TOPICS ON MULTIRESOLUTION SIGNAL PROCESSING AND BAYESIAN MODELING WITH APPLICATIONS IN BIOINFORMATICS

A Dissertation
Presented to
The Academic Faculty

By

Parisa Yousefi Zowj

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 2021

TOPICS ON MULTIRESOLUTION SIGNAL PROCESSING AND BAYESIAN MODELING WITH APPLICATIONS IN BIOINFORMATICS

Approved by:

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

Dr. David Goldsman, Co-Advisor
School of Industrial and Systems Engineering
Georgia Institute of Technology

Dr. Mirjana Milosevic Brockett
School of Biological Sciences
Georgia Institute of Technology

Dr. King Jordan
School of Biological Sciences
Georgia Institute of Technology

Dr. Jianjun Shi
School of Industrial and Systems Engineering
Georgia Institute of Technology

Dr. Yajun Mei
School of Industrial and Systems Engineering
Georgia Institute of Technology

Date Approved: December 11, 2020

To my beloved family:

My Parents (Khosrow Yousefi and Shahin Rahbar)

My Husband and Son (Kamran and Kian Paynabar)

ACKNOWLEDGEMENTS

First and foremost, I would like to express my appreciation to Dr. Brani Vidakovic for his support, extreme patience, kindness, and brilliance. I am very fortunate to have an opportunity to work under his advisement through my Ph.D.

I also would like to express my gratitude for my thesis committee members, Dr. David Goldsman, Dr. Mirjana Milosevic Brockett, Dr. King Jordan, and Dr. Jianjun Shi and Dr. Yajun Mei for their insightful comments and questions that are very helpful.
In addition, I am very thankful to all my teachers and professors who though me to embrace the education. I hope I can follow your path in my future carrier as a teacher.

Thank you to my amazing husband, Dr. Kamran Paynabar who has been very supportive during my graduate school experience. Thank you for always being there specially during the new and joyful parenthood experience. Kian and I will always be grateful of all you did for our family. A special thank to my most precious son, Kian, for adding joy to our life and being the easiest baby to raise with minimum late night crying and keeping us awake.

Last but not least, my sincere appreciation to my family members who give me endless love. My father and mother always believed in me and gave me all support to be able to follow my dreams. Special thank to my sister, Farnaz, who helped us with some babysitting during the COVID-19 pandemic while we could not use any daycare or outsiders help.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

Analysis of multi-resolution signals and time-series data has wide applications in biology, medicine, engineering, etc. In many cases, the large-scale (low-frequency) features of a signal including basic descriptive statistics, trends, smoothed functional estimates, do not carry useful information about the phenomenon of interest. On the other hand, the study of small-scale (high-frequency) features that look like noise may be more informative even though extracting such informative features are not always straightforward. In this dissertation we try to address some of the issues pertaining to high-frequency features extraction and denoising of noisy signals. Another topic studied in this dissertation is focused on the integration of genome data with transatlantic voyage data of enslaved people from Africa to determine the ancestry origin of Afro-Americans.

1. Assessment of Scaling by Auto-Correlation Shells. Scaling and extracting such high-frequency features, by analyzing the data in the time domain is impossible. To perform scaling a variety of tools such as Structure Functions, Spectrograms, Logscale Diagrams, q-th order Logscale Diagrams have effectively been used. Much of the literature in this area has focused on orthonormal bases because of their interesting properties including the simplicity of the implementation using numerical algorithms, and the capability of precisely detecting edges of signals. Although the analysis of scale-to-scale growth or decay of the orthonormal wavelet coefficients makes the estimation of the local behavior of signals possible, these coefficients are not shift-invariant. The orthonormal shells, on the other hand, are shift-invariant, but not symmetric.

In this chapter, we utilize the Auto-correlation (AC) Shell to propose a feature extraction method that can effectively capture small-scale information of a signal. The AC Shell is a redundant shift-invariant and symmetric representation of the signal that is obtained by using Auto-Correlation function of compactly supported wavelets. The small-scale fea-

tures are extracted by computing the energy of AC Shell coefficients at different levels of decomposition as well as the slope of the line fitted to these energy values.

We discuss the theoretical properties, and verify them using extensive simulations. We compare the extracted features from AC-Shell with those of Wavelets in terms of bias, variance, and mean square error (MSE). The results indicate that the AC-shell features tend to have smaller variance, hence more reliable. Moreover, to show its effectiveness, we validate our feature extraction method in the context of classification to identify patients with ovarian cancer through the analysis of their blood mass spectrum. For this study, we use the features extracted by AC Shell spectrogram along with a support vector machine classifier to distinguish control from cancer cases.

