorem, it follows:

$$Var\left(\frac{Y\phi_{Jk}^{per}(X_l)}{\hat{h}_n(\mathbf{X})}\right) \underset{n\to\infty}{\longrightarrow} Var\left(\frac{Y\phi_{Jk}^{per}(X_l)}{h(\mathbf{X})}\right) \tag{3.28}$$

where the last result holds since:

$$\mathbb{E}_{\mathbf{X}_1,\ldots,\mathbf{X}_n|\mathbf{X}=\mathbf{x}}\left[\frac{1}{\hat{h}_n(\mathbf{X})}|\mathbf{X}=\mathbf{x}\right] \underset{n\to\infty}{\longrightarrow} \frac{1}{h(\mathbf{x})}\,, \text{ and}$$
$$Var_{\mathbf{X}_1,\ldots,\mathbf{X}_n|\mathbf{X}=\mathbf{x}}\left(\frac{1}{\hat{h}_n(\mathbf{X})}|\mathbf{X}=\mathbf{x}\right) \underset{n\to\infty}{\longrightarrow} 0\,.$$

This implies,

$$\frac{1}{n}Var\left(\frac{Y\phi_{Jk}^{per}(X_l)}{\hat{h}_n(\mathbf{X})}\right) \underset{n\to\infty}{\longrightarrow} 0\,. \tag{3.29}$$

Lemma 3.2.1. *Let us suppose that conditions **(A1)**-**(A5)** and **(Ak1)**-**(Ak4)** hold. Then:*

$$\mathbb{E}\left[\left(\frac{Y\phi_{Jk}^{per}(X_l)}{h(\boldsymbol{X})}\right)^2\right] \leq C(\beta_0,p,\sigma^2,M_f)\cdot\left\{\frac{1}{\epsilon_h}\left(\lceil\log_2(\frac{1}{\epsilon_h})\rceil-1\right)+\frac{1}{\lceil\log_2(\frac{1}{\epsilon_h})\rceil}\right\}\,, \tag{3.30}$$

*where $C(\beta_0,p,\sigma^2,M_f)=(p\cdot M_f+|\beta_0|)^2+\sigma^2$. This result shows that $Var\left(\frac{Y\phi_{Jk}^{per}(X_l)}{h(\boldsymbol{X})}\right)$ is bounded from above, provided $p<\infty$, $\sigma^2<\infty$ and conditions **(A1)**-**(A5)** and **(Ak1)**-**(Ak4)** hold. Therefore,*

$$Var\left(\frac{Y\phi_{Jk}^{per}(X_l)}{\hat{h}_n(\boldsymbol{X})}\right) \underset{n\to\infty}{\to} Var\left(\frac{Y\phi_{Jk}^{per}(X_l)}{h(\boldsymbol{X})}\right) < \infty\,.$$

The proof can be found in Appendix B.4.

Similarly, as for V_{c1}, let's consider the behavior of $V_{c3} = Cov\left(\frac{1}{n}\sum_{i=1}^{n}\frac{Y_i\phi_{Jk}^{per}(X_{il})}{\hat{h}_n(\mathbf{X}_i)}\,,\,2^{-\frac{J}{2}}\hat{\beta}_0\right)$. Using the covariance definition and the iid assumption for the sample $\{y_i = f(\mathbf{x}_i), \mathbf{x}_i\}_{i=1}^n$, it follows that:

$$V_{c3} = \frac{2^{-\frac{J}{2}}}{n^2}\left\{\sum_{i=1}^{n} Cov\left(\frac{Y_i\phi_{Jk}^{per}(X_{il})}{\hat{h}_n(\mathbf{X}_i)}\,,\,\frac{Y_i}{\hat{h}_n(\mathbf{X}_i)}\right) + \sum_{\substack{i=1\\ i\neq j}}^{n}\sum_{j=1}^{n} Cov\left(\frac{Y_i\phi_{Jk}^{per}(X_{il})}{\hat{h}_n(\mathbf{X}_i)}\,,\,\frac{Y_j}{\hat{h}_n(\mathbf{X}_j)}\right)\right\}. \tag{3.31}$$

Lemma 3.2.2. *Let us suppose assumptions **(A1)**-**(A5)** and **(Ak1)**-**(Ak4)** are satisfied. The following results hold:*

$$Cov\left(\frac{1}{n}\sum_{i=1}^{n}\frac{Y_i\phi_{Jk}^{per}(X_{il})}{\hat{h}_n(\boldsymbol{X}_i)}\,,\,2^{-\frac{J}{2}}\hat{\beta}_0\right)\underset{n\to\infty}{\longrightarrow}0\,, \tag{3.32}$$

which further implies that for any fixed J, $l = 1, ..., p$, and $k = 0, ..., 2^J - 1$,

$$Cov\left(\hat{\beta}_0\,,\,\hat{c}_{Jk}^{(l)}\right)\underset{n\to\infty}{\longrightarrow}0\,. \tag{3.33}$$

The corresponding proofs can be found in Appendix B.5.

Putting together (3.22), (3.29) and (3.32) it follows that for a fixed J, $l = 1, ..., p$, and $k = 0, ..., 2^J - 1$:

$$Var\left(\hat{c}_{Jk}^{(l)}\right)\underset{n\to\infty}{\longrightarrow}0\,. \tag{3.34}$$

Finally, from (3.26) and (3.34) we get that for a fixed J, $l = 1, ..., p$, and $k = 0, ..., 2^J - 1$, $\hat{c}_{Jk}^{(l)}$ is consistent for $c_{Jk}^{(l)}$.

Unbiasedness and Consistency of $\hat{f}_J(\boldsymbol{x})$

From (4.5), we have that $f_J(\mathbf{x}) = \beta_0 + \sum_{l=1}^{p} \sum_{k=0}^{2^J-1} c_{Jk}^{(l)} \phi_{Jk}^{per}(x_l)$. If results (3.9) and (3.12) are substituted in the expression for $f_J(\mathbf{x})$, the data-driven estimator can be expressed as:

$$\hat{f}_J(\mathbf{x}) = \hat{\beta}_0 + \sum_{l=1}^{p} \sum_{k=0}^{2^J-1} \hat{c}_{Jk}^{(l)} \phi_{Jk}^{per}(x_l)\,.$$

Since both $\hat{\beta}_0$ and $\hat{c}_{Jk}^{(l)}$ are asymptotically unbiased, it follows:

$$\mathbb{E}\left[\hat{f}_J(\mathbf{x})\right] \underset{n\to\infty}{\to} f_J(\mathbf{x})\,, \text{ and} \tag{3.35}$$

$$Var\left(\hat{f}_J(\mathbf{x})\right) = Var\left(\hat{\beta}_0\right) + Var\left(\sum_{l=1}^{p} \sum_{k=0}^{2^J-1} \hat{c}_{Jk}^{(l)} \phi_{Jk}^{per}(x_l)\right) + 2Cov\left(\hat{\beta}_0\,,\, \sum_{l=1}^{p} \sum_{k=0}^{2^J-1} \hat{c}_{Jk}^{(l)} \phi_{Jk}^{per}(x_l)\right)\,. \tag{3.36}$$

In order to show that $Var\left(\hat{f}_J(\mathbf{x})\right) \underset{n\to\infty}{\to} 0$, we just need to prove that the second term of the expression (3.36) goes to zero as $n \to \infty$. This can be seen from (3.22) and (3.33).

Lemma 3.2.3. *For any $s \neq k$, $s, k = 0, ..., 2^J - 1$ and fixed J, under the stated assumptions:*

$$Cov\left(\hat{c}_{Jk}^{(l)}\,,\, \hat{c}_{Js}^{(l)}\right) \underset{n\to\infty}{\longrightarrow} 0\,. \tag{3.37}$$

The proof can be found in Appendix B.6.

From (3.37) it follows:

$$Var\left(\sum_{l=1}^{p} \sum_{k=0}^{2^J-1} \hat{c}_{Jk}^{(l)} \phi_{Jk}^{per}(x_l)\right) \underset{n\to\infty}{\longrightarrow} 0\,. \tag{3.38}$$

Finally, from (3.22), (3.33) and (3.38), it is clear that $Var\left(\hat{f}_J(\mathbf{x})\right) \underset{n\to\infty}{\to} 0$. This result together with (3.35) implies that:

$$\hat{f}_J(\mathbf{x}) \xrightarrow{\mathbb{P}} f_J(\mathbf{x})\,. \tag{3.39}$$

Therefore, the estimator $\hat{f}_J(\mathbf{x})$ is consistent for $f_J(\mathbf{x})$.

Remarks

(i) The results and derivations presented in lemmas 3.2.1-3.2.3, indicate that the estimator $\hat{f}_J(\mathbf{x})$ suffers from the course of dimensionality. In fact, the dependence from the dimension p of the random covariates $\mathbf{x}$ influence in both the convergence rate of the density estimator $\hat{h}_n(\mathbf{x})$ and the constant $C(\beta_0, p, \sigma^2, M_f)$.

(ii) As can be observed from this section results, one of the key assumptions used to show consistency of the estimates $\hat{f}_J(\mathbf{x})$, $\hat{c}_{Jk}^{(l)}$ and $\hat{\beta}_0$, is that the multiresolution index J is kept fixed. This ensures that $|\phi_{Jk}^{per}(x)| < \infty$, which enables the use of the dominated convergence theorem. Nonetheless, as it will be shown in the next section, it is possible to relax such assumption, enabling that $J = J(n)$ and furthermore, $J(n) \to \infty$ as $n \to \infty$.

3.2.3 $\mathbb{L}_2$ Risk Analysis of the Estimator $\hat{f}_J(\mathbf{x})$

In the last section, we showed that the estimates $\hat{f}_J(\mathbf{x})$, $\hat{c}_{Jk}^{(l)}$ and $\hat{\beta}_0$ are unbiased and consistent for $f_J(\mathbf{x})$, $c_{Jk}^{(l)}$ and β_0 respectively. In this section we provide a brief $\mathbb{L}_2$ risk analysis for the model estimate $f_J(\mathbf{x})$ and we show that $R(\hat{f}_J, f) = \mathbf{E}\left[||\hat{f}_J(\mathbf{x}) - f(\mathbf{x})||_2^2\right]$ converges to zero as $n \to \infty$.

As it will be demonstrated next, the rate of convergence of $\hat{f}_J(\mathbf{x})$ is influenced by the convergence properties of the kernel density estimator $\hat{h}_n(\mathbf{x})$ and the smoothness properties of the set $\left\{\phi_{J,k}^{per}(x), 0 \leq k \leq 2^J - 1\right\}$ generated by the scaling function $\phi(x)$, together with the functions $\{f_l(x)\}_{l=1}^p$ that define the additive model.

From the definition of $\hat{f}_J(\mathbf{x})$ and Cauchy-Schwartz inequality, it follows:

$$\mathbb{E}\left[||\hat{f}_J(\mathbf{x}) - f(\mathbf{x})||_2^2\right] \leq 2\left(\mathbb{E}\left[||\hat{f}_J(\mathbf{x}) - \mathbb{E}[\hat{f}_J(\mathbf{x})]||_2^2\right] + ||\mathbb{E}[\hat{f}_J(\mathbf{x})] - f(\mathbf{x})||_2^2\right) \tag{3.40}$$

Note that the first term on the rhs of (3.40) corresponds to the variance of the estimate $\hat{f}_J(\mathbf{x})$, while the second represents the square of the $bias(\hat{f}_J(\mathbf{x}))$.

Lemma 3.2.4. *Assume conditions **(A1)**-**(A5)** and **(Ak1)**-**(Ak4)** are satisfied. Then for $J = J(n)$ it follows:*

$$\mathbb{E}\left[||\hat{f}_J(\boldsymbol{x}) - \mathbb{E}[\hat{f}_J(\boldsymbol{x})]||_2^2\right] = \mathcal{O}\left(2^{J(n)}n^{-1}\right) . \tag{3.41}$$

The corresponding proof can be found in Appendix B.7.

Lemma 3.2.5. *In addition to conditions **(A1)**-**(A5)** and **(Ak1)**-**(Ak4)**, assume conditions 1-7 described in B.8 hold. Then:*

$$||\mathbb{E}[\hat{f}_J(\boldsymbol{x})] - f(\boldsymbol{x})||_2^2 = \mathcal{O}\left(2^{2J(n)}n^{-\frac{2\beta}{2\beta+p}} + 2^{-2J(n)(N+1)} + n^{-\frac{\beta}{2\beta+p}}2^{-J(n)(N+1)}\right) . \tag{3.42}$$

The corresponding proof can be found in Appendix B.8.

Lemma 3.2.6. *Define:*

$$\mathcal{F} = \left\{f \,|\, f_l \in L_2([0,1]),\, f_l \in W_2^{N+1}([0,1]),\, -\infty < m_l \leq f_l \leq M_l < \infty\right\},$$

where $f(\boldsymbol{x}) = \beta_0 + \sum_{l=1}^p f_l(x_l)$, and $W_2^{N+1}([0,1])$ represents the space of functions that

*are $N+1$-differentiable with $f_l^{(k)} \in \mathbb{L}_2([0,1])$, $k = 1, ..., N+1$. Suppose assumptions for lemmas 3.2.4 and 3.2.5 hold, and conditions **(A1)**-**(A5)** and **(Ak1)**-**(Ak4)** are satisfied. Then, it follows:*

$$\sup_{f \in \mathcal{F}} \left(\mathbb{E}\left[||\hat{f}_J(\boldsymbol{x}) - f(\boldsymbol{x})||_2^2 \right] \right) \leq \tilde{C} n^{-\left(\frac{2\beta}{2\beta+p}\right)\left(\frac{N+1}{N+3}\right)}, \tag{3.43}$$

provided (3.41) and (3.42), and $J = J(n)$ such that $2^{J(n)} \simeq n^{\frac{2\beta}{(2\beta+p)(N+3)}}$.

Also, it is possible to show that:

$$\mathbb{E}\left[||\hat{f}_J(\boldsymbol{x}) - \mathbb{E}[\hat{f}_J(\boldsymbol{x})]||_2^2 \right] = \mathcal{O}\left(n^{-\left(\frac{N+2}{N+3}\right)} n^{-\left(\frac{p}{2\beta+p}\right)} \right), \text{ and} \tag{3.44}$$

$$||\mathbb{E}[\hat{f}_J(\boldsymbol{x})] - f(\boldsymbol{x})||_2^2 = \mathcal{O}\left(n^{-\left(\frac{2\beta}{2\beta+p}\right)\left(\frac{N+1}{N+3}\right)} \right). \tag{3.45}$$

The corresponding proofs can be found in B.9.

Remarks and comments

(i) The additional assumptions described in B.8 are needed to use the wavelet approximation results presented in chapters 8-9 (Corollary 8.2) of [57].

(ii) As proposed in [57], the simplest way to obtain the wavelet approximation property utilized in the derivation of (3.42) is by selecting a bounded and compactly supported scaling function ϕ to generate $\left\{ \phi_{J,k}^{per}(x), 0 \leq k \leq 2^J - 1 \right\}$.

(iii) In the derivations for the convergence rate for the estimator $\hat{f}_J(\mathbf{x})$, the smoothness assumptions for the unknown functions f_l and the wavelet scaling function ϕ play a key role. In this sense, the index N corresponds to the minimum smoothness index among the unknown functions $\{f_1, ..., f_p\}$ and the scaling function ϕ.

(iv) From (3.44) and (3.45), it holds that the variance term of the estimator $\hat{f}_J(\mathbf{x})$, for large dimensions p is influenced primarily by the smoothness properties of the functional space that contains $\{f_l(x)\,,\, l=1,...,p\}$ and the wavelet basis $\left\{\phi^{per}_{J,k}(x), k=0,...,2^J-1\right\}$. Also, for n sufficiently large, the bias term dominates in the risk decomposition of $\hat{f}_J(\mathbf{x})$.

(v) As a result of the introduction of the density estimator $\hat{h}_n(\mathbf{x})$ in the model, $\hat{f}_J(\mathbf{x})$ suffers from the curse of dimensionality. In particular, it is interesting to note that this effect affects only the bias term, since as $p \to \infty$, $\mathbb{E}\left[||\hat{f}_J(\mathbf{x}) - \mathbb{E}[\hat{f}_J(\mathbf{x})]||_2^2\right] \to \mathcal{O}(n^{-\frac{7}{4}})$, for $N \geq 1$.

(vi) An alternative way to show the mean square consistency of the estimator $\hat{f}_J(\mathbf{x})$ is via Stone's theorem (details can be found in Theorem 4.1 [9]), by assuming a model with no intercept (i.e. $\beta_0 = 0$), and expressing the estimator as:

$$\hat{f}_J(\mathbf{x}) = \sum_{i=1}^{n} W_{n,i}(\mathbf{x}) \cdot y_i\,,$$

where $W_{n,i}(\mathbf{x}) = \sum_{l=1}^{p}\sum_{k=0}^{2^J-1}\left(\frac{\phi^{per}_{Jk}(X_{il})-2^{-\frac{J}{2}}}{n\cdot\hat{h}_n(\mathbf{X}_i)}\right)\phi^{per}_{Jk}(x_l)$. Then, the estimator is mean-square consistent provided the following conditions are satisfied:

i. For any n, $\exists\, c \in \mathbb{R}$ such that for every non-negative measurable function f satisfying $\mathbb{E}f(\mathbf{X}) < \infty$, $\mathbb{E}\left\{\sum_{i=1}^{n}|W_{n,i}(\mathbf{x})f(\mathbf{X}_i)|\right\} \leq c\,\mathbb{E}f(\mathbf{X})$.

ii. For all n, $\exists\, D \geq 1$ such that $\mathbb{P}\left\{\sum_{i=1}^{n}|W_{n,i}(\mathbf{x})| \leq D\right\} = 1$.

iii. For all $a > 0$, $\lim_{n\to\infty}\mathbb{E}\left\{\sum_{i=1}^{n}|W_{n,i}(\mathbf{x})|\,\mathbf{1}_{\{||\mathbf{X}_i-\mathbf{x}||>a\}}\right\} = 0$.

iv. $\sum_{i=1}^{n} W_{n,i}(\mathbf{x}) \xrightarrow[n\to\infty]{\mathbb{P}} 1$.

v. $\lim_{n\to\infty}\mathbb{E}\left\{\sum_{i=1}^{n} W_{n,i}(\mathbf{x})^2\right\} = 0$.

(vii) Indeed, for the estimator $\hat{f}_J(\mathbf{x})$ conditions (a)-(e) are satisfied, provided assumptions (**A1**)-(**A5**) and (**Ak1**)-(**Ak4**) hold, and for all $\mathbf{x} \in [0,1]^p$, $\lim_{n\to\infty}\left|1 - \frac{h(\mathbf{x})}{\hat{h}_n(\mathbf{x})}\right| = 0$.

3.2.4 Implementation illustration and considerations and comparison to other estimators.

Implementation Illustration

In this section, we illustrate the application of the proposed method in a controlled experiment, comparing the proposed estimator results with previously existing methodologies AMlet [1] and Back-fitting [42]. For this purpose, we choose the following functions for the construction of model (4.1):

$$
\begin{aligned}
f_1(x) &= \frac{1}{\sqrt{2}\sin(2\pi\, x)} \\
f_2(x) &= 1 - 4\,|x - \frac{1}{2}| \\
f_3(x) &= -\cos(4\pi\, x + 1) \\
f_4(x) &= 8\left(x - \frac{1}{2}\right)^2 - \frac{2}{3} \\
f_5(x) &= \frac{1}{\sqrt{2}}\cos(2\pi\, x) \\
f_6(x) &= \frac{1}{\sqrt{2}}\cos(4\pi\, x) \\
f_7(x) &= -0.5275 + 4\, e^{-500(x-0.23)^2} + 2\, e^{-2000(x-0.33)^2} \\
&\quad +4\, e^{-8000(x-0.47)^2} + 3\, e^{-16000(x-0.69)^2} + e^{-32000(x-0.83)^2} \\
f_8(x) &= 0.2\,\cos(4\pi\, x + 1) + 0.1\,\cos(24\pi\, x + 1) \\
f_9(x) &= -0.1744 + 2\, x^3\, \mathbf{1}_{(0.5<x\leq 0.8)} + 2\,(x-1)^3\, \mathbf{1}_{(0.8<x\leq 1)}
\end{aligned}
$$

The estimator $\hat{f}_J(\mathbf{x})$ was obtained using a box-type kernel with a bandwidth given by $\delta(n) = Kn^{-\frac{1}{4+p}}$, with K found via grid search. For the multiresolution space index J, we chose $J(n) = 5 + \lfloor 0.3\,\log_2(n) \rfloor$. The selection of the wavelet filter was Daubechies with 6 vanishing moments and the sample sizes used for this illustration were $n = 2048,\ 4096$ and 8192.

Similarly, the noise in the model was defined to be gaussian with zero mean and variance

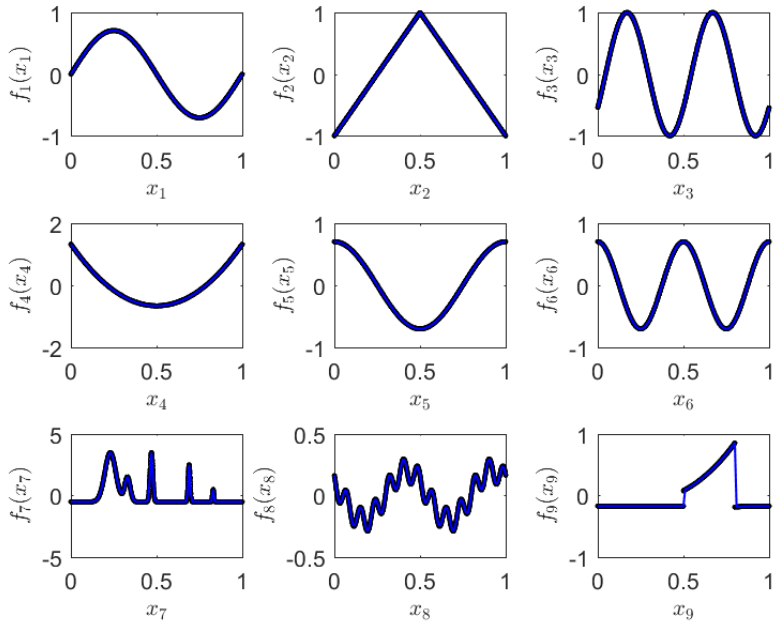


Figure 3.1: Graphic representation of the testing functions for the Additive Model.

given by $\sigma^2 = 0.15$. Finally, the joint distribution for the predictors $\mathbf{X}_1, ..., \mathbf{X}_n$ was generated by independent $\mathcal{U}(0,1)$ and a $Beta(3,3)$ random variables along each dimension. For the evaluation of the scaling functions ϕ_{Jk}^{per} we used Daubechies-Lagarias's algorithm.

The comparative results for the simulation study are shown in Figs. 3.2, 3.3, 3.4. Box plots with results for each function recovery RMSE are shown in Fig. 3.5. Average simulation times for the described setting are shown in Table 3.1.

Remarks and comments

(i) Choice of bandwidth for the density estimator $\hat{h}_n(\mathbf{x})$: During the implementation, we observed that results were highly sensitive to the choice of the bandwidth $\delta(n)$. We chose different values for a constant K in a bandwidth of the form $\delta(n) = K\, n^{-\frac{1}{2+p}}$.

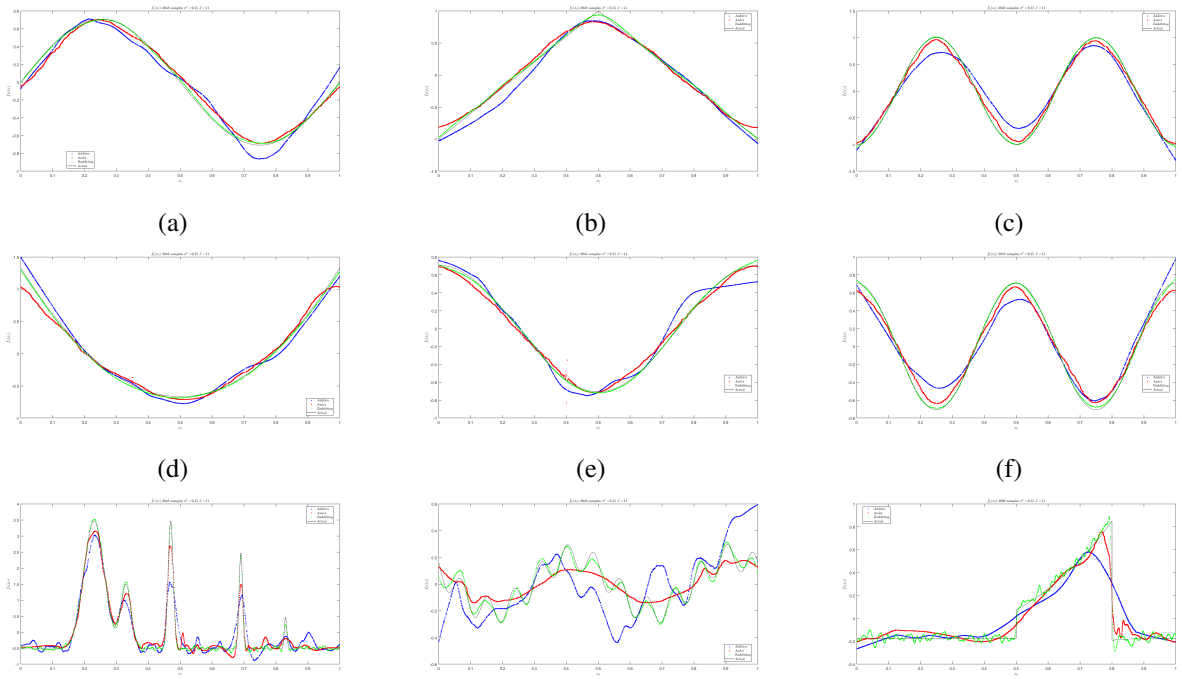


Figure 3.2: Functions estimation for $\mathcal{U}(0,1)$ designs, for $n = 2048$ samples. In red, the estimated function values at each sample point using AMlet; In black-dashed lines, the actual function shape; In blue lines, the smoothed version of the function values using lowess smoother and the wavelet-based method. In green, the estimated functions via backfitting.

Figures 3.2-3.4 show results obtained using K found via grid-search.

(ii) Sample size effect: As can be observed in 3.2-3.3, both the bias and the variance of the estimated functions show a decreasing behavior as n increases, which is consistent with theoretical results (3.43), (3.44) and (3.45).

(iii) Shadowing effect of the constant β_0: In some experiments, when the constant β_0 was too large with respect to the function effects, we observed that the method recovered the marginal densities of each predictor instead of the unknown functions. This effect can be explained from the expressions for the calculation of the empirical wavelet coefficients $\hat{c}_{Jk}^{(l)}$. For this reason, we recommend standardizing the response from the observed sample before fitting the model.

(iv) Sensitivity of the model to different random designs: In the case of design distributions

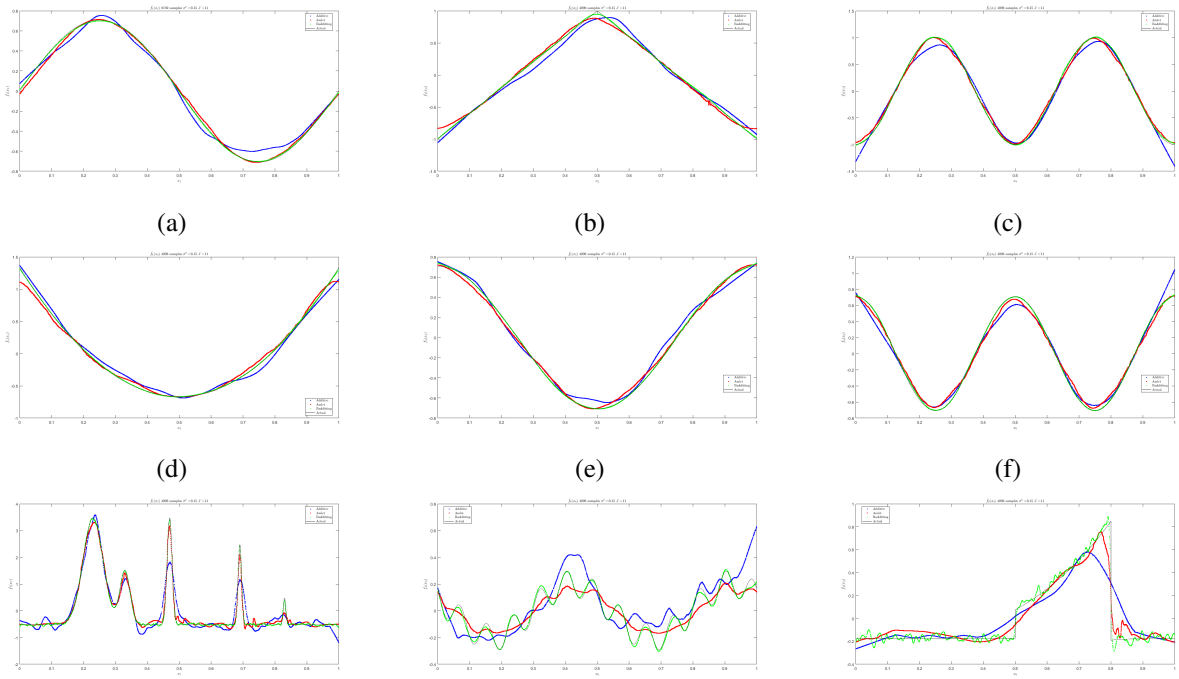
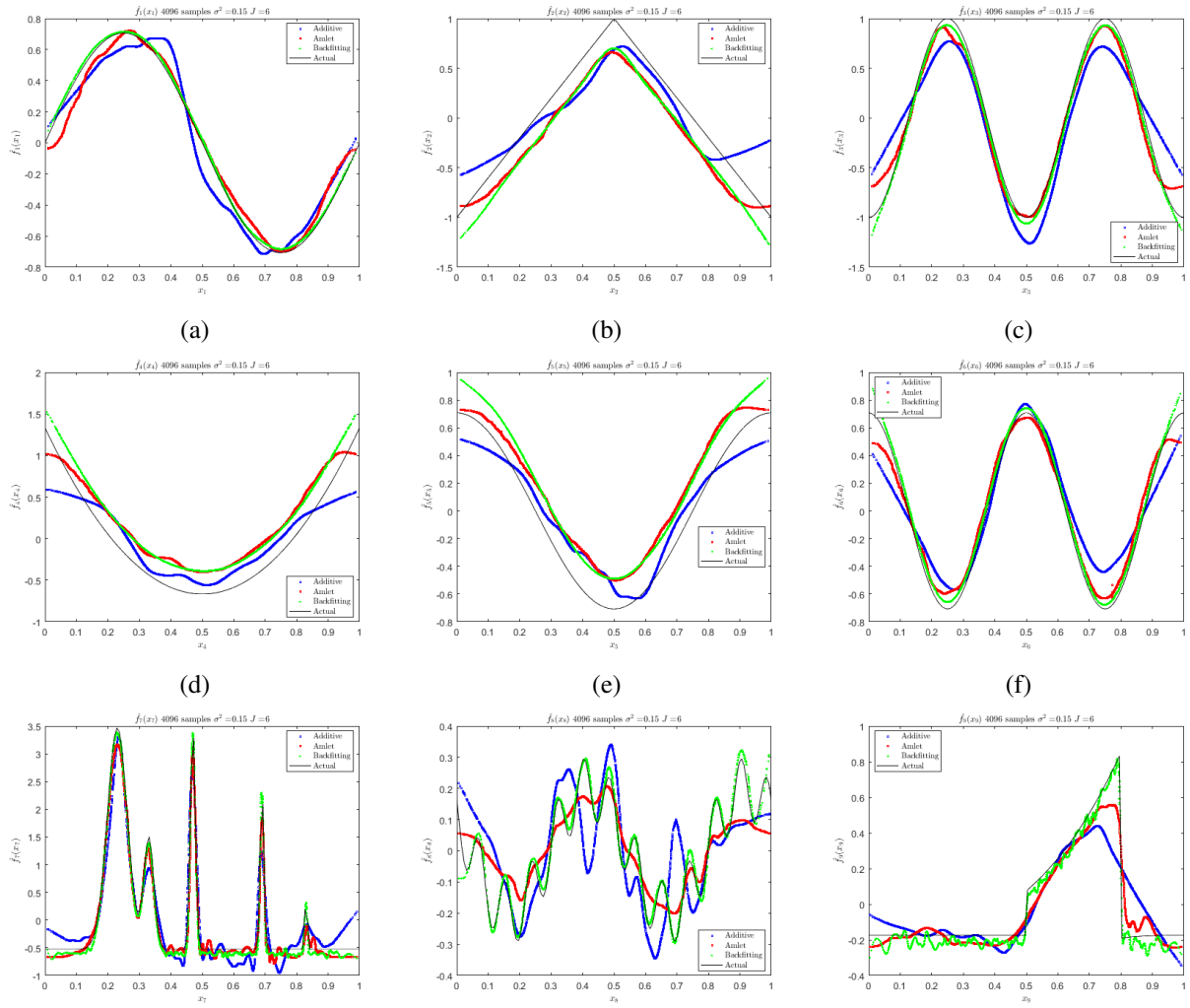


Figure 3.3: Functions estimation for $\mathcal{U}(0,1)$ designs, for $n = 4096$ samples. In red, the estimated function values at each sample point using AMlet; In black-dashed lines, the actual function shape; In blue lines, the smoothed version of the function values using lowess smoother and the wavelet-based estimator. In green, the estimated functions via backfitting.

that have fast decaying tails, problems were observed when there was no sufficient information for the estimation of the empirical coefficients in regions with low concentration of samples. Indeed, extremely large empirical wavelet coefficients were obtained in those cases, inflating the bias in the estimation.

(v) A possible remedial action for situation could be the use of the approach proposed in [52], by thresholding the density estimates according to some probabilistic rule, avoiding those samples for which $\hat{h}_n(\mathbf{x})$ is smaller than a suitably defined $\lambda_n > 0$.

(vi) Avoiding the curse of dimensionality: As noted in the previous sections, the proposed estimator suffers from the "curse of dimensionality". In particular, this effect arises as a result of the introduction of the non-parametric density estimator $\hat{h}_n(\mathbf{x})$ of the true density $h(\mathbf{x})$ of observed features $\mathbf{x}_i \in \mathbb{R}^p, i = 1, ..., n$. If instead, we assume that



(a) (b) (c)

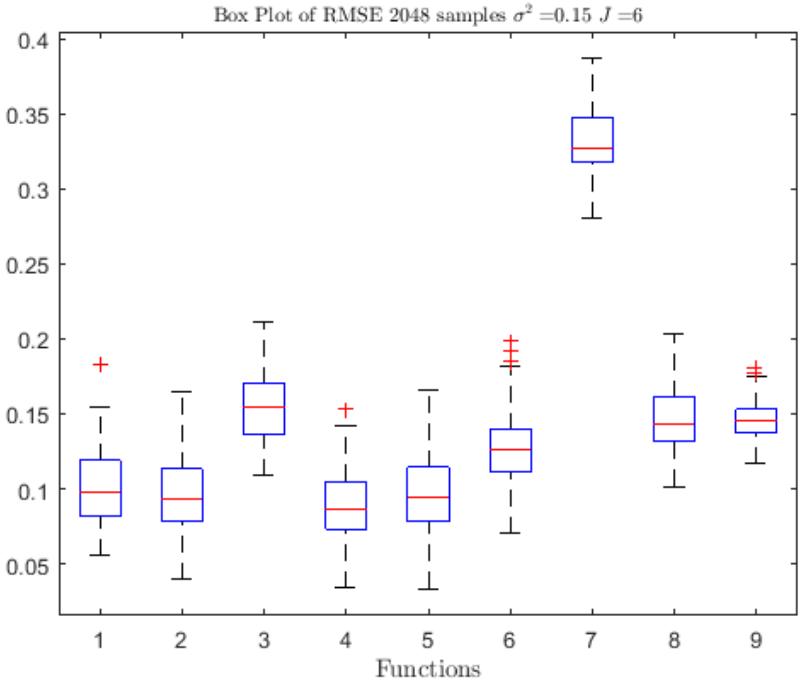
(d) (e) (f)

(g) (h) (i)

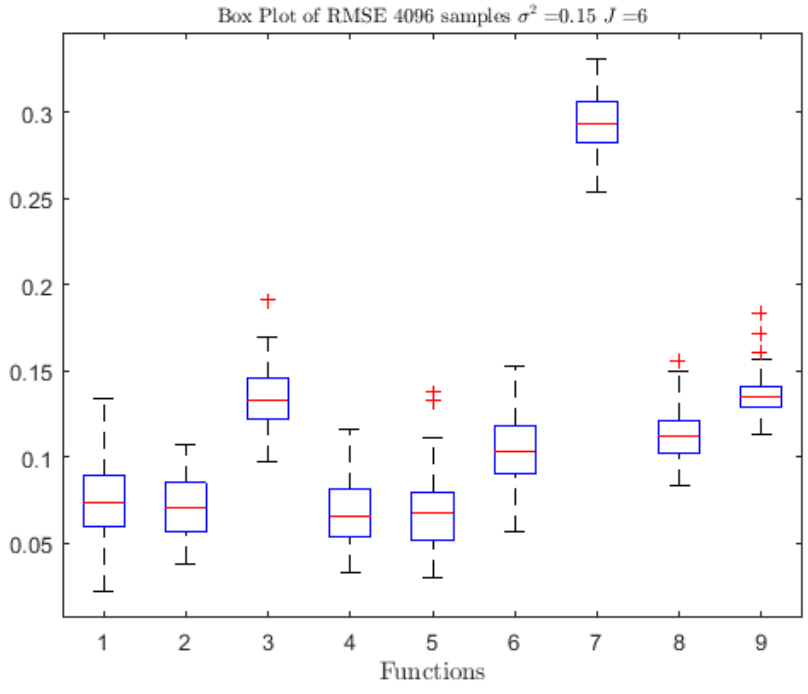
Figure 3.4: Functions estimation for $Beta(3,3)$ design, $n = 4096$ samples. In red, the estimated function values at each sample point via AMlet; In black-dashed lines, the actual function shape; In blue lines, the smoothed version of the function values using lowess smoother. In green, the estimated functions via back-fitting.

$h(\mathbf{x}) = \prod_{l=1}^{p} h_l(x_l)$, the resulting estimator converges at a rate $\sim n^{\frac{\beta}{2\beta+1}}$, which enables $\hat{f}_J(\mathbf{x})$ to achieve the rates shown in Proposition 7, with $p = 1$.

(vii) Computational run time: As observed in the implementation and simulation study, the proposed method is relatively fast for small sample sizes, however, since it is based on the evaluation of the scaling functions $\phi_{Jk}^{per}(x), 0 \leq k \leq 2^J - 1$ at every observed feature $x_{il}, i = 1, ..., n; \; l = 1, ..., p$ this process can be computationally intense for



(a)



(b)

Figure 3.5: Box plots for function recovery RMSE for (a) $n = 2048$ samples, and (b) $n = 4096$ samples using $\mathcal{U}(0, 1)$ design. 100 replications were used in the experiment.

large n and/or p. Using a 4-AMD cores, 16Gb RAM laptop, average computational times are the following: As it can be observed in Table 3.1, for large sample sizes the

Table 3.1: Average computational times

N	p	Filter	Time (sec.)
256	9	Daub6	5.6
2048	9	Daub6	67.7
8192	9	Daub6	419.9

computational cost of the algorithm is significant. This cost is mainly driven by the evaluations of the scaling functions at each sample point, for each dimension and shift. Once this data is available, the cost of the algorithm is significantly reduced.

3.2.5 Conclusions and Discussion

This Chapter introduced a wavelet-based method for the non-parametric estimation and prediction of non-linear additive regression models. Our estimator is based on data-driven wavelet coefficients computed using a locally weighted average of the observed samples, with weights defined by scaling functions obtained from an orthonormal periodic wavelet basis and a non-parametric density estimator $\hat{h}_n$. For this estimator, we showed mean-square consistency and illustrated practical results using theoretical simulations. In addition, we provided convergence rates and optimal choices for the tuning parameters for the algorithm implementation.

As was presented in this Chapter, the proposed estimator is completely data driven with only a few parameters of choice by the user (i.e. bandwidth $\delta(n)$, multiresolution index $J(n)$ and wavelet filter). Indeed, the nature of the estimator allows a block-matrix based implementation that introduces computational speed and makes the estimator suitable for real-life applications. In our implementation, Daubechies-Lagarias's algorithm was used to

evaluate the scaling functions ϕ_{Jk}^{per} at the observed sample points X_{ij}. From a computational viewpoint, this key component represents most of the computational cost of this algorithm.

Furthermore, we tested our method using different exemplary baseline functions and two random designs via a simulation study. In our experiments, the proposed method showed good performance identifying the unknown functions in the model, as compared to popular methods as back-fitting and AMlet, even though it suffers from the "curse of dimensionality"; Also, we observed that the estimator behaves accordingly to the large properties that were theoretically shown, which is an important feature for real-life applications.

In terms of some of the drawbacks, we can mention that our method does not offer automatic variable selection; however, this could be implemented by thresholding the obtained empirical wavelet coefficients in a post-estimation stage or by simple inspection, since a function that is zero over [0,1] maps to zero in the wavelet projection. Similarly, the proposed estimator was observed to be highly sensitive to the bandwidth choice $\delta(n)$, consequently, the use of cross-validation or grid-search during the estimation stage might be helpful to improve the accuracy of results.

Finally, in those design regions were the number of observed samples is small it is possible to obtain abnormally large wavelet coefficients; also as a result of the use of periodic wavelets, some problems may arise at the boundaries of the support for each function. Nonetheless, this can be fixed: using the idea developed by Pensky and Vidakovic (2001) [52], it is possible to avoid those samples that are associated with too-small density estimates $\hat{h}_n$, stabilizing the estimated wavelet coefficients and reducing the estimator bias.

Based on out theoretical analysis and numerical experiments, we can argue that our proposed method exhibits good statistical properties and is relatively easy to implement, which constitutes a good contribution in the statistical modeling field and in particular, in the analysis of the non-linear additive regression models.

CHAPTER 4

LEAST SQUARES WAVELET-BASED ESTIMATION FOR ADDITIVE REGRESSION MODELS USING NON EQUALLY-SPACED DESIGNS

As was mentioned in Chapter 3, Additive regression models are actively researched in the statistical field because of their usefulness in the analysis of responses determined by non-linear relationships with multivariate predictors. In this kind of statistical models, the response depends linearly on unknown functions of predictor variables and typically, the goal of the analysis is to make inference about these functions.

In this Chapter, we study the problem of additive regression using a very simple least squares approach based on a periodic orthogonal wavelets expansion on the interval [0,1]. For this estimator, we analyze its statistical properties, showing strong consistency (with respect to the $\mathbb{L}_2$ norm) characterized by optimal convergence rates up to a logarithmic factor, independent of the dimensionality of the problem. This is achieved by truncating the model estimates by a properly chosen parameter, and selecting the multiresolution level J used for the wavelet expansion, as a function of the sample size. In this approach, we obtain these results without the assumption of an equispaced design, a condition that is typically assumed in most wavelet-based procedures.

Finally, we show practical results obtained from a simulation study and a real life application, demonstrating the applicability of the proposed methods for the problem of non-linear robust additive regression models.

4.1 Introduction

Additive regression models are popular in the statistical field because of their usefulness in the analysis of responses determined by non-linear relationships involving multivariate predictors. In this kind of statistical models, the response depends linearly on unknown functions of the predictors and typically, the goal of the analysis is to make inferences about these functions. This model has been extensively studied through the application of piecewise polynomial approximations, splines, marginal integration, as well as back-fitting or functional principal components. Chapter 15 of [39], Chapter 22 of [9] and [40], [41] and [42] feature thorough discussions of the issues related to fitting such models and provide a comprehensive overview and analysis of various estimation techniques for this problem.

In general, the additive regression model relates a univariate response Y to predictor variables $\mathbf{X} \in \mathbb{R}^p$, $p \geq 1$, via a set of unknown non-linear functions $\{f_l \,|\, f_l : \mathbb{R} \to R\,,\, l = 1, ..., p\}$. The functions f_l may be assumed to have a specified parametric form (e.g. polynomial) or may be specified non-parametrically, simply as "smooth functions" that satisfy a set of constraints (e.g. belong to a certain functional space such as a Besov or Sobolev, Lipschitz continuity, spaces of functions with bounded derivatives, etc.). Though the parametric estimates may seem more attractive from the modeling perspective, they can have a major drawback: a parametric model automatically restricts the space of functions that is used to approximate the unknown regression function, regardless of the available data. As a result, when the elicited parametric family is not "close" to the assumed functional form the results obtained through the parametric approach can be misleading. For this reason, the non-parametric approach has gained more popularity in statistical research, providing a more general, flexible and robust approach in tasks of functional inference.

In this Chapter we study the problem of additive regression with random designs using a

simple least squares methodology based on a periodic orthogonal wavelet basis on the interval [0,1]. This, motivated by the goal of providing a different and more natural approach than the one provided in Chapter 3. In addition, given the simplicity of the Least Squares approach, we provide an in-depth theoretical analysis of its statistical properties of consistency and convergence rate.

Our results show that in this approach when is possible to choose the detail level $J = J(n)$ of the multiresolution space V_J such that an ill-conditioned design matrix is avoided, strongly consistent estimators (with respect to the $\mathbb{L}_2$ norm) can be obtained by truncating the estimated regression function using a suitable threshold parameter that depends on the sample size n. In this setting, we demonstrate that it is possible to achieve optimal convergence rates up to a logarithmic factor, independent of the dimensionality of the problem. Moreover, we obtain these results without the assumption of an equispaced design for the application of the wavelet procedures.

The choice of wavelets as an orthonormal basis is motivated by the fact that wavelets could be well localized in both time and scale (frequency), and possess superb approximation properties for signals with rapid local changes such as discontinuities, cusps, sharp spikes, etc.. Moreover, the representation of these signals in the form of wavelet decompositions can be accurately done using only a few wavelet coefficients, enabling sparsity and dimensionality reduction. This adaptivity does not, in general, hold for other standard orthonormal bases (e.g. Fourier basis) which may require many compensating coefficients to describe signal discontinuities or local bursts.

In addition, we show the potential of the proposed methodology via a simulation study and evaluate its performance using different exemplary functions and random designs, under different sample sizes. Here, we demonstrate that the proposed method is suitable for the problem of non-linear additive regression models and behave in coherence with the obtained the-

oretical results. Finally, we compare the results obtained through our proposed methodology against a previously published study, using a real life data set.

As it was mentioned in the previous chapter, additive regression models have been studied by many authors using a wide variety of approaches. The approaches include marginal integration, back-fitting, least squares (including penalized least squares), orthogonal series approximations, and local polynomials. Short descriptions of the most commonly used techniques can be found in Chapter 3.

4.2 Wavelet-based Estimation in Additive Regression Models

Suppose that instead of the typical linear regression model $y = \sum_{j=1}^{p} \beta_j x_j + \beta_0 + \epsilon$ which assumes linearity in the predictors $\mathbf{x} = (x_1, ..., x_p)$, we have the following:

$$\begin{aligned} f(\mathbf{x}) &= \beta_0 + f_A(\mathbf{x}) + \sigma \cdot \epsilon \\ &= \beta_0 + \sum_{j=1}^{p} f_j(x_j) + \sigma \cdot \epsilon\,, \end{aligned} \tag{4.1}$$

where ϵ, independent of $\mathbf{x}$, $\mathbb{E}[\epsilon] = 0$, $\mathbb{E}[\epsilon^2] = 1$, $\sigma > 0$, $\sigma < \infty$. Similarly, $\mathbf{x}_i \overset{\text{iid}}{\sim} h(\mathbf{x})$, an unknown design density of observations and $\{f_1(\cdot), ..., f_p(\cdot)\}$ are unknown real-valued functions to be estimated.

Suppose that we are able to observe a sample $\{y_i = f(\mathbf{x}_i), \mathbf{x}_i\}_{i=1}^{n}$ where $\mathbf{x}_1, ..., \mathbf{x}_n \overset{iid}{\sim} h(\mathbf{x})$. We are interested in estimating β_0 and $\{f_1(\cdot), ..., f_p(\cdot)\}$. For simplicity (without loss of generality) and identifiability, we assume:

(A1) The density $h(\mathbf{x})$ is of the continuous type and has support in $[0, 1]^p$. Also, we assume $\exists \epsilon_h > 0$ s.t. $h(\mathbf{x}) \geq \epsilon_h \ \forall \mathbf{x} \in [0, 1]^p$.

(A2) For $k = 1, ..., p$, $\int_0^1 f_k(x) dx_k = 0$.

(A3) For $k = 1, ..., p$, $\sup_{x\in[0,1]} |f_k(x)| \leq M_k < \infty$ and $\inf_{x\in[0,1]} \{f_k(x)\} \geq m_k > -\infty$. This implies that for $k = 1, ..., p$, $f_k \in \mathbb{L}_2([0, 1])$.

(A4) The density $h(\mathbf{x})$ is uniformly bounded in $[0, 1]^p$, that is, $\forall \mathbf{x} \in [0, 1]^p$, $|h(\mathbf{x})| \leq M$, $M < \infty$.

Furthermore, since as $J \to \infty$ the orthonormal set $\left\{\phi^{per}_{J,k}(x), 0 \leq k \leq 2^J - 1\right\}$ spans $\mathbb{L}_2([0, 1])$, each of the functions in (4.1) can be represented as:

$$f_l(x) = \lim_{j\to\infty} \sum_{k=0}^{2^j-1} c^{(l)}_{jk} \cdot \phi^{per}_{jk}(x), \quad l = 1, ..., p\,, \tag{4.2}$$

where $c^{(l)}_{jk}$ denotes the $j, k-$th wavelet coefficient of the $l-$th function in the model. Similarly, for some fixed J, $f_{l,J}(x)$, $l = 1, ..., p$ represents the orthogonal projection of $f_l(x)$ onto the multiresolution space V_J. Therefore, $f_{l,J}(x)$ can be expressed as:

$$f_{l,J}(x) = \sum_{k=0}^{2^J-1} c^{(l)}_{Jk} \cdot \phi^{per}_{Jk}(x), \quad l = 1, ..., p\,, \tag{4.3}$$

where:

$$c^{(l)}_{Jk} = \langle f_l(x), \phi^{per}_{Jk}(x)\rangle = \int_0^1 f_l(x)\phi^{per}_{Jk}(x)dx, \quad l = 1, ..., p\,. \tag{4.4}$$

Based on the model (4.1) and (4.3), it is possible to approximate $f(\mathbf{x})$ by an orthogonal projection $f_J(\mathbf{x})$ onto the multiresolution space spanned by the set of scaling functions:

$$\left\{\phi^{per}_{J,k}(x), 0 \leq k \leq 2^J - 1\right\},$$

by approximating each of the functions $f_l(\cdot)$, $l = 1, ..., p$ as described above. Therefore,

$f_J(\mathbf{x})$ can be expressed as:

$$f_J(\mathbf{x}) = \beta_0 + \sum_{l=1}^{p} \sum_{k=0}^{2^J-1} c_{Jk}^{(l)} \phi_{Jk}^{per}(x)\,. \tag{4.5}$$

Now, the goal is for a pre-specified multiresolution index J, to use the observed samples to estimate the unknown constant β_0 and the orthogonal projections of the functions $f_{l,J}(x)$, $l = 1, ..., p$.

Remarks

(i) Also, from the above conditions, the variance of the response $y(\mathbf{x})$ is bounded for every $\mathbf{x} \in \mathbb{R}^p$.

(ii) The assumption that the support of the random vector $\mathbf{X}$ is $[0, 1]^p$ can be always satisfied by carrying out appropriate monotone increasing transformations of each dimensional component, even in the case when the support before transformation is unbounded. In practice, it would be sufficient to transform the empirical support to $[0, 1]^p$.

4.3 A Least Squares approach for non-linear Additive model estimation using orthogonal wavelet basis

As it is shown in Chapter 22 of [9], it is possible to study the problem of additive regression using least squares. The empirical $\mathbb{L}_2$ risk is minimized over a linear spaced spanned by a defined orthogonal basis with dimension depending on the sample size. In this setting, consider the unknown functions $\{f_1, ..., f_p\}$ to be approximated by their respective orthogonal projections onto the multiresolution space V_J spanned by a given set of scaling functions $\left\{\phi_{J,k}^{per}(x), k = 0, ..., 2^J - 1\right\}$, for a given multiresolution index J. Consequently, the projection of the function $f_A(\mathbf{x}) = \sum_{j=1}^{p} f_j(x_j)$ onto V_J belongs to the linear space defined as:

$$\mathcal{F}_n = \left\{ f : [0,1]^p \to \mathbb{R} \,|\, f(\mathbf{x}) = \sum_{j=1}^{p} \sum_{k=0}^{2^{J(n)}-1} c^{(j)}_{J(n),k} \phi^{per}_{J(n),k}(x_j)\,,\ \mathbf{x} \in [0,1]^p \right\}\,, \tag{4.6}$$

where $x_j\,, j = 1, ..., p$ corresponds to the j-th component of the vector $\mathbf{x} \in [0,1]^p$. Thus, this projection of $f_A(\mathbf{x})$ onto $\mathcal{F}_n$ is defined by the set of coefficients:

$$\left\{ c^{(j)}_{J,k}\,,\ j = 1, ..., p\,;\ k = 0, ..., 2^{J(n)} - 1 \right\}.$$

As it is shown in [5], by the properties of MRA, $\cup_{j\geq 0} V_j$ is dense in $\mathbb{L}_2([0,1])$, where V_j is the space spanned by the orthonormal basis $\left\{ \phi^{per}_{J,k}(x), k = 0, ..., 2^J - 1\,; \right\}$ for some $J \geq 0$. Therefore, for any Lebesgue measure $\mu(\cdot)$ in $\mathbb{R}$ that is bounded away from zero and infinity in its support, we have that $\cup_{j\geq 0} V_j$ is dense in $\mathbb{L}_2\left(\mu([0,1])\right)$, thus the following result holds:

Lemma 4.3.1. *For any $f \in \mathbb{L}_2([0,1])$, $\epsilon > 0$ and bounded Lebesgue measure $\mu(\boldsymbol{x})$ in $\mathbb{R}^p$, $\exists \left\{ c^{(1)*}_{J,0}, ..., c^{(1)*}_{J,2^J-1}, ..., c^{(p)*}_{J,0}, ..., c^{(p)*}_{J,2^J-1} \right\}$ for which $J = J^*(n_0(\epsilon))$, such that:*

$$\int_{[0,1]^p} \left| \sum_{j=1}^{p} \left(\sum_{k=0}^{2^J-1} c^{(j)}_{J,k} \phi^{per}_{J,k}(x_j) - f_j(x_j) \right) \right|^2 \mu(d\boldsymbol{x}) \leq \epsilon\,. \tag{4.7}$$

The proof of the above assertion follows from the application of the inequality $(\sum_{j=1}^{d} a_j)^2 \leq d \cdot \sum_{j=1}^{d} a_j^2$, together with the fact that $\bigcup_{j\geq 0} V_j$ is dense in $\mathbb{L}_2\left(\mu([0,1])\right)$. This enables to find a multiresolution index J as a function of the sample size n sufficiently large, such that it is possible to approximate each of the functions f_j with a precision $\epsilon_j \leq \frac{\epsilon}{p \cdot ||\mu||_\infty}\,,\ j = 1, ..., p$, for $||\mu||_\infty$ defined as the infinity norm of the Lebesgue measure μ.

4.3.1 Least Squares problem formulation.

Following (4.1), suppose a model of the form:

$$y(\mathbf{x}) = f_A(\mathbf{x}) + \sigma \cdot \epsilon\,. \tag{4.8}$$

Assume conditions stated in 4.3 are satisfied. From (4.7), for a sample $\{(\mathbf{X}_i, Y_i)\}_{i=1}^n$, and a given multiresolution index $J = J(n)$, it is possible to define a least squares estimator of $f(\mathbf{x})$ over the space of functions defined by $\mathcal{F}_n$ in (4.6), as follows:

$$\begin{aligned}\hat{f}_{J(n)} &= \inf_{f\in\mathcal{F}_n} \frac{1}{n}\sum_{i=1}^{n} |f(\mathbf{X}_i) - Y_i|^2 \ , \\ &= \underset{\left\{c_{J,k}^{(j)},\, j=1,\ldots,p\,;\, k=0,\ldots,2^{J(n)}-1\right\}}{\arg\min} \frac{1}{n}\sum_{i=1}^{n}\left|\sum_{j=1}^{p}\sum_{k=0}^{2^J-1} c_{J,k}^{(j)}\phi_{J,k}^{per}(X_{ij}) - Y_i\right|^2 . \end{aligned} \tag{4.9}$$

Define:

$$\mathbf{c} = \begin{bmatrix} c_{J,0}^{(1)} \\ \vdots \\ c_{J,2^J-1}^{(1)} \\ \vdots \\ c_{J,0}^{(p)} \\ \vdots \\ c_{J,2^J-1}^{(p)} \end{bmatrix}_{p\cdot 2^{J(n)}\times 1} , \qquad \mathbf{B}(\mathbf{x}_i) = \begin{bmatrix} \phi_{J,0}^{per}(x_{i1}) \\ \vdots \\ \phi_{J,2^J-1}^{per}(x_{i1}) \\ \vdots \\ \phi_{J,0}^{per}(x_{ip}) \\ \vdots \\ \phi_{J,2^J-1}^{per}(x_{ip}) \end{bmatrix}_{p\cdot 2^{J(n)}\times 1} , \tag{4.10}$$

$$\mathbf{B} = \begin{bmatrix} \mathbf{B}(\mathbf{x}_1)^T \\ \vdots \\ \mathbf{B}(\mathbf{x}_n)^T \end{bmatrix}_{n\times p\cdot 2^{J(n)}} , \qquad \mathbf{Y} = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix}_{n\times 1} . \tag{4.11}$$

Then, it is possible to represent (4.9) as:

$$\hat{f}_{J(n)} = \underset{\mathbf{c}\in\mathbb{R}^{p\cdot 2^{J(n)}}}{\arg\min} \frac{1}{n}\left\|\mathbf{B}\cdot\mathbf{c} - \mathbf{Y}\right\|_2^2 . \tag{4.12}$$

Assuming that $\mathbf{X}_1, ..., \mathbf{X}_n$ have continuous joint distribution and $p \cdot 2^{J(n)} \leq n$, the matrix $\mathbf{B}$ is non-singular (since the event in which $\mathbf{X}_1, ..., \mathbf{X}_n$ are all distinct happens with probability 1). Therefore, the problem defined by (4.12) has a unique solution given by:

$$\mathbf{c}^* = \left(\mathbf{B}^T\mathbf{B}\right)^{-1}\mathbf{B}^T\mathbf{Y}\,. \tag{4.13}$$

Therefore, for a new observation $\mathbf{x}$, the estimator $\hat{f}_{J(n)}(\mathbf{x})$ can be represented as:

$$\hat{f}_{J(n)}(\mathbf{x}) = \mathbf{B}(\mathbf{x})^T\mathbf{c}^*\,. \tag{4.14}$$

4.3.2 Strong consistency of the Linear Least Squares Estimator.

In this section, we investigate the consistency property for the least squares estimator defined by equations (4.13) and (4.14). Throughout the analysis, we will use results and definitions contained in C.1 of the appendix, which have been previously introduced in the statistical literature.

Theorem 4.3.1. *Strong consistency of the Wavelet-based Least Squares Estimator*

Suppose an orthonormal basis $\left\{\phi_{j,k}^{per}(x), k = 0, ..., 2^J - 1, \right\}$ *which for* $J \to \infty$ *is dense in* $\mathbb{L}_2(\nu([0,1]))$ *for* $\nu \in \Upsilon$*, and let* Υ *be the set of bounded Lebesgue measures in* $[0,1]$*. Suppose* μ *is a bounded Lebesgue measure in* $[0,1]^p$*, and the following conditions are satisfied for the scaling function* ϕ*:*

(a) $\exists\,\Phi$*, bounded and non-increasing function in* $\mathbb{R}$ *such that* $\int \Phi(|u|)du < \infty$ *and* $|\phi(u)| \leq \Phi(|u|)$ *almost everywhere (a.e.).*

(b) In addition, $\int_{\mathbb{R}} |u|^{N+1}\Phi(|u|) < \infty$ *for some* $N \geq 0$*.*

(c) $\exists\,F$*, integrable, such that* $|K(x,y)| \leq F(x-y)$*,* $\forall x, y \in \mathbb{R}$*, for* $K(x,y) = \sum_k \phi(x-k)\phi(y-k)$*.*

(d) Suppose ϕ satisfies:

i. $\sum_k |\hat{\phi}(\xi + 2k\pi)|^2 = 1$*, a.e., where* $\hat{\phi}$ *denotes the Fourier transform of the scaling function* ϕ*.*

ii. $\hat{\phi}(\xi) = \hat{\phi}(\frac{\xi}{2}) m_0(\frac{\xi}{2})$*, where* $m_0(\xi)$ *is a* 2π*-periodic function and* $m_0 \in \mathbb{L}_2(0, 2\pi)$*.*

(e) $\int_{\mathbb{R}} x^k \psi(x) = 0$*, for* $k = 0, 1, ..., N$*,* $N \geq 1$ *where* ψ *is the mother wavelet corresponding to* ϕ*.*

(f) The functions $\{f_l\}_{l=1}^p$*, are such that* $f_l \in L_\infty([0,1])$ *and* $f_l \in W_\infty^{m+1}([0,1])$*,* $m \geq N$*, where* $W_\infty^m([0,1])$ *denotes the space of functions that are* m*-times weakly-differentiable and* $f_l^{(k)} \in L_\infty([0,1])$*,* $k = 1, ..., m$*.*

(g) $\theta_\phi(x) := \sum_k |\phi(x-k)|$ *such that* $||\theta_\phi||_\infty < \infty$*.*

According to corollary 8.2 [57], if $f \in W_\infty^{N+1}([0,1])$ *then* $||K_J f - f||_\infty^p = \mathcal{O}\left(2^{-pJ(N+1)}\right)$*,* $p \geq 1$*. Furthermore, assume condition (**A3**) is satisfied. Define the set of functions:*

$$\mathcal{F}_n = \left\{ f : [0,1]^p \to \mathbb{R} \,|\, f(\boldsymbol{x}) = \sum_{j=1}^{p} \sum_{k=0}^{2^J - 1} c_{Jk}^{(j)} \phi_{Jk}^{per}(x_j)\,;\; J = J(n) \right\}, \tag{4.15}$$

where x_j*,* $j = 1, ..., p$ *corresponds to the* j*-th component of the vector* $\boldsymbol{x} \in [0,1]^p$*. Also, let* $\beta_n > 0$ *be a parameter depending on the sample and assume* $\mathbb{E}[Y^2] < \infty$*. Define* $\hat{f}_{J(n)}$ *as in (4.12) and let* $f_{J(n)} = T_{\beta_n} \hat{f}_{J(n)} := \hat{f}_{J(n)} \mathbb{1}_{\{|\hat{f}_{J(n)}| \leq \beta_n\}} + sign(\hat{f}_{J(n)}) \beta_n \mathbb{1}_{\{|\hat{f}_{J(n)}| > \beta_n\}}$*,* $\mathcal{K}_n = 2^{J(n)}$*. Assume the following conditions hold:*

(i) $\beta_n \to \infty$ *as* $n \to \infty$*.*

(ii) $\frac{\mathcal{K}_n \beta_n^4 \log(\beta_n)}{n} \to 0$ *as* $n \to \infty$*.*

(iii) For some $\delta > 0$ *as* $n \to \infty$ $\frac{n^{1-\delta}}{\beta_n^4} \to \infty$*.*

Then:

$$\lim_{n\to\infty} \int \left| f_{J(n)}(\boldsymbol{x}) - f_A(\boldsymbol{x}) \right|^2 \mu(d\boldsymbol{x}) = 0 \quad \textit{(a.s.)}\,, \tag{4.16}$$

$$\lim_{n\to\infty} \mathbb{E}\left\{ \int \left| f_{J(n)}(\boldsymbol{x}) - f_A(\boldsymbol{x}) \right|^2 \mu(d\boldsymbol{x}) \right\} = 0\,. \tag{4.17}$$

The corresponding proof can be found in section C.2 of the appendix.

Remarks

(i) Note that the scaling function $\phi(x)$ for the wavelet basis $\left\{\phi_{J,k}^{per}(x), 0 \leq k \leq 2^J - 1\right\}$ for fixed J is absolutely integrable in $\mathbb{R}$. Therefore, $\int_{\mathbb{R}} |\phi(x)|dx = C_\phi < \infty$.

Corollary 4.3.1. *Note that if $|Y| \leq B$, $B < \infty$ (known), to guarantee strong consistency of the least squares estimator it suffices to verify the following conditions are satisfied:*

(a) For some $\delta > 0$, $n^{1-\delta} \to \infty$, as $n \to \infty$.

(b) $\frac{\mathcal{K}_n}{n} \to 0$, as $n \to \infty$.

Remarks and comments

(i) This theorem is similar to theorem 10.3 of [9]. In our case, we investigated the statistical properties possible to be obtained using a wavelet framework, in the set of functions $\mathcal{F}_n$ defined by (C.10), and assuming conditions stated in 4.3.1 for the scaling function ϕ hold, when the unknown regression function is additive and given by $m(\mathbf{x}) = \sum_{j=1}^{p} m_j(x_j)$.

(ii) From this theorem it is possible to conclude that the estimator defined in (4.12) results from the application of the wavelet framework directly to the NESD generated

by the observations $\mathbf{X}_1, ..., \mathbf{X}_n$. As was shown, this approach provides good statistical properties which suggests that it is possible to ignore the NESD condition without compromising the robustness and efficiency of the estimator.

(iii) As was presented, the strong consistency of (4.12) relies on parameters β_n and $\mathcal{K}_n = 2^{J(n)}$ that need to be selected. In the next section, optimal choices for both are proposed.

4.3.3 Convergence rate of the Wavelet-based Least Squares Estimator.

As was seen in the previous section, Theorem 4.3.1 shows that the least squares (LS) wavelet-based estimator is strongly consistent for all bounded Lebesgue measures in $[0,1]^p$ when the set of assumptions for the unknown functions and wavelet basis are satisfied. In this section, we investigate the convergence rates that are possible to attain with this estimator. In particular, we are interested in studying the rate at which:

$$\mathbb{E}\left[\int_{[0,1]^p} \left|f_{J(n)}(\mathbf{x}) - f_A(\mathbf{x})\right|^2 \mu(d\mathbf{x})\right] \underset{n\to\infty}{\longrightarrow} 0\,,$$

where $f_{J(n)} = T_{\beta_n}\hat{f}_{J(n)}$ for $\beta_n > 0$ and $\hat{f}_{J(n)}$ defined as in (4.12). Similarly as in the previous section, to investigate the convergence properties of the LS estimator, we use theorem C.1.6, introduced by Pollard (1984), detailed in C.1 of the appendix.

Lemma 4.3.2. *Suppose an orthonormal basis $\left\{\phi^{per}_{J,k}(x), k = 0, ..., 2^J - 1\right\}$ for a certain $J \geq 0$ which as $J \to \infty$ is dense in $\mathbb{L}_2(\nu([0,1]))$ for $\nu \in \Upsilon$, where Υ represents the set of bounded Lebesgue measures in $[0,1]$. Suppose μ is a bounded Lebesgue measure in $[0,1]^p$ and conditions stated in Theorem 4.3.1 for the scaling function ϕ, and assumptions **(A1)-(A4)** defined in 4.2 are satisfied. Define the set of functions $\mathcal{F}_n$ as in (C.10). Also, let $\beta_n > 0$ be a parameter depending on the sample and assume $\mathbb{E}[Y^2] < \infty$. Define $\hat{f}_{J(n)}$ as in (4.12) and let $f_{J(n)} = T_{\beta_n}\hat{f}_{J(n)}$, let $\mathcal{K}_n = p\,2^{J(n)}$. Furthermore, assume the following condition holds:*

(i) $\sum_{j=1}^{p} ||f_j||_\infty < L$, *for some* $L < \beta_n$.

Then:

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\left|f_{J(n)}(\boldsymbol{x}_i) - f_A(\boldsymbol{x}_i)\right|^2 \mid \boldsymbol{X}_1^n\right] \leq \min_{f\in\mathcal{F}_n}\left\{||f - f_A||_n^2\right\} + \frac{\sigma^2}{n}\mathcal{K}_n\,, \tag{4.18}$$

where $||f||_n^2 = \frac{1}{n}\sum_{i=1}^{n} f(x_i)^2$.

The corresponding proof can be found in Appendix C.3.

Lemma 4.3.3. *Suppose an orthonormal basis* $\left\{\phi_{J,k}^{per}(x), k = 0, ..., 2^J - 1\right\}$ *that for* $J \to \infty$ *is dense in* $\mathbb{L}_2(\nu([0,1]))$ *for* $\nu \in \Upsilon$, *where* Υ *represents the set of bounded Lebesgue measures in* $[0,1]$. *Suppose assumptions stated in Theorem 4.3.1 for the scaling function* ϕ, *and conditions* ***(A1)-(A4)*** *defined in 4.2 hold. Let the set of functions* $\mathcal{F}_n$ *to be defined as in (C.10).*

Then it follows:

$$\inf_{f\in\mathcal{F}_n}\int_{[0,1]^p}|f(\boldsymbol{x}) - f_A(\boldsymbol{x})|^2\,\mu(d\boldsymbol{x}) \leq p^2\,C_2^2\,2^{-2(N+1)\,J(n)}\,, \tag{4.19}$$

for a constant $C_2 > 0$, *independent of* n, J.

The corresponding proof can be found in Appendix C.4.

Theorem 4.3.2. *Consider assumptions stated for Lemma 4.3.2 and conditions (i)-(iii) for Theorem 4.3.2 hold . Define* $\hat{f}_{J(n)}$ *as in (4.12) and let* $f_{J(n)} = T_{\beta_n}\hat{f}_{J(n)}$, *let* $\mathcal{K}_n = 2^{J(n)}$. *Then:*

$$\mathbb{E}\left[\int_{[0,1]^p}\left|f_{J(n)}(\boldsymbol{x}) - f_A(\boldsymbol{x})\right|^2\mu(d\boldsymbol{x})\right] \leq \tilde{C}\max\left\{\beta_n^2, \sigma^2\right\}\frac{p\,2^{J(n)}}{n}\left(\log(n)+1\right) + 8\,C_2^2\,p^2\,2^{-2(N+1)J(n)}\,, \tag{4.20}$$

for proper constants $\tilde{C} > 0$ *and* $C_2 > 0$ *independent of* n, N, p.

The corresponding proof is based on the application of Lemma 4.3.2, Lemma 4.3.3 and Theorem P2 and can be found in Appendix C.5.

4.3.4 Optimal choice of Estimator parameters $J(n)$ and β_n.

In this section we propose choices for the parameters $J(n)$ and β_n used in the estimator. First, we look at the selection of the truncating parameter β_n.

Lemma 4.3.4. *Suppose a model of the form (4.8), with $0 < \sigma < \infty$. Assume ϵ is a sub-gaussian random variable independent of $\boldsymbol{x}$, such that $\mathbb{E}[\epsilon] = 0$, $\mathbb{E}[\epsilon^2] = 1$. Let $\{Y_1, ..., Y_n\}$ be the response observations in the sample $\{Y_i, \boldsymbol{X}_i\}_{i=1}^n$.*

Then, for $\beta_n = 4\sigma\sqrt{\log(n)}$ it follows:

$$\mathbb{P}\left\{\max\left\{Y_1, ..., Y_n\right\} > \beta_n\right\} = \mathcal{O}\left(\frac{1}{n}\right), \tag{4.21}$$

which implies that $\lim_{n\to\infty} \mathbb{P}\left\{\max\left\{Y_1, ..., Y_n\right\} > \beta_n\right\} \to 0$ at a rate $\frac{1}{n}$.

The corresponding proof can be found in Appendix C.6.

Remarks

(i) In practice, the value of σ is not known and it can be estimated by the sample variance $\hat{\sigma^2}$ of the response. Assuming independence between the random error ϵ and predictors $\mathbf{X}$, this is a suitable choice. However, this in practice could lead to a larger than optimal truncating parameter, since $Var(f(\mathbf{x})) \geq \sigma^2$.

(ii) Another possibility for choosing σ could be the one proposed by Donoho and Johnstone (1994), which is given by $\hat{\sigma} = \frac{\text{median}\left(\left\{\left|\hat{d}_{J-1,k}\right| : k=0,...,2^J-1\right\}\right)}{0.6745}$, where $\hat{d}_{J-1,k}$ are the discrete wavelet coefficients resulting from the DWT of the observed response $\mathbf{y}$.

Lemma 4.3.5. *Define $\hat{f}_{J(n)}$ as in (4.12) and let $f_{J(n)} = T_{\beta_n}\hat{f}_{J(n)}$. Suppose assumptions for Theorem 4.3.2 hold. Then, for $\beta_n = 4\sigma\sqrt{\log(n)}$ ($n \geq 2$), setting the multiresolution level $J(n)$ as:*

$$J^*(n) = \mathcal{K}_1 + \frac{1}{2\,N+3}\log_2\left(\frac{n}{\log(n)\,(\log(n)+1)}\right), \tag{4.22}$$

minimizes the $\mathbb{L}_2$-risk upper bound given by (4.20) and guarantees the strong consistency of the estimator $\hat{f}_{J(n)}$, where $\mathcal{K}_1 = \frac{1}{2\,N+3}\log_2\left(\frac{(N+1)\,C_2^2\,p}{\tilde{C}\,\sigma^2}\right)$.

The proof of this Lemma consists in the minimization of the upper bound (4.20) with respect $\tilde{\mathcal{K}}_n = 2^{J(n)}$. Note that the minimum exists and is unique due to the convexity of the objective function defined by (4.20). Similarly, it is possible to guarantee conditions (i)-(iii) of Theorem 4.3.1 are satisfied since:

$$\lim_{n\to\infty}\left(\frac{\log(n)^{\gamma+t}}{n^{\gamma}}\right) = 0\,,$$

$\forall\,\gamma \geq 1\,, t > 0$ (integers) which can be proved by applying L'Hopital's rule.

Theorem 4.3.3. *Suppose assumptions and results for Theorems 4.3.1, 4.3.2 and Lemmas 4.3.4 and 4.3.5 hold. Then, the estimator defined by in (4.12), and $f_{J(n)} = T_{\beta_n}\hat{f}_{J(n)}$ attains the following convergence rate for the $\mathbb{L}_2$-risk:*

$$\mathbb{E}\left[\int_{[0,1]^p}\left|f_{J(n)}(\boldsymbol{x}) - f_A(\boldsymbol{x})\right|^2\mu(d\boldsymbol{x})\right] \leq \tilde{K}\,\left(\frac{\beta_n^2\,\log(n)}{n}\right)^{\frac{2\gamma}{2\gamma+1}}, \tag{4.23}$$

where $\gamma = N+1$, $\tilde{K} = \left(2\,\gamma\,\tilde{C}\,p\right)^{\frac{2\gamma}{\gamma+1}}\,(8\,C_2^2\,p^2)^{\frac{1}{2\gamma+1}}$.

From (4.23), it is possible to distinguish 2 cases:

(i) From Corollary 4.3.1, if $|Y| \leq B$, $B < \infty$ (known) it follows:

$$\mathbb{E}\left[\left\|f_{J(n)} - f_A\right\|^2\right] = \mathcal{O}\left(\frac{\log(n)}{n}\right)^{\frac{2\gamma}{2\gamma+1}} . \tag{4.24}$$

(ii) If the upper bound of Y is not known, choosing β_n as in Lemma 4.3.4, the convergence rate takes the form of:

$$\mathbb{E}\left[\left\|f_{J(n)} - f_A\right\|^2\right] = \mathcal{O}\left(\frac{\log(n)^2}{n}\right)^{\frac{2\gamma}{2\gamma+1}} . \tag{4.25}$$

The proof of the above assertions follows from Lemmas 4.3.4 and 4.3.5 applied to Theorem 4.3.2.

Remarks

(i) Note that results (i) and (ii) show that the LS estimator defined by $\hat{f}_{J(n)}$ as in (4.12) does not suffer from the curse of dimensionality. Moreover, its convergence rate is optimal up to a logarithmic factor. This implies that is possible to apply the wavelet framework directly over non-equally spaced designs without compromising desirable statistical properties such as strong consistency and optimal $\mathbb{L}_2$ convergence rates.

4.3.5 Simulation Study

In the last section, we introduced a wavelet based least squares estimator for the additive regression model and proved its statistical properties. In this section, we investigate the performance of $\hat{f}_n(\mathbf{x})$ with respect to the ARMSE (Average Root Mean Squared Error) of estimation, via a simulation study. For this objective, we choose a set of exemplary baseline functions that combine different smoothness and spectral properties and are aimed to challenge the estimation process.

To simplify the implementation, we select specific functions that are supported in the [0,1] and also satisfy assumptions (**A1**)-(**A4**). These functions are defined as follows:

$$
\begin{aligned}
f_1(x) &= \frac{1}{\sqrt{2}\sin(2\pi\, x)} \\
f_2(x) &= 1 - 4\,|x - \frac{1}{2}| \\
f_3(x) &= -\cos(4\pi\, x + 1) \\
f_4(x) &= 8\left(x - \frac{1}{2}\right)^2 - \frac{2}{3} \\
f_5(x) &= \frac{1}{\sqrt{2}}\cos(2\pi\, x) \\
f_6(x) &= \frac{1}{\sqrt{2}}\cos(4\pi\, x) \\
f_7(x) &= -0.5275 + 4\, e^{-500(x-0.23)^2} + 2\, e^{-2000(x-0.33)^2} \\
&\quad +4\, e^{-8000(x-0.47)^2} + 3\, e^{-16000(x-0.69)^2} + e^{-32000(x-0.83)^2} \\
f_8(x) &= 0.2\,\cos(4\pi\, x + 1) + 0.1\,\cos(24\pi\, x + 1) \\
f_9(x) &= -0.1744 + 2\, x^3\, \mathbf{1}_{(0.5<x\leq 0.8)} + 2\,(x-1)^3\, \mathbf{1}_{(0.8<x\leq 1)}
\end{aligned}
$$

In this simulation study, we investigate the performance of the estimator for different sample sizes, noise variances σ^2, wavelet filters and distribution of the predictors $\mathbf{X}$. To quantify the estimator performance, we use the following global error measure:

$$
\widehat{ARMSE} = \sqrt{(\frac{1}{B}\sum_{b=1}^{B}\frac{1}{n}\sum_{i=1}^{n}\left(f(\mathbf{x}_i) - \hat{f}_{n,b}(\mathbf{x}_i)\right)^2)}\,, \tag{4.26}
$$

where B is the number of replications of the experiment and n is the number of samples. For all experiments we choose $B = 200$.

While implementing the simulations, we considered the following settings in a matlab-based

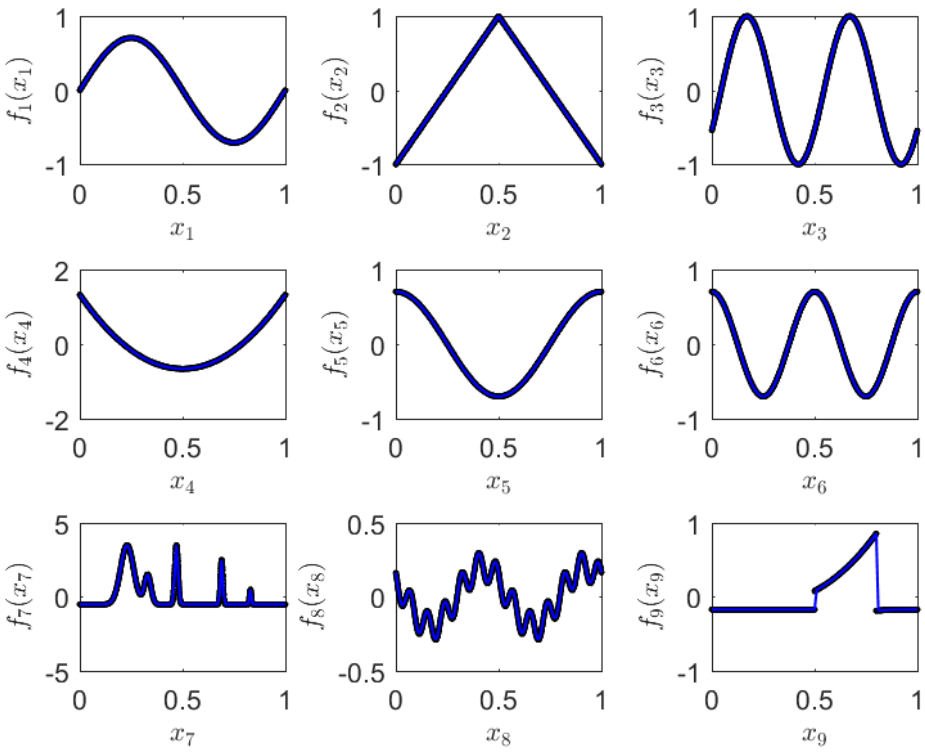


Figure 4.1: Graphic representation of the testing functions for the Additive Model.

script:

(i) We generated independent random numbers $\{\mathbf{X}_i\}_{i=1}^{N}$ from the $\{\mathcal{U}[0,1]\}^9$ and $\left\{Beta(\frac{3}{2},\frac{3}{2})\right\}^9$ joint distributions (satisfying assumptions (**A1**)-(**A4**)), and constructed the model defined in (4.8).

(ii) For the noise variance, we used $\sigma^2 = 0.75$ and $\sigma^2 = 0.25$, which produced different signal-to-noise ratios (SNR) used to assess the estimator robustness against noisy observations.

(iii) For the computation of the least squares estimator, we chose the scaling functions generated by the wavelet filters `Coiflets` and `Daubechies` with 24 and 4 coefficients respectively.

(iv) Both of the chosen wavelet filters satisfy conditions 1-6 listed in theorem 1. For Coiflets, the wavelet is near symmetric with compact support and has $N/3$ vanishing moments (N is the number of filter taps); in the case of Daubechies, the wavelet does not have the near-symmetry property but it has compact support and N vanishing moments.

(v) For the evaluation of the scaling functions ϕ_{Jk}^{per} (and construction of matrix B) we used Daubechies-Lagarias's algorithm.

(vi) The multiresolution level J was chosen to be $J(n) = 1+\lfloor \log_2(n) - \log_2(\log(n)(\log(n)+1)) \rfloor$.

(vii) The truncating parameter β_n was selected using the proposition detailed in remark (ii) of Lemma 4.3.4.

Simulation Results.

In this section, we summarize the simulation results obtained for the baseline distributions previously defined. In particular, we present the following:

(i) Tables 4.1 to 4.4 present details for ARMSE results obtained for each of the baseline distributions using a Uniform design $\{\mathcal{U}[0,1]\}^9$ for predictors. Similarly, in Tables 4.5 to 4.8 present details for RMSE results obtained for each of the baseline functions using a $\left\{Beta(\frac{3}{2},\frac{3}{2})\right\}^9$ design.

(ii) Figures 4.2a - 4.2b show the behavior of the RMSE for each of the functions $f_1, ..., f_9$ with respect to sample size and noise variance values $\sigma^2 = 0.75\,,\ 0.25$, for the Uniform design $\{\mathcal{U}[0,1]\}^9$ using Daubechies filter.

(iii) Figures 4.4a - 4.4b show the behavior of the RMSE for each of the functions $f_1, ..., f_9$ with respect to the sample size and the noise variance values $\sigma^2 = 0.75\,,\ 0.25$ for the Uniform design $\{\mathcal{U}[0,1]\}^9$ using Coiflets 24 filter.

(iv) Figures 4.5a - 4.13b show the recovered functions $f_1, ..., f_9$ for different sample sizes $n = 512, 1024, 4096$ and values of the noise variance $\sigma^2 = 0.25, 0.3$ for the Uniform design $\{\mathcal{U}[0,1]\}^9$ using a Coiflets 24 filter. The dashed lines (black) correspond to the actual function, computed at each data point x, whereas the magenta points show the estimated values of the function at each sample x. The red lines corresponds to a smoothed version of the estimated function values, computed using locally weighted scatterplot smoothing (lowess) with parameter 0.25 (this was done just for visualization purposes).

(v) Figures 4.14a - 4.14b show the behavior of the RMSE for each of the functions $f_1, ..., f_9$ with respect to the sample size and the noise variance values $\sigma^2 = 0.75\,,\ 0.25$ for the Beta design $\left\{Beta(\frac{3}{2}, \frac{3}{2})\right\}^9$ using Daubechies filter.

(vi) Figures 4.15a - 4.15b show the behavior of the RMSE for each of the functions $f_1, ..., f_9$ with respect to the sample size and the noise variance values $\sigma^2 = 0.75\,,\ 0.25$ for the Beta design $\left\{Beta(\frac{3}{2}, \frac{3}{2})\right\}^9$ using Coiflets 24 filter. In each figure, plots (b) and (d) correspond to zoomed in versions of plots (a) and (c) respectively.

(vii) Figures 4.16a - 4.24b show the recovered functions $f_1, ..., f_9$ for different sample sizes $n = 1024, 4096$ and values of the noise variance $\sigma^2 = 0.3$ for the Beta design $\left\{Beta(\frac{3}{2}, \frac{3}{2})\right\}^9$ using Coiflets 24 filter. The dashed lines (black) correspond to the actual function, computed at each data point x, whereas the magenta points show the estimated values of the function at each sample x. The red lines corresponds to a smoothed version of the estimated function values, computed using lowess smoothing with parameter 0.25 (this was done just for visualization purposes).

Table 4.1: RMSE results for Uniform distribution with $\sigma^2 = 0.25$ using Daubechies 4 wavelet filter.

	$n = 256$	$n = 512$	$n = 1024$	$n = 2048$	$n = 4096$
$f_1(x)$	0.0224	0.0143	0.0086	0.0035	0.002
$f_2(x)$	0.0227	0.0156	0.0089	0.0038	0.002
$f_3(x)$	0.0692	0.0174	0.0088	0.0038	0.002
$f_4(x)$	0.0241	0.0141	0.0086	0.0038	0.002
$f_5(x)$	0.0242	0.0148	0.0088	0.0036	0.002
$f_6(x)$	0.0391	0.0155	0.0087	0.0037	0.0021
$f_7(x)$	0.7327	0.1069	0.1051	0.1005	0.0533
$f_8(x)$	0.0289	0.0191	0.0103	0.0049	0.0021
$f_9(x)$	0.0543	0.0268	0.0143	0.0091	0.0029

Table 4.2: RMSE results for Uniform distribution with $\sigma^2 = 0.75$ using Daubechies 4 wavelet filter.

	$n = 256$	$n = 512$	$n = 1024$	$n = 2048$	$n = 4096$
$f_1(x)$	0.042	0.0362	0.0306	0.0126	0.0114
$f_2(x)$	0.0458	0.0345	0.0307	0.0121	0.0108
$f_3(x)$	0.0909	0.0382	0.0301	0.013	0.0109
$f_4(x)$	0.044	0.0342	0.0296	0.0127	0.0113
$f_5(x)$	0.0449	0.0341	0.0304	0.0125	0.0111
$f_6(x)$	0.064	0.0363	0.0305	0.0128	0.0113
$f_7(x)$	0.7577	0.1299	0.1283	0.1097	0.0624
$f_8(x)$	0.0478	0.0395	0.0322	0.0135	0.011
$f_9(x)$	0.0751	0.0468	0.0349	0.0177	0.0119

Results Discussion

As can be observed from the Figures and Tables illustrating the simulation results, the least squares methodology is able to provide accurate estimates for the simulated settings. In particular, for the smooth functions the estimates exhibit a small bias and a variance that decreases with the sample size as was theoretically shown.

For functions that are smooth and posses oscillations that are concentrated in small subintervals of the support (e.g. f_7 and f_9), when sample sizes are small the least squares estimates are likely to fail to detect the multimodality of the functions. However, as the sample size increases, the estimates are very accurate.

As seen from the illustrations of the behavior of the ARMSE with respect to the sample size,

Table 4.3: RMSE results for Uniform distribution with $\sigma^2 = 0.25$ using Coiflets 24 wavelet filter.

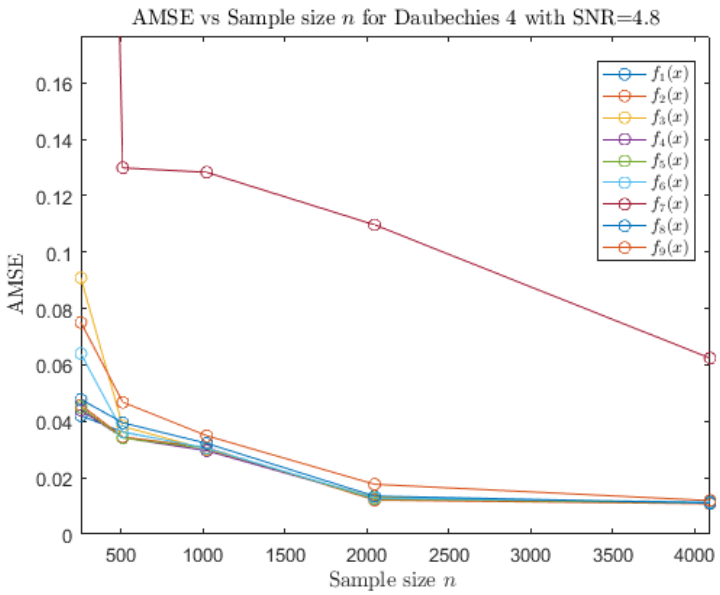
	$n = 256$	$n = 512$	$n = 1024$	$n = 2048$	$n = 4096$
$f_1(x)$	0.0193	0.0163	0.0058	0.0024	0.0013
$f_2(x)$	0.0191	0.0172	0.0057	0.0025	0.0013
$f_3(x)$	0.0198	0.0168	0.006	0.0025	0.0013
$f_4(x)$	0.0214	0.0177	0.0061	0.0025	0.0013
$f_5(x)$	0.0185	0.0165	0.0059	0.0024	0.0013
$f_6(x)$	0.0207	0.0177	0.0057	0.0025	0.0013
$f_7(x)$	0.7776	0.1946	0.0388	0.0353	0.0088
$f_8(x)$	0.0244	0.0222	0.0061	0.0027	0.0013
$f_9(x)$	0.0386	0.022	0.0083	0.0049	0.0032

Table 4.4: RMSE results for Uniform distribution with $\sigma^2 = 0.75$ using Coiflets 24 wavelet filter.

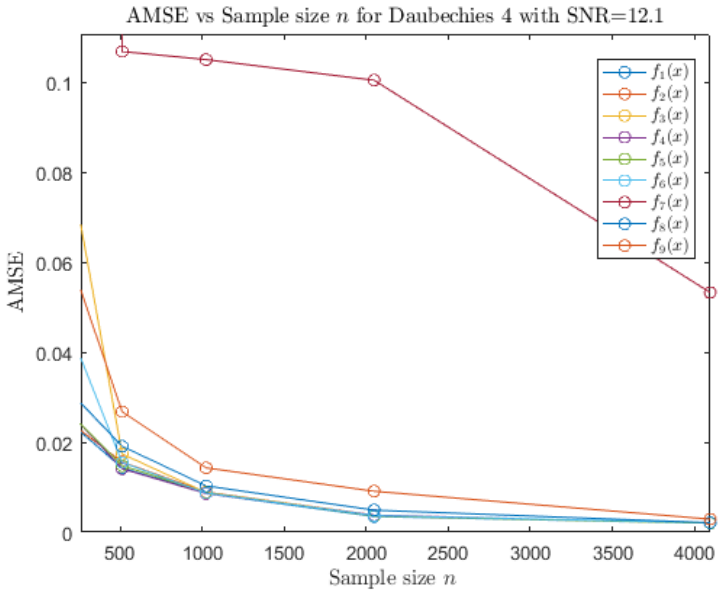
	$n = 256$	$n = 512$	$n = 1024$	$n = 2048$	$n = 4096$
$f_1(x)$	0.0369	0.0375	0.0259	0.0115	0.0102
$f_2(x)$	0.0407	0.0364	0.0268	0.0112	0.0103
$f_3(x)$	0.0377	0.0373	0.0269	0.0116	0.0102
$f_4(x)$	0.0417	0.0353	0.0266	0.0115	0.0104
$f_5(x)$	0.0395	0.0373	0.0265	0.0112	0.0101
$f_6(x)$	0.0397	0.0368	0.0268	0.0113	0.0105
$f_7(x)$	0.7796	0.2165	0.0598	0.0438	0.0178
$f_8(x)$	0.0438	0.0436	0.0273	0.0115	0.0103
$f_9(x)$	0.0571	0.0433	0.0289	0.0132	0.0121

for most of functions the least squares estimates behave similarly. Also, it is interesting to note the effect of the SNR on the estimation accuracy: although it was observed a decrease in the perfomance for small SNR, in general the estimates remain within a reasonable range of accuracy for practical applications. These two facts support the argument that the least squares estimator is robust enough for sufficiently smooth functions and a good range of SNR.

In the case of f_9 that corresponds to a non-smooth functions, even though the estimation performance was not as good as for the rest of the functions, it did a reasonably good job in detecting the average functional form with the simulated sample sizes. However, increasing the sample size would definitely lead to more accurate estimates (bias + variance), as suggested by the theoretical results.