The results show that for both region scenarios, the SVM classifier trained by using the AC Shell spectra slopes as features outperform the wavelet counterpart, in terms of sensitivity, specificity, precision, and accuracy. For example, the specificity of the AC-Shell SVM classifier is around .86, which is 17% higher than that of the Wavelet SVM classifier. The 4% difference in the sensitivity between the two methods, indicates that AC-Shell SVM classifier can outperform its Wavelet counterpart in detecting the cancer cases. The overall accuracy and precision of the AC-Shell SVM classifier are both around 0.93, which are 8% higher than that of Wavelet's. The main reason for this significant difference between the two methods is that AC Shell can generate more robust features with smaller variations that Wavelet.

2. Bayesian Binary Regressions in Wavelet-based Function Estimation. Wavelet shrinkage has been widely used in nonparametric statistics and signal processing for a variety of purposes including denoising noisy signals an images, dimension reduction, and variable/feature selection. Wavelet shrinkage follows a three-step procedure: 1) transformation of the original signals into the wavelet domain and obtaining wavelet coefficients; 2) shrinkage of the coefficients using a thresholding function; and 3) Transformation of the

shrunk coefficients back to the original domain, or utilization of of the low dimensional thresholded coefficients in building a regression/classification model.

Although the traditional wavelet shrinkage methods are effective and popular, they have one major drawback. In these methods the shrinkage process only relies on the information of the coefficient being thresholded and the information contained in the neighboring coefficients is ignored. Similarly, the standard AC Shell denoising methods shrink the empirical coefficients independently, by comparing their magnitudes with a threshold value. The information of other coefficients has no influence on behavior of a particular coefficients. However, due to redundant representation of signals and coefficients obtained by AC Shells, the dependency of neighboring coefficients and the amount of shared information between them increases. Therefore, it would be vital to propose a new thresholding approach for AC Shells coefficients that considers the information of neighboring coefficients.

In this chapter, we develop a new Bayesian denoising for AC Shell coefficients approach that integrates logistic regression, universal thresholding, and Bayesian inference. We validate the proposed method using extensive simulations with various types of smooth and non-smooth signals. The results indicate that for all signal types including the neighbor coefficients would improve the denoising process, resulting in lower MSEs. In all signal types neighboring methods have lowers MSEs than their non-neighboring counterparts. For example, for the Bumps signal the median MSE for Neighboring AC is around 19, while this value for AC Shell is more than 20. This indicates the value of including the neighboring information in the smoothing and denoising process.

Moreover, we validated the proposed methodology using a case study of denoising Atomic Force Microscopy (AFM) signals measuring the adhesion strength between two materials at the nano-newton scale, and correctly identifying the cantilever detachment point.

3. Bayesian Method in Combining Genetic and Historical Records of Transatlantic Slave Trade in the Americas. In the era between 1515 and 1865, more than 12 minions people were enslaved and forced to move from Africa to North and Latin America. The shipping documents have recorded the origin and disembarkation of enslaved people. However, over time due to slave trades they have been moved across North America. This makes identification of African American's origins particularly challenging.

Traditionally, genealogy study has been done via the exploration of historical records, family tress and birth certificates. Due to recent advancements in the field of genetics, genealogy has been revolutionized and become more accurate. Although these methods can provide continental differentiation, they have poor spatial resolution that makes it hard to localize ancestry assignment as these markers are distributed across different sub-continental regions.

To overcome the foregoing drawbacks, in this chapter, we propose a hybrid approach that combines the genetic markers results with the historical records of transatlantic voyage of enslaved people. Addition of the journey data can provide with substantially increased resolution in ancestry assignment, using a Bayesian modeling framework. The proposed Bayesian framework uses the the voyage data from historical records available in the transatlantic slave trade database as prior probabilities and combine them with genetic markers of Afro-Americans, considered as the likelihood information to estimated the posterior (updated) probabilities of their ancestry assignments to geographical regions in Africa.

We applied the proposed methodology to 60 Afro-American individuals, as well as to a group of 448 reference individuals. The results show that the prior information has increased the assignment probabilities obtained by the posterior distributions for some of the regions. The confusion matrices for both likelihood-based method (that does not consider the prior information) and posterior-based method clearly show the accuracy of the posterior-based assignment is more than that of the likelihood-based method. On average,

5 individuals that are mis-classified using the likelihoods are correctly classified using the posterior probabilities. This shows the importance of the prior information in making more accurate determination of one's origin.