(a) Daubechies filter, $\sigma^2 = 0.25$

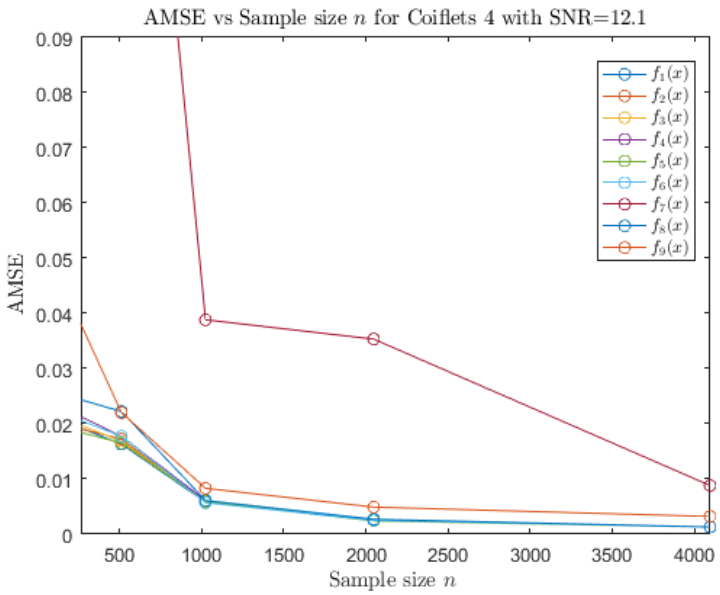


(b) Daubechies filter, $\sigma^2 = 0.75$

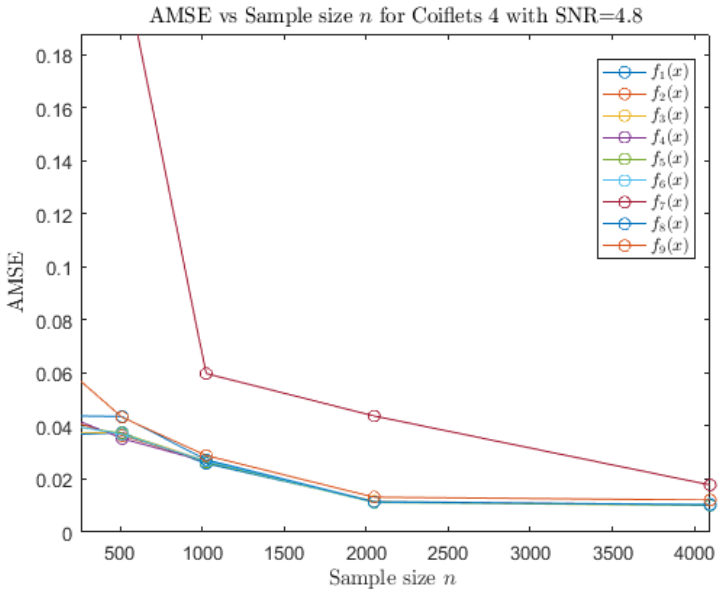
Figure 4.2: RMSE results for Uniform Design using Daubechies and Coiflets filter, for values of $\sigma^2 = 0.25 - 0.75$.

Remarks and comments

(i) Practical choice of $J(n)$. Since the optimal multiresolution index J was obtained up to and unknown additive constant $\mathcal{K}_1$ (see Lemma 4.3.5), for implementation purposes



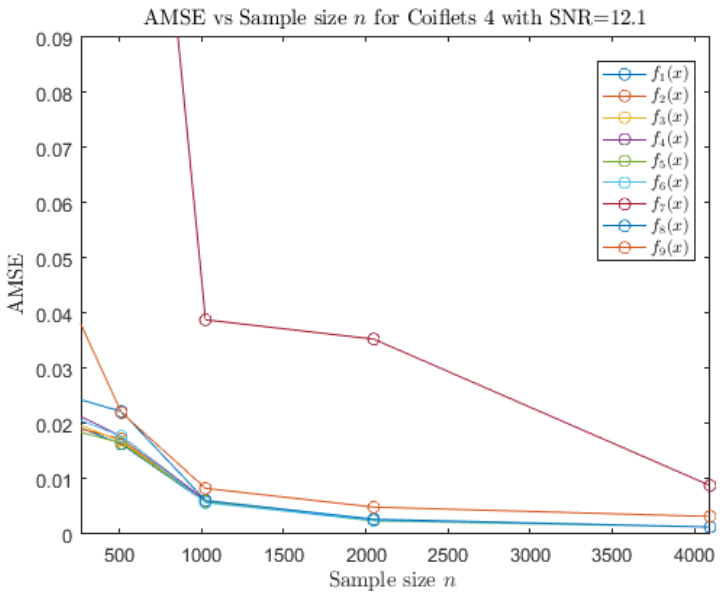
(a) Coiflets filter, $\sigma^2 = 0.25$



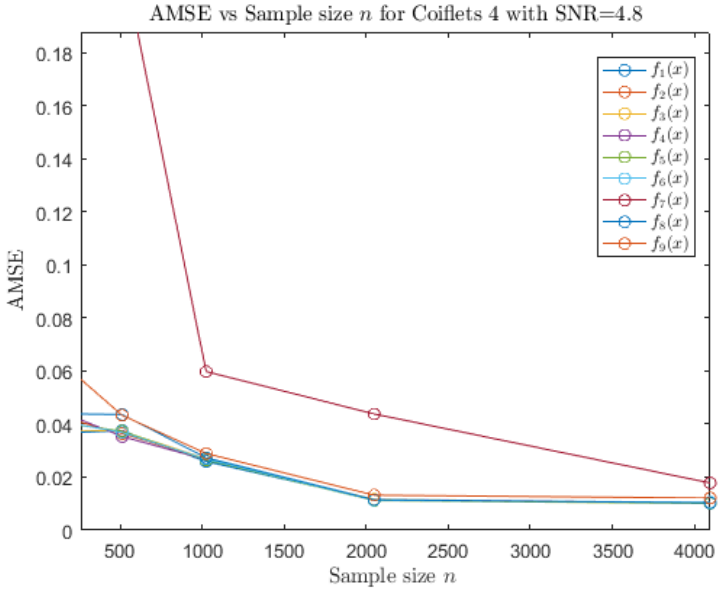
(b) Coiflets filter, $\sigma^2 = 0.75$

Figure 4.3: RMSE results for Uniform Design using Daubechies and Coiflets filter, for values of $\sigma^2 = 0.25 - 0.75$.

it is possible to replace it with a predefined integer. However, a large value for this constant would cause an undesired inflation of the estimator variance and also, increase the computational complexity of the algorithm.



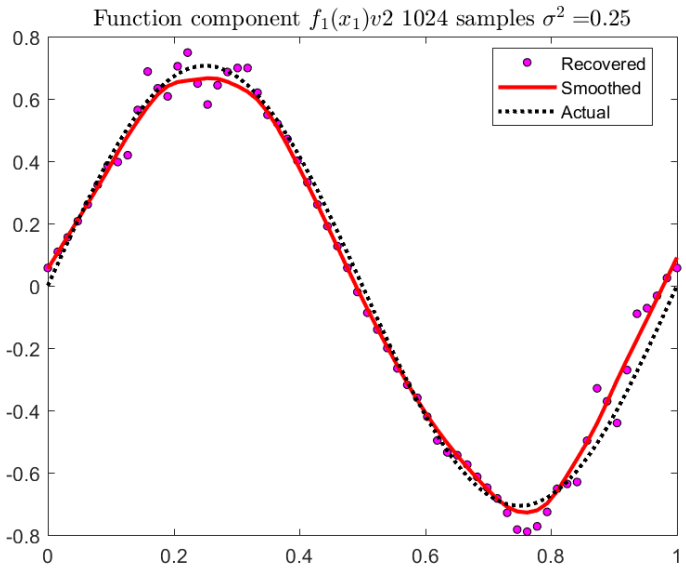
(a)



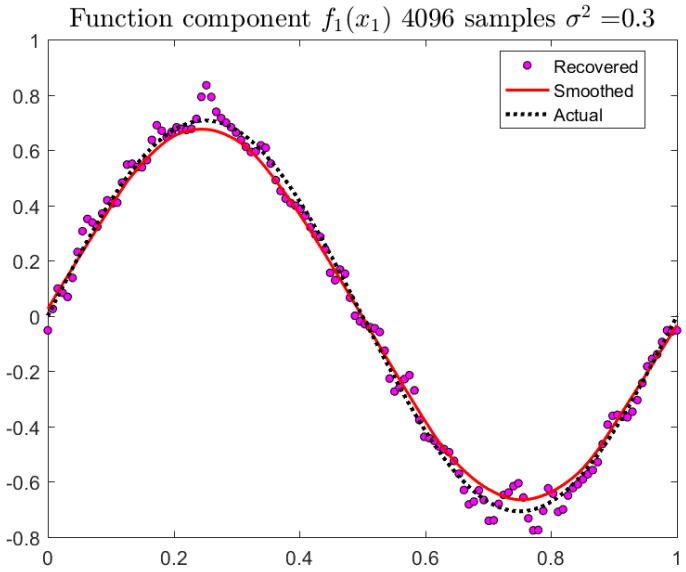
(b)

Figure 4.4: RMSE results for each function using Uniform Design and Coiflets 24 filter, for values of $\sigma^2 = 0.25\,,\ 0.75$.

(ii) In the case of densities with exponentially decaying tails (i.e. largely deviated form uniformity), large samples are needed in order to obtain accurate estimates. In fact, during the simulation study we observed cases where abnormally large wavelet coeffi-



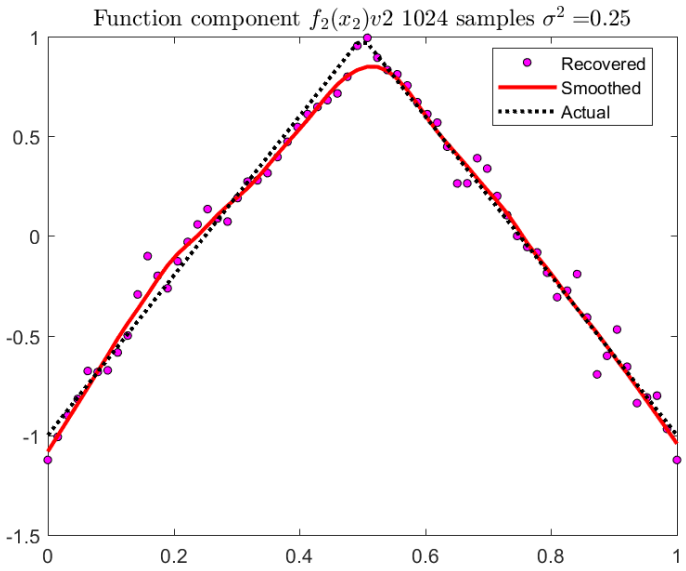
(a)



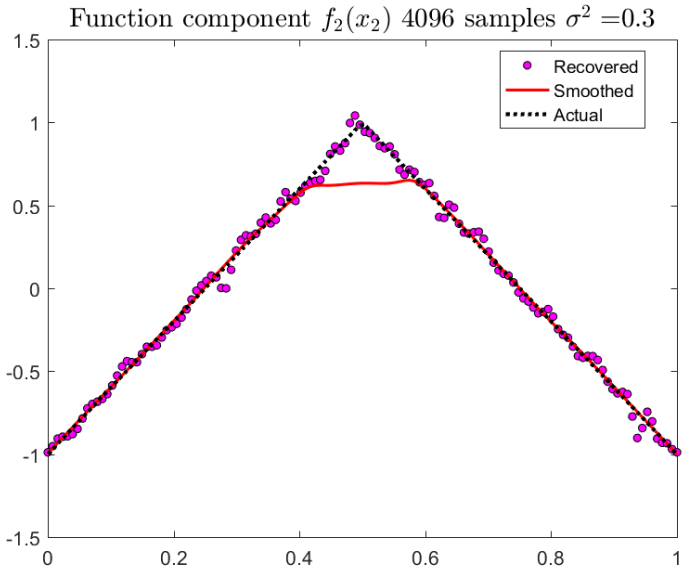
(b)

Figure 4.5: Estimated $f_1(x)$ using Uniform Design and Coiflets filter.

cients were obtained at the tails of the distribution (or regions with low density values). This was caused primarily due to possible violations of assumption (**A1**) and the lack of information available for a reasonable estimation of the coefficients in those regions.



(a)

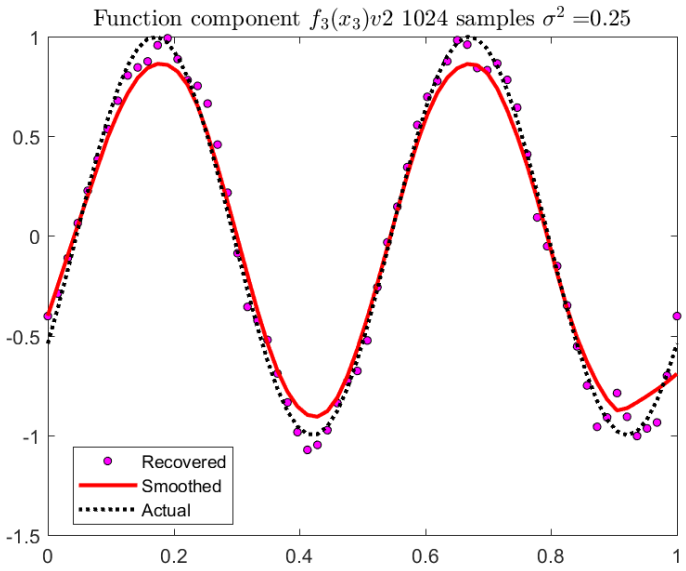


(b)

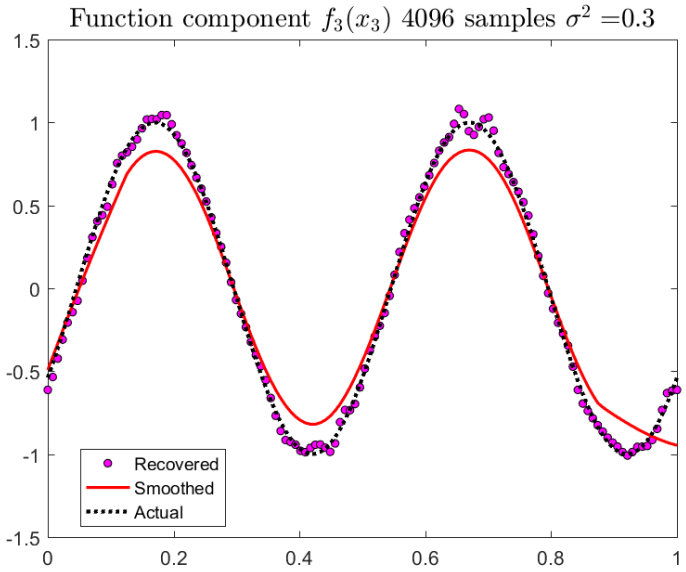
Figure 4.6: Estimated $f_2(x)$ using Uniform Design and Coiflets filter.

In this context, we suggest the following possible remedial actions:

(a) Restricting the domain of estimation to the 95% empirical quantiles along each of the dimensions of the predictors. This is a reasonable approach that can pre-



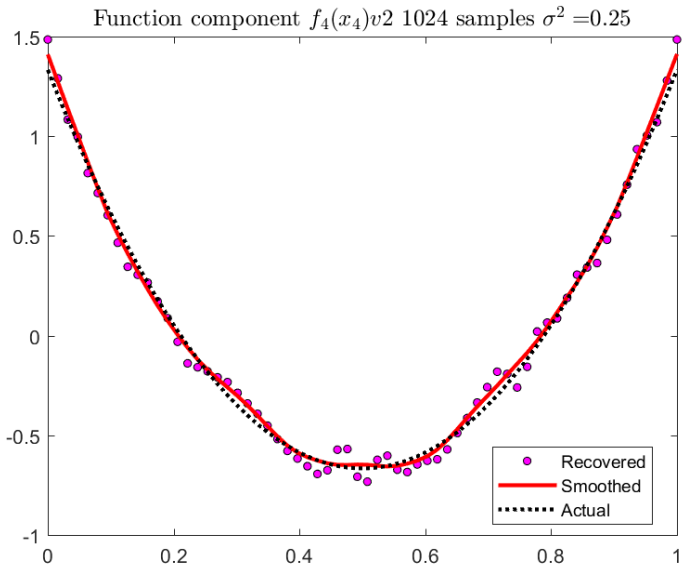
(a)



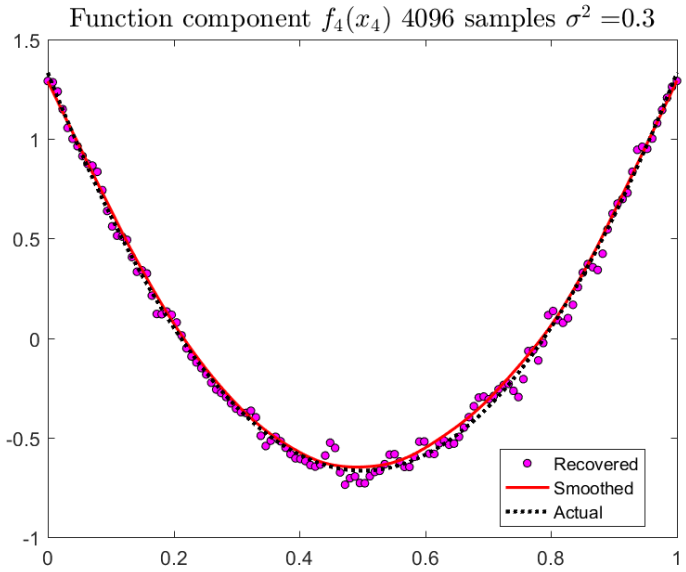
(b)

Figure 4.7: Estimated $f_3(x)$ using Uniform Design and Coiflets filter.

vent the generation of large coefficient that induce error in the function estimation procedure. However, this reduces the effective sample size and also, restricts the possibility of estimation of unlikely or rare cases.



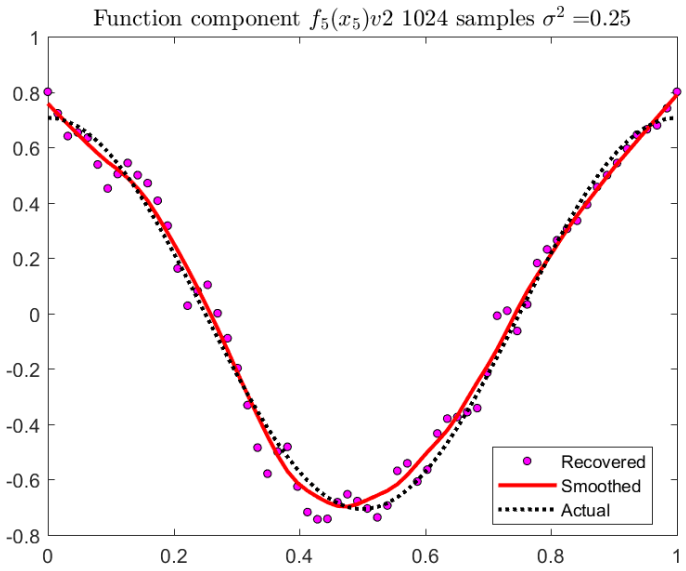
(a)



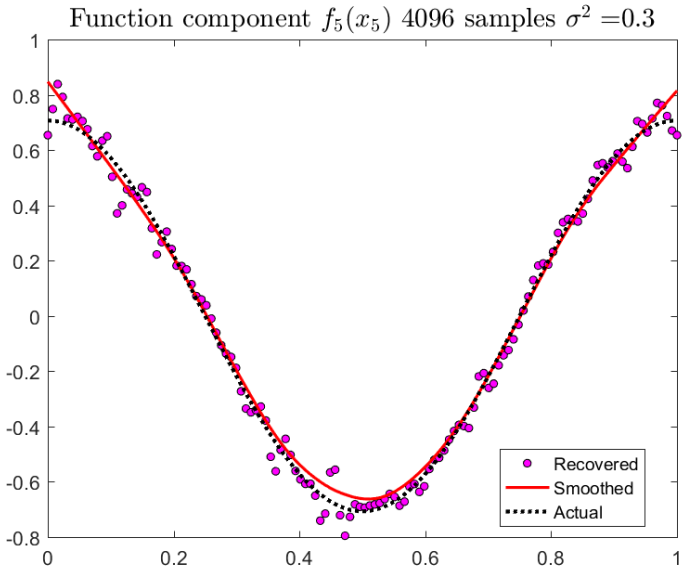
(b)

Figure 4.8: Estimated $f_4(x)$ using Uniform Design and Coiflets filter.

(b) Choosing parameter β_n via cross-validation to minimize the RMSE. Abnormally large wavelet coefficients would lead (in general) to large function estimates. This can be prevented by truncating the final estimates using β_n and the use of cross-



(a)

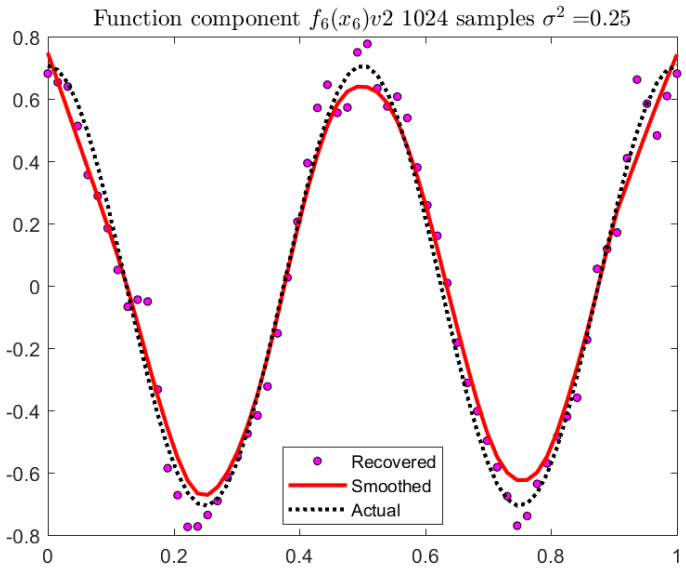


(b)

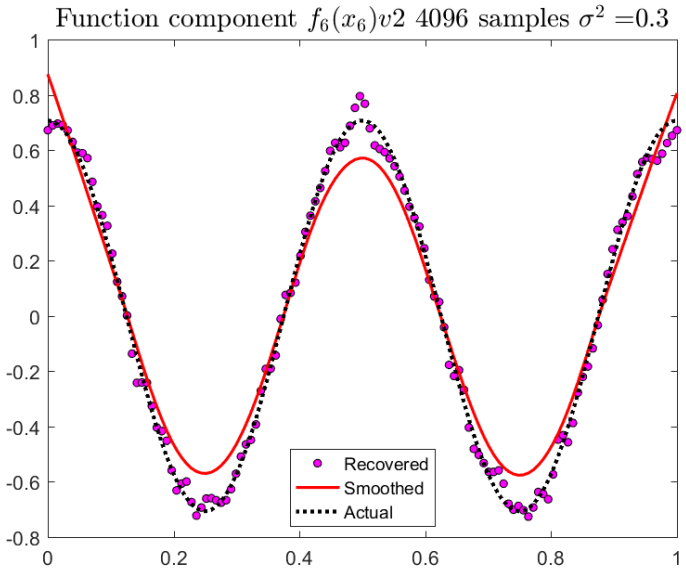
Figure 4.9: Estimated $f_5(x)$ using Uniform Design and Coiflets filter.

validation would allow an evidence-based selection of this parameter.

(iii) Model without β_0. Because of the strang-fix condition, the estimation of a model with a constant β_0 turned out to be unstable. For this reason, we recommend a pre-processing



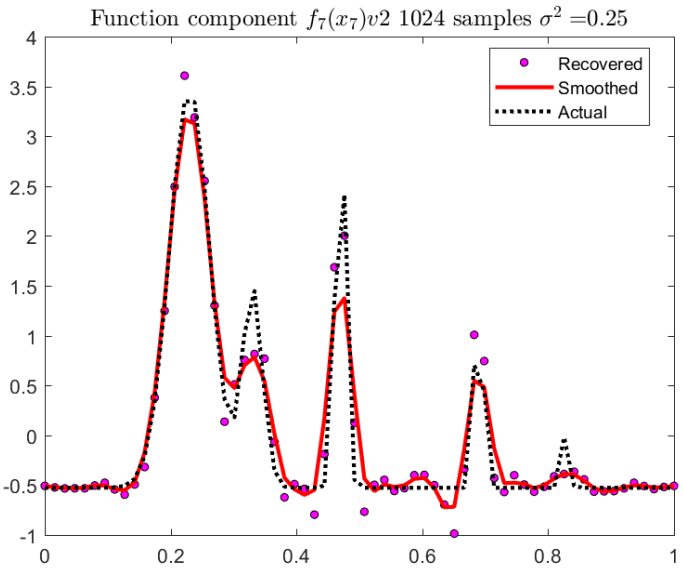
(a)



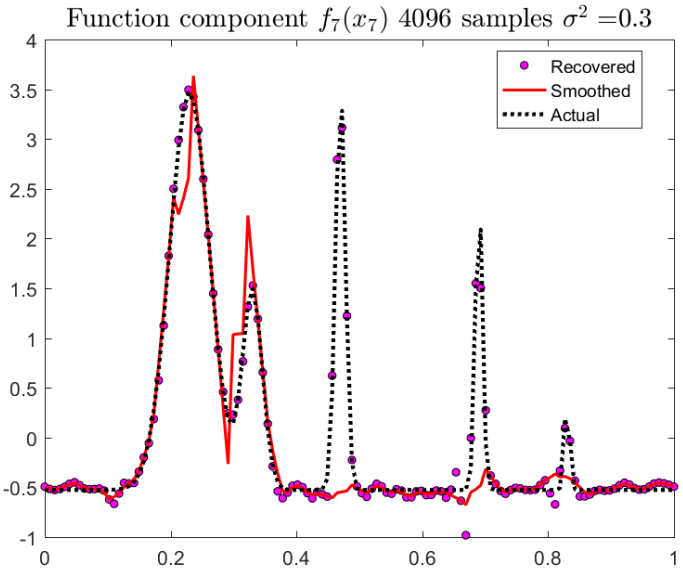
(b)

Figure 4.10: Estimated $f_6(x)$ using Uniform Design and Coiflets filter.

stage in which the response is standardized so that it has zero mean and a standard deviation of 1. This approach is a natural result if we modify assumption (**A1**) to be instead $\mathbb{E}\left[f_j(X_j)\right] = 0$ for $j = 1, ..., p$. Note that this does not alter at all the model



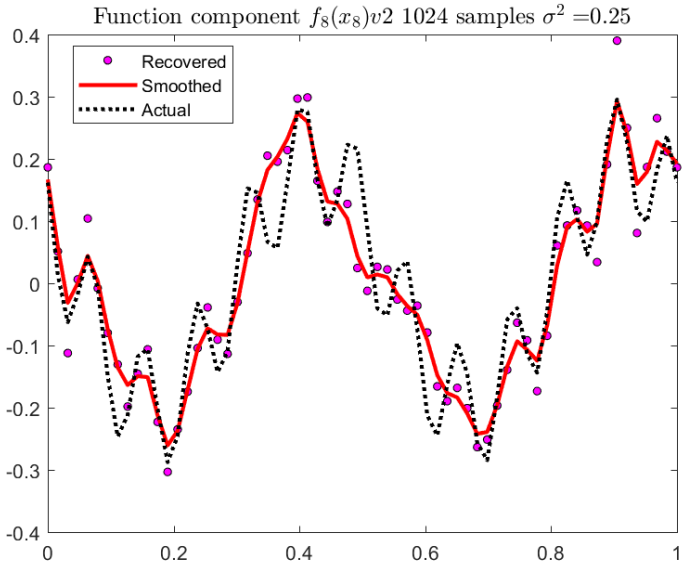
(a)



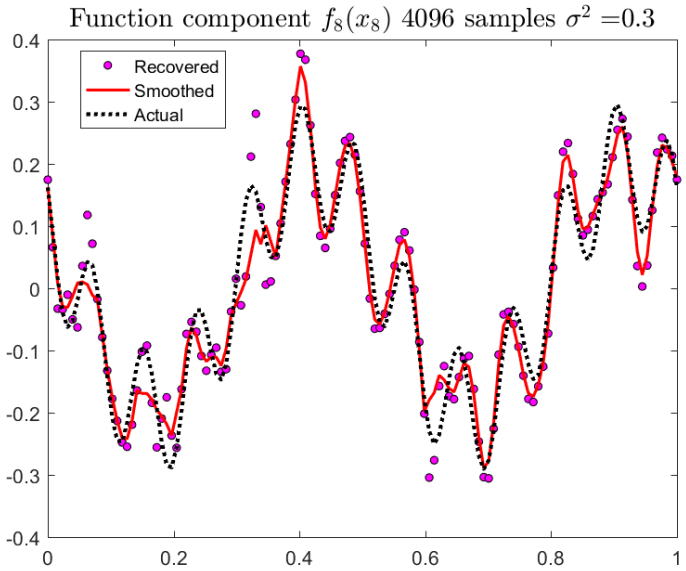
(b)

Figure 4.11: Estimated $f_7(x)$ using Uniform Design and Coiflets filter.

structure, estimation procedure or statistical properties. In this case the natural estimator of the intercept would be given by $\hat{\beta}_0 = \frac{1}{n}\sum_{i=1}^{n} y_i$.

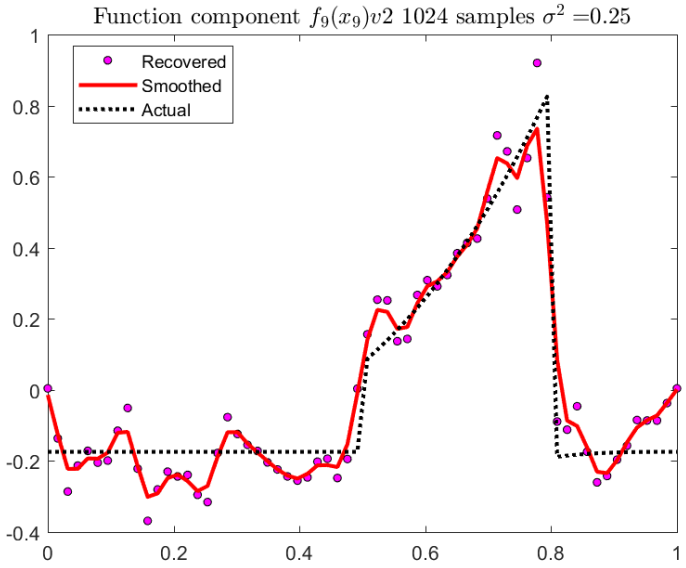


(a)

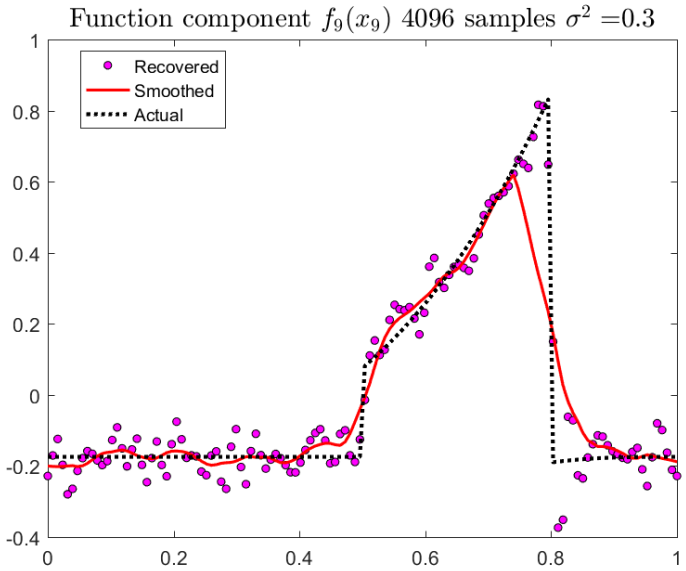


(b)

Figure 4.12: Estimated $f_8(x)$ using Uniform Design and Coiflets filter.



(a)



(b)

Figure 4.13: Estimated $f_9(x)$ using Uniform Design and Coiflets filter.

Table 4.5: RMSE results for $Beta(\frac{3}{2}, \frac{3}{2})$ distribution with $\sigma^2 = 0.25$ using Daubechies 4 wavelet filter.

	$n = 256$	$n = 512$	$n = 1024$	$n = 2048$	$n = 4096$
$f_1(x)$	0.0324	0.0246	0.0153	0.0058	0.0031
$f_2(x)$	0.0344	0.0212	0.0147	0.0057	0.003
$f_3(x)$	0.0971	0.026	0.0141	0.006	0.0031
$f_4(x)$	0.0325	0.0234	0.0143	0.0054	0.003
$f_5(x)$	0.0369	0.0237	0.0143	0.0054	0.0032
$f_6(x)$	0.0561	0.0248	0.0137	0.0061	0.003
$f_7(x)$	0.7254	0.1071	0.1072	0.101	0.0538
$f_8(x)$	0.0413	0.0273	0.0148	0.0071	0.0033
$f_9(x)$	0.067	0.0341	0.0194	0.0112	0.004

Table 4.6: RMSE results for $Beta(\frac{3}{2}, \frac{3}{2})$ distribution with $\sigma^2 = 0.75$ using Daubechies 4 wavelet filter.

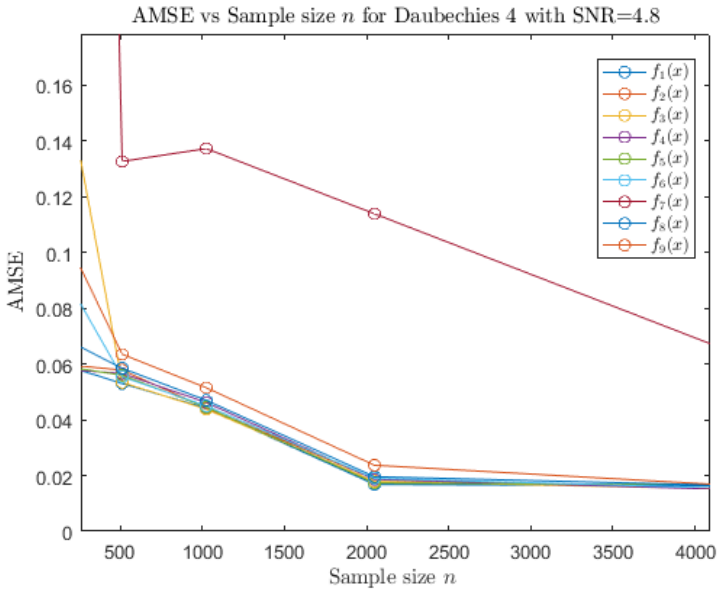
	$n = 256$	$n = 512$	$n = 1024$	$n = 2048$	$n = 4096$
$f_1(x)$	0.0578	0.053	0.0442	0.0168	0.0163
$f_2(x)$	0.0593	0.0578	0.0443	0.0187	0.0156
$f_3(x)$	0.1342	0.0534	0.0438	0.0179	0.0153
$f_4(x)$	0.0577	0.0566	0.0462	0.0186	0.0152
$f_5(x)$	0.0583	0.056	0.0445	0.0173	0.0167
$f_6(x)$	0.0819	0.0554	0.045	0.019	0.0156
$f_7(x)$	0.7534	0.1327	0.1373	0.1139	0.0672
$f_8(x)$	0.0662	0.0585	0.0470	0.0196	0.0166
$f_9(x)$	0.0949	0.0635	0.0515	0.0237	0.0169

Table 4.7: RMSE results for $Beta(\frac{3}{2}, \frac{3}{2})$ distribution with $\sigma^2 = 0.25$ using Coiflets 24 wavelet filter.

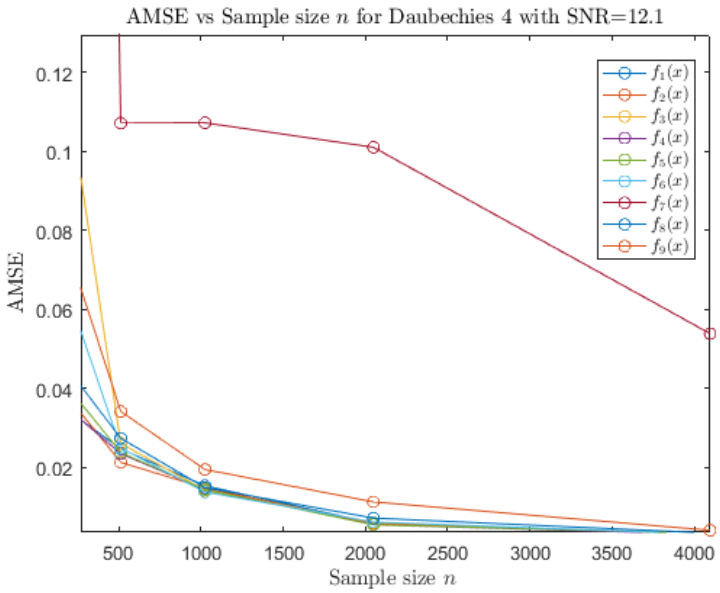
	$n = 256$	$n = 512$	$n = 1024$	$n = 2048$	$n = 4096$
$f_1(x)$	0.0284	0.0252	0.0091	0.0035	0.0017
$f_2(x)$	0.029	0.0258	0.0086	0.0036	0.0017
$f_3(x)$	0.0276	0.0248	0.0085	0.0034	0.0018
$f_4(x)$	0.0312	0.0246	0.0084	0.0036	0.0018
$f_5(x)$	0.0288	0.0246	0.0084	0.0036	0.0017
$f_6(x)$	0.0293	0.0245	0.0088	0.0034	0.0017
$f_7(x)$	0.757	0.1977	0.0398	0.0358	0.0091
$f_8(x)$	0.0347	0.0321	0.0081	0.0038	0.0017
$f_9(x)$	0.047	0.0313	0.011	0.0059	0.0035

Table 4.8: RMSE results for $Beta(\frac{3}{2}, \frac{3}{2})$ distribution with $\sigma^2 = 0.75$ using Coiflets 24 wavelet filter.

	$n = 256$	$n = 512$	$n = 1024$	$n = 2048$	$n = 4096$
$f_1(x)$	0.0488	0.0509	0.0346	0.0142	0.013
$f_2(x)$	0.0523	0.0511	0.0347	0.0144	0.0131
$f_3(x)$	0.0492	0.0467	0.0356	0.0149	0.0134
$f_4(x)$	0.0548	0.0493	0.037	0.0145	0.0133
$f_5(x)$	0.051	0.0511	0.0357	0.015	0.013
$f_6(x)$	0.0463	0.0523	0.036	0.015	0.013
$f_7(x)$	0.7911	0.2238	0.0678	0.0466	0.02060
$f_8(x)$	0.0563	0.0537	0.0351	0.0151	0.013
$f_9(x)$	0.0715	0.0574	0.0385	0.0175	0.0151

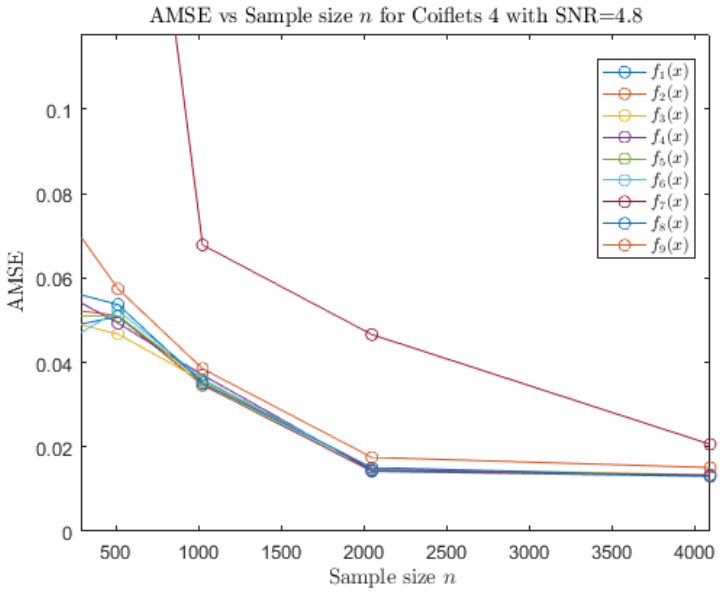


(a) Daubechies filter, $\sigma^2 = 0.25$

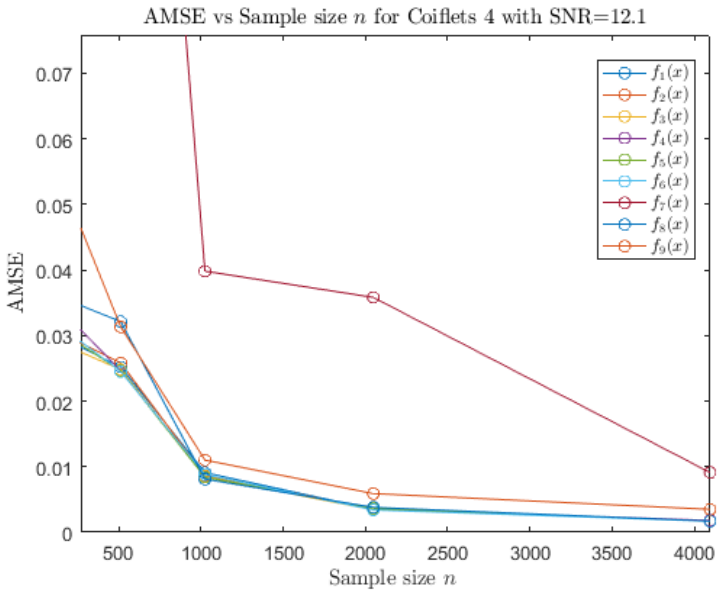


(b) Daubechies filter, $\sigma^2 = 0.75$

Figure 4.14: RMSE results for Beta Design using Daubechies filters, for values of $\sigma^2 = 0.25\,,\ 0.75$..

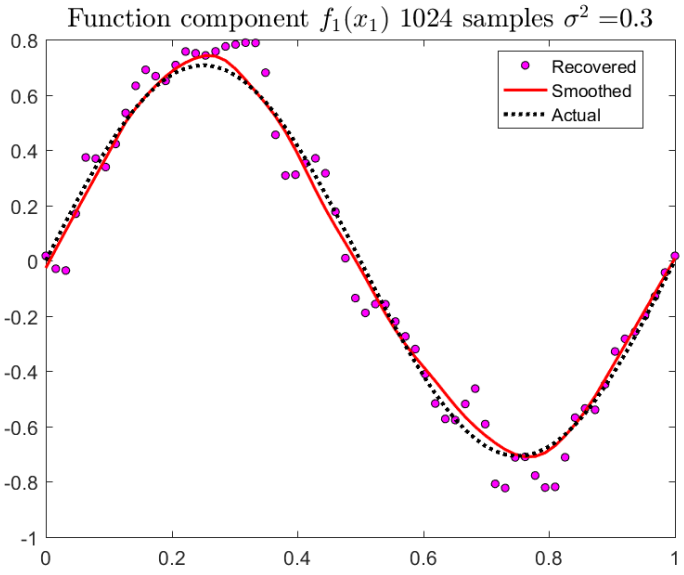


(a)

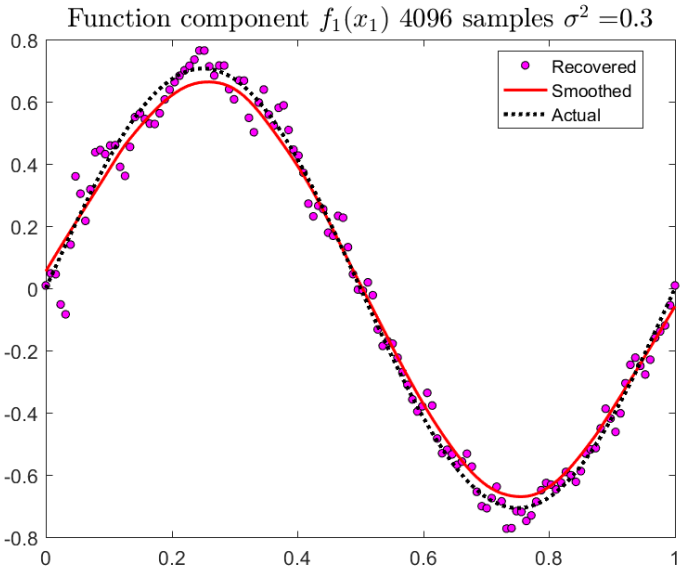


(b)

Figure 4.15: RMSE results for each function using Coiflets 24 filter, for values of $\sigma^2 = 0.25\,,\ 0.75$..

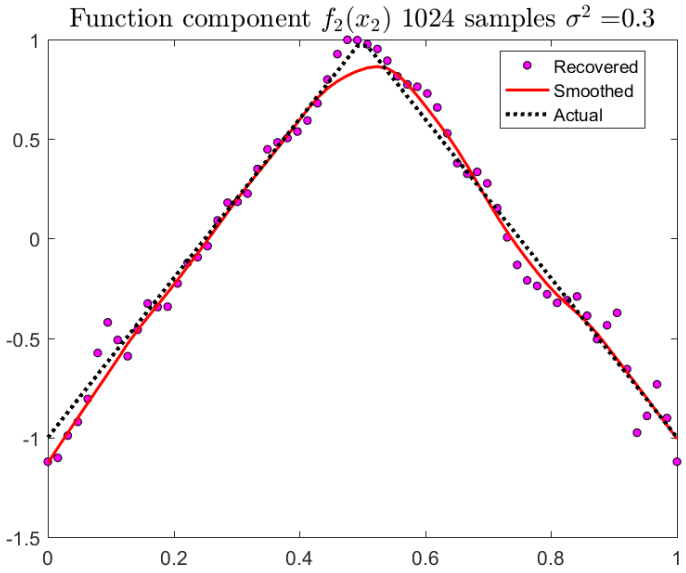


(a)

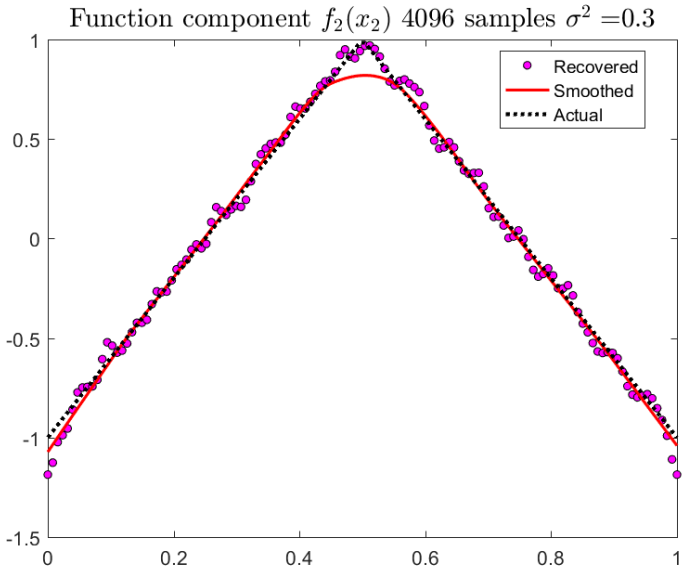


(b)

Figure 4.16: Estimated $f_1(x)$ using Beta Design and Coiflets filter.

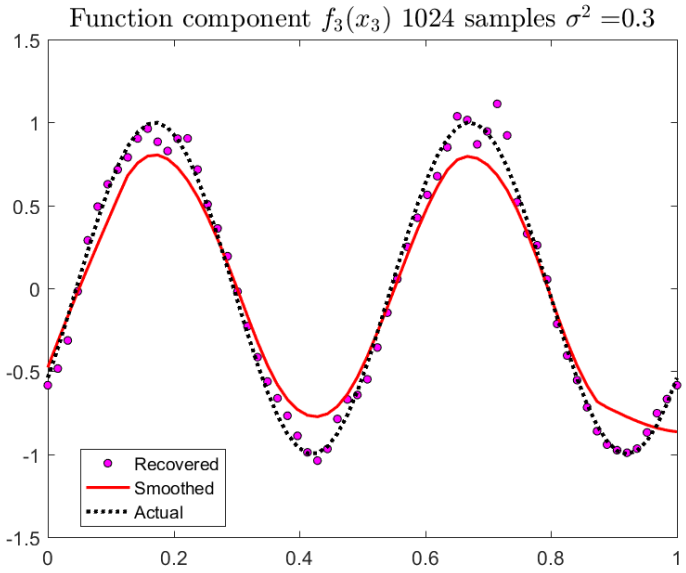


(a)

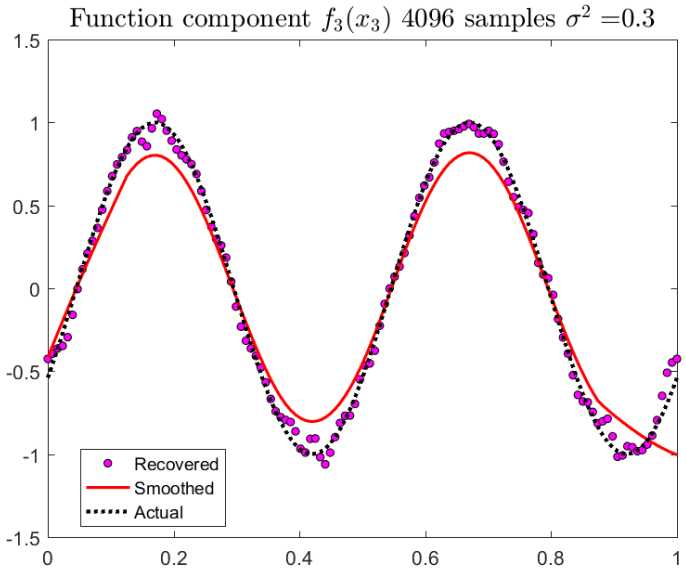


(b)

Figure 4.17: Estimated $f_2(x)$ using Beta Design and Coiflets filter.

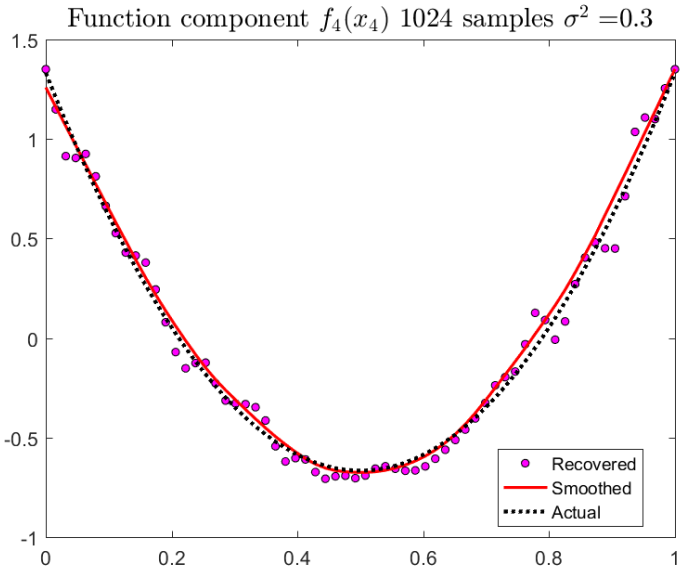


(a)

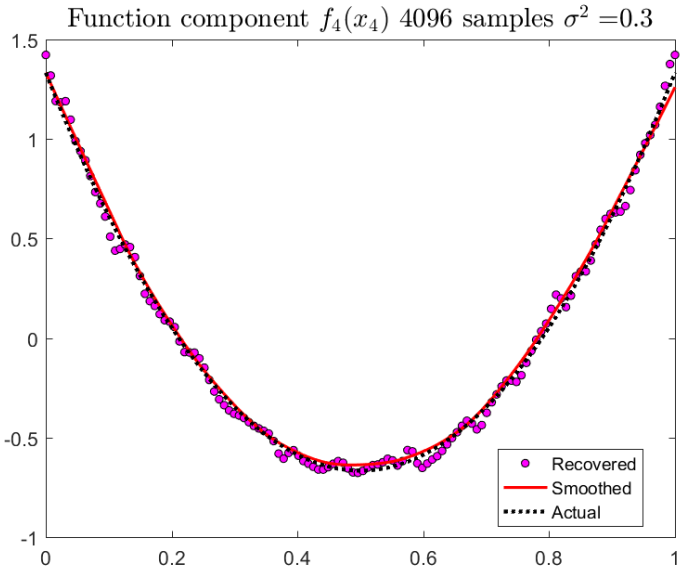


(b)

Figure 4.18: Estimated $f_3(x)$ using Beta Design and Coiflets filter.

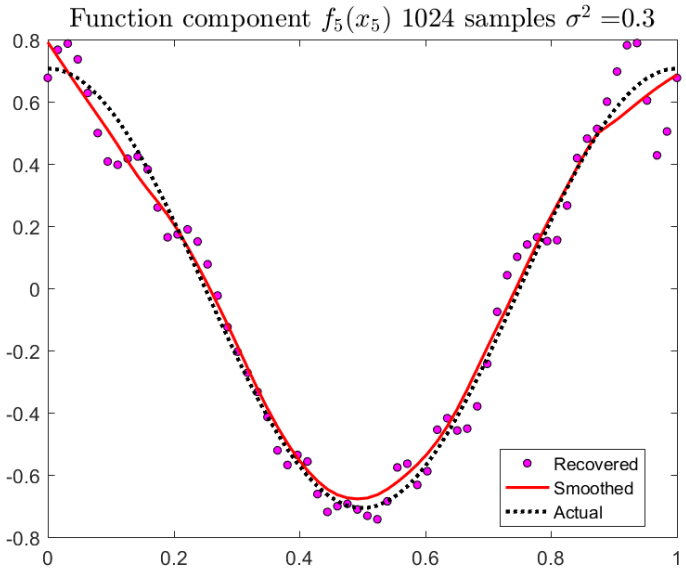


(a)

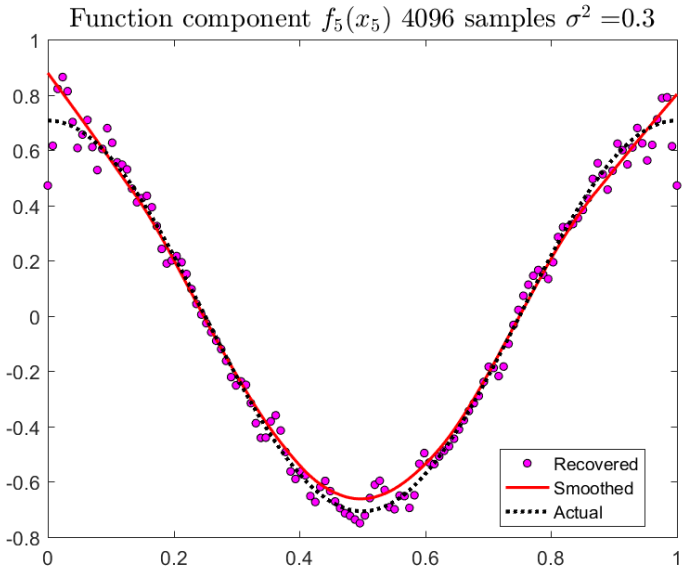


(b)

Figure 4.19: Estimated $f_4(x)$ using Beta Design and Coiflets filter.

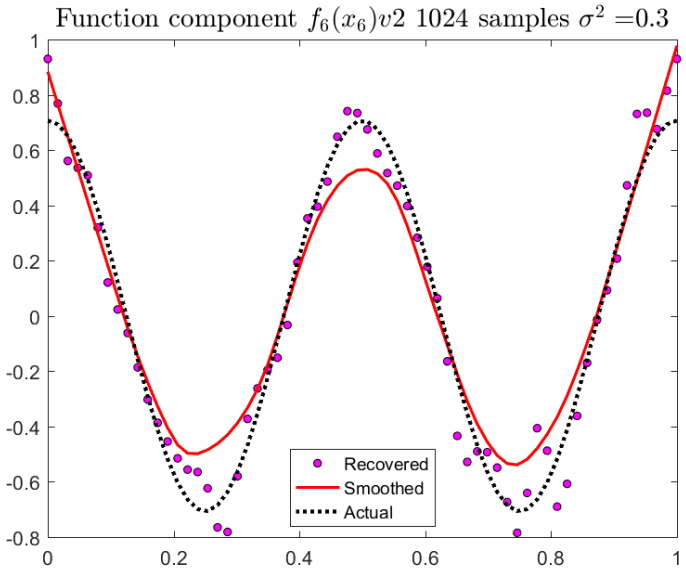


(a)

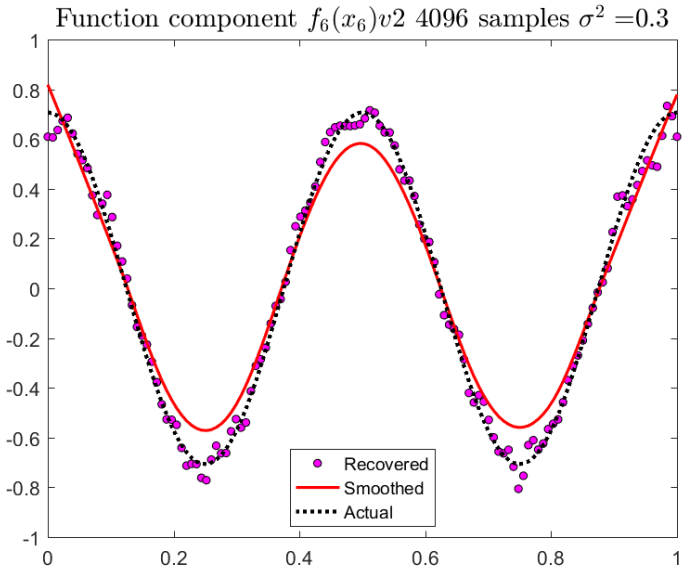


(b)

Figure 4.20: Estimated $f_5(x)$ using Beta Design and Coiflets filter.

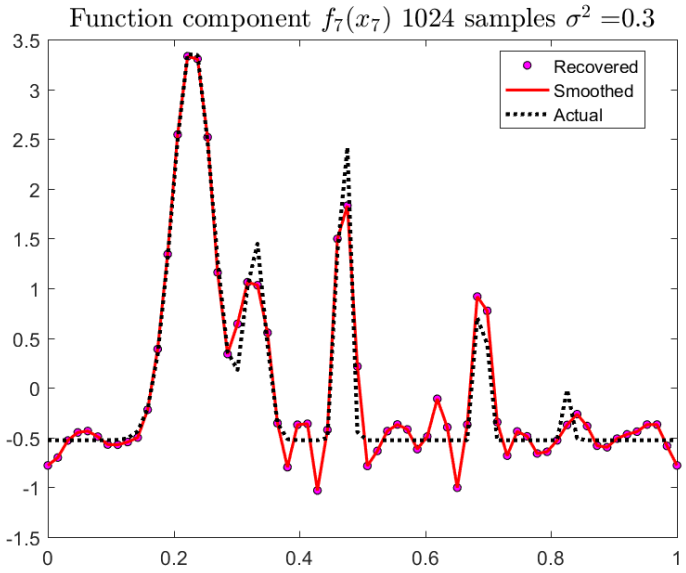


(a)

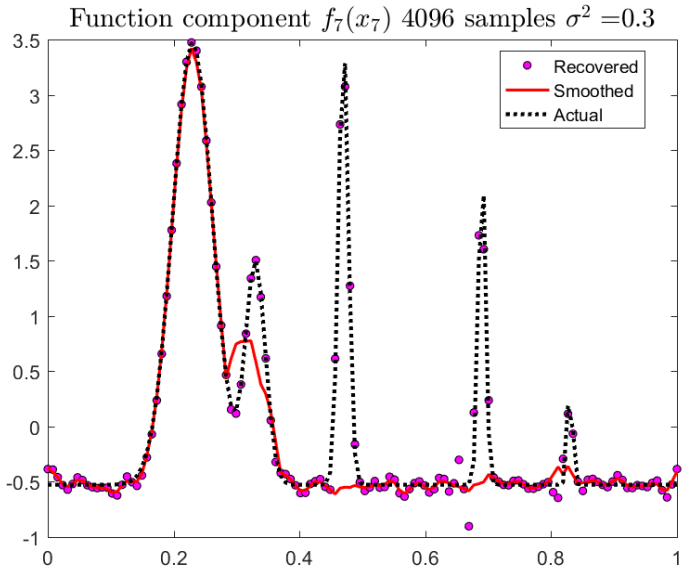


(b)

Figure 4.21: Estimated $f_6(x)$ using Beta Design and Coiflets filter.

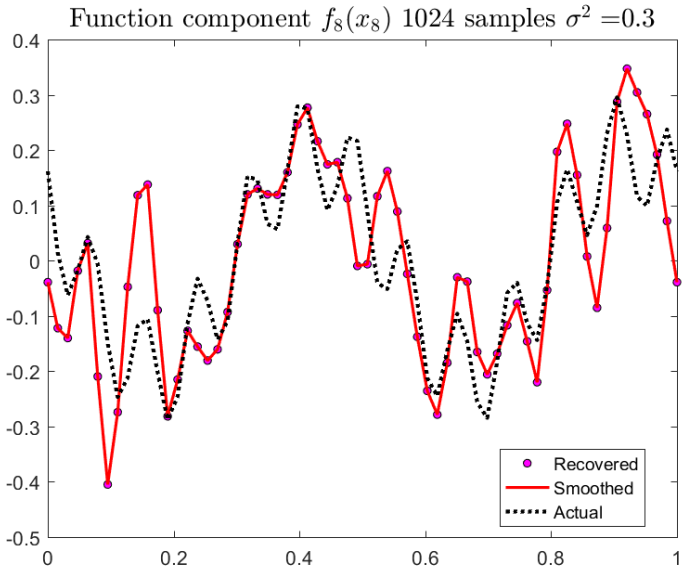


(a)

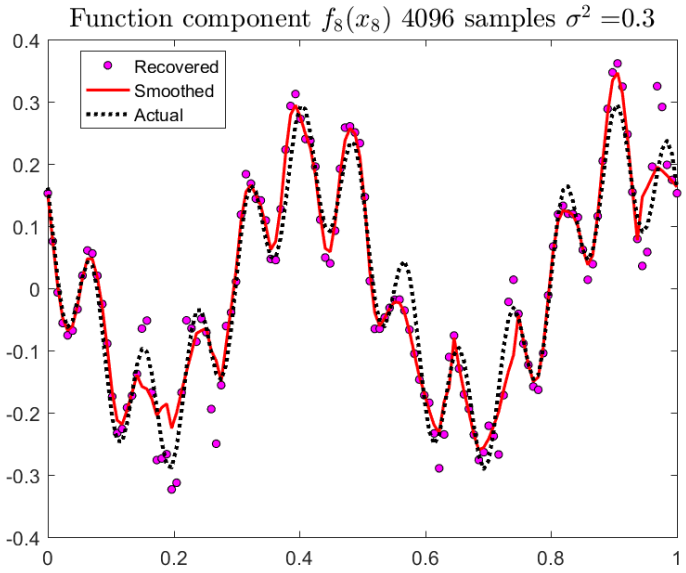


(b)

Figure 4.22: Estimated $f_7(x)$ using Beta Design and Coiflets filter.

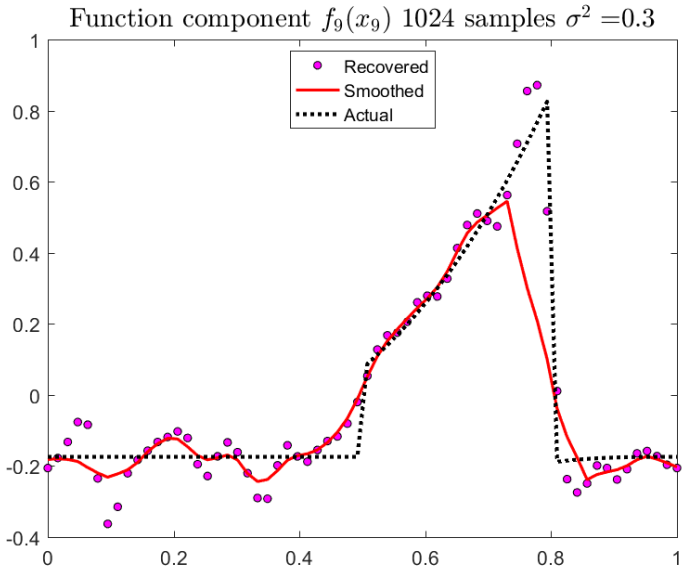


(a)

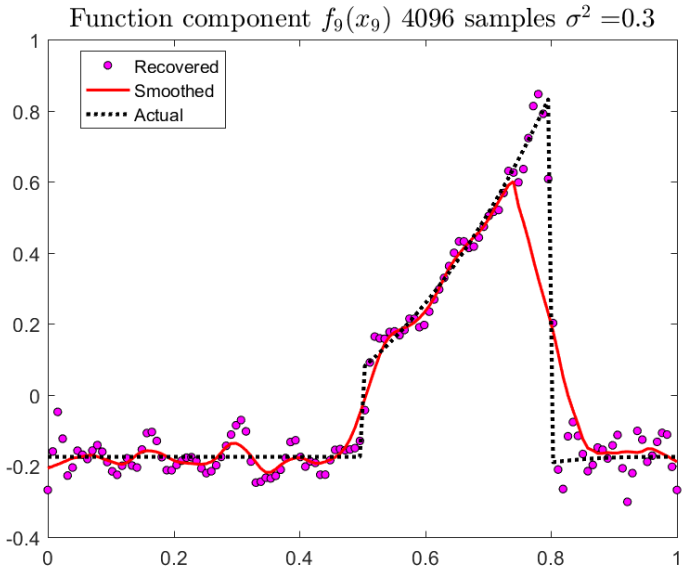


(b)

Figure 4.23: Estimated $f_8(x)$ using Beta Design and Coiflets filter.



(a)



(b)

Figure 4.24: Estimated $f_9(x)$ using Beta Design and Coiflets filter.

4.4 Practical Application of Wavelet based Least Squares Method

In this section we consider the implementation of our proposed estimator using a dataset available at the machine learning repository of UCI[1] concerning the study of hourly full load electrical output power (EP) of a combined cycle plant.

This data set was extensively analized by Tüfekci (2014)[58] using different statistical models, with the goal of predicting EP based on 4 available features. That research utilized a variety of predictive methods including: Simple Linear Regression (SLR), Multilayer Perceptron (MLP), Radial Basis Function Neural Network (RBF), Additive Regression (AR, using back-fitting), KStar (instance-based classifier), Locally Weighted Learning, Bagging REP Tree (BREP, Bootstrap based tree methods), Model Tree rules, Model Tress Regression (M5P), REP Trees, Support Vector Regression, Least Median Square (LMS), etc. A total of 15 statistical models were used and compared using 2-fold Crossvalidation after randomly shuffling the data 5 times. Then, prediction accuracy was evaluated using RMSE as an error metric.

Data set description

The dataset contains 9568 data points collected from a Combined Cycle Power Plant[2] over 6 years (2006-2011), when the power plant was set to work with full load. The features are used to predict the net hourly electrical energy output (EP) of the plant and consist of :

(a) Temperature (AT) : This input variable is measured in degrees Celsius and it varies

[1]UCI Machine Learning Repository `http://archive.ics.uci.edu/ml`. Irvine, CA: University of California, School of Information and Computer Science.

[2] A combined cycle power plant (CCPP) is composed of gas turbines (GT), steam turbines (ST) and heat recovery steam generators. In a CCPP, the electricity is generated by gas and steam turbines, which are combined in one cycle, and is transferred from one turbine to another. While the Vacuum is collected from and has effect on the Steam Turbine, he other three of the ambient variables effect the GT performance.

between 1.81C and 37.11C.

(b) Ambient Pressure (AP): This input variable is measured in millibar with an observed range from 992.89 to 1033.3 mbar.

(c) Relative Humidity (RH): This variable is measured as a percentage with an observed range from 25.56% to 100.16%.

(d) Exhaust Vacuum (V): This variable is measured in cm Hg with with an observed range from 25.36 to 81.56 cm Hg.

The characteristics of the data are the following: The EP is measured in mega watt with an observed range from 420.26 to 495.76 MW. Similarly, the general structure of the dataset can be summarized as:

Table 4.9: Application Data Set characteristics, obtained from [58].

Data Set characteristics	`Multivariate`
Number of samples	`9568`
Attribute characteristics	`Real`
Number of Attributes	5

More details about the data set and the problem in hand can be found in [58].

Implementation settings and results

For this analysis, we chose the following implementation settings:

(a) Daubechies 4 filter for the scaling functions.

(b) $J(n) = 1 + \lfloor \log_2(n) - \log_2(\log(n)(\log(n)+1)) \rfloor$.

(c) The response y was centered and standardized and the predictors $\mathbf{X}_1, ..., \mathbf{X}_n$ were re-scaled to $[0,1]^4$.

(d) To prevent unstable estimates at the tails of the marginal distributions of the predictors, we restricted the estimation range to the 95% empirical quantiles of the observed sample.

(e) The data was randomly split into training and testing over the samples belonging to the hypercube defined by the 95% empirical quantiles. 85% of the data was selected for training and the remaining 15% for testing purposes. The estimation process was repeated 100 times. The results for this procedure are illustrated in figures 4.26-4.27b.

(f) For comparison purposes (with results presented in Table 10 [58]), we also implemented the proposed method using 2-fold CV with Coiflets 24 filter. The process was replicated 10 times. In this case, the wavelet coefficients were obtained using the complete sample, without restricting the range of the estimation. Table 4.10 illustrates the differences in accuracy for the wavelet-based estimator and the best regression techniques used in [58].

The obtained results are summarized in the following figures and tables:

(i) Figure 4.26 shows the estimated unknown functions acting on each one of the problem features.

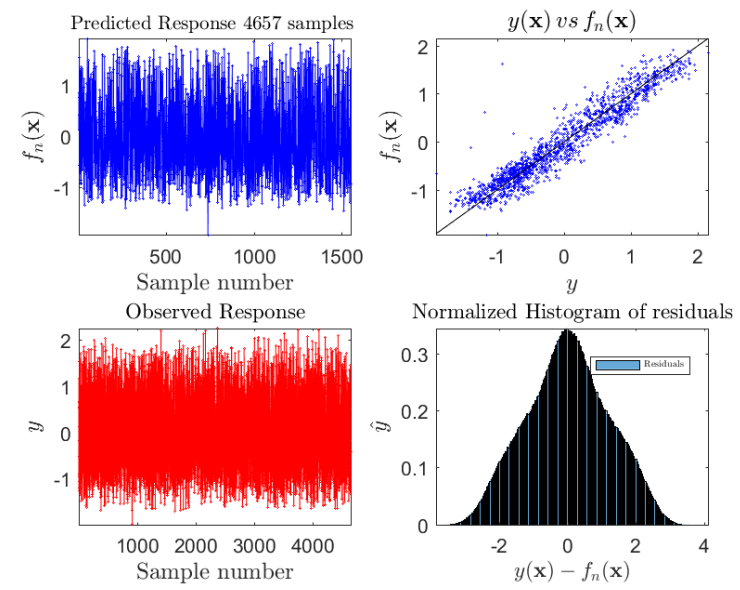
(ii) Figure 4.25a shows the estimated and actual standardized response, together with the $f_n(\mathbf{x})$ vs y plot and the residual plot $e_i = f_n(\mathbf{x}_i) - y_i$.

(iii) Table 4.10 shows RMSE for best methods in [58] and the Wavelet-based LS using 4 features.

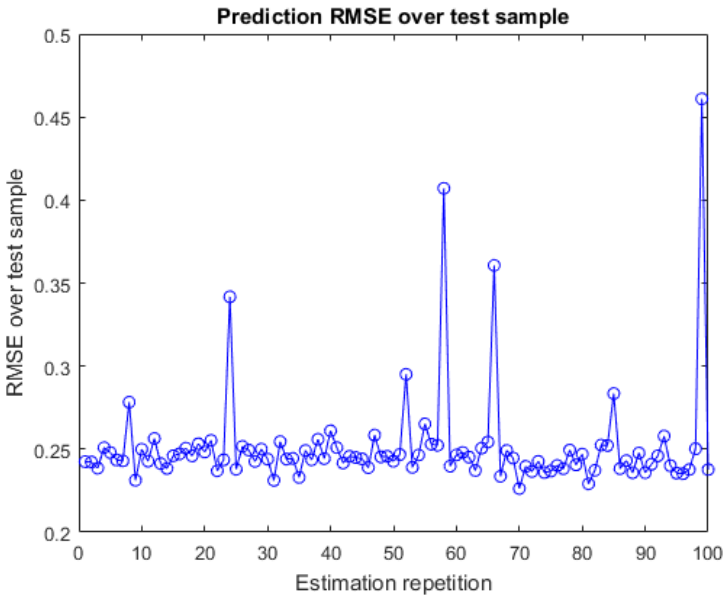
(iv) Table 4.11 shows RMSE for best methods in [58] and the Wavelet-based LS using 1 feature (AT).

(v) Table 4.12 shows RMSE for best methods in [58] and the Wavelet-based LS using 2 features (AT-V).

(vi) Table 4.13 shows RMSE for best methods in [58] and the Wavelet-based LS using 3 features (AT-V-RH).



(a)

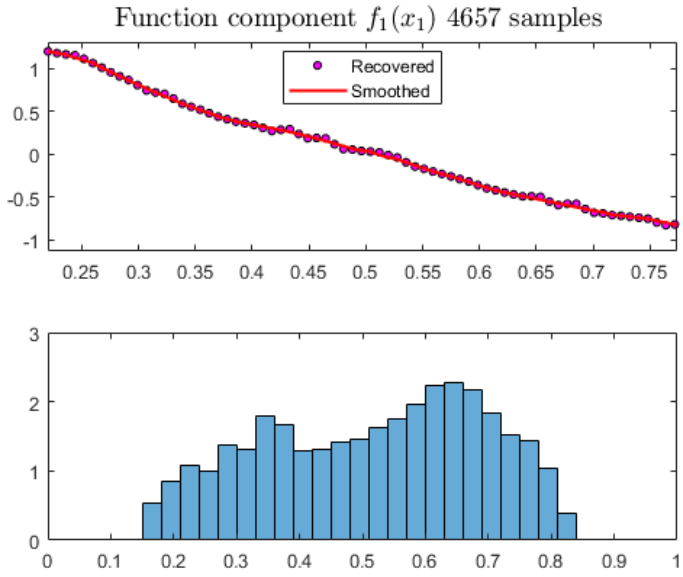


(b)

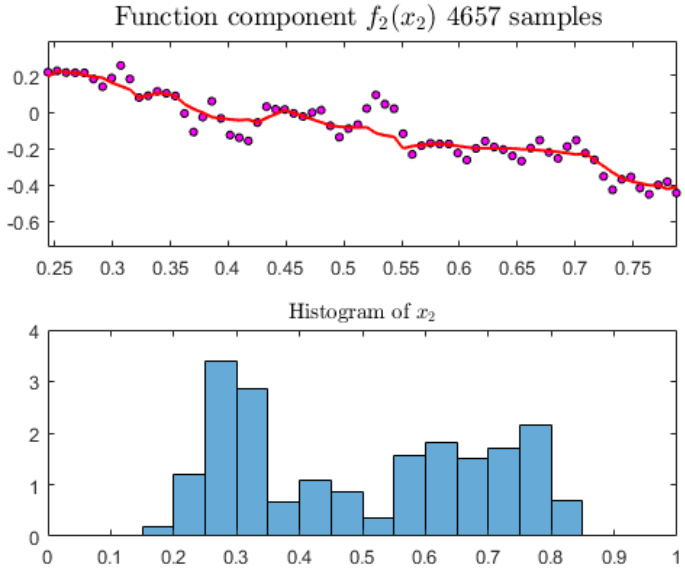
Figure 4.25: Estimaion result plots over the 95% empirical quantiles region and RMSE (computed using the standardized predictions) obtained over 100 replications.

Table 4.10: Comparison results for RMSE for best methods in [58] and the Wavelet-based LS using 4 features.

Kstar	BREP	M5P	MLP	RBF	LMS	SMOREg	M5R	REP	AR	Wavelet LS
3.861	3.787	4.087	5.339	8.487	4.572	4.563	4.128	4.211	5.556	4.325



(a) Estimated $f_1(x)$, corresponding to AT.

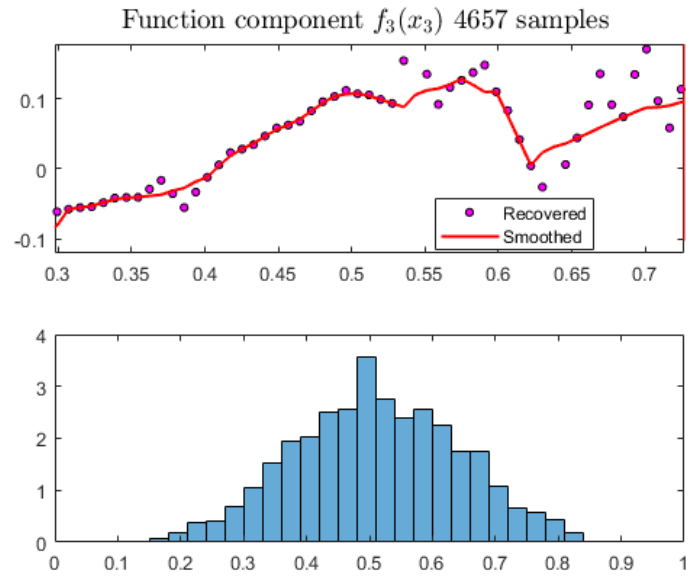


(b) Estimated $f_2(x)$, corresponding to AP.

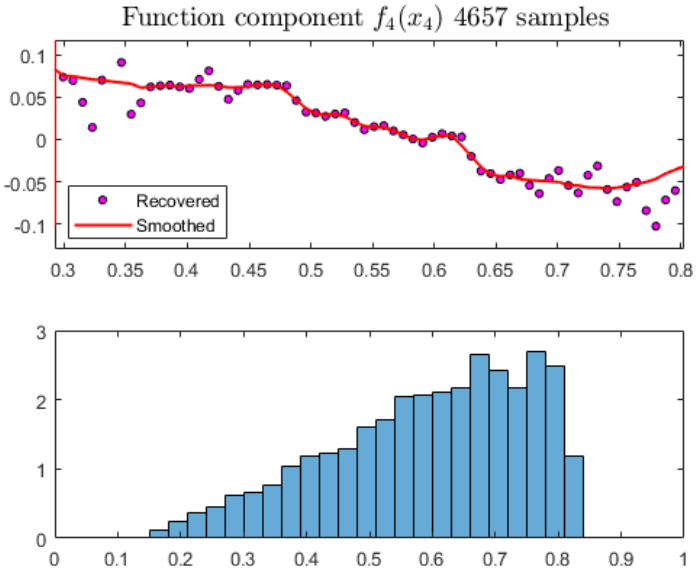
Figure 4.26: Estimated $f_1(x)$ and $f_2(x)$ over the 95% empirical quantiles region. The bottom panel illustrates the sample histograms for each considered feature, within the 95% empirical quantiles region.

Table 4.11: Comparison results for RMSE for best methods in [58] and the Wavelet-based LS using 1 feature (AT).

Kstar	BREP	M5P	LMS	SMOREg	M5R	REP	Wavelet LS
5.381	5.208	5.086	5.433	5.433	5.085	5.229	5.085



(a) Estimated $f_3(x)$, corresponding to RH



(b) Estimated $f_4(x)$, corresponding V.

Figure 4.27: Estimated $f_3(x)$ and $f_4(x)$ over the 95% empirical quantiles region. The bottom panel illustrates the sample histograms for each considered feature, within the 95% empirical quantiles region.

Table 4.12: Comparison results for RMSE for best methods in [58] and the Wavelet-based LS using 2 features (AT-V).

Kstar	BREP	M5P	LMS	SMOREg	M5R	REP	Wavelet LS
4.634	4.026	4.359	4.968	4.968	4.419	4.339	4.757

Table 4.13: Comparison results for RMSE for best methods in [58] and the Wavelet-based LS using 3 features (AT-V-RH).

Kstar	BREP	M5P	LMS	SMOREg	M5R	REP	Wavelet LS
4.331	3.934	4.178	4.580	4.585	4.217	4.291	4.776

Remarks and Comments

(i) From figures 4.26a-4.27b it is possible to observe that the wavelet-based estimator is able to capture the non-linear influences of each of the features considered in the model. From the plots it is possible to assess the significance of each one of the uncovered functions in the model; in particular, 4.26a shows an almost linear effect of the Temperature over EP with negative correlation. For the rest of the predictors, the effect on the response is almost negligible.

(ii) From figure 4.25a, we can conclude that the wavelet-based estimator is able to successfully predict the EP over the test sample. The predicted vs actual values lie in a straight line with no evident deviations apart from the noise in the data, showing a strong correlation between predicted and actual values.

(iii) In table 4.10, the average RMSE for the Wavelet-based LS method was 4.325 (non-standardized testing sample) which shows to be better than most of the results shown in Table 10 [58]. In particular, the best regression methods studied in such reference (i.e. Bagging REP Tree, KStar, Model Trees Regression) achieve mean RMSE of 3.861, 3.787 and 4.087 respectively which shows how suitable the wavelet-based least squares estimator is for the non-linear additive model setting. Even though it could be argued that our comparison is based on results that were obtained under different settings than the baseline experiments, the obtained RMSE shows competitive results for the wavelet-based model. Moreover, the estimation experiments conducted using 85% of the data for training and the remaining 15% for testing suggest that the prediction RMSE could

be even smaller than 4.17, which together with the simplicity of implementation positions the wavelet-based least squares method as a competitive for this kind of problems.

4.5 Conclusions and Discussion

This Chapter introduced a wavelet-based methodology for the non-parametric estimation and prediction of non-linear additive regression models with NESD. The proposed estimator is based on the projection of the unknown additive functions onto the space V_J generated by an orthonormal wavelet basis. In this setting, the data driven wavelet coefficients that define the model are obtained using a truncated least squares estimates.

For the proposed estimator, we showed its strong consistency and illustrated practical results via simulations using different exemplary baseline functions. Moreover, we provided convergence rates and optimal choices for the multiresolution index J and the truncation parameter β_n.

Our results show that our estimator doesn't suffer from the curse of dimensionality, and was observed to be robust with respect to sample size and noise variance in the model. In fact, our results show that the proposed method is able to successfully identify and predict the underlying model functions and response for relatively small sample sizes.

Moreover, the proposed estimators are completely data driven with only a few parameters of choice left to the user (multiresolution index J, wavelet filter and truncating parameter β_n). Also, the utilized matrix based structure introduces computational speed and makes the estimators suitable for real-life applications. In our model, we used Daubechies-Lagarias's algorithm for the evaluation of the scaling functions ϕ_{Jk}^{per} at the observed sample points X_{ij}. This stage of the estimation process corresponds to the bottleneck in terms of computationally speed, but for moderate sample sized, the cost of construction is relatively reasonable.

From a real data application viewpoint, in section 4.4 we tested the proposed least squares method using a real data set that was extensively analyzed by Tüfekci (2014) [58]. The obtained results show that the proposed estimators are capable of uncover the existing non-linear relationships between the response and predictors, while achieving a high predictive accuracy. In particular, the wavelet-based least squares method showed to be more accurate than the additive model based on backfitting used in [58].

In terms of some of the drawbacks observed throughout this research for the proposed method, it is possible to obtain abnormally large wavelet coefficients in those design regions were the number of samples is small (this is highly likely to occur at the tails of the design distribution); Also, some problems may arise at the boundaries of the support due to the periodic wavelets extension. Nonetheless, it is possible to adjust the truncating parameter β_n using cross-validation, which minimizes the effect of those large wavelet coefficients that induce errors in the prediction of the response and may contribute to reduce the effect predictors following exponentially decaying distributions.

In summary, based on the theoretical properties and results obtained in this Chapter, we can argue that the proposed estimators posses interesting interpretations and results and add value to practical data analysis: it has good asymptotic properties, is able to identify models that might be hard to do using other methods and also, it is relatively easy to implement which increases its potential to reach a wide variety of users.

Since the introduced methodology of this chapter relies in the assumption that the dimensionality of the design matrix $\mathbf{B}$ satisfies $p \cdot 2^{J(n)} \leq n$, it would be of high interest to investigate an alternative approach that relies in the regularization of the objective function 4.9, thus preventing an ill-conditioned least squares solution. For this reason, in the next Chapter an alternative approach that corresponds to a ridge-type least squares for the additive regression

problem is investigated.

CHAPTER 5

BAYESIAN APPROACH FOR NON-LINEAR ADDITIVE REGRESSION MODELS USING CONJUGATE $\mathcal{NIG}$ STRUCTURES

In this Chapter, a shrinkage-based estimator for the non-linear additive regression problem in the presence of gaussian noise is introduced. This shrinkage procedure results from the application of a Normal-Inverse-Gamma ($\mathcal{NIG}$) hierarchical model in three different settings: One general model in which it is assumed that the expansion coefficients follow conditionally a Normal distribution, with variance controlled by a single parameter τ. This approach is implemented using a backfitting methodology that allows the sequential estimation of each function in the model, choosing the parameter τ that minimizes the empirical MSE from the data.

Secondly, a more general modelling framework based on a mixture of two $\mathcal{NIG}$ models as joint prior on the expansion coefficients is introduced, enhancing the adaptability of the model to different degrees of smoothness of functions in the model. Similarly as for the general approach, this model is implemented using a backfitting approach, with prior parameters computed from the data, following the recommendations provided in Vidakovic and De Canditiis (2001)[59].

Next, a special case of the mixture model is introduced. Here, it is assumed that the expansion coefficients in the model are distributed by a point-mass contaminated Gaussian distribution, conditional on the noise variance σ^2. The point mass models non-energetic coefficients, while the Gaussian component represents the more "spread" distribution modeling large wavelet coefficients.

The conjugacy structure of the $\mathcal{NIG}$ model allows the derivation of closed-form expressions

for the shrinkage rule, that result in a explicit and fast estimation rules. These expressions are derived in the sequel, and algorithmic procedures for the estimation rules are proposed. Finally, both models are implemented and evaluated, illustrating their performance against a set of functions, and comparing the results with the Least Squares approach introduced in Chapter 4.

5.1 Introduction

Wavelet-based estimation procedures have shown to be appropriate for settings in which it is needed to estimate functions with unknown smoothness in and adaptive fashion. In particular, in section 1.1.5 the use of wavelet-based orthogonal basis as a characterization tool for functional spaces was discussed, showing the adaptability potential of this mathematical tool for problems in which unknown functions need to be estimated given a set of examples.

Over the course of the last two decades, a variety of non-parametric shrinkage methods have been proposed and studied. The works by Donoho and Johnstone (1994)[34], and by Donoho *et al.* (1995)[27] first introduced *RiskShrink, VisuShrink, SureShrink* and their modifications. These procedures, non-linear in nature, exploit the sparsity in the representation of signal in the wavelet domain, developing different data-dependent thresholding rules for the expansion coefficients, resulting in non-linear adaptive estimators.

In the non-Bayesian domain, Zhang and Wong (2003)[56] proposed a two-stage wavelet thresholding procedure using local polynomial fitting and marginal integration for the estimation of the additive components. Their method is adaptive to different degrees of smoothness of the components and has good asymptotic properties. Later on Sardy and Tseng (2004)[1] proposed a non-linear smoother and non-linear back-fitting algorithm that is based

on `WaveShrink`, modeling each function in the model as a parsimonious expansion in a wavelet basis that is further subjected to variable selection (i.e. which wavelets to use in the expansion) via non-linear shrinkage.

These methods (non-Bayesian) have been shown to possess excellent approximation properties, achieving minimax convergence rates over a variety of functional spaces such as Besov and Sobolev, and exploiting the speed of the discrete wavelet transform (DWT) that enables computational power when dealing with large dimensions/sample sizes. However, most of these methods are restricted to the univariate case and rely on equally-spaced observations of the model features, except for the two methodologies introduced by Zhang and Wong (2003)[56] and Sardy and Tseng (2004)[1].

Following the line of non-Bayesian methods for additive regression, in chapters 3 and 4, two different methodologies based on Wavelets that exploit their approximation capabilities were introduced. In particular, the Least Squares approach that was analyzed in Chapter 4 showed excellent asymptotic properties and estimation power even for small sample sizes. However, the theoretical guarantees and performance were subject to a restriction in the dimensions of the projection matrix generated by the evaluations of the scaling functions $\phi_{J,k}(\cdot)$ that generate the multiresolution space V_J, limiting its flexibility for real-life applications.

In the context of Bayesian methodologies, several approaches have been introduced in the literature since the seminal work by Donoho and Johnstone (1994) for the problem of functional estimation. The Bayesian paradigm has been proven to be suitable for this kind of statistical problems, since it allows the incorporation of prior information that is related to the underlying signal properties such as smoothness, periodicity, selfsimilarity, etc. Some ex-

amples of these procedures can be found in Vidakovic and Ruggeri (2001)[60], De Canditiis and Vidakovic (2001)[59], Hall, Kerkyacharian and Picard (1998,1999)[61], Chipman, Kolaczyk and McColloch (1997)[62], among others. These methodologies introduce different shrinkage rules (level-wise, block-based) that are obtained through closed-form expression, and are restricted either to the univariate and/or equally spaced designs.

In addition to these methodologies, Bayesian procedures that rely on Montecarlo Markov Chain (MCMC) approximations for the Bayesian inference and derivation of numerical shrinkage rules have been also proposed in the literature. For example, the work by Brezger and Lang (2006)[63], and Fahrmeir, Ludwig, et al.(2004)[64] provide procedures that are based on the use of P-splines and empirical Bayes, combined with MCMC simulations for the inference. Although these kind of methods have shown good estimation properties, their performance comes at the expense of computational costs and the challenging theoretical analysis of statistical properties.

The limitations of the existing Bayesian methodologies motivate the subject of this chapter: the development of a methodology that treats the non-linear additive regression problem in a more flexible way than the Least Squares approach, while capturing the adaptivity of the existing Bayesian procedures.

In fact, in this Chapter, we aim to extend the flexibility and estimation power of Bayesian methodologies for the problem of non-linear additive regression with non-equally spaced designs. At first, shrinkage-based estimator for the non-linear additive regression problem in the presence of gaussian noise is introduced. This shrinkage procedure results from the utilization of a Normal-Inverse-Gamma ($\mathcal{NIG}$) model in three different settings: One general

model in which the expansion coefficients are assumed to be independent and conditionally distributed as a Gaussian random vector, with variance controlled by a single parameter τ. This approach is implemented using a backfitting methodology that allows the sequential estimation of each function in the model, while choosing the parameter τ that minimizes the empirical MSE.

Secondly, a more general modelling framework based on a mixture of two $\mathcal{NIG}$ models as joint prior on the expansion coefficients is introduced, enhancing the adaptability of the model to different degrees of smoothness of the underlying functions in the model. Similarly as for the general approach, this methodology is implemented using a backfitting approach, with prior parameters computed from the data, following the recommendations provided in Vidakovic and De Canditiis (2001)[59].

Next, a special case of the mixture model is introduced. Here, it is assumed that the expansion coefficients in the model are distributed by a point-mass contaminated Gaussian distribution, conditional on the noise variance σ^2. The point mass models non-energetic coefficients, while the Gaussian component represents the more "spread" distribution modeling large wavelet coefficients. The goal of this model is to provide a more parsimonious estimation of the wavelet coefficients, as a result of a more strict shrinkage rule.

The conjugacy structure of the $\mathcal{NIG}$ model allows the derivation of closed-form expressions for the shrinkage rule, that result in a explicit and fast estimation rules. These expressions are derived in the sequel, and algorithmic procedures for the estimation rules are proposed.

Finally, the proposed models are implemented and evaluated, illustrating their performance

against a set of functions, and comparing the results with the Least Squares approach introduced in Chapter 4, and the procedure proposed by Sardy and Tseng (2004)[1].

5.2 Bayesian Extension of the Non-linear Additive Regression Problem Using Gaussian Conditional $\mathcal{NIG}$ Model.

Let us recall the additive regression model given by:

$$\begin{aligned} y(\mathbf{x}) &= f(\mathbf{x}) + \sigma \cdot \epsilon \,, \\ &= \sum_{l=1}^{p} f_l(x_l) + \sigma \cdot \epsilon \,. \end{aligned} \tag{5.1}$$

Here $\mathbf{x} \in [0,1]^p$, $f_l(x_l) \in \mathbb{L}_2([0,1])$, and $\epsilon \sim \mathcal{N}(0,1)$. Furthermore, for identifiability purposes, assume $\int_0^1 f_l(x_l)dx_l = 0$, for $l = 1 \ldots, p$. Note that this condition implies that:

$$||f(\mathbf{x})||_{\mathbb{L}_2([0,1]^p)} = \sum_{l=1}^{p} ||f_l(x_l)||_{\mathbb{L}_2([0,1])}.$$

In general, it follows that $||f(\mathbf{x})||_{\mathbb{L}_2([0,1]^p)} = \mathbb{1}^T \mathbf{G} \mathbb{1}$, where $\mathbf{1}^T = \Big[11...1\Big]_{1\times p}$, and $\mathbf{G}$ is a $p \times p$ matrix with entries given by:

$$\mathbf{G} = \begin{bmatrix} \beta_{11} & \beta_{12} & \ldots & \beta_{1p} \\ \beta_{21} & \beta_{22} & \ldots & \beta_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \beta_{p1} & \beta_{p2} & \ldots & \beta_{pp} \end{bmatrix}_{p\times p},$$

where the entries $\beta_{ij} = \int_0^1 \int_0^1 f_i(x_i) f_j(x_j) dx_i dx_j$, for $i, j = 1, ..., p$. It is clear that this matrix is symmetric.

Moreover, under the identifiability condition $\int_0^1 f_l(x_l)dx_l = 0$, for $l = 1 \ldots, p$, $\mathbf{G}$ is diagonal, with diagonal entries given by the $\mathbb{L}_2$ norm of the unknown functions in the model.

Now, suppose that each unknown function in the model (5.1) can be approximated by an element (i.e. a function) of a subspace V_J spanned by the wavelet orthonormal basis:

$$\left\{\phi_{00}, \psi_{jk}\ ,\ j = 0, ..., J-1;\ k = 0, ..., 2^{J-1}\right\},$$

for some multiresolution index J. Therefore, by the orthogonality principle, it follows that for $l = 1, ..., p$:

$$\begin{aligned} f_{l,J}(x) &= c_{00}^{(l)}\phi_{00}(x) + \sum_{j=0}^{J-1}\sum_{k=0}^{2^j-1} d_{jk}^{(l)}\psi_{jk}(x) \\ &= \langle f_l, \phi_{00}\rangle_{\mathbb{L}_2([0,1])}\phi_{00}(x) + \sum_{j=0}^{J-1}\sum_{k=0}^{2^j-1}\langle f_l, \psi_{jk}\rangle_{\mathbb{L}_2([0,1])}\psi_{jk}(x). \end{aligned} \tag{5.2}$$

Now, from (5.2) and (5.1), it is possible to represent $f_J(\mathbf{x})$ (i.e. the projection of the function $f(\mathbf{x})$ onto the space V_J) as follows:

$$f_J(\mathbf{x}) = \mathbf{c}_J^T\tilde{\mathbf{\Psi}}(\mathbf{x}), \tag{5.3}$$

where:

$$\mathbf{c}_J = \begin{bmatrix} \mathbf{c}_J^{(1)} \\ \mathbf{c}_J^{(2)} \\ \vdots \\ \mathbf{c}_J^{(p)} \end{bmatrix}_{p\cdot 2^J\times 1}, \quad \mathbf{c}_J^{(l)} = \begin{bmatrix} c_{00}^{(l)} \\ \mathbf{d}_J^{(l)} \end{bmatrix}_{2^J\times 1}, \tag{5.4}$$

where $c_{00}^{(l)},\ \mathbf{d}_J^{(l)}\ \ l = 1, ..., p$ are the expansion coefficients of $f_l(x_l)$ in (5.2). Similarly, $\tilde{\mathbf{\Psi}}(\mathbf{x})$

is given by:

$$\tilde{\mathbf{\Psi}}(\mathbf{x}) = \begin{bmatrix} \mathbf{\Psi}(x_1) \\ \mathbf{\Psi}(x_2) \\ \vdots \\ \mathbf{\Psi}(x_p) \end{bmatrix}_{p \cdots 2^J \times 1} , \quad \mathbf{\Psi}(x_l) = \begin{bmatrix} \phi_{00}(x_l) \\ \mathbf{\Psi}_J(x_l) \end{bmatrix}_{2^J \times 1} , \tag{5.5}$$

where $\phi_{00}(x_l),\ \mathbf{\Psi}_J(x_l),\ l = 1, ..., p$ are the scaling and wavelet functions evaluated at the l-coordinate of the feature vector $\mathbf{x}$, as shown in (5.2).

Consider now a sample of the form $\{(y_i, \mathbf{x}_i)\}_{i=1}^N$. Using definitions (5.4) and (5.5), it is possible to form the system:

$$\begin{aligned} y_1 &= \mathbf{c}_J^T \tilde{\mathbf{\Psi}}(\mathbf{x}_1) + \sigma \cdot \epsilon_1 \\ y_2 &= \mathbf{c}_J^T \tilde{\mathbf{\Psi}}(\mathbf{x}_2) + \sigma \cdot \epsilon_2 \\ &\vdots \\ y_N &= \mathbf{c}_J^T \tilde{\mathbf{\Psi}}(\mathbf{x}_N) + \sigma \cdot \epsilon_N. \end{aligned}$$

Putting this system into a matrix form, it follows:

$$\begin{aligned} \mathbf{y}_{N \times 1} &= \begin{bmatrix} \tilde{\mathbf{\Psi}}(\mathbf{x}_1)^T \\ \tilde{\mathbf{\Psi}}(\mathbf{x}_2)^T \\ \vdots \\ \tilde{\mathbf{\Psi}}(\mathbf{x}_N)^T \end{bmatrix}_{N \times p \cdot 2^J} \cdot \mathbf{c}^T_{J_{p \cdot 2^J \times 1}} + \sigma \cdot \boldsymbol{\epsilon}_{N \times 1} \\ \mathbf{y} &= \tilde{\tilde{\mathbf{\Psi}}} \mathbf{c}_J + \sigma \cdot \boldsymbol{\epsilon}. \end{aligned} \tag{5.6}$$

Now, since $\epsilon_i \overset{i.i.d}{\sim} \mathcal{N}(0,1)$, and $\sigma > 0$, it follows:

$$\mathbf{y} \mid \mathbf{X}, \sigma^2, \mathbf{c}_J \ \sim \ \mathcal{N}\left(\tilde{\tilde{\mathbf{\Psi}}}\mathbf{c}_J, \sigma^2\mathbf{I}_N\right), \tag{5.7}$$

where $\mathbf{X}$ is the matrix of observations $\mathbf{x}_1, ..., \mathbf{x}_N$.

Define, $\boldsymbol{\theta} = \tilde{\tilde{\mathbf{\Psi}}}\mathbf{c}_J$. Note that the parameter $\boldsymbol{\theta}$ in the model corresponds to a location parameter. Therefore, the model (5.7) becomes:

$$\mathbf{y} \mid \mathbf{X}, \sigma^2, \boldsymbol{\theta} \ \sim \ \mathcal{N}\left(\boldsymbol{\theta}, \sigma^2\mathbf{I}_N\right). \tag{5.8}$$

From (5.8), since $\tilde{\tilde{\mathbf{\Psi}}}$ is known, by estimating $\boldsymbol{\theta}$ it is possible to recover the expansion coefficients $\mathbf{c}_J$ in the model. Indeed:

$$\hat{\mathbf{c}_J} = \tilde{\tilde{\mathbf{\Psi}}}^{\dagger}\boldsymbol{\theta}, \tag{5.9}$$

where $\tilde{\tilde{\mathbf{\Psi}}}^{\dagger}$ denotes the pseudo-inverse of $\tilde{\tilde{\mathbf{\Psi}}}$.

Remarks:

(i) Note that $\tilde{\tilde{\mathbf{\Psi}}}$ is an $N \times p \cdot 2^J$ matrix. In general, since $\mathbf{x} \in [0,1]^p$ and is assumed to have a probability distribution of the continuous type, $rank(\tilde{\tilde{\mathbf{\Psi}}}) = \min(N, p \cdot 2^J)$. However, in the case of compactly supported wavelets, this matrix tends to sparse especially when the multiresolution index J is large, leading to ill-conditioning problems and numerical instabilities, for this reason, using some regularization technique is recommendable.

(ii) When $\tilde{\tilde{\mathbf{\Psi}}}$ is not a squared matrix, the use of a pseudo-inverse is needed in order to recover the empirical coefficients in the expansion (5.2). In fact, since typically $N < p \cdot 2^J$, the solution for the system $\tilde{\tilde{\mathbf{\Psi}}}^T\tilde{\tilde{\mathbf{\Psi}}}\hat{\mathbf{c}_J} = \tilde{\tilde{\mathbf{\Psi}}}^T\hat{\boldsymbol{\theta}}_{MAP}$ is not unique, meaning that for

any vector $\boldsymbol{\eta}$ within the null-space of $range(\tilde{\tilde{\boldsymbol{\Psi}}}^T)$ the vector $\hat{\mathbf{c}}_j + \alpha\boldsymbol{\eta}$, $\alpha \in \mathbb{R}$ is also a solution.

(iii) Nonetheless, in the case of multiple solutions, the estimate given by Eq.(5.9) corresponds to the one with minimum $\mathbb{L}_2$ norm.

Now, consider the case when $[\mathbf{c}_J, \sigma^2] \sim \mathcal{NIG}(\alpha, \delta, \boldsymbol{\mu}, \boldsymbol{\Sigma})$. Here, $\mathcal{NIG}$ stands for "Normal-Inverse-Gamma" distribution, and its parameters are given by the positive constants α, δ, the vector $\boldsymbol{\mu} \in \mathbb{R}^{p\cdot 2^J}$, and the covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{p\cdot 2^J \times p\cdot 2^J}$, which is assumed to be symmetric positive-definite.

Under the aforementioned joint distribution for parameters σ^2 (scale) and $\boldsymbol{\theta}$ (location), it follows:

$$\begin{aligned} \mathbf{c}_J|\sigma^2 &\sim \mathcal{N}(\boldsymbol{\mu}, \sigma^2\boldsymbol{\Sigma}), \\ \sigma^2 &\sim \mathcal{IG}(\alpha, \delta). \end{aligned}$$

Therefore, the model (5.8) takes the form:

$$\mathbf{y}|\mathbf{c}_J, \sigma^2 \sim \mathcal{N}(\tilde{\tilde{\boldsymbol{\Psi}}}\mathbf{c}_J, \sigma^2\mathbb{I}_N), \tag{5.10}$$

$$\mathbf{c}_J|\sigma^2 \sim \mathcal{N}(\boldsymbol{\mu}, \sigma^2\boldsymbol{\Sigma}), \tag{5.11}$$

$$\sigma^2 \sim \mathcal{IG}(\alpha, \delta). \tag{5.12}$$

Remarks:

(i) The "Normal-Inverse-Gamma" ($\mathcal{NIG}$) priors have been used previously in the wavelet framework because of their conjugate structure with respect to normal conditional models as in (5.8). See Vidakovic and Müller (1995)[65], Vanucci and Corradi (1999)[66] and De Canditiis and Vidakovic (2001)[59].

(ii) The conjugate structure allows for closed form solutions for the Bayes estimators under squared-error loss, which simplifies theoretical analysis, interpretation of results and practical implementation.

(iii) In addition to conjugacy, the $\mathcal{NIG}$ prior allows modeling the dependence between neighboring coefficients in the expansion (5.2). For these reasons, it is a reasonable choice for the analysis and inference of the additive model (5.1).

Now, based on the hierarchical model defined by Eqs.(5.10)-(5.12), our goal is to obtain:

$$\hat{\mathbf{c}_J} = \underset{\mathbf{c}_J \in \mathbb{R}^{p \cdot 2^J}}{\arg\max} \left(\pi(\mathbf{c}_J|\mathbf{y})\right). \tag{5.13}$$

5.2.1 Obtention of the posterior distribution $\pi(\mathbf{c}_J|\mathbf{y})$

Note that using the model defined in Eqs. (5.10)-(5.12), it follows:

$$\begin{aligned} \pi(\mathbf{c}_J, \sigma^2|\mathbf{y}) &= \frac{f(\mathbf{y}|\mathbf{c}_J, \sigma^2)\pi(\mathbf{c}_J|\sigma^2)g(\sigma^2)}{m(\mathbf{y})}, \text{ thus:} \\ \pi(\mathbf{c}_J|\mathbf{y}) &= \int_0^\infty \frac{f(\mathbf{y}|\mathbf{c}_J, \sigma^2)\pi(\mathbf{c}_J|\sigma^2)g(\sigma^2)d\sigma^2}{m(\mathbf{y})}. \end{aligned} \tag{5.14}$$

Here,

$$\begin{aligned} g(\sigma^2) &= \frac{\delta^\alpha}{\Gamma(\alpha)}\left(\frac{1}{\sigma^2}\right)^{\alpha+1} e^{-\frac{\delta}{\sigma^2}}, \\ \pi(\mathbf{c}_J|\sigma^2) &= \frac{1}{(2\pi)^{p\cdot 2^J/2}|\boldsymbol{\Sigma}|^{1/2}\sigma^N} e^{-\frac{1}{2\sigma^2}(\mathbf{c}_J-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{c}_J-\boldsymbol{\mu})}, \\ f(\mathbf{y}|\mathbf{c}_J, \sigma^2) &= \frac{1}{(2\pi)^{N/2}\sigma^N} e^{-\frac{1}{2\sigma^2}(\mathbf{y}-\tilde{\tilde{\boldsymbol{\Psi}}}\mathbf{c}_J)^T(\mathbf{y}-\tilde{\tilde{\boldsymbol{\Psi}}}\mathbf{c}_J)} \\ m(\mathbf{y}) &= \int_{\mathbb{R}^{p\cdot 2^J}}\int_{\mathbb{R}} f(\mathbf{y}|\boldsymbol{\theta}, \sigma^2)\pi(\mathbf{c}_J|\sigma^2)g(\sigma^2)d\sigma^2 d\mathbf{c}_J. \end{aligned}$$

Now, taking $f(\mathbf{y}|\mathbf{c}_J,\sigma^2)\pi(\mathbf{c}_J|\sigma^2)$, it follows by letting $M = p\cdot 2^J$:

$$\begin{aligned} f(\mathbf{y}|\mathbf{c}_J,\sigma^2)\pi(\mathbf{c}_J|\sigma^2) &= \frac{1}{(2\pi)^{\frac{N+M}{2}}|\boldsymbol{\Sigma}|^{1/2}(\sigma^2)^{\frac{N+M}{2}}} e^{-\frac{1}{2\sigma^2}(\mathbf{y}^T\mathbf{y}+\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})} \\ &\quad \times e^{-\frac{1}{2\sigma^2}\left(\mathbf{c}_J^T(\boldsymbol{\Sigma}^{-1}+\tilde{\tilde{\boldsymbol{\Psi}}}^T\tilde{\tilde{\boldsymbol{\Psi}}})\mathbf{c}_J - 2\mathbf{c}_J^T(\tilde{\tilde{\boldsymbol{\Psi}}}^T\mathbf{y}-\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})\right)}. \end{aligned} \tag{5.15}$$

Define:

$$\tilde{\boldsymbol{\Sigma}}^{-1} = \boldsymbol{\Sigma}^{-1} + \tilde{\tilde{\boldsymbol{\Psi}}}^T\tilde{\tilde{\boldsymbol{\Psi}}}, \tag{5.16}$$

$$\boldsymbol{\alpha} = \tilde{\tilde{\boldsymbol{\Psi}}}^T\mathbf{y} + \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}. \tag{5.17}$$

Similarly, note that $\mathbf{c}_J^T\tilde{\boldsymbol{\Sigma}}^{-1}\mathbf{c}_J - 2\mathbf{c}_J^T\boldsymbol{\alpha} = (\mathbf{c}_J-\tilde{\boldsymbol{\Sigma}}\boldsymbol{\alpha})^T\tilde{\boldsymbol{\Sigma}}^{-1}(\mathbf{c}_J-\tilde{\boldsymbol{\Sigma}}\boldsymbol{\alpha}) - \boldsymbol{\alpha}^T\tilde{\boldsymbol{\Sigma}}^{-1}\boldsymbol{\alpha}$. This implies that Eq. (5.15) takes the form:

$$\begin{aligned} f(\mathbf{y}|\mathbf{c}_J,\sigma^2)\pi(\mathbf{c}_J|\sigma^2) &= \frac{1}{(2\pi)^{\frac{N+M}{2}}|\boldsymbol{\Sigma}|^{1/2}(\sigma^2)^{\frac{N+M}{2}}} e^{-\frac{1}{2\sigma^2}(\mathbf{y}^T\mathbf{y}+\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}-\boldsymbol{\alpha}^T\tilde{\boldsymbol{\Sigma}}^{-1}\boldsymbol{\alpha})} \\ &\quad \times e^{-\frac{1}{2\sigma^2}(\mathbf{c}_J-\tilde{\boldsymbol{\Sigma}}\boldsymbol{\alpha})^T\tilde{\boldsymbol{\Sigma}}^{-1}(\mathbf{c}_J-\tilde{\boldsymbol{\Sigma}}\boldsymbol{\alpha})}. \end{aligned} \tag{5.18}$$

Note that Eq. (5.18) can be futher arranged as:

$$\begin{aligned} f(\mathbf{y}|\mathbf{c}_J,\sigma^2)\pi(\mathbf{c}_J|\sigma^2) &= \frac{|\tilde{\boldsymbol{\Sigma}}|^{1/2}}{(2\pi)^{N/2}|\boldsymbol{\Sigma}|^{1/2}}\left(\frac{1}{\sigma^2}\right)^{N/2} e^{-\frac{1}{2\sigma^2}(\mathbf{y}^T\mathbf{y}+\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}-\boldsymbol{\alpha}^T\tilde{\boldsymbol{\Sigma}}^{-1}\boldsymbol{\alpha})} \\ &\quad \times \frac{1}{(2\pi)^{M/2}|\tilde{\boldsymbol{\Sigma}}|^{1/2}(\sigma^2)^{M/2}} e^{-\frac{1}{2\sigma^2}(\mathbf{c}_J-\tilde{\boldsymbol{\Sigma}}\boldsymbol{\alpha})^T\tilde{\boldsymbol{\Sigma}}^{-1}(\mathbf{c}_J-\tilde{\boldsymbol{\Sigma}}\boldsymbol{\alpha})}. \end{aligned}$$

Using this last result, it follows that $f(\mathbf{y}|\boldsymbol{\theta},\sigma^2)\pi(\boldsymbol{\theta}|\sigma^2)g(\sigma^2)$ is given by:

$$\begin{aligned} f(\mathbf{y}|\boldsymbol{\theta},\sigma^2)\pi(\boldsymbol{\theta}|\sigma^2)g(\sigma^2) &= \frac{|\tilde{\boldsymbol{\Sigma}}|^{1/2}\delta^\alpha}{(2\pi)^{N/2}|\boldsymbol{\Sigma}|^{1/2}\Gamma(\alpha)}\left(\frac{1}{\sigma^2}\right)^{\frac{N}{2}+\alpha+1} e^{-\frac{1}{\sigma^2}(\delta+\frac{\mathbf{y}^T\mathbf{y}}{2}+\frac{\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}}{2}-\frac{\boldsymbol{\alpha}^T\tilde{\boldsymbol{\Sigma}}^{-1}\boldsymbol{\alpha}}{2})} \\ &\quad \times \frac{1}{(2\pi)^{M/2}|\tilde{\boldsymbol{\Sigma}}|^{1/2}(\sigma^2)^{M/2}} e^{-\frac{1}{2\sigma^2}(\mathbf{c}_J-\tilde{\boldsymbol{\Sigma}}\boldsymbol{\alpha})^T\tilde{\boldsymbol{\Sigma}}^{-1}(\mathbf{c}_J-\tilde{\boldsymbol{\Sigma}}\boldsymbol{\alpha})}. \end{aligned}$$

Define:

$$\delta^* = \frac{\mathbf{y}^T\mathbf{y}}{2} + \frac{\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}}{2} - \frac{\boldsymbol{\alpha}^T\tilde{\boldsymbol{\Sigma}}^{-1}\boldsymbol{\alpha}}{2}, \tag{5.19}$$

$$\alpha^* = \frac{N}{2} + \alpha. \tag{5.20}$$

Thus:

$$\begin{aligned}
f(\mathbf{y}|\mathbf{c}_J,\sigma^2)\pi(\mathbf{c}_J|\sigma^2)g(\sigma^2) &= \frac{|\tilde{\boldsymbol{\Sigma}}|^{1/2}\delta^\alpha}{(2\pi)^{N/2}|\boldsymbol{\Sigma}|^{1/2}\Gamma(\alpha)}\left(\frac{1}{\sigma^2}\right)^{\alpha^*+1} e^{-\frac{1}{\sigma^2}(\delta+\delta^*)} \\
&\quad \times \frac{1}{(2\pi)^{M/2}|\tilde{\boldsymbol{\Sigma}}|^{1/2}(\sigma^2)^{M/2}} e^{-\frac{1}{2\sigma^2}(\mathbf{c}_J-\tilde{\boldsymbol{\Sigma}}\boldsymbol{\alpha})^T\tilde{\boldsymbol{\Sigma}}^{-1}(\mathbf{c}_J-\tilde{\boldsymbol{\Sigma}}\boldsymbol{\alpha})}, \\
&= \frac{|\tilde{\boldsymbol{\Sigma}}|^{1/2}\delta^\alpha\Gamma(\alpha^*)}{(2\pi)^{N/2}|\boldsymbol{\Sigma}|^{1/2}\Gamma(\alpha)(\delta+\delta^*)^{\alpha^*}} \\
&\quad \times \left[\frac{(\delta+\delta^*)^{\alpha^*}}{\Gamma(\alpha^*)}\left(\frac{1}{\sigma^2}\right)^{\alpha^*+1} e^{-\frac{1}{\sigma^2}(\delta+\delta^*)}\right] \\
&\quad \times \frac{1}{(2\pi)^{M/2}|\tilde{\boldsymbol{\Sigma}}|^{1/2}(\sigma^2)^{M/2}} e^{-\frac{1}{2\sigma^2}(\mathbf{c}_J-\tilde{\boldsymbol{\Sigma}}\boldsymbol{\alpha})^T\tilde{\boldsymbol{\Sigma}}^{-1}(\mathbf{c}_J-\tilde{\boldsymbol{\Sigma}}\boldsymbol{\alpha})}, \\
&= \frac{|\tilde{\boldsymbol{\Sigma}}|^{1/2}\delta^\alpha\Gamma(\alpha^*)}{(2\pi)^{N/2}|\boldsymbol{\Sigma}|^{1/2}\Gamma(\alpha)(\delta+\delta^*)^{\alpha^*}} \\
&\quad \times \mathcal{IG}_{\sigma^2}\left(\alpha^*\,,\,\delta+\delta^*\right)\mathcal{N}_{\mathbf{c}_J}\left(\tilde{\boldsymbol{\Sigma}}\boldsymbol{\alpha}\,,\,\sigma^2\tilde{\boldsymbol{\Sigma}}\right).
\end{aligned}$$

Note that the last result implies that:

$$m(\mathbf{y}) = \frac{|\tilde{\boldsymbol{\Sigma}}|^{1/2}\delta^\alpha\Gamma(\alpha^*)}{(2\pi)^{N/2}|\boldsymbol{\Sigma}|^{1/2}\Gamma(\alpha)(\delta+\delta^*)^{\alpha^*}}, \tag{5.21}$$

where the rhs of the last expression depends on $\mathbf{y}$ through the parameter δ^*.

Therefore, it follows that conditionally on the observation of $\mathbf{y}$, the pair $[\mathbf{c}_J, \sigma^2]$ has distribu-

tion given by:

$$\pi(\mathbf{c}_J, \sigma^2|\mathbf{y}) = \mathcal{NIG}\left(\alpha^*, \beta^*, \boldsymbol{\mu}^*, \tilde{\boldsymbol{\Sigma}}\right), \tag{5.22}$$

where:

$$\begin{aligned}
\alpha^* &= \alpha + \frac{N}{2}, && (5.23)\\
\beta^* &= \delta + \frac{\mathbf{y}^T\mathbf{y}}{2} + \frac{\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}}{2} - \frac{\boldsymbol{\alpha}^T\tilde{\boldsymbol{\Sigma}}^{-1}\boldsymbol{\alpha}}{2}, && (5.24)\\
\boldsymbol{\mu}^* &= \tilde{\boldsymbol{\Sigma}}\boldsymbol{\alpha}, && (5.25)\\
\tilde{\boldsymbol{\Sigma}} &= (\boldsymbol{\Sigma}^{-1} + \tilde{\tilde{\boldsymbol{\Psi}}}^T\tilde{\tilde{\boldsymbol{\Psi}}})^{-1}. && (5.26)
\end{aligned}$$

Now, using the last results, $\pi(\mathbf{c}_J|\mathbf{y})$ is given by:

$$\pi(\mathbf{c}_J|\mathbf{y}) = \frac{(\delta+\delta^*)^{\frac{N}{2}+\alpha}\Gamma(\frac{M}{2}+\alpha^*)}{\Gamma(\frac{N}{2}+\alpha)(2\pi)^{M/2}|\tilde{\boldsymbol{\Sigma}}|^{1/2}(\delta+\delta^*+\frac{1}{2}h(\mathbf{c}_J))^{\frac{M}{2}+\alpha^*}}, \tag{5.27}$$

where $h(\mathbf{c}_J) = (\mathbf{c}_J - \tilde{\boldsymbol{\Sigma}}\boldsymbol{\alpha})^T\tilde{\boldsymbol{\Sigma}}^{-1}(\mathbf{c}_J - \tilde{\boldsymbol{\Sigma}}\boldsymbol{\alpha}))$, and $M = p \cdot 2^J$. Here, the dependence on the observed vector $\mathbf{y}$ is given by the parameter δ^* defined in Eq.(5.19). Furthermore, since the matrix $\tilde{\boldsymbol{\Sigma}}$ is symmetric positive semi-definite, it follows that $\forall\ \mathbf{w} \in \mathbb{R}^N$, $\mathbf{w}^T\tilde{\boldsymbol{\Sigma}}\mathbf{w} \geq 0$.

5.2.2 Connection between posterior distribution $\pi(\mathbf{c}_J|\mathbf{y})$ and the Multivariate $t-$ distribution

Suppose $\mathbf{w} \in \mathbb{R}^N$. Then, it is said that $\mathbf{w}$ follows a multivariate $t-$ distribution with ν degrees of freedom and parameters $\boldsymbol{\mu} \in \mathbb{R}^N$ and $\boldsymbol{\Sigma} \in \mathbb{R}^{N\times N}$, (symmetric positive definite), if its density function is given by:

$$\pi(\mathbf{w}|\nu, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{\Gamma(\frac{\nu+N}{2})}{\Gamma(\frac{\nu}{2})\nu^{N/2}\pi^{N/2}|\boldsymbol{\Sigma}|^{1/2}}\left(1 + \frac{1}{\nu}(\mathbf{w}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{w}-\boldsymbol{\mu})\right)^{-(\frac{\nu+N}{2})} \tag{5.28}$$

Therefore, it is possible to observe that after some algebraic manipulations, it follows:

$$[\mathbf{c}|\mathbf{y}] \sim t_{2\alpha^*}\left(\tilde{\boldsymbol{\Sigma}}\boldsymbol{\alpha}\ ,\ \frac{\delta+\delta^*}{\alpha^*}\tilde{\boldsymbol{\Sigma}}\right).$$

It is a well known result in statistics that, under aforementioned probability model it follows that:

$$\begin{aligned}\mathbb{E}[\mathbf{w}|\nu,\boldsymbol{\mu},\boldsymbol{\Sigma}] &= \tilde{\boldsymbol{\Sigma}}\boldsymbol{\alpha}, &(5.29)\\ Var(\mathbf{w}|\nu,\boldsymbol{\mu},\boldsymbol{\Sigma}) &= \left(\frac{\nu}{\nu-2}\right)\left(\frac{\delta+\delta^*}{\alpha^*}\right)\tilde{\boldsymbol{\Sigma}}, &(5.30)\end{aligned}$$

where $\nu > 2$. Moreover, since the multivariate $t-$ distribution is elliptical, it follows that $Mode(\mathbf{w}|\nu,\boldsymbol{\mu},\boldsymbol{\Sigma}) = \mathbb{E}[\mathbf{w}|\nu,\boldsymbol{\mu},\boldsymbol{\Sigma}]$.

5.2.3 Obtention of the Bayes Estimator $\hat{\mathbf{c}}_J$

Using the results from the last section, given that $[\mathbf{c}|\mathbf{y}] \sim t_{2\alpha^*}\left(\tilde{\boldsymbol{\Sigma}}\boldsymbol{\alpha}\ ,\ \frac{\delta+\delta^*}{\alpha^*}\tilde{\boldsymbol{\Sigma}}\right)$, it follows that: Using the last result, by observing Eq. (5.27), it follows that:

$$\begin{aligned}\mathbf{c}_{\hat{J_{MAP}}} &= \tilde{\boldsymbol{\Sigma}}\boldsymbol{\alpha},\\ &= \left(\boldsymbol{\Sigma}^{-1}+\tilde{\tilde{\boldsymbol{\Psi}}}^T\tilde{\tilde{\boldsymbol{\Psi}}}\right)^{-1}(\tilde{\tilde{\boldsymbol{\Psi}}}^T\mathbf{y}+\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}), &(5.31)\end{aligned}$$

where $\boldsymbol{\Sigma}$ and $\boldsymbol{\mu}$ are hyper-parameters of the prior distribution of the location parameter $[\mathbf{c}_J|\sigma^2]$. Define $\mathbf{c}_J^{LS}$ as the least-squares solution of the system:

$$\min_{\mathbf{c}_J\in\mathbb{R}^{p\cdot 2^J}} ||\mathbf{y}-\tilde{\tilde{\boldsymbol{\Psi}}}\mathbf{c}_J||_2^2.$$

Thus, $\mathbf{c}_J^{LS} = \left(\tilde{\tilde{\boldsymbol{\Psi}}}^T\tilde{\tilde{\boldsymbol{\Psi}}}\right)^{-1}\tilde{\tilde{\boldsymbol{\Psi}}}^T\mathbf{y}$.

Similarly, define $\mathbf{H} = \left(\boldsymbol{\Sigma}^{-1} + \tilde{\tilde{\boldsymbol{\Psi}}}^T \tilde{\tilde{\boldsymbol{\Psi}}}\right)^{-1} \tilde{\tilde{\boldsymbol{\Psi}}}^T \tilde{\tilde{\boldsymbol{\Psi}}}$. Therefore, from Eq.(5.31) it is possible to obtain:

$$\mathbf{c}_{\hat{J_{MAP}}} = \mathbf{H}\mathbf{c}_J^{LS} + (\mathbf{I} - \mathbf{H})\,\boldsymbol{\mu}, \tag{5.32}$$

since $\mathbf{I} - \mathbf{H} = \left(\boldsymbol{\Sigma}^{-1} + \tilde{\tilde{\boldsymbol{\Psi}}}^T \tilde{\tilde{\boldsymbol{\Psi}}}\right)^{-1} \boldsymbol{\Sigma}^{-1}$.

From the last result, it follows that under the $\mathcal{NIG}$ model, the MAP estimator of the empirical wavelet coefficients of the unknown functions is a weighted average of the least-squares estimator, and the coefficients location $\boldsymbol{\mu}$. In particular, when $\boldsymbol{\mu} = \mathbf{0}$, the estimator $\mathbf{c}_{\hat{J_{MAP}}}$ reduces to a ridge-type regularized estimator.

Note that this approach allows the use of efficient numerical algorithms such as `Conjugate Gradient Descent` or `Steepest Descent` which enable fast computations, even in the case of large sample sizes.

5.2.4 Prior Parameter Selection

From the definition of the expansion coefficients in Eq. (5.4), the simplest possible prior selection on the location parameter $\mathbf{c}_J$ is the following:

$$\boldsymbol{\mu} = \mathbf{0}, \tag{5.33}$$

$$\boldsymbol{\Sigma} = \tau^2 \mathbf{I}_{p \cdot 2^J}. \tag{5.34}$$

This selection for the prior parameter $\boldsymbol{\mu}$ results from the fact that typically the wavelet coefficients $\mathbf{c}_j$ are concentrated around zero.

Similarly, choosing $\mathbf{\Sigma} = \tau^2 \mathbf{I}_{p \cdot 2^J}$ assumes that the wavelet coefficients for the model are uncorrelated, with equal variance $\tau^2 > 0$. Even though this last assumption can be argued to be too strong, given the simplicity of the resulting model, it is worth to be analyzed in light of the balance between practical implementation and accuracy of results.

Note that the aforementioned model depends only on the parameter $\tau > 0$; in order to enforce robustness in the estimation process, it is possible to find its optimal value via line-search.

5.2.5 Bayesian Model Implementation

Using the prior selection for the hyper-parameters detailed in Eqs. (5.33), (5.34), the estimator takes the form:

$$\hat{\mathbf{c}_J} = \left(\tilde{\tilde{\mathbf{\Psi}}}^T \tilde{\tilde{\mathbf{\Psi}}} + \tau^{-2} \mathbf{I}_N \right)^{-1} \tilde{\tilde{\mathbf{\Psi}}}^T \mathbf{y}, \tag{5.35}$$

which is the solution of the optimization program:

$$\min_{\mathbf{c}_J \in \mathbb{R}^{p \cdot 2^J}} ||\mathbf{y} - \tilde{\tilde{\mathbf{\Psi}}} \mathbf{c}_J||_2^2 + \tau^{-2} ||\mathbf{c}_J||_2^2,$$

which corresponds to an l_2- regularized least-squares program. Furthermore, from the last equation, it is possible to observe that as $\tau \to \infty$, the solution converges to the usual least-squares problem:

$$\min_{\mathbf{c}_J \in \mathbb{R}^{p \cdot 2^J}} ||\mathbf{y} - \tilde{\tilde{\mathbf{\Psi}}} \mathbf{c}_J||_2^2.$$

Using the singular value decomposition (SVD) of $\tilde{\tilde{\mathbf{\Psi}}}$ we can express it as:

$$\tilde{\tilde{\mathbf{\Psi}}}_{N\times p\cdot 2^J} = \mathbf{U}_{N\times R}\mathbf{\Sigma}_{R\times R}\mathbf{V}^T_{R\times p\cdot 2^J},$$

where, R corresponds to $rank(\tilde{\tilde{\mathbf{\Psi}}}) \leq \min(N, p\cdot 2^J)$. Assuming $\tilde{\tilde{\mathbf{\Psi}}}^T$ is full column rank, it follows that $\mathbf{U}\mathbf{U}^T = \mathbf{U}^T\mathbf{U} = \mathbf{I}_N$. Similarly, it holds that $\mathbf{V}^T\mathbf{V} = \mathbf{I}_N$, and $\mathbf{\Sigma} = diag(\sigma_1, ..., \sigma_N)$ for $\sigma_1 \geq \sigma_2 \geq ... \geq \sigma_N > 0$.

Therefore, the SVD enables to express the first term of the rhs of Eq.(5.35) as follows:

$$\left(\tilde{\tilde{\mathbf{\Psi}}}^T\tilde{\tilde{\mathbf{\Psi}}} + \tau^{-2}\mathbf{I}_N\right)^{-1} \;=\; \left(\mathbf{V}\mathbf{\Sigma}^2\mathbf{V}^T + \tau^{-2}\mathbf{I}_N\right)^{-1}.$$

Using the matrix inversion lemma (i.e. Woodbury matrix identity), and letting $M = p\cdot 2^J$, we obtain:

$$\left(\tilde{\tilde{\mathbf{\Psi}}}^T\tilde{\tilde{\mathbf{\Psi}}} + \tau^{-2}\mathbf{I}_N\right)^{-1} \;=\; \tau^2\mathbf{I}_M - \mathbf{V}\mathbf{S}\mathbf{V}^T,$$

for $\mathbf{S} = diag(\frac{\tau^4\sigma_r^2}{1+\tau^2\sigma_r^2})$, $r = 1, ..., R$. Thus Eq.(5.35) becomes:

$$\hat{\mathbf{c}_J} \;=\; \mathbf{V}\mathbf{D}\mathbf{U}^T\mathbf{y}, \tag{5.36}$$

where:

$$\mathbf{D} \;=\; diag\left(\frac{\tau^2\sigma_r}{1+\tau^2\sigma_r^2}\right),$$

for $r = 1, ..., R$. This last result implies that:

$$\hat{\mathbf{c}_J} \;=\; \sum_{l=1}^{R}\frac{\tau^2\sigma_l}{1+\tau^2\sigma_l^2}\langle\mathbf{y}, \mathbf{u}_l\rangle\cdot\mathbf{v}_l, \tag{5.37}$$

where $\{\mathbf{u}_1, ..., \mathbf{u}_R\}$ and $\{\mathbf{v}_1, ..., \mathbf{v}_R\}$ are the columns of the matrices **U** and **V** respectively. Note that the last expression shows that the empirical estimator for the expansion coefficients corresponds to linear combinations of the columns of the matrix **V**, with linear coefficients given by the weighted projection of the observed response **y** onto the column space of **U**, with weights defined by the singular values and stabilization parameter τ.

5.2.6 Iterative Solution of the Model via Backfitting

If we observe the model defined in Eq.(5.1), the estimation of the wavelet coefficients for each function in the model can be done using an iterative fashion, by using as response the residuals over the corresponding dimension, assuming that the remaining functions are known. This is the idea of backfitting[41], and it can be illustrated as follows:
Define $r_l(x_l) = \mathbf{y}(\mathbf{x}) - \sum_{m \neq l} \hat{f}_m(x_m)$, where $\hat{f}_m(x_m)$ are estimated of the unknown functions in the model. Thus, it follows that:

$$r_l(x_l) \quad = \quad f_l(x_l) + \sigma \cdot \epsilon \;\; l = 1, ..., p.$$

Under this definition, using the hierarchical structure of the Bayesian model as in Eqs.(5.10-5.12), it follows:

$$\mathbf{r}_l | \mathbf{c}_J^{(l)}, \sigma^2 \quad \sim \quad \mathcal{N}(\tilde{\tilde{\mathbf{\Psi}}}_l \mathbf{c}_J^{(l)}, \sigma^2 \mathbb{I}_N), \tag{5.38}$$

$$\mathbf{c}_J^{(l)} | \sigma^2 \quad \sim \quad \mathcal{N}(\boldsymbol{\mu}_l, \sigma^2 \mathbf{\Sigma}_l), \tag{5.39}$$

$$\sigma^2 \quad \sim \quad \mathcal{IG}(\alpha, \delta). \tag{5.40}$$

Using this model and the prior parameter selection presented 5.2.4, from Eq.(5.9), it follows that:

$$\hat{\mathbf{c}}_J^{(l)} \quad = \quad \left(\tilde{\tilde{\mathbf{\Psi}}}_l^T \tilde{\tilde{\mathbf{\Psi}}}_l + \tau^{-2}\mathbf{I}_N\right)^{-1} \tilde{\tilde{\mathbf{\Psi}}}_l^T \mathbf{r}_l.$$

Since the functions in the model are not known, using the backfitting approach we can estimate the model using the following algorithm:

Algorithm 1 Computation of Bayes Estimator for $\mathcal{NIG}$ model

procedure BAYES ESTIMATOR USING $\mathcal{NIG}$ MODEL
  $J = \lfloor \log_2(n/p) \rfloor$, $\mathbf{h}_{filt}$, vanishing moments.
  Compute $\tilde{\tilde{\mathbf{\Psi}}} = [\tilde{\tilde{\mathbf{\Psi}}}_1 \dots \tilde{\tilde{\mathbf{\Psi}}}_p]$ via Daubechies-Lagarias.
  $\mathbf{R} = [\mathbf{r}_1 ... \mathbf{r}_p] = \mathbf{0}_{N\times p}$
  $\hat{\mathbf{c}}_J = [\hat{\mathbf{c}}_J^{(1)} \dots \hat{\mathbf{c}}_J^{(p)}] = \mathbf{0}_{2^J\times p}$
  while $||\hat{\mathbf{y}} - \mathbf{y}||_2/||\mathbf{y}||_2 > \delta$ **do**
    for $l = 1, ..., p$ **do**
      $\mathbf{y}_l = \mathbf{y} - \sum_{m\neq l} \mathbf{r}_m$
      Obtain* $\hat{\mathbf{c}}_J^{(l)} = \left(\tilde{\tilde{\mathbf{\Psi}}}_l^T \tilde{\tilde{\mathbf{\Psi}}}_l + \hat{\tau}^{-2}\mathbf{I}_N\right)^{-1} \tilde{\tilde{\mathbf{\Psi}}}_l^T \mathbf{y}_l$
      $\mathbf{r}_l = \tilde{\tilde{\mathbf{\Psi}}}_l^T \tilde{\tilde{\mathbf{\Psi}}}_l \hat{\mathbf{c}}_J^{(l)}$
      $\mathbf{r}_l = \mathbf{r}_l - \bar{\mathbf{r}}_l$
    $\hat{\mathbf{f}} = \mathbf{R}$ each column corresponds to the estimated functions in the model
    $\hat{\mathbf{y}} = \sum_{l=1}^p \mathbf{r}_l$ estimated response

Remarks

(i) $\hat{\mathbf{c}}_J^{(l)}$ is obtained by solving the system $\left(\tilde{\tilde{\mathbf{\Psi}}}_l^T \tilde{\tilde{\mathbf{\Psi}}}_l + \hat{\tau}^{-2}\mathbf{I}_N\right) \hat{\mathbf{c}}_J^{(l)} = \tilde{\tilde{\mathbf{\Psi}}}_l^T \mathbf{y}_l$ using the method of conjugate gradients.

(ii) $\hat{\tau}$ is obtained via grid-search, choosing the value that minimizes $||\mathbf{y}_l - \mathbf{r}_l(\tau)||_2^2$.

(iii) $\delta > 0$ is a tolerance parameter that controls the number of inner iterations of the algorithms along each of the coordinates $l = 1, ..., p$.

(iv) Since the columns of $\mathbf{R}$ contain the raw estimated functions in the model which contain

some noise, it is possible to apply a linear smoother to improve prediction accuracy. This will be illustrated in the next section via a simulation study.

5.2.7 Simulation Results

In this section, we investigate the performance of Bayesian $\mathcal{NIG}$ model and compare its results with respect to Least Squares estimator previously introduced in section 4.2. The error measure utilized is the ARMSE (Average Root Mean Squared Error) of estimation.

$$AMSE(\hat{f}_l) = \frac{1}{B}\sum_{b=1}^{B} MSE(\hat{f}_{l,b}), \quad MSE(\hat{f}_{l,b}) = \frac{1}{N}\sum_{i=1}^{N}(\hat{f}_{l,b}(x_i) - f_l(x_i))^2$$

For this objective, we choose the same set of exemplary baseline functions used to investigate the performance of the Least Squares estimator (see section 4.3.5).

In this simulation study, we use a `Daubechies` filter with 6 vanishing moments, and sample sizes $N = 512, 1024, 2048, 4096$. The random noise variance for the different trials was set to $\sigma = 0.39$. The predictors were drawn from a $\mathcal{U}[0,1]^p$ distribution. The predictions were obtained at 100 equally spaced points along each of the model dimensions.

The following tables (5.1)-(5.4), illustrate the estimation results for this approach:

Table 5.1: ARMSE comparison between Bayes estimator and Least Squares, for $B = 50$ replications, Daubechies 6 filter and $\sigma = 0.39$.

$N = 512$	$f_1(x)$	$f_2(x)$	$f_3(x)$	$f_4(x)$	$f_5(x)$	$f_6(x)$	$f_7(x)$	$f_8(x)$	$f_9(x)$
Bayes	0.01664	0.01402	0.01245	0.01221	0.01366	0.01108	0.35583	0.01721	0.02174
Least Squares	0.01594	0.01684	0.01524	0.01562	0.01698	0.01468	0.36407	0.02105	0.02581

Table 5.2: ARMSE comparison between Bayes estimator and Least Squares, for $B = 50$ replications, Daubechies 6 filter and $\sigma = 0.39$.

$N = 1024$	$f_1(x)$	$f_2(x)$	$f_3(x)$	$f_4(x)$	$f_5(x)$	$f_6(x)$	$f_7(x)$	$f_8(x)$	$f_9(x)$
Bayes	0.00412	0.00190	0.00211	0.00198	0.00174	0.00205	0.22512	0.00630	0.00809
Least Squares	0.00384	0.00396	0.00382	0.00377	0.00379	0.00398	0.06580	0.00456	0.00719

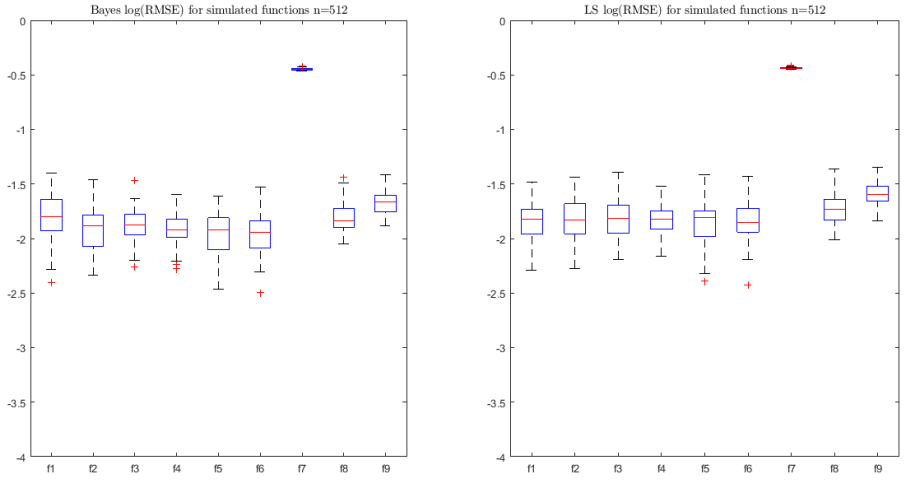
Table 5.3: ARMSE comparison between Bayes estimator and Least Squares, for $B = 50$ replications, Daubechies 6 filter and $\sigma = 0.39$.

$N = 2048$	$f_1(x)$	$f_2(x)$	$f_3(x)$	$f_4(x)$	$f_5(x)$	$f_6(x)$	$f_7(x)$	$f_8(x)$	$f_9(x)$
Bayes	0.00187	0.00055	0.00063	0.00052	0.00049	0.00049	0.22595	0.00502	0.00691
Least Squares	0.00169	0.00156	0.00162	0.00167	0.00163	0.00165	0.06373	0.002486	0.00497

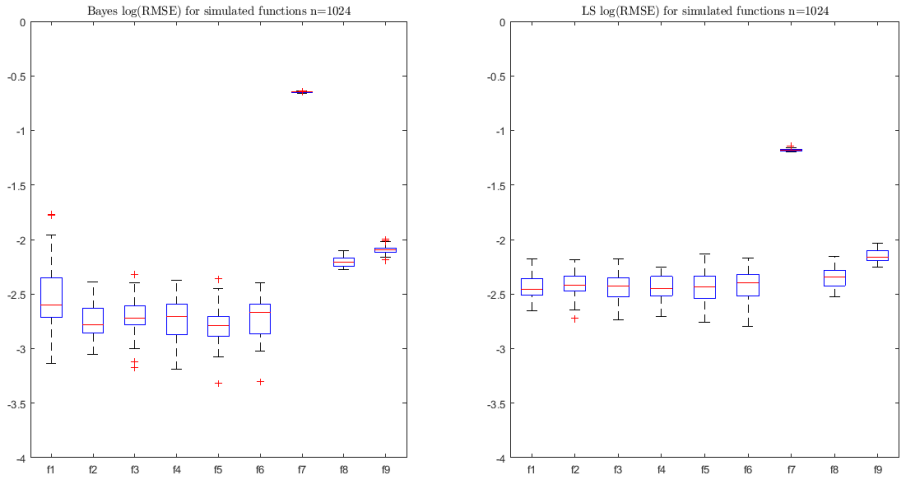
Similarly, in Figs.(5.1)-(5.2) show RMSE boxplots for $B = 50$ replications of the estimation process:

Table 5.4: ARMSE comparison between Bayes estimator and Least Squares, for $B = 50$ replications, Daubechies 6 filter and $\sigma = 0.39$.

$N = 4096$	$f_1(x)$	$f_2(x)$	$f_3(x)$	$f_4(x)$	$f_5(x)$	$f_6(x)$	$f_7(x)$	$f_8(x)$	$f_9(x)$
Bayes	0.000895	0.00027	0.00031	0.00025	0.00019	0.00022	0.22534	0.00481	0.00663
Least Squares	0.00061	0.00061	0.00062	0.00063	0.00063	0.00062	0.00639	0.00065	0.00203



(a) $N = 512$ samples



(b) $N = 1024$ samples

Figure 5.1: Estimation result box-plots for the $\log_{10}(ARMSE)$ computed for both Bayes and least squares procedures, using $B = 50$ replications, for each of the testing functions $f_1(x), ..., f_9(x)$.

The following Figures illustrate the estimation performance of this approach:

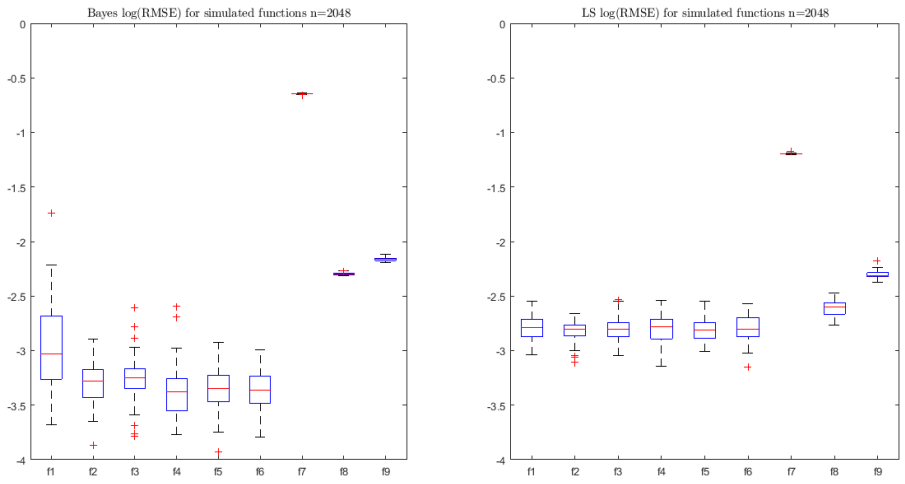
Remarks

(i) As it can be observed in Figs.(5.1) and (5.2), the Bayesian estimates show a slightly higher variance than the estimates obtained via Least Squares. However, this behavior is compensated with its superior performance in the MSE sense. On average, the Bayesian approach show a reduction of nearly 50% in the MSE of estimation when compared to Least Squares.

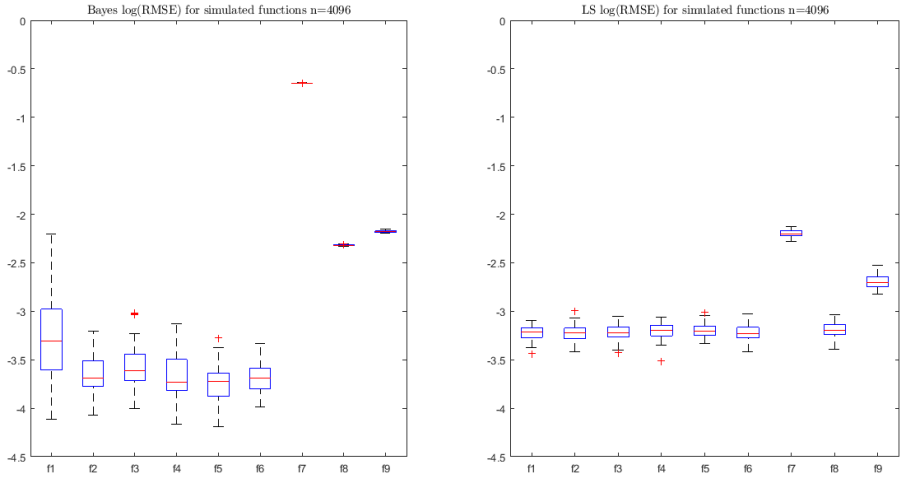
(ii) In addition, as the sample size increases, it is possible to observe the reduction of the MSE of estimation.

(iii) During the implementation, it was observed that it is possible to further improve the estimator performance by smoothing each of the functions via the application of a local linear smoother such as `lowess`. This is due to the fact that the recovered expansion coefficients exhibit sometimes a noisy behavior (depending on the observed sample) which inflates the variance of the estimates. For this reason, applying a denoising scheme is beneficial.

(iv) An additional approach to denoise the estimated expansion coefficients could also be the use of wavelet shrinkage. In fact, the expansion coefficients can be interpreted as the resulting wavelet decomposition of an unknown function observed at equally spaced points. Through the application of the procedure proposed by Donoho et al. (1995)[26], the obtained coefficients can be smoothed, leading to a more parsimonious representation of the recovered function.

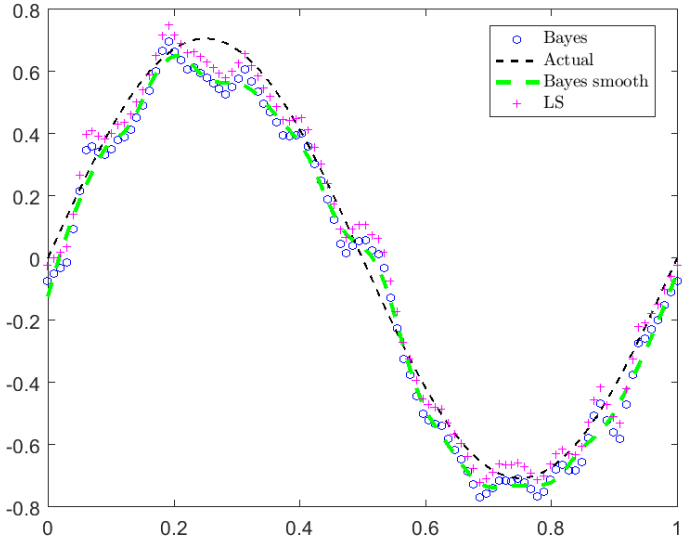


(a) $N = 2048$ samples

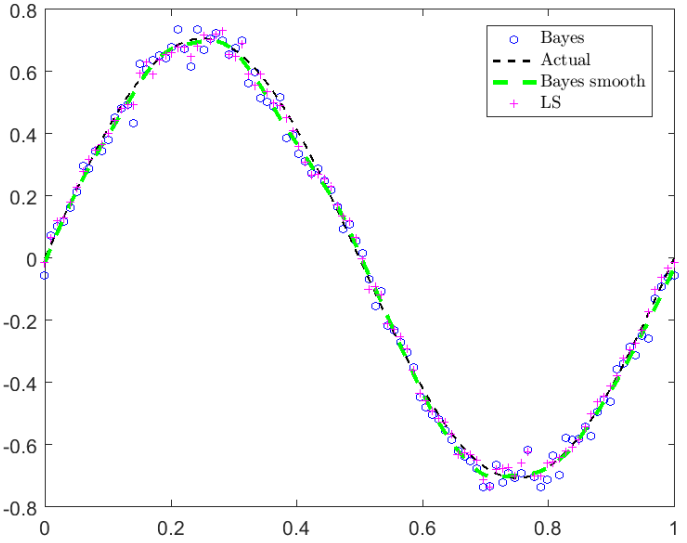


(b) $N = 4096$ samples

Figure 5.2: Estimation result box-plots for the $\log_{10}(ARMSE)$ computed for both Bayes and least squares procedures, using $B = 50$ replications, for each of the testing functions $f_1(x), ..., f_9(x)$.

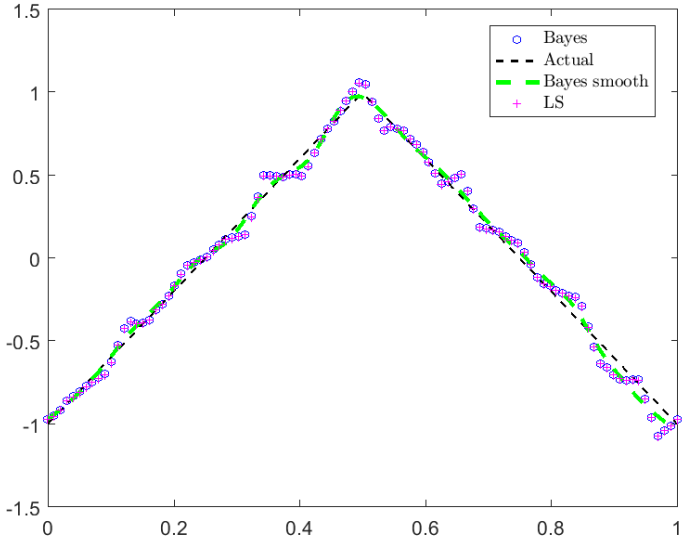


(a) $f_1(x)$, $N = 1024$

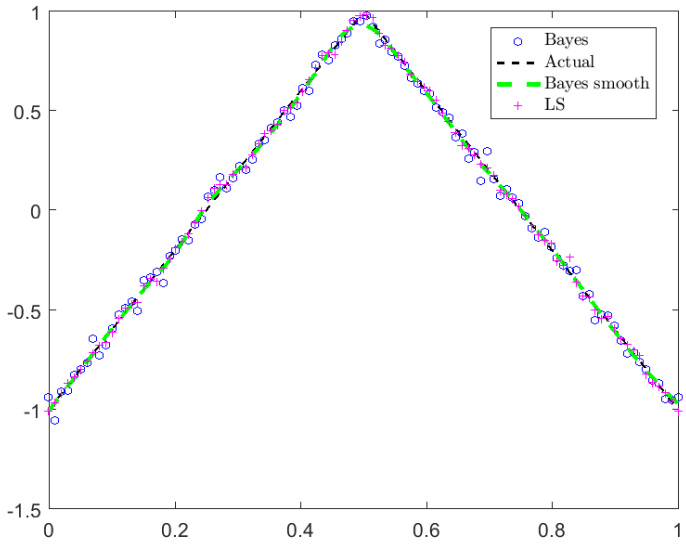


(b) $f_1(x)$, $N = 4096$

Figure 5.3: Estimated function $f_1(x)$ for $N = 1024, 4096$ samples.

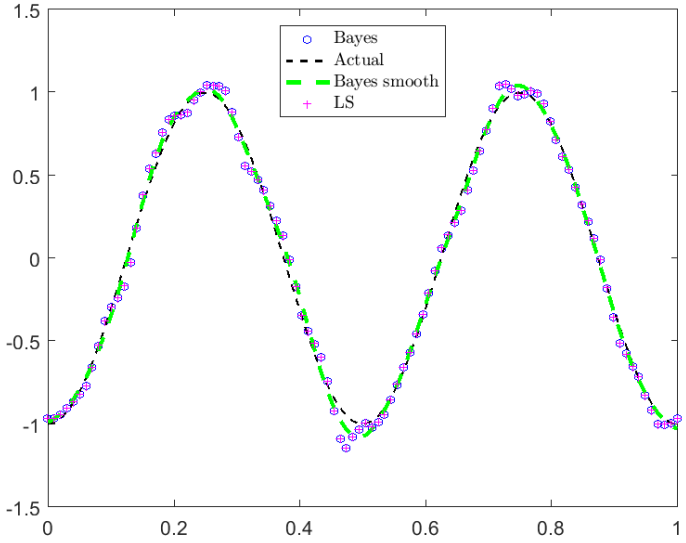


(a) $f_2(x)$, $N = 1024$

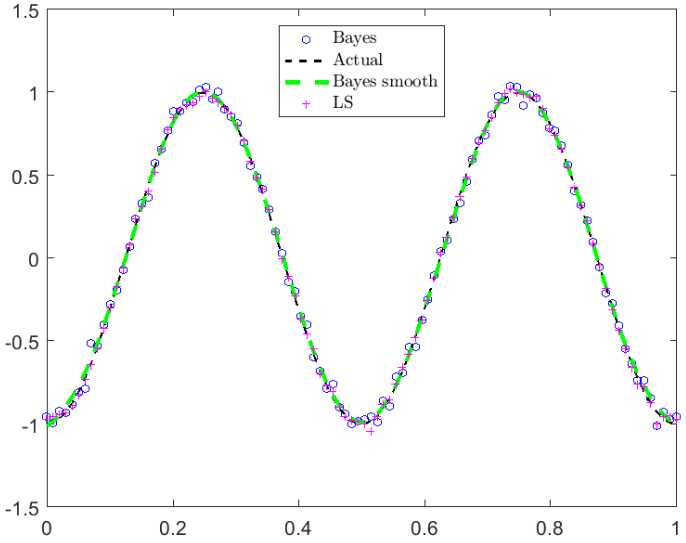


(b) $f_2(x)$, $N = 4096$

Figure 5.4: Estimated function $f_2(x)$ for $N = 1024, 4096$ samples.

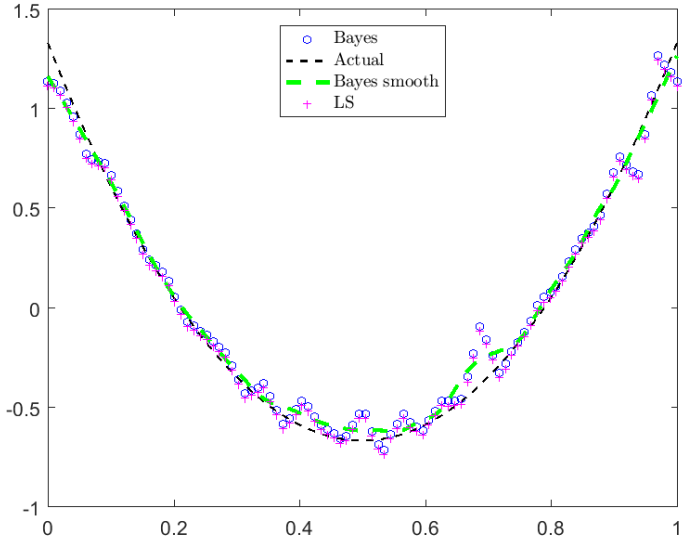


(a) $f_3(x)$, $N = 1024$

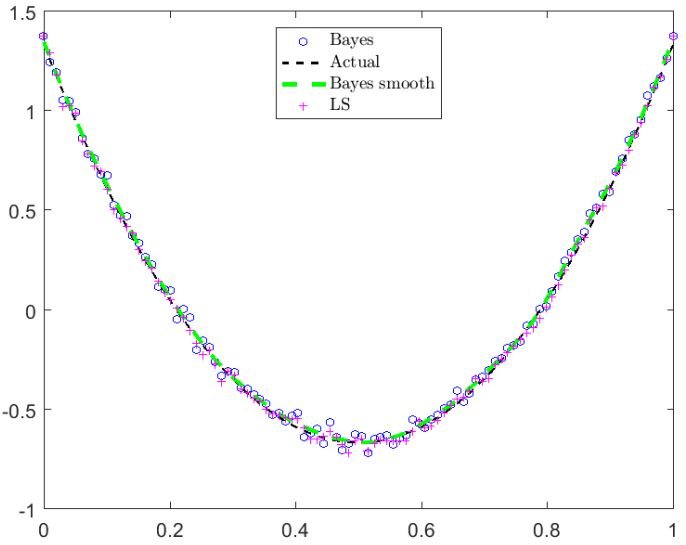


(b) $f_3(x)$, $N = 4096$

Figure 5.5: Estimated function $f_3(x)$ for $N = 1024, 4096$ samples.

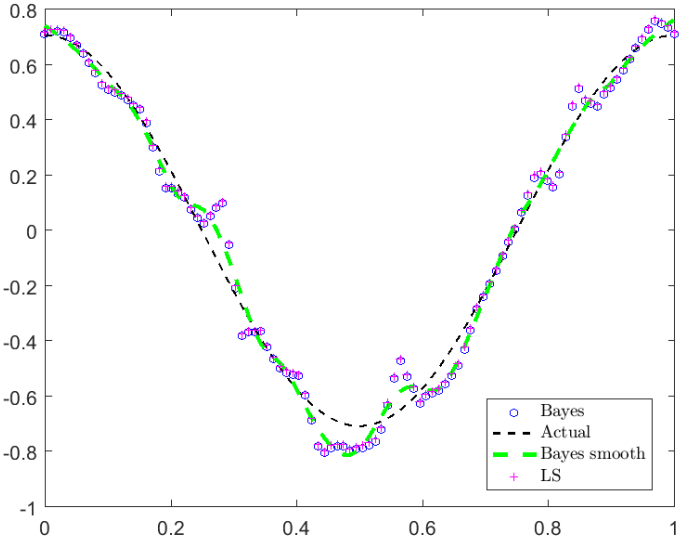


(a) $f_4(x)$, $N = 1024$

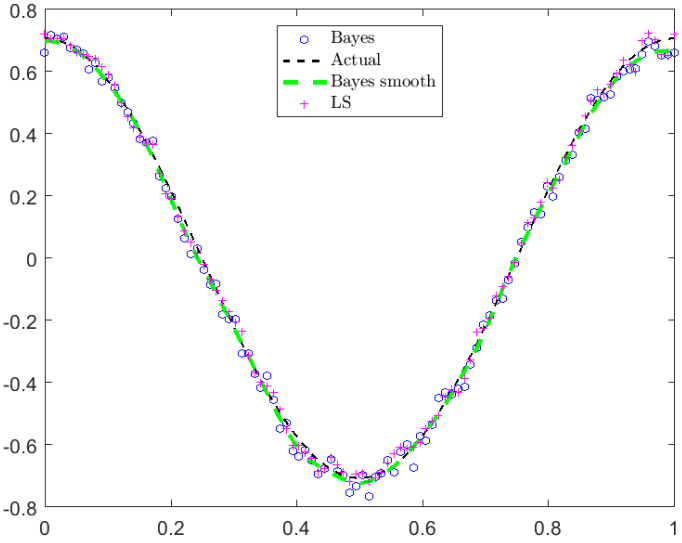


(b) $f_4(x)$, $N = 4096$

Figure 5.6: Estimated function $f_4(x)$ for $N = 1024, 4096$ samples.

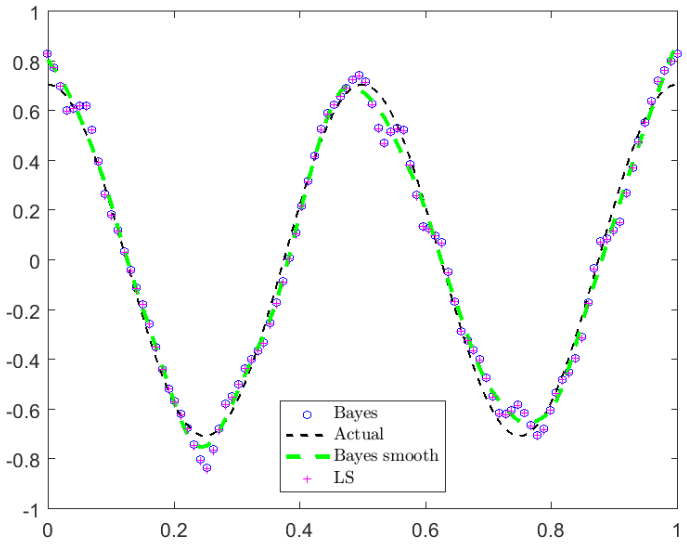


(a) $f_5(x)$, $N = 1024$

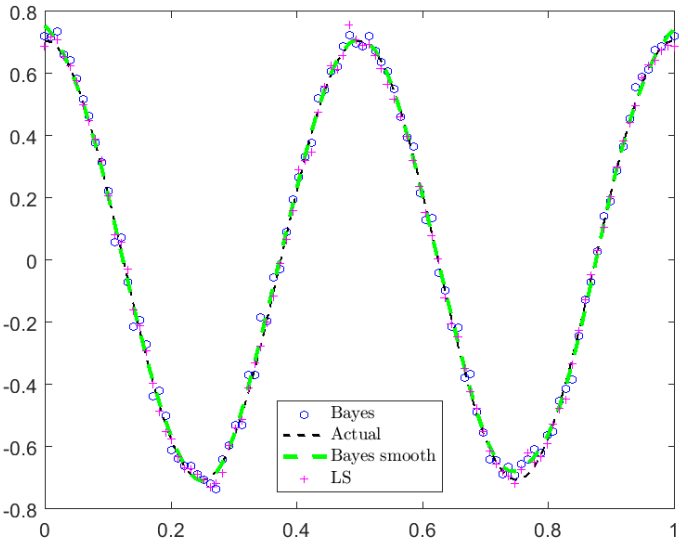


(b) $f_5(x)$, $N = 4096$

Figure 5.7: Estimated function $f_5(x)$ for $N = 1024, 4096$ samples.

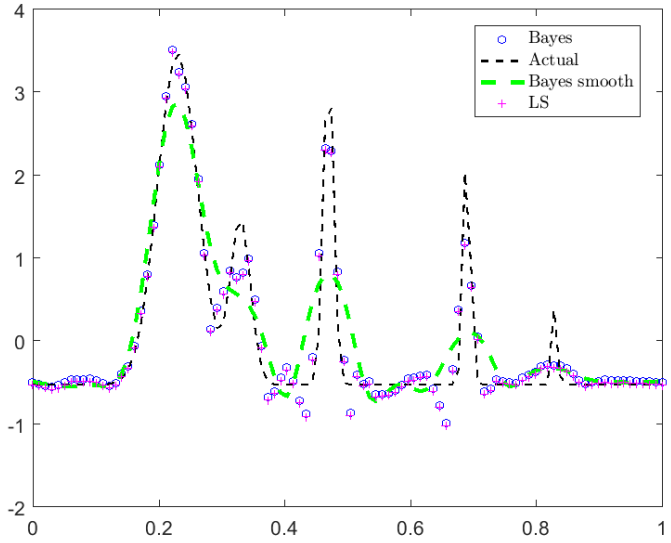


(a) $f_6(x)$, $N = 1024$

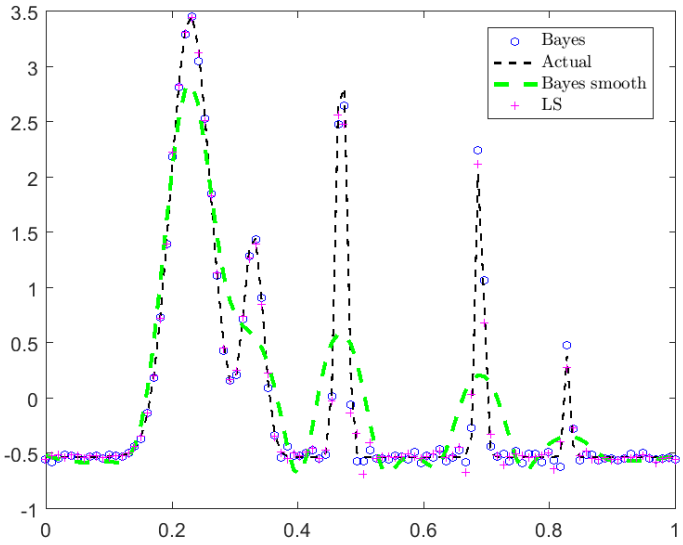


(b) $f_6(x)$, $N = 4096$

Figure 5.8: Estimated function $f_6(x)$ for $N = 1024, 4096$ samples.

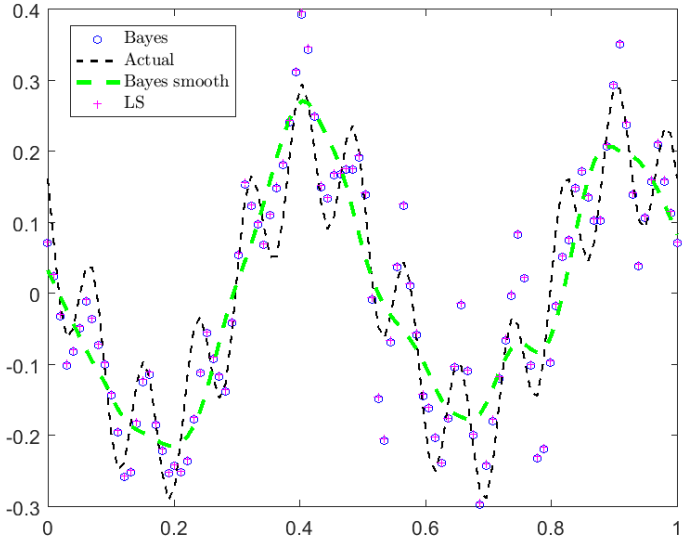


(a) $f_7(x)$, $N = 1024$

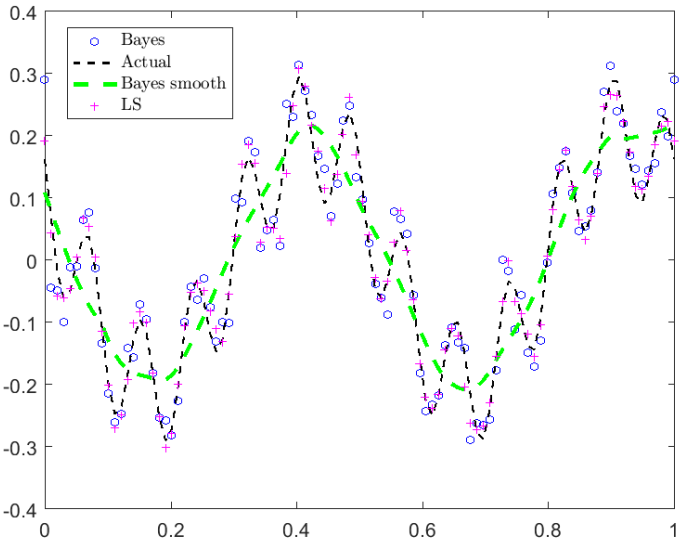


(b) $f_7(x)$, $N = 4096$

Figure 5.9: Estimated function $f_7(x)$ for $N = 1024, 4096$ samples.

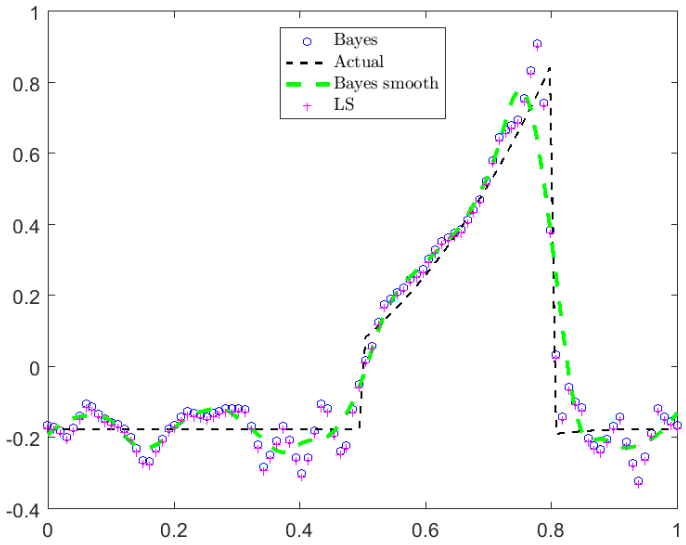


(a) $f_8(x)$, $N = 1024$

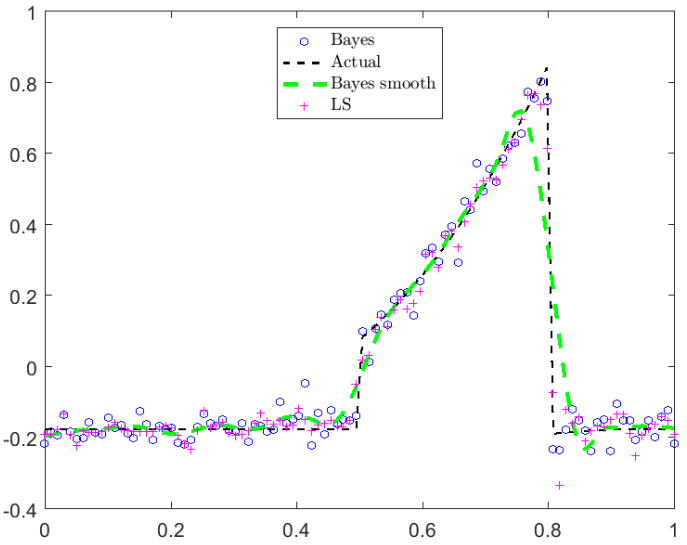


(b) $f_8(x)$, $N = 4096$

Figure 5.10: Estimated function $f_8(x)$ for $N = 1024, 4096$ samples.



(a) $f_9(x)$, $N = 1024$



(b) $f_9(x)$, $N = 4096$

Figure 5.11: Estimated function $f_9(x)$ for $N = 1024, 4096$ samples.

5.3 Bayesian Estimation using a Mixture $\mathcal{NIG}$ Model.

In the context of and non-linear additive regression model, suppose a hierarchical structure of the form:

$$\begin{aligned}
\mathbf{y}|\mathbf{c}_J, \sigma^2 &\sim \mathcal{N}(\tilde{\tilde{\mathbf{\Psi}}}\mathbf{c}_J, \sigma^2\mathbb{I}_N), && (5.41)\\
\mathbf{c}_J|\sigma^2, \gamma &\sim \gamma\mathcal{N}(\boldsymbol{\mu}, \sigma^2\mathbf{\Sigma}) + (1-\gamma)\mathcal{N}(\boldsymbol{\mu}, \sigma^2\mathbf{\Delta}), && (5.42)\\
\sigma^2 &\sim \mathcal{IG}(\alpha, \delta), && (5.43)\\
\gamma &\sim Bernoulli(q). && (5.44)
\end{aligned}$$

Here, $0 < q < 1$, $\mathbf{\Sigma} = \lambda^2\mathbf{I}$, $\mathbf{\Delta} = \tau^2\mathbf{I}$, and $\boldsymbol{\mu} = \mathbf{0}$. The motivation for choosing this kind of hierarchical model has to do with its flexibility to model functions with different proportions of large wavelet and small coefficients, which depends on the function smoothness. This makes this modeling strategy suitable to adapt better to models defined by functions that differ in their smoothness degree, while exploiting the conjugacy of the $\mathcal{NIG}$ model to obtain closed form solutions for the Bayes estimator (under squared error loss).

The first component in the model defined by Eq.(5.42) corresponds to a spread distribution that models large coefficients (i.e. $\lambda \gg 1$), whereas the second component describes small magnitude coefficients, hence τ is small.

Similarly, the term γ models the proportion for large and small coefficients in the wavelet expansion. This proposed model can be interpreted as an extension of the model proposed by Vidakovic and Canditiis (2001).

5.3.1 Derivation of the Bayes Estimator and Shrinkage Rule

Using results from section 5.2.1, and the model defined in Eqs.(5.41)-(5.44), it is possible to obtain:

$$\begin{aligned}\pi(\mathbf{c}_J, \sigma^2|\mathbf{y}) &= \frac{1}{m(\mathbf{y})} q \left(f(\mathbf{y}|\mathbf{c}_J, \sigma^2)\pi(\mathbf{c}_J|\sigma^2, \gamma = 1)g(\sigma^2) \right) + \\ &\quad \frac{1}{m(\mathbf{y})}(1-q) \left(f(\mathbf{y}|\mathbf{c}_J, \sigma^2)\pi(\mathbf{c}_J|\sigma^2, \gamma = 0)g(\sigma^2) \right).\end{aligned}$$

Note that each of the terms in the above expression can be exactly matched to Eqs.(5.14)-(5.21). Therefore, following the same procedure used to obtain those equations, it is possible to define:

$$\alpha^* = \alpha + \frac{N}{2}, \tag{5.45}$$

$$\boldsymbol{\alpha} = \tilde{\tilde{\boldsymbol{\Psi}}}^T \mathbf{y}, \tag{5.46}$$

$$\tilde{\boldsymbol{\Sigma}} = \left(\boldsymbol{\Sigma}^{-1} + \tilde{\tilde{\boldsymbol{\Psi}}}^T \tilde{\tilde{\boldsymbol{\Psi}}} \right)^{-1}, \tag{5.47}$$

$$\tilde{\boldsymbol{\Delta}} = \left(\boldsymbol{\Delta}^{-1} + \tilde{\tilde{\boldsymbol{\Psi}}}^T \tilde{\tilde{\boldsymbol{\Psi}}} \right)^{-1}, \tag{5.48}$$

$$\delta^{A*} = \frac{1}{2} \left(||\mathbf{y}||_2^2 - \boldsymbol{\alpha}^T \tilde{\boldsymbol{\Sigma}} \boldsymbol{\alpha} \right), \tag{5.49}$$

$$\delta^{B*} = \frac{1}{2} \left(||\mathbf{y}||_2^2 - \boldsymbol{\alpha}^T \tilde{\boldsymbol{\Delta}} \boldsymbol{\alpha} \right). \tag{5.50}$$

Furthermore, it is possible to obtain:

$$\begin{aligned} m(\mathbf{y}) &= q \cdot \frac{|\tilde{\boldsymbol{\Sigma}}|^{1/2} \delta^{\alpha} \Gamma(\alpha^*)}{(2\pi)^{N/2} |\boldsymbol{\Sigma}|^{1/2} \Gamma(\alpha)(\delta + \delta^{A*})^{\alpha^*}} \\ &\quad + (1-q) \cdot \frac{|\tilde{\boldsymbol{\Delta}}|^{1/2} \delta^{\alpha} \Gamma(\alpha^*)}{(2\pi)^{N/2} |\boldsymbol{\Delta}|^{1/2} \Gamma(\alpha)(\delta + \delta^{B*})^{\alpha^*}}. \end{aligned} \tag{5.51}$$

Similarly, the pair $[\mathbf{c}_J, \sigma^2|\mathbf{y}]$ has posterior distribution given by:

$$\begin{aligned}[\mathbf{c}_J, \sigma^2|\mathbf{y}] \;\sim\; & \frac{q \cdot A}{q \cdot A + (1-q) \cdot B} \mathcal{NIG}\left(\alpha^*, \delta^{A*}, \tilde{\boldsymbol{\Sigma}}\boldsymbol{\alpha}, \tilde{\boldsymbol{\Sigma}}\right) \\ & + \frac{(1-q) \cdot B}{q \cdot A + (1-q) \cdot B} \mathcal{NIG}\left(\alpha^*, \delta^{B*}, \tilde{\boldsymbol{\Delta}}\boldsymbol{\alpha}, \tilde{\boldsymbol{\Delta}}\right), \end{aligned} \tag{5.52}$$

where:

$$\begin{aligned} A &= \frac{|\tilde{\boldsymbol{\Sigma}}|^{1/2} \delta^{\alpha} \Gamma(\alpha^*)}{(2\pi)^{N/2} |\boldsymbol{\Sigma}|^{1/2} \Gamma(\alpha) (\delta + \delta^{A*})^{\alpha^*}}, \\ B &= \frac{|\tilde{\boldsymbol{\Delta}}|^{1/2} \delta^{\alpha} \Gamma(\alpha^*)}{(2\pi)^{N/2} |\boldsymbol{\Delta}|^{1/2} \Gamma(\alpha) (\delta + \delta^{B*})^{\alpha^*}}. \end{aligned}$$

Therefore, if we define:

$$w = \frac{q \cdot A}{q \cdot A + (1-q) \cdot B} = \frac{q}{q + (1-q) \cdot \frac{B}{A}},$$

it follows that:

$$[\mathbf{c}_J, \sigma^2|\mathbf{y}] \;\sim\; w \mathcal{NIG}\left(\alpha^*, \delta^{A*}, \tilde{\boldsymbol{\Sigma}}\boldsymbol{\alpha}, \tilde{\boldsymbol{\Sigma}}\right) + (1-w) \mathcal{NIG}\left(\alpha^*, \delta^{B*}, \tilde{\boldsymbol{\Delta}}\boldsymbol{\alpha}, \tilde{\boldsymbol{\Delta}}\right).$$

This implies:

$$\begin{aligned} \pi(\mathbf{c}_J|\mathbf{y}) &= w \cdot \mathbf{t}_{2\alpha^*}\left(\tilde{\boldsymbol{\Sigma}}\boldsymbol{\alpha}\,,\; \frac{\delta + \delta^{A*}}{\alpha^*} \tilde{\boldsymbol{\Sigma}}\right) \\ & \quad + (1-w) \cdot \mathbf{t}_{2\alpha^*}\left(\tilde{\boldsymbol{\Delta}}\boldsymbol{\alpha}\,,\; \frac{\delta + \delta^{B*}}{\alpha^*} \tilde{\boldsymbol{\Delta}}\right), \end{aligned} \tag{5.53}$$

where $\mathbf{t}_{\nu}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ corresponds to a multivariate t distribution, defined as in Eq.(5.28).

Therefore, under the squared error loss, the Bayes estimator is given by:

$$\begin{aligned}\hat{\mathbf{c}}_J &= \mathbb{E}[\mathbf{c}|\mathbf{y}] = w\cdot\tilde{\boldsymbol{\Sigma}}\boldsymbol{\alpha}+(1-w)\cdot\tilde{\boldsymbol{\Delta}}\boldsymbol{\alpha} \qquad (5.54)\\ &= w\cdot\left(\boldsymbol{\Sigma}^{-1}+\tilde{\tilde{\boldsymbol{\Psi}}}^T\tilde{\tilde{\boldsymbol{\Psi}}}\right)^{-1}\tilde{\tilde{\boldsymbol{\Psi}}}^T\mathbf{y}+(1-w)\left(\boldsymbol{\Delta}^{-1}+\tilde{\tilde{\boldsymbol{\Psi}}}^T\tilde{\tilde{\boldsymbol{\Psi}}}\right)^{-1}\tilde{\tilde{\boldsymbol{\Psi}}}^T\mathbf{y}.\end{aligned}$$

Since it is assumed that $0<q<1$, $\boldsymbol{\Sigma}=\lambda^2\mathbf{I}$, $\boldsymbol{\Delta}=\tau^2\mathbf{I}$, using the same argument that led to Eq.(5.37), the Bayes estimator becomes:

$$\begin{aligned}\hat{\mathbf{c}}_J &= w\cdot\left(\lambda^{-2}\mathbf{I}+\tilde{\tilde{\boldsymbol{\Psi}}}^T\tilde{\tilde{\boldsymbol{\Psi}}}\right)^{-1}\tilde{\tilde{\boldsymbol{\Psi}}}^T\mathbf{y}+(1-w)\left(\tau^{-2}\mathbf{I}+\tilde{\tilde{\boldsymbol{\Psi}}}^T\tilde{\tilde{\boldsymbol{\Psi}}}\right)^{-1}\tilde{\tilde{\boldsymbol{\Psi}}}^T\mathbf{y}, \qquad (5.55)\\ &= \sum_{l=1}^{R}\left(w\frac{\lambda^2\sigma_l}{1+\lambda^2\sigma_l^2}+(1-w)\frac{\tau^2\sigma_l}{1+\tau^2\sigma_l^2}\right)\langle\mathbf{y},\mathbf{u}_l\rangle\cdot\mathbf{v}_l, \qquad (5.56)\end{aligned}$$

where, for $M=p\cdot 2^J$ it follows:

$$\begin{aligned}w &= \frac{q}{q+(1-q)\cdot\frac{B}{A}},\\ &= \frac{q}{q+(1-q)\cdot\frac{|\tilde{\boldsymbol{\Delta}}|^{1/2}|\boldsymbol{\Sigma}|^{1/2}(\delta+\delta^{A*})^{\alpha^*}}{|\tilde{\boldsymbol{\Sigma}}|^{1/2}|\boldsymbol{\Delta}|^{1/2}(\delta+\delta^{B*})^{\alpha^*}}},\\ &= \frac{q}{q+(1-q)\cdot\left(\frac{\tau}{\lambda}\right)^M\frac{\left|\lambda^{-2}\mathbf{I}+\tilde{\tilde{\boldsymbol{\Psi}}}^T\tilde{\tilde{\boldsymbol{\Psi}}}\right|^{1/2}}{\left|\tau^{-2}\mathbf{I}+\tilde{\tilde{\boldsymbol{\Psi}}}^T\tilde{\tilde{\boldsymbol{\Psi}}}\right|^{1/2}}\left(\frac{\delta+\frac{1}{2}\left(||\mathbf{y}||_2^2-||\tilde{\mathbf{D}}_\lambda^{1/2}\mathbf{U}^T\mathbf{y}||_2^2\right)}{\delta+\frac{1}{2}\left(||\mathbf{y}||_2^2-||\tilde{\mathbf{D}}_\tau^{1/2}\mathbf{U}^T\mathbf{y}||_2^2\right)}\right)^{\alpha+N/2}},\end{aligned}$$

where $\tilde{\mathbf{D}}_\tau=diag\left(\frac{\tau^2\sigma_r^2}{1+\tau^2\sigma_r^2}\right)$, and $\tilde{\mathbf{D}}_\lambda=diag\left(\frac{\lambda^2\sigma_r^2}{1+\lambda^2\sigma_r^2}\right)$, $r=1,...,R$.

Assuming that $\tilde{\tilde{\boldsymbol{\Psi}}}$ is full column rank, it follows that $\mathbf{V}\mathbf{V}^T=\mathbf{I}$. Therefore:

$$\left|\tau^{-2}\mathbf{I}+\tilde{\tilde{\boldsymbol{\Psi}}}^T\tilde{\tilde{\boldsymbol{\Psi}}}\right|^{1/2} = \left|\mathbf{V}(\tau^{-2}\mathbf{I}+\mathbf{S}^2)\mathbf{V}^T\right|^{1/2}=\left|\tau^{-2}\mathbf{I}+\mathbf{S}^2\right|^{1/2}=\prod_{m=1}^{M}\tau^{-1}\sqrt{1+\tau^2\sigma_m^2}.$$

Using this result, it follows that the mixing weight w, takes the form:

$$w = \frac{q}{q + (1-q) \cdot \left(\frac{\tau^2}{\lambda^2}\right)^M \prod_{m=1}^{M} \sqrt{\frac{1+\lambda^2\sigma_m^2}{1+\tau^2\sigma_m^2}} \left(\frac{\delta + \frac{1}{2}\left(||\mathbf{y}||_2^2 - ||\tilde{\mathbf{D}}_\lambda^{1/2}\mathbf{U}^T\mathbf{y}||_2^2\right)}{\delta + \frac{1}{2}\left(||\mathbf{y}||_2^2 - ||\tilde{\mathbf{D}}_\tau^{1/2}\mathbf{U}^T\mathbf{y}||_2^2\right)} \right)^{\alpha + N/2}} \quad (5.57)$$

This shows that using the mixture $\mathcal{NIG}$ model, the Bayes estimator of the expansion coefficients lives in the column space of the matrix $\mathbf{V}$, with coefficients that are weighted versions of the orthogonal projections of the observed response $\mathbf{y}$ onto the row space of $\mathbf{U}$, with weights depending on the prior parameters in the model, and the singular values of $\tilde{\tilde{\mathbf{\Psi}}}$.

5.3.2 Selection of Hyper-parameters

In order to propose an estimation procedure that enforces robustness against the wide range of possible functions in a model, the selection of hyper-parameters must depend on the data. In particular, our proposed model requires the specification of the following:

(a) (α, δ), that specify the prior distribution of the prior knowledge of noise variability σ^2.

(b) λ^2 and τ^2, that model the concentration of large and small expansion coefficients, respectively.

(c) q, that models the prior probability that the coefficients have high variance (i.e. high energy).

Selection of Prior Parameters (α, δ)

Following the recommendations proposed in Vidakovic and De Canditiis (2001), since σ^2 is modelled via and $\mathcal{IG}(\alpha, \delta)$ it is possible to set:

$$\frac{\alpha}{\delta + 2} = \underset{0 \le k \le 2^{J-1}}{\text{median}} \frac{|d_{Jk}|}{0.6745}, \quad (5.58)$$

where d_{Jk} are the detail wavelet coefficients resulting from the DWT of the observed response vector $\mathbf{y}$, i.e. $\mathbf{d} = \mathbf{W}\mathbf{y}$, where $\mathbf{W}$ is an orthogonal wavelet matrix.

Assuming that $N = 2^J$ the DWT applied to the data vector $\mathbf{y}$ generates a vector $\mathbf{d} = \mathbf{W}_k \cdot \mathbf{y}$ which has the following structure:

$$\mathbf{d} = [\mathbf{c}^{J-k}; \mathbf{d}^{J-k}; \mathbf{d}^{J-k+1}; ...; \mathbf{d}^{J-2}; \mathbf{d}^{J-1}] \tag{5.59}$$

In the last expression k corresponds to the number of steps in the DWT (usually, $k = J$). Also, it is important to mention that due to the decimated nature of the chosen DWT, the size of the vector $\mathbf{d}$ is also N (as in the original data vector $\mathbf{y}$). In Eq.(5.59), $\mathbf{c}^{J-k}$ corresponds to the smooth coefficients at scale level $J - k$; similarly, $\mathbf{d}^{J-k}$ corresponds to the set of detail coefficients at the scale level $J - k$.

Since there are infinite number of pairs (α, δ) that are a solution of Eq.(5.58), setting $3 \leq \alpha \leq 12$ will allow the estimates to remain within the robust region of the Bayes estimator with respect to α, as shown in [59].

Selection of Variance Parameters $\lambda^2,\ \tau^2$

Here, we consider the recommendation stated in Vidakovic and De Canditiis (2001), in which it is suggested that:

$$\lambda^2 = 3 \max\left\{\mathbf{d}^{J-1}\right\}, \tag{5.60}$$

$$\tau^2 = \max\left\{10^{-6} \max|\mathbf{d}^{J-1}|, \min|\mathbf{d}^{J-1}|\right\}. \tag{5.61}$$

Selection of Mixing Probability q

In this case, it is possible to observe the behavior of the wavelet coefficients $\mathbf{d}$ resulting from the DWT of the observed response $\mathbf{y}$, with respect to a certain threshold. This means, that we can define:

$$q = \frac{1}{2^J} \sum_{k=0}^{2^J-1} \mathbb{1}_{\{|d_{jk}|>\delta\}},$$

where $\mathbb{1}_A$ is the indicator function that has value equal to 1 if A is true, and zero if it is not. Similarly, $\delta > 0$ is a threshold that is properly defined.

Note that since we assume that $\mathbf{d}$ is centered at zero, and normally distributed with variance $\sigma^2\mathbf{I}_N$, it is possible to show that:

$$\mathbb{P}\left[|d_{Jk}| > \delta\right] \leq 2 \cdot e^{-\frac{\delta^2}{2\sigma^2}}.$$

This follows from the application of Markov's inequality and the utilization of the moment generating function of the Normal distribution.

If we want to find a bound that decays linearly with respect to $N = 2^J$, it is possible to show that:

$$\mathbb{P}\left[|d_{Jk}| > \sqrt{2\sigma^2 \log(2^J)}\right] \leq 2^{1-J}.$$

Therefore, the more number of coefficients $|d_{Jk}|$ that are greater than equal to this bound, the more likely the underlying function has wavelet coefficients with locations corresponding to

the spread distribution. Therefore:

$$q = \frac{1}{2^J}\sum_{k=0}^{2^J-1} \mathbb{1}_{\left\{|d_{jk}|>\sqrt{2\hat{\sigma}^2\log(2^J)}\right\}}, \tag{5.62}$$

where $\hat{\sigma}^2 = \underset{0\leq k\leq 2^{J-1}}{\text{median}} \frac{|d_{Jk}|}{0.6745}$.