CHAPTER 1

BACKGROUND ON SELF-SIMILAR PROCESSES, WAVELETS AND SCALING

Theoretical analysis of self-similar processes such as fractional Brownian motion, which are intrinsically invariant to changes in scale are becoming an fundamental tool for modeling a wide range of real-world phenomena in engineering, physics, medicine, biology, economics, geology, chemistry, and so on.

Time series can be examined in two complementary domains: time and scale/frequency domain. Multiscale methods including wavelets, orthogonal shells, autocorrelation shells, and general time/frequency representations, provide tools and environments to analyze and model such time-series in both time and scale and to unify several related phenomena including self-similarity, and long range dependence.

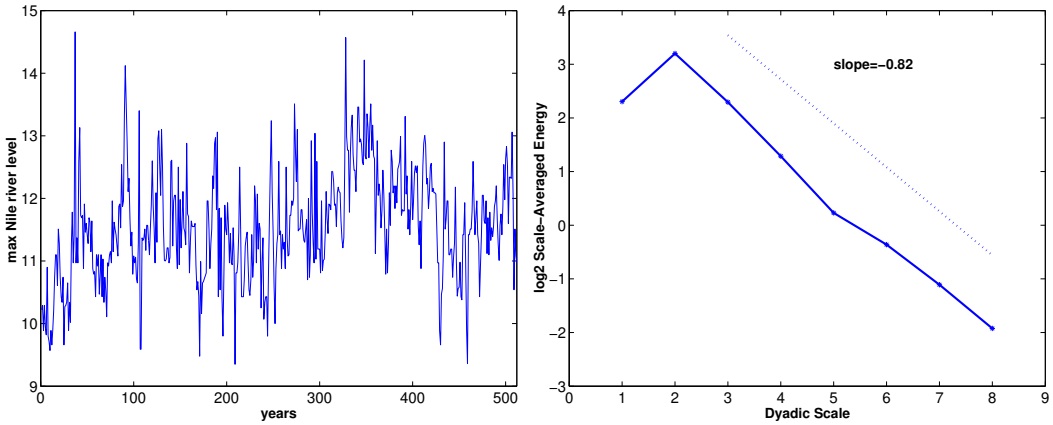
In this Chapter we introduce self-similar processes and its properties, review fractional Brownian motion (fBm) process as one of the most popular self-similar processes, and discuss the basic principles of wavelet modeling and decomposition.

1.1 Self-similar Processes

One of the early mentions of the self-similar processes was done by Harold Edwin Hurst who discovered the Hurst exponent. He was trying to find an optimal reservoir capacity R such that it can accept the river flow in N units of time, $X_1, X_2, ..., X_N$, and have a constant withdrawal of $\bar{X}$ per unit time. By inspecting historical data on Nile River, Hurst discovered an interesting phenomenon that is now referred to as the Hurst effect. He realized that the adjusted range (the ratio of range and standard deviation R/S) scales as N^H for data

ranging from 0.46 to 0.93, with mean 0.73 and standard deviation of 0.09. H was later called the Hurst exponent.

In contrast to Hurst's findings, Feller proved that the theoretical value of R/S was 1/2 for independent and identically distributed random variables with a finite second moment (Feller, 1951). It was assumed that strong Markovian dependence was responsible for the deviation, which Hursts results showed. Later on Barnard (1956) proved that $H = 1/2$ holds for Markovian dependence cases.



(a) (b)

Figure 1.1: (a) Nile yearly minimal level data; (b) its Wavelet log spectra

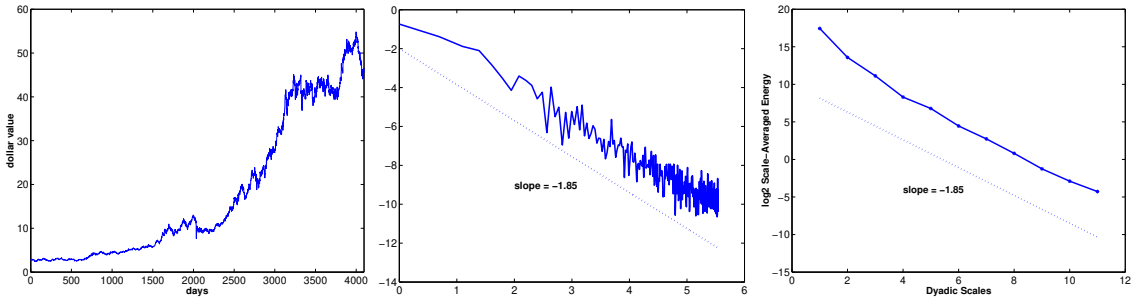
1.1.1 Examples of Self-Similar Processes

Other examples of scaling data and self-similar processes in a variety of applications are discussed in the following subsections.