Remarks

(a) The proposed settings for the prior parameters in the model are aimed to exploit the information contained in the data, enforcing the Empirical Bayes approach.

(b) In addition to enhance a data-driven approach, these prior parameter setting are simple to obtain. In particular, the DWT via Mallat's algorithm has a low computational complexity, which improves this method efficiency.

(c) As an alternative way to specify the values for the parameters $\lambda^2,\ \tau^2$, it is possible to use a grid-search methodology to identify the values that minimize the MSE of estimation. However, this introduces an additional layer of computational complexity in the algorithm that may not be beneficial in light of the potential improvements in the prediction accuracy.

5.4 Bayesian Estimation Using γ-Contaminated $\mathcal{NIG}$ Structures

As was observed in sections 5.2 and 5.3, it is possible to approach the Additive regression problem from a Bayesian perspective, which introduces regularization and shrinkage in the expansion coefficient estimates.

In this section, we introduce an alternative model that enhances the shrinkage procedure as a way provide more adaptive expansion coefficients. This was motivated by the behavior

observed during the implementation of the previous models, stated in remarks 5.2.7, and the methodology proposed in [59].

Consider a model of the form:

$$\begin{aligned}
\mathbf{y}|\mathbf{c}_J, \sigma^2 &\sim \mathcal{N}(\tilde{\tilde{\mathbf{\Psi}}}\mathbf{c}_J, \sigma^2 \mathbb{I}_N), && (5.63)\\
\mathbf{c}_J|\sigma^2, \gamma &\sim \gamma\boldsymbol{\delta}_{(\mathbf{c}_J - \mathbf{0})} + (1-\gamma)\mathcal{N}(\boldsymbol{\mu}, \sigma^2\mathbf{\Sigma}), && (5.64)\\
\sigma^2 &\sim \mathcal{IG}(\alpha, \delta), && (5.65)\\
\gamma &\sim Bernoulli(q). && (5.66)
\end{aligned}$$

This model, is a special case of the mixture $\mathcal{NIG}$ structure introduced in section 5.2. Here, we place a point mass at $\mathbf{0}$ for the expansion coefficients, which enhances the shrinkage in the case of low-energy signals. This point-mass can be interpreted as a degenerate multivariate Gaussian distribution, with location at $\mathbf{0}$ and covariance matrix given by $\epsilon\mathbf{I}_{p\cdot 2^J}$, for $\epsilon \to 0$. In practical terms, this model is expected to enforce sparsity in the estimation, making it more suitable for variable selection.

5.4.1 Derivation of the Estimator and Point-Mass Shrinkage Rule

Using results from section 5.2.1 and the model equations defined in Eqs.(5.63)-(5.66), it is possible to show that under squared error loss, the bayes estimator of the expansion coefficients is given by:

$$\hat{\mathbf{c}}_J^S = (1-w)\left(\mathbf{\Sigma}^{-1} - \tilde{\tilde{\mathbf{\Psi}}}^T\tilde{\tilde{\mathbf{\Psi}}}\right)^{-1}\tilde{\tilde{\mathbf{\Psi}}}^T\mathbf{y}, \quad (5.67)$$

since the posterior distribution of $[\mathbf{c}_J|\mathbf{y}]$ has the form:

$$\pi(\mathbf{c}_J|\mathbf{y}) \;=\; w\cdot\boldsymbol{\delta}_{(\mathbf{c}_J-\mathbf{0})}+(1-w)\mathbf{t}_{2\alpha^*}\left(\tilde{\boldsymbol{\Sigma}}\boldsymbol{\alpha}\;,\;\frac{\delta+\delta^*}{\alpha^*}\tilde{\boldsymbol{\Sigma}}\right) \tag{5.68}$$

where:

$$\begin{aligned}
w &= \frac{q\cdot m_0(\mathbf{y})}{q\cdot m_0(\mathbf{y})+(1-q)\cdot m_1(\mathbf{y})}, && (5.69)\\
m_0(\mathbf{y}) &= \frac{\delta^\alpha\Gamma(\alpha+N/2)}{(2\pi)^{N/2}\Gamma(\alpha)\left(\delta+\frac{1}{2}||\mathbf{y}||_2^2\right)^{\alpha+N/2}}, && (5.70)\\
m_1(\mathbf{y}) &= \frac{|\tilde{\boldsymbol{\Sigma}}|^{1/2}\delta^\alpha\Gamma(\alpha^*)}{(2\pi)^{N/2}|\boldsymbol{\Sigma}|^{1/2}\Gamma(\alpha)\left(\delta+\delta^*\right)^{\alpha+N/2}}, && (5.71)\\
\tilde{\boldsymbol{\Sigma}} &= \left(\boldsymbol{\Sigma}^{-1}+\tilde{\tilde{\boldsymbol{\Psi}}}^T\tilde{\tilde{\boldsymbol{\Psi}}}\right)^{-1}, && (5.72)\\
\delta^* &= \frac{1}{2}||\mathbf{y}||_2^2-\frac{1}{2}\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}-\frac{1}{2}\boldsymbol{\alpha}^T\tilde{\boldsymbol{\Sigma}}^{-1}\boldsymbol{\alpha}, && (5.73)\\
\boldsymbol{\alpha} &= \tilde{\tilde{\boldsymbol{\Psi}}}^T\mathbf{y}+\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}. && (5.74)
\end{aligned}$$

It is clear that the proposed estimator (5.67) is a shrunken version of the least squares estimator $\mathbf{c}_J^{LS}=(\tilde{\tilde{\boldsymbol{\Psi}}}^T\tilde{\tilde{\boldsymbol{\Psi}}})^{-1}\tilde{\tilde{\boldsymbol{\Psi}}}^T\mathbf{y}$. In fact, by letting $\mathbf{H}=\left(\boldsymbol{\Sigma}^{-1}+\tilde{\tilde{\boldsymbol{\Psi}}}^T\tilde{\tilde{\boldsymbol{\Psi}}}\right)^{-1}\tilde{\tilde{\boldsymbol{\Psi}}}^T\tilde{\tilde{\boldsymbol{\Psi}}}$, it follows that:

$$\hat{\mathbf{c}}_J^S \;=\; (1-w)\mathbf{H}\mathbf{c}_J^{LS}.$$

Observe that as $w\to 0$ the estimator converges to the $\mathcal{NIG}$ estimator (which corresponds to an l_2-regularized least squares solution). On the contrary, when $w\to 1$ (meaning that the expansion coefficients correspond to a low energy function), the Bayes estimator $\hat{\mathbf{c}}_J^S$ is close to $\mathbf{0}$.

From Eq.(5.68), it follows that the posterior distribution of the expansion coefficients, given

the data and the model corresponds to a $w-$contaminated multivariate t distribution (as shown in Eq.(5.28) with parameters $\nu = 2\alpha^*$, $\boldsymbol{\mu} = \tilde{\boldsymbol{\Sigma}}\boldsymbol{\alpha}$, and $\boldsymbol{\Sigma} = \tilde{\boldsymbol{\Sigma}}$. Furthermore, from the properties of this distribution, it follows that any linear transformation given by a matrix $\mathbf{A}$ preserves the distributional structure. In fact,

$$\mathbf{A} \cdot [\mathbf{c}_J|\mathbf{y}] \;\; = \;\; w \cdot \mathbf{A}\boldsymbol{\delta}_{(\mathbf{c}_J-\mathbf{0})} + (1-w)\mathbf{A} \cdot \mathbf{t}_{2\alpha^*}\left(\tilde{\boldsymbol{\Sigma}}\boldsymbol{\alpha}\;,\; \frac{\delta+\delta^*}{\alpha^*}\tilde{\boldsymbol{\Sigma}}\right),$$

where the second term on the right is distributed as $\mathbf{t}_{2\alpha^*}\left(\mathbf{A}\tilde{\boldsymbol{\Sigma}}\boldsymbol{\alpha}\;,\; \frac{\delta+\delta^*}{\alpha^*}\mathbf{A}\tilde{\boldsymbol{\Sigma}}\mathbf{A}^T\right)$.

Finally, in the backfitting context, the proposed model for each functional component is defined by the following parameters:

$$\begin{aligned}
w_l &= \frac{q_l}{q_l + (1-q_l)\cdot\frac{m_{1,l}(\mathbf{r}_l)}{m_{0,l}(\mathbf{r}_l)}}, && (5.75)\\
m_{0,l}(\mathbf{r}_l) &= \frac{\delta_l^{\alpha_l}\Gamma(\alpha_l + N/2)}{(2\pi)^{N/2}\Gamma(\alpha_l)\left(\delta_l + \frac{1}{2}||\mathbf{r}_l||_2^2\right)^{\alpha_l+N/2}}, && (5.76)\\
m_{1,l}(\mathbf{r}_l) &= \frac{|\tilde{\boldsymbol{\Sigma}}_l|^{1/2}\delta_l^{\alpha_l}\Gamma(\alpha_l^*)}{(2\pi)^{N/2}|\boldsymbol{\Sigma}_l|^{1/2}\Gamma(\alpha_l)\left(\delta_l + \delta_l^*\right)^{\alpha_l+N/2}}, && (5.77)\\
\tilde{\boldsymbol{\Sigma}}_l &= \left(\boldsymbol{\Sigma}_l^{-1} + \tilde{\tilde{\boldsymbol{\Psi}}}_l^T\tilde{\tilde{\boldsymbol{\Psi}}}_l\right)^{-1}, && (5.78)\\
\delta_l^* &= \frac{1}{2}||\mathbf{r}_l||_2^2 - \frac{1}{2}\boldsymbol{\alpha}_l^T\tilde{\boldsymbol{\Sigma}}_l^{-1}\boldsymbol{\alpha}_l, && (5.79)\\
\boldsymbol{\alpha}_l &= \tilde{\tilde{\boldsymbol{\Psi}}}_l^T\mathbf{r}_l. && (5.80)
\end{aligned}$$

Here, $0 < q_{m,l} < 1$, $\boldsymbol{\Sigma}_l = \lambda_l^2\mathbf{I}_{2^J}$, and $\boldsymbol{\mu}_l = \mathbf{0}$. Similarly, as proposed in section 5.2.6, $r_l(x_l) = \mathbf{y}(\mathbf{x}) - \sum_{m\neq l}\hat{f}_m(x_m)$, where $\hat{f}_m(x_m)$ are estimates of the unknown functions in the model.

As an alternative for the parameter covariance matrix $\boldsymbol{\Sigma}_l$, instead of modeling it as an iden-

tity matrix (assuming independence among the expansion coefficients $\mathbf{c}_J$), it is possible to use Zellner's prior. Indeed, under such approach, it follows that $\mathbf{\Sigma}_l = g \cdot (\tilde{\tilde{\mathbf{\Psi}}}_l^T \tilde{\tilde{\mathbf{\Psi}}}_l)^{-1}$, for $g > 0$ given by possible choices: $n,\ p^2,\ \max\{n, p^2\}$.

This prior is traditionally called Zellner's g-prior in the Bayesian literature due to the use of the constant g introduced by Zellner (1986)[67] in front of Fisher's information matrix $(\tilde{\tilde{\mathbf{\Psi}}}_l^T \tilde{\tilde{\mathbf{\Psi}}}_l)^{-1}$. Its motivation is that, it avoids the specification of a whole prior covariance matrix by using the information matrix as a global scale. Also, it allows for a specification of the constant g in terms of observational units, or empirical bayes, which introduces flexibility in the model fitting (at the expense of computational cost).

5.4.2 Elicitation of Hyper parameters

Suppose a fixed $l \in \{1, ..., p\}$. In order to obtain the model hyper parameters, we follow the same suggestions and approach described in section 5.3.2.

Similarly for the mixture model, this block-shrinkage approach requires the specification of the following:

(a) (α_l, δ_l), that specify the prior distribution of the noise variability σ^2.

(b) In the case of independent expansion coefficients: $\lambda_{m,l}^2$, that models the concentration of non-zero expansion coefficients, for each function component and block. On the other hand, when using Zellner's prior, the constant $g > 0$ needs to be specified.

(c) q_l, that models the prior probability that the coefficients for function f_l are zero.

Selection of Prior Parameters (α, δ)

Following the recommendations proposed in Vidakovic and De Canditiis (2001), since σ^2 is modelled via and $\mathcal{IG}(\alpha_l, \delta_l)$ it is possible to set:

$$\frac{\alpha_l}{\delta_l + 2} = \underset{0 \leq k \leq 2^{J-1}}{\text{median}} \frac{|d_{Jk}^{(l)}|}{0.6745}, \tag{5.81}$$

where d_{Jk} are the detail wavelet coefficients resulting from the DWT of the observed residual vector $\mathbf{r}_l$, i.e. $\mathbf{d} = \mathbf{W} \cdot \mathbf{r}_l$, where $\mathbf{W}$ is an orthogonal wavelet matrix. Since there are infinite number of pairs (α_l, δ_l) that are a solution of Eq.(5.81), setting $3 \leq \alpha_l \leq 12$ will allow the estimates to remain within the robust region of the Bayes estimator with respect to α_l, as shown in [59].

Selection of Variance Parameter λ_l^2

Here, we consider the recommendation stated in Vidakovic and De Canditiis (2001)[59], in which it is suggested that:

$$\lambda_l^2 = 3 \max \left\{ \mathbf{d}^l \right\}. \tag{5.82}$$

Here, $\mathbf{d}^l$ corresponds to the DWT of the residual vector $\mathbf{r}_l$.

Selection of Mixing Probability q_l

In this case, it is possible to observe the behavior of the wavelet coefficients $\mathbf{d}^{(l)}$ resulting from the DWT of the observed residual $\mathbf{r}_l$, with respect to a certain threshold. This means,

that we can define:

$$q_l \quad = \quad \frac{1}{2^J} \sum_{k=0}^{2^J-1} \mathbb{1}_{\left\{|d_k^{(l)}|>\delta_l\right\}},$$

where $\mathbb{1}_A$ is the indicator function that has value equal to 1 if A is true, and zero if it is not. Similarly, $\delta_l > 0$ is a threshold that is properly defined.

Note that since it is possible to assume that $\mathbf{d}^{(l)}$ is centered at zero, and normally distributed with variance $\sigma^2\mathbf{I}_N$, it can be shown that:

$$\mathbb{P}\left[|d_k^{(l)}| > \delta_l\right] \leq 2 \cdot e^{-\frac{\delta_l^2}{2\sigma^2}}.$$

This follows from the application of Markov inequality and the utilization of the moment generating function of the Normal distribution.

If we want to find a bound that decays linearly with respect to $L = 2^J$, it is possible to show that:

$$\mathbb{P}\left[|d_k^{(l)}| > \sqrt{2\sigma^2 \log(2^J)}\right] \leq 2^{1-J}.$$

Therefore, the more number of coefficients $|d_k^{(l)}|$ that are greater than equal to this bound, the more likely the underlying function will have expansion coefficients with locations corresponding to the non-zero values. Therefore:

$$q_l \quad = \quad \frac{1}{2^J} \sum_{k=0}^{2^J-1} \mathbb{1}_{\left\{|d_k^{(l)}|>\sqrt{2\hat{\sigma}^2 \log(2^J)}\right\}}, \tag{5.83}$$

where $\hat{\sigma}^2 = \underset{0\leq k\leq 2^{J-1}}{\text{median}} \frac{|d_k^{(l)}|}{0.6745}$.

5.5 Simulation Study

In this section we investigate the finite sample performances of the proposed Bayesian methodologies via simulation. All the estimators are implemented using MATLAB®, and estimation results are compared with previously published methodologies AMlet (Sardy and Tseng, 2004)[1], and the Wavelet-based Least Squares, introduced in Chapter 4.

For the simulation, we consider standard conditions: the observed response is corrupted by Gaussian additive noise, and the features in the model are uniformly distributed in $[0,1]^p$. The model used for this analysis is given by:

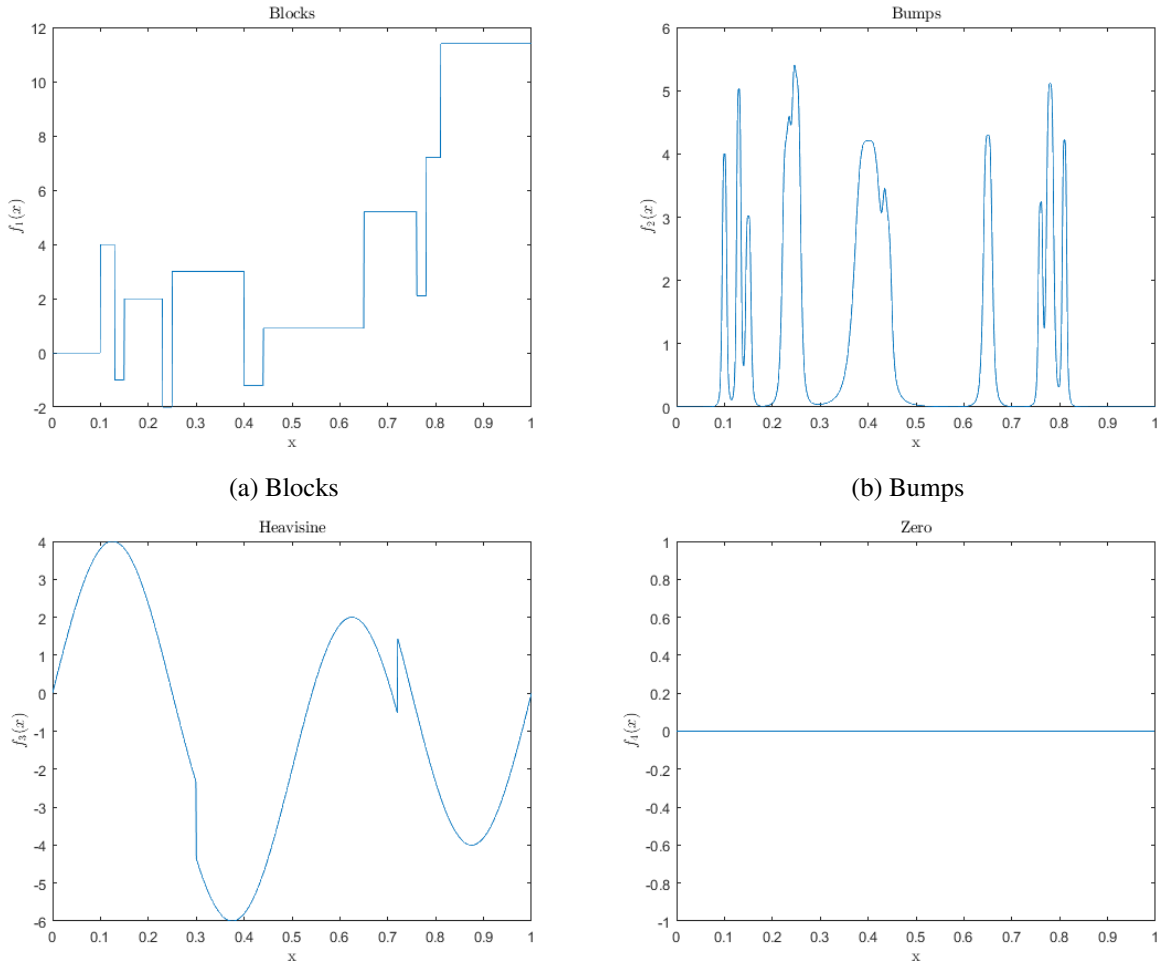
$$y(\mathbf{x}) \quad = \quad f_1(X_1) + f_2(X_2) + f_3(X_3) + f_4(X_4) + \sigma \cdot \epsilon.$$

Here, X_l are independent $\mathcal{U}(0,1)$, and $\epsilon \sim \mathcal{N}(0,1)$. Similarly, each of the functions in the model is given by:

- $f_1(X_1)$ is the piecewise constant `blocks` function (Donoho and Johnstone, 1994)[34].
- $f_2(X_2)$ is the continuous but erratic `bumps` function (Donoho and Johnstone, 1994)[34].
- $f_3(X_3)$ is the relatively smooth `heavisine` function (Donoho and Johnstone, 1994)[34].
- $f_4(X_4)$ is the `zero` function (Sardy and Tseng, 2004)[1], representing a non-significant feature.

In order to adjust the simulation to the settings used by Sardy and Tseng (2004)[1], and obtain results that can be compared against those published in [1], the non-zero functions are scaled and centered to have a standard error equal to 3:

$$\int_0^1 (f_l(x) - \bar{f})^2 dx = 3^2, \quad \bar{f} = \int_0^1 f_l(x) dx.$$



(a) Blocks

(b) Bumps

(c) Heavisine

(d) Zero

Figure 5.12: Functions used in the simulated additive model.

The standard deviation of the additive noise $\sigma = 0.05$, enabling a large signal-to-noise ratio of the observed response. The wavelet filter used for the expansion is the `Daubechies` with 8 vanishing moments. The multiresolution index for the expansion is set as $J = \log_2(N)$. The estimator performance is measured using the $AMSE$, similarly as in section 5.2.7. Here, we set $B = 50$ and $N = 512,\ 1024,\ 2048,\ 4096,\ 8192$.

The simulation results are displayed by the following instances:

(i) Figures 5.13 to 5.20 illustrate typical estimates for each function in the model, for each sample size. Plots for each model, in addition to the Least Squares (LS) estimates are displayed.

(ii) Tables 5.5 to 5.7 illustrate the empirical MSE of estimation, obtained from the simulation for sample sizes $N = 1024,\ 2048,\ 4096$. Values for each Bayes model, LS and AMlet estimates are displayed.

(iii) Figures 5.21 to 5.28 provide box plots for the MSE estimates for each Bayes model, LS and AMlet estimates, for all sample sizes $N = 512,\ 1024,\ 2048,\ 4096,\ 8192$.

Table 5.5: AMSE(standard deviation) results for Functions in the model, for $N = 1024$. In blue the minimum average MSE, in magenta, the corresponding minimum standard deviation of MSE.

	Block	Bumps
Bayes Total	0.52016 (0.03557)	1.0619867 (0.02450)
Bayes Mixture	0.52051 (0.03569)	1.06234 (0.02513)
Bayes Point	0.57980 (0.04351)	1.26478 (0.08206)
Least Squares	0.95203 (0.02675)	2.09584 (0.02760)
AMlet[1]	0.97790 (0.14014)	1.31650 (0.20801)
	Heavisine	Zero
Bayes Total	0.13965 (0.02475)	0.10913 (0.00126)
Bayes Mixture	0.07369 (0.01649)	0.03824 (0.01030)
Bayes Point	0.50730 (0.13925)	0.20343 (0.04192)
Least Squares	0.12967 (0.02566)	0.09834 (0.02725)
AMlet[1]	0.29501 (0.06532)	0.00461 (0.00393)

5.5.1 Remarks and Comments

(i) As can be observed from Tables x - x and Figures 5.21-5.28 the Bayesian methodologies present a large sample behavior which is similar to AMlet and LS, meaning that the MSE decreases as a function of N. In particular, it can be observed that the empirical results suggest that our proposed procedures, together with the LS present a smaller $\mathbb{L}_2$-risk that Amlet, for all sample sizes included in the study.

Table 5.6: AMSE(standard deviation) results for Functions in the model, for $N = 2048$. In blue the minimum average MSE, in magenta, the corresponding minimum standard deviation of MSE.

	Block	Bumps
Bayes Total	0.27177 (0.02117)	0.22830 (0.008338)
Bayes Mixture	0.27183 (0.02120)	0.22835 (0.00836)
Bayes Point	0.45494 (0.04079)	0.63394 (0.06133)
Least Squares	0.89767 (0.01446)	2.05133 (0.01560)
AMlet[1]	0.33718 (0.04888)	0.35967 (0.06058)
	Heavisine	Zero
Bayes Total	0.04165 (0.00876)	0.03153 (0.005162)
Bayes Mixture	0.04169 (0.00878)	0.00575 (0.00149)
Bayes Point	0.83454 (0.11631)	0.31326 (0.03674)
Least Squares	0.07540 (0.01185)	0.04543 (0.01506)
AMlet[1]	0.09442 (0.01739)	0.00105 (0.00079)

(ii) As can be observed in Figures 5.13 to 5.20, the Bayesian methods are able to accurately estimate the functions in the model, automatically adapting to each of the functions irregularities. This is particularly interesting in the case of the Blocks and Heavisine functions, for which the estimators nicely capture the rapid variations and discontinuities at the different scales.

(iii) From a visual perspective, it is evident from the simulations that as the sample size increases, the estimates are more stable and accurate which indicates that the Bias and Variance monotonically decreases with respect to the sample size. In particular, it can be observed that the Bayesian methods exhibit (in general) a smaller variability than AMlet and LS, indicating good finite sample behavior in the MSE sense.

(iv) When sample sizes are relatively small, it can be observed that the estimates are noisy but centered around the true function values. For this reason, the estimation accuracy could be improved by introducing a post-processing stage in which the function estimates are smoothed by introducing local-linear smoothers, or by thresholding the expansion coefficients in the same way as proposed by Donoho et al. (1994)[34].

(v) Note that in the first case (local linear smoothing), the estimator remains linear, meaning

Table 5.7: AMSE(standard deviation) results for Functions in the model, for $N = 4096$. In blue the minimum average MSE, in magenta, the corresponding minimum standard deviation of MSE.

	Block	Bumps
Bayes Total	0.17242 (0.01804)	0.02450 (0.00233)
Bayes Mixture	0.17244 (0.01805)	0.02451 (0.00234)
Bayes Point	0.60721 (0.05843)	0.53388 (0.04364)
Least Squares	0.41730 (0.00565)	0.97852 (0.00797)
AMlet[1]	0.06861 (0.010672)	0.05828 (0.00968)
	Heavisine	Zero
Bayes Total	0.01084 (0.00339)	0.00787 (0.00126)
Bayes Mixture	0.01085 (0.00339)	0.00076 (0.00022)
Bayes Point	0.89621 (0.09760)	0.70119 (0.06212)
Least Squares	0.03990 (0.00429)	0.02122 (0.00358)
AMlet[1]	0.02045 (0.00462)	0.00035 (0.00025)

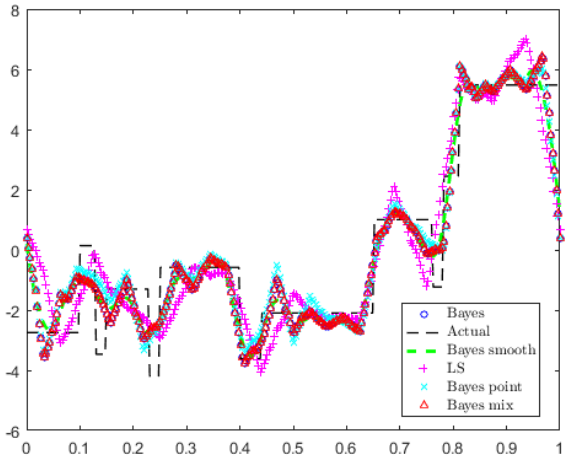
that it is a linear combination of the observed response, resulting from the application of an appropriate matrix. This alternative is more adequate for smooth functions. On the other hand, after applying thresholding the estimates become non-linear, which is especially suitable for irregular functions.

(vi) Even though the obtained results correspond to the Daubechies 6 filter (used for the expansion of all functions in the model), during the implementation phase we tested different filter such as Symmlets and Coiflets (refer to [3] for technical details), obtaining comparable results. However, since Daubechies-Lagarias algorithm is used for the construction of the design matrix, there is a trade-off between computational speed and accuracy of estimation, meaning that choosing a filter too large (in terms of number of taps) may inflate the computational cost for the calculations of the design matrix, without improving the estimation significantly enough as compared to the use of a shorter filter. However, the filter choice is a matter of subjective opinion and part of the art of statistical modeling.

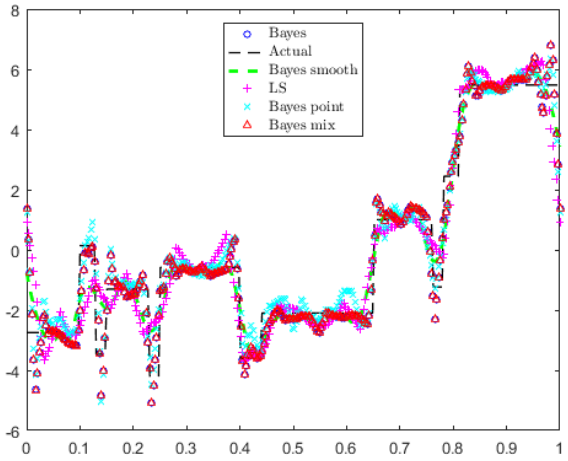
(vii) In the same line as the above argument, the proposed algorithm is capable of allowing the use of different filters for each feature in the model, meaning that the construction

of the design matrix can result from the use of multiple wavelet filters. For example, the simulation results shown in this section could be significantly improved if the Haar filter was utilized instead of Daubechies 6 for Blocks and Zero functions. Haar basis spans piece-wise constant functions in $\mathbb{L}_2([0,1])$, so it fits Blocks and Zero almost perfectly. This fact illustrates the flexibility of the proposed methodologies to introduce expert or previous knowledge about the problem to inform the estimation, and allow experimentation.

(viii) Regarding computational costs and efficiency, once the design matrix is constructed (using Daubechies-Lagarias), computations are extremely efficient since the estimate structure enables the use conjugate gradients, avoiding the explicit computation of matrix inversions, thus making the methods competitive when the sample size is moderate.

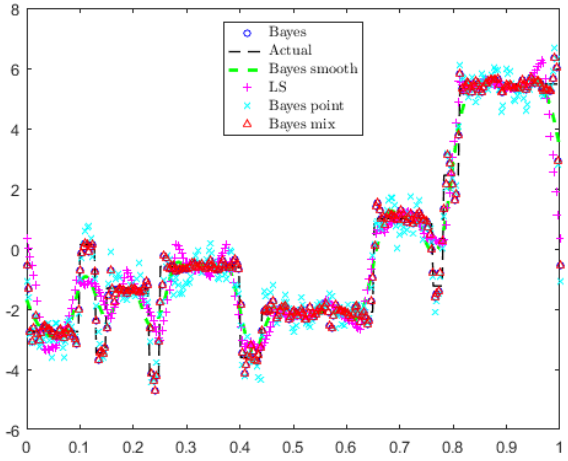


(a) Blocks, $N = 512$

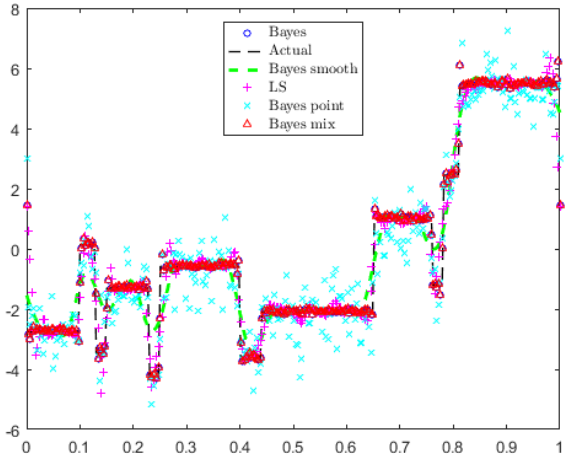


(b) Blocks, $N = 1024$

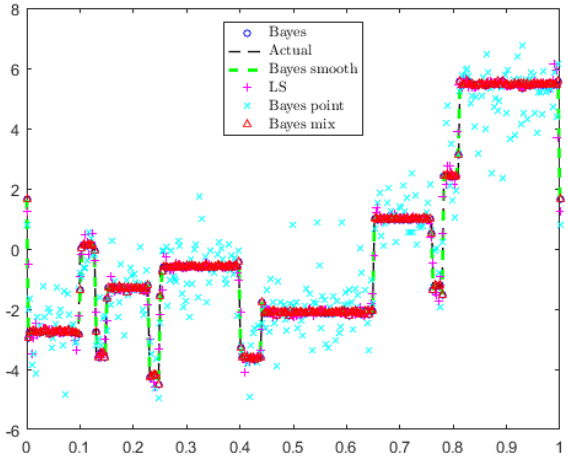
Figure 5.13: Typical estimated function Blocks for $N = 512,\ 1024$ samples.



(a) Blocks, $N = 2048$

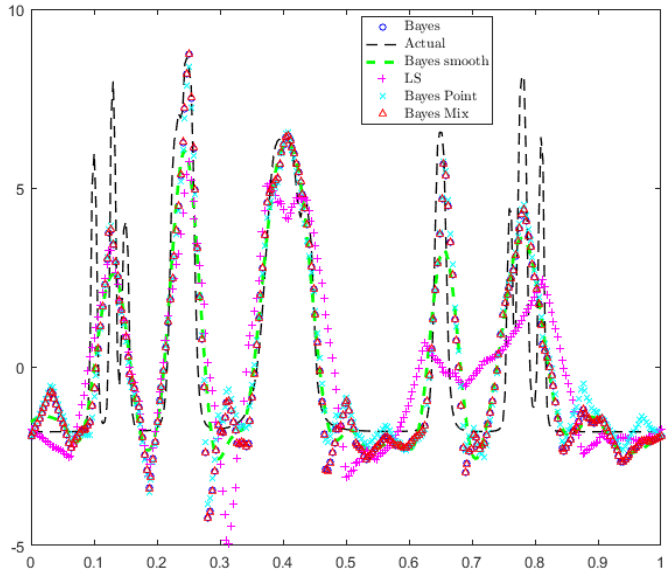


(b) Blocks, $N = 4096$

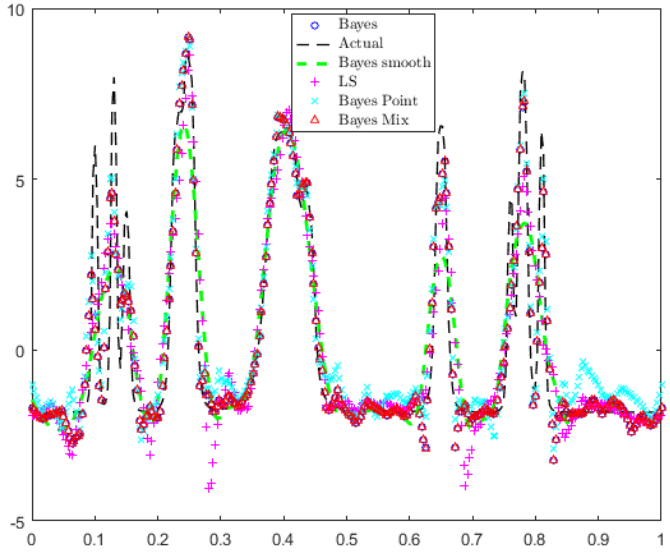


(c) Blocks, $N = 8192$

Figure 5.14: Typical estimated function Blocks for $N = 2048,\ 4096,\ 8192$ samples.

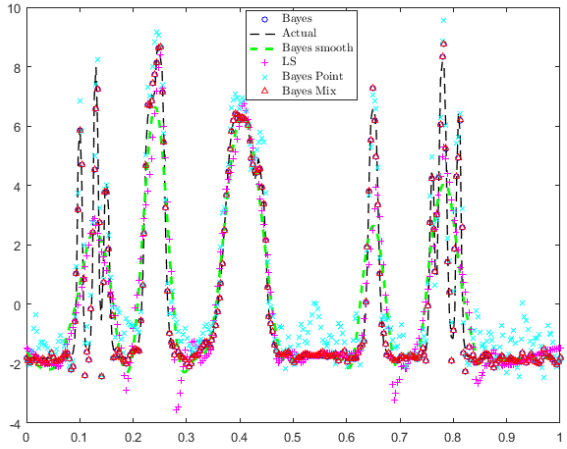


(a) Bumps, $N = 512$

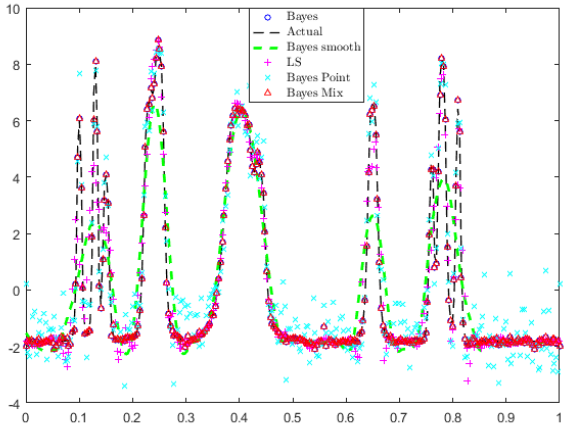


(b) Bumps, $N = 1024$

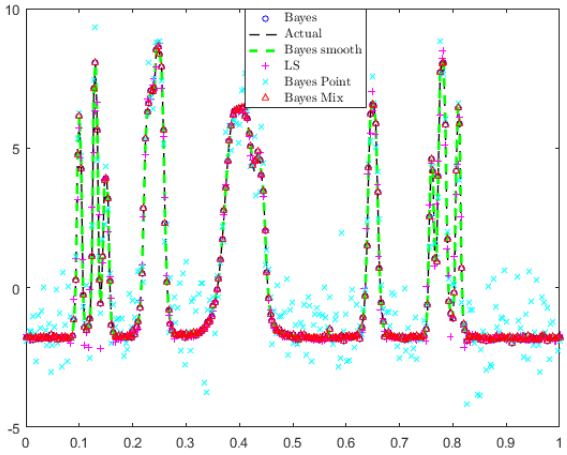
Figure 5.15: Typical estimated function Bumps for $N = 512,\ 1024$ samples.



(a) Bumps, $N = 2048$

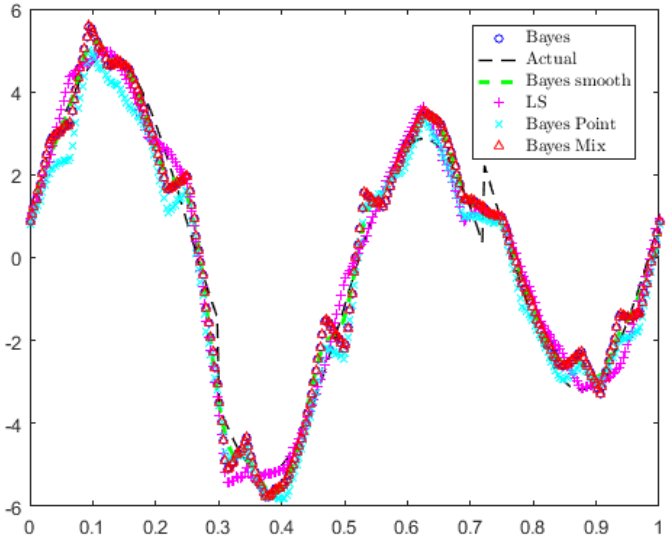


(b) Bumps, $N = 4096$

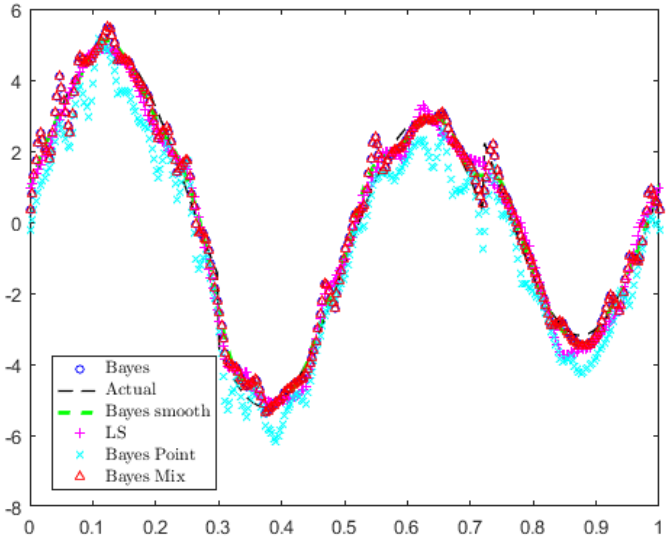


(c) Bumps, $N = 8192$

Figure 5.16: Typical estimated function Bumps for $N = 2048,\ 4096,\ 8192$ samples.

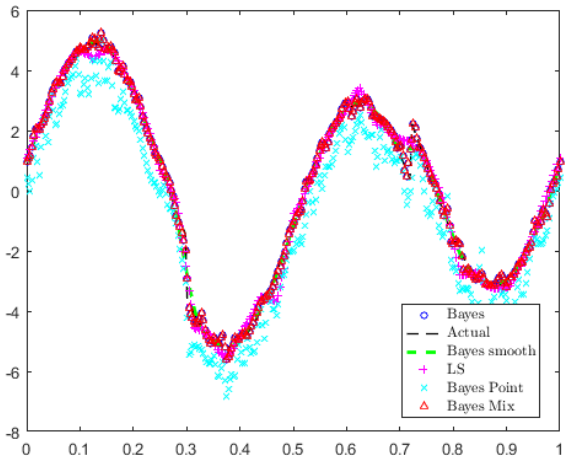


(a) Heavisine, $N = 512$

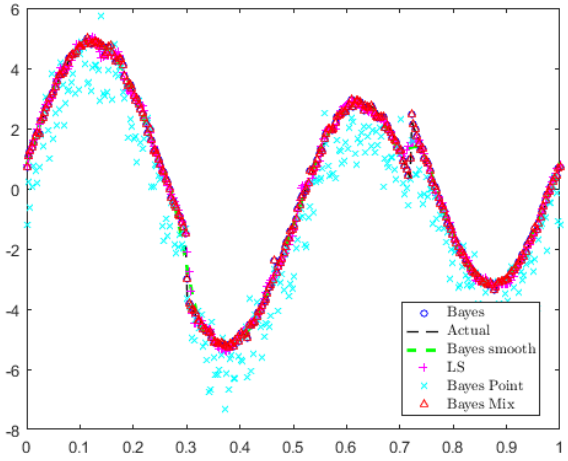


(b) Blocks, $N = 1024$

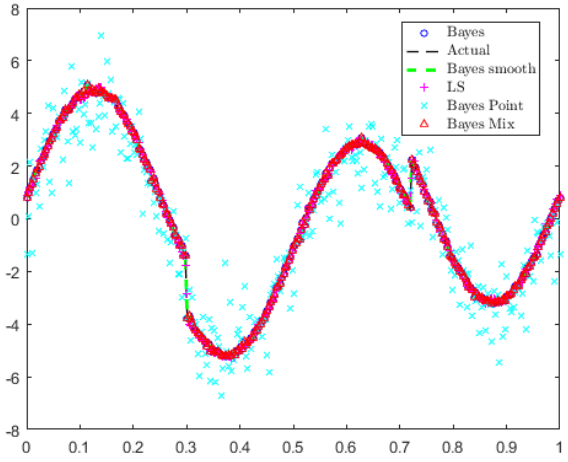
Figure 5.17: Typical estimated function Heavisine for $N = 512,\ 1024$ samples.



(a) Heavisine, $N = 2048$

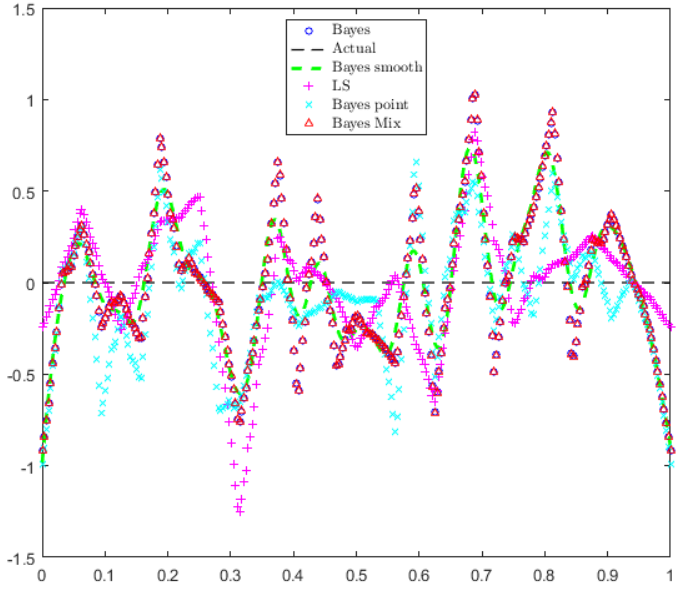


(b) Heavisine, $N = 4096$

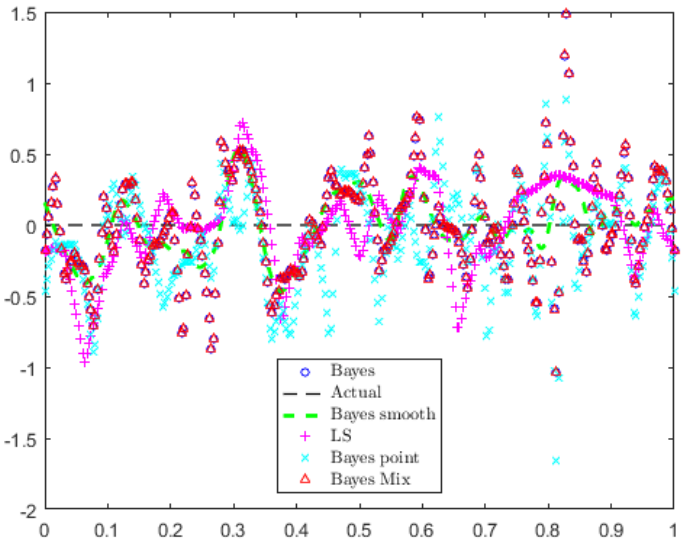


(c) Heavisine, $N = 8192$

Figure 5.18: Typical estimated function Heavisine for $N = 2048,\ 4096,\ 8192$ samples.

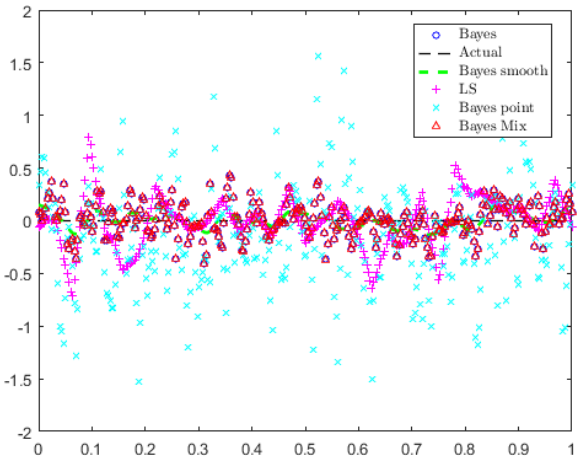


(a) Zero, $N = 512$

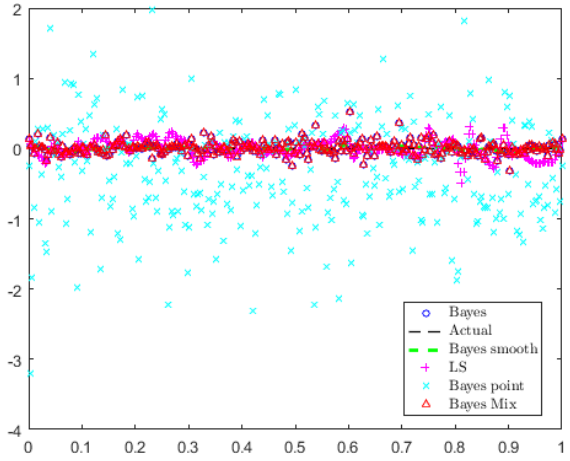


(b) Zero, $N = 1024$

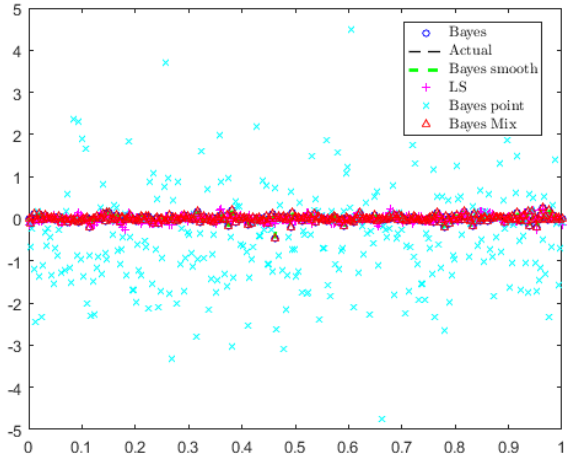
Figure 5.19: Typical estimated function Zero for $N = 512,\ 1024$ samples.



(a) Zero, $N = 2048$

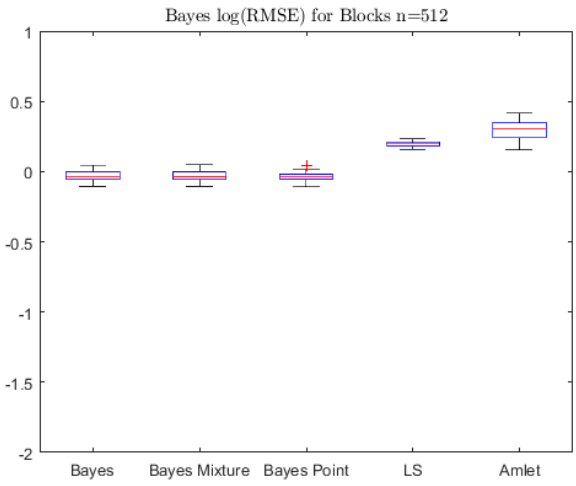


(b) Zero, $N = 4096$

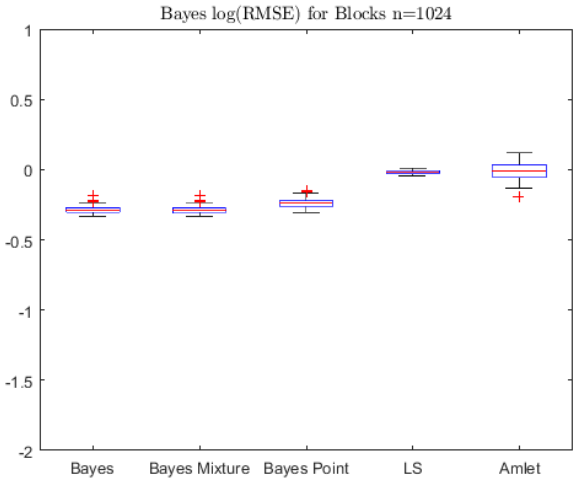


(c) Zero, $N = 8192$

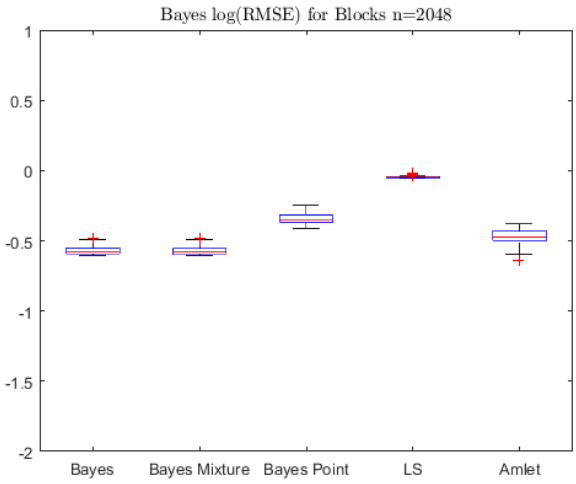
Figure 5.20: Typical estimated function zero for $N = 2048,\ 4096,\ 8192$ samples.



(a) Blocks, $N = 512$

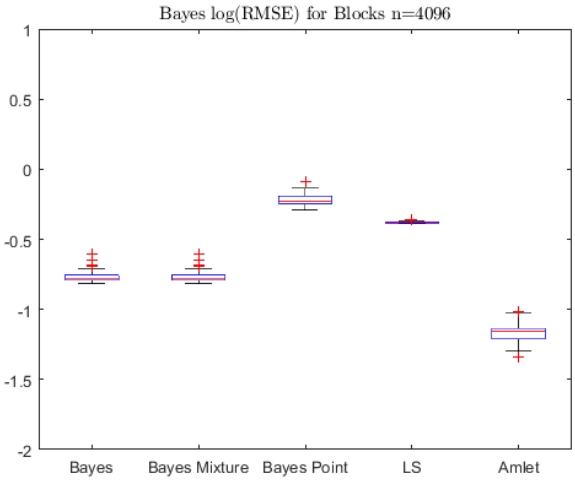


(b) Blocks, $N = 1024$

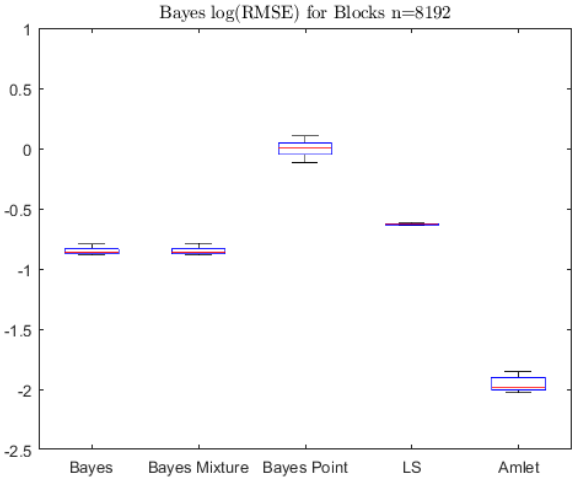


(c) Blocks, $N = 2048$

Figure 5.21: Empirical MSE for Blocks on a $\log_{10}$ scale. In each panel from left to right: Bayes Total, Bayes Mixture, Bayes Point Mass, Least Squares, Amlet.

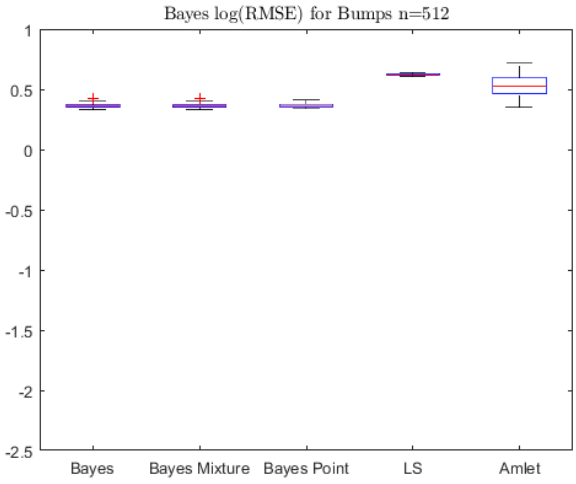


(a) Blocks, $N = 4096$

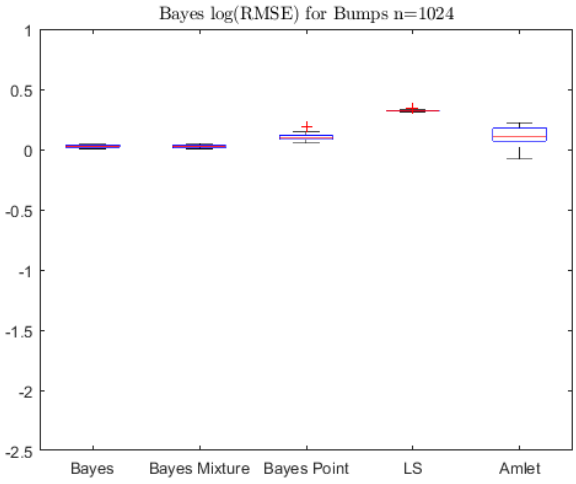


(b) Blocks, $N = 8192$

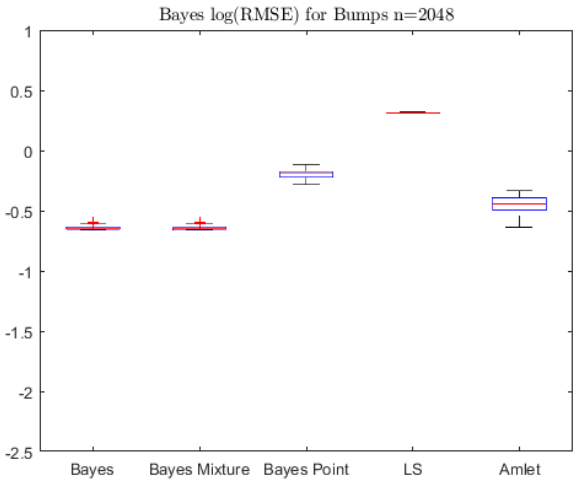
Figure 5.22: Empirical MSE for Blocks on a $\log_{10}$ scale. In each panel from left to right: Bayes Total, Bayes Mixture, Bayes Point Mass, Least Squares, Amlet.



(a) Bumps, $N = 512$

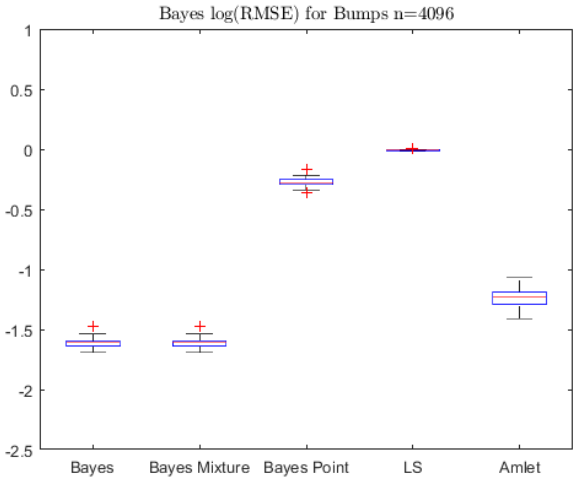


(b) Bumps, $N = 1024$

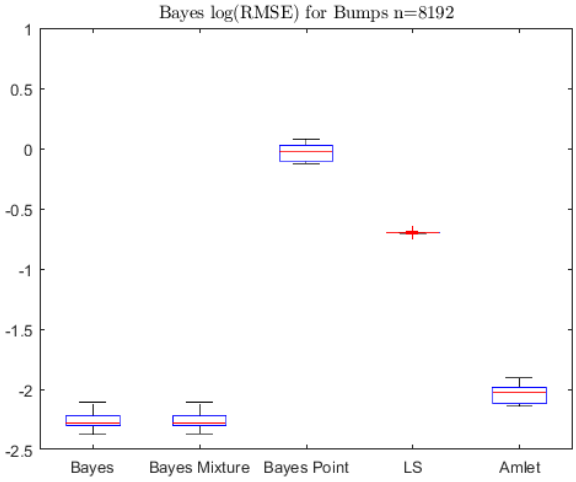


(c) Bumps, $N = 2048$

Figure 5.23: Empirical MSE for Bumps on a $\log_{10}$ scale. In each panel from left to right: Bayes Total, Bayes Mixture, Bayes Point Mass, Least Squares, Amlet.

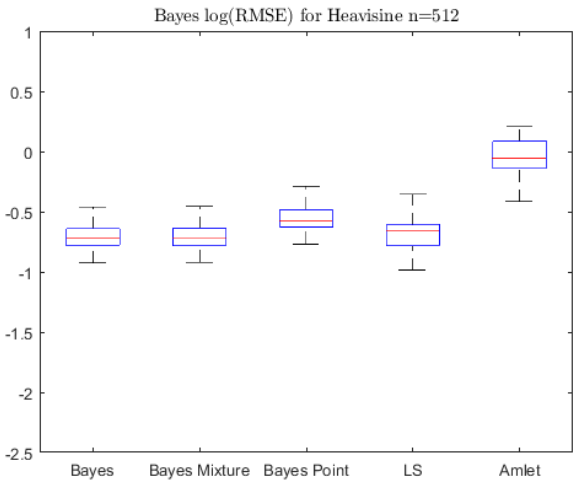


(a) Bumps, $N = 4096$

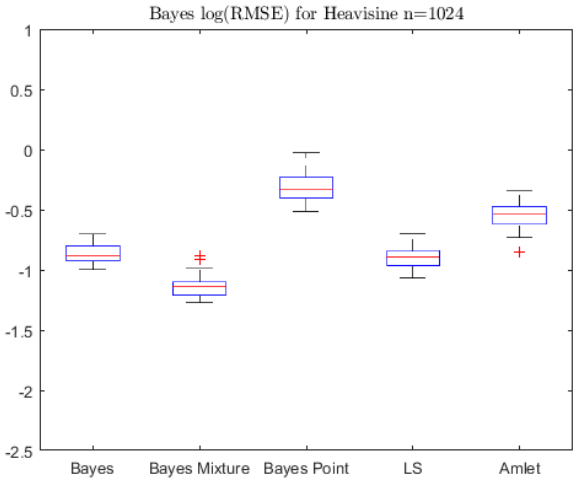


(b) Bumps, $N = 8192$

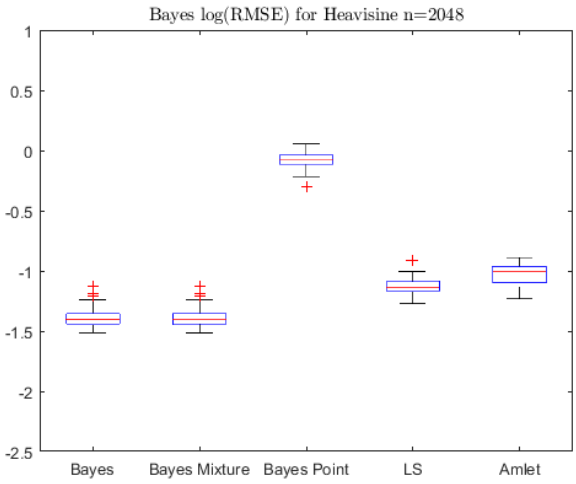
Figure 5.24: Empirical MSE for Bumps on a $\log_{10}$ scale. In each panel from left to right: Bayes Total, Bayes Mixture, Bayes Point Mass, Least Squares, Amlet.



(a) Heavisine, $N = 512$

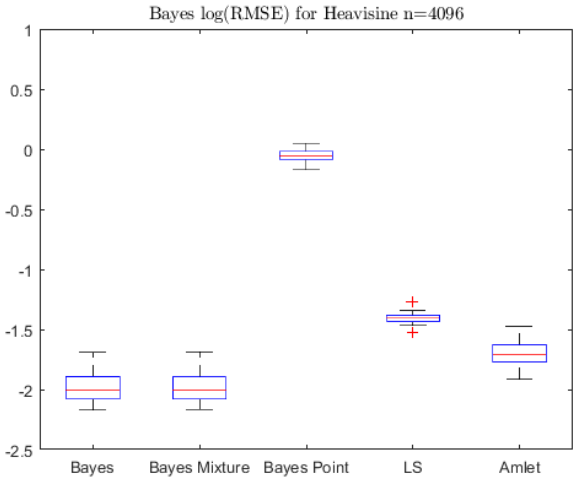


(b) Heavisine, $N = 1024$

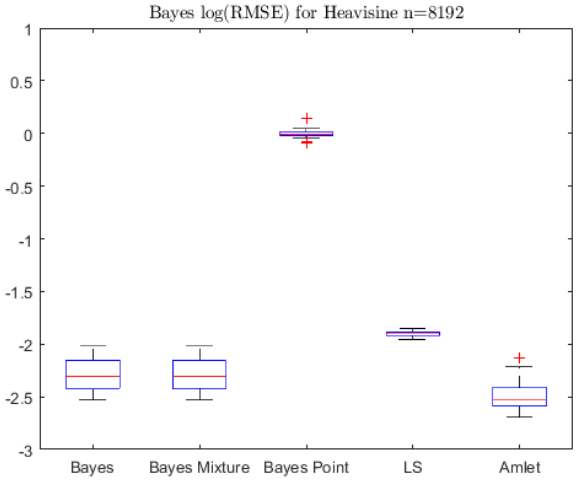


(c) Heavisine, $N = 2048$

Figure 5.25: Empirical MSE for Heavisine on a $\log_{10}$ scale. In each panel from left to right: Bayes Total, Bayes Mixture, Bayes Point Mass, Least Squares, Amlet.

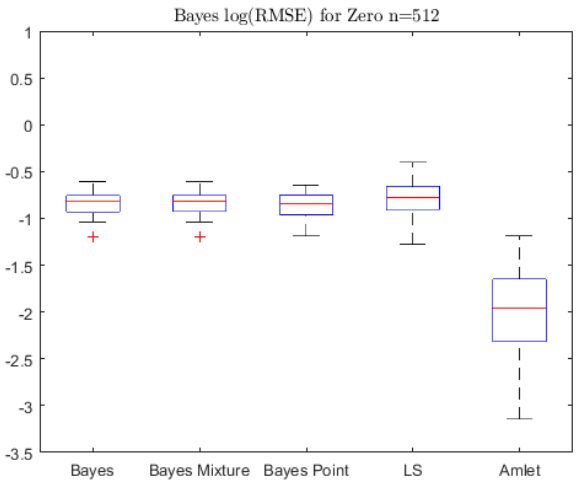


(a) Heavisine, $N = 4096$

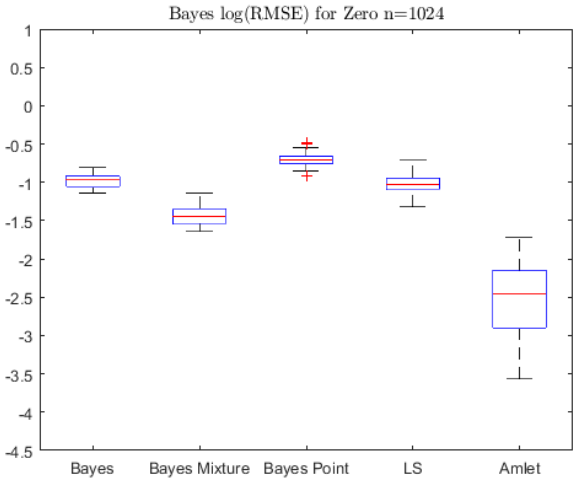


(b) Heavisine, $N = 8192$

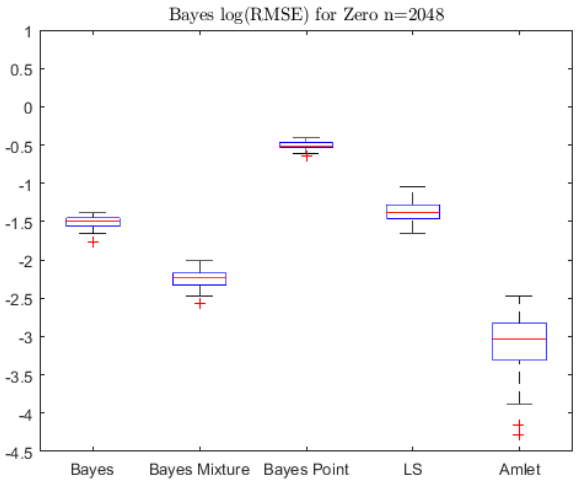
Figure 5.26: Empirical MSE for Heavisine on a $\log_{10}$ scale. In each panel from left to right: Bayes Total, Bayes Mixture, Bayes Point Mass, Least Squares, Amlet.



(a) Zero, $N = 512$

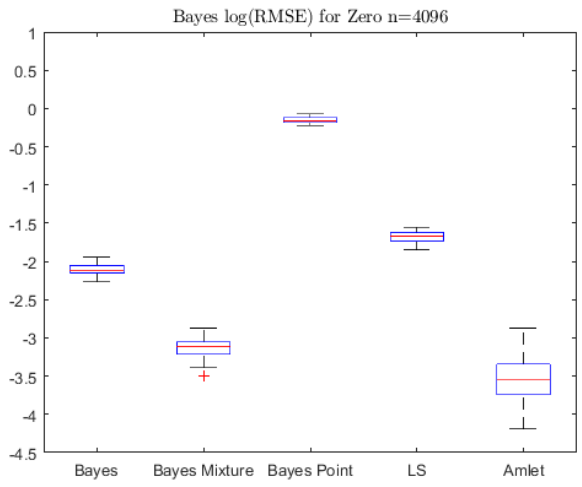


(b) Zero, $N = 1024$

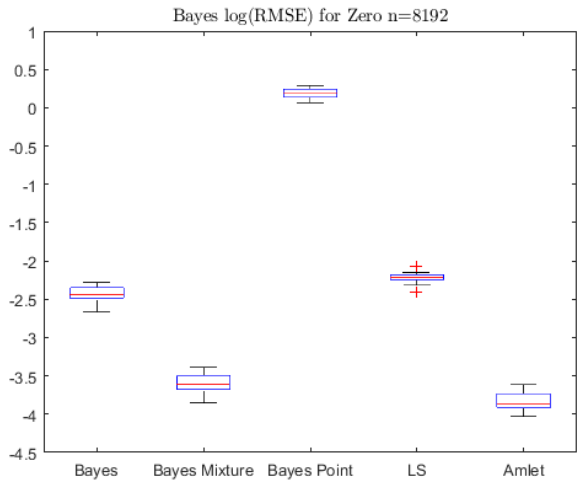


(c) Zero, $N = 2048$

Figure 5.27: Empirical MSE for Zero on a $\log_{10}$ scale. In each panel from left to right: Bayes Total, Bayes Mixture, Bayes Point Mass, Least Squares, Amlet.



(a) Zero, $N = 4096$



(b) Zero, $N = 8192$

Figure 5.28: Empirical MSE for Zero on a $\log_{10}$ scale. In each panel from left to right: Bayes Total, Bayes Mixture, Bayes Point Mass, Least Squares, Amlet.

5.6 Conclusions

In this chapter, we proposed and explored three different wavelet-based shrinkage methods for the adaptive estimation of additive regression models, exploiting conjugate structures that enable simple implementations and relatively efficient estimation using backfitting.

For each method, a complete derivation of the marginal and posterior distributions, model parameters and Bayes estimators was provided, and the linear nature of the estimates and shrinkage procedures is illustrated.

The proposed Bayes procedures are flexible and adaptive, capable of modeling dependency between the expansion coefficients, while introducing regularization in the matrix inversion needed for the computation of the shrinkage rules. In addition, the hyper-parameters needed for the specification of the prior distributions are computed from the data by following empirical Bayes, or extracting the values directly from the projection of the observed response into the wavelet domain, as proposed in [59].

Finally, the performance of each proposed methodology was assessed through a simulation study using benchmark functions that have been widely used in the literature, comparing results to those obtained by the Least Squares estimator introduced in Chapter 4, and the procedure proposed by Sardy and Tseng (2004)[1]. The simulation settings utilized resembled those used by previous authors, enabling a reasonable comparison with their methods.

Based on the obtained results, we can argue that the proposed Bayes procedures offer a competitive advantage against existing methodologies: they tend to outperform AMlet (without

the fast computations) and LS for small sample sizes and smooth functions, exhibiting good asymptotic properties and $\mathbb{L}_2$ risk behavior. On top of that, the methods are completely data-driven and fairly flexible and simple to implement. Moreover, when sample size and number of predictors is moderate, the estimation process is relatively fast which increases it potential applicability in real-life scenarios.

CHAPTER 6

MULTISCALE CORRELATION ANALYSIS IN THE WAVELET DOMAIN

In this Chapter, we exploit the linearity of the Discrete Wavelet Transformation (DWT) for the analysis of sample correlation. The usual Pearson's sample correlation coefficient is decomposed as a weighted sum of correlations between wavelet coefficients at different scale levels.

This representation enables a more detailed representation of the correlation structure in the data, revealing linear relationships at scales finer than the one utilized for the data collection, assessing how existing linear relationships are decomposed across different scales. This alternative way to express correlation can lead to very useful insights such as identifying sampling rates that lead to orthogonality between signals, capture the maximal information between samples or account for the observed relationship between sequences.

This Chapter introduces a formal and novel definition of the wavelet based correlation, discussing some of its characteristics and properties and providing simulation based examples that aim to illustrate possible scenarios expected to occur in real-life. In addition, two test-statistics that exploit the whitening properties of the DWT are proposed for the assessment of the statistical significance of the observed scale-wise correlations. Furthermore, these methods are evaluated in terms of their type I and II errors, comparing their performance with popular parametric and non-parametric statistical tests, using simulated data generated from stationary MA(1), AR(1) and ARMA(1,1) processes.

6.1 Introduction

The superb capabilities of wavelets for the decomposition of processes at different time scales while preserving time localization have translated into a growing popularity of wavelet-based methodologies for the analysis of correlation between signals. This is specially noticeable within the fields of economics, finance and physical sciences. Gencay et al. (2001)[68] provides a detailed description of the application of wavelets in economics and finance, with most results based on the application of the maximal overlap discrete wavelet transform (MODWT). This tools was introduced in 2000 by Percival and Walden (2000)[69], and is a non-orthogonal modified version of the DWT. Among its main differences with the DWT, it possess the flexibility to handle sequences of any lenght N, and it is shift-invariant, meaning that a shift in the signal does not change the pattern of the wavelet transformed coefficients.

Regarding some of the existing research that makes use of wavelet-based correlations, the following works provide good references: Grinsted et al. (2004)[70] proposed a methodology for the analysis of coherence between signal in a certain state spaces, Capobianco (2004)[71] applied wavelet methods to the multiresolution analysis of high frequency Nikkei stock index data, estimating periodic effects in those signals. Later on, Fernández-Macho (2012)[72] proposed a method for multiple correlation analysis using wavelet transformations, Benhmad (2013)[73] analyzed the cross-contamination between stock markets, showing a scale dependency via the use of wavelet transforms.

In the atmospheric and physical sciences context, Hudgins (1992)[74] introduced the concepts of wavelets cross spectrum and wavelet cross correlation using the CWT, applying those concept in the analysis of atmospheric turbulence (Hudgins et al. 1993). Liu (1994)[75] defined a wavelet cross spectrum, similar to Hudgins but using complex wavelets instead. Later on, Lindsay et al. (1996)[76] utilized the DWT to define the wavelet covariance, along with large-sample based confidence intervals for the analysis of surface temperatures in the

Beaufort sea. In 2000 Whitcher, Guttorp and Percival [77] extended the notion of wavelet covariance for the MODWT, defining wavelet cross covariance and cross correlation. More recently, Pering et al. (2014)[78] introduced the use of CWT for the analysis of correlation in the geosciences domain by combining the scale-wise representations and Spearman's rank correlation (see Spearman 1904[79]). In this same context Casagrande et. al (2015)[80] utilized wavelet-based cross-correlations using CWT with complex wavelets, showing a method able to capture the dynamics of the soil moisture-temperature coupling over a wide range of temporal scales.

Even though the idea of applying wavelets to generate a multiscale version of the correlation analysis is not new, most of the existing literature focuses mainly on its applications, with restricted attention to the tool itself and some of its properties for stationary sequences. In particular, based on the available information that was possible to gather for the elaboration of this Chapter, it was observed that few results about the performance of commonly used tests for the assessment of statistical significance of correlation findings for different types of stationary processes have proposed in the literature.

For this reason, in this chapter our goal is to introduce a definition of the wavelet based correlation procedure using an orthogonal DWT resulting from compactly supported wavelets, showing that the additive structure of the sample covariance that leads to a weighted sum of level-wise correlations between expansion coefficients in the wavelet domain. In addition, some interesting results regarding some of its distributional and statistical properties is provided for specific types of stationary processes, aiming to build intuition and provide a framework over which different statistical tools could be built.

Along this line, our second goal for this Chapter consists of proposing a test statistic that exploits the whitening property of wavelets for the assessment of the statistical significance of the observed level-wise correlations. In order to study its expected performance and provide intuition about the effect of different types of stationary processes over the performance of

commonly used statistical tests, a simulation-based comparison with the well-known Pearson's $t-$test and some other non-parametric statistical procedures is provided, utilizing simulated stationary processes that aim to illustrate possible scenarios that are expected to occur in real-life.

A a result of this, our findings suggest that the proposed test statistic based on the condition number of the sample covariance matrix of level-wise wavelet coefficients exhibits a significantly smaller type I error than popular statistical tests used as benchmark. In particular, when the analyzed signals are uncorrelated and exhibit short-time high oscillations the proposed methodology significantly outperforms the other tests, which tend to significantly increase their false rejection rates reaching in some cases, average rates higher than 30%. Similarly, in terms of the type II error, the proposed test is in general, as good as the other tests showing a consistent behavior across the different models and setting that were tested.

Finally, and application use-case that illustrates the applicability of the proposed tools in a data set that studies daily average temperatures in the cities of Atlanta and Athens, GA. is presented, and a brief discussion is provided.

6.2 Scale-wise Representation of Sample Correlation via DWT

Consider two real-valued random sequences $X_1, ..., X_N$, and $Y_1, ..., Y_N$ resulting from observations of the random variables X and Y with assumed joint distribution given by $f_{X,Y}(x, y)$. Denote as $\mathbf{x} = [X_1 \dots X_N]^T$ and $\mathbf{y} = [Y_1 \dots Y_N]^T$ each of the observed sequences respectively. WLOG, assume that $\mathbb{E}[X] = \mathbb{E}[Y] = 0$, and $Var(X) = \sigma_X^2 < \infty$, $Var(Y) = \sigma_Y^2 < \infty$. These assumptions will be utilized throughout the sequel for all derivations and results, in most cases without explicitly mentioning them.

As it is known in the statistical domain, the correlation is a measure of linear relationship

between X, Y, and as proposed by Pearson [81], can be computed as:

$$\rho_{X,Y} = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}} = \frac{\mathbb{E}[XY]}{\sigma_X \sigma_Y}. \tag{6.1}$$

Note that, if X and Y are independent, meaning $f_{X,Y}(x,y) = f_X(x) f_Y(y)$, then $\rho_{X,Y} = 0$. Moreover, to compute the correlation we need to obtain the numerator of Eq.(6.1) which represents the Covariance between X and Y (under the assumption of zero mean random variables).

In most practical situations knowledge of the underlying probability density of the observed samples is not available, therefore $Cov(X,Y)$ cannot be computed directly. In such cases, natural estimators of $\mathbb{E}[XY]$ and $\rho_{X,Y}$ are given by:

$$\widehat{Cov}(X,Y) = \frac{1}{N}\sum_{i=1}^{N} X_i Y_i = \frac{1}{N}\mathbf{x}^T\mathbf{y} = \frac{1}{N}\langle \mathbf{x}, \mathbf{y}\rangle \tag{6.2}$$

$$\hat{\rho}_{X,Y} = \frac{\langle \mathbf{x}, \mathbf{y}\rangle}{\sqrt{\langle \mathbf{x}, \mathbf{x}\rangle \langle \mathbf{y}, \mathbf{y}\rangle}}. \tag{6.3}$$

Here $\langle \cdot, \cdot \rangle$ denotes the standard inner product in $\mathbb{R}^N$. Since each of the sequences can be interpreted as observations resulting from equally spaced measurements of a certain process (e.g. hourly stock prices, weekly average temperatures, distributed sensors, etc.), it is possible to represent them in the wavelet domain via the DWT.

In this context, suppose an orthogonal wavelet matrix of the decimated type, denoted by $\mathbf{W}$. Let $\mathbf{d}_X = \mathbf{W}\mathbf{x}$ and $\mathbf{d}_Y = \mathbf{W}\mathbf{y}$ be the resulting vectors of wavelet coefficients from the DWT of $\mathbf{x}$ and $\mathbf{y}$ respectively. Then, it follows:

$$\frac{1}{N}\mathbf{d}_X^T\mathbf{d}_Y = \frac{1}{N}\mathbf{x}^T\mathbf{W}^T\mathbf{W}\mathbf{y} = \frac{1}{N}\langle \mathbf{x}, \mathbf{y}\rangle = \widehat{Cov}(X,Y), \tag{6.4}$$

$$\mathbf{d}_X^T\mathbf{d}_X = \mathbf{x}^T\mathbf{W}^T\mathbf{W}\mathbf{x} = \langle \mathbf{x}, \mathbf{x}\rangle. \tag{6.5}$$

The last result indicates that due to the orthogonality of $\mathbf{W}$ and the linearity of the DWT, energy is preserved, so it is possible to analyze the correlation of the sequences in the wavelet domain, preserving its structure. In fact, the application of the orthogonal DWT can be interpreted as a special rotation of the original sequences that preserves length and allows the decomposition of their dimensionality into disjoint subspaces where they can be analyzed separately.

Assuming that $N = 2^J$, the DWT applied to the data sequences $\mathbf{x}$ and $\mathbf{y}$ generates vectors of expansion coefficients that have the following structure:

$$\mathbf{d} = \begin{bmatrix} \mathbf{c}^{(J-k)} \\ \mathbf{d}^{(J-k)} \\ \mathbf{d}^{(J-k+1)} \\ \vdots \\ \mathbf{d}^{(J-2)} \\ \mathbf{d}^{(J-1)} \end{bmatrix}_{2^J} . \tag{6.6}$$

In the last expression k corresponds to the number of steps or depth in the DWT (usually, $k = J$). Also, it is important to mention that due to the decimated nature of the chosen DWT, the size of the vector $\mathbf{d}$ is also N (as in the original data vector $\mathbf{x}$). In Eq.(6.6), $\mathbf{c}^{(J-k)}$ corresponds to the smooth coefficients at scale level $J - k$; similarly, $\mathbf{d}^{(J-k)}$ corresponds to the set of detail coefficients at the scale level $J-k$. Assuming $k = J$, each component $\mathbf{d}^{(J-l)}$

of $\mathbf{d}$ in Eq.(6.6) is given by:

$$\mathbf{d}^{(J-l)} = \begin{bmatrix} d_{J-l,0} \\ d_{J-l,1} \\ \vdots \\ d_{J-l,2^{J-l}-1} \end{bmatrix}_{2^{J-l}\times 1} \quad , \text{ for } l = J, J-1, ..., 1.$$

Here $d_{j,m},\ c_{j,m}$ correspond to the discrete wavelet coefficients in the wavelet expansion of a function $f \in V_j \bigcup W_j$, as described in Chapter 1.

Therefore, it follows that the transformed sequences are structured as:

$$\mathbf{d}_X = \begin{bmatrix} c_X^{(0)} \\ d_X^{(0)} \\ \mathbf{d}_X^{(1)} \\ \vdots \\ \mathbf{d}_X^{(2)} \\ \mathbf{d}_X^{(J-1)} \end{bmatrix}_{2^J} \quad , \quad \mathbf{d}_Y = \begin{bmatrix} c_Y^{(0)} \\ d_Y^{(0)} \\ \mathbf{d}_Y^{(1)} \\ \vdots \\ \mathbf{d}_Y^{(2)} \\ \mathbf{d}_Y^{(J-1)} \end{bmatrix}_{2^J} \quad , \tag{6.7}$$

which implies that:

$$\widehat{Cov}(X,Y) = \langle \mathbf{d}_X, \mathbf{d}_Y \rangle = c_X^{(0)} c_Y^{(0)} + \sum_{j=0}^{J-1} \langle \mathbf{d}_X^{(j)}, \mathbf{d}_Y^{(j)} \rangle. \tag{6.8}$$

Note that $\langle \mathbf{d}_X^{(j)}, \mathbf{d}_Y^{(j)} \rangle = \sum_{k=0}^{2^j-1} d_{j,k}^{(X)} d_{j,k}^{(Y)}$, for $j = 0, ..., J-1$. This last fact, together with Eq.(6.8) shows that the sample covariance between the sequences $\mathbf{x}$ and $\mathbf{y}$ can be decomposed as the summation of the level-wise inner products of wavelet coefficients.

Now, from the last set of results, it is possible to express the sample correlation $\hat{\rho}_{X,Y}$ as:

$$\hat{\rho}_{X,Y} = \frac{c_X^{(0)} c_Y^{(0)} + \sum_{j=0}^{J-1} \langle \mathbf{d}_X^{(j)}, \mathbf{d}_Y^{(j)} \rangle}{\sqrt{\left(c_X^{(0)} c_X^{(0)} + \sum_{j=0}^{J-1} \langle \mathbf{d}_X^{(j)}, \mathbf{d}_X^{(j)} \rangle \right) \left(c_Y^{(0)} c_Y^{(0)} + \sum_{j=0}^{J-1} \langle \mathbf{d}_Y^{(j)}, \mathbf{d}_Y^{(j)} \rangle \right)}}. \tag{6.9}$$

Define for $j = 0, ..., J-1$:

$$w_j = \sqrt{w_X^{(j)} w_Y^{(j)}} = \sqrt{\frac{||\mathbf{d}_X^{(j)}||_2^2 \cdot ||\mathbf{d}_Y^{(j)}||_2^2}{\left(||c_X^{(0)}||_2^2 + \sum_{j=0}^{J-1} ||\mathbf{d}_X^{(j)}||_2^2 \right) \left(||c_Y^{(0)}||_2^2 + \sum_{j=0}^{J-1} ||\mathbf{d}_Y^{(j)}||_2^2 \right)}}. \tag{6.10}$$

Here, $w_X^{(j)} = \frac{||\mathbf{d}_X^{(j)}||_2^2}{||\mathbf{d}_X||_2^2}$, and $w_Y^{(j)} = \frac{||\mathbf{d}_Y^{(j)}||_2^2}{||\mathbf{d}_Y||_2^2}$. Clearly, $0 \leq w_j \leq 1$, for $j = 0, ..., J-1$. Therefore, Eq.(6.9) becomes:

$$\hat{\rho}_{X,Y} = \frac{c_X^{(0)} c_Y^{(0)}}{||\mathbf{d}_X||_2 ||\mathbf{d}_Y||_2} + \sum_{j=0}^{J-1} w_j \hat{\rho}_{X,Y}^{(j)}, \quad \text{and} \tag{6.11}$$

$$\hat{\rho}_{X,Y}^{(j)} = \frac{\langle \mathbf{d}_X^{(j)}, \mathbf{d}_Y^{(j)} \rangle}{||\mathbf{d}_X^{(j)}||_2 ||\mathbf{d}_Y^{(j)}||_2}, \quad j = 0, ..., J-1. \tag{6.12}$$

Note that from Eq.(6.12), it is clear that for $j = 0, ..., J-1$ the terms $\hat{\rho}_{X,Y}^{(j)}$ satisfy $|\hat{\rho}_{X,Y}^{(j)}| \leq 1$. Thus, expression (6.11) indicates that the sample correlation $\hat{\rho}_{X,Y}$ can be expressed as the weighted sum of the level-wise correlation between the expansion coefficients resulting from the DWT of each of the signals. By this representation, it is possible to assess the individual contributions of each of the scales to the overall correlation between two signals.

This last fact enables to relate the scale-wise correlations to the original measurement scale, thus identifying linear relations that may exist at scales that are finer than the one utilized for the data collection, providing additional information and insights about the data under study.

Remarks

(a) From expression (6.11), another interpretation of the usual correlation coefficient be-

tween two signals is that it corresponds to the aggregated effect of multiple-scales interactions. Therefore, the multiscale representation offers a richer source of information about the existing relationships

(b) Note that for a fixed j, $w_j = 0$ if either $||\mathbf{d}_X^{(j)}||_2 = 0$ or $||\mathbf{d}_Y^{(j)}||_2 = 0$. This means that low-energy levels are prone to have small weights. However, smooth signals tend to concentrate most of its energy in low scale levels, so it can be expected that high weights could be observed in those cases.

(c) Since $w_X^{(j)} = \sqrt{\frac{||\mathbf{d}_X^{(j)}||_2^2}{||\mathbf{d}_X||_2^2}}$, it follows that:

$$\frac{(c_X^{(0)})^2}{||\mathbf{d}_X||_2^2} + \sum_{j=0}^{J-1}(w_j^{(X)})^2 = 1.$$

(d) Based on the previous definition, since the energy distribution across different scale levels is directly related to the signal smoothness and stochastic characteristics (e.g. self-similarity, highly localized oscillations, etc.) the observation of the weights distribution could be a good source of information for the assessment of those signal features.

(e) From a geometrical viewpoint, from Eq.(6.12) it is possible to observe that:

$$\hat{\rho}_{X,Y}^{(j)} = cos(\theta_{X,Y}^{(j)}),$$

where $cos(\theta_{X,Y}^{(j)})$ is the angle formed between vectors $\mathbf{d}_X^{(j)}$ and $\mathbf{d}_Y^{(j)}$ in $\mathbb{R}^{2^j}$. Therefore:

$$\hat{\rho}_{X,Y} = \frac{c_X^{(0)} c_Y^{(0)}}{||\mathbf{d}_X||_2||\mathbf{d}_Y||_2} + \sum_{j=0}^{J-1} w_j cos(\theta_{X,Y}^{(j)}).$$

6.3 Some Interesting Correlation Relationships Between Signals and Properties of Wavelet Coefficients

In this section, some interesting correlation relationships between signals are studied. This aims to the generation of insights and the identification of special structures or properties that arise in such cases for the multiscale approach, generating a framework for further analysis, testing and interpretation of results. Each of the cases that are studied are complemented with simulation-based examples that illustrate the properties under discussion.

6.3.1 Case 1: Perfect Correlation between $\mathbf{x}$ and $\mathbf{y}$

This is a trivial situation that assumes that $Y = aX$, for an arbitrary $a \neq 0$. Due to the linearity and homogeneity of the DWT, this linear relationship translates into a representation of the sequence $\mathbf{y}$ in the wavelet domain that is just a re-scaled version of the wavelet representation of the signal $\mathbf{x}$. This implies that:

$$\begin{aligned} \hat{\rho}_{X,Y}^{(j)} &= sign(a),\ j = 0, ..., J-1, \\ w_j &= 1,\ j = 0, ..., J-1, \\ \hat{\rho}_{X,Y} &= sign(a). \end{aligned}$$

This result follows directly from Eqs.(6.11)-(6.12), and suggest that strong linear relationships between signals are likely to be evenly spread out into the level-wise correlations.

6.3.2 Case 2: Perfect correlation between $\mathbf{x}$ and $\mathbf{y}$ at a particular multiresolution level j_0.

Suppose that for a fixed $0 \leq j_0 \leq J-1$, and $\beta \neq 0$, it holds:

$$\mathbf{d}_Y^{(j_0)} = \beta \cdot \mathbf{d}_X^{(j_0)}.$$

This relation implies that at the multiresolution level j_0, samples $\mathbf{x}$ and $\mathbf{y}$ are linearly dependent. Note that under this condition, it follows:

$$\begin{aligned}
\hat{\rho}_{X,Y}^{(j_0)} &= sign(\beta), \\
w_{j_0} &= \frac{|\beta|}{||\mathbf{d}_X||_2 \sqrt{C_{j_0} + \beta^2}}, \\
C_{j_0} &= \frac{1}{||\mathbf{d}_X^{(j_0)}||_2^2} \left(||c_Y^{(0)}||_2^2 + \sum_{j=0,\ j \neq j_0}^{J-1} ||\mathbf{d}_Y^{(j)}||_2^2 \right).
\end{aligned}$$

Here, $C_{j_0} > 0$ represents the ratio of energies contained in all levels but j_0 in signal $\mathbf{y}$ and the energy contained at level j_0 in signal $\mathbf{x}$. Note that $C_{j_0} = 0$ when all energy in the signal is concentrated in level j_0. Similarly, $C_{j_0} \to \infty$ when no energy is contained at scale j_0.

Suppose for a fixed $C_{j_0} > 0$, an perfect linear relation between both signals at level j_0 exists and $||\mathbf{d}_X||_2 = 1$. Then $w_{j_0} = w_{j_0}(\beta, C_{j_0})$ is a non-negative symmetric function of (β, C_{j_0}), that as $\beta \to \infty$, $w_{j_0}(\beta, C_{j_0}) \to 1$. In the presence of linear dependence, the value of C_{j_0} determines how fast the weight w_{j_0} converges to 1. Fig. 6.1 illustrates this behavior:

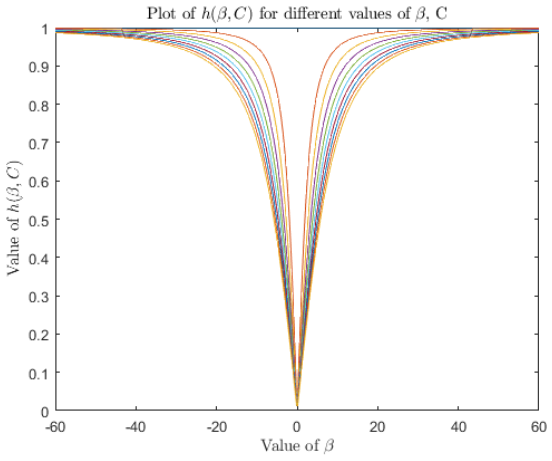


Figure 6.1: Plot of the correlation weights w_j for different values of $C_j > 0$. Each of the colored lines represent the shape of $w_j(\beta, C)$ for a fixed value of C. In the plot, C ranges from 0 to 100. The larger C, the smaller the slope of the curve around zero, and the slower it reaches the asymptotic value of 1.

Figure 6.1 shows that a perfect correlation at a fixed scale j_0 does not necessarily imply high weights. In the extreme case when the transformed signals are orthogonal for every scale level, except for j_0, then $\hat{\rho}_{X,Y}^{(j)} = 0,\ \forall\, j \neq j_0$. Thus, in such scenario the sample correlation takes the form:

$$\begin{aligned}\hat{\rho}_{X,Y} &= \frac{c_X^{(0)} c_Y^{(0)}}{||\mathbf{d}_X||_2 ||\mathbf{d}_Y||_2} + w_{j_0} sign(\beta), \\ &= \frac{c_X^{(0)} c_Y^{(0)}}{||\mathbf{d}_X||_2 ||\mathbf{d}_Y||_2} + \frac{\beta}{||\mathbf{d}_X||_2 \sqrt{C_{j_0} + \beta^2}}\end{aligned}$$

Note that because of the effect of the weight w_{j_0}, this value can be small (and even close to zero), meaning that significant correlation at a scale level does not necessarily reflects on the overall correlation in the original domain.

Simulation-based examples

In this section the case of correlation between signals at a specific scales is exemplified via simulation, and some graphical illustrations are provided. The methodology used for this purpose is the following:

- Generate two independent random sequences $X_1, ..., X_N$, $Y_1, ..., Y_N$ for $N = 2^J$,and $X, Y \sim \mathcal{N}(0, \sigma^2)$. Construct and orthogonal wavelet matrix $\mathbb{W}$ by choosing a appropriate wavelet filter (e.g. Daubechies 6).
- Obtain $\mathbf{d}_X = \mathbf{W}\mathbf{x}$, and $\mathbf{d}_Y = \mathbf{W}\mathbf{y}$ using the wavelet matrix $\mathbf{W}$.
- For a chosen multiresolution level j_0 and $\beta \neq 0$, set $\mathbf{d}_Y^{(j_0)} \leftarrow \beta \cdot \mathbf{d}_X^{(j_0)}$. This generates a new transformed signal $\tilde{\mathbf{d}_Y}$.
- Return back to the original domain of both signals, meaning $\mathbf{x} = \mathbf{W}^T \mathbf{d}_X$, and $\mathbf{y} = \mathbf{W}^T \tilde{\mathbf{d}_Y}$.

- Compute the sample correlation in the usual way (Pearson's sample correlation coefficient).

This process is then repeated multiple times, following a montecarlo methodology. The following tables and figures illustrate the obtained results for different values of N, β and σ^2:

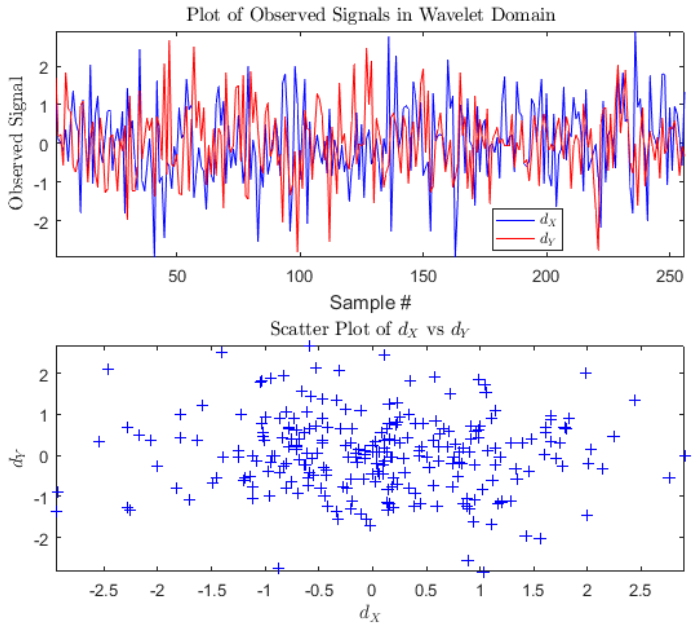
(i) In Figs. 6.2a-6.3b it is possible to observe that perfect correlation in the wavelet domain may not be visually evident in the time domain. However, in the wavelet domain that behavior is clear.

(ii) In Fig. 6.4 the effect of correlation on an individual scale level on the sample correlation at the original domain is exemplified, using $N = 256$, $\beta = 1$ and $\sigma = 0.1$. It is interesting to observe that the coarser the scale, the less significant the effect of the correlation. This is particularly evident for panels (a)-(c), whereas for (d)-(e) there is no significant different between the two samples.

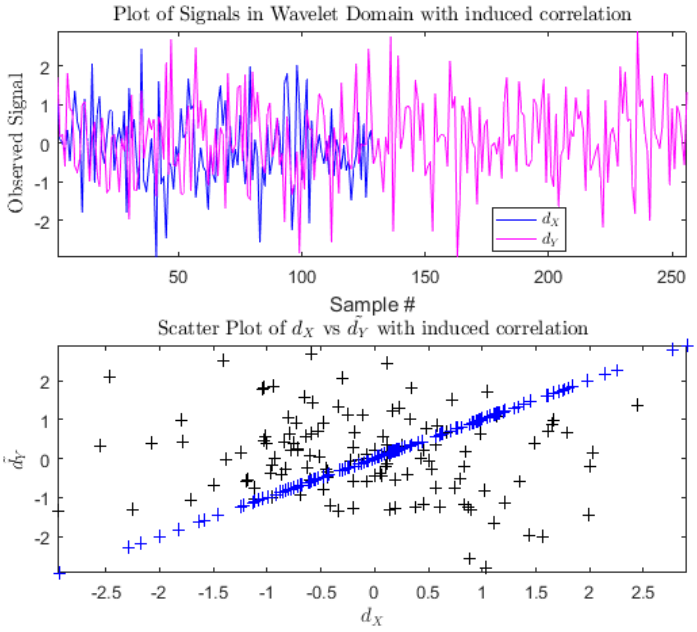
(iii) Also, from Fig. 6.4 in the case of no correlation between wavelet coefficients at each scale, the scale-correlations are almost symmetric around zero, with variability that is monotonically decreasing as the detail level increases.

(iv) In Fig. 6.6 the effects of perfect correlation are enhanced due to the artificial nature of the example. However, when using signals from real applications it can be expected that significant departures from zero of the median level-wise sample correlation would occur, which will allow the detection of the presence of correlation between the signals at a particular scale.

(v) In Fig. 6.6 the effect of perfect correlation a each scale level on the sample behavior of the scale correlation is exemplified, using $N = 256$, $\beta = 1$ and $\sigma = 0.1$. It is interesting to observe that the coarser the scale, the less significant the effect of the correlation.

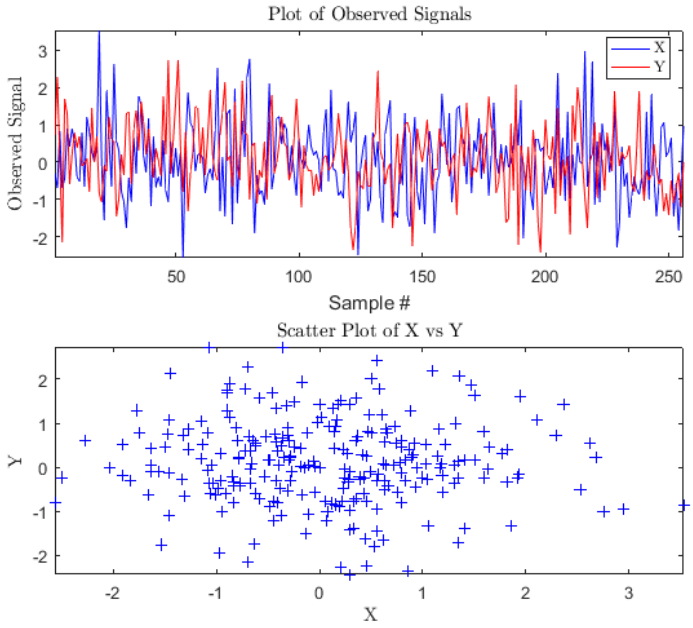


(a) Uncorrelated signals in the wavelet domain

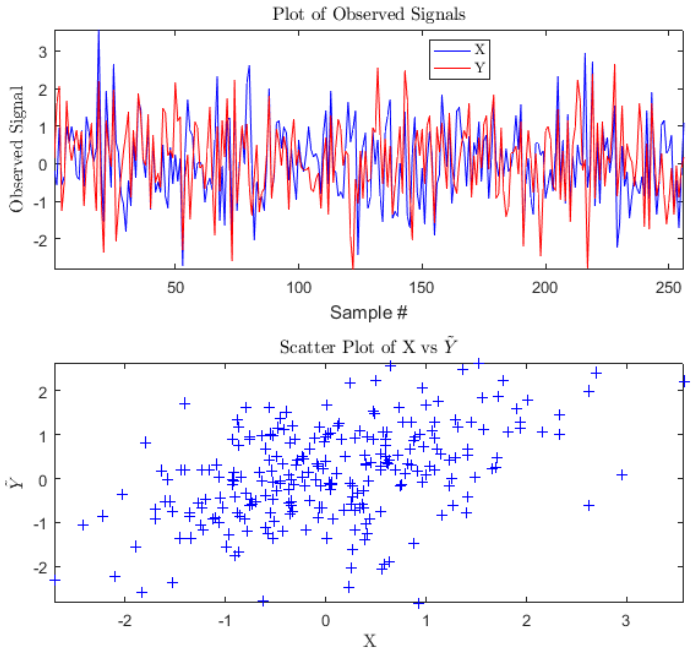


(b) Correlated signals in the wavelet domain, at scale level=7

Figure 6.2: Comparative Plots of uncorrelated (a) vs. correlated (b) signals in the wavelet domain.

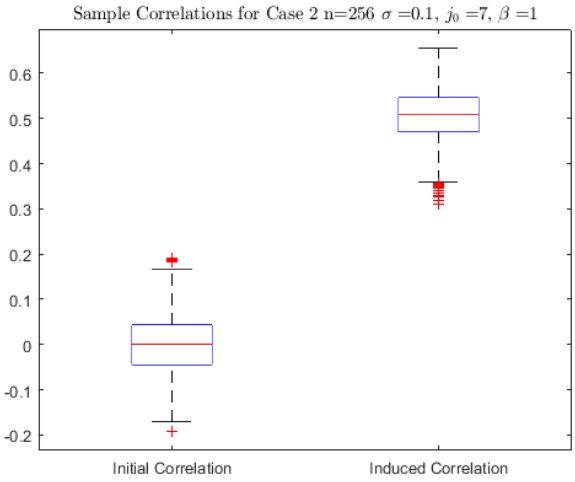


(a) Uncorrelated signals in the time domain

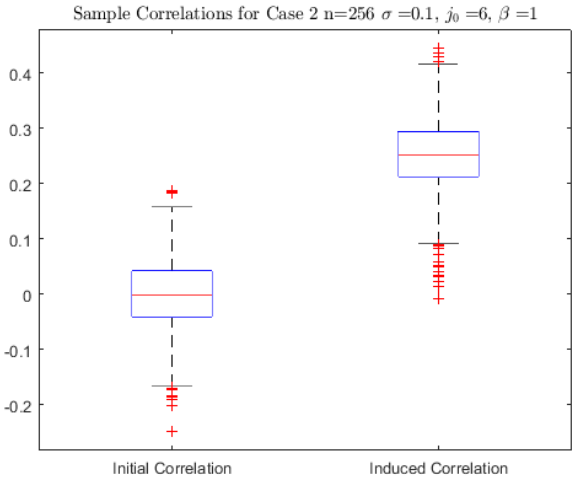


(b) Correlated signals in the time domain.

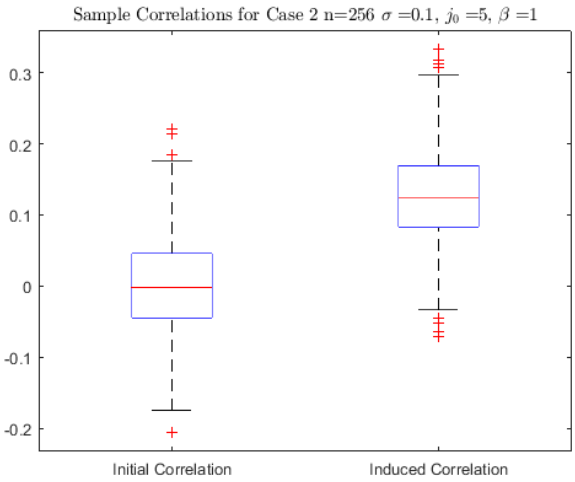
Figure 6.3: Comparative Plots of uncorrelated (a) vs. correlated (b) signals in the time domain, with perfect correlation at scale-level 7.



(a) Perfect correlation at $J = 7$

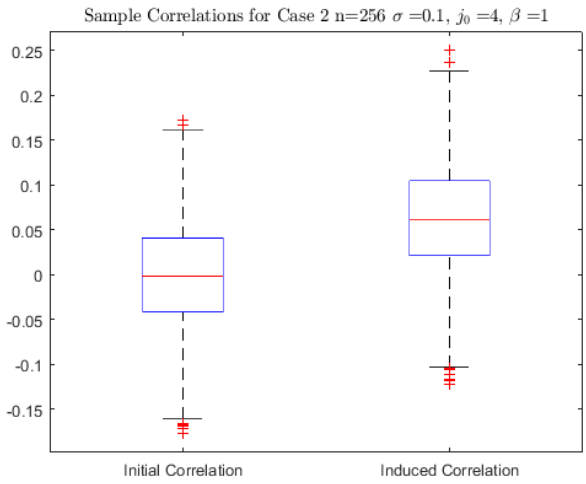


(b) Perfect correlation at $J = 6$

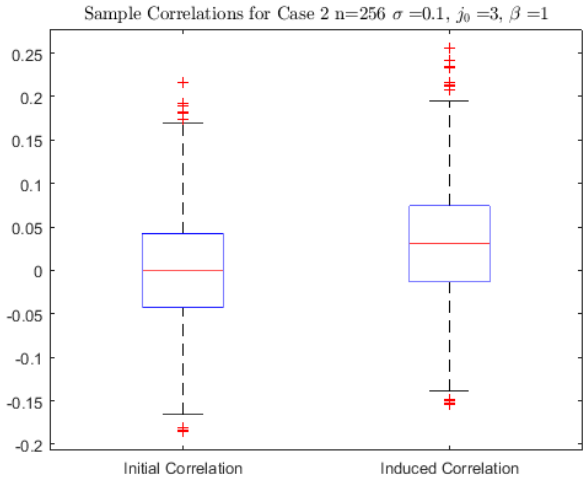


(c) Perfect correlation at $J = 5$

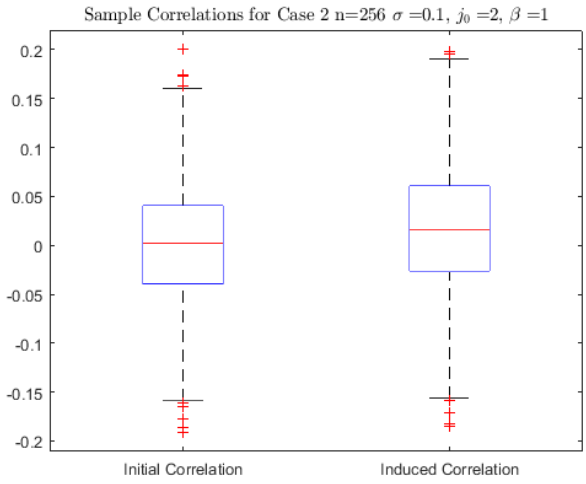
Figure 6.4: Comparative boxplots of the typical effects on overall correlation at the original domain, given perfect correlation at the wavelet domain. The experiments were replicated 1000 times.



(a) Perfect correlation at $J = 4$

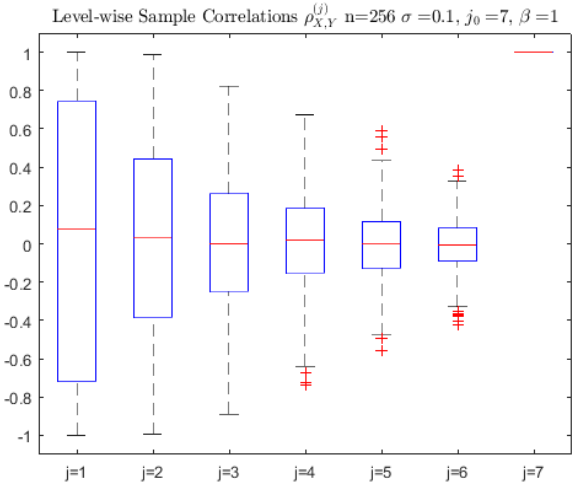


(b) Perfect correlation at $J = 3$

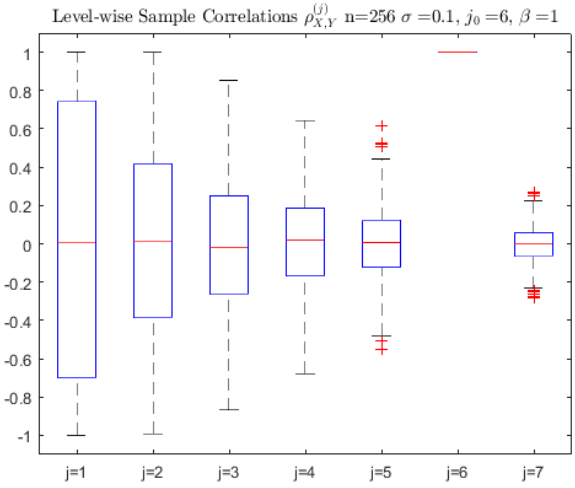


(c) Perfect correlation at $J = 2$

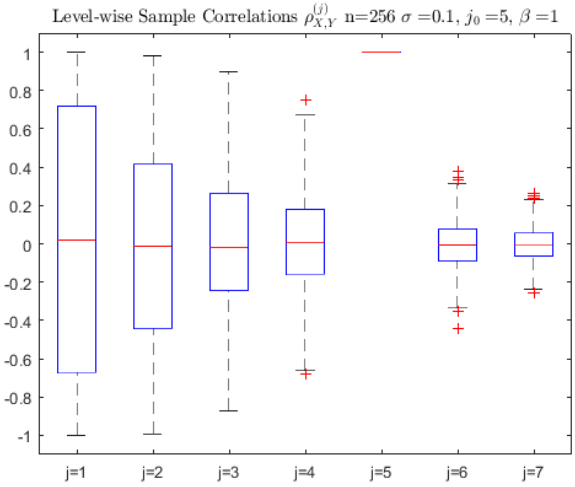
Figure 6.5: Comparative boxplots of the typical effects on overall correlation at the original domain, given perfect correlation at the wavelet domain. The experiments were replicated 1000 times.



(a) Perfect correlation at $J = 7$

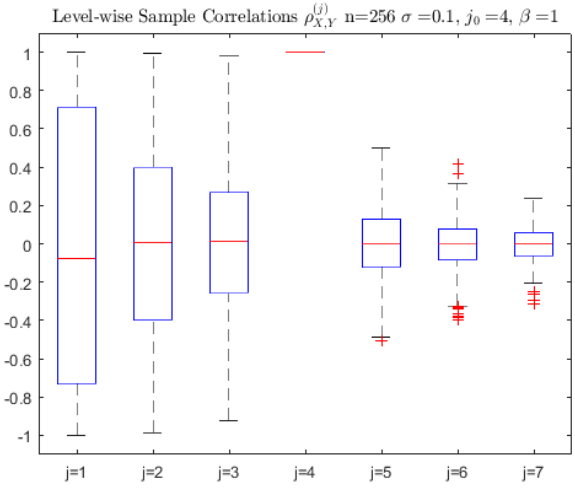


(b) Perfect correlation at $J = 6$

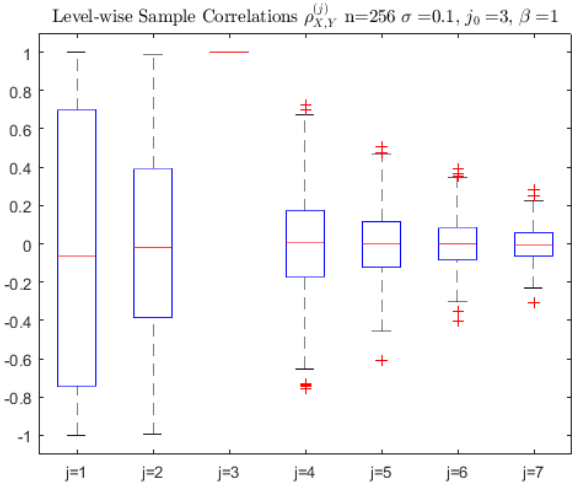


(c) Perfect correlation at $J = 5$

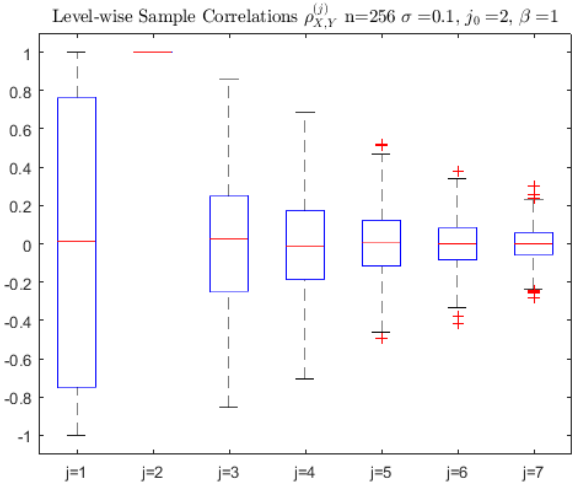
Figure 6.6: Comparative boxplots of the typical effects on scale correlations in the wavelet domain, given perfect correlation at each scale level. The experiments were replicated 1000 times.



(a) Perfect correlation at $J = 4$



(b) Perfect correlation at $J = 3$



(c) Perfect correlation at $J = 2$

Figure 6.7: Comparative boxplots of the typical effects on scale correlations in the wavelet domain, given perfect correlation at each scale level. The experiments were replicated 1000 times.

6.3.3 Case 3: Perfect correlation between $\mathbf{x}$ and $\mathbf{y}$ at a all multiresolution levels $0 \leq j \leq J-1$, and its translation into the original signal domain correlation.

This case is a generalization of the previous situation. Here, instead of just one level perfectly correlated, it is assumed that all levels are perfectly correlated between the two signal.

Suppose that for a fixed $j = 0, ..., J-1$, there exists $\beta_j \neq 0$, such that:

$$\mathbf{d}_Y^{(j)} = \beta_j \cdot \mathbf{d}_X^{(j)}.$$

This relation implies that at the multiresolution level j, samples $\mathbf{x}$ and $\mathbf{y}$ are linearly dependent. Note that under this condition, it follows for $j = 0, ..., J-1$:

$$\begin{aligned} \hat{\rho}_{X,Y}^{(j)} &= sign(\beta_j), \\ w_j &= \frac{|\beta_j|}{||\mathbf{d}_X||_2 \sqrt{C_j + \beta_j^2}}, \\ C_j &= \frac{1}{||\mathbf{d}_X^{(j)}||_2^2} \left(||c_Y^{(0)}||_2^2 + \sum_{j=0,\ j \neq j_0}^{J-1} ||\mathbf{d}_Y^{(j)}||_2^2 \right). \end{aligned}$$

Therefore, the sample correlation takes the form:

$$\hat{\rho}_{X,Y} = \frac{c_X^{(0)} c_Y^{(0)}}{||\mathbf{d}_X||_2 ||\mathbf{d}_Y||_2} + \frac{1}{||\mathbf{d}_X||_2} \sum_{j=0}^{J-1} \frac{\beta_j}{\sqrt{C_j + \beta_j^2}} \tag{6.13}$$

Here, similarly as in the previous case, $C_j > 0$ represents the ratio of energies contained in all levels but j in signal $\mathbf{y}$ and the energy contained at level j in signal $\mathbf{x}$. Note that $C_j = 0$ when all energy in the signal is concentrated in level j, and $C_j \to \infty$ when no energy is contained at scale j.

Using the same methodology as in Case 2, the following Figures illustrate typical qualitative behavior of the correlation structure of the signals:

As it can be observed in Fig. 6.8 it is interesting to note the fact that although all scales are perfectly correlated, the overall correlation (b) is not exactly 1. This is due to the fact that $\sum_{j=0}^{J-1} w_j \neq 1$. However, comparing the scale-correlations between the original uncorrelated signals with the perfectly-scale correlated samples, it is clear that there is a significant difference between the samples. This is evident after inspecting Fig. 6.8b.

6.3.4 Some Theoretical Properties of wavelet coefficients for stationary, finite energy processes.

In this section, some of the properties of wavelet coefficients obtained from the orthogonal DWT for stationary, finite energy sequences are studied. In particular, we analyze stationarity and dependency between the expansion coefficients resulting from an orthogonal DWT. Before begining with this analysis, the following definitions are needed:

Definition 6.3.1. *A stochastic process* $\{X(t),\ t \in \mathbb{R}\}$ *is called* ***strictly stationary*** *if, for every* $n \in \mathbb{Z}$*, every permutation of* $t_1, ..., t_n\ \in \mathbb{R}$*, and every lag* $\tau \in \mathbb{R}$*, it holds:*

$$(X(t_1 + \tau), ..., X(t_n + \tau)) \overset{\mathcal{D}}{=} (X(t_1), ..., X(t_n))\,. \tag{6.14}$$

Here $\overset{\mathcal{D}}{=}$ denotes "equal in distribution". Clearly, a sequence $X_{t_1}, ..., X_{t_n}$ of iid random variables, satisfying $\mathbb{E}[X] = \mu$, and $Var(X) = \sigma^2 < \infty$, is strictly stationary. Similarly, note that if the following conditions are satisfied:

(i) $\mathbb{E}[|X(t)|^2] < \infty$.

(ii) $\mathbb{E}[X(t)] = \mu$.

(iii) For all $s, t\ \in \mathbb{R}$, $Cov(X(t), X(s)) = \gamma_X(s - t) < \infty$.

Then, the process $\{X(t),\ t \in \mathbb{R}\}$ is said to be **weakly and second order stationary**. Note that these properties do not necessarily imply (6.14).

Definition 6.3.2. *Consider two second order, real-valued, zero mean, weakly stationary processes* $\{X(t),\ t \in \mathbb{R}\}$, $\{Y(t),\ t \in \mathbb{R}\}$. *Then if for all* $t,\ s,\ \tau\ \in \mathbb{R}$*:*

$$\mathbb{E}[X(t+\tau)Y(s+\tau)] = \mathbb{E}[X(t)Y(s)] = \gamma_{XY}(t-s), \tag{6.15}$$

then the processes are said to be ***cross (weakly) stationary****.*

Definition 6.3.3. *Suppose a process* $\{X(t),\ t \in \mathbb{R}\}$ *that has mean zero and is at least weakly stationary. Then, if for every* n*, integer, finite,* $\tau \in \mathbb{R}$*, and any permutation* $(t_1, ..., t_n)$ *where* $t_1 < t_2 < \cdots < t_n$ *it holds:*

$$\sum_{k=1}^{n} |X(t_k + \tau)|^2 < \infty, \tag{6.16}$$

then, the process is said to have finite energy.

Note that this condition states that at every window of finite length, resulting from any scale of observations, the energy contained in the signal is finite. In particular, this condition must be satisfied in order for the multiscale correlation decomposition to be well-defined. In fact, since the proposed method is based on the orthogonal DWT, energies are preserved.

Definition 6.3.4. *A process* $\{X(t),\ t \in \mathbb{R}\}$ *is said t have* ***stationary increments*** *if for every vector* $\mathbf{h} \in \mathbb{R}^K,\ K < \infty$*, it holds:*

$$(X(t+h_1) - X(t), ..., X(t+h_K) - X(t)) \overset{\mathcal{D}}{=} (X(h_1) - X(0), ..., X(h_K) - X(0))\,. \tag{6.17}$$

Definition 6.3.5. *Suppose a stochastic process* $\{X(t),\ t \in \mathbb{R}\}$ *that is weakly stationary. Then, if* $\forall\ t \in \mathbb{R}$*:*

$$\lim_{|h| \to \infty} \gamma_X(h) = \lim_{|h| \to \infty} \mathbb{E}[X(t)X(t-h)] = 0, \tag{6.18}$$

*then, it has **finite memory**. This implies that it exists $h_0 \in \mathbb{R}$ large enough such that $\forall\ h > h_0, \forall\ t \in \mathbb{R}$ the random variables $X(t),\ X(t+h)$ can be considered uncorrelated.*

Definition 6.3.6 (Averkamp and Houdre (2000)[82])**.** *Let $\{X(t),\ t \in \mathbb{R}\}$ that is weakly stationary process, with auto-correlation function $\gamma_X(t,s)$. The DWT of $X(t)$ is a discrete random field given by:*

$$\{d_{j,k},\ j,k\ \in \mathbb{Z}\} = \left\{ \int_{\mathbb{R}} X(t)\psi_{jk}(t)dt,\ j,k\ \in \mathbb{Z} \right\}, \tag{6.19}$$

which is well-defined if the above path integrals are well defined (i.e. the integral converges with probability one), and (as noted in [82]):

$$\int_{\mathbb{R}} \sqrt{\gamma_X(t,t)}|\psi_{jk}(t)|dt < \infty.$$

Note that the coefficients $d_{j,k}$ contain the information about contiguous scales centered around scale 2^j, and time instant $k \cdot 2^j$ being the discretization of the CWT discussed in Chapter 1. If (6.19) is well-defined, then:

$$\mathbb{E}[d_{j,k}d_{j',k'}] = \int_{\mathbb{R}} \gamma_X(t,s)\psi_{j,k}(t)\psi_{j',k'}(s)dtds, \tag{6.20}$$

is well defined as well.

Some properties of wavelet coefficients resulting from stationary processes.

Lemma 6.3.1. *Suppose a process $\{X(t),\ t \in \mathbb{R}\}$ that has stationary increments. Then, the sequence of wavelet coefficient $\{d_{jk},\ k = 0, ..., 2^j - 1\}$ is stationary.*

Proof. Suppose a wavelet function $\psi(t)$. Each wavelet coefficient d_{jk} is obtained as:

$$\begin{aligned} d_{jk} &= \int X(t)\psi_{jk}(t)dt, \\ &= \int X(t-\delta)\psi_{jk}(t-\delta)dt, \ \text{ set } \delta = 2^{-j}v, \ v \in \mathbb{Z}, \\ &= 2^{j/2}\int X(t-2^{-j}v)\psi(2^j t-(k+v))dt. \end{aligned}$$

Since $\int \psi(t)dt = 0$ and $(X(t-2^{-j}v) - X(2^{-j}v)) \stackrel{\mathcal{D}}{=} (X(t) - X(0))$, it follows:

$$d_{jk} \stackrel{\mathcal{D}}{=} d_{j,k+v}, \ v \in \mathbb{Z},$$

which implies that $\{d_{jk}, \ k = 0, ..., 2^j - 1\}$ is stationary. Note that this result does not necessarily hold in the reverse direction. □

Note that as illustrated in Chapter 9, Vidakovic (1999)[3], the following two results relate the aforementioned Lemma to more general classes of processes:

(a) Lemma 9.2.1 [3]. If $X(t), \ t \in \mathbb{R}$ is a weakly stationary process, for $l, \ n \in \mathbb{Z}$ and $j \geq l$, the random sequence $\left\{d_{j,2^{j-l}k+n}, \ k \in \mathbb{Z}\right\}$ is weakly stationary as well.

(b) Theorem 9.2.1 [3]. If $X(t), \ t \in \mathbb{R}$ is a second order stationary process for which $\gamma_X(s,t)$ is bounded and continuous in $\mathbb{R}^2$, then the sequence $\{d_{j,k}, \ k \in \mathbb{Z}\}$ is weakly stationary iff $X(t)$ is weakly stationary. In particular, if the wavelet basis is compactly supported, then the condition on the boundedness of $\gamma_X(s,t)$ can be relaxed. The proof of this Theorem can be found in Averkamp (2000)[82].

Lemma 6.3.2. *Suppose a White Noise (WN) process* $\{X(t), \ t \in \mathbb{R}\}$*, where* $X(t) \stackrel{iid}{\sim} \mathcal{N}(0, \sigma^2)$*, for* $\sigma^2 < \infty$*. Then for a fixed* j*,* $k = 0, ..., 2^j - 1$ *the wavelet coefficients* d_{jk} *satisfy::*

$$d_{j,k} \stackrel{\mathcal{D}}{=} 2^{-j/2}d_{0,k} \stackrel{\mathcal{D}}{=} 2^{-j/2}d_{0,0},$$

where the wavelet function $\psi(t)$ is compactly supported and satisfies $|\psi(t)| < \infty$, for every $t \in \mathbb{R}$.

Proof. Since the WN process $\{X(t),\ t \in \mathbb{R}\}$ is strictly stationary, by the definition of wavelet coefficients, it follows:

$$\begin{aligned} d_{j,k} &= \int X(t)\psi_{j,k}(t)dt, \\ &= 2^{-j/2}\int X(2^{-j}u)\psi(u-k)du, \\ &\stackrel{\mathcal{D}}{=} 2^{-j/2}\int X(u)\psi(u-k)du = 2^{-j/2}d_{0,k}, \\ &= 2^{-j/2}\int X(z+k)\psi(z)dz, \\ &\stackrel{\mathcal{D}}{=} 2^{-j/2}\int X(z)\psi(z)dz \ = \ 2^{-j/2}d_{0,0}, \end{aligned}$$

where the last result holds from the fact that for every $u \in \mathbb{R}$, $X(2^{-j}u) \stackrel{\mathcal{D}}{=} X(u)$. □

Lemma 6.3.3. *Suppose a WN process sampled at regularly spaced integer-valued intervals $\{X(n),\ n \in \mathbb{N}\}$, where $X(n) \stackrel{iid}{\sim} \mathcal{N}(0, \sigma^2)$. Then, the sequence of wavelet coefficients resulting from the DWT using an orthogonal wavelet family satisfy:*

(i) $d_{j,k} \stackrel{\mathcal{D}}{=} d_{0,k} \stackrel{\mathcal{D}}{=} d_{0,0}$.

(ii) $d_{0,0} \sim \mathcal{N}(0, \sigma^2)$

Proof. Note that by the definition of the DWT (see section 1.1.8), the scaling and wavelet filters $\mathbf{h} = [h(n)],\ n = 1, ..., L$, $\mathbf{g} = [g(n)],\ n = 1, ..., L$ satisfy:

$$\begin{aligned} \sum_n h(n) = \sqrt{2}, \quad & \sum_n h(n)^2 = 1, \\ \sum_n g(n) = 0, \quad & \sum_n g(n)^2 = 1. \end{aligned}$$

Since the DWT of the vector $\mathbf{X} = [X(1) \dots X(N)]^T \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_N)$ can be obtained by the linear transformation $\mathbf{d} = W \cdot \mathbf{x}$, where $W_{N \times N}$ an orthogonal matrix, it follows:

$$W \cdot \mathbf{x} \sim \mathcal{N}(\mathbf{0}, \sigma^2 W \cdot W^T),$$

which implies (due to the orthogonality condition of W), that (i) and (ii) follow. □

Lemma 6.3.4. *Consider a second order, zero mean weakly stationary process* $\{X(t),\ t \in \mathbb{R}\}$ *with bounded* $\gamma_X(s,t)$*, for every* $s,\ t\ \in \mathbb{R}$*. Then for* $j = 0, ..., J-1$*, the following results hold:*

(i) $\mathbb{E}[d_{j,k}] = 0$ *for* $k = 0, ..., 2^j - 1$.

(ii) $Var(d_{j,k}) = \mathbb{E}[d_{j,k}^2] = C_{\psi,X}^{(j)} < \infty$*, for* $k = 0, ..., 2^j - 1$,

provided the coefficients $\{d_{j,k}\ k \in \mathbb{Z}\}$ *are well defined.*

Proof. Assuming the process $\{X(t),\ t \in \mathbb{R}\}$ is second order stationary, and the wavelet basis $\{\psi_{j,k},\ j,k \in \mathbb{Z}\}$ is orthonormal with compact support, it follows from the dominated convergence theorem, and the fact that $\int \psi(t)dt = 0$:

$$\mathbb{E}[d_{jk}] = \int \mathbb{E}[X(t)]\psi_{jk}(t)dt = 0.$$

This result shows that (i) holds. Also, it implies that $Var(d_{j,k}) = \mathbb{E}[d_{j,k}^2]$. Now, from the definition of the wavelet coefficients and the orthogonality of the basis, it follows:

$$\begin{aligned} \mathbb{E}[d_{j,k}^2] &= \int\int \gamma_X(s-t)\psi_{j,k}(t)\psi_{j,k}(u)dtdu, \\ &= 2^{-j}\int\int \gamma_X(2^{-j}(s-w))\psi(s)\psi(w)dsdw = C_{\psi,X}^{(j)}, \end{aligned}$$

where the last equation results from a change of variables in the integration and does not

depend on k. The expression on the rhs is a function only of the scale level j and $\gamma_X(\cdot)$, which implies that for $k = 0, ..., 2^j - 1$ the variance of the wavelet coefficients takes the same form. Thus, (ii) follows. □

Theorem 6.3.1 (Walter (1994)[83], Vidakovic (1999)[3]). *Let $\{X(t),\ t \in \mathbb{R}\}$ be a stationary process and let $X_J(t)$ be its projection onto the multiresolution space V_J spanned by $\{\phi_{J,k}(t),\ k \in \mathbb{Z}\}$. If the scaling function is $r-$regular, then:*

$$\begin{aligned} \mathbb{E}[|X(t) - X_J(t)|^2] &\rightarrow 0, \text{ when } J \rightarrow \infty, \text{ and} \\ \mathbb{E}[d_{j,k}d_{j',k'}] &= \frac{1}{2\pi}\int_{\mathbb{R}} \hat{\gamma_X}(w)\Psi\left(\frac{w}{2^j}\right)\Psi\left(\frac{\bar{w}}{2^{j'}}\right)e^{-iwk2^{-j}}e^{-iw'k2^{-j'}}2^{-(j+j')/2}dw, \end{aligned}$$

where $\hat{\gamma_X}$ and Ψ are the Fourier transformations of γ_X and ψ respectively. Using this last result, if the wavelet basis is of the Meyer type (see [3]), such that both $\hat{\gamma_X}$ and Ψ are in the space $\mathbb{C}^p,\ p > 1$ (i.e. the space of $p-$times continuously differentiable functions), then the coefficients defined in (6.19) satisfy:

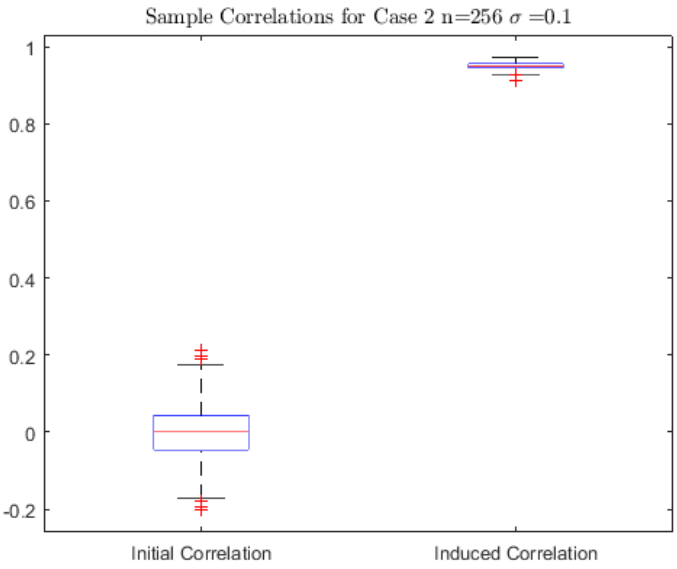
(i) If $|j - j'| > 1$, $d_{j,k}$ and $d_{j',k'}$ are uncorrelated.

(ii) If $|j - j'| = 1$, $d_{j,k}$ and $d_{j',k'}$ have arbitrarily small correlation.

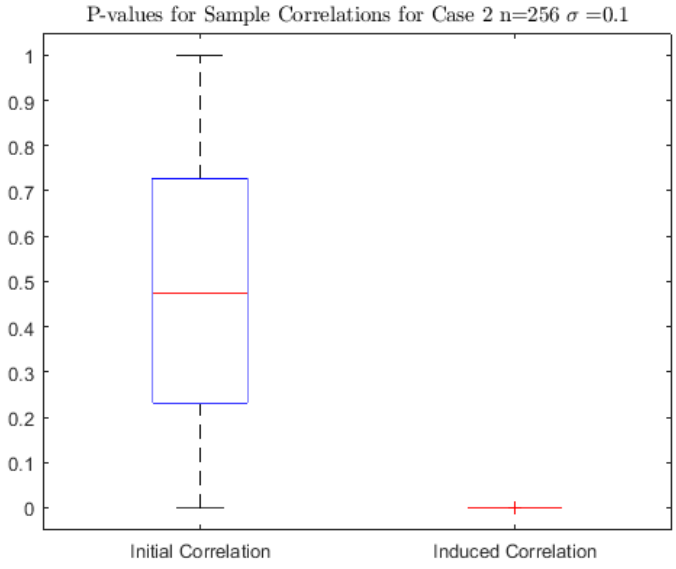
(iii) If $j = j'$, then $d_{j,k}$ and $d_{j',k'}$ have correlation that is of the order $\mathcal{O}(|k - k'|^{-p})$.

The proof of this theorem can be found in [83].

In the next section, the aforementioned results will be used as the foundation for the development of test statistics that can be utilized in the assessment of the significance of level-wise correlations resulting from its multiscale representation.



(a) Perfect correlation at $j = 1, ..., 7$



(b) Perfect correlation at $j = 1, ..., 7$

Figure 6.8: Comparative box-plots of the typical effects on scale correlations in the wavelet domain and time domain, given perfect correlation at each scale level. (a) illustrates the usual sample correlation in the time domain, (b) shows the corresponding p-values for the test-statistic. The experiments were replicated 1000 times. Values for β_j were chose at random.

6.4 Statistical Tests for Multi-scale Correlation in the Wavelet Domain Based on the Whitening Property of DWT

In a similar way to their time-domain counterparts, multiscale correlations can be tested for significance on scale-dependent basis. In this section, we propose two scale-dependent test statistics designed to assess the significance of the obtained sample correlations in the wavelet domain. These tests are constructed from both a parametric and non-parametric perspective and their performance is compared with well known test statistics such as: *T-test*, *Spearman rank correlation* and *Kendall's rank correlation* using a simulation study.

In particular, the performance comparison is made in terms of the estimated probability of type I and type II error for stationary models of the kind: $AR(1)$, $MA(1)$, $ARMA(1,1)$. Even though the number of different models that can be encountered in practice is extremely large, we restrict the simulation study to these models because they tend to cover a range of stochastic behavior that is likely to be observed in real life situations. Also, the methodology presented can be easily extended to more sophisticated models, which allows the interested reader to implement and extend these results beyond what is contained in this section.

6.4.1 Student Test for Normally Distributed Random Variables

Lemma 6.4.1 (Student (1908)[84], Kendall (1938)[85])**.** *Assume the observed sequences satisfy:*

$X_1, ..., X_N \overset{iid}{\sim} \mathcal{N}(0, \sigma_X^2)$*, and* $Y_1, ..., Y_N \overset{iid}{\sim} \mathcal{N}(0, \sigma_Y^2)$ *and are uncorrelated. Then, for* $j = 0, ..., J-1$ *the level-wise correlations* $\hat{\rho}_{X,Y}^{(j)}$ *it is possible to define sample statistics* V_j *such that:*

$$V_j = \frac{2^{j/2}}{\sqrt{1-(\hat{\rho}_{X,Y}^{(j)})^2}} \hat{\rho}_{X,Y}^{(j)} \sim t_{2^j}, \tag{6.21}$$

where t_{2^j} *denotes a* $t-$*distribution with* 2^j *degrees of freedom.*

This result follows from the properties of the Normal distribution and the orthogonality nature of the DWT. Also, it enables the use of standard hypothesis testing:

$$\begin{aligned} H_0 &: \hat{\rho}^{(j)}_{X,Y} = 0 \\ H_1 &: \hat{\rho}^{(j)}_{X,Y} \neq 0 \end{aligned}$$

Thus, for a fixed $0 < \alpha < 1$, if $|V_j| > t_{\frac{\alpha}{2},2^j}$ then the null hypothesis is rejected at the $(1-\alpha)-$level of significance. In Fig. (6.9) different shapes of the $t-$distribution are depicted,illustrating the behavior of its tails with respect to the number of degrees of freedom ($\nu > 0$). Note that as ν grows, the distribution approximates to a standard normal. This implies, that for values of $j > 30$ it is possible to utilize the usual $Z-$statistic instead.

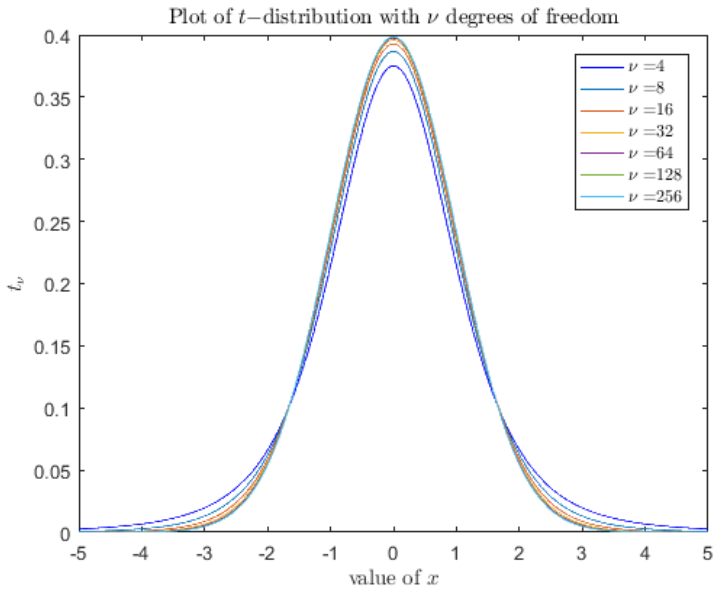


Figure 6.9: Plot of $t-$distribution for 2^j degrees of freedom, $j = 1, ..., J-1$. Note that $\nu > 30$ the distribution closely approximates to a $\mathcal{N}(0,1)$ distribution.

6.4.2 A Local Test Statistic Based on the Distributional Structure of Wavelet Coefficients for Stationary Sequences, Assuming Normality

Assume that the sequence of wavelet coefficients $\{d_{j,k},\ k = 0, ..., 2^j - 1\}$ resulting from the DWT of a stationary process $\{X(t),\ t \in \mathbb{R}\}$ is distributed as:

$$d_{j,k} \sim \mathcal{N}(0, \sigma_j^2),\ k = 0, ..., 2^j - 1,\ j = 0, ..., J - 1. \tag{6.22}$$

Here, the assumption of first and second moment being the same for all coefficients in the same scale results from Lemma 6.3.4. Now, the assumption about normality although it can be considered too strong (especially in the case of heavy-tailed processes) it can be argued to be reasonable due to the following facts:

(i) In the case of zero mean, second order stationary gaussian processes (i.e. constant variance as a function of time), the application of the DWT produces a multivariate gaussian vector with zero mean and covariance structure given by $W\Sigma W^T$, where Σ is the covariance matrix of the random vector $\mathbf{X} = [X(1) \dots X(N)]^T$. As seen from Theorem 6.3.1, under certain conditions, these matrix is highly likely to be close to diagonal.

(ii) Similarly, a process that is stationary with zero mean when viewed in terms of its empirical distribution, under certain conditions (e.g. symmetry and rate of decay of the tails) it can be approximated by a normal distribution with variance equivalent to the process variance within the observed window. This approximation, however is not adequate in the case when the process is not symmetric around its mean and/or when its tails decay either faster or slower (especially slower) than the Normal distribution (as in the case of the Double Exponential, t-distribution or Cauchy).

(iii) An example of this situation can be found in processes that are zero mean, station-

ary and exhibit a high oscillatory behavior during short-time intervals. A time series generated from an AR(1) process with parameter $\phi = 0.95$ has this kind of features (see Figs.6.16a and 6.16b for a proper illustration of this behavior). In these cases, the wavelet coefficients at high scale levels exhibit an empirical behavior that suggests a significant departure from normality, exhibiting asymmetry and heavy tails.

(iv) The normality assumption can also find a more rigorous ground in the results shown by Cohen et al. (2015)[86], in which wavelet coefficients resulting from the orthogonal DWT of iid processes generated from different distributions (e.g. Uniform, Exponential, Gamma, Weibull, Rayleigh) exhibit normality with constant mean and variance. This implies that for signals that are close to white noise (WN), it is relatively safe to assume normality of the wavelet coefficients.

Now, as shown in Theorem 6.3.1, for a process with autocovariance function $\gamma_X(h)$ and a wavelet function $\psi(t)$ with sufficiently smooth spectral densities, then the wavelet coefficients corresponding to the same scale levels can be considered uncorrelated, provided the distance between them with respect to the shift k is sufficiently large. The smoother the respective Fourier transforms, the faster the decay of the correlation between contiguous wavelet coefficients.

Consider two sequences $\{X(n),\ n \in \mathbb{N}\}$ and $\{Y(n),\ n \in \mathbb{N}\}$ that are zero mean, second order stationary. Provided condition (6.22) holds and $\{X(n),\ n \in \mathbb{N}\}$ and $\{Y(n),\ n \in \mathbb{N}\}$ are uncorrelated in the wavelet domain, it is possible to define test statistics $T_{1,j,k},\ T_{2,j,k}$ for $j = 0, ..., J-1$, $k = 0, ..., 2^j - 1$:

$$T_{1,j,k} = 2^{j/2} \frac{\sigma_j^{(X)}}{\sigma_j^{(Y)}} \frac{d_{j,k}^{(Y)}}{||\mathbf{d}_j^{(X)}||_2}, \tag{6.23}$$

$$T_{2,j,k} = 2^{j/2} \frac{\sigma_j^{(Y)}}{\sigma_j^{(X)}} \frac{d_{j,k}^{(X)}}{||\mathbf{d}_j^{(Y)}||_2}. \tag{6.24}$$

Under the assumption of normality, no correlation (independence) and $\sigma_j^{(X)},\ \sigma_j^{(Y)}$ known, by Cochran's Theorem (1934)[87], it follows that $T_{1,j,k}$ and $T_{2,j,k}$ are distributed as t_{2^j} (i.e. t distribution with 2^j degrees of freedom).

Now, for a pre-specified confidence level $0 < \alpha < 1$, under the null hypothesis H_0 of no-correlation between the wavelet sequences, it is possible to reject H_0 if $|T_{1,j,k}| > t_{1-\frac{\alpha}{2},2^j}$, where this last term corresponds to the $1-\alpha/2$ quantile of the $t-$ distribution with 2^j degrees of freedom. This definition of the test statistic has the advantage that for each level it is possible to obtain multiple test statistics that can be used to assess the significance of the observed sample correlation in the wavelet domain. Another advantage is the fact that since wavelets capture local behavior, it may be possible that correlation exists only between coefficients belonging to a particular subset of the shifts $k = 0, ..., 2^j - 1$. For this reason, using a statistic that utilizes all the coefficients combined (e.g. the average) could lead to the loss of locality and therefore, a loss in sensitivity of the test.

Using definitions (6.23) and (6.24), the critical value $t_{1-\frac{\alpha}{2},2^j}$, it is possible to define:

$$\begin{aligned}
\mathbf{I}_{1,j} &= [\mathbf{1}_{1,j,0} \dots \mathbf{1}_{1,j,2^j-1}], \\
\mathbf{I}_{2,j} &= [\mathbf{1}_{2,j,0} \dots \mathbf{1}_{2,j,2^j-1}], \ \text{where} \\
\mathbf{1}_{1,j,k} &= \begin{cases} 1 & \text{if } |T_{1,j,k}| > t_{1-\frac{\alpha}{2},2^j} \\ 0 & \text{if } |T_{1,j,k}| \leq t_{1-\frac{\alpha}{2},2^j} \end{cases}, \ \text{for } k = 0, ..., 2^j - 1.
\end{aligned}$$

In particular, note that for $k = 0, ..., 2^j - 1$ the random variables $\mathbf{1}_{1,j,k}$ are iid $Bernoulli(p_{j,\alpha})$, where $p_{j,\alpha} = \mathbb{P}\left(|T_{1,j,k}| > t_{1-\frac{\alpha}{2},2^j}\right)$. Moreover, since the two random vectors $\mathbf{I}_{1,j}$ and $\mathbf{I}_{2,j}$ are not independent, it is possible to expect a certain level of agreement between them. This means that for each scale level $j = 0, ..., J-1$ it is possible to construct a table of the form: Therefore, the decision about rejection (or fail to reject) the null hypothesis H_0 can be

Table 6.1: Agreement table for local significance test

H_0	0	1
0	$\sum_{k=0}^{2^j-1} \mathbf{1}_{(\mathbf{1}_{1,j,k}=0,\mathbf{1}_{2,j,k}=0)}$	$\sum_{k=0}^{2^j-1} \mathbf{1}_{(\mathbf{1}_{1,j,k}=0,\mathbf{1}_{2,j,k}=1)}$
1	$\sum_{k=0}^{2^j-1} \mathbf{1}_{(\mathbf{1}_{1,j,k}=1,\mathbf{1}_{2,j,k}=0)}$	$\sum_{k=0}^{2^j-1} \mathbf{1}_{(\mathbf{1}_{1,j,k}=1,\mathbf{1}_{2,j,k}=1)}$

made by inspecting the entries of Table (6.1). For two wavelet coefficient sequences that are uncorrelated, we would expect that $\sum_{k=0}^{2^j-1} \mathbf{1}_{(\mathbf{1}_{1,j,k}=0,\mathbf{1}_{2,j,k}=0)}$ is large and close to 2^j. Similarly, for two signals that exhibit correlation at some shifts k, it can be expected that $\sum_{k=0}^{2^j-1} \mathbf{1}_{(\mathbf{1}_{1,j,k}=0,\mathbf{1}_{2,j,k}=0)} + \mathbf{1}_{(\mathbf{1}_{1,j,k}=1,\mathbf{1}_{2,j,k}=1)}$ is larger than a certain threshold. In particular, if the correlation in the wavelet domain for a certain scale is well spread across all shifts k, it can be expected that this entry has a value that is close to 2^j.

This, last statement constitutes just a hypothesis that needs to be validated via a proper simulational study, since a theoretical derivation of the distribution of the entries of Table 6.1 under the alternative hypothesis for stationary stochastic processes seems, at a first glance, an extremely challenging task that even though interesting in itself, may not offer any significant advantages for practical purposes over the insights obtained from a simulation study.

In the case of the entries of (6.1) related to pairs $(\mathbf{1}_{1,j,k}, \mathbf{1}_{2,j,k})$ that are not concordant, since both indicator variables are dependent (because of the construction of the test statistic), it can be expected under H_0 that their magnitudes should be similar and small, because they most likely result from random effects. In fact, assuming H_0 holds, $\mathbb{P}(T_{1,j,k} = 0, T_{2,j,k} = 1|H_0) \leq \alpha$. For this reason, it can be expected that the empirical distributions of these entries should be fairly similar.

As can be expected, in practical applications even though the signals may not be correlated, due to violations of the normality assumption of the wavelet coefficients and/or due to the possible discontinuities of the spectral densities of the processes $\{X(t),\ t \in \mathbb{R}\}$, $\{Y(t),\ t \in \mathbb{R}\}$, together with possible numerical effects, the entries of Table (6.1) can have exhibit a certain degree of variability that needs to be accounted for. For this reason, the definition a proper

decision threshold using an empirical approach seems adequate. In particular, due to the wide variety of stochastic processes that could observed in reality, a comprehensive simulation study would be extremely challenging. However, restricting the type of processes to those that are encountered more frequently could be a good starting point towards this goal.

Simulation-Based Elicitation of a Decision Threshold for Testing H_0 against H_1

As was previously mentioned, due to the underlying randomness of data in real life, and the wide variety of stochastic processes that can be observed, defining empirical-based critical values seems a reasonable approach. In particular, we will focus on decision values for processes of the form AR(1), MA(1) and ARMA(1,1); nonetheless, the methodology utilized can be easily extended to different kinds of models such as ARMA(p,q), ARIMA, etc.
The goal of this empirical-based study will be to analyze the empirical behavior of the entries of Table 6.1, under the hypothesis that no correlation between signals in the wavelet domain exists, and utilizing as critical value $t_{1-\frac{\alpha}{2},2^j}$ with $\alpha = 0.05$. For this purpose, a simulation study was conducted in which several replications of each model with different parameters were run (the details of the models and the utilized parameters is shown in Section 6.4.4.

The following plots depict representative empirical distributions for each one of the entries of Table 6.1:

From Fig. 6.10, it is possible to observe the following:

(i) From panel 6.10a, the proportion of pairs of the type (0,0) exceeds 95% with high probability.

(ii) From panel 6.11a, the proportion of pairs of the type (1,1) is negligible, with a 0.3% observed in less than 1% of the replications. This suggests that a small critical value could be utilized for the test.

(iii) From panels 6.10b and 6.10c, it is possible to observe that the statistical behavior of the proportion of pairs of the type (0,1)-(0,1) is significantly similar. Both empirical histograms (out of 50000 replications) exhibit the same modes, quantiles and distributional forms, as was hypothesized in the construction of the test statistic.

Using these results, we proposed the following critical values for the rejection of the null hypothesis H_0 (see Table 6.2): Here, $\hat{p}^*_{11}$ corresponds to the proportion of entries of Table 6.1 of

Table 6.2: Proposed critical values for the count test statistic.

j	4	5	6	7	8+
$\hat{p}^*_{11}$	0.05	0.05	0.05	0.05	0.05

the type (1,1). These values will be utilized in the simulation-based performance comparison of type I and II errors, which is presented in the following sections.

An important question that needs to be addressed, in addition with the finding presented in this section has to do with the distribution of the entries of Table 6.1 under the alternative hypothesis H_1. This is a very important aspect of the test, since the goal is to provide a statistical test that achieves the lowest possible type I and type II errors. Choosing a critical value that minimizes the type I of a test error could severely affect its type II error. For this reason, a threshold that achieves a good balance between the two errors is desired.

6.4.3 A Non-Parametric Significance Test Based on the Geometry of the Wavelet Coefficient Sequences.

In the previous section, a significance test based on the normality assumption of wavelet coefficients was proposed. As argued, this assumption could be too strong in some cases, and could lead to wrong statistical conclusions. In this section a non-parametric test that exploits

the geometry of the wavelet coefficient sequences is introduced, aiming to enhance robustness when departures from normality are present in the wavelet decomposition.

Consider now two sequences $\{X(n),\ n \in \mathbb{N}\}$ and $\{Y(n),\ n \in \mathbb{N}\}$ that are zero mean, second order stationary. Assume that the corresponding wavelet coefficients resulting from the orthogonal DWT of each sequence, at each level are uncorrelated. This implies that, if we look at the sample covariance matrix of the pairs:

$$\left\{ \begin{bmatrix} \tilde{d}_{j,0}^{(X)} \\ \tilde{d}_{j,0}^{(Y)} \end{bmatrix}, ..., \begin{bmatrix} \tilde{d}_{j,2^j-1}^{(X)} \\ \tilde{d}_{j,2^j-1}^{(Y)} \end{bmatrix} \right\},\ j = 0, ..., J-1$$

that is defined as:

$$\widehat{\mathbf{\Sigma}}_j = 2^{-j} \sum_{k=0}^{2^j-1} \mathbf{d}_{X,Y}^{(j,k)} \mathbf{d}_{X,Y}^{(j,k)^T}, \tag{6.25}$$

where, $\mathbf{d}_{X,Y}^{(j,k)} = \begin{bmatrix} \tilde{d}_{j,k}^{(X)} \\ \tilde{d}_{j,k}^{(Y)} \end{bmatrix} \in \mathbb{R}^2\ ,\ k = 0, ..., 2^j - 1$, and $\tilde{d}_{j,k} = d_{j,k}/||\mathbf{d}_X^{(j)}||_2$, from the definition of level-wise correlations $\widehat{\mathbf{\Sigma}}_j$ satisfies:

$$\widehat{\mathbf{\Sigma}}_j = 2^{-j} \begin{bmatrix} \sum_{k=0}^{2^j-1} \tilde{d}_{j,k}^{(X)2} & \sum_{k=0}^{2^j-1} \tilde{d}_{j,k}^{(X)} \tilde{d}_{j,k}^{(Y)} \\ \sum_{k=0}^{2^j-1} \tilde{d}_{j,k}^{(X)} \tilde{d}_{j,k}^{(Y)} & \sum_{k=0}^{2^j-1} \tilde{d}_{j,k}^{(Y)2} \end{bmatrix} = \begin{bmatrix} 1 & \hat{\rho}_{XY}^{(j)} \\ \hat{\rho}_{XY}^{(j)} & 1 \end{bmatrix}.$$

This implies that the eigenvalues of $\widehat{\mathbf{\Sigma}}_j$ are given by:

$$\lambda_1^{(j)} = 1 + |\hat{\rho}_{XY}^{(j)}|, \text{ and } \lambda_2^{(j)} = 1 - |\hat{\rho}_{XY}^{(j)}|, \text{ for } j = 0, ..., J-1.$$

Thus, the corresponding condition number is given by:

$$\kappa(\widehat{\mathbf{\Sigma}}_j) = \frac{\lambda_1^{(j)}}{\lambda_2^{(j)}} = \frac{1 + |\hat{\rho}_{XY}^{(j)}|}{1 - |\hat{\rho}_{XY}^{(j)}|} \geq 1.$$

This implies that $\kappa(\widehat{\mathbf{\Sigma}}_j)|H_0 = 1$ and $\kappa(\widehat{\mathbf{\Sigma}}_j)|H_1 = \infty$. Here, similarly as in the previous section, H_0 corresponds to the null hypothesis of no correlation between the wavelet coefficient sequences.

As can be expected, in practical applications even though the signals may not be correlated, due to randomness and numerical effects the computed level-wise correlation can be different than zero, causing a departure of the condition number $\kappa(\widehat{\mathbf{\Sigma}}_j)$ from 1. For this reason, it is necessary to define a proper threshold that, assuming H_0 true, could help us making a decision about whether or not to reject/fail to reject the null hypothesis.

Elicitation of a Decision Threshold for Testing H_0 against H_1

Consider the following plot of $\kappa(\widehat{\mathbf{\Sigma}}_j)$ as a function of $\hat{\rho}_{XY}^{(j)}$:

As it can be observed, the condition number remains relatively stable for $0 < |\hat{\rho}_{XY}^{(j)}| < 0.5$, which implies robustness for numerical or random effects that may cause artificial inflation of the condition number. This can be directly linked to a good performance in terms of the type I error of the test, aspect that is investigated in the next section.

On the other hand, when $|\hat{\rho}_{XY}^{(j)}| > 0.5$ the condition number is very sensible to slight variations of the sample correlation magnitude, which suggests that this test statistic could exhibit good performance in terms of the type II error.

Along the same line of the above argument, under the assumption that H_0 holds, and based on the fact that $0 \leq |\hat{\rho}_{XY}^{(j)}| \leq 1$, it is possible to assume that the distribution of the level-wise correlations $|\hat{\rho}_{XY}^{(j)}|$, for $|\hat{\rho}_{XY}^{(j)}| > \rho^*$ can be bounded from above by another distribution with parameters that can be adjusted to capture the randomness in the samples that cause artificial inflation of the condition number, while achieving a polynomial rate of decay of the tails. In other words, if we approximate the distribution of $|\hat{\rho}_{XY}^{(j)}|$ by another density that behaves

roughly in the same way up to a certain value ρ^*, but has heavier tails (for $|\hat{\rho}_{XY}^{(j)}| > \rho^*$), and we conduct the hypothesis test based on this approximate density, as long as the quantile that corresponds to the significance level α used for the test is greater than equal than ρ^*, the probability of type I error obtained via this approach will be an upper bound of the true distribution's probability of type I error. This argument can be summarized in the following Theorem:

Theorem 6.4.1. *Suppose that under H_0, $|\hat{\rho}_{XY}^{(j)}| \sim \mathcal{F}_0$. Let $\mathcal{G}$ be another density function class such that:*

(i) $supp(\mathcal{F}_0) \subseteq supp(\mathcal{G})$.

(ii) For $\rho > \rho^*$, $f_0(\rho) \leq g(\rho)$, $\forall\, g \in \mathcal{G}$.

Then, if for an arbitrary $0 < \alpha < 1$, there exists $\rho_{\mathcal{G},\alpha} \geq \rho^*$, such that $\mathbb{P}_{\mathcal{G}}(T > \rho_{\mathcal{G},\alpha}) \leq \alpha$, it follows:

$$\mathbb{P}_{\mathcal{F}_0}(T > \rho_{\mathcal{G},\alpha}) \leq \mathbb{P}_{\mathcal{G}}(T > \rho_{\mathcal{G},\alpha}) \leq \alpha.$$

This result implies that defining a significance level using the density function $\mathcal{G}$ is equivalent to define an upper bound for the significance level corresponding to the density function $\mathcal{F}_0$. Moreover, assuming that the density functions $\mathcal{F}_0$ and $\mathcal{G}$ are continuous, then by the monotonicity of $\mathbb{P}_{\mathcal{F}_0}(T > t)$, it follows that $\rho_{\mathcal{G},\alpha} \geq \rho_{\mathcal{F}_0,\alpha}$, meaning that the obtained critical value using the surrogate distribution is also an upper bound of the critical value corresponding to the true distribution.

Corollary 6.4.1. *Suppose conditions and results of Theorem 6.4.1 hold. Assume there exists a transformation $h : supp(\mathcal{G}) \to \mathcal{S}$, that is continuous, strictly increasing and invertible. Then, for an arbitrary $t_{\mathcal{H},\alpha} \in \mathcal{S}$ such that:*

$$\mathbb{P}_{\mathcal{H}}(U > t_{\mathcal{H},\alpha}) \leq \alpha,$$

where $U \sim \mathcal{H}$, and $\mathcal{H}$ corresponds to the probability distribution generated by the transformation $h(\mathcal{G})$, it follows:

$$\mathbb{P}_{\mathcal{F}_0}(T > \rho_{\mathcal{G},\alpha}) \leq \mathbb{P}_{\mathcal{G}}(T > \rho_{\mathcal{G},\alpha}) = \mathbb{P}_{\mathcal{H}}(h(T) > t_{\mathcal{H},\alpha}) \leq \alpha,$$

where $h^{-1}(t_{\mathcal{H},\alpha}) = \rho_{\mathcal{G},\alpha}$.

This result implies that it is possible to define a critical value for the transformed variable $h(T)$ and that would be equivalent to the definition of a critical value under the original distribution, guaranteeing a performance as good as the one measured by using the transformation.

Putting these last two results in the context of the condition number definition, it follows that $\kappa(\widehat{\mathbf{\Sigma}}_j)$ results from a continuous, strictly increasing transformation of $|\hat{\rho}_{XY}^{(j)}|$. Thus, the proposed procedure of defining critical values based on a surrogate distribution would lead, in theory, to valid statistical conclusions without making any specific assumptions about the distributional form of $|\hat{\rho}_{XY}^{(j)}|$ under H_0.

For example, setting as surrogate of $|\hat{\rho}_{XY}^{(j)}| \, |H_0$ the $\mathcal{B}eta(1,1) = \mathcal{U}(0,1)$ distribution would be extremely conservative. In fact, under this setting, the distribution of the condition number takes the form:

$$\kappa(\widehat{\mathbf{\Sigma}}_j)|H_0, \mathcal{D}_\rho \sim f(\kappa|H_0, \mathcal{D}_\rho) = \frac{2}{(\kappa+1)^2}\mathbf{1}_{\{\kappa \geq 1\}}.$$

Here, $\mathcal{D}_\rho$ denotes the assumed surrogate distribution for $|\hat{\rho}_{XY}^{(j)}| \, |H_0$. Using this distribution, it is possible to show that for $\kappa^* > 39$, $\mathbb{P}\left(\kappa(\widehat{\mathbf{\Sigma}}_j) > \kappa^* | H_0, \mathcal{D}_\rho\right) \leq 0.05$. This threshold value, even though it would guarantee a very small type I error of the test, it can severely impact its type II error.

Along this line of reasoning, suppose now that we use as surrogate for $|\hat{\rho}_{XY}^{(j)}| \, |H_0$ a $\mathcal{B}eta(1, M)$,

$M \in \mathbb{N}$ distribution. Then, it follows:

$$\kappa(\widehat{\mathbf{\Sigma}}_j)|H_0, \mathcal{D}_\rho \sim f(\kappa|H_0, \mathcal{D}_\rho) = \frac{M \cdot 2^M}{(\kappa+1)^{M+1}} \mathbf{1}_{\{\kappa \geq 1\}},$$

and for $\kappa^* \geq 1$:

$$\mathbb{P}\left(\kappa(\widehat{\mathbf{\Sigma}}_j) \leq \kappa^* | H_0, \mathcal{D}_\rho\right) = 1 - \frac{2^M}{(\kappa^*+1)^M}.$$

These results follow from the application of a transformation using the definition of $\kappa(\widehat{\mathbf{\Sigma}}_j)$, and then computing the integral corresponding to the cumulative density function.

This approximation of the probability density of $|\hat{\rho}_{XY}^{(j)}|$ via $\mathcal{B}eta(1, M)$, in addition to produce polynomial rate of decay of the tails, allows for closed form expressions for both the probability density of the condition number $\kappa(\widehat{\mathbf{\Sigma}}_j)$, and its cumulative density function. This facilitates the analysis and empirical definition of critical values for the statistical test.

For example, let the approximating density of $|\hat{\rho}_{XY}^{(j)}|\ |H_0$ be a $\mathcal{B}eta(1, 5)$ distribution. This would imply that the probability of the absolute level-wise correlation exceeding 0.5 would be less than 0.05, which is reasonable assuming H_0 holds. Under this setting, the distribution of the condition number takes the form:

$$\kappa(\widehat{\mathbf{\Sigma}}_j)|H_0, \mathcal{D}_\rho \sim f(\kappa|H_0, \mathcal{D}_\rho) = \frac{128}{(\kappa+1)^6} \mathbf{1}_{\{\kappa \geq 1\}}.$$

Using this distribution, it is possible to show that for $\kappa^* > 2.6$, $\mathbb{P}\left(\kappa(\widehat{\mathbf{\Sigma}}_j) > \kappa^* | H_0, \mathcal{D}_\rho\right) \leq 0.05$.

In order to achieve a good balance between type I and II errors, it may be reasonable to assume parameters for the approximating distribution of $|\hat{\rho}_{XY}^{(j)}|\ |H_0$ that generate a rate of decay that resembles the empirical evidence for certain type of processes. In particular, as can be observed in Figs. 6.6 and 6.13 , for signals that correspond to either WN, AR(1) or MA(1) processes, it is possible to assume that rate of decay of $|\hat{\rho}_{XY}^{(j)}|\ |H_0$ as it approaches

1 can be modeled as polynomial. The use of a polynomial rate instead of an exponential would be beneficial from a robustness viewpoint, since it will lead to larger critical values, thus accounting for more variability in the data sources. Of course, this improvement in robustness could be at the expense of an undesired increase in the type II error of the test. For this reason, it is possible to follow the same methodology previously proposed based on the results of Theorem 6.4.1 and Corollary 6.4.1 to majorize the left tail of the distribution of the $|\hat{\rho}_{XY}^{(j)}|$ under the alternative hypothesis H_1.

Suppose that under H_1, $|\hat{\rho}_{XY}^{(j)}| \sim \mathcal{F}_1$. Let $\mathcal{D}$ be another density function class such that:

(i) $supp(\mathcal{F}_1) \subseteq supp(\mathcal{D})$.

(ii) For $\rho \leq \rho^{**}$, $f_1(\rho) \leq g_{\mathcal{D}}(\rho)$, $\forall\, g_{\mathcal{D}} \in \mathcal{D}$.

Then, if for an arbitrary $0 < \beta < 1$, there exists $\rho_{\mathcal{D},\beta} \leq \rho^{**}$ such that $\mathbb{P}_{\mathcal{D}}(T \leq \rho_{\mathcal{D},\beta}) \leq 1 - \beta$, it implies:

$$\mathbb{P}_{\mathcal{F}_1}(T \leq \rho_{\mathcal{D},\beta}) \leq \mathbb{P}_{\mathcal{D}}(T \leq \rho_{\mathcal{D},\beta}) \leq 1 - \beta.$$

From Corollary 6.4.1, for a continuous, strictly increasing and invertible transformation $h : supp(\mathcal{D}) \to \mathcal{S}$, and an arbitrary $t_{\mathcal{H},\beta}$ such that:

$$\mathbb{P}_{\mathcal{H}}(U \leq t_{\mathcal{H},\beta}) \leq 1 - \beta,$$

it follows:

$$\mathbb{P}_{\mathcal{F}_1}(T \leq \rho_{\mathcal{D},\beta}) \leq \mathbb{P}_{\mathcal{D}}(T \leq \rho_{\mathcal{D},\beta}) = \mathbb{P}_{\mathcal{H}}(h(T) \leq t_{\mathcal{H},\beta}) \leq 1 - \beta,$$

where, $h^{-1}(t_{\mathcal{H},\beta}) = \rho_{\mathcal{D},\beta}$.

Now, suppose that we want to determine a critical value $\tilde{\rho} \in [0,1]$ that for a predefined $0 < \lambda < 1$, solves:

$$\min_{0 \leq t \leq 1} \left(\lambda \cdot \mathbb{P}_{\mathcal{F}_0}(T > t) + (1-\lambda) \cdot \mathbb{P}_{\mathcal{F}_1}(T \leq t) \right). \tag{6.26}$$

As can be observed, the solution to 6.26 corresponds to a decision threshold $\tilde{\rho}$ that minimizes the weighted sum of type I and II errors. In the objective, the parameter λ corresponds to the imputed relative average cost of each type of error, and can be chosen in accordance to the problem nature. Similarly, another possible interpretation for this parameter could be the prior probability that H_0 is true.

Using the last set of results related to Theorem 6.4.1 and Corollary 6.4.1, it is possible to solve:

$$\min_{h(\rho^*) \leq z \leq h(\rho^{**})} \left(\lambda \cdot \mathbb{P}_{\mathcal{H}(\mathcal{G})}(Z > z) + (1-\lambda) \cdot \mathbb{P}_{\mathcal{H}(\mathcal{D})}(Z \leq z) \right). \tag{6.27}$$

Here, it is assumed that $0 < \rho^* < \tilde{\rho} < \rho^{**} < 1$. Setting $\mathcal{G} \sim \mathcal{B}eta(1, M)$ and $\mathcal{D} \sim \mathcal{B}eta(M, 1)$ the optimization problem 6.27 becomes:

$$\min_{z \geq 1} \left(\lambda \cdot \frac{2^M}{(z+1)^M} + (1-\lambda) \cdot e^{M \log\left(\frac{z-1}{z+1}\right)} \right), \tag{6.28}$$

with solution given by:

$$z^* = 1 + \left(\frac{\lambda \cdot 2^{M-1}}{1-\lambda} \right)^{\frac{1}{M-1}}. \tag{6.29}$$

In (6.28), the constraint on $h(\rho^*) \leq z \leq h(\rho^{**})$ is relaxed since it is possible to assume that $|\rho^* - \rho^{**}| > \epsilon$, for $\epsilon = \epsilon(M)$, and under the transformation h, that absolute distance is significantly increased.

Note that in (6.29), if $\lambda = 0.5$, it follows that $z^* = 3$ which is independent of the parameter M. This implies that if equal weights are allocated to each type of error, then by utilizing surrogate distributions $\mathcal{B}eta(1, M)$ and $\mathcal{B}eta(M, 1)$ the optimal critical value for the condition

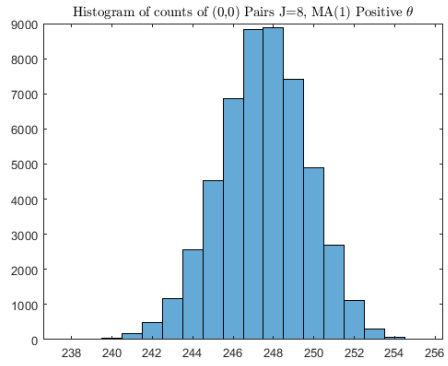
number is independent of the parameter M that controls how fast the tails of the surrogate distributions decay.

Based on the aforementioned discussion, and the observed statistical behavior of $|\hat{\rho}_{XY}^{(j)}| \; |H_0$ for processes of the form AR(1), MA(1) and WN (see Fig. 6.13), we propose the following critical values according to the corresponding scale level (see Table 6.3):

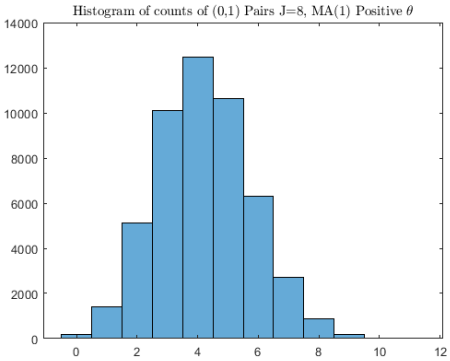
Table 6.3: Proposed critical values for the condition number test statistic.

j	4	5	6	7	8+
κ^*	3.3	2.6	2.3	2.0	2.0

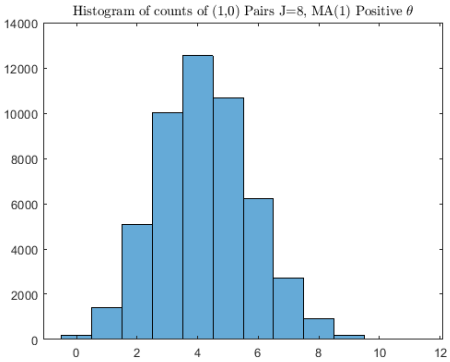
In the next section, a simulation-based comparative study between the introduced test statistics and other popular statistical tests is introduced in order to analyze their performance in terms of the type I and II errors, and validate the proposed critical values.



(a) Histogram of $\sum_{k=0}^{2^j-1} \mathbf{1}_{(\mathbf{1}_{1,j,k}=0,\mathbf{1}_{2,j,k}=0)}$ for $J = 8$

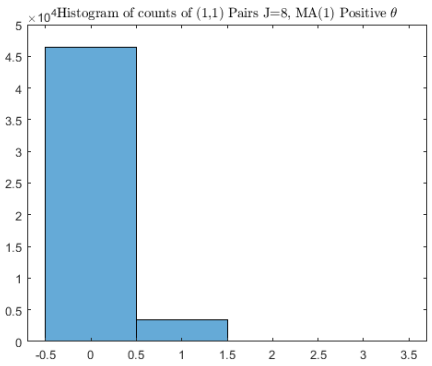


(b) Histogram of $\sum_{k=0}^{2^j-1} \mathbf{1}_{(\mathbf{1}_{1,j,k}=0,\mathbf{1}_{2,j,k}=1)}$ for $J = 8$



(c) Histogram of $\sum_{k=0}^{2^j-1} \mathbf{1}_{(\mathbf{1}_{1,j,k}=1,\mathbf{1}_{2,j,k}=0)}$ for $J = 8$

Figure 6.10: Typical Histograms of the entries of Table 6.3 for an MA(1) process with parameter $\theta = 0.9$. The experiments were replicated 50000 times. Similar behavior were observed for for the rest of detail levels fir AR(1), WN and ARMA(1,1) model, with no significant differences for the empirical quantiles.



(a) Histogram of $\sum_{k=0}^{2^j-1} \mathbf{1}_{(\mathbf{1}_{1,j,k}=1,\mathbf{1}_{2,j,k}=1)}$ for $J = 8$

Figure 6.11: Typical Histograms of the entries of Table 6.3 for an MA(1) process with parameter $\theta = 0.9$. The experiments were replicated 50000 times. Similar behavior were observed for for the rest of detail levels fir AR(1), WN and ARMA(1,1) model, with no significant differences for the empirical quantiles.

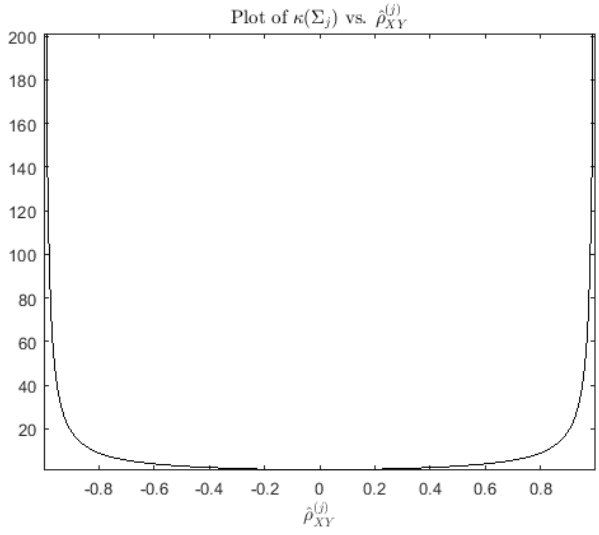
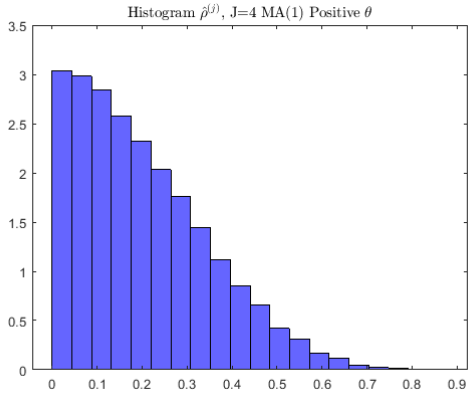
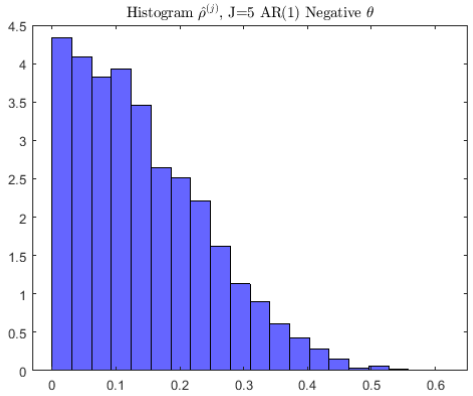


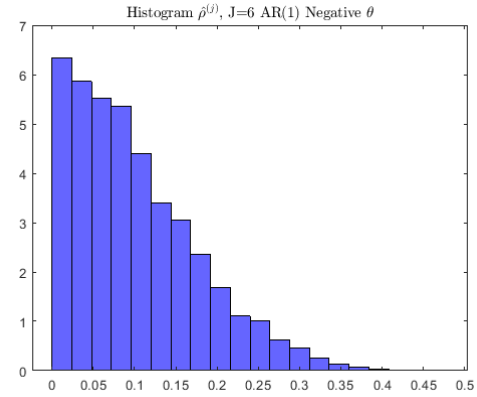
Figure 6.12: Plot of the condition number $\kappa(\widehat{\mathbf{\Sigma}}_j)$ as a function of $\hat{\rho}_{XY}^{(j)}$.



(a) Histogram of $|\hat{\rho}_{XY}^{(j)}|$ for $J = 4$

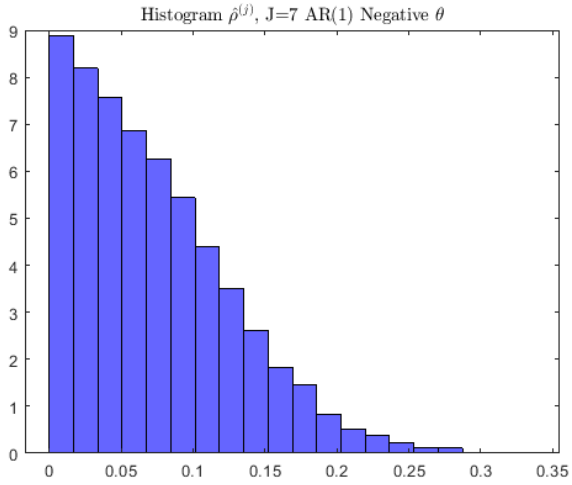


(b) Histogram of $|\hat{\rho}_{XY}^{(j)}|$ for $J = 5$

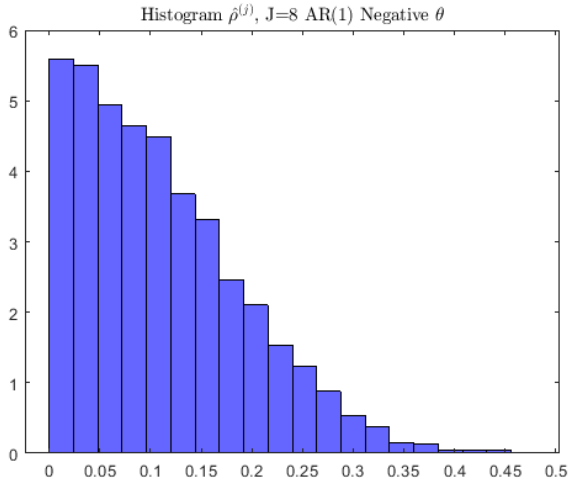


(c) Histogram of $|\hat{\rho}_{XY}^{(j)}|$ for $J = 6$

Figure 6.13: Typical Histograms of the level-wise correlations for uncorrelated sequences of an AR(1) process with $\theta < 0$. The experiments were replicated 50000 times. Similar behavior were observed for MA(1) and WN processes, with no significant differences for the type of decay and empirical quantiles.



(a) Histogram of $|\hat{\rho}_{XY}^{(j)}|$ for $J = 7$



(b) Histogram of $|\hat{\rho}_{XY}^{(j)}|$ for $J = 8$

Figure 6.14: Typical Histograms of the level-wise correlations for uncorrelated sequences of an AR(1) process with $\theta < 0$. The experiments were replicated 50000 times. Similar behavior were observed for MA(1) and WN processes, with no significant differences for the type of decay and empirical quantiles.

6.4.4 Simulation Study of the Probability of Type I Error for Uncorrelated Stationary Sequences.

In this section we investigate the performance of the proposed statistical tests via simulation. All the estimators are implemented using MATLAB®, and estimation results are compared with previously published statistical methodologies: Pearson's $t-$test [79], Spearman's rank correlation [81] and Kendall's [85].

For the simulation, we consider the following models:

$$X_t = \theta \cdot \epsilon_{t-1} + \epsilon_t, \text{ (MA(1) model)}, \tag{6.30}$$

$$X_t = \phi \cdot X_{t-1} + \epsilon_t, \text{ (AR(1) model)}, \tag{6.31}$$

$$X_t = \phi \cdot X_{t-1} + \theta \cdot \epsilon_{t-1} + \epsilon_t \text{ (ARMA(1,1) model)}, \tag{6.32}$$

where $t \in \mathbb{N}$ and $\epsilon_t \overset{iid}{\sim} \mathcal{N}(0,1)$. Here, for stability conditions (stationary models) it is assumed that $|\phi| < 1$, $|\theta| < 1$. For the simulation, the following parameters were utilized:

(a) MA(1) model: $\theta = \{-0.9,\ -0.7,\ -0.5,\ 0.5,\ 0.7,\ 0.9\}$.

(b) AR(1) model: $\phi = \{-0.9,\ -0.7,\ -0.5,\ 0.5,\ 0.7,\ 0.9\}$.

(c) ARMA(1,1) model: $(\phi, \theta) = \{(-0.8, 0.1),\ (-0.5, 0.1),\ (-0.2, 0.1),\ (0.1, -0.8), ...\}$ $\{..., \ (-0.9, 0.9),\ (0.1, -0.2)\}$.

(d) $C^* = 0.05$, $\kappa^* = 2.6$ for all scale levels $j = 1, ..., J-1$.

(e) Significance level for Pearson's $t-$test [79], Spearman's rank correlation [81] and Kendall's [85] was set to $\alpha = 0.05$.

(f) Sequence length $N = 512$, which allowed a wavelet decomposition up to level $J - 1 = 8$.

(g) Wavelet filter was set to Symmlet with 10 vanishing moments. This choice is motivated by the fact that since no wavelet system (except Haar) can be compactly supported and

symmetric at the same time [5], Symmlets are considered "close to symmetric". In addition, the corresponding filter can be set such that the wavelet function $\psi(t)$ satisfies the vanishing moments condition, which is desirable to enhance the whitening properties of the DWT for stationary sequences.

The simulations implementation consisted of 100 batches of $B = 1000$ replications for each one of the models given in Eqs. (6.30), (6.31) and (6.32). At each replication, two independent samples of $N = 512$ were generated, computing the corresponding DWT, and test statistics at each scale level $j = 1, ..., 8$. Test decisions were made based upon the parameters previously defined, and the following results were collected:

(i) Box plots of the empirical probability of type I error, computed as:

$$\hat{p}_{Test_m} = \frac{1}{B} \sum_{b=1}^{B} \mathbf{1}_{\{T_{m,b} > T^*\}}.$$

Here, $Test_m$ refers to the utilized statistical test, $Test_{m,b}$ corresponds to the test statistic m resulting from the $b-$th sample, and T^* corresponds to the critical value. As was previously mentioned, $B = 1000$ and the experiments were repeated 100 times.

(ii) Summary tables with the averages and standard deviations of $\hat{p}_{Test_m}$ obtained from all replications of the experiment.

Since for the simulated processes the obtained results for certain parameter setting exhibit relatively similar behavior, in this section we include only the most representative cases. For those models and parameter settings that were omitted, the observed results are almost identical quantitatively and qualitatively as the ones that are presented.

The reported Tables contain only the sample averages, since in most cases, all tests exhibit

similar standard deviations that can be observed in the included Box plots in Appendix D.[1]

Table 6.4: Obtained results for average $\hat{p}_{Test_m}$ for AR(1) with parameter $\phi = -0.9$.

j	4	5	6	7	8
Condition number	0.07576	0.01016	0.0003	0	0.00054
Counts	0.00004	0	0	0	0
T-test	0.05952	0.05482	0.05374	0.05538	0.38382
Kendall	0.04164	0.05028	0.04962	0.05212	0.3677
Spearman	0.0502	0.05084	0.05084	0.05284	0.36946

Table 6.5: Obtained results for average $\hat{p}_{Test_m}$ for AR(1) with parameter $\phi = -0.7$.

j	4	5	6	7	8
Condition number	0.07416	0.0096	0.0004	0	0
Counts	0	0	0	0	0
T-test	0.05904	0.0552	0.05082	0.05522	0.14582
Kendall	0.04286	0.0487	0.04742	0.05338	0.13912
Spearman	0.05066	0.0494	0.04826	0.05462	0.13948

Table 6.6: Obtained results for average $\hat{p}_{Test_m}$ for AR(1) with parameter $\phi = 0.9$.

j	4	5	6	7	8
Condition number	0.08338	0.01436	0.00072	2.00E-05	0
Counts	2.00E-05	0	0	0	0
T-test	0.0663	0.07166	0.0738	0.07486	0.05924
Kendall	0.04774	0.06352	0.06716	0.06942	0.05594
Spearman	0.05624	0.06402	0.0672	0.06948	0.05626

Table 6.7: Obtained results for average $\hat{p}_{Test_m}$ for MA(1) with parameter $\theta = 0.9$.

j	4	5	6	7	8
Condition number	0.07288	0.00916	0.00032	0	0
Counts	0	0	0	0	0
T-test	0.05748	0.0532	0.05114	0.0549	0.1369
Kendall	0.04128	0.04864	0.04936	0.05394	0.13084
Spearman	0.04962	0.0494	0.04978	0.05426	0.13152

Remarks and Comments

(i) As was mentioned, from the simulations that were implemented according to the proposed methodology, in most cases (i.e. when the time series did not exhibit short-time

[1] In each of the included Boxplots there is an entry with a $C2$ denomination, that corresponds to a test statistic based on the weighted differences of two Chi-square distributions, that was included in the simulation but for the purpose of this Chapter objectives has no influence due to its observed performance.

Table 6.8: Obtained results for average $\hat{p}_{Test_m}$ for MA(1) with parameter $\theta = 0.7$.

j	4	5	6	7	8
Condition number	0.07418	0.00998	0.0002	0	0
Counts	0	0	0	0	0
T-test	0.05882	0.05434	0.049	0.05454	0.11436
Kendall	0.04148	0.04878	0.04838	0.05162	0.11016
Spearman	0.04978	0.04942	0.04918	0.05224	0.11008

Table 6.9: Obtained results for average $\hat{p}_{Test_m}$ for MA(1) with parameter $\theta = -0.9$.

j	4	5	6	7	8
Condition number	0.09086	0.01664	0.00056	0	0
Counts	0	0	0	0	0
T-test	0.0732	0.07558	0.0699	0.06668	0.05642
Kendall	0.05166	0.06598	0.06596	0.06352	0.0552
Spearman	0.06188	0.06712	0.06752	0.06394	0.05514

Table 6.10: Obtained results for average $\hat{p}_{Test_m}$ for ARMA(1,1) with parameters $\phi = -0.8$, $\theta = 0.1$.

j	4	5	6	7	8
Condition number	0.07634	0.00936	0.00038	0	0
Counts	0	0	0	0	0
T-test	0.0599	0.0535	0.05168	0.05478	0.21726
Kendall	0.04406	0.0491	0.04924	0.05228	0.20614
Spearman	0.05256	0.04992	0.04968	0.05264	0.20726

Table 6.11: Obtained results for average $\hat{p}_{Test_m}$ for ARMA(1,1) with parameters $\phi = -0.9$, $\theta = 0.9$.

j	4	5	6	7	8
Condition number	0.07664	0.00964	0.0002	0	0
Counts	6.00E-05	0	0	0	0
T-test	0.06032	0.05502	0.0525	0.05204	0.05026
Kendall	0.04272	0.04998	0.05004	0.05022	0.0499
Spearman	0.04986	0.04936	0.0513	0.05022	0.04958

high oscillations) the T-test, Spearman's and Kendall's tests performed as expected. The observed average empirical type I error was close to the established significance level of $\alpha = 0.05$, with a standard deviation that was within 1% range on average.

(ii) In the cases of highly oscillatory signals (i.e. when $-0.9 < \phi < -0.5$, $0.5 < \theta < 0.9$ for AR(1) and MA(1), respectively), these tests showed a significant increment in the average type I error, with average values that were on the range between 14% to 36%. This behavior could be explained by the fact that these kinds of processes tend to produce large wavelet coefficients that depart normality at high scales. For a proper illus-

tration of this statement, consider Fig. 6.16 where a typical time series for AR(1) and MA(1) with high oscillations are depicted. As can be observed, in panels 6.16b and 6.17b, at scale levels $j = 7$ and $j = 8$, the wavelet coefficients exhibit an irregular behavior with high variability, which is linked to the short time high oscillations that can be observed in the corresponding time series. Coupling this with results of Tables 6.4 and 6.7, it can be inferred that the inflated probabilities of type I error can be associated with the observed behavior of the wavelet coefficients. In particular, the departure from normality in terms of heavier tails. In Fig. 6.18a, a typical empirical distribution for wavelet coefficients resulting from an AR(1) process with parameter $\phi = -0.9$ is shown. In it, the departure from normality is evident.

(iii) In such situations, the proposed tests (Local and condition number) remained very stable, exhibiting average type I errors less than the pre-defined significance level $\alpha = 0.05$.

(iv) As can be observed from Tables 6.4 to 6.11, in most scenarios the proposed tests (6.4.2 and 6.4.3) outperform the other statistical procedures used as benchmark, leading to an average type I error probability that is significantly smaller than the other tests. This enhances their reliability in terms of the false rejection rates that can be expected if utilized in practice. Nonetheless, this feature needs to be combined with a good performance in terms of the type II error, aspect that is investigated in the next section. A summary plot that illustrates these results is shown in Fig. 6.15.

(v) In particular, it is interesting to note that in most cases, the count-based tests yields a zero probability of type I error. This could suggest that the utilized decision threshold could be too large, and could potentially cause poor performance in terms of the type II error. This aspect is investigated in the next section.

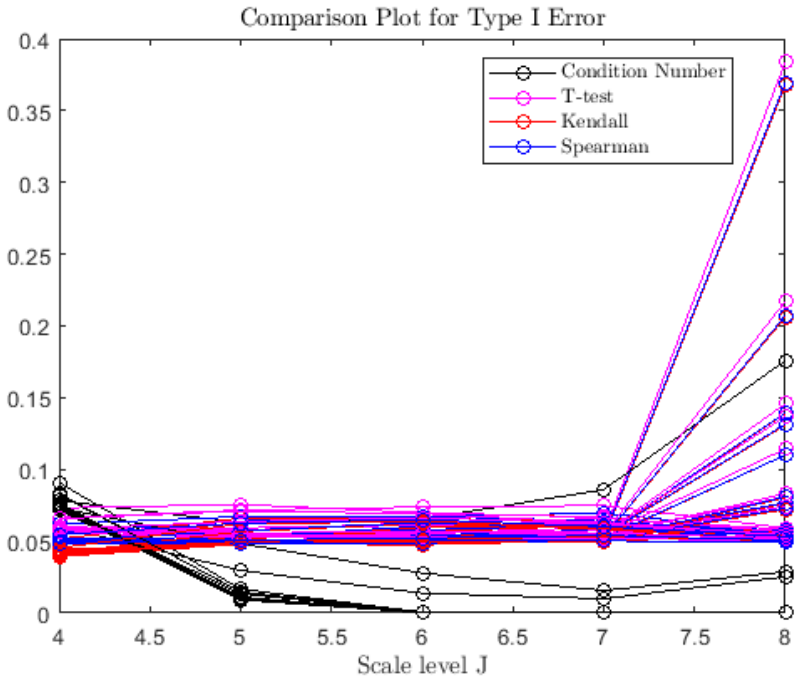
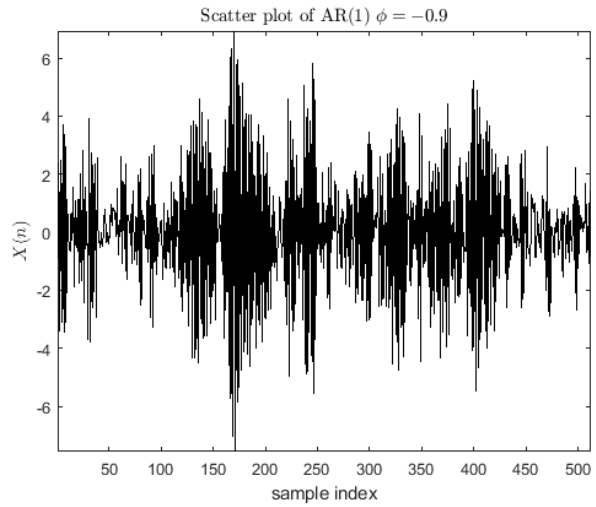
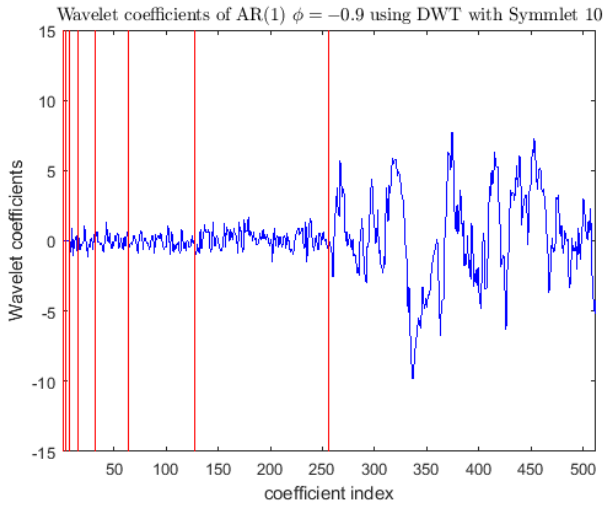


Figure 6.15: Summary plot of average type I error probability for each test statistic

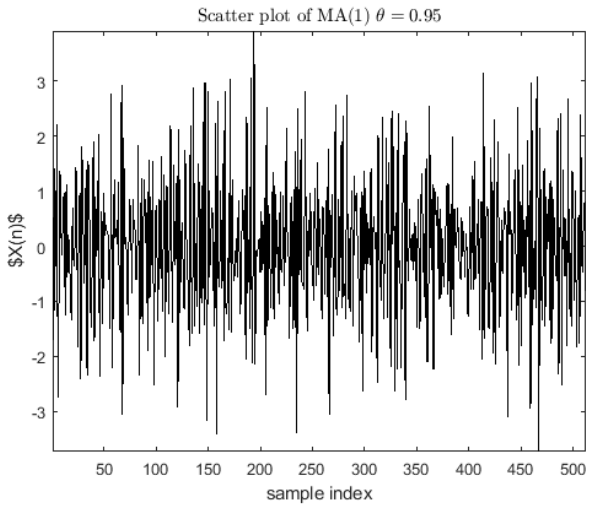


(a) Scatter plot of AR(1) process with $\phi = -0.9$

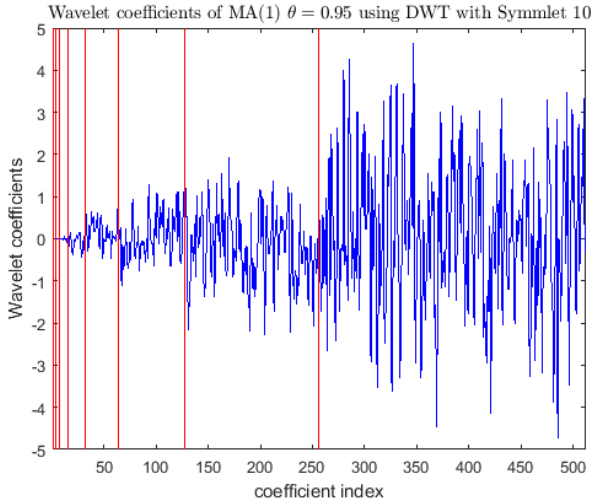


(b) Plot of wavelet coefficients generated using DWT with Symmlet 10 of AR(1) shown in panel 6.16a.

Figure 6.16: Scatter plots of typical AR(1) with high oscillatory behavior and their respective wavelet coefficients generated from orthogonal DWT using Symmlet 10. The red lines indicate the separation between consecutive scale levels, arranged in an increasing order.

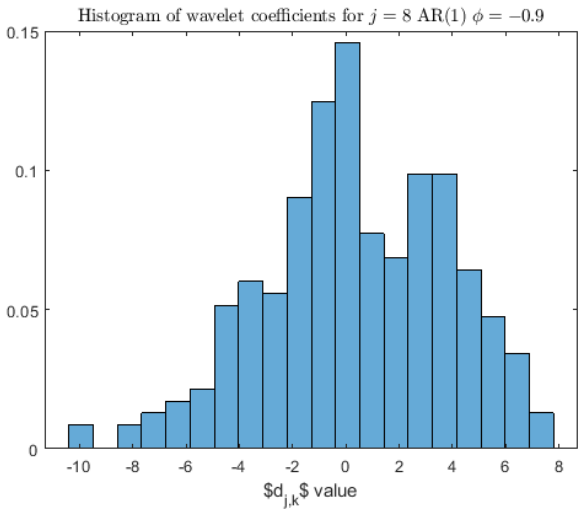


(a) Scatter plot of MA(1) process with $\theta = 0.9$



(b) Plot of wavelet coefficients generated using DWT with Symmlet 10 of MA(1) shown in panel 6.17a.

Figure 6.17: Scatter plots of typical MA(1) with high oscillatory behavior and their respective wavelet coefficients generated from orthogonal DWT using Symmlet 10. The red lines indicate the separation between consecutive scale levels, arranged in an increasing order.



(a) Typical histogram of wavelet coefficients for level $j = 8$ of an AR(1) process with $\phi = -0.9$, obtained by DWT using Symmlet 10. As seen from the Figure, the coefficients exhibit heavier tails than the Gaussian distribution, as well as a certain degree of asymmetry with respect the origin.

Figure 6.18

6.4.5 Simulation Study of the Probability of Type II Error for Correlated Stationary Sequences.

In this section we study the performance of the proposed tests (6.4.2 and 6.4.3) in terms of the type II error, comparing them with previously published methodologies in a similar way as in the previous section. This analysis is necessary to complement the assessment of the expected performance in terms of type I error of the proposed estimators that was introduced in the last section. Ideally, an optimal statistical test should have both type I and type II errors as small as possible, having a high probability of detecting a meaningful difference assuming the relationship between the signals did exist. In practice this is extremely difficult due to the adversarial nature between the two errors.

In this study, the statistical models and settings are the same as the ones utilized in Section 6.4.4. However, a slightly different methodology was utilized in order to produce sequences that were correlated in the wavelet domain:

(i) For each model and parameter setting, a sequence of $N = 512$ samples was generated, and the wavelet decomposition was obtained via DWT using a Symmlet 10 filter.

(ii) At each scale level $j = 1, ..., 8$ the following model was used:

$$d_{j,k}^{(Y)} = \beta_j \cdot d_{j,k}^{(X)} + \epsilon_k,\ k = 0, ..., 2^j - 1.$$

Here, ϵ_k are iid $\mathcal{N}(0, \sigma^2)$ random variables. The values of β_j were chosen within the range $-0.25 < \beta_j < -0.1$ and $0.1 < \beta_j < 0.25$; similarly, the noise variance σ^2 was set to 30% of the variability of the coefficients $\mathbf{d}_X^{(j)}$. These values allow the generation of conditions that impose an adequate degree of complexity for the detection of correlation. High values of β_j and/or small values of σ^2 will facilitate detection, as was illustrated in Section 6.3.2.

(iii) The simulations implementation consisted of 100 batches of $B = 1000$ replications for

each one of the models given in Eqs. (6.30), (6.31) and (6.32). At each replication, two independent samples of $N = 512$ were generated, computing the corresponding DWT, and test statistics at each scale level $j = 1, ..., 8$. Test decisions were made based upon the parameters previously defined, and the following results were collected:

i. Box plots of the empirical probability of type II error, computed as:

$$\hat{p}_{Test_m} = \frac{1}{B}\sum_{b=1}^{B}(1 - \mathbf{1}_{\{T_{m,b} > T^*\}}).$$

Here, $Test_m$ refers to the utilized statistical test, $Test_{m,b}$ corresponds to the test statistic m resulting from the $b-$th sample, and T^* corresponds to the critical value. As was previously mentioned, $B = 1000$ and the experiments were repeated 100 times.

ii. The critical values for the proposed tests were set as: $C^* = 0.0039$ (for count-based test), $\kappa^* = 2.6$ for all scale levels $j = 1, ..., J - 1$ (for condition number test).

iii. Significance level for Pearson's $t-$test [79], Spearman's rank correlation [81] and Kendall's [85] was set to $\alpha = 0.05$.

iv. Summary tables with the averages and standard deviations of $\hat{p}_{Test_m}$ obtained from all replications of the experiment.

In the following Tables results of the most representative cases are included. For those models and parameter settings that were omitted, the observed results are almost identical quantitatively and qualitatively as the ones that are presented. The reported values contain only the sample averages, since in most cases, all tests exhibit similar standard deviations that can be observed in the included Box plots in Appendix D.[2]

[2]In each of the included Boxplots there is an entry with a $C2$ denomination, that corresponds to a test statistic based on the weighted differences of two Chi-square distributions, that was included in the simulation study but for the purpose of this Chapter objectives has no influence due to its observed performance.

Remarks and Comments

(i) As it can be observed in Tables 6.12 to 6.18, the counts-based test introduced in Section 6.4.2 exhibits a poor performance in terms of its ability to effectively detect existing correlations between the signals. This could be explained due to the fact that when correlation is present, the variation on the empirical histograms for the entries of Table 6.1 is subtle, with just a slight increment of in the modes for the counts of disagreeing pairs, and pairs of the type (1,1). A possible remedial action would be the investigation of empirical quantiles for each entry, defining a multiple hypothesis test that allows the control of the type I error as a whole, via a Bonferroni-type test.

(ii) Similarly, it is possible to observe that for scale level $j = 4$ the non-parametric tests Kendall and Spearman rank correlation show a significant rate (approx. 24%) of tests that fail to reject H_0. This rate significantly decreases as the scale level increases, which shows a sample size effect on the type II error that is expected due to the nature of the test statistics utilized in the study. This feature if shared by all the implemented tests, which is evident from the examination of the entries of Tables 6.12 to 6.18.

(iii) Among all the implemented statistical tests, in terms of the expected probability of type II error, the T-test shows the best performance across all models and scales, with the exception of scale level $j = 4$ in which the Condition number test achieves the best results.

(iv) In general, it can be noticed that the Condition number test introduced in Section 6.4.3 shows a performance in terms of the expected probability of type II error that even though is not strictly better than the benchmark tests (except for the scale level $j = 4$), achieves values that are small enough to be considered competitive from a practical viewpoint. Moreover, its observed performance is consistent across all model setting and scale levels, which supports its reliability and robustness for real life applications.

A summary plot that illustrates these results is shown in Fig. 6.19.

(v) This fact, added to the significantly better performance of the expected type I error illustrated in Section 6.4.4 and its behavior observed for small scale levels suggest that it can be utilized in practical applications.

Table 6.12: Obtained results for average $\hat{p}_{Test_m}$ for AR(1) with parameter $\phi = -0.9$, $\beta = 0.25$.

j	4	5	6	7	8
Condition number	0.0859	0.03172	0.00542	6.00E-05	0
Counts	9.99E-01	0.99996	1	1	1
T-test	0.10828	0.00398	0	0	0
Kendall	0.2731	0.01748	6.00E-05	0	0
Spearman	0.24096	0.01656	6.00E-05	0	0

Table 6.13: Obtained results for average $\hat{p}_{Test_m}$ for AR(1) with parameter $\phi = -0.7$, $\beta = 0.25$.

j	4	5	6	7	8
Condition number	0.08436	0.03098	0.00446	1.00E-04	0
Counts	0.99902	0.97208	0.87174	0.64876	0.70946
T-test	0.10806	0.00426	0	0	0
Kendall	0.27126	0.0171	2.00E-05	0	0
Spearman	0.23976	0.017	2.00E-05	0	0

Table 6.14: Obtained results for average $\hat{p}_{Test_m}$ for AR(1) with parameter $\phi = 0.9$, $\beta = 0.25$.

j	4	5	6	7	8
Condition number	0.09206	0.03368	0.0049	8.00E-05	0
Counts	0.999	0.97204	0.8733	0.63738	0.31556
T-test	0.1154	0.00454	0	0	0
Kendall	0.27038	0.01724	6.00E-05	0	0
Spearman	0.2372	0.01654	6.00E-05	0	0

Table 6.15: Obtained results for average $\hat{p}_{Test_m}$ for MA(1) with parameter $\theta = 0.9$, $\beta = 0.25$.

j	4	5	6	7	8
Condition number	0.08444	0.03152	0.0054	0.00016	0
Counts	0.99884	0.97278	0.87236	0.6455	0.31878
T-test	0.10698	0.00416	0	0	0
Kendall	0.2699	0.017	8.00E-05	0	0
Spearman	0.23762	0.01624	8.00E-05	0	0

Table 6.16: Obtained results for average $\hat{p}_{Test_m}$ for MA(1) with parameter $\theta = 0.7$, $\beta = 0.25$.

j	4	5	6	7	8
Condition number	0.08732	0.03362	0.0045	1.00E-04	0
Counts	0.99884	0.97274	0.87264	0.65222	0.31496
T-test	0.1104	0.00442	0	0	0
Kendall	0.27784	0.01806	2.00E-05	0	0
Spearman	0.2449	0.0178	4.00E-05	0	0

Table 6.17: Obtained results for average $\hat{p}_{Test_m}$ for MA(1) with parameter $\theta = -0.9$, $\beta = 0.25$.

j	4	5	6	7	8
Condition number	0.08324	0.0278	0.00446	0.00014	0
Counts	0.99914	0.9683	0.8727	0.64704	0.32194
T-test	0.10452	0.0033	0	0	0
Kendall	0.27156	0.0175	0	0	0
Spearman	0.24244	0.01722	0	0	0

Table 6.18: Obtained results for average $\hat{p}_{Test_m}$ for ARMA(1,1) with parameters $\phi = -0.9$ and $\theta = 0.9$, $\beta = 0.25$.

j	4	5	6	7	8
Condition number	0.08788	0.03162	0.0054	0.00018	0
Counts	9.99E-01	0.97154	0.8713	0.64974	0.32064
T-test	0.10854	0.0043	2.00E-05	0	0
Kendall	0.27024	0.01726	4.00E-05	0	0
Spearman	0.23968	0.01694	4.00E-05	0	0

Table 6.19: Obtained results for average $\hat{p}_{Test_m}$ for ARMA(1,1) with parameters $\phi = -0.8$ and $\theta = 0.1$, $\beta = 0.25$.

j	4	5	6	7	8
Condition number	0.08726	0.03112	0.0048	0.00014	0
Counts	0.9987	0.97074	0.86844	0.64848	0.3608
T-test	0.10926	0.00398	0	0	0
Kendall	0.26844	0.01694	4.00E-05	0	0
Spearman	0.2386	0.01638	4.00E-05	0	0

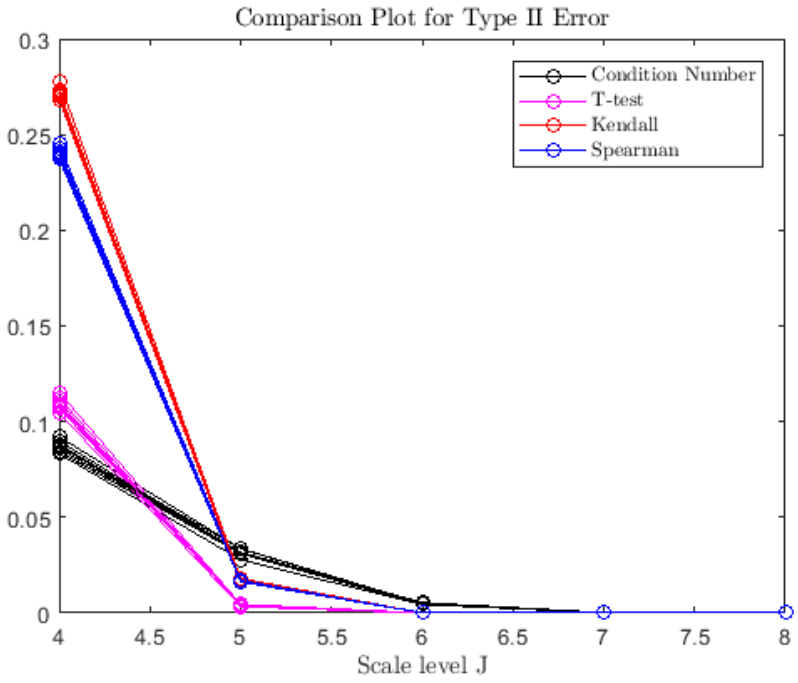
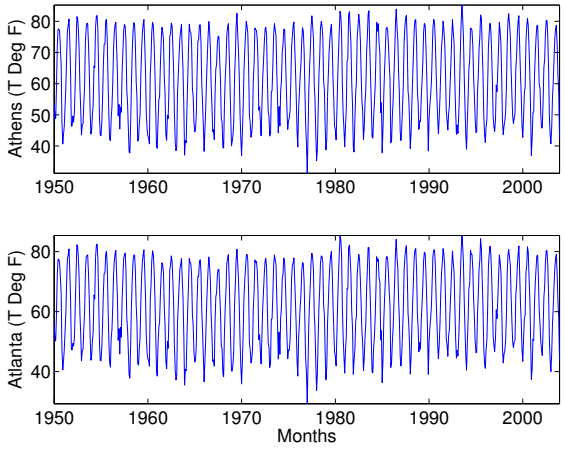


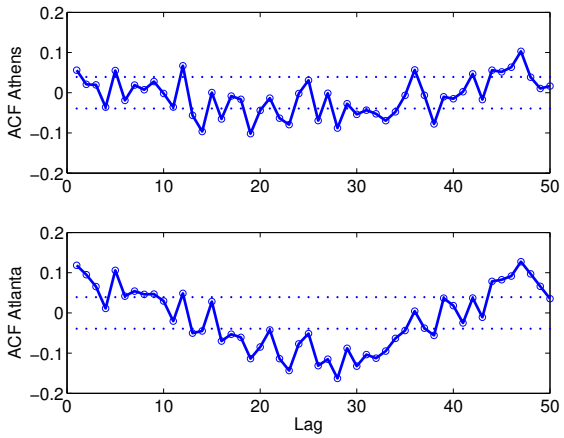
Figure 6.19: Summary plot of average type II error probability for each test statistic

6.5 Application Example: Monthly Temperatures Atlanta-Athens, GA.

We close this Chapter with a study of temperatures from Athens and Atlanta, Georgia, USA. Figure 6.20a plots monthly averaged temperatures (averaged over day of month) for these two stations during the period Jan 1950 — Dec 2003. There are 648 observations in each series[3]. Athens and Atlanta both lie in the Piedmont region of north Georgia and are approximately 60 miles apart.



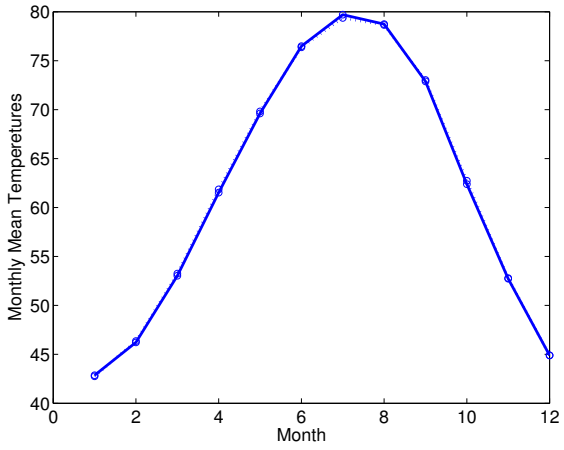
(a) Athens and Atlanta monthly averaged temperatures ($^{\circ}F$)



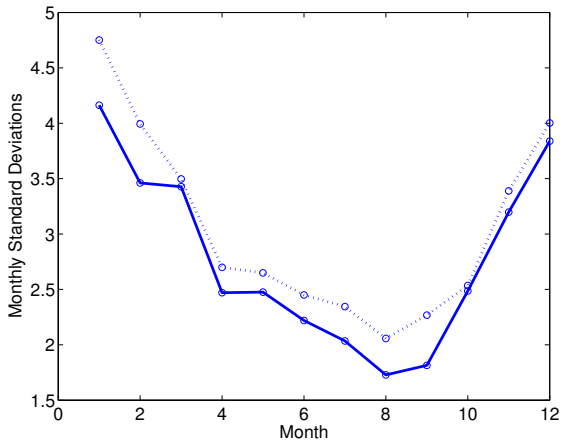
(b) Athens and Atlanta Temperature Sample Autocovariances

As seasonality arises in temperature series taken from temperate zone latitudes (winter tem-

[3]Data was obtained from `https://www.iweathernet.com/atlanta-weather-records`



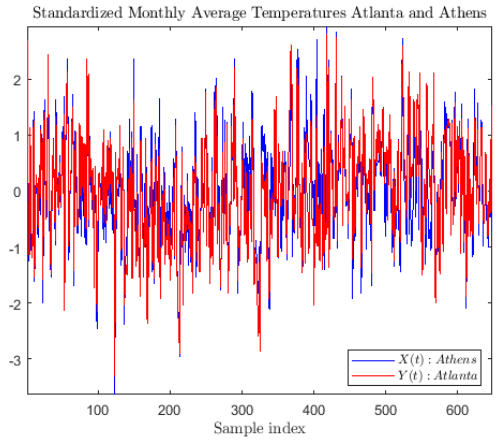
(a) Average monthly Temperatures for the Sample



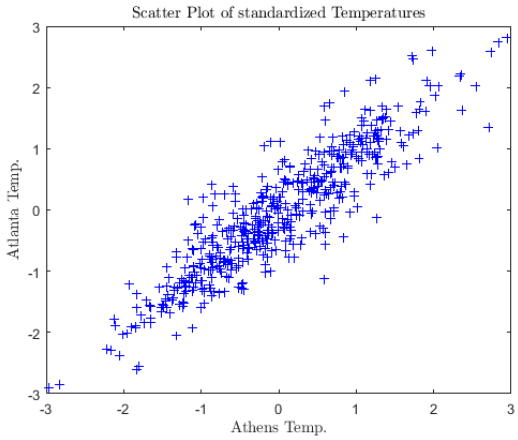
(b) Monthly standard deviations of Temperatures for the Sample

Figure 6.21: Monthly averages and standard deviations of temperatures for both Atlanta (solid blue) and Athens (dashed blue), computed across the samples shown in Fig. 6.20a.

peratures are colder and more variable than summer temperatures), we first standardize each series by month via subtracting a monthly sample mean and then dividing by a monthly sample standard deviation (see Figs. 6.21 and 6.20b). Lund *et al.* (1995)[88] explains more on the stationarizing effects of seasonal standardizations. The sample autocovariance functions for the Athens and Atlanta seasonally standardized series are displayed in Figure 6.20b. The dashed lines here are 95% confidence bounds (pointwise) for white noise.



(a) Athens and Atlanta standardized Temperature Samples



(b) Scatter plot of Athens and Atlanta standardized Temperature Samples

Figures 6.20a and 6.20b give credence to local folklore that Athens and Atlanta enjoy similar weather. As the seasonal mean and standard deviations from the two sites are also very similar, the two towns are indeed similar climatologically. In fact, as seen in Fig. 6.22b, both standardized samples are indeed highly correlated with Pearson's correlation coefficient of $\hat{\rho} = 0.9051$ and a corresponding p-value$= 1.6056 \times 10^{-191}$. Implications of this are that one site could serve as a reference station for the other. This is very useful should a new gauge need to be calibrated, a forecast of future series values need to be made, the quality/legitimacy of future values at one location be questioned.

For this reason, it would be very interesting to analyze the multiscale correlation patterns between these two signals, gathering insights about how the cross-related influences between them boil down into different time scales.

Using the methodology introduced in the previous sections, the following results summarize the findings of the multiscale correlation analysis of the average daily temperatures between Athens and Atlanta (Table 6.20 and Fig. 6.23):

Table 6.20: Estimation results for standardized monthly averages temperatures Athens and Atlanta, GA. This results were obtained using the wavelet filter Symmlet 10.

	$j=4$	$j=5$	$j=6$	$j=7$	$j=8$
$\hat{\rho}_{XY}^{(j)}$	0.8143	0.859	0.9435	0.951	0.9647
w_j	0.0518	0.0942	0.2079	0.2389	0.3765
$\hat{\rho}_{kendall}$	0.6333	0.7419	0.7887	0.8093	0.8439
$\hat{\rho}_{spearman}$	0.8029	0.9069	0.9247	0.9498	0.9637
Condition number	9.7678	13.1894	34.3966	39.7983	55.6653
T-test	5.6108	9.4932	22.7774	34.7901	58.61

Remarks and Comments

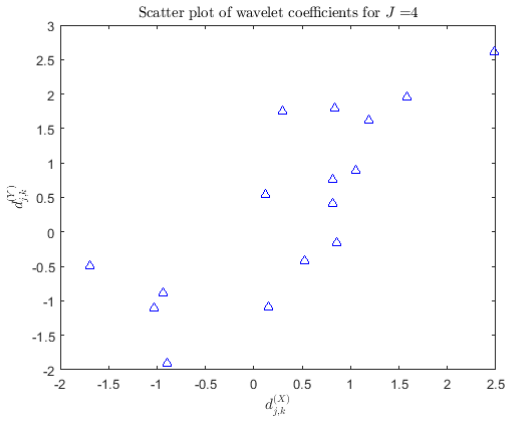
(i) It is interesting to note that the observed correlation in the measuring time scale (days) is evenly spread among the scale levels of the decomposition. Both signals at all scale levels exhibit a significant linear relationships that can be observed in Fig. 6.23. This behavior is clearly captured by the wavelet correlation coefficients, with statistics that show high significance measured by the condition number approach introduced in Section 6.3.1. 6.4.3. These results suggest that there is a strong linear relationship between the two temperature sequences, resembling what was illustrated in

(ii) As can be observed in Table 6.20, the results obtained by using the condition number test statistic are concordant with the other statistical tests utilized as benchmark. This is consistent with the finding presented in sections 6.4.4 and 6.4.5.

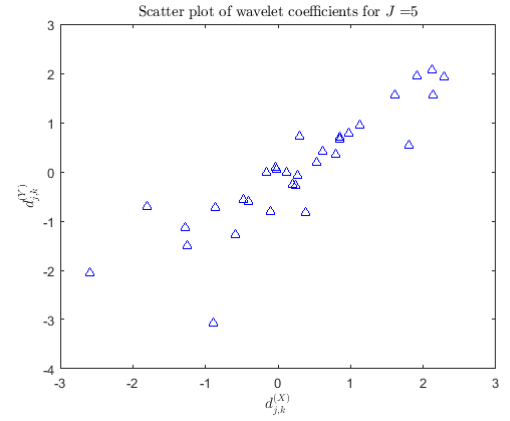
(iii) From the analysis, it follows that the multiscale correlations show that temperatures in Athens and Atlanta are in-phase across multiple scales. This means that temperatures are similar at multiple time resolutions that can be directly linked to the scale levels of the decomposition. For example, for $j = 4$ to $j = 8$ correlations in the wavelet domain are statistically significant, meaning that for time resolutions ranging from 6 minutes to 90 minutes, average temperatures measured during those intervals are statistically extremely similar for the two cities.

This application example although very simple, is very illustrative for the extra insights given by the multiscale correlation analysis via wavelets as compared with the usual sample correlation. Its simplicity of implementation allows an easy extension to more complicated problems such as:

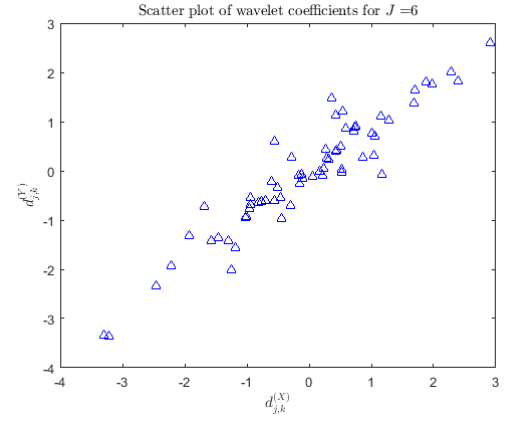
- Correlation analysis of multiple sites (e.g. spatially-distributed sensors)
- Time varying correlation between two signals (e.g. application of this methodology for time rolling windows to capture non-stationary behaviors)
- Multiscale correlation analysis at different time-shifts.



(a) $J = 4$

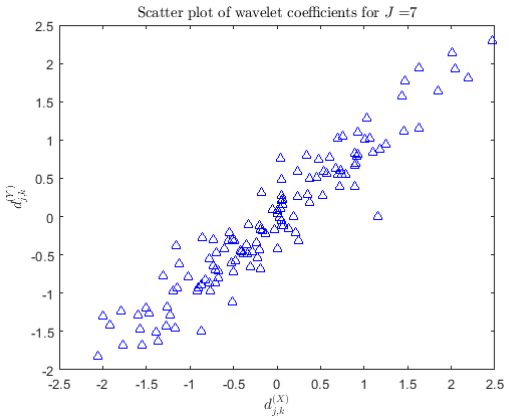


(b) $J = 5$

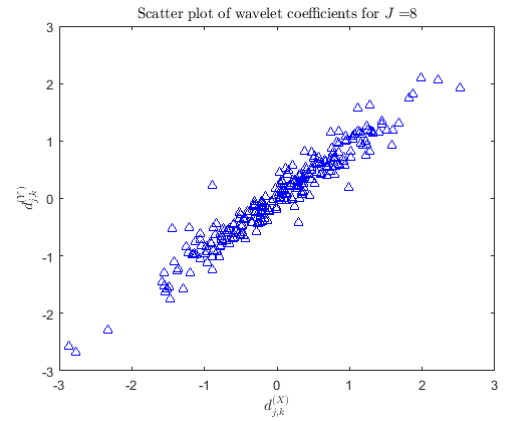


(c) $J = 6$

Figure 6.23: Scatter plots of the wavelet coefficients for each scale, corresponding to the DWT of the Athens (x-axis) and Atlanta (y-axis) daily average temperatures. This results were obtained using the wavelet filter Symmlet 10.



(a) $J = 7$



(b) $J = 8$

Figure 6.24: Scatter plots of the wavelet coefficients for each scale, corresponding to the DWT of the Athens (x-axis) and Atlanta (y-axis) daily average temperatures. This results were obtained using the wavelet filter Symmlet 10.

6.6 Conclusions

In this chapter a wavelet based correlation decomposition using an orthogonal DWT was introduced and analyzed. One important feature of this approach is that it breaks down the sample covariance into an additive structure that leads to a weighted sum of level-wise correlations between expansion coefficients in the wavelet domain. Thus, it enables a scale-by-scale analysis of the existing linear relationships between two signals.

In addition, some interesting distributional and statistical properties of wavelet coefficients were provided for certain types of stationary processes, building a theoretical background that was used for the development of different test statistics. In this context, two statistical tests that exploit the whitening property of wavelets were proposed and analyzed via a simulation-based study. Their performance was compared with the well-known Pearson's $t-$test and non-parametric statistical procedures such as Spearman's rank correlation and Kendalls's tau, by using simulated stationary processes aimed to resemble possible scenarios that are expected to occur in real-life.

As can be observed from Tables 6.4 to 6.11, in most scenarios the proposed test statistics tend outperform the other statistical procedures used as benchmark, leading to a significantly smaller average type I error probability than the other tests. Similarly, for the expected probability of type II error, it was noticed that the Condition number test introduced in Section 6.4.3 showed a performance that even though not strictly better than the benchmark tests (except for the scale level $j = 4$), achieves values that are small enough to be considered competitive from a practical viewpoint.

Also, as a by-product of the simulation study, it was possible to observe that when the analyzed signals exhibit high oscillations concentrated in short time spans, the usual test statistics perform poorly in terms of an increased false rejection rate of the no correlation hypothesis (between 14% and 30%+). These results were obtained for stationary processes of the type

AR(1), MA(1) and ARMA(1,1), so it may not hold true for other kinds of stochastic processes.

In summary, in this Chapter a novel and competitive tool for the significance analysis of multiscale correlation was introduced, analyzed and evaluated, hence contributing to the existing methodologies in the scientific community for the correlation analysis of stationary time series.

Appendices

APPENDIX A

APPENDIX CHAPTER 2

A.1 Derivation of the unbiased partial-data estimator.

In this section we provide the derivation for the partial-data estimator proposed in 2.2.2. From (2.36) and (2.37), it follows:

$$\mathbb{E}(\hat{f}_J(x)) = \sum_{k=0}^{2^J-1} \mathbb{E}\left[\tilde{c_{Jk}}\right] \cdot \phi_{J,k}^{per}(x)\,. \tag{A.1}$$

Using (2.35), the expectation in the left hand side (lhs) of (A.1) is given by:

$$\mathbb{E}\left[\tilde{c_{Jk}}\right] = \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}\frac{\phi_{J,k}^{per}(Y_{(i)})}{1-\hat{G}(Y_{(i)})}\right] - \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}\frac{\mathbb{1}_{(\delta_i=0)}(1-\hat{F}(Y_{(i)}))}{1-\hat{G}(Y_{(i)})}\phi_{J,k}^{per}(Y_{(i)})\right]\,. \tag{A.2}$$

Assuming iid samples and $G(y)$ known, the first expectation on the rhs of (A.2) can be obtained as:

$$\mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}\frac{\phi_{J,k}^{per}(Y_{(i)})}{1-\hat{G}(Y_{(i)})}\right] = \mathbb{E}_Y\left[\frac{\phi_{J,k}^{per}(Y)}{1-G(Y)}\right]\,. \tag{A.3}$$

Similarly, provided iid samples, and both $F(y)$ and $G(y)$ known, the expectation of the second term in the rhs of (A.2) can be obtained as:

$$\mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}\frac{\mathbb{1}_{(\delta_i=0)}(1-\hat{F}(Y_{(i)}))}{1-\hat{G}(Y_{(i)})}\phi_{J,k}^{per}(Y_{(i)})\right] = \mathbb{E}_{Y,\delta=0}\left[\frac{(1-F(Y))\phi_{J,k}^{per}(Y)}{1-G(Y)}\right]\,. \tag{A.4}$$

Since $f_{Y,\delta}(y,\delta=0)=g(y)(1-F(y))$, it follows:

$$\mathbb{E}_{Y,\delta=0}\left[\frac{(1-F(Y))\phi^{per}_{J,k}(Y)}{1-G(Y)}\right]=\mathbb{E}_T\left[\frac{(1-F(T))\phi^{per}_{J,k}(T)}{1-G(T)}\right]-\mathbb{E}_T\left[\frac{F(T)(1-F(T))\phi^{per}_{J,k}(T)}{1-G(T)}\right]. \tag{A.5}$$

Finally, combining (A.3) and (A.5), it follows:

$$\mathbb{E}\left[\tilde{c_{Jk}}\right]=\mathbb{E}_Y\left[\frac{\phi^{per}_{J,k}(Y)}{1-G(Y)}\right]-\mathbb{E}_T\left[\frac{(1-F(T))\phi^{per}_{J,k}(T)}{1-G(T)}\right]+\mathbb{E}_T\left[\frac{F(T)(1-F(T))\phi^{per}_{J,k}(T)}{1-G(T)}\right]. \tag{A.6}$$

Using (2.27) and (A.6), (A.6) takes the form:

$$\mathbb{E}\left[\tilde{c_{Jk}}\right]=c_{Jk}+\mathbb{E}_T\left[\frac{F(T)(1-F(T))\phi^{per}_{J,k}(T)}{1-G(T)}\right], \tag{A.7}$$

which further implies that for (A.1), it follows:

$$\mathbb{E}(\hat{f}_J(x))=f_J(x)+\sum_{k=0}^{2^J-1}\mathbb{E}_T\left[\frac{F(T)(1-F(T))\phi^{per}_{J,k}(T)}{1-G(T)}\right]\phi^{per}_{J,k}(x)\,. \tag{A.8}$$

To facilitate notation, define $b_{J,k}=\mathbb{E}_T\left[\frac{F(T)(1-F(T))\phi^{per}_{J,k}(T)}{1-G(T)}\right]$. Thus, (A.1) can be represented as:

$$\mathbb{E}(\hat{f}_J(x))=f_J(x)+\sum_{k=0}^{2^J-1}b_{J,k}\cdot\phi^{per}_{J,k}(x)\,. \tag{A.9}$$

Using the same approach as in (2.29), $b_{J,k}$ (i.e. the wavelet coefficient that define the bias of $\hat{f}_J(x)$ can be estimated from the sample as follows:

$$\tilde{b}_{J,k}=\frac{1}{N}\sum_{i=1}^{N}\mathbb{1}_{(\delta_i=0)}\frac{\hat{F}(Y_i)(1-\hat{F}(Y_i))\phi^{per}_{J,k}(Y_i)}{1-\hat{G}(Y_i)}\,. \tag{A.10}$$

Therefore, the biased-corrected version of the estimator can be represented as:

$$\hat{f}_J^*(x) = \hat{f}_J(x) - \sum_{k=0}^{2^J-1} \tilde{b}_{J,k} \cdot \phi_{J,k}^{per}(x)\,, \tag{A.11}$$

$$\hat{f}_J^*(x) = \sum_{k=0}^{2^J-1} \tilde{c}_{J,k}^* \cdot \phi_{J,k}^{per}(x)\,, \tag{A.12}$$

where:

$$\tilde{c}_{J,k}^* = \tilde{c}_{J,k} - \tilde{b}_{J,k} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{1-\hat{G}(Y_{(i)})} - \frac{\mathbb{1}_{(\delta_{(i)}=0)}(1-\hat{F}(Y_{(i)}))}{1-\hat{G}(Y_{(i)})} - \frac{\mathbb{1}_{(\delta_{(i)}=0)}\hat{F}(Y_{(i)})(1-\hat{F}(Y_{(i)}))}{1-\hat{G}(Y_{(i)})}\right)\cdot \phi_{J,k}^{per}(Y_{(i)})\,. \tag{A.13}$$

Note that (A.13) can be further simplified into:

$$\tilde{c}_{J,k}^* = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{1-\mathbb{1}_{(\delta_{(i)}=0)}(1-\hat{F}(Y_{(i)}))(1+\hat{F}(Y_{(i)}))}{1-\hat{G}(Y_{(i)})}\right)\phi_{J,k}^{per}(Y_{(i)})\,. \tag{A.14}$$

Computing the expectation of the bias-correction coefficient $\tilde{b}_{Jk}$, it follows:

$$\mathbb{E}_Y\left[\tilde{b}_{Jk}\right] = b_{Jk} - \mathbb{E}_T\left[\frac{F(T)^2(1-F(T))}{1-G(T)}\phi_{Jk}^{per}(T)\right]\,. \tag{A.15}$$

Therefore, the bias of $\tilde{b}_{Jk}$ can be corrected by defining $\tilde{b}_{Jk}^* = \tilde{b}_{Jk} + \mathbb{E}_T\left[\frac{F(T)^2(1-F(T))}{1-G(T)}\phi_{Jk}^{per}(T)\right]$. Using the empirical argument as in (A.10), $\tilde{b}_{Jk}^*$ can be estimated by:

$$\tilde{b}_{Jk}^* = \tilde{b}_{Jk} + \frac{1}{N}\sum_{i=1}^{N}\frac{\mathbb{1}_{(\delta_{(i)}=0)}F(Y_i)^2(1-F(Y_i))}{1-G(Y_i)}\phi_{Jk}^{per}(Y_i)\,. \tag{A.16}$$

This implies that the updated bias-corrected estimator of b_{Jk} can be represented as:

$$\tilde{b}_{Jk}^* = \frac{1}{N}\sum_{i=1}^{N}\frac{\mathbb{1}_{(\delta_{(i)}=0)}F(Y_i)(1-F(Y_i))(1+F(Y_i))}{1-G(Y_i)}\phi_{Jk}^{per}(Y_i)\,. \tag{A.17}$$

Taking the expectation of $\tilde{b}^{*}_{Jk}$, it follows:

$$\mathbb{E}_Y\left[\tilde{b}^{*}_{Jk}\right] = b_{Jk} - \mathbb{E}_T\left[\frac{F(T)^3(1-F(T))}{1-G(T)}\phi^{per}_{Jk}(T)\right]. \tag{A.18}$$

Following the same methodology used to derive (A.17), an updated bias-corrected estimate of $\tilde{b}^{*}_{Jk}$, denoted by $\tilde{b}^{**}_{Jk}$ can be represented as:

$$\tilde{b}^{**}_{Jk} = \frac{1}{N}\sum_{i=1}^{N}\frac{\mathbb{1}_{(\delta_{(i)}=0)}F(Y_i)(1-F(Y_i))(1+F(Y_i)+F(Y_i)^2))}{1-G(Y_i)}\phi^{per}_{Jk}(Y_i)\,. \tag{A.19}$$

Taking the expectation of $\tilde{b}^{**}_{Jk}$, it follows:

$$\mathbb{E}_Y\left[\tilde{b}^{**}_{Jk}\right] = b_{Jk} - \mathbb{E}_T\left[\frac{F(T)^4(1-F(T))}{1-G(T)}\phi^{per}_{Jk}(T)\right]. \tag{A.20}$$

This implies that the bias-corrected estimate of b_{Jk} represented as $\tilde{b}^{***}_{Jk} = \tilde{b}^{**}_{Jk} + \mathbb{E}_T\left[\frac{F(T)^4(1-F(T))}{1-G(T)}\phi^{per}_{Jk}(T)\right]$ can be iteratively updated. Thus, following the same process as before, it follows:

$$\tilde{b}^{***}_{Jk} = \frac{1}{N}\sum_{i=1}^{N}\frac{\mathbb{1}_{(\delta_{(i)}=0)}F(Y_i)(1-F(Y_i))(1+F(Y_i)+F(Y_i)^2+F(Y_i)^3))}{1-G(Y_i)}\phi^{per}_{Jk}(Y_i)\,. \tag{A.21}$$

From the last set of equations, it follows that this process can be repeated sequentially, infinitely many times. This implies that:

$$\tilde{\tilde{b}}_{Jk} = \frac{1}{N}\sum_{i=1}^{N}\frac{\mathbb{1}_{(\delta_{(i)}=0)}F(Y_i)(1-F(Y_i))\sum_{l=0}^{\infty}F(Y_i)^l}{1-G(Y_i)}\phi^{per}_{Jk}(Y_i)\,, \tag{A.22}$$

provided $0 < F(Y) < 1$. Therefore, it follows that $\sum_{l=0}^{\infty}F(Y_i)^l$ is a convergent series. In fact, it is a geometric power series that satisfies:

$$\sum_{l=0}^{\infty}F(Y_i)^l = \frac{1}{1-F(Y_i)}\,. \tag{A.23}$$

Therefore, this implies that (A.22) takes the form:

$$\tilde{\tilde{b}}_{Jk} = \frac{1}{N}\sum_{i=1}^{N} \frac{\mathbb{1}_{(\delta_{(i)}=0)} F(Y_i)}{1-G(Y_i)} \phi_{Jk}^{per}(Y_i)\,. \tag{A.24}$$

Clearly, $\tilde{\tilde{b}}_{Jk}$ is an unbiased estimate of b_{Jk}. Therefore, we conclude that the unbiased estimate of the c_{Jk} coefficient, denoted by $\tilde{\tilde{c}}_{Jk}$ is given by:

$$\tilde{\tilde{c}}_{Jk} = \tilde{c}_{Jk} - \tilde{\tilde{b}}_{Jk} = \frac{1}{N}\sum_{i=1}^{N} \frac{\mathbb{1}_{(\delta_{(i)}=1)}}{1-G(Y_i)} \phi_{Jk}^{per}(Y_i)\,, \tag{A.25}$$

thus, it is possible to define the partial-data density estimator $\hat{f}^{PD}(x)$ as:

$$\hat{f}^{PD}(x) = \sum_{k=0}^{2^J-1} \tilde{c}_{Jk} \cdot \phi_{J,k}^{per}(x)\,, \tag{A.26}$$

where:

$$\tilde{c}_{Jk} = \frac{1}{N}\sum_{i=1}^{N} \frac{\mathbb{1}_{(\delta_{(i)}=1)}}{1-\hat{G}(Y_i)} \phi_{Jk}^{per}(Y_i)\,, \tag{A.27}$$

which is unbiased for $f_J(x)$, provided $G(y)$ is known and $0 < F(Y) < 1$.

A.2 Proof of Lemma 2.2.1

Assume the following conditions are satisfied:

(i) The scaling function ϕ that generates the orthonormal set $\left\{\phi_{Jk}^{per}, 0 \le k \le 2^J\right\}$ has compact support and satisfies $||\theta_\phi(x)||_\infty = C < \infty$, for $\theta_\phi(x) := \sum_{r\in\mathbb{Z}} |\phi(x-r)|$.

(ii) $\exists\ F \in \mathbb{L}_2(\mathbb{R})$ such that $|K(x,y)| \le F(x-y)$, for all $x, y \in \mathbb{R}$, where $K(x,y) = \sum_{k\in\mathbb{Z}} \phi(x-k)\phi(y-k)$.

(iii) For $s = m + 1$, $m \geq 1$, integer, $\int |x|^s F(x)dx < \infty$.

(iv) $\int (y - x)^l K(x, y)dy = \delta_{0,l}$ for $l = 0, ..., s$.

(v) The density f belongs to the s-sobolev space $W_2^s([0, 1])$, $s > 1$ defined as:

$$W_2^s([0,1]) = \left\{ f \mid f \in \mathbb{L}_2([0,1]),\, \exists\, f^{(1)}, ..., f^{(s)} \text{ s.t. } f^{(l)} \in \mathbb{L}_2([0,1]),\, l = 1, ..., s \right\}.$$

Then, it follows:

$$\sup_{f \in W_2^s([0,1])} \mathbb{E}\left[||\hat{f}^{PD}(x) - f(x)||_2^2 \right] \leq C_1 \frac{2^J}{N} + C_2 2^{-2sJ}\,, \text{and} \tag{A.28}$$

for $J = \lfloor \log_2(N) - \log_2(\log(N)) \rfloor$:

$$\sigma_J^2(x) = \mathcal{O}(\log(N)^{-1})\,, \tag{A.29}$$

$$\mathbb{E}\left[\parallel f(x) - \hat{f}^{PD}(x) \parallel_2^2 \right] \leq \mathcal{O}(N^{-s} \log(N)^s) \tag{A.30}$$

for $C_1 > 0$, $C_2 > 0$ independent of J and N, provided $\exists\, \alpha_1 \mid 0 < \alpha_1 < \infty$, $C_T \in (0, 1)$ such that $(1 - G(y)) \geq C_T e^{-\alpha_1 y}$ for $y \in [0, 1)$, and $0 \leq F(y) \leq 1\ \forall y \in [0, 1]$.

Proof

Note that $\hat{f}^{PD}(x)$ can be expressed as follows:

$$\hat{f}^{PD}(x) = \frac{1}{N} \sum_{i=1}^{N} w_i K_J(Y_i, x)\,, \tag{A.31}$$

where $w_i = \frac{\delta_i}{1 - G(Y_i)}$, and $K_J(x, Y_i) = 2^J \sum_{k \in \mathbb{Z}} \phi(2^J x - k)\phi(2^J y - k)$, for $i = 1, ..., N$. Since it is assumed that $\exists\, \alpha_1 \mid 0 < \alpha_1 < \infty$, $C_T \in (0, 1)$ such that $(1 - G(y)) \geq C_T e^{-\alpha_1 y}$

for $y \in [0,1)$, this implies that $0 \leq w_i \leq \frac{e^{\alpha_1}}{C_T}$, for $i = 1, ..., N$.

Also, it is possible to bound the $\mathbb{L}_2$ risk of the estimator $\hat{f}^{PD}(x)$ as follows:

$$\mathbb{E}\left[||\hat{f}^{PD}(x) - f(x)||_2^2\right] \leq 2\left\{\mathbb{E}\left[||\hat{f}^{PD}(x) - \mathbb{E}[\hat{f}^{PD}(x)]||_2^2\right] + ||\mathbb{E}[\hat{f}^{PD}(x)] - f(x)||_2^2\right\}, \tag{A.32}$$

where the first term in the rhs of (A.31) corresponds to $Var(\hat{f}^{PD}(x))$ and the second, to $bias(\hat{f}^{PD}(x))$.

Bound for $\mathbb{E}\left[||\hat{f}^{PD}(x) - \mathbb{E}[\hat{f}^{PD}(x)]||_2^2\right]$

From (A.31), it follows:

$$\hat{f}^{PD}(x) - \mathbb{E}[\hat{f}^{PD}(x)] = \frac{1}{N}\sum_{i=1}^{N}\left(w_i K_J(x, Y_i) - \mathbb{E}[w_i K_J(x, Y_i)]\right) .$$

Define $Z_i(x) = w_i K_J(x, Y_i) - \mathbb{E}[w_i K_J(x, Y_i)]$ and $\tilde{Z}_i(x) = K_J(x, Y_i) - \mathbb{E}[K_J(x, Y_i)]$. Clearly, $\mathbb{E}[Z_i(x)] = \mathbb{E}[\tilde{Z}_i(x)] = 0$. This implies:

$$|\hat{f}^{PD}(x) - \mathbb{E}[\hat{f}^{PD}(x)| \leq \frac{e^{\alpha_1}}{C_T}\frac{1}{N}\left|\sum_{i=1}^{N}\tilde{Z}_i(x)\right| ,$$

since $0 \leq w_i \leq \frac{e^{\alpha_1}}{C_T}$, for $i = 1, ..., N$. Therefore, it follows:

$$\begin{aligned} |\hat{f}^{PD}(x) - \mathbb{E}[\hat{f}^{PD}(x)|^2 &\leq \frac{e^{2\alpha_1}}{C_T^2}\frac{1}{N^2}\left|\sum_{i=1}^{N}\tilde{Z}_i(x)\right|^2 \\ \mathbb{E}\left[\int_0^1 |\hat{f}^{PD}(x) - \mathbb{E}[\hat{f}^{PD}(x)|^2 dx\right] &\leq \frac{e^{2\alpha_1}}{C_T^2}\frac{1}{N^2}\mathbb{E}\left[\int_0^1\left|\sum_{i=1}^{N}\tilde{Z}_i(x)\right|^2 dx\right] . \end{aligned}$$

From conditions (i) and (ii), Fubini's thorem implies:

$$\begin{aligned}\mathbb{E}\left[\int_0^1 |\hat{f}^{PD}(x) - \mathbb{E}[\hat{f}^{PD}(x)|^2 dx\right] &\leq \frac{e^{2\alpha_1}}{C_T^2}\frac{1}{N^2}\int_0^1 \mathbb{E}\left[\left|\sum_{i=1}^N \tilde{Z}_i(x)\right|^2\right] dx \\ &\leq \frac{e^{2\alpha_1}}{C_T^2}\frac{1}{N}\int_0^1 \mathbb{E}[\tilde{Z}_1(x)^2] dx\,, \qquad \text{(A.33)}\end{aligned}$$

where (A.33) follows from the fact that $\tilde{Z}_i(x)$ are iid, with $\mathbb{E}[\tilde{Z}_i(x)] = 0$, and $\mathbb{E}[\tilde{Z}_i(x)^2] < \infty$. This, together with the application of Rosenthal's inequality implies $\mathbb{E}\left[\left|\sum_{i=1}^N \tilde{Z}_i(x)\right|^2\right] \leq \sum_{i=1}^N \mathbb{E}[\tilde{Z}_i(x)^2] = N\,\mathbb{E}[\tilde{Z}_1(x)^2]$.

Since $\mathbb{E}[\tilde{Z}_1(x)^2] = \mathbb{E}[K_J(x, Y_1)^2] - (K_J f_Y(x))^2 \leq \mathbb{E}[K_J(x, Y_1)^2]$, where $K_J f_Y(x) = \int_0^1 K_J(x,u) f_Y(u) du$, and the fact that $|K_J(x,y)| = 2^J |K(2^J x, 2^J y)|$, it follows from (A.33) and condition (ii):

$$\mathbb{E}\left[||\hat{f}^{PD}(x) - \mathbb{E}[\hat{f}^{PD}(x)]||_2^2\right] \leq \frac{e^{2\alpha_1}}{C_T^2}\frac{1}{N}\int_0^1 \mathbb{E}[K_J(x, Y_1)^2] dx$$

$$\begin{aligned}\int_0^1 \mathbb{E}[K_J(x, Y_1)^2] dx &\leq 2^J \int_0^1 \left[\int_{-2^J y}^{2^J(1-y)} F^2(v) dv\right] f_Y(y) dy \\ &\leq 2^J ||F||_2^2\,. \qquad \text{(A.34)}\end{aligned}$$

Therefore, substituting (A.34) into (A.33), it follows:

$$\mathbb{E}\left[||\hat{f}^{PD}(x) - \mathbb{E}[\hat{f}^{PD}(x)]||_2^2\right] \leq \frac{||F||_2^2 e^{2\alpha_1}}{C_T^2}\frac{2^J}{N}\,. \qquad \text{(A.35)}$$

Bound for $||\mathbb{E}[\hat{f}^{PD}(x)] - f(x)||_2^2$

According to corollary 8.2 [57], if $f \in W_2^s([0,1])$ then $||K_J f - f||_2^2 = \mathcal{O}\left(2^{-2Js}\right)$. Furthermore, assume conditions (i)-(iv) are satisfied. Since $\mathbb{E}[\hat{f}^{PD}(x)] = K_J f(x)$, it follows:

$$||\mathbb{E}[\hat{f}^{PD}(x)] - f(x)||_2^2 \leq C_2\, 2^{-2Js}\,. \tag{A.36}$$

Finally, putting together (A.35) and (A.36), it follows:

$$\sup_{f \in W_2^s([0,1])} \mathbb{E}\left[||\hat{f}^{PD}(x) - f(x)||_2^2\right] \leq C_1 \frac{2^J}{N} + C_2 2^{-2sJ}\,, \tag{A.37}$$

as desired, for $C_1 = \frac{||F||_2^2 e^{2\alpha_1}}{C_T^2}$ and $C_2 > 0$, independent of N and J.

From (A.37), by choosing $J = \lfloor \log_2(N) - \log_2(\log(N)) \rfloor$, it follows that $\sigma_J^2(x) = \mathcal{O}(\log(N)^{-1})$. Furthermore, this also implies that $\sup_{f \in W_2^s([0,1])} \mathbb{E}\left[||\hat{f}^{PD}(x) - f(x)||_2^2\right] = \mathcal{O}(N^{-s}\log(N)^2$, which completes the proof.

Remarks

Note that from (A.37), it is possible to choose the multiresolution level J such that the upper bound for the $\mathbb{L}_2$ risk is minimized. In this context, it is possible to show that $J^*(N) = \frac{1}{2s+1}\log_2\left(\frac{2s\,C_2}{C_1}\right) + \frac{1}{2s+1}\log_2(N)$ achieves that result. Moreover, under this choice of J, it follows:

$$\sup_{f \in W_2^s([0,1])} \mathbb{E}\left[||\hat{f}^{PD}(x) - f(x)||_2^2\right] \leq \tilde{C} N^{-\frac{2s}{2s+1}}\,.$$

A.3 Proof of Lemma 2.2.2

Under the assumptions and definitions stated in 2.2.1 and 2.2.4, and choosing $J = \lfloor \log_2(N) - \log_2(\log(N)) \rfloor$, it follows:

$$\sup_{f \in W_2^s([0,1])} \mathbb{E}\left[\| f(x) - \hat{f}^{PD}(x) \|_2^2 \right] = \mathcal{O}(N^{-s} \log(N)^s) . \tag{A.38}$$

Proof

Assume conditions (i)-(iv) established in A.2 are satisfied. Furthermore, assume $\exists\, \gamma > 0$ and a constant $C \in (0,1)$ such that $1 - \hat{G}(y) \geq Ce^{-\gamma y}$, for $y \in [0,1)$. Note that $\hat{f}^{PD}(x)$ can be expressed as follows:

$$\hat{f}^{PD}(x) = \frac{1}{N} \sum_{i=1}^{N} w_i K_J(Y_i, x) , \tag{A.39}$$

where $w_i = \frac{\delta_i}{1-\hat{G}(Y_i)}$, and $K_J(x, Y_i) = 2^J \sum_{k \in \mathbb{Z}} \phi(2^J x - k)\phi(2^J y - k)$, for $i = 1, ..., N$. Since it is assumed that $\exists\, \gamma > 0$ and a constant $C \in (0,1)$ such that $1 - \hat{G}(y) \geq Ce^{-\gamma y}$, for $y \in [0,1)$, this implies that $0 \leq w_i \leq \frac{e^\gamma}{C}$, for $i = 1, ..., N$. Thus, following the same methodology as in A.2, it follows that by choosing $J = \lfloor \log_2(N) - \log_2(\log(N)) \rfloor$:

$$\sup_{f \in W_2^s([0,1])} \mathbb{E}\left[\| f(x) - \hat{f}^{PD}(x) \|_2^2 \right] = \mathcal{O}(N^{-s} \log(N)^s) . \tag{A.40}$$

Remarks

(i) Observe that by following the same methodology as in A.2, it is possible to obtain:

$$\sup_{f \in W_2^s([0,1])} \mathbb{E}\left[||\hat{f}^{PD}(x) - f(x)||_2^2 \right] \leq C_1 \frac{2^J}{N} + C_2 2^{-2sJ} ,$$

for $C_1 = \frac{||F||_2^2 e^{2\gamma}}{C^2}$ and $C_2 > 0$, independent of N and J.

(ii) The last result implies that by choosing $J^*(N) = \frac{1}{2s+1}\log_2\left(\frac{2s\,C_2}{C_1}\right) + \frac{1}{2s+1}\log_2(N)$, the $\mathbb{L}_2$ risk of the estimator $\hat{f}^{PD}(x)$ when G is unknown is also mean square consistent, and achieves a convergence rate of the order $\sim N^{-\frac{2s}{2s+1}}$.

A.4 Proof of Lemma 2.2.3

From (2.62), and for N large it follows that the rhs of (2.63) corresponds to the sum of normally distributed random variables $\sim N(0, \sigma_{Jk}^2)$ which is indeed a normally distributed random variable. To obtain its variance, it can be used the fact that:

$$Cov\left(\sqrt{N}(\tilde{c}_{Jk} - c_{Jk})\,, \sqrt{N}(\tilde{c}_{Jl} + c_{Jl})\right) = N\,\mathbb{E}\left[(\tilde{c}_{Jk} - c_{Jk})(\tilde{c}_{Jl} - c_{Jl})\right]. \tag{A.41}$$

Thus, (2.55) implies:

$$\mathbb{E}\left[N(\tilde{c}_{Jk} - c_{Jk})(\tilde{c}_{Jl} - c_{Jl})\right] = N\,\left(\mathbb{E}\left[\tilde{c}_{Jk}\tilde{c}_{Jl}\right] - c_{Jk}c_{Jl}\right) - (c_{Jk} - c_{Jl})\mathcal{O}(\log(N))\,. \tag{A.42}$$

Using (2.46), it follows:

$$\tilde{\tilde{c}}_{Jk}\tilde{\tilde{c}}_{Jl} = A_1 + A_2 + A_3 + A_4 + A_5 + A_6 + A_7 + A_8 + A_9\,, \tag{A.43}$$

where:

$$A_1 = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\frac{\delta_i\delta_j\phi_{Jk}^{per}(Y_i)\phi_{Jl}^{per}(Y_j)}{(1-G_T(Y_i))(1-G_T(Y_j))} \tag{A.44}$$

$$A_2 = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\frac{\delta_i\phi_{Jk}^{per}(Y_i)U_{jl}}{1-G_T(Y_i)} \tag{A.45}$$

$$A_3 = \frac{1}{N}R_{Nl}\sum_{i=1}^{N}\frac{\delta_i\phi_{Jk}^{per}(Y_i)}{1-G_T(Y_i)} \tag{A.46}$$

$$A_4 = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\frac{\delta_j\phi_{Jl}^{per}(Y_j)U_{ik}}{1-G_T(Y_j)} \tag{A.47}$$

$$A_5 = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}U_{ik}U_{jl} \tag{A.48}$$

$$A_6 = \frac{1}{N}R_{Nl}\sum_{i=1}^{N}U_{ik} \tag{A.49}$$

$$A_7 = \frac{1}{N}R_{Nk}\sum_{i=1}^{N}\frac{\delta_j\phi_{Jl}^{per}(Y_j)}{1-G_T(Y_j)} \tag{A.50}$$

$$A_8 = \frac{1}{N}R_{Nk}\sum_{i=1}^{N}U_{il} \tag{A.51}$$

$$A_9 = R_{Nk}R_{Nl}\,. \tag{A.52}$$

From the last set of equations, it is possible to observe that the following pairs have the same structure (i.e. they are symmetric counter parts of each other) (A_2, A_4), (A_3, A_7) and (A_6, A_8).

Now, assuming that $\mathbb{E}\left[\frac{\delta^2\phi_{Jk}^{per}(Y)\phi_{Jl}^{per}(Y)}{(1-G(Y))^2}\right]$ is finite (provided (2.47), (2.48), and the assumptions stated above) for A_1, it follows:

$$\begin{aligned}\mathbb{E}[A_1] &= \frac{1}{N^2}\mathbb{E}\left[\sum_{i=1}^{N}\sum_{j=1}^{N}\frac{\delta_i\phi_{Jk}^{per}(Y_i)U_{jl}}{1-G_T(Y_i)}\right] \\ &= \frac{1}{N}\mathbb{E}\left[\frac{\delta^2\phi_{Jk}^{per}(Y)\phi_{Jl}^{per}(Y)}{(1-G(Y))^2}\right] + \frac{N-1}{N}c_{Jk}c_{Jl}\,.\end{aligned} \tag{A.53}$$

Consider possible upper bounds for $\gamma_{1,Jk}(x)$ and $\gamma_{2,Jk}(x)$. Using the corresponding definitions stated in 2.2.4, it follows:

$$\begin{aligned}
\gamma_{1,Jk}(x) &= \frac{1}{(1-F_X(x))(1-G_T(x))}\int_x^1 \phi_{Jk}^{per}(u)f_X(u)du \\
&\leq \frac{\parallel f_X \parallel_\infty M\, 2^{-\frac{J}{2}}}{c(1-G_T(x))^{\beta+1}} \qquad \text{(A.54)} \\
&\leq \frac{e^{\frac{\alpha_1(\beta+1)}{2}} \parallel f_X \parallel_\infty M\, 2^{-\frac{J}{2}}}{c\, C_T^{\frac{\beta+1}{2}}}\,. \qquad \text{(A.55)}
\end{aligned}$$

Similarly, for $\gamma_{2,Jk}(x)$, it follows:

$$\begin{aligned}
\gamma_{2,Jk}(x) &\leq \int_0^1 \frac{|\phi_{Jk}^{per}(u)|f_X(u)du}{(1-F_X(u))(1-G_T(u))} \\
&\leq \int_0^1 \frac{|\phi_{Jk}^{per}(u)|f_X(u)du}{c(1-G_T(u))^{\beta+1}} \\
&\leq \frac{e^{\frac{\alpha_1(\beta+1)}{2}} \parallel f_X \parallel_\infty M\, 2^{-\frac{J}{2}}}{c\, C_T^{\frac{\beta+1}{2}}}\,. \qquad \text{(A.56)}
\end{aligned}$$

Therefore, the last result implies that for $k, l = 0, ..., 2^J - 1$ and $\tilde{i} \in \{0, 1\}$:

$$\begin{aligned}
\gamma_{\tilde{i},Jk}(x)\gamma_{\tilde{i},Jl}(x) &\leq \frac{\parallel f_X \parallel_\infty^2 M^2\, 2^{-J}}{c^2(1-G_T(x))^{2(\beta+1)}} \\
&\leq \frac{e^{\alpha_1(\beta+1)} \parallel f_X \parallel_\infty^2 M^2\, 2^{-J}}{c^2 C_T^{\beta+1}} \\
&\leq \mathcal{O}(N^{-1}\log(N))\,. \qquad \text{(A.57)}
\end{aligned}$$

Using the last result,it follows:

$$\begin{aligned}
\mathbb{E}\left[(1-\delta)\gamma_{1,Jk}(Y)\gamma_{2,Jl}(Y)\right] &\leq \frac{e^{\alpha_1(\beta+1)} \parallel f_X \parallel_\infty^2 M^2\, 2^{-J}}{c^2 C_T^{\beta+1}} \int_0^1 (1-G(u))f_X(u)du \\
&\leq \frac{e^{\alpha_1(\beta+1)} \parallel f_X \parallel_\infty^2 M^2\, 2^{-J}}{c^2 C_T^{\beta+1}}\,. \qquad \text{(A.58)}
\end{aligned}$$

Clearly, from the last result the same upper bound holds for $\mathbb{E}\left[(1-\delta)^2\gamma_{1,Jk}(Y)\gamma_{1,Jl}(Y)\right]$ and $\mathbb{E}\left[\gamma_{2,Jk}(Y)\gamma_{2,Jl}(Y)\right]$.

Now, for the pair (A_2, A_4), it follows:

$$\begin{aligned}
\mathbb{E}\left[A_2\right] &= \frac{1}{N^2}\mathbb{E}\left[\sum_{i=1}^{N}\sum_{j=1}^{N}\frac{\delta_i\phi_{Jk}^{per}(Y_i)U_{jl}}{1-G_T(Y_i)}\right] \\
&= -\frac{1}{N}\mathbb{E}\left[\frac{\delta\phi_{Jk}^{per}(Y)\gamma_{2,Jk}(Y)}{1-G_T(Y)}\right] \\
&\leq \frac{1}{N}\frac{e^{\frac{\alpha_1(\beta+1)}{2}}\parallel f_X\parallel_\infty\ M\,2^{-\frac{J}{2}}}{c\,C_T^{\frac{\beta+1}{2}}}\int_0^1|\phi_{Jk}^{per}(u)|c(1-G_T(u))^{\beta-1}g_T(u)du \\
&\leq \frac{1}{N}\frac{e^{\frac{\alpha_1(\beta+1)}{2}}\parallel f_X\parallel_\infty\parallel g_T\parallel_\infty\ M^2\,2^{-J}}{C_T^{\frac{\beta+1}{2}}} \\
&\leq \mathcal{O}(N^{-2}\log(N))\,, \qquad \text{(A.59)}
\end{aligned}$$

In the case of the pair (A_3, A_7) we have:

$$\begin{aligned}
\mathbb{E}\left[A_3\right] &= \frac{1}{N}\mathbb{E}\left[R_{Nl}\sum_{i=1}^{N}\frac{\delta_i\phi_{Jk}^{per}(Y_i)}{1-G_T(Y_i)}\right] \\
&\leq \mathcal{O}(N^{-1}\log(N))c_{Jk} \qquad \text{(A.60)}
\end{aligned}$$

For the term A_5 we have the following:

$$\begin{aligned}
\mathbb{E}\left[A_5\right] &= \frac{1}{N^2}\mathbb{E}\left[\sum_{i=1}^{N}\sum_{j=1}^{N}U_{ik}U_{jl}\right] \\
&= \frac{1}{N}\mathbb{E}\left[U_kU_l\right]
\end{aligned}$$

Therefore, using the definition of U_k:

$$\mathbb{E}\left[A_5\right] = \frac{1}{N}\mathbb{E}\left[(1-\delta)^2\gamma_{1,Jk}(Y)\gamma_{1,Jl}(Y) - (1-\delta)\gamma_{1,Jk}(Y)\gamma_{2,Jl}(Y) - (1-\delta)\gamma_{1,Jl}(Y)\gamma_{2,Jk}(Y) + \gamma_{2,Jk}(Y)\gamma_{2,Jl}(Y)\right]$$

From the last result and (A.57), it is clear that:

$$\mathbb{E}\left[A_5\right] \leq \mathcal{O}(N^{-2}\log(N)) \tag{A.61}$$

Now, for the pair (A_6, A_8) it is clear from the zero mean condition of U_k and the fact that $R_N = \mathcal{O}(N^{-1}\log(N))$ that:

$$\begin{aligned} \mathbb{E}\left[A_6\right] &\leq \mathcal{O}(N^{-2}\log(N)) && \text{(A.62)} \\ \mathbb{E}\left[A_9\right] &\leq \mathcal{O}(N^{-2}\log(N)^2) && \text{(A.63)} \end{aligned}$$

Putting together (A.53)-(A.63) in (A.43) we get:

$$\mathbb{E}\left[\tilde{\tilde{c}}_{Jk}\tilde{\tilde{c}}_{Jl}\right] \leq \frac{1}{N}\mathbb{E}\left[\frac{\delta^2\phi_{Jk}^{per}(Y)\phi_{Jl}^{per}(Y)}{(1-G(Y))^2}\right] + \frac{N-1}{N}c_{Jk}c_{Jl} + \mathcal{O}(N^{-2}\log(N)) + \mathcal{O}(N^{-2}\log(N)^2) + \mathcal{O}(N^{-1}\log(N))(c_{Jk}+c_{Jl}) \tag{A.64}$$

Therefore, (A.42) becomes:

$$\mathbb{E}\left[N(\tilde{\tilde{c}}_{Jk} - c_{Jk})(\tilde{\tilde{c}}_{Jl} - c_{Jl})\right] \leq \mathbb{E}\left[\frac{\delta^2\phi_{Jk}^{per}(Y)\phi_{Jl}^{per}(Y)}{(1-G(Y))^2}\right] - c_{Jk}c_{Jl} + \mathcal{O}(N^{-1}\log(N)^2) \tag{A.65}$$

Therefore, for N large the last result suggests that:

$$Cov\left(\sqrt{N}(\tilde{\tilde{c}}_{Jk}-c_{Jk}),\sqrt{N}(\tilde{\tilde{c}}_{Jl}+c_{Jl})\right)\approx\mathbb{E}\left[\frac{\delta^2\phi_{Jk}^{per}(Y)\phi_{Jl}^{per}(Y)}{(1-G(Y))^2}-c_{Jk}c_{Jl}\right] \tag{A.66}$$

Finally, in light of the last result and the properties of the Normal Distribution, result (2.64) follows. Therefore,

$$\hat{f}^{PD}(x)\overset{app.}{\sim}N\left(f(x),\tfrac{1}{N}\sum_{k=0}^{2^J-1}\sigma_{Jk}^2(\phi_{Jk}^{per}(x))^2+\tfrac{2}{N}\sum_{k<l}\mathbb{E}\left[\tfrac{\delta^2\phi_{Jk}^{per}(Y)\phi_{Jl}^{per}(Y)}{(1-G(Y))^2}-c_{Jk}c_{Jl}\right]\phi_{Jk}^{per}(x)\phi_{Jl}^{per}(x)\right) \tag{A.67}$$

APPENDIX B

APPENDIX CHAPTER 3

B.1 Proof of $\int_0^1 \phi_{jk}^{per}(x)dx = 2^{-\frac{j}{2}}$.

For $j \leq 0$, the Strang-Fix condition (see [89]) gives $\phi_{jk}(x) \equiv 2^{-j/2}$, so the claim is trivial. In the case of $j > 0$, it follows:

$$\begin{aligned}
\int_0^1 \phi_{jk}^{per}(x)dx &= \sum_{m\in Z} \int_0^1 \phi_{jk}(x+m)dx \\
&= \sum_{m\in Z} \int_0^1 2^{j/2}\phi(2^j(x+m)-k)dx \\
&\quad [2^j(x+m) = t] \\
&= \sum_{m\in Z} \int_{m2^j}^{(m+1)2^j} 2^{j/2}2^{-j}\phi(t-k)dt \\
&= 2^{-j/2} \int_R \phi(t-k)dt = 2^{-j/2}\,, \qquad \text{(B.1)}
\end{aligned}$$

which shows the desired result.

B.2 Important results from Multivariate Taylor Series expansion.

In this section we provide definitions and results that will be needed for the derivation of the density estimator $\hat{h}_n(\mathbf{x})$ properties.

Define $\boldsymbol{\alpha} := (\alpha_1, ..., \alpha_p)$, $\boldsymbol{\beta} := (\beta_1, ..., \beta_p)$, $|\boldsymbol{\alpha}| := \sum_{j=1}^p \alpha_j$, $|\boldsymbol{\beta}| := \sum_{j=1}^p \beta_j$ and $\boldsymbol{\alpha}! =$

$\prod_{j=1}^{p} \alpha_j!$. Similarly, let:

$$\mathbf{x}^{\boldsymbol{\alpha}} := \prod_{j=1}^{p} x_j^{\alpha_j},\ \mathbf{x} \in \mathbb{R}^p\,, \tag{B.2}$$

$$\partial^{\boldsymbol{\alpha}} f := \partial_1^{\alpha_1} \cdot \ldots \cdot \partial_p^{\alpha_p} f = \frac{\partial^{|\boldsymbol{\alpha}|} f}{\partial x_1^{\alpha_1} \cdot \ldots \cdot \partial x_p^{\alpha_p}}\,. \tag{B.3}$$

From the multinomial theorem, it follows that for any $\mathbf{x} \in \mathbb{R}^p$, and any integer $k > 0$:

$$\begin{aligned} |\mathbf{x}|^k &= \sum_{\alpha_1} \sum_{\alpha_2} \cdot \ldots \cdot \sum_{\alpha_p} \frac{k!}{\alpha_1! \cdot \ldots \cdot \alpha_p!} x_1^{\alpha_1} \cdot \ldots \cdot x_p^{\alpha_p},\ s.t.\ |\boldsymbol{\alpha}| = k\,, \\ &= \sum_{|\boldsymbol{\alpha}|=k} \frac{k!}{\boldsymbol{\alpha}!} \mathbf{x}^{\boldsymbol{\alpha}}\,. \end{aligned} \tag{B.4}$$

Now, suppose a function $f : \mathbb{R}^p \to \mathbb{R}$, such that $f \in \mathbb{C}^k$ on a convex open set $\mathbb{S} \subset \mathbb{R}^p$. We are interested in the Taylor series expansion of $f(\mathbf{x})$ around a point $\mathbf{x}_0 \in \mathbf{S}$.

If we look at the behavior of $f()$ over the points that are in the line between $\mathbf{x}$ and $\mathbf{x}_0$, it follows that any of those points $\mathbf{x}^*$ can be contained in a set defined as:

$$L(\mathbf{x}, \mathbf{x}_0) = \{\mathbf{x}^* \in \mathbf{S}\ \ s.t. \forall t \in [0,1]\, \mathbf{x}^* = \mathbf{x}_0 + t(\mathbf{x} - \mathbf{x}_0)\}\ .$$

Using the last definition, we have that $\forall \mathbf{x} \in L(\mathbf{x}, \mathbf{x}_0)$, $f(\mathbf{x}^*) = f(\mathbf{x}_0 + t(\mathbf{x} - \mathbf{x}_0)) = g(t)$. Define $\mathbf{v} = \mathbf{x} - \mathbf{x}_0$, therefore, for $1 \leq l \leq k$, it follows:

$$g^{(l)}(t) = (\mathbf{v} \bullet \nabla)^l \cdot f(\mathbf{x}_0 + t \cdot \mathbf{v})\,,$$

where

$$
\begin{aligned}
(\mathbf{v} \bullet \nabla)^{(l)} f &= (v_1 \frac{\partial}{\partial x_1} + ... + v_p \frac{\partial}{\partial x_p})^l f \,, \\
&= \sum_{|\boldsymbol{\alpha}|=l} \frac{l!}{\boldsymbol{\alpha}!} v_1^{\alpha_1} \cdot ... \cdot v_p^{\alpha_p} \frac{\partial^{\alpha_1}}{\partial x_1^{\alpha_1}} \cdot ... \cdot \frac{\partial^{\alpha_p}}{\partial x_p^{\alpha_p}} f \,, \\
&= \sum_{|\boldsymbol{\alpha}|=l} \frac{l!}{\boldsymbol{\alpha}!} v_1^{\alpha_1} \cdot ... \cdot v_p^{\alpha_p} \partial^{\boldsymbol{\alpha}} f \,. \qquad \text{(B.5)}
\end{aligned}
$$

If we now make a Taylor series expansion of $g(t)$ around a point t_0, for $\delta \in [t, t_0]$ it follows:

$$
g(t) = \sum_{l=0}^{k-1} \frac{g^{(l)}(t_0)}{l!} (t - t_0)^l + \frac{g^{(k)}(\delta)(t - t_0)^k}{k!}
$$

Letting $t_0 \to 0$ and $t \to 1$, we have that $g^{(l)}(t_0) \to \sum_{|\boldsymbol{\alpha}|=l} \frac{l!}{\boldsymbol{\alpha}!} v_1^{\alpha_1} \cdot ... \cdot v_p^{\alpha_p} \partial^{\boldsymbol{\alpha}} f(\mathbf{x}_0)$ and $g(t) \to f(\mathbf{x})$.

Therefore, the Taylor series expansion of f around $\mathbf{x}_0$ is given by:

$$
f(\mathbf{x}) = \sum_{l=0}^{k-1} \frac{(\mathbf{v} \bullet \nabla)^{(l)} f(\mathbf{x}_0)}{l!} + \frac{(\mathbf{v} \bullet \nabla)^{(k)} f(\mathbf{x}_0 + \delta \mathbf{v})}{k!} \,. \qquad \text{(B.6)}
$$

Define the Taylor series expansion of $f()$ around $\mathbf{x}_0$ of order k and its remainder term as as:

$$
\begin{aligned}
f_{\mathbf{x}_0,k}(\mathbf{x}) &= \sum_{l=0}^{k-1} \frac{(\mathbf{v} \bullet \nabla)^{(l)} f(\mathbf{x}_0)}{l!} \,, \\
R_{\mathbf{x}_0,k}(\mathbf{v}) &= \frac{(\mathbf{v} \bullet \nabla)^{(k)} f(\mathbf{x}_0 + \delta \mathbf{v})}{k!} \,.
\end{aligned}
$$

. Then, by Taylor's theorem and (B.4), it follows:

$$
|R_{\mathbf{x}_0,k}(\mathbf{v})| \leq \frac{M_h}{(k+1)!} ||\mathbf{v}||_1^{(k+1)} \,, \qquad \text{(B.7)}
$$

provided assumption (**A4**) holds. Finally, from results (B.6) and (B.7), it follows that:

$$f(\mathbf{x}) - f_{\mathbf{x}_0,k}(\mathbf{x}) = R_{\mathbf{x}_0,k}(\mathbf{v})\,. \tag{B.8}$$

B.3 Consistency of the Kernel density estimator.

In this section, we provide an overview of the asymptotic properties of the density estimator $\hat{h}_n()$, which are needed later to show the consistency of the estimates $\hat{\beta}_0$ and $\hat{c}_{Jk}^{(l)}$. See [90] for a detailed discussion of the Kernel Density estimator properties.

Consider a kernel-type density estimator given by:

$$\hat{h}_n(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{\delta^p} K\left(\frac{\mathbf{x}-\mathbf{x}_i}{\delta}\right)\,, \tag{B.9}$$

where $\frac{1}{\delta^p}K\left(\frac{\mathbf{x}-\mathbf{x}_i}{\delta}\right) := K_\delta(\mathbf{x},\mathbf{x}_i)$ and $\delta = \delta(n) > 0$ is a proper bandwidth, and $K(\mathbf{x}) > 0$ is the kernel function. This last condition guarantees that $\hat{h}_n(\mathbf{x})$ is non-negative and continuous as a finite sum of positive and continuous functions.

From (3.9) and (3.12) it is clear that we need a kernel function such that $\hat{h}_n(\mathbf{x}) > 0$ and bounded in the support of $h()$. Assume that the chosen kernel satisfies:

(Ak1) $K(\mathbf{x})$ is real-valued, Borel measurable function with $||K||_\infty < \infty$.

(Ak2) $K(\mathbf{x})$ has $\beta - 1$ ($\beta \geq 2$) vanishing moments, i.e. $\int K(\mathbf{v})||\mathbf{v}||_1^s d\mathbf{v} = 0,\ s = 1, ..., \beta - 1$.

(Ak3) $K(\mathbf{x})$ belongs to $\mathbb{L}_2(\mathbb{R}^p)$.

(Ak4) $K(\mathbf{x})$ satisfies $\int K(\mathbf{v})d\mathbf{v} = 1$ and $\int K(\mathbf{v})||\mathbf{v}||_1^{\beta} d\mathbf{v} = M_{k,\beta} < \infty$.

(Ak5) $\sup_{\mathbf{x},\mathbf{y}\in[0,1]^p} |K_\delta(\mathbf{x},\mathbf{y})| \leq C_1\delta^{-p}$, for $\delta = \delta(n) > 0$, $C_1 > 0$.

(Ak6) $\sup_{\mathbf{x}\in[0,1]^p} \mathbb{E}\left[(K_\delta^2(\mathbf{x},\mathbf{x}_i)\right)] \leq C_2\delta^{-p}$, for $\delta = \delta(n) > 0$, $C_1 > 0$, $C_2 > 0$.

Lemma B.3.1. *Consider a kernel that satisfies **(Ak1)-(Ak6)** and a random variable **X** defined on a probability space $(\Omega, \Im, \mathbb{P})$ with density $h()$. Assume **(A1)** and **(A5)** are satisfied, then $\hat{h}_n()$ is consistent, provided $n\delta^p \to \infty$ and $\delta^p \to 0$ as $n \to \infty$.*

This means that $\forall \boldsymbol{x} \in [0,1]^p$ for which $\mathbb{P}\{\omega \in \Omega \,|\, \boldsymbol{X}(\omega) = \boldsymbol{x}\} > 0$, it follows:

$$\hat{h}_n(\boldsymbol{x}) \xrightarrow{\mathbb{P}} h(\boldsymbol{x}) \tag{B.10}$$

Proof. Consider an iid sample $\{y_i, \mathbf{x}_i\}_{i=1}^n$. It follows that the expectation of the density estimator (B.9) takes the form:

$$\mathbb{E}[\hat{h}_n(\mathbf{x})] = \int K(\mathbf{v})h(\mathbf{x} + \delta\mathbf{v})d\mathbf{v}$$

If we subtract $h(\mathbf{x})$ from the above expression, we get:

$$\begin{aligned}
\mathbb{E}[\hat{h}_n(\mathbf{x}) - h(\mathbf{x})] &= \int K(\mathbf{v})\left[h(\mathbf{x}+\delta\mathbf{v}) - h(\mathbf{x})\right] d\mathbf{v}\,, \\
&= \int K(\mathbf{v})\left[h(\mathbf{x}+\delta\mathbf{v}) - h_{\mathbf{x},\beta}(\mathbf{x}+\delta\mathbf{v}) + h_{\mathbf{x},\beta}(\mathbf{x}+\delta\mathbf{v}) - h(\mathbf{x})\right] d\mathbf{v}\,, \\
&= \int K(\mathbf{v})\left[h(\mathbf{x}+\delta\mathbf{v}) - h_{\mathbf{x},\beta}(\mathbf{x}+\delta\mathbf{v})\right] d\mathbf{v} + \\
&\quad + \int K(\mathbf{v})\left[h_{\mathbf{x},\beta}(\mathbf{x}+\delta\mathbf{v}) - h(\mathbf{x})\right] d\mathbf{v}\,,
\end{aligned}$$

provided assumption (**Ak4**) holds.

From (B.6) that in the second term of (B.11): $h(\mathbf{x}+\delta\mathbf{v})_{\mathbf{x},\beta} - h(\mathbf{x}) = \sum_{l=1}^{k-1} \frac{(\mathbf{v}\bullet\nabla)^{(l)} f(\mathbf{x}_0)}{l!}$. Moreover, by assumption (**Ak2**):

$$\int K(\mathbf{v}) \left[h_{\mathbf{x},\beta}(\mathbf{x}+\delta\mathbf{v}) - h(\mathbf{x})\right] d\mathbf{v} = 0\,. \tag{B.11}$$

Similarly, the first term of the rhs of (B.11) can be expressed as: $h(\mathbf{x}+\delta\mathbf{v}) - h_{\mathbf{x},\beta}(\mathbf{x}+\delta\mathbf{v}) = R_{\mathbf{x},\beta}(\delta\mathbf{v})$, provided (B.8). Therefore, from (B.7), it follows:

$$\begin{aligned}
\mathbb{E}[\hat{h}_n(\mathbf{x}) - h(\mathbf{x})] &= \int K(\mathbf{v}) R_{\mathbf{x},\beta}(\delta\mathbf{v}) d\mathbf{v}\,, \\
|\mathbb{E}[\hat{h}_n(\mathbf{x}) - h(\mathbf{x})]| &\leq \int K(\mathbf{v}) |R_{\mathbf{x},\beta}(\delta\mathbf{v})| d\mathbf{v}\,, \\
&\leq \frac{M_h \delta^\beta}{\beta!} \int K(\mathbf{v}) ||\mathbf{v}||_1^\beta d\mathbf{v}\,, \\
|bias(\hat{h}_n)| &\leq C(h,\beta)\delta^\beta\,,
\end{aligned} \tag{B.12}$$

where $C(h,\beta) = \frac{M_h M_{k,\beta}}{\beta!}$. Also, from the last set of equations, it is possible to obtain:

$$\sup_{\mathbf{x}\in[0,1]^p} \left|\mathbb{E}[\hat{h}_n(\mathbf{x}) - h(\mathbf{x})]\right| \leq C(h,\beta)\delta^\beta\,. \tag{B.13}$$

Now, for a fixed $\mathbf{x}$, the variance of $\hat{h}_n(\mathbf{x})$, can be expressed and bounded as follows:

$$
\begin{aligned}
Var\left(\hat{h}_n(\mathbf{x})\right) &= \frac{1}{n\delta^{2p}} Var\left(K\left(\frac{\mathbf{x}-\mathbf{X}_1}{\delta}\right)\right), \\
&\leq \frac{1}{n\delta^{2p}} \mathbb{E}\left[K\left(\frac{\mathbf{x}-\mathbf{X}_1}{\delta}\right)^2\right], \\
&\leq \frac{1}{n\delta^{p}} \int K(\mathbf{v})^2 h(\mathbf{x}+\delta\mathbf{v}) d\mathbf{v}, \\
&\leq \frac{M \cdot C}{n\delta^{p}}, \qquad \text{(B.14)} \\
\sup_{\mathbf{x}\in[0,1]^p} \mathbb{E}\left[\left(\hat{h}_n(\mathbf{x}) - h(\mathbf{x})\right)^2\right] &\leq \frac{M \cdot C}{n\delta^{p}}, \qquad \text{(B.15)}
\end{aligned}
$$

provided assumptions (**A6**) and (**Ak3**) hold, for $C = \int K(\mathbf{v})^2 d\mathbf{v}$.

From the above results, it is possible to express the $\mathbb{L}_2$ risk of the estimator $\hat{h}_n(\mathbf{x})$ as:

$$
\mathbf{R}\left(\hat{h}_n, h\right) = Var\left(\hat{h}_n(\mathbf{x})\right) + bias(\hat{h}_n(\mathbf{x}))^2 .
$$

Using results (B.12) and (B.15), we get that:

$$
\mathbf{R}\left(\hat{h}_n, h\right) \leq \frac{M \cdot C}{n\delta^{p}} + C(h,\beta)^2 \delta^{2\beta} \qquad \text{(B.16)}
$$

Clearly, as $n \to \infty$, if $n\delta^p \to \infty$ and $\delta^p \to 0$, it follows that $\mathbf{R}\left(\hat{h}_n, h\right) \to 0$. Therefore, $\hat{h}_n(\mathbf{x})$ is mean-square consistent, which automatically implies:

$$
\hat{h}_n(\mathbf{x}) \xrightarrow{\mathbb{P}} h(\mathbf{x}) .
$$

If we ignore the constants (with respect to n) in (B.16), it is possible to show that the bandwidth $\delta(n)$ that minimizes $\mathbf{R}\left(\hat{h}_n, h\right)$ is given by $\delta^* \sim n^{-\frac{1}{2\beta+p}}$ (up to a constant) and thus,

$\mathbf{R}\left(\hat{h}_n, h\right)^* \geq C \cdot n^{-\frac{2\beta}{2\beta+p}}$. Similarly, under this optimal bandwidth, we have that (B.15) becomes:

$$\sup_{\mathbf{x}\in[0,1]^p} \mathbb{E}\left[\left(\hat{h}_n(\mathbf{x}) - h(\mathbf{x})\right)^2\right] \leq M \cdot C n^{-\frac{2\beta}{2\beta+p}}\,. \tag{B.17}$$

□

B.4 Derivation of an upper bound for $\mathbb{E}\left[\left(\frac{Y\phi_{Jk}^{per}(X_l)}{h(\mathbf{X})}\right)^2\right]$.

Consider a sequence of constant positive piecewise functions $\{g_b,\, b \geq 1\}$ that satisfy:

(i) $0 < g_b(\mathbf{x}) \leq h(\mathbf{x}), \forall\, \mathbf{x} \in [0,1]^p$.

(ii) $g_b(\mathbf{x}) \leq g_{b+1}(\mathbf{x}), \forall\, \mathbf{x} \in [0,1]^p$.

(iii) $g_b(\mathbf{x}) \uparrow h(\mathbf{x})$ as $b \to \infty$.

Define $g_b(\mathbf{x})$ for $b \geq \lfloor \log_2\left(\frac{1}{\epsilon_h}\right) \rfloor$ as follows:

$$g_b(\mathbf{x}) = \begin{cases} \frac{r}{2^b} & \frac{r}{2^b} \leq h(\mathbf{x}) \leq \frac{r+1}{2^b} \quad r = 1, ..., b\cdot 2^b - 1 \\ b & h(\mathbf{x}) > b \end{cases}$$

Therefore, we can express $g_b(\mathbf{x})$ as:

$$g_b(\mathbf{x}) = \sum_{r=1}^{b\cdot 2^b-1} \left(\frac{r}{2^b}\right) \mathbf{1}_{\left\{\mathbf{x}:\, \frac{r}{2^b} \leq h(\mathbf{x}) \leq \frac{r+1}{2^b}\right\}} + b \cdot \mathbf{1}_{\{\mathbf{x}:\, h(\mathbf{x}) > b\}}\,. \tag{B.18}$$

From (B.18), for a fixed b define:

$$\begin{aligned} \Omega_{rb} &= \left\{\mathbf{x}:\, \frac{r}{2^b} \leq h(\mathbf{x}) \leq \frac{r+1}{2^b}\right\}, r = 1, ..., b\cdot 2^b - 1\,, \\ \Omega_b &= \{\mathbf{x}:\, h(\mathbf{x}) > b\}\,. \end{aligned}$$

This partitions the support of the random vector $\mathbf{X}$ into $b \cdot 2^b$ disjoints subsets for which $\bigcup_{r=1}^{b\cdot 2^b-1} \{\Omega_{rb}\} \bigcup \{\Omega_b\} = [0,1]^p$. Similarly, the sequence of functions $\{g_b,\, b \geq 1\}$ approximate $h(\mathbf{x})$ from below, in a quantization fashion. Therefore:

$$\mathbb{E}\left[\left(\frac{Y\phi_{Jk}^{per}(X_l)}{h(\mathbf{X})}\right)^2\right] = \sum_{r=1}^{b\cdot 2^b-1} \mathbb{E}\left[\left(\frac{Y\phi_{Jk}^{per}(X_l)}{h(\mathbf{X})}\right)^2 \mathbf{1}_{\left\{\mathbf{x}:\, \frac{r}{2^b} \leq h(\mathbf{x}) \leq \frac{r+1}{2^b}\right\}}\right] + \mathbb{E}\left[\left(\frac{Y\phi_{Jk}^{per}(X_l)}{h(\mathbf{X})}\right)^2 \mathbf{1}_{\{\mathbf{x}: h(\mathbf{x})>b\}}\right],$$

$$\mathbb{E}\left[\left(\frac{Y\phi_{Jk}^{per}(X_l)}{h(\mathbf{X})}\right)^2\right] \leq \left((p\cdot M_f + |\beta_0|)^2 + \sigma^2\right)\left(\sum_{r=1}^{b\cdot 2^b-1} \mathbb{E}\left[\frac{\phi_{Jk}^{per}(X_l)^2 \mathbf{1}_{\left\{\mathbf{x}:\, \frac{r}{2^b} \leq h(\mathbf{x}) \leq \frac{r+1}{2^b}\right\}}}{h(\mathbf{X})^2}\right] + \mathbb{E}\left[\frac{\phi_{Jk}^{per}(X_l)^2 \mathbf{1}_{\{\mathbf{x}:\, h(\mathbf{x})>b\}}}{h(\mathbf{X})^2}\right]\right),$$

$$\begin{aligned}
\mathbb{E}\left[\left(\frac{Y\phi_{Jk}^{per}(X_l)}{h(\mathbf{X})}\right)^2\right] &\leq \left((p\cdot M_f + |\beta_0|)^2 + \sigma^2\right)\left(\sum_{r=1}^{b\cdot 2^b-1} \int_{\Omega_{rb}} \frac{\phi_{Jk}^{per}(X_l)^2}{h(\mathbf{X})} d\mathbf{x} + \int_{\Omega_b} \frac{\phi_{Jk}^{per}(X_l)^2}{h(\mathbf{X})} d\mathbf{x}\right), \\
&\leq \left((p\cdot M_f + |\beta_0|)^2 + \sigma^2\right)\left(\sum_{r=1}^{b\cdot 2^b-1} \int_{\Omega_{rb}} \frac{\phi_{Jk}^{per}(X_l)^2}{g_b(\mathbf{X})} d\mathbf{x} + \int_{\Omega_b} \frac{\phi_{Jk}^{per}(X_l)^2}{g_b(\mathbf{X})} d\mathbf{x}\right), \\
&\leq \left((p\cdot M_f + |\beta_0|)^2 + \sigma^2\right)\left(\sum_{r=1}^{b\cdot 2^b-1} \frac{2^b}{r} \int_{\Omega_{rb}} \phi_{Jk}^{per}(X_l)^2 d\mathbf{x} + \frac{1}{b}\int_{\Omega_b} \phi_{Jk}^{per}(X_l)^2 d\mathbf{x}\right), \\
&\leq \left((p\cdot M_f + |\beta_0|)^2 + \sigma^2\right)\left(2^b(b2^b-1) + \frac{1}{b}\right), \\
&\leq \left((p\cdot M_f + |\beta_0|)^2 + \sigma^2\right)\left\{\inf_{b \geq \lfloor \log_2\left(\frac{1}{\epsilon_h}\right)\rfloor}\left(2^b(b2^b-1) + \frac{1}{b}\right)\right\}, \\
&\leq \left((p\cdot M_f + |\beta_0|)^2 + \sigma^2\right)\left\{\frac{1}{\epsilon_h}\left(\lceil \log_2(\frac{1}{\epsilon_h})\rceil - 1\right) + \frac{1}{\lceil \log_2(\frac{1}{\epsilon_h})\rceil}\right\}, \qquad \textbf{(B.19)}
\end{aligned}$$

where the last result holds since the function $f(b) = 2^b(b2^b-1) + \frac{1}{b}$ is strictly increasing in b and $b \geq \lfloor \log_2\left(\frac{1}{\epsilon_h}\right)\rfloor$.

Remarks

Note that this bound could be further improved if instead of piecewise constant functions, we use a different approximation technique. Nonetheless, obtaining tight bounds is not the intention of this derivations, but instead showing that the second moment of the random variable $\frac{Y\phi_{Jk}^{per}(X_l)}{h(\mathbf{X})}$ is bounded under suitable conditions.

B.5 Asymptotic correlation between $\frac{Y_i\phi_{Jk}^{per}(X_{il})}{\hat{h}_n(\mathbf{X}_i)}$ and $\hat{\beta}_0$.

Similarly as for V_{c1} in (3.27), consider the asymptotic behavior of $V_{c3} = Cov\left(\frac{1}{n}\sum_{i=1}^{n}\frac{Y_i\phi_{Jk}^{per}(X_{il})}{\hat{h}_n(\mathbf{X}_i)}\,,\,2^{-\frac{J}{2}}\hat{\beta}_0\right)$ assuming conditions (**A1**)-(**A5**) and (**Ak1**)-(**Ak6**) hold. Using the covariance properties and the iid sample $\{y_i = f(\mathbf{x}_i), \mathbf{x}_i\}_{i=1}^n$, it follows:

$$V_{c3} = \frac{2^{-\frac{J}{2}}}{n^2}\left\{\sum_{i=1}^{n} Cov\left(\frac{Y_i\phi_{Jk}^{per}(X_{il})}{\hat{h}_n(\mathbf{X}_i)}\,,\,\frac{Y_i}{\hat{h}_n(\mathbf{X}_i)}\right) + \sum_{\substack{i=1\\ i\neq j}}^{n}\sum_{j=1}^{n} Cov\left(\frac{Y_i\phi_{Jk}^{per}(X_{il})}{\hat{h}_n(\mathbf{X}_i)}\,,\,\frac{Y_j}{\hat{h}_n(\mathbf{X}_j)}\right)\right\}. \tag{B.20}$$

Case $i = j$

We have for $i = j$, $i = 1, ..., n$:

$$Cov\left(\frac{Y_i\phi_{Jk}^{per}(X_{il})}{\hat{h}_n(\mathbf{X}_i)}\,,\,\frac{Y_i}{\hat{h}_n(\mathbf{X}_i)}\right) \;=\; \mathbb{E}\left[\frac{Y^2\phi_{Jk}^{per}(X_l)}{\hat{h}_n(\mathbf{X})^2}\right] - \mathbb{E}\left[\frac{Y\phi_{Jk}^{per}(X_l)}{\hat{h}_n(\mathbf{X})}\right]\mathbb{E}\left[\frac{Y}{\hat{h}_n(\mathbf{X})}\right].$$

Using conditional expectation in the same way as in 3.16 and applying dominated convergence, it follows:

$$Cov\left(\frac{Y_i\phi_{Jk}^{per}(X_{il})}{\hat{h}_n(\mathbf{X}_i)}\,,\,\frac{Y_i}{\hat{h}_n(\mathbf{X}_i)}\right)\underset{n\to\infty}{\to}\mathbb{E}\left[\frac{Y^2\phi_{Jk}^{per}(X_l)}{h(\mathbf{X})^2}\right]-\beta_0\mathbb{E}\left[\frac{Y\phi_{Jk}^{per}(X_l)}{h(\mathbf{X})}\right]. \qquad \text{(B.21)}$$

Case $i \neq j$

For $i \neq j$, $i, j = 1, ..., n$, it is possible to obtain:

$$Cov\left(\frac{Y_i\phi_{Jk}^{per}(X_{il})}{\hat{h}_n(\mathbf{X}_i)}\,,\,\frac{Y_j}{\hat{h}_n(\mathbf{X}_j)}\right) \;=\; \mathbb{E}\left[\frac{Y_iY_j\phi_{Jk}^{per}(X_{il})}{\hat{h}_n(\mathbf{X}_i)\hat{h}_n(\mathbf{X}_j)}\right]-\mathbb{E}\left[\frac{Y\phi_{Jk}^{per}(X_l)}{\hat{h}_n(\mathbf{X})}\right]\mathbb{E}\left[\frac{Y}{\hat{h}_n(\mathbf{X})}\right].$$

From the definition of $\hat{h}_n(\mathbf{X})$ in (B.9), it follows:

$$\hat{h}_n(\mathbf{X}_i) \;=\; \frac{K(\mathbf{0})}{n\delta^p}+\frac{n-1}{n}\hat{h}_{n-1}(\mathbf{X}_i)\,,$$

therefore, for n sufficiently large:

$$\hat{h}_n(\mathbf{X}_i)\approx\hat{h}_{n-1}^{(-i)}(\mathbf{X}_i)\,,$$

provided $n\delta^p$ uniformly goes to ∞, where $\hat{h}_{n-1}^{(-i)}(\mathbf{X}_i)$ corresponds to the kernel density estimator computed without the $i-$th sample, evaluated at $\mathbf{X}_i$.

Let $\mathbf{X}^{(-i,-j)}$ denote the sample $\{\mathbf{X}_1, ..., \mathbf{X}_n\}$ without $\mathbf{X}_i, \mathbf{X}_j$. Therefore, using conditional expectation and for n sufficiently large:

$$\begin{aligned}
\mathbb{E}\left[\frac{Y_iY_j\phi_{Jk}^{per}(X_{il})}{\hat{h}_n(\mathbf{X}_i)\hat{h}_n(\mathbf{X}_j)}\right] &= \mathbb{E}_{\mathbf{X}_i,\mathbf{X}_j}\left[\mathbb{E}_{\mathbf{X}^{(-i,-j)}|\mathbf{X}_i,\mathbf{X}_j}\left[\frac{Y_iY_j\phi_{Jk}^{per}(X_{il})}{\hat{h}_n(\mathbf{X}_i)\hat{h}_n(\mathbf{X}_j)}|\mathbf{X}_i,\mathbf{X}_j\right]\right], \\
&= \mathbb{E}_{\mathbf{X}_i,\mathbf{X}_j}\left[Y_iY_j\phi_{Jk}^{per}(X_{il})\cdot\mathbb{E}_{\mathbf{X}^{(-i,-j)}|\mathbf{X}_i,\mathbf{X}_j}\left[\frac{1}{\hat{h}_n(\mathbf{X}_i)\hat{h}_n(\mathbf{X}_j)}|\mathbf{X}_i,\mathbf{X}_j\right]\right], \\
&\approx \mathbb{E}_{\mathbf{X}_i,\mathbf{X}_j}\left[Y_iY_j\phi_{Jk}^{per}(X_{il})\cdot\mathbb{E}_{\mathbf{X}^{(-i,-j)}|\mathbf{X}_i,\mathbf{X}_j}\left[\frac{1}{\hat{h}_{n-1}^{(-i)}(\mathbf{X}_i)\hat{h}_{n-1}^{(-j)}(\mathbf{X}_j)}|\mathbf{X}_i,\mathbf{X}_j\right]\right].
\end{aligned}$$

Using the last result and dominated convergence, it follows:

$$\begin{aligned}
\mathbb{E}\left[\frac{Y_iY_j\phi_{Jk}^{per}(X_{il})}{\hat{h}_n(\mathbf{X}_i)\hat{h}_n(\mathbf{X}_j)}\right] &\underset{n\to\infty}{\to} \mathbb{E}_{\mathbf{X}_i,\mathbf{X}_j}\left[\frac{Y_iY_j\phi_{Jk}^{per}(X_{il})}{h(\mathbf{X}_i)h(\mathbf{X}_j)}\right], \\
&\underset{n\to\infty}{\to} \beta_0\cdot\mathbb{E}\left[\frac{Y\phi_{Jk}^{per}(X_l)}{h(\mathbf{X})}\right], \qquad \text{(B.22)}
\end{aligned}$$

provided the iid condition of the observed sample. Finally,

$$\begin{aligned}
Cov\left(\frac{Y_i\phi_{Jk}^{per}(X_{il})}{\hat{h}_n(\mathbf{X}_i)}, \frac{Y_j}{\hat{h}_n(\mathbf{X}_j)}\right) &\underset{n\to\infty}{\to} \beta_0\cdot\mathbb{E}\left[\frac{Y\phi_{Jk}^{per}(X_l)}{h(\mathbf{X})}\right] - \beta_0\cdot\mathbb{E}\left[\frac{Y\phi_{Jk}^{per}(X_l)}{h(\mathbf{X})}\right], \\
&\underset{n\to\infty}{\to} 0\,. \qquad \text{(B.23)}
\end{aligned}$$

Therefore, using (B.21) and (B.23) in (B.20), it follows:

$$
\begin{aligned}
V_{c3} &= 2^{-\frac{J}{2}} \left\{ nCov\left(\frac{Y_i \phi_{Jk}^{per}(X_{il})}{\hat{h}_n(\mathbf{X}_i)} \,,\, \frac{Y_i}{\hat{h}_n(\mathbf{X}_i)} \right) + n(n-1)Cov\left(\frac{Y_i \phi_{Jk}^{per}(X_{il})}{\hat{h}_n(\mathbf{X}_i)} \,,\, \frac{Y_j}{\hat{h}_n(\mathbf{X}_j)} \right) \right\}, \\
&= 2^{-\frac{J}{2}} \left\{ \frac{1}{n} Cov\left(\frac{Y_i \phi_{Jk}^{per}(X_{il})}{\hat{h}_n(\mathbf{X}_i)} \,,\, \frac{Y_i}{\hat{h}_n(\mathbf{X}_i)} \right) + \frac{n(n-1)}{n^2} Cov\left(\frac{Y_i \phi_{Jk}^{per}(X_{il})}{\hat{h}_n(\mathbf{X}_i)} \,,\, \frac{Y_j}{\hat{h}_n(\mathbf{X}_j)} \right) \right\}.
\end{aligned}
$$

This last result implies:

$$
Cov\left(\frac{1}{n} \sum_{i=1}^{n} \frac{Y_i \phi_{Jk}^{per}(X_{il})}{\hat{h}_n(\mathbf{X}_i)} \,,\, 2^{-\frac{J}{2}} \hat{\beta}_0 \right) \underset{n\to\infty}{\to} 0 \,. \tag{B.24}
$$

As a corollary, we can see that from (B.24), it follows that $Cov\left(\hat{\beta}_0 \,,\, \hat{c}_{Jk}^{(l)}\right) \underset{n\to\infty}{\to} 0$. In fact, note that $Cov\left(\hat{\beta}_0 \,,\, \hat{c}_{Jk}^{(l)}\right)$ can be expressed as:

$$
Cov\left(\hat{\beta}_0 \,,\, \hat{c}_{Jk}^{(l)}\right) = Cov\left(\hat{\beta}_0 \,,\, \frac{1}{n} \sum_{i=1}^{n} \frac{Y_i \phi_{Jk}^{per}(X_{il})}{\hat{h}_n(\mathbf{X}_i)} \right) - 2^{-\frac{J}{2}} Var\left(\hat{\beta}_0\right) \,.
$$

Therefore, from (3.22) and (B.24), it is clear that $Cov\left(\hat{\beta}_0 \,,\, \hat{c}_{Jk}^{(l)}\right) \underset{n\to\infty}{\to} 0$ as desired.

Finally, this asertion also implies that:

$$
Cov\left(\hat{\beta}_0 \,,\, \sum_{l=1}^{p} \sum_{k=0}^{2^J-1} \hat{c}_{Jk}^{(l)} \phi_{Jk}^{per}(x_l) \right) \underset{n\to\infty}{\to} 0 \,, \tag{B.25}
$$

by the properties of the covariance function.

B.6 Asymptotic convergence of $Cov\left(\hat{c}_{Jk}^{(l)} \,,\, \hat{c}_{Js}^{(l)}\right)$.

For any $s \neq k$, $s, k = 0, ..., 2^J - 1$ and fixed J, assuming conditions (**A1**)-(**A5**) and (**Ak1**)-(**Ak6**) hold, it follows:

$$
\begin{aligned}
Cov\left(\hat{c}_{Jk}^{(l)}\,,\,\hat{c}_{Js}^{(l)}\right) &= Cov\left(\frac{1}{n}\sum_{i=1}^{n}\frac{Y_i\phi_{Jk}^{per}(X_{il})}{\hat{h}_n(\mathbf{X}_i)} - 2^{-\frac{J}{2}}\hat{\beta}_0\,,\,\frac{1}{n}\sum_{i=1}^{n}\frac{Y_i\phi_{Js}^{per}(X_{il})}{\hat{h}_n(\mathbf{X}_i)} - 2^{-\frac{J}{2}}\hat{\beta}_0\right), \\
&= \mathbb{E}\left[\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{Y_iY_j\phi_{Jk}^{per}(X_{il})\phi_{Js}^{per}(X_{jl})}{\hat{h}_n(\mathbf{X}_i)\hat{h}_n(\mathbf{X}_j)}\right] - 2^{-\frac{J}{2}}\mathbb{E}\left[\hat{\beta}_0\frac{1}{n}\sum_{i=1}^{n}\frac{Y_i\phi_{Jk}^{per}(X_{il})}{\hat{h}_n(\mathbf{X}_i)}\right] \\
&\quad -2^{-\frac{J}{2}}\mathbb{E}\left[\hat{\beta}_0\frac{1}{n}\sum_{i=1}^{n}\frac{Y_i\phi_{Js}^{per}(X_{il})}{\hat{h}_n(\mathbf{X}_i)}\right] + 2^{-J}\mathbb{E}\left[\hat{\beta}_0^2\right] - \mathbb{E}\left[\hat{c}_{Jk}^{(l)}\right]\mathbb{E}\left[\hat{c}_{Js}^{(l)}\right], \\
&= \mathbb{E}\left[\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{Y_iY_j\phi_{Jk}^{per}(X_{il})\phi_{Js}^{per}(X_{jl})}{\hat{h}_n(\mathbf{X}_i)\hat{h}_n(\mathbf{X}_j)}\right] - 2^{-\frac{J}{2}}Cov\left(\hat{\beta}_0\,,\,\frac{1}{n}\sum_{i=1}^{n}\frac{Y_i\phi_{Js}^{per}(X_{il})}{\hat{h}_n(\mathbf{X}_i)}\right) \\
&\quad -2^{-\frac{J}{2}}\mathbb{E}\left[\hat{\beta}_0\right]\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\frac{Y_i\phi_{Js}^{per}(X_{il})}{\hat{h}_n(\mathbf{X}_i)}\right] - 2^{-\frac{J}{2}}Cov\left(\hat{\beta}_0\,,\,\frac{1}{n}\sum_{i=1}^{n}\frac{Y_i\phi_{Jk}^{per}(X_{il})}{\hat{h}_n(\mathbf{X}_i)}\right) \\
&\quad -2^{-\frac{J}{2}}\mathbb{E}\left[\hat{\beta}_0\right]\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\frac{Y_i\phi_{Jk}^{per}(X_{il})}{\hat{h}_n(\mathbf{X}_i)}\right] + 2^{-J}\mathbb{E}\left[\hat{\beta}_0^2\right] - \mathbb{E}\left[\hat{c}_{Jk}^{(l)}\right]\mathbb{E}\left[\hat{c}_{Js}^{(l)}\right], \\
&= \mathbb{E}\left[\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{Y_iY_j\phi_{Jk}^{per}(X_{il})\phi_{Js}^{per}(X_{jl})}{\hat{h}_n(\mathbf{X}_i)\hat{h}_n(\mathbf{X}_j)}\right] - 2^{-\frac{J}{2}}Cov\left(\hat{\beta}_0\,,\,\frac{1}{n}\sum_{i=1}^{n}\frac{Y_i\phi_{Js}^{per}(X_{il})}{\hat{h}_n(\mathbf{X}_i)}\right) \\
&\quad 2^{-\frac{J}{2}}Cov\left(\hat{\beta}_0\,,\,\frac{1}{n}\sum_{i=1}^{n}\frac{Y_i\phi_{Jk}^{per}(X_{il})}{\hat{h}_n(\mathbf{X}_i)}\right) + 2^{-J}Var\left(\hat{\beta}_0\right) \\
&\quad -\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\frac{Y_i\phi_{Jk}^{per}(X_{il})}{\hat{h}_n(\mathbf{X}_i)}\right]\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\frac{Y_i\phi_{Js}^{per}(X_{il})}{\hat{h}_n(\mathbf{X}_i)}\right].
\end{aligned}
$$

Using the same argument that led to (B.22), for $i \neq j$, it follows:

$$
\mathbb{E}\left[\frac{Y_iY_j\phi_{Jk}^{per}(X_{il})\phi_{Js}^{per}(X_{jl})}{\hat{h}_n(\mathbf{X}_i)\hat{h}_n(\mathbf{X}_j)}\right] \underset{n\to\infty}{\to} \mathbb{E}\left[\frac{Y\phi_{Jk}^{per}(X_l)}{h(\mathbf{X})}\right]\mathbb{E}\left[\frac{Y\phi_{Js}^{per}(X_l)}{h(\mathbf{X})}\right].
$$

Similarly, for $i = j$:

$$
\mathbb{E}\left[\frac{Y_i^2\phi_{Jk}^{per}(X_{il})\phi_{Js}^{per}(X_{il})}{\hat{h}_n(\mathbf{X}_i)^2}\right] \underset{n\to\infty}{\to} \mathbb{E}\left[\frac{Y^2\phi_{Jk}^{per}(X_l)\phi_{Js}^{per}(X_l)}{h(\mathbf{X})^2}\right].
$$

Therefore, it follows that:

$$Cov\left(\hat{c}_{Jk}^{(l)}\,,\,\hat{c}_{Js}^{(l)}\right)\underset{n\to\infty}{\to}0\,,$$

as desired.

B.7 Proof of Lemma 3.2.4.

Let's assume conditions (**A1**)-(**A5**) and (**Ak1**)-(**Ak4**) are satisfied. For $i=1,...,n$, define:

$$\begin{aligned}
K_J(x,y) &= 2^J\sum_k \phi(2^Jx-k)\phi(2^Jy-k),\\
Z_i(\mathbf{x}) &= \frac{y_i}{\hat{h}_n(\mathbf{x}_i)}\left(\sum_{l=1}^p K_J(X_{il},x_l)\right)-\mathbb{E}\left[\frac{y_1}{\hat{h}_n(\mathbf{x}_1)}\left(\sum_{l=1}^p K_J(X_{1l},x_l)\right)\right].
\end{aligned}$$

Since $\mathbf{X}_1,...,\mathbf{X}_n$ are iid, $Z_i(\mathbf{x}),\ i=1,...,n$ are iid with $\mathbb{E}[Z_i(\mathbf{x})]=0$. From the definition of $\hat{f}_J(\mathbf{x})$ and $Z_i(\mathbf{x})$, after some algebra it is possible to get:

$$\begin{aligned}
\mathbb{E}\left[||\hat{f}_J(\mathbf{x})-\mathbb{E}[\hat{f}_J(\mathbf{x})]||_2^2\right] &\leq \mathbb{E}\left[\int_{[0,1]^p}\left\{\mid(\hat{\beta}_0-\mathbb{E}[\hat{\beta}_0])\left(1-2^{-\frac{J}{2}}\sum_{k=0}^{2^J-1}\sum_{l=1}^p\phi_{Jk}^{per}(x_l)\right)\mid\right\}\right]\\
&\quad+\mathbb{E}\left[\int_{[0,1]^p}\mid\frac{1}{n}\sum_{i=1}^n Z_i(\mathbf{x})\mid^2 d\mathbf{x}\right],\\
&\leq 2\mathbb{E}\left[(\hat{\beta}_0-\mathbb{E}[\hat{\beta}_0])^2\right]\int_{[0,1]^p}\left(1-2^{-\frac{J}{2}}\sum_{k=0}^{2^J-1}\sum_{l=1}^p\phi_{Jk}^{per}(x_l)\right)^2 d\mathbf{x}\\
&\quad+\frac{2}{n^2}\int_{[0,1]^p}\mathbb{E}\left[\mid\sum_{i=1}^n Z_i(\mathbf{x})\mid^2\right]d\mathbf{x}\,.
\end{aligned}$$

Denote:

$$\begin{aligned} S_{f1} &= \int_{[0,1]^p} \left(1 - 2^{-\frac{J}{2}} \sum_{k=0}^{2^J-1} \sum_{l=1}^{p} \phi_{Jk}^{per}(x_l)\right)^2 d\mathbf{x}\,, \\ S_{f2} &= \mathbb{E}\left[(\hat{\beta}_0 - \mathbb{E}[\hat{\beta}_0])^2\right] = Var\left(\hat{\beta}_0\right)\,, \\ S_{f3} &= \frac{2}{n^2} \int_{[0,1]^p} \mathbb{E}\left[\mid \sum_{i=1}^{n} Z_i(\mathbf{x}) \mid^2\right]. \end{aligned}$$

Computations for S_{f1}

Expanding the squared argument for S_{f1}, it follows:

$$\begin{aligned} S_{f1} &= \int_{[0,1]^p} \left(1 - 2^{1-\frac{J}{2}} \sum_{l=1}^{p} \sum_{k=0}^{2^J-1} \phi_{Jk}^{per}(x_l) + \sum_{l=1}^{p} \sum_{k_1=0}^{2^J-1} \sum_{m=1}^{p} \sum_{k_2=0}^{2^J-1} \phi_{Jk_1}^{per}(x_l)\phi_{Jk_2}^{per}(x_m)\right) d\mathbf{x}\,, \\ &= 1 - 2^{1-\frac{J}{2}} \sum_{l=1}^{p} \sum_{k=0}^{2^J-1} \int_0^1 \phi_{Jk}^{per}(x_l)dx_l + \sum_{l=1}^{p} \sum_{k_1=0}^{2^J-1} \sum_{m=1}^{p} \sum_{k_2=0}^{2^J-1} \int_0^1 \int_0^1 \phi_{Jk_1}^{per}(x_l)\phi_{Jk_2}^{per}(x_m)dx_l dx_m\,. \end{aligned}$$

Since $\int_0^1 \mid \phi_{Jk}^{per}(x_l) \mid dx_l \leq C_\phi 2^{-\frac{J}{2}}$ and $\left\{\phi_{J,k}^{per}(x), k = 0, ..., 2^J - 1\right\}$ are orthonormal, it follows:

$$S_{f1} = (p-1)^2 + p^2\left(2^J - 1\right) = \mathcal{O}\left(2^J\right)\,. \tag{B.26}$$

Computations for S_{f2}

Using the identity $Var(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$, since $\hat{\beta}_0 = \frac{1}{n}\sum_{i=1}^{n} \frac{y_i}{\hat{h}_n(\mathbf{x}_i)}$ it is possible to show:

$$
\begin{aligned}
\mathbb{E}\left[\hat{\beta}_0^2\right] &= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{Y_iY_j}{\hat{h}_n(\mathbf{X}_i)\hat{h}_n(\mathbf{X}_j)}\,, \\
&\leq \frac{(|\beta_0|+pM_f)^2+\sigma^2}{n^2}\sum_{i=1}^{n}\mathbb{E}\left[\frac{1}{\hat{h}_n(\mathbf{X}_i)^2}\right]+\frac{2}{n^2}\sum_{i<j}^{n}\mathbb{E}\left[\frac{Y_iY_j}{\hat{h}_n(\mathbf{X}_i)\hat{h}_n(\mathbf{X}_j)}\right], \\
&\leq \frac{(|\beta_0|+pM_f)^2+\sigma^2}{n}\mathbb{E}\left[\frac{1}{\hat{h}_n(\mathbf{X})^2}\right]+\frac{2}{n^2}\sum_{i<j}^{n}\mathbb{E}\left[\frac{Y_iY_j}{\hat{h}_n(\mathbf{X}_i)\hat{h}_n(\mathbf{X}_j)}\right], \\
&\leq \frac{(|\beta_0|+pM_f)^2+\sigma^2}{n}\mathbb{E}\left[\frac{1}{\hat{h}_n(\mathbf{X})^2}-\frac{1}{h(\mathbf{X})^2}\right]+\frac{(|\beta_0|+pM_f)^2+\sigma^2}{n}\mathbb{E}\left[\frac{1}{h(\mathbf{X})^2}\right] \\
&\quad +\frac{2}{n^2}\sum_{i<j}^{n}\mathbb{E}\left[\frac{Y_iY_j}{\hat{h}_n(\mathbf{X}_i)\hat{h}_n(\mathbf{X}_j)}\right].
\end{aligned}
$$

Now, since $|\mathbb{E}\left[\frac{1}{\hat{h}_n(\mathbf{X})^2}-\frac{1}{h(\mathbf{X})^2}\right]| \leq Cn^{-\frac{\beta}{2\beta+p}}$ and $h(\mathbf{x}) > \epsilon_h$, it follows:

$$
\mathbb{E}\left[\hat{\beta}_0^2\right] \leq C_1 n^{-\frac{3\beta+p}{2\beta+p}} + C_2 n^{-1} + \frac{2}{n^2}\sum_{i<j}^{n}\mathbb{E}\left[\frac{Y_iY_j}{\hat{h}_n(\mathbf{X}_i)\hat{h}_n(\mathbf{X}_j)}\right],
$$

for $C_1 = C\cdot(|\beta_0|+pM_f)^2+\sigma^2$ and $C_2 = \frac{\left(|\beta_0|+pM_f\right)^2+\sigma^2}{\epsilon_h^2}$.

Since $n\delta^p$ uniformly converges to ∞, $\hat{h}_n(\mathbf{X}_i) \approx \hat{h}_{n-1}^{(-i)}(\mathbf{X}_i)$, for n large. The notation $\approx$ means that the ratio between the lhs and the rhs terms goes to 1 as $n \to \infty$. Also, since we have an iid sample, it holds:

$$
\begin{aligned}
\mathbb{E}\left[\frac{Y_i Y_j}{\hat{h}_n(\mathbf{X}_i)\hat{h}_n(\mathbf{X}_j)}\right] &= \mathbb{E}_{\mathbf{X}_i,\mathbf{X}_j}\left[\mathbb{E}_{\mathbf{X}^{(-i,-j)}|\mathbf{X}_i,\mathbf{X}_j}\left(\frac{Y_i Y_j}{\hat{h}_{n-1}^{(-i)}(\mathbf{X}_i)\hat{h}_{n-1}^{(-j)}(\mathbf{X}_j)}|\mathbf{X}_i,\mathbf{X}_j\right)\right], \\
&= \mathbb{E}_{\mathbf{X}_i,\mathbf{X}_j}\left[\mathbb{E}_{\mathbf{X}^{(-i,-j)}}\left(\frac{Y_i}{\hat{h}_{n-1}^{(-i)}(\mathbf{X}_i)}|\mathbf{X}_i,\mathbf{X}_j\right)\mathbb{E}_{\mathbf{X}^{(-i,-j)}}\left(\frac{Y_j}{\hat{h}_{n-1}^{(-j)}(\mathbf{X}_j)}|\mathbf{X}_i,\mathbf{X}_j\right)\right], \\
&\approx \mathbb{E}_{\mathbf{X}_i,\mathbf{X}_j}\left[\mathbb{E}_{\mathbf{X}^{(-i,-j)}}\left(\frac{Y_i}{\hat{h}_n(\mathbf{X}_i)}\right)\mathbb{E}_{\mathbf{X}^{(-i,-j)}}\left(\frac{Y_j}{\hat{h}_n(\mathbf{X}_j)}\right)\right], \\
&\approx \left(\mathbb{E}\left[\frac{Y}{\hat{h}_n(\mathbf{X})}\right]\right)^2.
\end{aligned}
$$

This implies:

$$
\mathbb{E}\left[\hat{\beta}_0^2\right] \leq C^*\left(n^{-\frac{3\beta+p}{2\beta+p}} + n^{-1}\right) + \frac{n(n-1)}{n^2}\left(\mathbb{E}\left[\frac{Y}{\hat{h}_n(\mathbf{X})}\right]\right)^2, \tag{B.27}
$$

for some $C^* > \max\{C_1, C_2\} > 0$. Similarly, it follows:

$$
\begin{aligned}
\mathbb{E}\left[\hat{\beta}_0\right] &= \mathbb{E}\left[\frac{\beta_0 + \sum_{l=1}^p f_l(x_l) + \epsilon}{\hat{h}_n(\mathbf{X})}\right], \\
&= \mathbb{E}\left[\frac{Y}{\hat{h}_n(\mathbf{X})}\right].
\end{aligned}
$$

The last result, together with (B.27) imply:

$$\begin{aligned}
\mathbb{E}\left[\hat{\beta}_0^2\right] - \left(\mathbb{E}\left[\hat{\beta}_0\right]\right)^2 &\leq C^*\left(n^{-\frac{3\beta+p}{2\beta+p}} + n^{-1}\right) + \frac{n(n-1)}{n^2}\left(\mathbb{E}\left[\frac{Y}{\hat{h}_n(\mathbf{X})}\right]\right)^2 - \left(\mathbb{E}\left[\frac{Y}{\hat{h}_n(\mathbf{X})}\right]\right)^2, \\
&\leq C^*\left(n^{-\frac{3\beta+p}{2\beta+p}} + n^{-1}\right) - \frac{1}{n}\left(\mathbb{E}\left[\frac{Y}{\hat{h}_n(\mathbf{X})}\right]\right)^2, \\
&\leq C^*\left(n^{-\frac{3\beta+p}{2\beta+p}} + n^{-1}\right), \\
S_{f2} &= \mathcal{O}\left(n^{-1}\right). \qquad \text{(B.28)}
\end{aligned}$$

Thus, from (B.26) and (B.28), it follows that:

$$S_{f1}S_{f2} = \mathcal{O}\left(2^J n^{-1}\right).$$

Computations for S_{f3}

From Rosenthal's inequality, $\exists\, C(2) > 0$ such that:

$$\begin{aligned}
\frac{2}{n^2}\int_{[0,1]^p} \mathbb{E}\left[\left|\sum_{i=1}^n Z_i(\mathbf{x})\right|^2\right] &\leq \frac{4C(2)}{n^2}\int_{[0,1]^p}\sum_{i=1}^n \mathbb{E}\left[Z_i(\mathbf{x})^2\right] d\mathbf{x}, \\
&\leq \frac{4C(2)}{n^2}\sum_{i=1}^n \int_{[0,1]^p} \mathbb{E}\left[Z_i(\mathbf{x})^2\right] d\mathbf{x}.
\end{aligned}$$

By the definition of $Z_i(\mathbf{x})$, it follows:

$$\int_{[0,1]^p} \mathbb{E}\left[Z_i(\mathbf{x})^2\right] d\mathbf{x} \leq \sum_{l=1}^p \sum_{k_1=0}^{2^J-1}\sum_{m=1}^p\sum_{k_2=0}^{2^J-1} \mathbb{E}\left[\frac{Y_i^2\phi_{Jk_1}^{per}(X_{il})\phi_{Jk_2}^{per}(X_{im})}{\hat{h}_n(\mathbf{X}_i)^2}\right]\int_{[0,1]^p}\phi_{Jk_1}^{per}(x_l)\phi_{Jk_2}^{per}(x_m)d\mathbf{x}.$$

From the orthonormality of the scaling functions $\left\{\phi_{J,k}^{per}(x), k = 0, ..., 2^J - 1\right\}$ and (B.1), it follows:

$$\int_{[0,1]^p} \phi_{Jk_1}^{per}(x_l)\phi_{Jk_2}^{per}(x_m)d\mathbf{x} = \begin{cases} 1 & k_1 = k_2 \quad l = m \\ 0 & k_1 \neq k_2 \quad l = m \\ 2^{-J} & k_1 = k_2 \quad l \neq m \\ 2^{-J} & k_1 \neq k_2 \quad l \neq m \end{cases}$$

Therefore,

$$\begin{aligned} \int_{[0,1]^p} \mathbb{E}\left[Z_i(\mathbf{x})^2\right] d\mathbf{x} \quad \leq \quad & \sum_{l=1}^{p}\sum_{k=0}^{2^J-1}\left(\mathbb{E}\left[\frac{Y_i^2\phi_{Jk}^{per}(X_{il})^2}{\hat{h}_n(\mathbf{X}_i)^2}\right]\right) \\ & +2^{-J}\sum_{l\neq m}^{p}\sum_{k=0}^{2^J-1}\left(\mathbb{E}\left[\frac{Y_i^2\phi_{Jk}^{per}(X_{il})\phi_{Jk}^{per}(X_{im})}{\hat{h}_n(\mathbf{X}_i)^2}\right]\right) \\ & +2^{-J}\sum_{l\neq m}^{p}\sum_{k_1\neq k_2}^{2^J-1}\left(\mathbb{E}\left[\frac{Y_i^2\phi_{Jk_1}^{per}(X_{il})\phi_{Jk_2}^{per}(X_{im})}{\hat{h}_n(\mathbf{X}_i)^2}\right]\right). \end{aligned}$$

Since $\sup_{\mathbf{x}\in[0,1]^p}\{\beta_0 + \sum_{l=1}^{p} f_l(x_l)\} \leq (|\beta_0| + pM_f)$, we can show:

$$\begin{aligned} \mathbb{E}\left[\frac{Y_i^2\phi_{Jk}^{per}(X_{il})^2}{\hat{h}_n(\mathbf{X}_i)^2}\right] \quad & \leq \quad \left((|\beta_0| + pM_f)^2 + \sigma^2\right)\mathbb{E}\left[\frac{\phi_{Jk}^{per}(X_{il})^2}{\hat{h}_n(\mathbf{X}_i)^2}\right], \\ & \leq \quad C_1\mathbb{E}\left[\phi_{Jk}^{per}(X_{il})^2\left(\frac{1}{\hat{h}_n(\mathbf{X}_i)^2} - \frac{1}{h(\mathbf{X})^2}\right)\right], \\ & \leq \quad C_1 \cdot C \cdot n^{-\frac{\beta}{2\beta+p}}\mathbb{E}\left[\phi_{Jk}^{per}(X_{il})^2\right], \\ & \leq \quad C_1 \cdot C \cdot Mn^{-\frac{\beta}{2\beta+p}} \end{aligned} \tag{B.29}$$

for $C_1 = \left((|\beta_0| + pM_f)^2 + \sigma^2\right)$ and M as the upper bound of the density $h(\mathbf{x})$ from assumption (**A5**). Similarly, when $l \neq m$, it follows:

$$\begin{aligned}
\mathbb{E}\left[\frac{Y_i^2\phi_{Jk}^{per}(X_{il})\phi_{Jk}^{per}(X_{im})}{\hat{h}_n(\mathbf{X}_i)^2}\right] &\leq \left((|\beta_0| + pM_f)^2 + \sigma^2\right)\mathbb{E}\left[\frac{\phi_{Jk}^{per}(X_{il})\phi_{Jk}^{per}(X_{im})}{\hat{h}_n(\mathbf{X}_i)^2}\right], \\
&\leq C_1\mathbb{E}\left[\phi_{Jk}^{per}(X_{il})\phi_{Jk}^{per}(X_{im})\left(\frac{1}{\hat{h}_n(\mathbf{X}_i)^2} - \frac{1}{h(\mathbf{X})^2}\right) + \frac{\phi_{Jk}^{per}(X_{il})\phi_{Jk}^{per}(X_{im})}{h(\mathbf{X})^2}\right] \\
&\leq C_1 \cdot C \cdot n^{-\frac{\beta}{2\beta+p}}\mathbb{E}\left[\phi_{Jk}^{per}(X_{il})\phi_{Jk}^{per}(X_{im})\right] + \frac{C_1}{\epsilon_h^2}\mathbb{E}\left[\phi_{Jk}^{per}(X_{il})\phi_{Jk}^{per}(X_{im})\right], \\
&\leq C_1 \cdot C \cdot M2^{-J}n^{-\frac{\beta}{2\beta+p}} + \frac{C_1}{\epsilon_h^2}M2^{-J}.
\end{aligned}$$

In the case $k_1 \neq k_2 \quad l \neq m$, it is possible to show:

$$\begin{aligned}
\mathbb{E}\left[\frac{Y_i^2\phi_{Jk_1}^{per}(X_{il})\phi_{Jk_2}^{per}(X_{im})}{\hat{h}_n(\mathbf{X}_i)^2}\right] &\leq C_1 \cdot \mathbb{E}\left[\frac{\phi_{Jk_1}^{per}(X_{il})\phi_{Jk_2}^{per}(X_{im})}{\hat{h}_n(\mathbf{X}_i)^2}\right], \\
&\leq C_1\mathbb{E}\left[\phi_{Jk_1}^{per}(X_{il})\phi_{Jk_2}^{per}(X_{im})\left(\frac{1}{\hat{h}_n(\mathbf{X}_i)^2} - \frac{1}{h(\mathbf{X})^2}\right) + \frac{\phi_{Jk_1}^{per}(X_{il})\phi_{Jk_2}^{per}(X_{im})}{h(\mathbf{X})^2}\right] \\
&\leq C_1 \cdot C \cdot n^{-\frac{\beta}{2\beta+p}}\mathbb{E}\left[\phi_{Jk_1}^{per}(X_{il})\phi_{Jk_2}^{per}(X_{im})\right] + \frac{C_1}{\epsilon_h^2}\mathbb{E}\left[\phi_{Jk_1}^{per}(X_{il})\phi_{Jk_2}^{per}(X_{im})\right], \\
&\leq C_1 \cdot C \cdot M2^{-J}n^{-\frac{\beta}{2\beta+p}} + \frac{C_1}{\epsilon_h^2}M2^{-J}.
\end{aligned}$$

The last set of results imply:

$$\begin{aligned}
\int_{[0,1]^p} \mathbb{E}\left[Z_i(\mathbf{x})^2\right] d\mathbf{x} &\leq p \cdot 2^J \cdot C_1 \cdot C \cdot M n^{-\frac{\beta}{2\beta+p}} \\
&\quad + p(p-1)\left\{C_1 \cdot C \cdot M 2^{-J} n^{-\frac{\beta}{2\beta+p}} + \frac{C_1}{\epsilon_h^2} M 2^{-J}\right\} \\
&\quad + p(p-1)(2^J-1)\left\{C_1 \cdot C \cdot M 2^{-J} n^{-\frac{\beta}{2\beta+p}} + \frac{C_1}{\epsilon_h^2} M 2^{-J}\right\}, \\
&\leq p \cdot 2^J \cdot C_1 \cdot C \cdot M n^{-\frac{\beta}{2\beta+p}} + p(p-1)\left\{C_1 \cdot C \cdot M \cdot n^{-\frac{\beta}{2\beta+p}} + \frac{C_1}{\epsilon_h^2} M\right\}, \\
&\leq C^* \left(2^J n^{-\frac{\beta}{2\beta+p}} + n^{-\frac{\beta}{2\beta+p}} + 1\right),
\end{aligned}$$

for $C^* = \max\left\{p\, C_1\, C\, M\, ,\, p(p-1)\, C_1\, C\, M\, ,\, p(p-1)\frac{C_1}{\epsilon_h^2} M\right\} > 0$. Finally, we obtain:

$$\begin{aligned}
\frac{4C(2)}{n^2} \sum_{i=1}^{n} \int_{[0,1]^p} \mathbb{E}\left[Z_i(\mathbf{x})^2\right] d\mathbf{x} &\leq \frac{4C(2)}{n} C^* \left(2^J n^{-\frac{\beta}{2\beta+p}} + n^{-\frac{\beta}{2\beta+p}} + 1\right), \\
&\leq C^{**} \left(2^J n^{-\frac{3\beta+p}{2\beta+p}} + n^{-\frac{3\beta+p}{2\beta+p}} + n^{-1}\right), \\
S_{f3} &= \mathcal{O}\left(2^J n^{-\frac{3\beta+p}{2\beta+p}} + n^{-\frac{3\beta+p}{2\beta+p}} + n^{-1}\right), \qquad \text{(B.30)}
\end{aligned}$$

for $C^{**} = 4C(2)\, C^* > 0$.

Finally, from (B.26),(B.27) and (B.30), it follows:

$$\begin{aligned}
\mathbb{E}\left[||\hat{f}_J(\mathbf{x}) - \mathbb{E}[\hat{f}_J(\mathbf{x})]||_2^2\right] &\leq \mathcal{O}\left(2^J n^{-1}\right) + \mathcal{O}\left(2^J n^{-\frac{3\beta+p}{2\beta+p}} + n^{-1}\right) \\
&\leq \mathcal{O}\left(2^J n^{-1}\right). \qquad \text{(B.31)}
\end{aligned}$$

which completes the proof.

B.8 Proof of Lemma 3.2.5.

Suppose that in addition to assumptions (**A1**)-(**A5**) and (**Ak1**)-(**Ak4**), the following conditions are satisfied:

(a) $\exists\, \Phi$, bounded and non-increasing function in $\mathbb{R}$ such that $\int \Phi(|u|)du < \infty$ and $|\phi(u)| \leq \Phi(|u|)$ almost everywhere (a.e.).

(b) In addition, $\int_{\mathbb{R}} |u|^{N+1}\Phi(|u|)du < \infty$ for some $N \geq 0$.

(c) $\exists\, F$, integrable, such that $|K(x,y)| \leq F(x-y)$, $\forall x, y \in \mathbb{R}$.

(d) Suppose ϕ satisfies:

i. $\sum_k |\hat{\phi}(\xi + 2k\pi)|^2 = 1$, a.e., where $\hat{\phi}$ denotes the Fourier transform of the scaling function ϕ.

ii. $\hat{\phi}(\xi) = \hat{\phi}(\frac{\xi}{2})m_0(\frac{\xi}{2})$, where $m_0(\xi)$ is a 2π-periodic function and $m_0 \in \mathbb{L}_2(0, 2\pi)$.

(e) $\int_{\mathbb{R}} x^k \psi(x)dx = 0$, for $k = 0, 1, ..., N$, $N \geq 1$ where ψ is the mother wavelet corresponding to ϕ.

(f) The functions $\{f_l\}_{l=1}^p$, are such that $f_l \in W_\infty^{m+1}([0,1])$, $m \geq N$, where $W_\infty^m([0,1])$ denotes the space of functions that are m-times weakly-differentiable and $f_l^{(k)} \in L_\infty([0,1])$, $k = 0, ..., m$.

(g) $\theta_\phi(x) := \sum_k |\phi(x-k)|$ such that $||\theta_\phi||_\infty < \infty$. Under this condition, it follows that $f_l^{(k)} \in L_p([0,1])$, $k = 0, ..., m$ for $p \geq 1$.

Then under Corollary 8.2 [57], if $f \in W_\infty^{N+1}([0,1])$ then $||K_J f - f||_\infty^p = \mathcal{O}\left(2^{-pJ(N+1)}\right)$, $p \geq 1$. This implies:

$$||\mathbb{E}[\hat{f}_J(\mathbf{x})] - f(\mathbf{x})||_2^2 = \mathcal{O}\left(2^{2J} n^{-\frac{2\beta}{2\beta+p}} + 2^{-2J(N+1)} + n^{-\frac{\beta}{2\beta+p}} 2^{-J(N+1)}\right), \tag{B.32}$$

for $f(\mathbf{x}) = \beta_0 + \sum_{l=1}^{p} f_l(x_l)$.

Proof. Define $f_{lJ}(x_l) := K_J f_l(x_l) = \int_0^1 f_l(u) K_J(x_l, u)\, du$. Suppose a fixed $\mathbf{x}$, then:

$$\mathbb{E}[\hat{f}_J(\mathbf{x})] - f(\mathbf{x}) \;=\; bias\left(\hat{\beta}_0\right) + \sum_{l=1}^{p}\sum_{k=0}^{2^J-1} bias\left(\hat{c}_{Jk}^{(l)}\right)\phi_{Jk}^{per}(x_l) + \sum_{l=1}^{p}\left(f_{lJ}(x_l) - f_l(x_l)\right) .$$

Furthermore, since $\mathbb{E}\left[\frac{\sum_{l=1}^{p} f_l(X_l)}{h(\mathbf{X})}\right] = 0$, it follows:

$$\begin{aligned} bias\left(\hat{\beta}_0\right) &\leq |\beta_0| C n^{-\frac{\beta}{2\beta+p}} + \mathbb{E}_{\mathbf{X}}\left[\sum_{l=1}^{p} f_l(x_l)\mathbb{E}_{\mathbf{X}_1,\ldots,\mathbf{X}_n}\left(\frac{1}{\hat{h}_n(\mathbf{X})} - \frac{1}{h(\mathbf{X})}\right)\right], \\ &\leq (|\beta_0| + pM_f)\, C n^{-\frac{\beta}{2\beta+p}} . \end{aligned}$$

Similarly, following the same argument for $bias\left(\hat{c}_{Jk}^{(l)}\right)$, it is possible to show:

$$bias\left(\hat{c}_{Jk}^{(l)}\right) \;\leq\; (|\beta_0| + pM_f)\, C 2^{-\frac{J}{2}} n^{-\frac{\beta}{2\beta+p}} .$$

Therefore, this implies:

$$
\begin{aligned}
\mathbb{E}[\hat{f}_J(\mathbf{x})] - f(\mathbf{x}) &\leq C_1^* n^{-\frac{\beta}{2\beta+p}} + C_1^* 2^{-\frac{J}{2}} n^{-\frac{\beta}{2\beta+p}} \sum_{l=1}^{p} \sum_{k=0}^{2^J-1} |\phi_{Jk}^{per}(x_l)| + \sum_{l=1}^{p} |K_J f_l(x_l) - f_l(x_l)| \ , \\
\left(\mathbb{E}[\hat{f}_J(\mathbf{x})] - f(\mathbf{x})\right)^2 &\leq C_1^{**} n^{-\frac{2\beta}{2\beta+p}} + C_1^{**} 2^{-J} n^{-\frac{2\beta}{2\beta+p}} \sum_{l=1}^{p} \sum_{k_1=0}^{2^J-1} \sum_{m=1}^{p} \sum_{k_2=0}^{2^J-1} \left|\phi_{Jk_1}^{per}(x_l)\right| \left|\phi_{Jk_2}^{per}(x_m)\right| \\
&\quad + \left(\sum_{l=1}^{p} |K_J f_l(x_l) - f_l(x_l)| \right)^2 + 2C_1^* 2^{-\frac{J}{2}} n^{-\frac{2\beta}{2\beta+p}} \sum_{l=1}^{p} \sum_{k=0}^{2^J-1} |\phi_{Jk}^{per}(x_l)| \\
&\quad + 2C_1^* n^{-\frac{\beta}{2\beta+p}} \sum_{l=1}^{p} |K_J f_l(x_l) - f_l(x_l)| \\
&\quad + 2C_1^* 2^{-\frac{J}{2}} n^{-\frac{\beta}{2\beta+p}} \sum_{l=1}^{p} \sum_{k=0}^{2^J-1} \sum_{m=1}^{p} |\phi_{Jk}^{per}(x_l)| \, |K_J f_m(x_m) - f_m(x_m)| \ ,
\end{aligned}
$$

where $C_1^* = (|\beta_0| + pM_f)\, C$ and $C_1^{**} = (|\beta_0| + pM_f)^2\, C^2$ are positive constants independent of J and n. Furthermore, since $\int_0^1 |\phi_{Jk}^{per}(u)|\, du = 2^{-\frac{J}{2}} C_\phi$, C_ϕ, it follows:

$$
\int_{[0,1]^p} |\phi_{Jk_1}^{per}(x_l)||\phi_{Jk_2}^{per}(x_m)| d\mathbf{x} = \begin{cases} 1 & k_1 = k_2 \quad l = m \\ 2^J ||\theta_\phi||_\infty^2 & k_1 \neq k_2 \quad l = m \\ 2^{-J} C_\phi^2 & k_1 = k_2 \quad l \neq m \\ 2^{-J} C_\phi^2 & k_1 \neq k_2 \quad l \neq m \end{cases}
$$

Using the last set of equations, we obtain:

$$\begin{aligned}\int_{[0,1]^p}\left(\mathbb{E}[\hat{f}_J(\mathbf{x})]-f(\mathbf{x})\right)^2 d\mathbf{x} \;\leq\; & C_1^{**}n^{-\frac{2\beta}{2\beta+p}}+pC_1^{**}n^{-\frac{2\beta}{2\beta+p}}+pC_1^{**}n^{-\frac{2\beta}{2\beta+p}}2^J(2^J-1)||\theta_\phi||_\infty^2\\ &+C_\phi^2C_1^{**}n^{-\frac{2\beta}{2\beta+p}}p(p-1)+\left\|\sum_{l=1}^p|K_Jf_l(x_l)-f_l(x_l)|\right\|_2^2\\ &+2pC_\phi C_1^*n^{-\frac{2\beta}{2\beta+p}}+2C_1^*n^{-\frac{\beta}{2\beta+p}}\left\|\sum_{l=1}^p|K_Jf_l(x_l)-f_l(x_l)|\right\|_1\\ &+2C_1^*2^{-\frac{J}{2}}n^{-\frac{\beta}{2\beta+p}}\sum_{l=1}^p\sum_{k=0}^{2^J-1}\sum_{m=1}^p\int_{[0,1]^p}|\phi_{Jk}^{per}(x_l)|\,|K_Jf_m(x_m)-f_m(x_m)|\,dx\end{aligned}$$

Using the properties of L_p norms and Corollary 8.2 [57], it follows:

$$\begin{aligned}\left\|\sum_{l=1}^p|K_Jf_l(x_l)-f_l(x_l)|\right\|_2^2 &\leq 2\sum_{l=1}^p\|K_Jf_l(x_l)-f_l(x_l)\|_2^2\leq C_*2^{-2J(N+1)}\,,\\ \left\|\sum_{l=1}^p|K_Jf_l(x_l)-f_l(x_l)|\right\|_1 &\leq \sum_{l=1}^p\|K_Jf_l(x_l)-f_l(x_l)\|_1\leq C_{**}2^{-J(N+1)}\,.\end{aligned}$$

Therefore, this implies:

$$
\begin{aligned}
\int_{[0,1]^p} \left(\mathbb{E}[\hat{f}_J(\mathbf{x})] - f(\mathbf{x})\right)^2 d\mathbf{x} \;\leq\; & C_1^{**} n^{-\frac{2\beta}{2\beta+p}} + pC_1^{**} n^{-\frac{2\beta}{2\beta+p}} + pC_1^{**} n^{-\frac{2\beta}{2\beta+p}} 2^J (2^J - 1) ||\theta_\phi||_\infty^2 \\
& + C_\phi^2 C_1^{**} n^{-\frac{2\beta}{2\beta+p}} p(p-1) + C_* 2^{-2J(N+1)} + 2pC_\phi C_1^{**} n^{-\frac{2\beta}{2\beta+p}} \\
& + 2C_1^* n^{-\frac{\beta}{2\beta+p}} C_{**} 2^{-J(N+1)} \\
& + 2C_1^* 2^{-\frac{J}{2}} n^{-\frac{\beta}{2\beta+p}} \sum_{l=1}^{p} \sum_{k=0}^{2^J-1} \int_{[0,1]^p} |\phi_{Jk}^{per}(x_l)| \, |K_J f_l(x_l) - f_l(x_l)| \, d\mathbf{x} \\
& + 2C_1^* 2^{-\frac{J}{2}} n^{-\frac{\beta}{2\beta+p}} \sum_{l \neq m}^{p} \sum_{k=0}^{2^J-1} \int_{[0,1]^p} |\phi_{Jk}^{per}(x_l)| \, |K_J f_m(x_m) - f_m(x_m)| \, d\mathbf{x} \,, \\
\leq \; & C^{***} \left\{ n^{-\frac{2\beta}{2\beta+p}} + 2^{2J} n^{-\frac{2\beta}{2\beta+p}} + 2^{-2J(N+1)} + n^{-\frac{2\beta}{2\beta+p}} 2^{-J(N+1)} \right\} \\
& + C^{***} \left\{ 2^{-\frac{J}{2}} n^{-\frac{\beta}{2\beta+p}} \sum_{l=1}^{p} \sum_{k=0}^{2^J-1} \int_{[0,1]^p} |\phi_{Jk}^{per}(x_l)| \, |K_J f_l(x_l) - f_l(x_l)| \, d\mathbf{x} \right\} \\
& + C^{***} \left\{ 2^{-\frac{J}{2}} n^{-\frac{\beta}{2\beta+p}} \sum_{l \neq m}^{p} \sum_{k=0}^{2^J-1} \int_{[0,1]^p} |\phi_{Jk}^{per}(x_l)| \, |K_J f_m(x_m) - f_m(x_m)| \, d\mathbf{x} \right.
\end{aligned}
$$

for $C^{***} = \max\left\{p\, C_1^{**} \,,\, p\, C_1^{**} ||\theta_\phi||_\infty^2 \,,\, 2pC_\phi^2 C_1^{**} \,,\, C_* \,,\, 2C_1^* C_{**}\right\} > 0$, independent of J and n.

Assumption 7 and Corollary 8.2 [57] imply:

$$
\int_{[0,1]^p} |\phi_{Jk}^{per}(x_l)| \, |K_J f_m(x_m) - f_m(x_m)| \, d\mathbf{x} \leq \begin{cases} C 2^{-\frac{J}{2}} ||\theta_\phi||_\infty 2^{-J(N+1)} & l = m \\ C \cdot C_\phi 2^{-\frac{J}{2}} 2^{-J(N+1)} & l \neq m \end{cases}
$$

Therefore:

$$\begin{aligned}\int_{[0,1]^p}\left(\mathbb{E}[\hat{f}_J(\mathbf{x})]-f(\mathbf{x})\right)^2 d\mathbf{x} &\leq \tilde{C}^{***}\left\{n^{-\frac{2\beta}{2\beta+p}}+2^{2J}n^{-\frac{2\beta}{2\beta+p}}+2^{-2J(N+1)}+n^{-\frac{2\beta}{2\beta+p}}2^{-J(N+1)}+n^{-\frac{\beta}{2\beta+p}}2^{-}\right.\\ &\leq \tilde{C}^{***}\left\{2^{2J}n^{-\frac{2\beta}{2\beta+p}}+2^{-2J(N+1)}+n^{-\frac{\beta}{2\beta+p}}2^{-J(N+1)}\right\},\end{aligned}$$

for $\tilde{C}^{***}=\max\left\{C^{***},\, C||\theta_\phi||_\infty,\, C\cdot C_\phi\right\}>0$. Thus,

$$\left\|\mathbb{E}[\hat{f}_J(\mathbf{x})]-f(\mathbf{x})\right\|_2^2=\mathcal{O}\left(2^{2J}n^{-\frac{2\beta}{2\beta+p}}+2^{-2J(N+1)}+n^{-\frac{\beta}{2\beta+p}}2^{-J(N+1)}\right), \tag{B.33}$$

which completes the proof. □

Remarks

Note that assumptions ii(d)i and ii(d)ii are automatically satisfied by choosing the orthonormal basis $\left\{\phi_{J,k}^{per}(x), k=0,...,2^J-1\right\}$. These are explicitly stated to be consistent with results presented in [57] that were used to obtain the estimator approximation properties.

B.9 Proof of Lemma 3.2.6.

Define $\mathcal{F}=\left\{f \mid f_l \in L_2([0,1]),\, f_l \in W_2^{N+1}([0,1]),\, -\infty<m_l\leq f_l\leq M_l<\infty,\, l=1,...,p\right\}$ where $W_2^{N+1}([0,1])$ is the space of functions that are $N+1$-times differentiable, and $f^{(k)}(x)\in L_2([0,1]),\, k=0,...,N+1$. For $f(\mathbf{x})=\beta_0+\sum_{l=1}^p f_l(x_l)$ consider that assumptions 1-7 from Proposition 6 and conditions (**A1**)-(**A5**), and (**Ak1**)-(**Ak4**) are satisfied. Then:

$$\sup_{f\in\mathcal{F}}\left(\mathbb{E}\left[||\hat{f}_J(\mathbf{x})-f(\mathbf{x})||_2^2\right]\right)\leq \tilde{C}n^{-\left(\frac{\beta}{2\beta+p}\right)\left(\frac{N+1}{N+3}\right)}, \tag{B.34}$$

provided (3.41) and (3.42), for $J=J(n)$ such that $2^{J(n)}\simeq n^{\frac{2\beta}{(2\beta+p)(N+3)}}$.

Proof. For $C > 0$ sufficiently large it follows:

$$\mathbb{E}||\hat{f}_J(\mathbf{x}) - f(\mathbf{x})||_2^2 \leq C\left(2^{2J}n^{-\frac{2\beta}{2\beta+p}} + 2^{-2J(N+1)} + n^{-\frac{\beta}{2\beta+p}}2^{-J(N+1)}\right),$$

from (B.31) and (B.33).

The last result implies that it is possible to choose $J = J(n)$ such that the upper bound of the Risk is minimized. Consequently, (ignoring constants) it is possible to show that $2^{J(n)} \simeq n^{\frac{2\beta}{(2\beta+p)(N+3)}}$ provides such optimal result. Moreover, since the upper bound is valid $\forall f \in \mathcal{F}$:

$$\sup_{f\in\mathcal{F}}\left(\mathbb{E}\left[||\hat{f}_J(\mathbf{x}) - f(\mathbf{x})||_2^2\right]\right) \leq \tilde{C}n^{-\left(\frac{2\beta}{2\beta+p}\right)\left(\frac{N+1}{N+3}\right)} \tag{B.35}$$

, □

which completes the proof.

Under the optimal choice of $J(n)$, it follows:

$$\mathbb{E}\left[\left\|\hat{f}_J(\mathbf{x}) - \mathbb{E}[\hat{f}_J(\mathbf{x})]\right\|_2^2\right] = \mathcal{O}\left(n^{-\left(\frac{N+2}{N+3}\right)}n^{-\left(\frac{p}{2\beta+p}\right)}\right), \tag{B.36}$$

$$\left\|\mathbb{E}[\hat{f}_J(\mathbf{x})] - f(\mathbf{x})\right\|_2^2 = \mathcal{O}\left(n^{-\left(\frac{2\beta}{2\beta+p}\right)\left(\frac{N+1}{N+3}\right)}\right). \tag{B.37}$$

As can be observed in (B.36) and (B.37), the variance term of the estimator $\hat{f}_J(\mathbf{x})$ is influenced primarily by the properties of the functional space that contains $\{f_l(x)\,,\, l = 1, ..., p\}$ and the wavelet basis $\left\{\phi_{J,k}^{per}(x), k = 0, ..., 2^J - 1\right\}$. Similarly, for n sufficiently large, the bias effect dominates in the risk decomposition and is responsible for the average approximation error of the estimator.

APPENDIX C

APPENDIX CHAPTER 4

C.1 Previous Theorems and definitions

In this section, we provide important definitions and results previously published that are used to derive the theoretical properties of the proposed estimators.

C.1.1 Theorem P1 (Pollard 1984)

Consider a class of functions $\mathcal{G} = \{g\,,\, g : \mathbb{R}^p \to [0, B]\}$, then for any $n \in \mathbb{N}$ and any $\epsilon > 0$:

$$\mathbb{P}\left\{\sup_{g\in\mathcal{G}}\left|\frac{1}{n}\sum_{i=1}^{n} g(\mathbf{Z}_i) - \mathbb{E}[g(\mathbf{Z})]\right| > \epsilon\right\} \leq 8 \cdot \mathbb{E}\left[\mathcal{N}_1\left(\frac{\epsilon}{8}, \mathcal{G}, \mathbf{z}_1^n\right)\right] \cdot e^{-\frac{n\cdot\epsilon^2}{128B^2}}\,, \tag{C.1}$$

where $B < \infty$ (i.e. the functions g are uniformly bounded over the class $\mathcal{G}$), $\{\mathbf{Z}, \mathbf{Z}_i\}_{i=1}^n$ is an iid sample of random variables in $\mathbb{R}^p$, $\mathcal{N}_1\left(\frac{\epsilon}{8}, \mathcal{G}, \mathbf{z}_1^n\right)$ is the $\mathbb{L}_1$ $\frac{\epsilon}{8}$-covering number of $\mathcal{G}$ on $\mathbf{z}_1^n = \{\mathbf{Z}_i\}_{i=1}^n$. This is the smallest $N \in \mathbb{N}$ such that for every function $g \in \mathcal{G}$ and a given probability measure μ on $\mathbb{R}^p$ and $s \geq 1$ there exists a $j = j(g) \in \{1, ..., N\}$ for which $||g - g_j||_{\mathbb{L}_1(\mu)} < \epsilon$, for $||g||_{\mathbb{L}_1(\mu)} := \left(\int |f(z)|d\mu_n\right) = \left(\frac{1}{n}\sum_{i=1}^n |g(\mathbf{z}_i) - g_j(\mathbf{z}_i)|^s\right)^{\frac{1}{s}}$.

A detailed proof of this theorem and a illustrative discussion about covering numbers can be found in [91] and [9].

C.1.2 Lemma G1 (Györfi et al. 2002)

Consider a probability measure μ on $\mathbb{R}^p$, $s \geq 1$, $\epsilon > 0$ and a class of functions $\mathcal{G}$ on $\mathbb{R}^p$. Then:

$$\mathcal{M}\left(2\epsilon, \mathcal{G}, ||\cdot||_{\mathbb{L}_s(\mu)}\right) \leq \mathcal{N}\left(\epsilon, \mathcal{G}, ||\cdot||_{\mathbb{L}_s(\mu)}\right) \leq \mathcal{M}\left(\epsilon, \mathcal{G}, ||\cdot||_{\mathbb{L}_s(\mu)}\right) . \tag{C.2}$$

Here, $\mathcal{M}\left(\epsilon, \mathcal{G}, ||\cdot||_{\mathbb{L}_s(\mu)}\right)$ represents the size of the largest ϵ-packing of $\mathcal{G}$ with respect to $||\cdot||_{\mathbb{L}_s(\mu)}$. This is the largest $N \in \mathbb{N}$ such that the collection of functions $\{g_1, ..., g_N\} \in \mathcal{G}$ satisfy $||g_j - g_l||_{\mathbb{L}_s(\mu)} \geq \epsilon$, for $||g||_{\mathbb{L}_s(\mu)} := \left(\int |f(z)|^s d\mu\right)^{\frac{1}{s}}$.

A detailed proof of this Lemma, together with definitions and details about covering and packing numbers can be found in section 9 of [9].

C.1.3 Theorem G2 (Györfi et al. 2002)

Before stating this theorem, consider the following definitions:

Definitions G2.1 Consider a class of subsets of $\mathbb{R}^p$ denoted by $\mathcal{A}$. Let $n \in \mathbb{N}$. Then,

(i) For a sample $\mathbf{z}_1, ..., \mathbf{z}_n \in \mathbb{R}^p$, define $s\left(\mathcal{A}, \{\mathbf{z}_1, ..., \mathbf{z}_n\}\right)$ as the number of different subsets of $\{\mathbf{z}_1, ..., \mathbf{z}_n\}$ that can be expressed as sets of the form $A \cap \{\mathbf{z}_1, ..., \mathbf{z}_n\}$ for $A \in \mathcal{A}$. This is $s\left(\mathcal{A}, \{\mathbf{z}_1, ..., \mathbf{z}_n\}\right) = |A \cap \{\mathbf{z}_1, ..., \mathbf{z}_n\} : A \in \mathcal{A}|$.

(ii) If for a set $H \subseteq \mathbb{R}^p$ $s\left(\mathcal{A}, H\right) = 2^n$ (i.e. every subset of H can be represented as $A \cap H$ for $A \in \mathcal{A}$), then we say that $\mathcal{A}$ shatters H.

(iii) The n-th shatter coefficient of $\mathcal{A}$ given a sample containing n points is the maximal number of different subsets of the n points that are contained by sets in $\mathcal{A}$, therefore,

they can be represented as $A \cap H$ for $A \in \mathcal{A}$. We denote the n-th shatter coefficient of $\mathcal{A}$ as $S(\mathcal{A}, n)$. Note that for all $n > k$ we have that $S(\mathcal{A}, k) < 2^k$ implies $S(\mathcal{A}, n) < 2^n$.

(iv) Suppose that $\mathcal{A} \subseteq \mathbb{R}^p \neq \emptyset$, the VC dimension (Vapnis-Chervonenkis dimension) $V_{\mathcal{A}}$ of $\mathcal{A}$ corresponds to the largest integer n such that there exists a set of n points in $\mathbb{R}^p$ that can be shattered by $\mathcal{A}$. This is $V_{\mathcal{A}} = \sup \{n \in \mathbb{N} : \ S(\mathcal{A}, n) = 2^n\}$.

(v) Suppose $\mathcal{G}$ is a class of functions in $\mathbb{R}^p$ such that $\forall g \in \mathcal{G}\,, g : \mathbb{R}^p \to [0, B]$. Let's define the set $\mathcal{G}^+ := \{(\mathbf{z}, t) \in \mathbb{R}^p \times \mathbb{R}\,; t \leq (\mathbf{z})\,; g \in \mathcal{G}\}$. This set corresponds to the set of all sub-graphs of the functions contained in the set $\mathcal{G}$.

Now, consider a class of functions $\mathcal{G}$ in $\mathbb{R}^p$ such that $\forall g \in \mathcal{G}\,, g : \mathbb{R}^p \to [0, B]$ with $V_{\mathcal{G}^+} \geq 2$. Let $s \geq 1$ and μ a probability measure on $\mathbb{R}^p$ and let $0 < \epsilon < \frac{B}{4}$; then:

$$\mathcal{M}\left(\epsilon, \mathcal{G}, ||\cdot||_{\mathbb{L}_s(\mu)}\right) \leq 3\left(\frac{2eB^s}{\epsilon^s}\log\left(\frac{3eB^s}{\epsilon^s}\right)\right)^{V_{\mathcal{G}^+}}. \tag{C.3}$$

A detailed proof of this Theorem, together with definitions and details about shattering numbers and VC dimension can be found in section 9 of [9].

C.1.4 Theorem G3 (Györfi et al. 2002)

This theorem provides an upper bound on the VC dimension for r-dimensional vector spaces. Consider $\mathcal{G}$ to be a r-dimensional vector space of real functions defined on $\mathbb{R}^p$. Let $\mathcal{A} = \{\mathbf{z} : g(\mathbf{x}) \geq 0 : g \in \mathcal{G}\}$. Then:

$$V_{\mathcal{A}} \leq r\,. \tag{C.4}$$

A detailed proof of this Theorem can be found in section 9.4 of [9].

C.1.5 Theorem G4 (Györfi et al. 2002)

This theorem provides necessary and sufficient conditions for the consistency of least squares estimators. Consider $\mathcal{F}_n = \mathcal{F}_n\left(\{(Y_i, \mathbf{X}_i\}_{i=1}^n\right)$ a class of functions $f : \mathbb{R}^p \to \mathbb{R}$. Let β_n be a parameter depending on the sample size n such that $\beta_n \to \infty$ as $n \to \infty$. Let $\hat{f}_{J(n)}$ be defined as in (4.12) and $f_{J(n)} = T_{\beta_n}\hat{f}_{J(n)}$ (i.e. the truncated version of $\hat{f}_{J(n)}$) and μ be a Lebesgue measure in $\mathbb{R}^p$; Then :

(i) If for all $L > 0$ the following conditions hold:

$$\lim_{n\to\infty} \inf_{f\in\mathcal{F}_n : ||f||_\infty \le \beta_n} \int |f(\mathbf{x}) - f_A(\mathbf{x})|^2 \, \mu(d\mathbf{x}) \;=\; 0 \;(a.s.)\,, \quad \text{(C.5)}$$

$$\lim_{n\to\infty} \sup_{f\in T_{\beta_n}\mathcal{F}_n} \left| \frac{1}{n}\sum_{i=1}^n |f(\mathbf{X}_i) - Y_{i,L}|^2 - \mathbb{E}\left[(f(\mathbf{X}) - Y_L)^2\right] \right| \;=\; 0 \;(a.s.)\,, \quad \text{(C.6)}$$

then:

$$\lim_{n\to\infty} \int \left|f_{J(n)}(\mathbf{x}) - f_A(\mathbf{x})\right|^2 \mu(d\mathbf{x}) = 0, \text{ almost surely (a.s.).}$$

Here, $Y_L = T_L Y = \left\{ \begin{array}{ll} Y & |Y| \le \beta_n \\ \beta_n \cdot \text{sign}(Y) & |Y| > \beta_n \end{array} \right\}$.

(ii) If for all $L > 0$ the following conditions hold:

$$\lim_{n\to\infty} \mathbb{E}\left\{ \inf_{f\in\mathcal{F}_n : ||f||_\infty \le \beta_n} \int |f(\mathbf{x}) - f_A(\mathbf{x})|^2 \, \mu(d\mathbf{x}) \right\} \;=\; 0\,, \quad \text{(C.7)}$$

$$\lim_{n\to\infty} \mathbb{E}\left\{ \sup_{f\in T_{\beta_n}\mathcal{F}_n} \left| \frac{1}{n}\sum_{i=1}^n |f(\mathbf{X}_i) - Y_{i,L}|^2 - \mathbb{E}\left[(f(\mathbf{X}) - Y_L)^2\right] \right| \right\} \;=\; 0\,, \quad \text{(C.8)}$$

then:

$$\lim_{n\to\infty} \mathbb{E}\left\{ \int \left|f_{J(n)}(\mathbf{x}) - f_A(\mathbf{x})\right|^2 \mu(d\mathbf{x}) \right\} = 0.$$

A detailed proof of this Theorem can be found in section 10.1 of [9].

This theorem shows that strong consistency is achieved for any least squares estimator obtained over a data-dependent class of functions $\mathcal{F}_n$, truncated by a suitable parameter β_n that depends on the sample size and converges to ∞, and provided that the approximation error (C.5) converges to zero a.s. (i.e. for every $\omega \in \Omega$ such that $\mathbb{P}(\omega) \neq 0$, $f_n(\omega) \to f_A$ with probability 1), and that the empirical $\mathbb{L}_2$ norm uniformly converges to the $\mathbb{L}_2(\mu)$ norm over the set of functions $T_{\beta_n}\mathcal{F}_n$.

C.1.6 Theorem P2 (Pollard 1984)

Suppose $\mathcal{F}$ is a class of functions $f : \mathbb{R}^p \to \mathbb{R}$ such that $\forall\, \mathbf{x} \in \mathbb{R}^p$, $|f(\mathbf{x})| < B$, for $0 < B < \infty$. Then, for $\epsilon > 0$ (arbitrary) it follows:

$$\mathbb{P}\left\{\exists\, f \in \mathcal{F} :\, ||f|| - 2||f||_n > \epsilon\right\} \leq 3 \cdot \mathbb{E}\left[\mathcal{N}_2\left(\frac{\sqrt{2}}{24}\epsilon,\, \mathcal{F}, \mathbf{X}_1^{2n}\right)\right] e^{-\frac{n\epsilon^2}{288\, B^2}}\,, \tag{C.9}$$

where $||g||^2 = \int_{\mathbb{R}^p} |g(\mathbf{x})|^2 d\mathbf{x}$ and $||g||_n^2 = \frac{1}{n}\sum_{i=1}^n |g(\mathbf{x}_i)|^2$. A detailed proof of this Lemma, together with definitions and details about covering and packing numbers can be found in section 11 of [9].

C.2 Proof of Theorem 4.3.1.

Suppose an orthonormal set of functions $\left\{\phi_{J,k}^{per}(x), k = 0, ..., 2^J - 1\right\}$ which as $J \to \infty$ is dense in $\mathbb{L}_2(\nu([0,1]))$ for $\nu \in \Upsilon$, and let Υ be the set of bounded Lebesgue measures in $[0,1]$. Suppose μ is a bounded Lebesgue measure in $[0,1]^p$, and the following conditions are satisfied for the scaling function ϕ:

(a) $\exists\, \Phi$, bounded and non-increasing function in $\mathbb{R}$ such that $\int \Phi(|u|)du < \infty$ and $|\phi(u)| \leq \Phi(|u|)$ almost everywhere (a.e.).

(b) In addition, $\int_{\mathbb{R}} |u|^{N+1}\Phi(|u|)du < \infty$ for some $N \geq 0$.

(c) $\exists\, F$, integrable, such that $|K(x,y)| \leq F(x-y)$, $\forall x, y \in \mathbb{R}$, for $K(x,y) = \sum_k \phi(x-k)\phi(y-k)$.

(d) Suppose ϕ satisfies:

i. $\sum_k |\hat{\phi}(\xi + 2k\pi)|^2 = 1$, a.e., where $\hat{\phi}$ denotes the Fourier transform of the scaling function ϕ.

ii. $\hat{\phi}(\xi) = \hat{\phi}(\frac{\xi}{2}) m_0(\frac{\xi}{2})$, where $m_0(\xi)$ is a 2π-periodic function and $m_0 \in \mathbb{L}_2(0, 2\pi)$.

(e) $\int_{\mathbb{R}} x^k \psi(x) dx = 0$, for $k = 0, 1, ..., N$, $N \geq 1$ where ψ is the mother wavelet corresponding to ϕ.

(f) The functions $\{f_l\}_{l=1}^p$, are such that $f_l \in L_\infty([0,1])$ and $f_l \in W_\infty^{m+1}([0,1])$, $m \geq N$, where $W_\infty^m([0,1])$ denotes the space of functions that are m-times weakly-differentiable and $f_l^{(k)} \in L_\infty([0,1])$, $k = 1, ..., m$.

(g) $\theta_\phi(x) := \sum_k |\phi(x-k)|$ such that $||\theta_\phi||_\infty < \infty$.

Under Corollary 8.2 [57], if $f \in W_\infty^{N+1}([0,1])$ then $||K_J f - f||_\infty^p = \mathcal{O}\left(2^{-pJ(N+1)}\right)$, $p \geq 1$. Furthermore, assume condition (**A3**) is satisfied. Define the set of functions:

$$\mathcal{F}_n = \left\{ f : [0,1]^p \to \mathbb{R} \,|\, f(\mathbf{x}) = \sum_{j=1}^{p} \sum_{k=0}^{2^J - 1} c_{Jk}^{(j)} \phi_{Jk}^{per}(x_j)\,;\; J = J(n) \right\}, \tag{C.10}$$

where $x_j, j = 1, ..., p$ corresponds to the j-th component of the vector $\mathbf{x} \in [0,1]^p$. Also, let $\beta_n > 0$ be a parameter depending on the sample and assume $\mathbb{E}[Y^2] < \infty$. Define $\hat{f}_{J(n)}$ as in (4.12) and let $f_{J(n)} = T_{\beta_n} \hat{f}_{J(n)} := \hat{f}_{J(n)} \mathbb{1}_{\{|\hat{f}_{J(n)}| \leq \beta_n\}} + \text{sign}(\hat{f}_{J(n)}) \beta_n \mathbb{1}_{\{|\hat{f}_{J(n)}| > \beta_n\}}$, $\mathcal{K}_n = 2^{J(n)}$. Assume the following conditions hold:

(i) $\beta_n \to \infty$ as $n \to \infty$.

(ii) $\frac{\mathcal{K}_n \beta_n^4 \log(\beta_n)}{n} \to 0$ as $n \to \infty$.

(iii) For some $\delta > 0$ as $n \to \infty$ $\frac{n^{1-\delta}}{\beta_n^4} \to \infty$.

Then:

$$\lim_{n\to\infty} \int \left| f_{J(n)}(\mathbf{x}) - f_A(\mathbf{x}) \right|^2 \mu(d\mathbf{x}) = 0 \quad \text{(a.s.)}\,, \tag{C.11}$$

$$\lim_{n\to\infty} \mathbb{E}\left\{ \int \left| f_{J(n)}(\mathbf{x}) - f_A(\mathbf{x}) \right|^2 \mu(d\mathbf{x}) \right\} = 0\,. \tag{C.12}$$

Proof

The proof for this theorem is based on the application of Theorem G4 (Györfi et al. 2002) described in C.1.5, checking conditions (C.5)-(C.8) are satisfied.

This proof is composed of 2 parts: the first shows that conditions (C.5) and (C.7) are implied by assumption (i). The second part shows that assumptions (ii) and (iii) imply conditions (C.6) and (C.8) of Theorem C.1.5.

Part 1

Consider an arbitrary $\epsilon > 0$. Then for $f \in \mathcal{F}_n$, it follows:

$$\begin{aligned}
\int_{[0,1]^p} |f(\mathbf{x}) - f_A(\mathbf{x})|^2 \mu(d\mathbf{x}) &= \int_{[0,1]^p} \left| \sum_{j=1}^{p} \left(\sum_{k=0}^{2^J-1} c_{Jk}^{(j)} \phi_{Jk}^{per}(x_j) - f_j(x_j) \right) \right|^2 \mu(d\mathbf{x}) \\
&\leq p \cdot \sum_{j=1}^{p} \int_{[0,1]^p} \left(\sum_{k=0}^{2^J-1} c_{Jk}^{(j)} \phi_{Jk}^{per}(x_j) - f_j(x_j) \right)^2 \mu(d\mathbf{x}) \\
&\leq p \cdot \sum_{j=1}^{p} \int_{0}^{1} \left(\sum_{k=0}^{2^J-1} c_{Jk}^{(j)} \phi_{Jk}^{per}(x_j) - f_j(x_j) \right)^2 \nu_j(dx_j)
\end{aligned} \tag{C.13}$$

where $\nu_1, ..., \nu_p$ are bounded Lebesgue measures on $[0,1]$ (since μ is a bounded Lebesgue measure in $[0,1]^p$). Since $\left\{ \phi_{j,k}^{per}(x), k = 0, ..., 2^j - 1,\ j \geq 0 \right\}$ is dense in $\mathbb{L}_2(\nu([0,1]))$, by

Proposition 1 in 4.3.1:

$$\exists \ \left\{ c_{J,0}^{(1)*}, ..., c_{J,2^J-1}^{(1)*}, ..., c_{J,0}^{(p)*}, ..., c_{J,2^J-1}^{(p)*} \right\},$$

for which $J = J^*(n_0(\epsilon))$ such that:

$$\int_{[0,1]^p} \left| \sum_{j=1}^{p} \left(\sum_{k=0}^{2^J-1} c_{J,k}^{(j)} \phi_{J,k}^{per}(x_j) - f_j(x_j) \right) \right|^2 \mu(d\mathbf{x}) \leq \epsilon \,. \tag{C.14}$$

Therefore, for a given $\epsilon > 0$, it is possible to find $n_0(\epsilon)$ such that for $J^* = J(n_0(\epsilon))$ (C.14) holds.

Now for a fixed $n = n_0(\epsilon)$ the set $\mathcal{F}_n$ is composed of functions that are uniformly bounded by a parameter depending on the sample size. In fact, it is possible to show that $||f||_\infty \leq ||\theta_\phi||_\infty ||f_j^*||_\infty \cdot 2^{\frac{J(n_0(\epsilon))}{2}}$, where $||f_j^*||_\infty = \max\limits_{j=1,...,p} ||f_j||_\infty$. Therefore, for an arbitrary $\epsilon > 0$, and for all $n \leq n_0(\epsilon)$, $\exists\, \beta_n > 0$ such that:

$$\sum_{j=1}^{p} \sum_{k=0}^{2^{J(n)}-1} c_{Jk}^{(j)*} \phi_{Jk}^{per}(x_j) \ \in \ \left\{ f \in \mathcal{F}_n \,|\, ||f||_\infty \leq \beta_{n_0(\epsilon)} \right\} .$$

From this last result and (C.13),(C.14), for $n \geq n_0(\epsilon)$ it follows:

$$\inf_{\{f \in \mathcal{F}_n \,|\, ||f||_\infty \leq \beta_n\}} \int_{[0,1]^p} |f(\mathbf{x}) - f_A(\mathbf{x})|^2 \, \mu(d\mathbf{x}) \leq \epsilon \,. \tag{C.15}$$

Since $\epsilon > 0$ is arbitrary, (C.15) implies:

$$\lim_{n \to \infty} \left\{ \inf_{\{f \in \mathcal{F}_n \,|\, ||f||_\infty \leq \beta_n\}} \int_{[0,1]^p} |f(\mathbf{x}) - f_A(\mathbf{x})|^2 \, \mu(d\mathbf{x}) \right\} = 0 \,, \tag{C.16}$$

which shows that as $J = J(n) \to \infty$ $(n \to \infty)$ and $\beta_n \to \infty$ $(n \to \infty)$, (C.5) is satisfied.

From (C.15) and the last result, the dominated convergence theorem implies:

$$\lim_{n\to\infty} \mathbb{E}\left\{ \inf_{\{f\in\mathcal{F}_n \,|\, ||f||_\infty \leq \beta_n\}} \int_{[0,1]^p} |f(\mathbf{x}) - f_A(\mathbf{x})|^2 \, \mu(d\mathbf{x}) \right\} = 0\,, \tag{C.17}$$

therefore, (C.7) is also implied, provided $J = J(n) \to \infty$ $(n \to \infty)$ and $\beta_n \to \infty$ $(n \to \infty)$.

Part 2

In this part, we use results provided in section C.1.5 of the appendix. Consider $L > 0$ arbitrary and assume (wlog) that $L < \beta_n$. Define $\mathbf{Z} = (\mathbf{X}, Y)$ and $\mathbf{Z}_i = (\mathbf{X}_i, Y_i)$ for $i = 1, ..., n$. Also, define the set of functions:

$$\mathcal{G}_n = \left\{ g, : \, [0,1]^p \times \mathbb{R} \to \mathbb{R} \, : \, \exists \, f \in T_{\beta_n}\mathcal{F}_n \text{ s.t. } g(\mathbf{X}, y) = |f(\mathbf{X}) - T_L Y|^2 \right\} .$$

Note that the last definition implies that $\sup_{f\in T_{\beta_n}\mathcal{F}_n} \left| \frac{1}{n}\sum_{i=1}^n |f(\mathbf{X}_i) - Y_{i,L}|^2 - \mathbb{E}\left[(f(\mathbf{X}) - Y_L)^2\right] \right|$ is equivalent to:

$$\sup_{g\in\mathcal{G}_n} \left| \frac{1}{n}\sum_{i=1}^n g(\mathbf{Z}_i) - \mathbb{E}\left[g(\mathbf{Z})\right] \right| .$$

Moreover, since it is assumed that $L < \beta_n$, every function $g \in \mathcal{G}_n$ satisfies $0 \leq g(\mathbf{Z}) \leq 4\beta_n^2$. This allows the application of Theorem P1 (Pollard 1984) as follows:

For an arbitrary $\epsilon > 0$, it follows:

$$\mathbb{P}\left\{ \sup_{g\in\mathcal{G}_n} \left| \frac{1}{n}\sum_{i=1}^n g(\mathbf{Z}_i) - \mathbb{E}\left[g(\mathbf{Z})\right] \right| > \epsilon \right\} \leq 8 \cdot \mathbb{E}\left[\mathcal{N}_1\left(\frac{\epsilon}{8}, \mathcal{G}_n, \mathbf{z}_1^n\right)\right] e^{-\frac{n\epsilon^2}{2048\beta_n^4}} . \tag{C.18}$$

Lemma G1 shows that $\mathcal{N}_1\left(\frac{\epsilon}{8}, \mathcal{G}_n, \mathbf{z}_1^n\right) \leq \mathcal{M}_1\left(\frac{\epsilon}{8}, \mathcal{G}_n, \mathbf{z}_1^n\right)$. Therefore, a relation between $\mathcal{M}_1\left(\frac{\epsilon}{8}, \mathcal{G}_n, \mathbf{z}_1^n\right)$ and $\mathcal{M}_1\left(\lambda, T_{\beta_n}\mathcal{F}_n, \mathbf{X}_1^n\right)$ needs to be established for some $\lambda = \lambda(\epsilon) > 0$.

Consider $g_1, g_2 \in \mathcal{G}_n$ (i.e. $\exists \, f_1, f_2 \, \in T_{\beta_n}\mathcal{F}_n$ s.t. $g(\mathbf{X}, y) = |f(\mathbf{X}) - T_L Y|^2$), then if $\{g_1, ..., g_M\}$

is an $\mathbb{L}_1$-$\frac{\epsilon}{8}$ packing of $\mathcal{G}_n$ on $\mathbf{z}_1^n$, $\forall 1 \leq j < m \leq M$ it holds:

$$\frac{1}{n}\sum_{i=1}^{n}|g_j(\mathbf{z}_i) - g_m(\mathbf{z}_i)| \;\; \geq \;\; \frac{\epsilon}{8}\,.$$

Using the definition of $\mathcal{G}_n$, it follows:

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}|g_1(\mathbf{z}_i) - g_2(\mathbf{z}_i)| &= \frac{1}{n}\sum_{i=1}^{n}\left||f_1(\mathbf{X}_i) - T_L Y_i|^2 - |f_2(\mathbf{X}_i) - T_L Y_i|^2\right| \\
&= \frac{1}{n}\sum_{i=1}^{n}\left(|f_1(\mathbf{X}_i) - f_2(\mathbf{X}_i)|\,|f_1(\mathbf{X}_i) + f_2(\mathbf{X}_i) - 2T_L Y_i|\right) \\
&\leq \frac{1}{n}\sum_{i=1}^{n}|f_1(\mathbf{X}_i) - f_2(\mathbf{X}_i)| \cdot 4\beta_n \\
\frac{\epsilon}{32\beta_n} &\leq \sum_{i=1}^{n}|f_1(\mathbf{X}_i) - f_2(\mathbf{X}_i)|\,.
\end{aligned}$$

Therefore, if $\{g_1, ..., g_M\}$ is an $\mathbb{L}_1$-$\frac{\epsilon}{8}$ packing of $\mathcal{G}_n$ on $\mathbf{z}_1^n$, then $\{f_1, ..., f_M\}$ is an $\mathbb{L}_1$-$\frac{\epsilon}{32\beta_n}$ packing of $T_{\beta_n}\mathcal{F}_n$ on $\mathbf{X}_1^n$. Thus this result implies:

$$\mathcal{M}_1\left(\frac{\epsilon}{8}, \mathcal{G}_n, \mathbf{z}_1^n\right) \leq \mathcal{M}_1\left(\frac{\epsilon}{32\beta_n}, T_{\beta_n}\mathcal{F}_n, \mathbf{X}_1^n\right) \tag{C.19}$$

Substituting the last result in (C.18), leads to:

$$\mathbb{P}\left\{\sup_{g\in\mathcal{G}_n}\left|\frac{1}{n}\sum_{i=1}^{n}g(\mathbf{Z}_i) - \mathbb{E}\left[g(\mathbf{Z})\right]\right| > \epsilon\right\} \leq 8\cdot\mathbb{E}\left[\mathcal{M}_1\left(\frac{\epsilon}{32\beta_n}, T_{\beta_n}\mathcal{F}_n, \mathbf{X}_1^n\right)\right] e^{-\frac{n\epsilon^2}{2048\beta_n^4}}\,. \tag{C.20}$$

Now, applying Theorem G2, for $0 < \epsilon < \frac{\beta_n}{4}$ it follows:

$$\mathcal{M}_1\left(\frac{\epsilon}{32\beta_n}, T_{\beta_n}\mathcal{F}_n, \mathbf{X}_1^n\right) \leq 3\left(\frac{128\,e\,\beta_n^2}{\epsilon}\log\left(\frac{192\,e\,\beta_n^2}{\epsilon}\right)\right)^{V_{T_{\beta_n}\mathcal{F}_n^+}}\,. \tag{C.21}$$

Since $T_{\beta_n}\mathcal{F}_n^+ = \{(\mathbf{x},t) \in [0,1]^p \times \mathbb{R} : t \leq f(\mathbf{x}),\ f \in T_{\beta_n}\mathcal{F}_n\}$, for $t > \beta_n$ the pair $(\mathbf{x},t) \notin T_{\beta_n}\mathcal{F}_n^+$. On the contrary, when $t \leq \beta_n$ since $\forall f \in T_{\beta_n}\mathcal{F}_n\ \beta_n \leq f \leq \beta_n$, every pair $(\mathbf{x},t) \in T_{\beta_n}\mathcal{F}_n^+$. This implies:

$$V_{T_{\beta_n}\mathcal{F}_n^+} \leq V_{\mathcal{F}_n^+}\,. \tag{C.22}$$

Similarly, since $dim(\mathcal{F}_n) = p \cdot 2^J$, Theorem G3 implies:

$$V_{\mathcal{F}_n^+} \leq p \cdot 2^J + 1\,. \tag{C.23}$$

Combining (C.21), (C.22), and (C.23), it is possible to express (C.20) as:

$$\begin{aligned}
\mathbb{P}\left\{\sup_{g\in\mathcal{G}_n}\left|\frac{1}{n}\sum_{i=1}^{n} g(\mathbf{Z}_i) - \mathbb{E}\left[g(\mathbf{Z})\right]\right| > \epsilon\right\} &\leq 24\cdot\left(\left(\frac{128\,e\,\beta_n^2}{\epsilon}\log\left(\frac{192\,e\,\beta_n^2}{\epsilon}\right)\right)^{(p\cdot 2^J+1)}\right) e^{-\frac{n\epsilon^2}{2048\beta_n^4}} \\
&\leq 24\cdot e^{2(p\cdot 2^J+1)\log\left(\frac{192\,e\,\beta_n^2}{\epsilon}\right)-\frac{n\epsilon^2}{2048\beta_n^4}}\,.
\end{aligned}$$

Finally, it follows:

$$\begin{aligned}
\mathbb{P}\left\{\sup_{g\in\mathcal{G}_n}\left|\frac{1}{n}\sum_{i=1}^{n} g(\mathbf{Z}_i) - \mathbb{E}\left[g(\mathbf{Z})\right]\right| > \epsilon\right\} &\leq \sum_{n=1}^{\infty}\mathbb{P}\left\{\sup_{g\in\mathcal{G}_n}\left|\frac{1}{n}\sum_{i=1}^{n} g(\mathbf{Z}_i) - \mathbb{E}\left[g(\mathbf{Z})\right]\right| > \epsilon\right\} \\
&\leq \sum_{n=1}^{\infty} 24\cdot e^{2(p\cdot 2^J+1)\log\left(\frac{192\,e\,\beta_n^2}{\epsilon}\right)-\frac{n\epsilon^2}{2048\beta_n^4}} \\
&\leq \sum_{n=1}^{\infty} 24\cdot e^{\left\{-n^{\delta}\frac{n^{1-\delta}}{\beta_n^4}\left(\frac{\epsilon^2}{2048}-\frac{2(p\cdot 2^J+1)\beta_n^4}{n}\log\left(\frac{192\,e\,\beta_n^2}{\epsilon}\right)\right)\right\}}
\end{aligned} \tag{C.24}$$

Notice that if for some $\delta > 0$ the following conditions hold:

(a) $\frac{n^{1-\delta}}{\beta_n^4} \longrightarrow \infty$ as $n \to \infty$,

(b) $\frac{2(p\cdot 2^J+1)\beta_n^4}{n}\log\left(\frac{192\,e\,\beta_n^2}{\epsilon}\right) \longrightarrow \infty$ as $n \to \infty$,

then the series (C.24) is absolutely convergent. Denote $\mathcal{K}_n = p\cdot 2^J$ and observe that condition

(b) can be bounded as:

$$\begin{aligned}\frac{2(\mathcal{K}_n+1)\beta_n^4}{n}\log\left(\frac{192\,e\,\beta_n^2}{\epsilon}\right) &\leq \frac{4(\mathcal{K}_n+1)\beta_n^4\log(\beta_n)}{n}+\frac{C_1(\mathcal{K}_n+1)\beta_n^4}{n}\\ &\leq C_2\frac{\mathcal{K}_n\beta_n^4\log(\beta_n)}{n}\,, \end{aligned} \tag{C.25}$$

for a constant $C_2>0$ independent of n.

Therefore, if $\frac{\mathcal{K}_n\beta_n^4\log(\beta_n)}{n}\longrightarrow\infty$ as $n\to\infty$, then we get condition (b) satisfied by assumption (ii). This implies that the terms in the series (C.24) go to zero. Therefore:

$$\sum_{n=1}^{\infty}24\cdot e^{\left\{-n^{\delta}\frac{n^{1-\delta}}{\beta_n^4}\left(\frac{\epsilon^2}{2048}-\frac{2(p\cdot 2^J+1)\beta_n^4}{n}\log\left(\frac{192\,e\,\beta_n^2}{\epsilon}\right)\right)\right\}}<\infty\,.$$

This result implies that $\exists\, n_0(\epsilon)$ such that for $n>n_0(\epsilon)$, it follows:

$$\mathbb{P}\left\{\sup_{g\in\mathcal{G}_n}\left|\frac{1}{n}\sum_{i=1}^{n}g(\mathbf{Z}_i)-\mathbb{E}\left[g(\mathbf{Z})\right]\right|>\epsilon\right\}\longrightarrow 0\ \ (n\to\infty)\,. \tag{C.26}$$

Similarly, for $\epsilon>0$ it follows:

$$\begin{aligned}\mathbb{E}\left\{\sup_{g\in\mathcal{G}_n}\left|\frac{1}{n}\sum_{i=1}^{n}g(\mathbf{Z}_i)-\mathbb{E}\left[g(\mathbf{Z})\right]\right|\right\} &= \int_0^{\infty}\mathbb{P}\left\{\sup_{g\in\mathcal{G}_n}\left|\frac{1}{n}\sum_{i=1}^{n}g(\mathbf{Z}_i)-\mathbb{E}\left[g(\mathbf{Z})\right]\right|>t\right\}dt\\ &\leq \epsilon+\int_{\epsilon}^{\infty}\mathbb{P}\left\{\sup_{g\in\mathcal{G}_n}\left|\frac{1}{n}\sum_{i=1}^{n}g(\mathbf{Z}_i)-\mathbb{E}\left[g(\mathbf{Z})\right]\right|>t\right\}dt\\ &\leq \epsilon+\int_{\epsilon}^{\infty}24\cdot\left(\left(\frac{192\,e\,\beta_n^2}{t}\right)^{2(\mathcal{K}_n+1)}\right)e^{-\frac{n\,t^2}{2048\beta_n^4}}dt\\ &\leq \epsilon+24\frac{2048\beta_n^4}{n\epsilon}e^{2(\mathcal{K}_n+1)\log\left(\frac{192\,e\,\beta_n^2}{\epsilon}\right)-\frac{n\epsilon^2}{2048\beta_n^4}}\\ &\leq \epsilon+24\cdot 2048\frac{1}{n^{\delta}}\frac{\beta_n^4}{n^{1-\delta}}e^{-n^{\delta}\frac{n^{1-\delta}}{\beta_n^4}\left(\frac{\epsilon^2}{2048}-\frac{2(\mathcal{K}_n+1)\beta_n^4}{n}\log\left(\frac{192\,e\,\beta_n^2}{\epsilon}\right)\right)}\,. \end{aligned}$$

Clearly, since condition (a) and (b) are satisfied by assumptions (ii) and (iii), the second term of the above equation goes to zero as $n \to \infty$. Since ϵ is arbitrary, this implies:

$$\mathbb{E}\left\{\sup_{g\in\mathcal{G}_n}\left|\frac{1}{n}\sum_{i=1}^{n} g(\mathbf{Z}_i) - \mathbb{E}\left[g(\mathbf{Z})\right]\right|\right\} \longrightarrow 0 \;\; (n \to \infty)\,. \tag{C.27}$$

By the Borel-Cantelli Lemma, (C.27) and (C.26) show assumptions (ii) and (iii) imply conditions (C.6) and (C.8) of Theorem C.1.5. This, together with results from Part 1 show that (C.11) and (C.12) hold, and the Theorem is proved.

C.3 Proof of Lemma 4.3.2.

Suppose an orthonormal set of functions $\left\{\phi_{j,k}^{per}(x), k = 0, ..., 2^j - 1,\, j \geq 0\right\}$ which is dense in $\mathbb{L}_2(\nu([0,1]))$ for $\nu \in \Upsilon$, which represents the set of bounded lebesgue measures in $[0,1]$. Suppose μ is a bounded lebesgue measure in $[0,1]^p$ and that conditions stated in Theorem 1 for the scaling function ϕ, and assumptions (**A1**)-(**A4**) presented in 4.2 hold.

Define the set of functions $\mathcal{F}_n$ as in (C.10). Also, let $\beta_n > 0$ be a parameter depending on the sample and assume $\mathbb{E}\left[Y^2\right] < \infty$. Define $\hat{f}_{J(n)}$ as in (4.12) and let $f_{J(n)} = T_{\beta_n}\hat{f}_{J(n)}$, let $\mathcal{K}_n = p\, 2^{J(n)}$.

Furthermore, assume the following condition holds:

(i) $\sum_{j=1}^{p} ||f_j||_\infty < L$, for some $L < \beta_n$.

Then:

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\left|f_{J(n)}(\mathbf{x}_i) - f_A(\mathbf{x}_i)\right|^2 \mid \mathbf{X}_1^n\right] \leq \min_{f\in\mathcal{F}_n}\left\{||f - f_A||_n^2\right\} + \frac{\sigma^2}{n}\mathcal{K}_n \tag{C.28}$$

Proof

First, note that $||f_A||_\infty < \beta_n$ (from condition (i)), implies that $||f_{J(n)} - f_A||_n^2 \leq ||\hat{f}_{J(n)} - f_A||_n^2$. Therefore, this further implies:

$$\begin{aligned}
\mathbb{E}\left[||f_{J(n)} - f_A||_n^2 \mid \mathbf{X}_1^n\right] &\leq \mathbb{E}\left[||\hat{f}_{J(n)} - f_A||_n^2 \mid \mathbf{X}_1^n\right] \\
&\leq \mathbb{E}\left[\left\|\hat{f}_{J(n)} - \mathbb{E}\left[\hat{f}_{J(n)} \mid \mathbf{X}_1^n\right] + \mathbb{E}\left[\hat{f}_{J(n)} \mid \mathbf{X}_1^n\right] - f_A\right\|_n^2 \mid \mathbf{X}_1^n\right] \\
&\leq \mathbb{E}\left[\left\|\hat{f}_{J(n)} - \mathbb{E}\left[\hat{f}_{J(n)} \mid \mathbf{X}_1^n\right]\right\|_n^2 \mid \mathbf{X}_1^n\right] + \mathbb{E}\left[\left\|\mathbb{E}\left[\hat{f}_{J(n)} \mid \mathbf{X}_1^n\right] - f_A\right\|_n^2 \mid \mathbf{X}_1^n\right] \\
&\quad +2\mathbb{E}\left\{\frac{1}{n}\sum_{i=1}^{n}\left(\hat{f}_{J(n)}(\mathbf{X}_i) - \mathbb{E}\left[\hat{f}_{J(n)} \mid \mathbf{X}_1^n\right]\right)\left(\mathbb{E}\left[\hat{f}_{J(n)} \mid \mathbf{X}_1^n\right] - f_A(\mathbf{X}_i)\right) \mid \mathbf{X}_1^n\right. \\
&\leq \mathbb{E}\left[\left\|\hat{f}_{J(n)} - \mathbb{E}\left[\hat{f}_{J(n)} \mid \mathbf{X}_1^n\right]\right\|_n^2 \mid \mathbf{X}_1^n\right] + \left\|\mathbb{E}\left[\hat{f}_{J(n)} \mid \mathbf{X}_1^n\right] - f_A\right\|_n^2, \qquad \text{(C.2}
\end{aligned}$$

where the last result follows since the last term in the third inequality is zero. From definitions (4.10), (4.13), and (4.14), for any $i \in \{1, ..., n\}$ it follows:

$$\begin{aligned}
\mathbb{E}\left[\hat{f}_{J(n)}(\mathbf{X}_i) \mid \mathbf{X}_1^n\right] &= \mathbb{E}\left[\mathbf{B}^T(\mathbf{X}_i)\,\mathbf{c}^* \mid \mathbf{X}_1^n\right] \\
&= \mathbf{B}^T(\mathbf{X}_i)\,\left(\mathbf{B}^T\mathbf{B}\right)^{-1}\mathbf{B}^T\,\mathbb{E}\left[\mathbf{Y} \mid \mathbf{X}_1^n\right] \\
&= \mathbf{B}^T(\mathbf{X}_i)\,\left(\mathbf{B}^T\mathbf{B}\right)^{-1}\mathbf{B}^T \begin{bmatrix} f_A(\mathbf{X}_1) \\ \vdots \\ f_A(\mathbf{X}_n) \end{bmatrix} \\
&= \mathbf{B}^T(\mathbf{X}_i)\,\left(\mathbf{B}^T\mathbf{B}\right)^{-1}\mathbf{B}^T\,\mathbf{f}. \qquad \text{(C.30)}
\end{aligned}$$

Now, from the last set of equations, it follows that $\mathbb{E}\left[\mathbf{c}^* \mid \mathbf{X}_1^n\right] = \left(\mathbf{B}^T\mathbf{B}\right)^{-1}\mathbf{B}^T\,\mathbf{f}$, which implies:

$$\frac{1}{n}\left(\mathbf{B}^T\mathbf{B}\right)\mathbb{E}\left[\mathbf{c}^* \mid \mathbf{X}_1^n\right] = \frac{1}{n}\mathbf{B}^T\,\mathbf{f}.$$

Therefore, $\mathbb{E}\left[\mathbf{c}^* \mid \mathbf{X}_1^n\right]$ is the least squares solution for the problem: $\min_{\mathbf{a}\in\mathbb{R}^{\mathcal{K}_n}} \left\{\|\mathbf{B}\,\mathbf{a}-\mathbf{f}\|_n^2\right\}$. This implies that $\left\|\mathbb{E}\left[\hat{f}_{J(n)} \mid \mathbf{X}_1^n\right] - f_A\right\|_n^2 = \min_{f\in\mathcal{F}_n} ||f-f_A||_n^2$.

Therefore, this result and (C.29), imply:

$$\mathbb{E}\left[||f_{J(n)}-f_A||_n^2 \mid \mathbf{X}_1^n\right] \leq \mathbb{E}\left[\left\|\hat{f}_{J(n)} - \mathbb{E}\left[\hat{f}_{J(n)} \mid \mathbf{X}_1^n\right]\right\|_n^2 \mid \mathbf{X}_1^n\right] + \min_{f\in\mathcal{F}_n} ||f-f_A||_n^2 .$$

For a fixed $\mathbf{x}$, from definitions (4.10), (4.13) and (4.14), it follows:

$$\begin{aligned}
\mathbb{E}\left[\left|\hat{f}_{J(n)}(\mathbf{x}) - \mathbb{E}\left[\hat{f}_{J(n)}(\mathbf{x}) \mid \mathbf{X}_1^n\right]\right|^2 \mid \mathbf{X}_1^n\right] &= \mathbb{E}\left[\left|\mathbf{B}(\mathbf{x})^T\,\mathbf{c}^* - \mathbf{B}(\mathbf{x})^T\mathbb{E}\left[\mathbf{c}^* \mid \mathbf{X}_1^n\right]\right|^2 \mid \mathbf{X}_1^n\right] \\
&= \mathbb{E}\left[\left|\mathbf{B}(\mathbf{x})^T\,\left(\mathbf{B}^T\mathbf{B}\right)^{-1}\mathbf{B}^T\left(\mathbf{Y}-\mathbf{f}\right)\right|^2 \mid \mathbf{X}_1^n\right] \\
&= \mathbf{B}(\mathbf{x})^T\,\mathbf{H}\,\mathbb{E}\left[\left(\mathbf{Y}-\mathbf{f}\right)\left(\mathbf{Y}-\mathbf{f}\right)^T\right]\mathbf{H}^T\mathbf{B}(\mathbf{x}) \qquad \text{(C.31)}
\end{aligned}$$

where $\mathbf{H} = \left(\mathbf{B}^T\mathbf{B}\right)^{-1}\mathbf{B}^T$. By the assumptions of model (4.1), it follows that $\mathbb{E}\left[\left(\mathbf{Y}-\mathbf{f}\right)\left(\mathbf{Y}-\mathbf{f}\right)^T\right] = \sigma^2\mathbf{I}_{\mathcal{K}_n}$. Therefore, (C.31) can be expressed as:

$$\mathbb{E}\left[\left|\hat{f}_{J(n)}(\mathbf{x}) - \mathbb{E}\left[\hat{f}_{J(n)}(\mathbf{x}) \mid \mathbf{X}_1^n\right]\right|^2 \mid \mathbf{X}_1^n\right] = \sigma^2\mathbf{B}(\mathbf{x})^T\left(\mathbf{B}^T\mathbf{B}\right)^{-1}\mathbf{B}(\mathbf{x}) .$$

By substituting this result in $\mathbb{E}\left[\left\|\hat{f}_{J(n)} - \mathbb{E}\left[\hat{f}_{J(n)} \mid \mathbf{X}_1^n\right]\right\|_n^2 \mid \mathbf{X}_1^n\right]$, it follows:

$$\mathbb{E}\left[||f_{J(n)}-f_A||_n^2 \mid \mathbf{X}_1^n\right] \leq \min_{f\in\mathcal{F}_n} ||f-f_A||_n^2 + \frac{\sigma^2}{n}\sum_{i=1}^{n}\mathbf{B}(\mathbf{x}_i)^T\left(\mathbf{B}^T\mathbf{B}\right)^{-1}\mathbf{B}(\mathbf{x}_i) . \qquad \text{(C.32)}$$

Notice that:

$$\begin{aligned}
\sum_{i=1}^{n} \mathbf{B}(\mathbf{x}_i)^T \left(\mathbf{B}^T\mathbf{B}\right)^{-1} \mathbf{B}(\mathbf{x}_i) &= \text{trace}\left\{\sum_{i=1}^{n} \mathbf{B}(\mathbf{x}_i)^T \left(\mathbf{B}^T\mathbf{B}\right)^{-1} \mathbf{B}(\mathbf{x}_i)\right\} \\
&= \sum_{i=1}^{n} \text{trace}\left\{\mathbf{B}(\mathbf{x}_i)^T \left(\mathbf{B}^T\mathbf{B}\right)^{-1} \mathbf{B}(\mathbf{x}_i)\right\} \\
&= \sum_{i=1}^{n} \text{trace}\left\{\left(\mathbf{B}^T\mathbf{B}\right)^{-1} \mathbf{B}(\mathbf{x}_i)\mathbf{B}(\mathbf{x}_i)^T\right\} \\
&= \text{trace}\left\{\left(\mathbf{B}^T\mathbf{B}\right)^{-1} \sum_{i=1}^{n} \mathbf{B}(\mathbf{x}_i)\mathbf{B}(\mathbf{x}_i)^T\right\} \\
&= \text{trace}\left\{\mathbf{I}_{\mathcal{K}_n}\right\} = \mathcal{K}_n\,, \qquad \text{(C.33)}
\end{aligned}$$

where the last 2 equalities follow from definitions (4.10) and (4.13). In fact, it is possible to observe that $\sum_{i=1}^{n} \mathbf{B}(\mathbf{x}_i)\mathbf{B}(\mathbf{x}_i)^T = \mathbf{B}^T\mathbf{B}$. Therefore, this result applied to (C.32) implies:

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\left|f_{J(n)}(\mathbf{x}_i) - f_A(\mathbf{x}_i)\right|^2 \mid \mathbf{X}_1^n\right] \leq \min_{f\in\mathcal{F}_n}\left\{||f - f_A||_n^2\right\} + \frac{\sigma^2}{n}\mathcal{K}_n\,.$$

which proves assertion (C.28).

C.4 Proof of Lemma 4.3.3.

Suppose an orthonormal basis $\left\{\phi_{j,k}^{per}(x), k = 0, ..., 2^j - 1,\ j \geq 0\right\}$ which is dense in $\mathbb{L}_2(\nu([0,1]))$ for $\nu \in \Upsilon$, which represents the set of bounded lebesgue measures in $[0,1]$. Suppose assumptions stated in Theorem 1 for the scaling function ϕ, and conditions (**A1**)-(**A4**) defined in 4.2 hold. Let the set of functions $\mathcal{F}_n$ to be defined as in (C.10).

Then it follows:

$$\inf_{f\in\mathcal{F}_n}\int_{[0,1]^p}\left|f(\mathbf{x}) - f_A(\mathbf{x})\right|^2 \mu(d\mathbf{x}) \leq p^2\, C_2^2\, 2^{-2(N+1)\, J(n)}\,. \qquad \text{(C.34)}$$

Proof

Denote $f_j^J = \sum_{k=0}^{2^J-1} c_{Jk}^{(j)} \phi_{Jk}^{per}$. Consider:

$$\begin{aligned}
\inf_{f\in\mathcal{F}_n} \int_{[0,1]^p} \left|f(\mathbf{x}) - f_A(\mathbf{x})\right|^2 \mu(d\mathbf{x}) &= \inf_{f\in\mathcal{F}_n} \int_{[0,1]^p} \left|\sum_{j=1}^{p} \left(f_j^J(x_j) - f_j(x_j)\right)\right|^2 \mu(d\mathbf{x}) \\
&\leq p \inf_{f\in\mathcal{F}_n} \int_{[0,1]^p} \sum_{j=1}^{p} \left|f_j^J(x_j) - f_j(x_j)\right|^2 \mu(d\mathbf{x}) \\
&\leq p \inf_{f\in\mathcal{F}_n} \sum_{j=1}^{p} \sup_{x_j\in[0,1]} \left|f_j^J(x_j) - f_j(x_j)\right|^2 \\
&\leq p \inf_{f\in\mathcal{F}_n} \sum_{j=1}^{p} \left(\sup_{x_j\in[0,1]} \left|f_j^J(x_j) - f_j(x_j)\right|\right)^2 .
\end{aligned}$$

By corollary 8.2 of [57], it follows that $\sup_{x_j\in[0,1]} \left|f_j^J(x_j) - f_j(x_j)\right| = \mathcal{O}\left(2^{-J\,(N+1)}\right)$; therefore, $\exists\, C_2$ independent of $n,$ and J such that $\sup_{x_j\in[0,1]} \left|f_j^J(x_j) - f_j(x_j)\right| \leq C_2\, 2^{-J\,(N+1)}$. Thus:

$$\inf_{f\in\mathcal{F}_n} \int_{[0,1]^p} \left|f(\mathbf{x}) - f_A(\mathbf{x})\right|^2 \mu(d\mathbf{x}) \leq p^2\, C_2^2\, 2^{-2(N+1)\,J(n)} ,$$

as desired.

C.5 Proof of Theorem 4.3.2.

This proof follows the same methodology as in section 10 of [9]. Consider assumptions stated for Lemma 1 and conditions (i)-(iii) for Theorem 1 hold . Then:

$$\mathbb{E}\left[\int_{[0,1]^p} \left|f_{J(n)}(\mathbf{x}) - f_A(\mathbf{x})\right|^2 \mu(d\mathbf{x})\right] \leq \tilde{C} \max\left\{\beta_n^2, \sigma^2\right\} \frac{p\, 2^{J(n)}}{n} \left(\log(n) + 1\right) , +8\, C_2^2\, p^2\, 2^{-2(N+1)J(n)} , \tag{C.35}$$

for proper constants $\tilde{C} > 0$ and $C_2 > 0$ independent of n, N, p.

Proof

Note that $||f_{J(n)} - f_A||^2 = \int_{[0,1]^p} \left|f_{J(n)}(\mathbf{x}) - f_A(\mathbf{x})\right|^2 \mu(d\mathbf{x})$ can be expressed as follows:

$$\begin{aligned}
||f_{J(n)} - f_A||^2 &= \left(||f_{J(n)} - f_A|| - 2\,||f_{J(n)} - f_A||_n + 2\,||f_{J(n)} - f_A||_n\right)^2 \\
&\leq \left(\max\left\{0\,,\ ||f_{J(n)} - f_A|| - 2\,||f_{J(n)} - f_A||_n\right\} + 2\,||f_{J(n)} - f_A||_n\right)^2 \\
&\leq 2\left(\max\left\{0\,,\ ||f_{J(n)} - f_A|| - 2\,||f_{J(n)} - f_A||_n\right\}\right)^2 + 8\,||f_{J(n)} - f_A||_n^2\,, \\
&\leq S_{1,n} + 8\,S_{2,n}\,.
\end{aligned}$$

Observe that $\mathbb{E}\left[S_{2,n}\right] = \mathbb{E}_{\mathbf{X}_1^n}\left[\mathbb{E}\left(||f_{J(n)} - f_A||_n^2 \mid \mathbf{X}_1^n\right)\right]$. Similarly, from the definition of $f_{J(n)}$ and condition (i) of Lemma 1, it follows that $||f_{J(n)} - f_A||_n^2 \leq ||\hat{f}_{J(n)} - f_A||_n^2$. These 2 results and Lemma 1 imply:

$$\begin{aligned}
\mathbb{E}\left[S_{2,n}\right] &\leq \mathbb{E}_{\mathbf{X}_1^n}\left[\min_{f\in\mathcal{F}_n}\left\{||f - f_A||_n^2\right\}\right] + \frac{\sigma^2}{n}\mathcal{K}_n \\
&\leq \mathbb{E}_{\mathbf{X}_1^n}\left[\min_{f\in\mathcal{F}_n}\left\{\frac{1}{n}\sum_{i=1}^{n}|f(\mathbf{x}_i) - f_A(\mathbf{x}_i)|^2\right\}\right] + \frac{\sigma^2}{n}\mathcal{K}_n \\
&\leq \inf_{f\in\mathcal{F}_n}\left\{\int_{[0,1]^p}|f(\mathbf{x}) - f_A(\mathbf{x})|^2\,\mu(d\mathbf{x})\right\} + \frac{\sigma^2}{n}\mathcal{K}_n\,, \qquad \text{(C.36)}
\end{aligned}$$

where the last inequality follows from the properties of the expected value and the iid condition of the sample $\mathbf{X}_1^n = (\mathbf{X}_1, ..., \mathbf{X}_n)$. Now, for $S_{1,n}$, define:

$$\mathcal{G}_n = \left\{g_n\ :\ [0,1]^p \to \mathbb{R}\,;\ g_n = f_{J(n)} - f_A\ \mid\ f_{J(n)} \in\ T_{\beta_n}\mathcal{F}_n\right\}\,.$$

Observe that $\forall g \in \mathcal{G}_n \ |g_n| \leq 2\,\beta_n$. Consider $u > 0$ (arbitrary) and:

$$\begin{aligned}
\mathbb{P}\left\{S_{1,n} > u\right\} &= \mathbb{P}\left\{2\,\left(\max\left\{0\,,\,||f_{J(n)} - f_A|| - 2\,||f_{J(n)} - f_A||_n\right\}\right)^2 > u\right\} \\
&= \mathbb{P}\left\{\max\left\{0\,,\,||f_{J(n)} - f_A|| - 2\,||f_{J(n)} - f_A||_n\right\} > \sqrt{\frac{u}{2}}\right\} \\
&\leq \mathbb{P}\left\{\max\left\{0\,,\,||f_{J(n)} - f_A|| - 2\,||f_{J(n)} - f_A||_n\right\} > \sqrt{\frac{u}{2}}\right\}\,.
\end{aligned}$$

From Theorem P2, it follows:

$$\begin{aligned}
\mathbb{P}\left\{S_{1,n} > u\right\} &\leq 3\,\mathbb{E}\left[\mathcal{N}_2\left(\frac{\sqrt{2}}{24}\sqrt{\frac{u}{2}}\,,\mathcal{G}_n,\mathbf{X}_1^{2n}\right)\right]\,e^{-\frac{n\,\frac{u}{2}}{288\,(2\beta_n)^2}} \\
&\leq 3\,\mathbb{E}\left[\mathcal{N}_2\left(\frac{\sqrt{u}}{24}\,,\mathcal{G}_n,\mathbf{X}_1^{2n}\right)\right]\,e^{-\frac{n\,u}{576\cdot 4\beta_n^2}}\,.
\end{aligned}$$

Lemma G1 implies that $\mathcal{N}_2\left(\frac{\sqrt{u}}{24}\,,\mathcal{G}_n,\mathbf{X}_1^{2n}\right) \leq \mathcal{M}_2\left(\frac{\sqrt{u}}{24}\,,\mathcal{G}_n,\mathbf{X}_1^{2n}\right)$. Similarly, from Theorem G2, it follows that $\mathcal{M}_2\left(\frac{\sqrt{u}}{24}\,,\mathcal{G}_n,\mathbf{X}_1^{2n}\right) \leq 3\,\left(\frac{2\,e\,4\beta_n^2}{\left(\frac{\sqrt{u}}{24}\right)^2}\log\left(\frac{3\,e\,4\beta_n^2}{\left(\frac{\sqrt{u}}{24}\right)^2}\right)\right)^{V_{\mathcal{G}_n^+}}$. Using the same argument as in the proof of Theorem 1, Theorem G3 implies that $V_{\mathcal{G}_n^+} \leq \mathcal{K}_n + 1$. Therefore:

$$\mathbb{P}\left\{S_{1,n} > u\right\} \leq 3\,\left(\frac{24^2\,12\,e\,\beta_n^2}{u}\right)^{2(\mathcal{K}_n+1)}\,e^{-\frac{n\,u}{576\cdot 4\beta_n^2}}\,.$$

Note that for $u > \frac{576\,\beta_n^2}{n}$, $\frac{24^2\,12\,e\,\beta_n^2}{u} \leq 12\,e\,n$; Therefore, it follows:

$$\mathbb{P}\left\{S_{1,n} > u\right\} \leq 3\,\left(12\,e\,n\right)^{2(\mathcal{K}_n+1)}\,e^{-\frac{n\,u}{576\cdot 4\beta_n^2}}\,.$$

Now, consider $\delta > 0$. For $u > \frac{576\,\beta_n^2}{n}$, $\mathbb{E}\left[S_{1,n}\right]$ can be bounded as follows:

$$
\begin{aligned}
\mathbb{E}\left[S_{1,n}\right] &\leq \int_0^\infty \mathbb{P}\left\{S_{1,n} > t\right\} dt \\
&\leq \delta + \int_\delta^\infty \mathbb{P}\left\{S_{1,n} > t\right\} dt \\
&\leq \delta + 3\,\left(12\,e\,n\right)^{2(\mathcal{K}_n+1)} \int_\delta^\infty e^{-\frac{n\,t}{576\cdot 4\beta_n^2}} dt \\
&\leq \delta + 3\,\left(12\,e\,n\right)^{2(\mathcal{K}_n+1)} \left(\frac{2304\,\beta_n^2}{n}\right) e^{-\frac{n\,\delta}{576\cdot 4\beta_n^2}}\,. \qquad \text{(C.37)}
\end{aligned}
$$

Observe that the rhs of (C.37) is continuous for $\delta > 0$. Therefore, it is possible to obtain a value of δ that minimizes the upper bound. In this context, it is possible to show that $\delta^* = \frac{2304\,\beta_n^2}{n} \log\left(9 \cdot \left(12\,e\,n\right)^{2(\mathcal{K}_n+1)}\right)$ is the aforementioned minimizer. Using this result, it follows:

$$
\mathbb{E}\left[S_{1,n}\right] \leq \frac{2304\,\beta_n^2}{n} \log\left(9 \cdot \left(12\,e\,n\right)^{2(\mathcal{K}_n+1)}\right) + \frac{2304\,\beta_n^2}{n}\,.
$$

After some algebra, the last expression takes the form:

$$
\mathbb{E}\left[S_{1,n}\right] \leq \frac{\tilde{C}\,\beta_n^2\,\mathcal{K}_n\left(\log(n)+1\right)}{n}\,, \qquad \text{(C.38)}
$$

for $\tilde{C} = 4608\log(12)$. This, together with (C.36) imply:

$$
\mathbb{E}\left\{||f_{J(n)} - f_A||^2\right\} \leq \frac{\tilde{c}\,\beta_n^2\,\mathcal{K}_n\left(\log(n)+1\right)}{n} + 8\,\inf_{f\in\mathcal{F}_n}\left\{\int_{[0,1]^p} |f(\mathbf{x}) - f_A(\mathbf{x})|^2\,\mu(d\mathbf{x})\right\} + \frac{8\sigma^2}{n}\mathcal{K}_n\,.
$$

Finally, from Lemma 2 it follows:

$$
\mathbb{E}\left\{||f_{J(n)} - f_A||^2\right\} \leq \tilde{C}\max\left\{\beta_n^2, \sigma^2\right\}\frac{p\,2^{J(n)}}{n}\left(\log(n)+1\right)\,, +8\,C_2^2\,p^2\,2^{-2(N+1)J(n)}, \qquad \text{(C.39)}
$$

which proves the desired result.

C.6 Proof of Lemma 4.3.4.

Suppose a model of the form (4.8). Assume ϵ is a sub-gaussian random variable independent of $\mathbf{X}$ such that $\mathbb{E}[\epsilon] = 0, \mathbb{E}[\epsilon^2] = 1, 0 < \sigma < \infty$. Let $\{Y_1, ..., Y_n\}$ be the response observations from the iid sample $\{(Y_i, \mathbf{X}_i)\}_{i=1}^n$ and suppose $||f_A||_\infty \leq L$.

Then, for $\beta_n = 4\sigma\sqrt{\log(n)}$ it follows:

$$\mathbb{P}\left\{\max\left\{Y_1, ..., Y_n\right\} > \beta_n\right\} = \mathcal{O}\left(\frac{1}{n}\right) . \tag{C.40}$$

Proof

Denote $Y_{(n)} = \max\left\{Y_1, ..., Y_n\right\}$. For some $\delta > 0$ it holds:

$$\begin{aligned}
\mathbb{P}\left\{Y_{(n)} > \beta_n\right\} &\leq \mathbb{P}\left\{\cup_{i=1}^n Y_i > \beta_n\right\} \\
&\leq \sum_{i=1}^n \mathbb{P}\left\{Y_i > \beta_n\right\} \\
&\leq n \int_{[0,1]^p} \mathbb{P}\left\{f_A(\mathbf{u}) + \sigma\,\epsilon > \beta_n \mid \mathbf{X} = \mathbf{u}\right\} h(\mathbf{u}) d\mathbf{u} \\
&\leq n \int_{[0,1]^p} \mathbb{P}\left\{|\epsilon| > \frac{\beta_n - L}{\sigma} \mid \mathbf{X} = \mathbf{u}\right\} h(\mathbf{u}) d\mathbf{u} \\
&\leq n\,\mathbb{P}\left\{|\epsilon| > \frac{\beta_n - L}{\sigma}\right\} .
\end{aligned}$$

Since ϵ is assumed to be sub-gaussian ($\mathbb{E}[\epsilon] = 0$, $\mathbb{E}[\epsilon^2] = 1$, $0 < \sigma < \infty$), we have that $\forall s \in \mathbb{R}$, $\mathbb{E}\left[e^{s\,\epsilon}\right] \leq e^{\frac{s^2}{2}}$. Consequently, it is possible to show that $\mathbb{P}\left\{|\epsilon| > t\right\} \leq 2\,e^{-\frac{t^2}{2}}$. Using this result in the last equation, it follows:

$$\mathbb{P}\left\{Y_{(n)} > \beta_n\right\} \leq 2\,n\,e^{-\frac{(\beta_n - L)^2}{2\sigma^2}} .$$

Suppose it is possible to choose β_n in such a way that $2\,n\,e^{-\frac{(\beta_n - L)^2}{2\sigma^2}} \leq \frac{1}{n}$. This implies that $Y_{(n)}$ it's bounded in probability. Under this setting, assuming that for n large enough $\sqrt{2}\,\sigma\,\sqrt{\log(n)} > L$, it follows:

$$\mathbb{P}\left\{\max\{Y_1, ..., Y_n\} > \sqrt{2}\,\sigma\,\sqrt{\log(n)}\right\} = \mathcal{O}\left(\frac{1}{n}\right)\,,$$

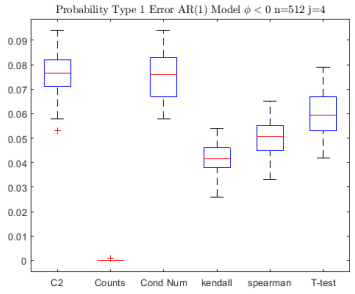
which shows (C.40) holds.

APPENDIX D

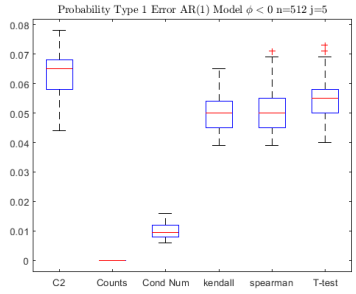
APPENDIX CHAPTER 6

D.1 Additional Results For Type I and II Error Simulation-Based Performance Studies

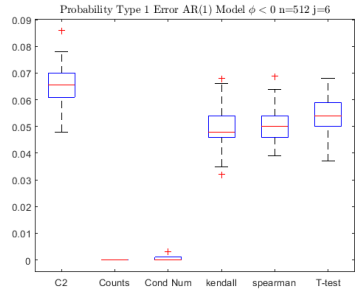
D.1.1 Box Plots for Type I Error Simulation Study



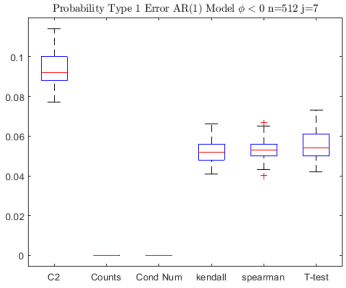
(a) Box plot of $\hat{p}_{Test_m}$ for $J = 4$



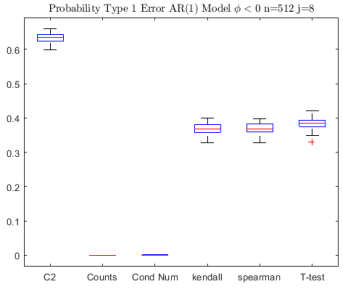
(b) Box plot of $\hat{p}_{Test_m}$ for $J = 5$



(c) Box plot of $\hat{p}_{Test_m}$ for $J = 6$

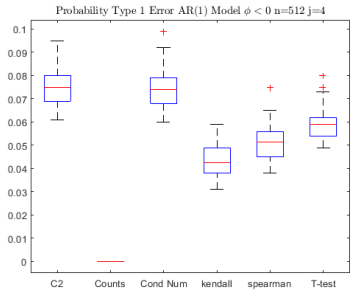


(d) Box plot of $\hat{p}_{Test_m}$ for $J = 7$

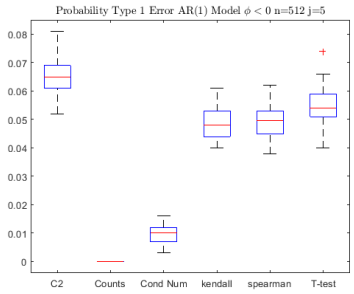


(e) Box plot of $\hat{p}_{Test_m}$ for $J = 8$

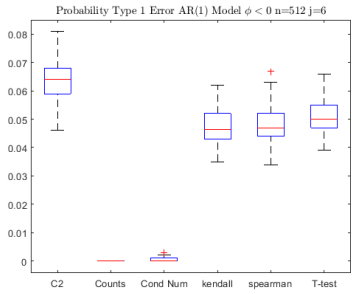
Figure D.1: Box plots of $\hat{p}_{Test_m}$ for AR(1) with parameter $\phi = -0.9$. For scale levels $J = 4$ to $J = 7$, most of tests remains within the expected 5% type I error; however, for $J = 8$ the tests $C2$, Kendall, Spearman's and T-test exhibit a significantly large deviation from the expected error, with an average of approximately 38%. This implies that on average, for this kind of stochastic processes, wavelet coefficient corresponding to short time scales depart from normality and exhibit heavier tails, which causes an artificial inflation of the likelihood of a false positive classification.



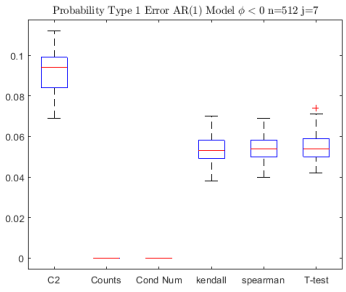
(a) Box plot of $\hat{p}_{Test_m}$ for $J = 4$



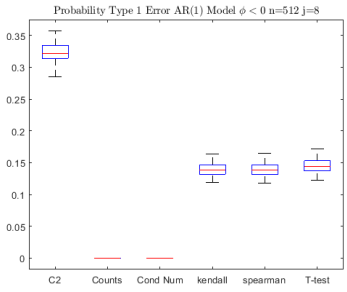
(b) Box plot of $\hat{p}_{Test_m}$ for $J = 5$



(c) Box plot of $\hat{p}_{Test_m}$ for $J = 6$

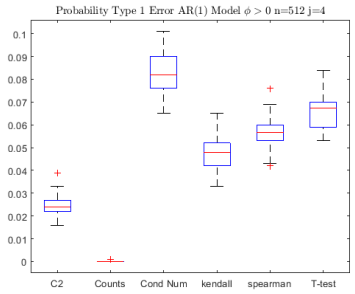


(d) Box plot of $\hat{p}_{Test_m}$ for $J = 7$

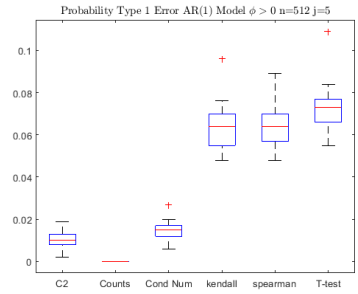


(e) Box plot of $\hat{p}_{Test_m}$ for $J = 8$

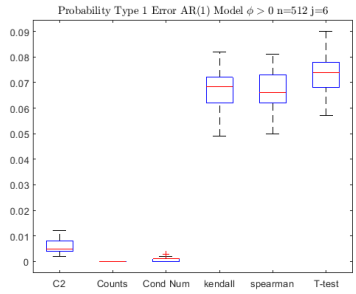
Figure D.2: Box plots of $\hat{p}_{Test_m}$ for AR(1) with parameter $\phi = -0.7$. For scale levels $J = 4$ to $J = 7$, most of tests remains within the expected 5% type I error; however, for $J = 8$ the tests $C2$, Kendall, Spearman's and T-test exhibit a significantly large deviation from the expected error, with an average of approximately 14%. This implies that on average, for this kind of stochastic processes, wavelet coefficient corresponding to short time scales depart from normality and exhibit heavier tails, which causes an artificial inflation of the likelihood of a false positive classification.



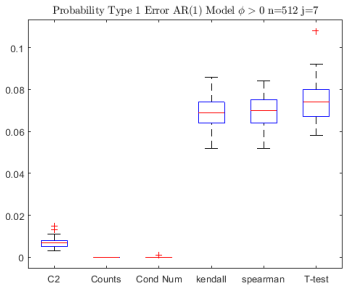
(a) Box plot of $\hat{p}_{Test_m}$ for $J = 4$



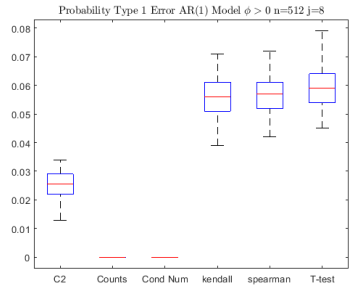
(b) Box plot of $\hat{p}_{Test_m}$ for $J = 5$



(c) Box plot of $\hat{p}_{Test_m}$ for $J = 6$

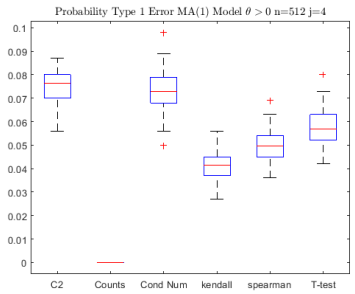


(d) Box plot of $\hat{p}_{Test_m}$ for $J = 7$

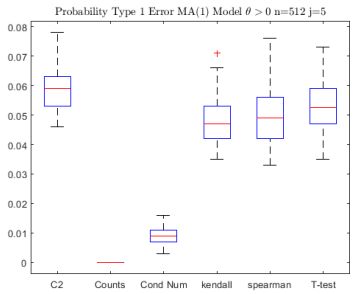


(e) Box plot of $\hat{p}_{Test_m}$ for $J = 8$

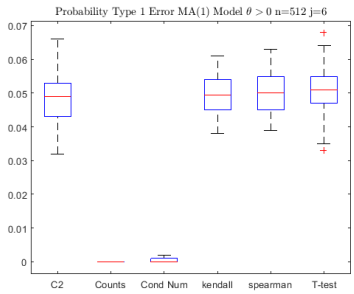
Figure D.3: Box plots of $\hat{p}_{Test_m}$ for AR(1) with parameter $\phi = 0.9$. For all scale levels most of tests remains within the expected 5% type I error.



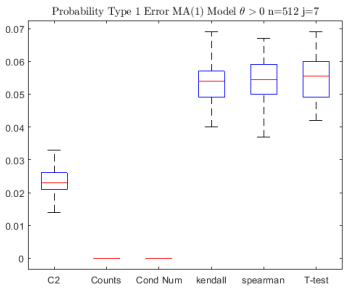
(a) Box plot of $\hat{p}_{Test_m}$ for $J = 4$



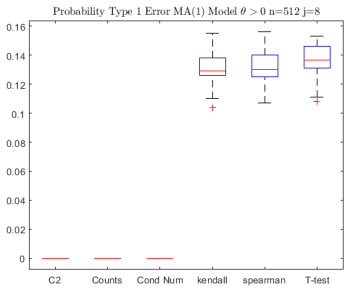
(b) Box plot of $\hat{p}_{Test_m}$ for $J = 5$



(c) Box plot of $\hat{p}_{Test_m}$ for $J = 6$

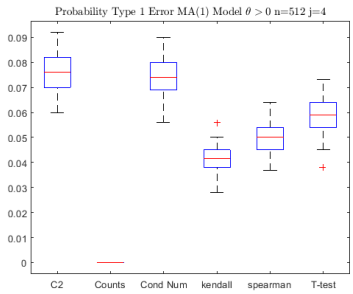


(d) Box plot of $\hat{p}_{Test_m}$ for $J = 7$

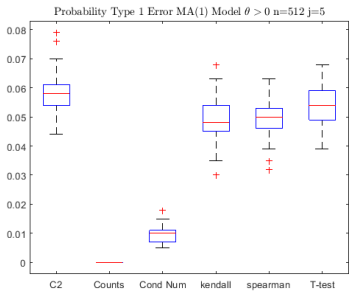


(e) Box plot of $\hat{p}_{Test_m}$ for $J = 8$

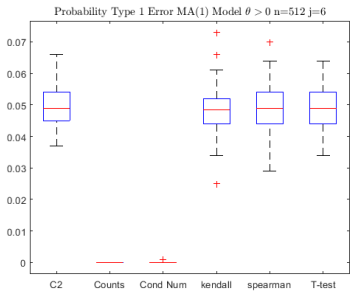
Figure D.4: Box plots of $\hat{p}_{Test_m}$ for MA(1) with parameter $\theta = 0.9$. For scale levels $J = 4$ to $J = 7$, most of tests remains within the expected 5% type I error. However, for $J = 8$, the tests $C2$, Kendall, Spearman's and T-test exhibit a significantly large deviation from the expected error, with an average of approximately 13%.



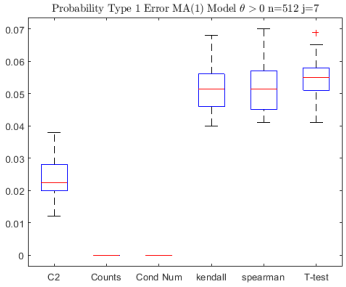
(a) Box plot of $\hat{p}_{Test_m}$ for $J = 4$



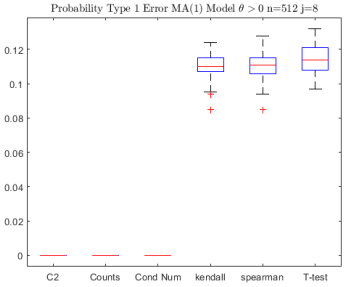
(b) Box plot of $\hat{p}_{Test_m}$ for $J = 5$



(c) Box plot of $\hat{p}_{Test_m}$ for $J = 6$

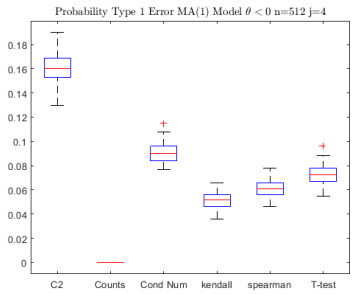


(d) Box plot of $\hat{p}_{Test_m}$ for $J = 7$

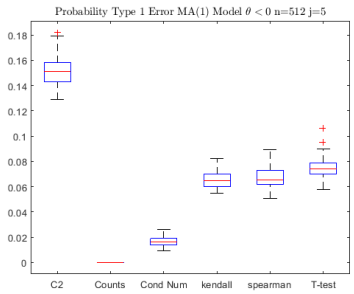


(e) Box plot of $\hat{p}_{Test_m}$ for $J = 8$

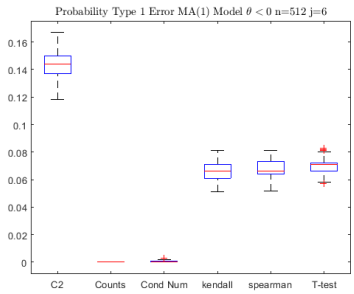
Figure D.5: Box plots of $\hat{p}_{Test_m}$ for MA(1) with parameter $\theta = 0.7$. For scale levels $J = 4$ to $J = 7$, most of tests remains within the expected 5% type I error. However, for $J = 8$, the tests $C2$, Kendall, Spearman's and T-test exhibit a significantly large deviation from the expected error, with an average of approximately 11%.



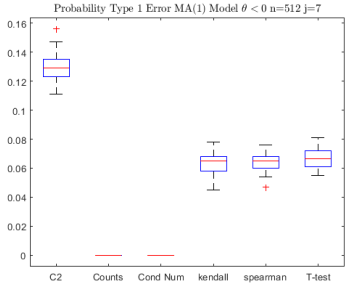
(a) Box plot of $\hat{p}_{Test_m}$ for $J = 4$



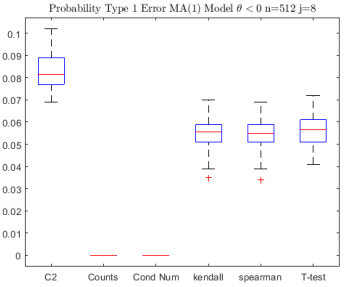
(b) Box plot of $\hat{p}_{Test_m}$ for $J = 5$



(c) Box plot of $\hat{p}_{Test_m}$ for $J = 6$

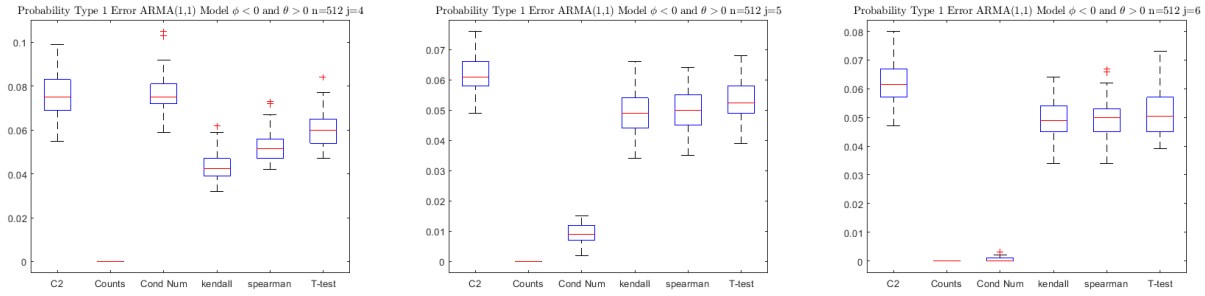


(d) Box plot of $\hat{p}_{Test_m}$ for $J = 7$

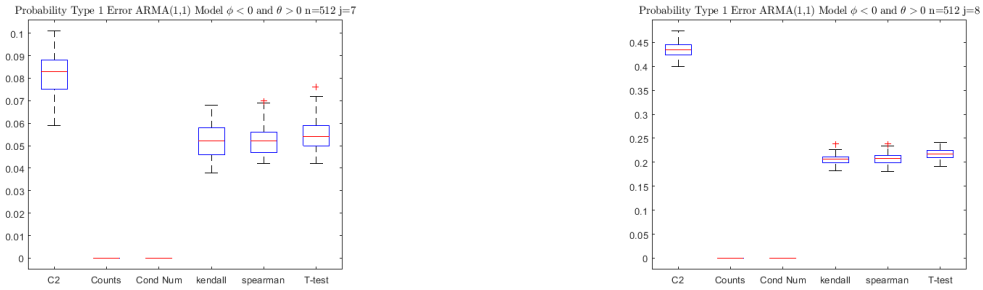


(e) Box plot of $\hat{p}_{Test_m}$ for $J = 8$

Figure D.6: Box plots of $\hat{p}_{Test_m}$ for MA(1) with parameter $\theta = -0.9$. For all scale levels, most of tests remains within the expected 5% type I error. In particular, the tests $C2$, Kendall, Spearman's and T-test exhibit a slight deviation from the expected error, with an average of approximately 6.5%.

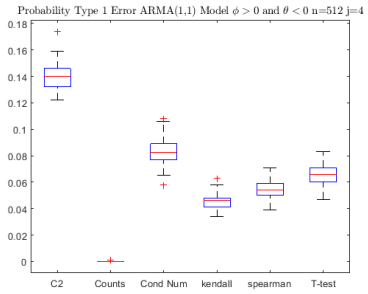


(a) Box plot of $\hat{p}_{Test_m}$ for $J = 4$ (b) Box plot of $\hat{p}_{Test_m}$ for $J = 5$ (c) Box plot of $\hat{p}_{Test_m}$ for $J = 6$

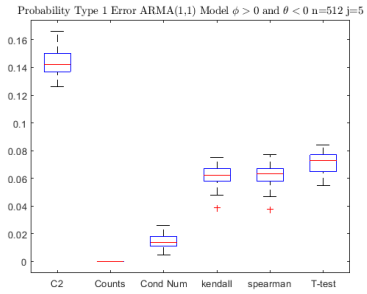


(d) Box plot of $\hat{p}_{Test_m}$ for $J = 7$ (e) Box plot of $\hat{p}_{Test_m}$ for $J = 8$

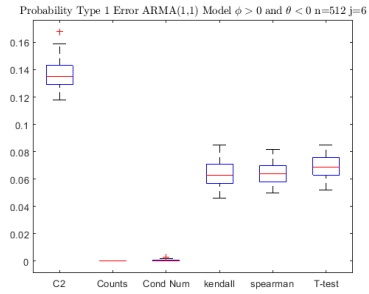
Figure D.7: Box plots of $\hat{p}_{Test_m}$ for ARMA(1,1) with parameters $\phi = -0.8, \theta = 0.1$. For scale levels $J = 4$ to $J = 7$, most of tests remains within the expected 5% type I error; however, for $J = 8$ the tests $C2$, Kendall, Spearman's and T-test exhibit a significantly large deviation from the expected error. This implies that on average, for this kind of stochastic processes, wavelet coefficient corresponding to short time scales depart from normality, which causes an artificial inflation of the likelihood of a false positive classification.



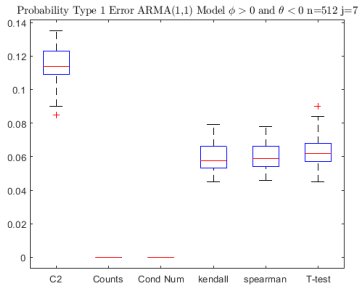
(a) Box plot of $\hat{p}_{Test_m}$ for $J = 4$



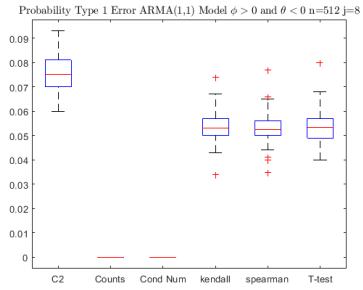
(b) Box plot of $\hat{p}_{Test_m}$ for $J = 5$



(c) Box plot of $\hat{p}_{Test_m}$ for $J = 6$



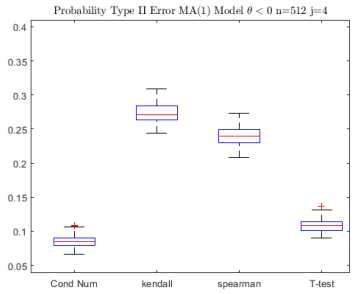
(d) Box plot of $\hat{p}_{Test_m}$ for $J = 7$



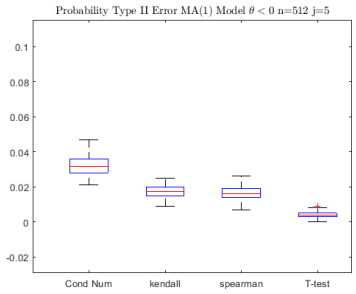
(e) Box plot of $\hat{p}_{Test_m}$ for $J = 8$

Figure D.8: Box plots of $\hat{p}_{Test_m}$ for ARMA(1,1) with parameters $\phi = -0.9$, $\theta = 0.9$. For scale levels $J = 4$ to $J = 7$, most of tests remains within the expected 5% type I error.

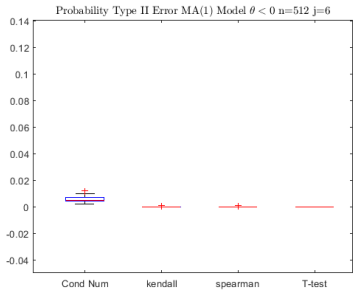
D.1.2 Box Plots for Type II Error Simulation Study



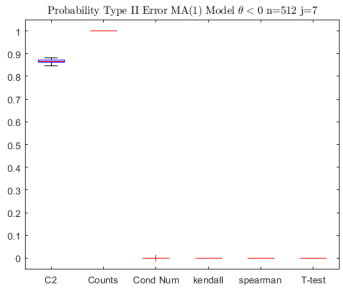
(a) Box plot of $\hat{p}_{Test_m}$ for $J = 4$



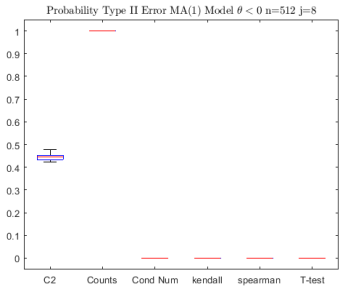
(b) Box plot of $\hat{p}_{Test_m}$ for $J = 5$



(c) Box plot of $\hat{p}_{Test_m}$ for $J = 6$

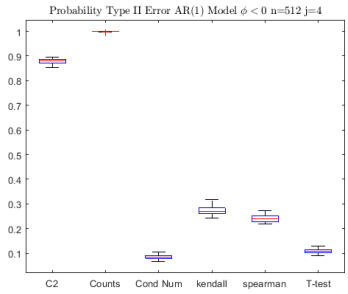


(d) Box plot of $\hat{p}_{Test_m}$ for $J = 7$

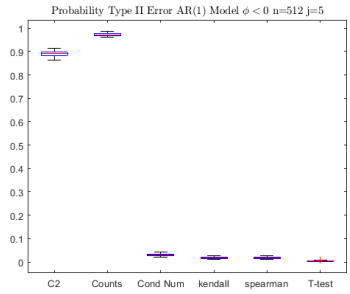


(e) Box plot of $\hat{p}_{Test_m}$ for $J = 8$

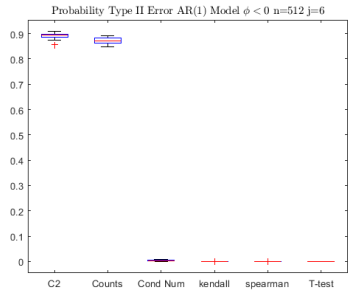
Figure D.9: Box plots of $\hat{p}_{Test_m}$ (average probability of type II error) for AR(1) with parameter $\phi = -0.9$. For scale levels $J = 5$ to $J = 8$, most of tests remains within the 5% average type II error; however, for $J = 4$ the tests Kendall and Spearman's exhibit a significantly large deviation from the expected error, with an average of approximately 24%. Also, the observed performance of the Condition number test can be considered as good as the statistical tests used as benchmark.



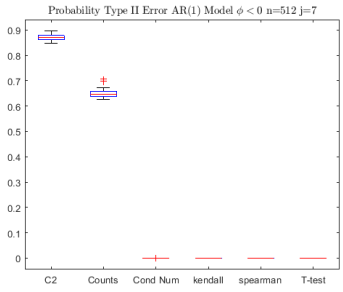
(a) Box plot of $\hat{p}_{Test_m}$ for $J = 4$



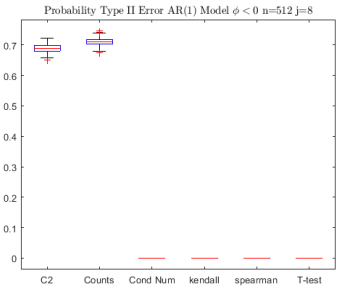
(b) Box plot of $\hat{p}_{Test_m}$ for $J = 5$



(c) Box plot of $\hat{p}_{Test_m}$ for $J = 6$

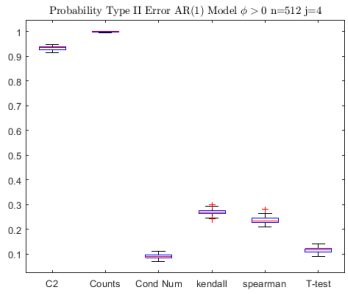


(d) Box plot of $\hat{p}_{Test_m}$ for $J = 7$

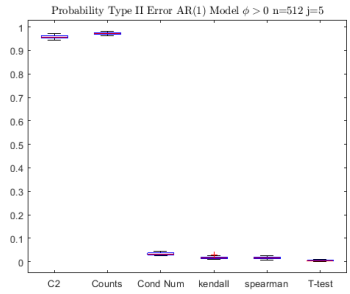


(e) Box plot of $\hat{p}_{Test_m}$ for $J = 8$

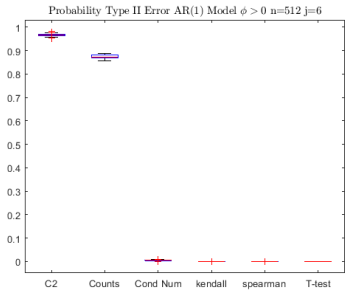
Figure D.10: Box plots of $\hat{p}_{Test_m}$ (average probability of type II error) for AR(1) with parameter $\phi = -0.7$. For scale levels $J = 5$ to $J = 8$, most of tests remains within the 5% average type II error; however, for $J = 4$ the tests Kendall and Spearman's exhibit a significantly large deviation from the expected error, with an average of approximately 24%. Also, the observed performance of the Condition number test can be considered as good as the statistical tests used as benchmark.



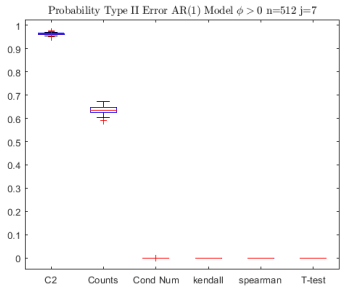
(a) Box plot of $\hat{p}_{Test_m}$ for $J = 4$



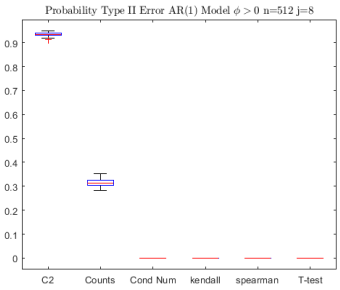
(b) Box plot of $\hat{p}_{Test_m}$ for $J = 5$



(c) Box plot of $\hat{p}_{Test_m}$ for $J = 6$

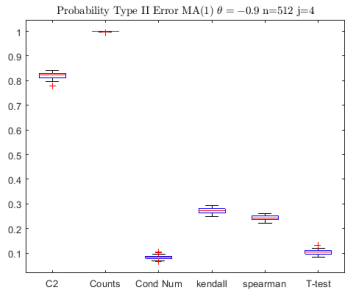


(d) Box plot of $\hat{p}_{Test_m}$ for $J = 7$

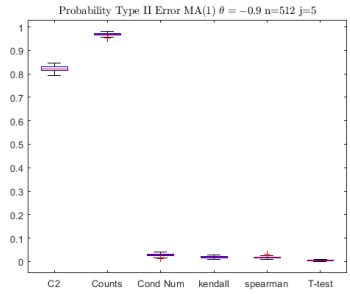


(e) Box plot of $\hat{p}_{Test_m}$ for $J = 8$

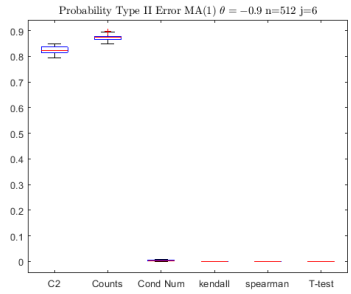
Figure D.11: Box plots of $\hat{p}_{Test_m}$ (average probability of type II error) for AR(1) with parameter $\phi = 0.9$. For scale levels $J = 5$ to $J = 8$, most of tests remains within the 5% average type II error; however, for $J = 4$ the tests Kendall and Spearman's exhibit a significantly large deviation from the expected error, with an average of approximately 24%. Also, the observed performance of the Condition number test can be considered as good as the statistical tests used as benchmark.



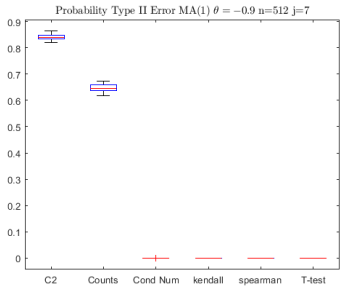
(a) Box plot of $\hat{p}_{Test_m}$ for $J = 4$



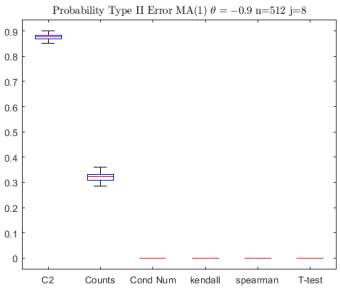
(b) Box plot of $\hat{p}_{Test_m}$ for $J = 5$



(c) Box plot of $\hat{p}_{Test_m}$ for $J = 6$

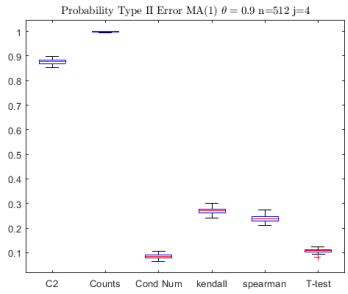


(d) Box plot of $\hat{p}_{Test_m}$ for $J = 7$

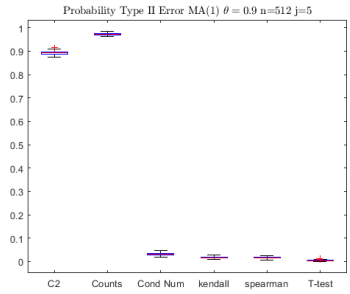


(e) Box plot of $\hat{p}_{Test_m}$ for $J = 8$

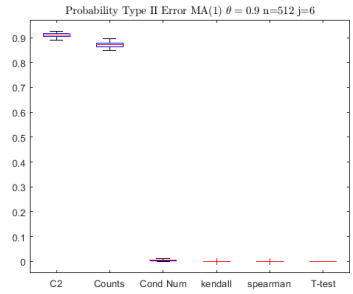
Figure D.12: Box plots of $\hat{p}_{Test_m}$ (average probability of type II error) for MA(1) with parameter $\theta = -0.9$. For scale levels $J = 5$ to $J = 8$, most of tests remains within the 5% average type II error; however, for $J = 4$ the tests Kendall and Spearman's exhibit a significantly large deviation from the expected error, with an average of approximately 24%. Also, the observed performance of the Condition number test can be considered as good as the statistical tests used as benchmark.



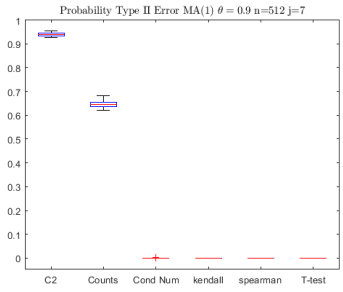
(a) Box plot of $\hat{p}_{Test_m}$ for $J = 4$



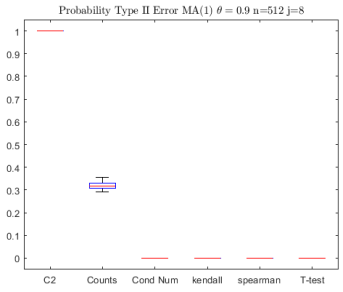
(b) Box plot of $\hat{p}_{Test_m}$ for $J = 5$



(c) Box plot of $\hat{p}_{Test_m}$ for $J = 6$

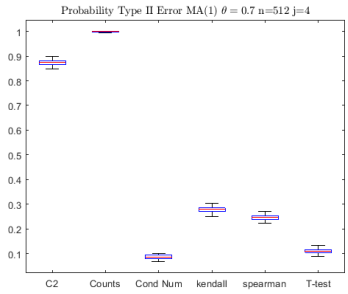


(d) Box plot of $\hat{p}_{Test_m}$ for $J = 7$

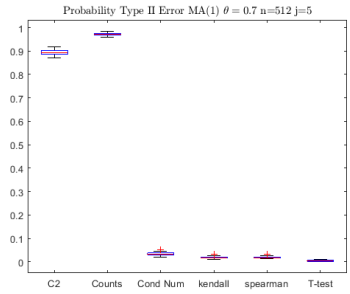


(e) Box plot of $\hat{p}_{Test_m}$ for $J = 8$

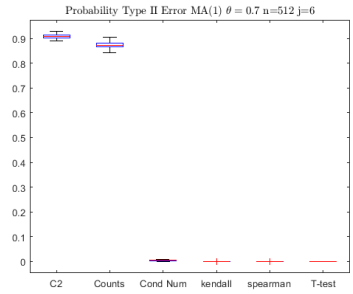
Figure D.13: Box plots of $\hat{p}_{Test_m}$ (average probability of type II error) for MA(1) with parameter $\theta = 0.9$. For scale levels $J = 5$ to $J = 8$, most of tests remains within the 5% average type II error; however, for $J = 4$ the tests Kendall and Spearman's exhibit a significantly large deviation from the expected error, with an average of approximately 24%. Also, the observed performance of the Condition number test can be considered as good as the statistical tests used as benchmark.



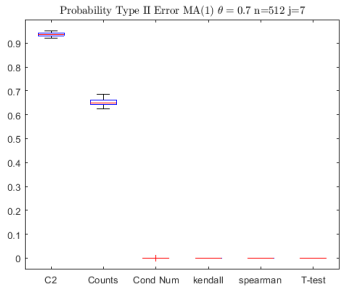
(a) Box plot of $\hat{p}_{Test_m}$ for $J = 4$



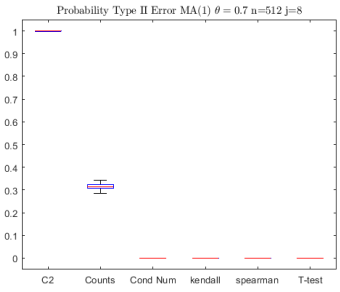
(b) Box plot of $\hat{p}_{Test_m}$ for $J = 5$



(c) Box plot of $\hat{p}_{Test_m}$ for $J = 6$

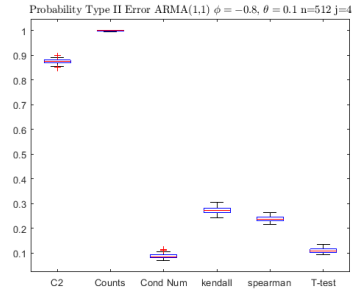


(d) Box plot of $\hat{p}_{Test_m}$ for $J = 7$

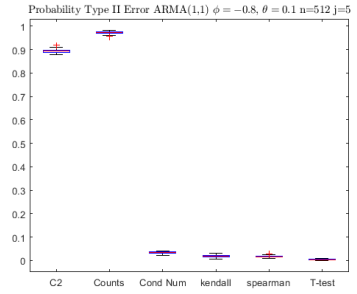


(e) Box plot of $\hat{p}_{Test_m}$ for $J = 8$

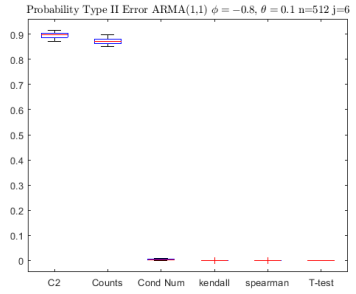
Figure D.14: Box plots of $\hat{p}_{Test_m}$ (average probability of type II error) for MA(1) with parameter $\theta = 0.7$. For scale levels $J = 5$ to $J = 8$, most of tests remains within the 5% average type II error; however, for $J = 4$ the tests Kendall and Spearman's exhibit a significantly large deviation from the expected error, with an average of approximately 24%. Also, the observed performance of the Condition number test can be considered as good as the statistical tests used as benchmark.



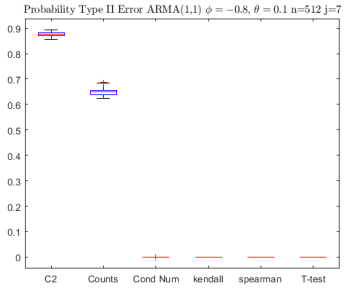
(a) Box plot of $\hat{p}_{Test_m}$ for $J = 4$



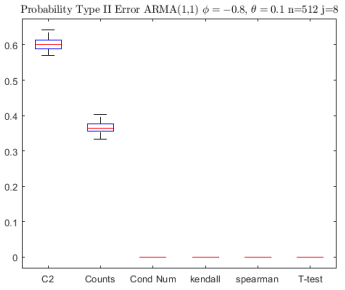
(b) Box plot of $\hat{p}_{Test_m}$ for $J = 5$



(c) Box plot of $\hat{p}_{Test_m}$ for $J = 6$

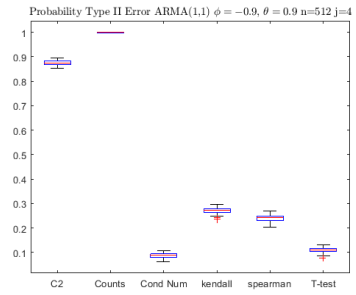


(d) Box plot of $\hat{p}_{Test_m}$ for $J = 7$

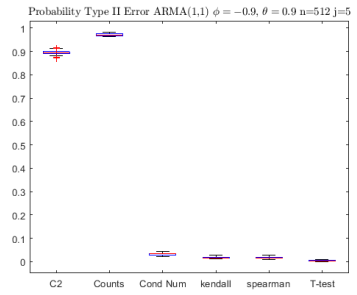


(e) Box plot of $\hat{p}_{Test_m}$ for $J = 8$

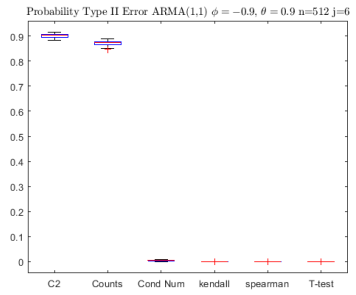
Figure D.15: Box plots of $\hat{p}_{Test_m}$ (average probability of type II error) for ARMA(1) with parameters $\phi = -0.8$ $\theta = 0.1$. Induced linear relationship given by $\beta = 0.25$.



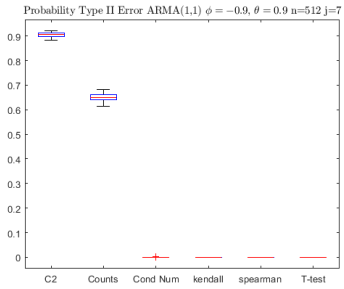
(a) Box plot of $\hat{p}_{Test_m}$ for $J = 4$



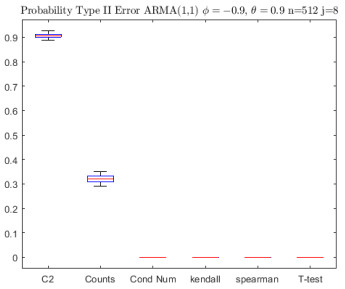
(b) Box plot of $\hat{p}_{Test_m}$ for $J = 5$



(c) Box plot of $\hat{p}_{Test_m}$ for $J = 6$



(d) Box plot of $\hat{p}_{Test_m}$ for $J = 7$



(e) Box plot of $\hat{p}_{Test_m}$ for $J = 8$

Figure D.16: Box plots of $\hat{p}_{Test_m}$ (average probability of type II error) for ARMA(1) with parameters $\phi = -0.9$ $\theta = 0.9$. Induced linear relationship given by $\beta = 0.25$.

REFERENCES

[1] S. Sardy and P. Tseng, "Amlet, ramlet, and gamlet: Automatic nonlinear fitting of additive models, robust and generalized, with wavelets", *Journal of Computational and Graphical Statistics*, 2004.

[2] D. L. Donoho, R. C. Liu, and B. MacGibbon, "Minimax risk over hyperrectangles, and implications", *Ann. Statist.*, vol. 18, no. 3, pp. 1416–1437, Sep. 1990.

[3] B. Vidakovic, *Statistical Modeling by Wavelets*. Wiley, New York, 1999.

[4] A. Antoniadis, G. Gregoire, and P. Vial, "Random designs wavelet curve smoothing", *Statistics and Probability Letters*, 1997.

[5] I. Daubechies, "Ten lectures on wavelets", *CBMS-NSF regional conferences series in applied mathematics*, 1992.

[6] J. Morlet, G. Arens, E. Fourgeau, and D. Giard, "Wave propagation and sampling theory; part i, complex signal and scattering in multilayered media", *Geophysics*, vol. 47, no. 2, p. 203, 1982. eprint: `/gsw/content_public/journal/geophysics/47/2/10.1190_1.1441328/5/203.pdf`.

[7] A. Grossmann, J. Morlet, and T. Paul, "Transforms associated to square integrable group representations. I. General results", *Journal of Mathematical Physics*, vol. 26, pp. 2473–2479, Oct. 1985.

[8] I. Daubechies, "Orthonormal bases of compactly supported wavelets", *Communications on Pure and Applied Mathematics*, vol. 41, no. 7, pp. 909–996, 1988. eprint: `https://onlinelibrary.wiley.com/doi/pdf/10.1002/cpa.3160410705`.

[9] L. Györfi, M. Kohler, A. Krzyzak, and H. Walk, *A Distribution-Free Theory of Nonparametric Regression*. Springer, 2002.

[10] S. Jaffard, "Pointwise smoothness, two-microlocalization and wavelet coefficients", vol. 35, pp. 155–168, 1991, Exported from https://app.dimensions.ai on 2019/01/04.

[11] S. Jaffard and P. Laurençot, "Wavelets: A tutorial in theory and applications", in, C. K. Chui, Ed., San Diego, CA, USA: Academic Press Professional, Inc., 1992, ch. Or-

thonormal Wavelets, Analysis of Operators, and Applications to Numerical Analysis, pp. 543–601, ISBN: 0-12-174590-2.

[12] I. Daubechies and J. Lagarias, "Two-scale difference equations. i. existence and global regularity of solutions", *SIAM Journal on Mathematical Analysis*, vol. 22, no. 5, pp. 1388–1410, 1991. eprint: https://doi.org/10.1137/0522089.

[13] S. G. Mallat, "A theory for multiresolution signal decomposition: The wavelet representation", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 11, no. 7, pp. 674–693, 1989.

[14] W. Stute, "The central limit theorem under random censorship", *The Annals of Statistics*, vol. 23, pp. 422–439, 1995.

[15] L. Deroye and L. Györfi, *Nonparametric Density Estimation*. John Wiley & Sons, 1985.

[16] E. Parzen, "On estimation of a probability density function and mode", *The Annals of Statistics*, vol. 33, pp. 1065–1073, 1962.

[17] M. Rossenblat, "Remarks on some nonparametric estimates of a density function", *The Annals of Mathematical Statistics*, vol. 27, pp. 832–837, 1956.

[18] A. Antoniadis, "Wavelets in statistics: A review", University of Joseph Fourier, Laboratorie IMAG-LMC, 38041 Grenoble Cedex 9, France, Tech. Rep., 1997.

[19] S. Efromovich, *Nonparametric Curve Estimation, Methods, Theory and Applications*, First, ser. Springer Series in Statistics. Springer, 1999.

[20] G. G. Antoniadis A. and G. Nason, "Density and hazard rate estimation for right-censored data by using wavelets methods.", *J. Roy. Statist. Soc.*, vol. 61, pp. 63–84, 1999.

[21] P. Doukhan and J. R. León, "Déviation quadratique destimateurs de densité par projections orthogonales", *Comptes Rendus de lAcadémie des Sciences-Series I-Mathematics*, vol. 310, pp. 425–430, 1990.

[22] A. Antoniadis and R. Carmona, "Multiresolution analyses and wavelets for density estimation", Tech. Rep., 1991.

[23] G. Kerkyacharian and D. Picard, "Density estimation in besov spaces", *Statistics Probability Letters*, vol. 13, no. 1, pp. 15–24, 1992.

[24] G. G. Walter, “A sampling theorem for wavelet subspaces”, *IEEE Transactions on Information Theory*, vol. 38, no. 2, pp. 881–884, 1992.

[25] Y. Meyer, *Wavelets and Operators*. Cambridge: Cambridge University Press, 1992.

[26] D. Donoho, “Nonlinear wavelet methods for recovery of signals, densities and spectra from indirect and noisy data”, *Proceedings of Symposia in Applied Mathematics*, vol. 47, pp. 173–205, 1993.

[27] D. Donoho, I. Johnstone, G. Kerkyacharian, and D. Picard, “Wavelet shrinkage: Asymptopia?”, *Journal of Royal Statistical Society*, 1995.

[28] D. D., J. I.M., K. G., and P. D., “Density estimation by wavelets thresholding”, *The Annals of Statistics*, vol. 2, pp. 508–539, 1996.

[29] A. Pinheiro and B. Vidakovic, “Estimating the square root of a density via compactly supported wavelets”, *Computational Statistic and Data Analysis*, vol. 25, pp. 399–415, 1997.

[30] M. Vanucci, “Nonparametric density estimation using wavelets”, Department of Statistics, Texas A&M University, Duke University, U.S.A., Discussion paper 95-26, 1998.

[31] L. Li, “Non-linear wavelet-based density estimator under random censorship”, *Journal of Statistical planning and Inference*, vol. 117, pp. 35–58, 2003.

[32] ——, “On the minimax optimality of wavelet estimators with censored data”, *Journal of Statistical planning and Inference*, vol. 137, pp. 1138–1150, 2007.

[33] Y.-Y. Zou and H.-Y. Liang, “Wavelet estimation of density for censored data with censoring indicator missing at random”, *A Journal of Theoretical and Applied Statistics*, 2017.

[34] D. L. Donoho and I. M. Johnstone, “Ideal spatial adaptation by wavelet shrinkage”, *Biometrika*, vol. 81, no. 3, pp. 425–455, 1994. eprint: `/oup/backfile/content_public/journal/biomet/81/3/10.1093_biomet_81.3.425/1/81.3.425.pdf`.

[35] J. Restrepo, G. Leaf, and G. Schlossnagle, “Periodized daubechies wavelets”, Mathematics and Computer Science Division, Argonne, National Laboratory, Argonne, IL 60439, U.S.A., Tech. Rep., 1996.

[36] A. Antoniadis, J. Bigot, and T. Sapatinas, "Wavelet estimators in nonparametric regression: A comparative simulation study", *Journal of Statistical Software, Articles*, vol. 6, no. 6, pp. 1–83, 2001.

[37] W. Stute, "Strong and weak representation of cumulative hazard function and kaplan-meier estimator on increasing sets", *Journal of Statistical Planning and Inference*, vol. 42, pp. 315–329, 1994.

[38] A. DasGupta, *Asymptotic Theory of Probability and Statistics*, I. O. G. Casella S. Fienberg, Ed. Springer, 2008.

[39] J. Ramsay and B. Silverman, *Functional Data Analysis*. Springer, 2005.

[40] E. Mammen and J. Nielsen, "Generalised structured models.", *Biometrika*, 2003.

[41] A. Buja, T. Hastie, and R. Tibshirani, "Linear smoothers and additive models", *The Annal of Statistics*, 1989.

[42] T. Hastie and R. Tibshirani, *Generalized Additve Models*. John Wiley and sons, 1990.

[43] D. Tjostheim and B. Auestad, "Nonparametric identification of nonlinear time series", *Journal of the American Statistical Association*, 1994.

[44] J. P. Nielsen and O. B. Linton, "Kernel estimation in a nonparametric marker dependent hazard model", *The Annals of Statistics*, 1995.

[45] C. R., H. W. L. O.B., and S.-L. E., "Nonparametric estimation of additive separable regression models.", *Statistical Theory and Computational Aspects of Smoothing. Contributions to Statistics.*, 1996.

[46] S. Sperlich, O. Linton, and W. Härdle, "Integration and backfitting methods in additive models-finite sample properties and comparison.", *Test*, vol. 8, pp. 419–458, 1999.

[47] A. Buja, T. Hastie, and R. Tibshirani, "Linear smoothers and additive models (with discussion).", *The Annals of Statistics*, 1989.

[48] T. Hastie and R. Tibshirani, *Generalized Additive Models*. Chapman and Hall, 1990.

[49] J. Opsomer and D. Rupert, "Fitting a bivariate additive model by local polynomial regression", *The Annals of Statistics*, 1997.

[50] E. Mammen, O. Linton, and J. Nielsen, "The existence and asymptotic properties of a backfitting projection algorithm under weak conditions", *The Annals of Statistics*, 1999.

[51] E. Mammen and B. Park, "A simple smooth backfitting method for additive models", *The Annals of Statistics*, 2006.

[52] M. Pensky and B. Vidakovic, "On non-equally spaced wavelet regression", *The Annals of Statistics*, 2001.

[53] A. Antoniadis, G. Gregoire, and I. W. McKeague, "Wavelet methods for curve estimation", *Journal of the American Statistical Association*, vol. 89, no. 428, pp. 1340–1353, 1994.

[54] U. Amato and A. Antoniadis, "Adaptive wavelet series estimation in separable nonparametric regression models", *Statistics and Computing*, 2001.

[55] D. Donoho, I. Johnstone, J. Hoch, and J. Stern, "Maximum entropy and the nearly-black object", *Journal of the Royal Statistical Society*, 1992.

[56] S. Zhang and M.-Y. Wong, "Wavelet threshold estimation for additive regression models", *The Annals of Statistics*, vol. 31, no. 1, pp. 152–173, Feb. 2003.

[57] W. Hardle, G. Kerkyacharian, D. Picard, and A. Tsybakov, *Wavelets, Approximation, and Statistical Applications*. Springer, 1998.

[58] P. Tüfekci, "Prediction of full load electrical power output of a base load operated combined cycle power plant using machine learning methods", *International Journal of Electrical Power & Energy Systems*, vol. Volume 60, pp. 126–140, Sep. 2014.

[59] D. D. Canditiis and B. Vidakovic, "Wavelet bayesian block shrinkage via mixtures of normal inverse gamma priors", *Technical Report, Georgia Institute of Technology*, 2001.

[60] B. Vidakovic and F. Ruggeri, "Bams method: Theory and simulations", *Sankhy: The Indian Journal of Statistics, Series B (1960-2002)*, vol. 63, no. 2, pp. 234–249, 2001.

[61] P. Hall, G. Kerkyacharian, and D. Picard, "Block threshold rules for curve estimation using kernel and wavelet methods", *Ann. Statist.*, vol. 26, no. 3, pp. 922–942, Jun. 1998.

[62] H. Chipman, E. Kolaczyk, and R. McCulloch, "Adaptive bayesian wavelet shrinkage", *Journal of the American Statistical Association*, vol. 92, no. 440, pp. 1413–1421, Dec. 1997.

[63] A. Brezger and S. Lang, "Generalized structured additive regression based on bayesian p-splines", *Computational Statistics Data Analysis*, vol. 50, no. 4, pp. 967–991, 2006.

[64] L. Fahrmeir, T. Kneib, and S. Lang, "Penalized structured additive regression for space-time data: A bayesian perspective", *Statistica Sinica*, vol. 14, no. 3, pp. 731–761, 2004.

[65] B. Vidakovic and P. Müller, "Wavelet shrinkage with affine bayes rules with applications", *Institute of Statistics and Decision Sciences*, vol. 34, no. 95, pp. 152–173, 1995.

[66] M. Vannucci and F. Corradi, "Covariance structure of wavelet coefficients: Theory and models in a bayesian perspective", *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 61, no. 4, pp. 971–986, eprint: https://rss.onlinelibrary.wiley.com/doi/pdf/10.1111/1467-9868.00214.

[67] A. Zellner, "Bayesian estimation and prediction using asymmetric loss functions", *Journal of the American Statistical Association*, vol. 81, no. 394, pp. 446–451, 1986. eprint: https://www.tandfonline.com/doi/pdf/10.1080/01621459.1986.10478289.

[68] R. Gençay, F. Selçuk, and B. Whitcher, "5 - wavelets and stationary processes", in *An Introduction to Wavelets and Other Filtering Methods in Finance and Economics*, R. Gençay, F. Selçuk, and B. Whitcher, Eds., San Diego: Academic Press, 2002, pp. 161–201, ISBN: 978-0-12-279670-8.

[69] D. B. Percival and A. T. Walden, *Wavelet Methods for Time Series Analysis*, ser. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2000.

[70] A. Grinsted, J. C. Moore, and S. Jevrejeva, "Application of the cross wavelet transform and wavelet coherence to geophysical time series", *Nonlinear Processes in Geophysics*, vol. 11, no. 5/6, pp. 561–566, 2004.

[71] E. Capobianco, "Multiscale analysis of stock index return volatility", *Computational Economics*, vol. 23, no. 3, pp. 219–237, 2004.

[72] J. Fernández-Macho, "Wavelet multiple correlation and cross-correlation: A multiscale analysis of Eurozone stock markets", *Physica A: Statistical Mechanics and its Applications*, vol. 391, no. 4, pp. 1097–1104, 2012.

[73] F. Benhmad, "Bull or bear markets: A wavelet dynamic correlation perspective", *Economic Modelling*, vol. 32, pp. 576 –591, 2013.

[74] L. H. Hudgins, "Wavelet Analysis of Atmospheric Turbulence", PhD thesis, UNIVERSITY OF CALIFORNIA, IRVINE., 1992.

[75] P. C. Liu, "Wavelet spectrum analysis and ocean wind waves", in *Wavelets in Geophysics*, ser. Wavelet Analysis and Its Applications, E. Foufoula-Georgiou and P. Kumar, Eds., vol. 4, Academic Press, 1994, pp. 151 –166.

[76] R. W. Lindsay, D. B. Percival, and D. A. Rothrock, "The discrete wavelet transform and the scale analysis of the surface properties of sea ice", *IEEE Transactions on Geoscience and Remote Sensing*, vol. 34, no. 3, pp. 771–787, 1996.

[77] B. Whitcher, P. Guttorp, and D. B. Percival, "Wavelet analysis of covariance with application to atmospheric time series", *Journal of Geophysical Research: Atmospheres*, vol. 105, no. D11, pp. 14 941–14 962, 2000. eprint: `https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2000JD900110`.

[78] T. Pering, G. Tamburello, A. McGonigle, E. Hanna, and A. Aiuppa, "Correlation of oscillatory behaviour in matlab using wavelets", *Computers & Geosciences*, vol. 70, pp. 206 –212, 2014, © 2014 Elsevier Ltd. This is an author produced version of a paper subsequently published in Computers & Geosciences. Uploaded in accordance with the publisher's self-archiving policy.

[79] C. Spearman, "The proof and measurement of association between two things", *The American Journal of Psychology*, vol. 15, no. 1, pp. 72–101, 1904.

[80] E. Casagrande, B. Mueller, D. G. Miralles, D. Entekhabi, and A. Molini, "Wavelet correlations to reveal multiscale coupling in geophysical systems", *Journal of Geophysical Research: Atmospheres*, vol. 120, no. 15, pp. 7555–7572, 2015. eprint: `https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1002/2015JD023265`.

[81] E. C. FIELLER, H. O. HARTLEY, and E. S. PEARSON, "Tests for rank correlation coefficients. i", *Biometrika*, vol. 44, no. 3-4, pp. 470–481, 1957. eprint: `/oup/backfile/content_public/journal/biomet/44/3-4/10.1093/biomet/44.3-4.470/2/44-3-4-470.pdf`.

[82] R. Averkamp and C. Houdre, "A note on the discrete wavelet transform of second-order processes", *IEEE Transactions on Information Theory*, vol. 46, no. 4, pp. 1673–1676, 2000.

[83] G. Walter, *Wavelets and other Orthogonal Systems with Applications*, 3261. Boca Raton a.o. CRC Press, 1994.

[84] Student, "The probable error of a mean", *Biometrika*, vol. 6, no. 1, pp. 1–25, 1908.

[85] M. G. KENDALL, "A new measure of rank correlation", *Biometrika*, vol. 30, no. 1-2, pp. 81–93, 1938. eprint: /oup/backfile/content_public/journal/biomet/30/1-2/10.1093/biomet/30.1-2.81/2/30-1-2-81.pdf.

[86] A Cohen, T Tiplica, and A Kobi, "Statistical process control for ar(1) or non-gaussian processes using wavelets coefficients", *Journal of Physics: Conference Series*, vol. 659, no. 1, 2015.

[87] W. G. Cochran, "The distribution of quadratic forms in a normal system, with applications to the analysis of covariance", *Mathematical Proceedings of the Cambridge Philosophical Society*, vol. 30, no. 2, 178191, 1934.

[88] R. Lund, H. Hurd, P. Bloomfield, and R. Smith, "Climatological time series with periodic correlation", *Journal of Climate*, vol. 8, no. 11, pp. 2787–2809, 1995. eprint: https://doi.org/10.1175/1520-0442(1995)008<2787:CTSWPC>2.0.CO;2.

[89] P. A. Morettin, A. Pinheiro, and B. Vidakovic, *Wavelets in Functional Data Analysis*. Springer, 2017.

[90] D. Wied and R. Weisbach, "Consistency of the kernel density estimator - a survey", 2010.

[91] D. Pollard, *Covergence of Stochastic Processes*. Springer-Verlag, 1984.

VITA

German A. Schnaidt Grez received his B.Sc. in Electrical Engineering from the Naval Polytechnic Academy in Chile. He served as a Navy Officer in the Chilean Navy from 2004-2016, completing different assignments such as Commanding Officer of a Search and Rescue ship, Weapons Engineering Officer of a type 23 frigate, and as an Operations Research Analyst and Systems Analysis Division manager for the surface fleet. During 2014-2015 he earned a M.Sc. in Operations Research from the Georgia Institute of Technology. He is currently a Ph.D. candidate in Industrial Engineering, investigating Machine Learning tools based on the use of wavelets. During his doctoral studies, he interned as a Data Scientist at Amazon, developing predictive algorithms using statistical modeling methodologies in the context of transportation and logistics.