1.1.1.1 Stock Market Prices and Exchange Rates

Many economic time series, such as stock market prices, exchange rate and asset return exhibit scaling laws and long range dependence. This is in empirical contradiction to several economic theories (random walk theory for stock market, perfect markets, etc) and gave rise to several theories and models describing the scaling and LRD (such as ARFIMA, fGn, fBm, GARCH, etc).

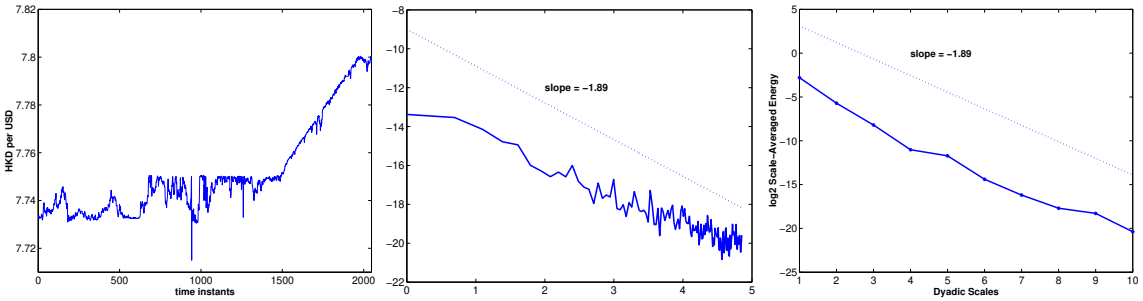
Two example data sets discussed here include Coca Cola stock market prices and rates of exchange between Hong Kong Dollar (HKD) and USDollar (USD).



(a) (b) (c)

Figure 1.2: (a) Coke Stock Market Prices; (b) scaling behavior in the Fourier domain, and (c) scaling behavior in the wavelet domain.

The rates of exchange between Hong Kong Dollar (HKD) and USDollar (USD) as reported by the ONADA Company between 24 March 1995 and 1 November 2000.



(a) (b) (c)

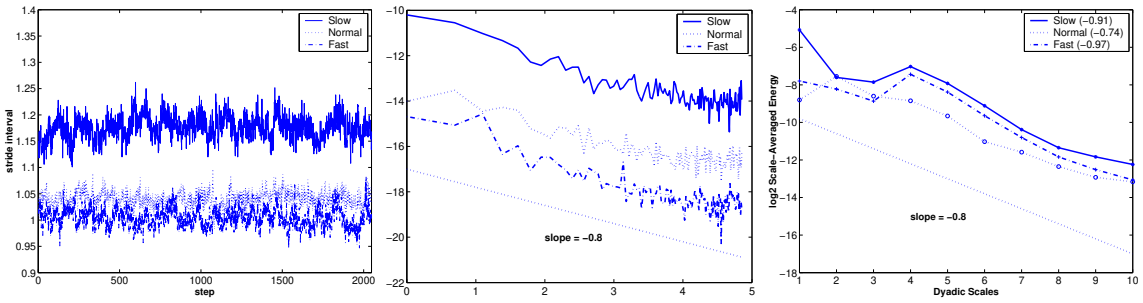
Figure 1.3: (a) Exchange Rates HKD per US$; (b) scaling behavior in the Fourier domain, and (c) scaling behavior in the wavelet domain.

1.1.1.2 Gait Data

Scaling laws were recently detected in the apparently ”noisy” variations in the stride interval (duration of the gait cycle) of human walking. Dynamic analysis of these step-to-step fluctuations revealed a self-similar pattern: fluctuations at one time scale are statistically similar to those at multiple other time scales, at least over hundreds of steps, while healthy subjects walk at their normal rate. The experimental data consist of measurements on a

healthy subject who walked for 1 hour at his usual, slow and fast paces. The stride interval fluctuations exhibited long-range correlations with power-law decay for up to a thousand strides at all three walking rates.

It is curious that during metronomically-paced walking, these long-range correlations disappeared; variations in the stride interval were anti-correlated. Experiments confirm that scaling behavior of spontaneous stride interval are normally quite robust and intrinsic to the locomotor system. Furthermore, this fractal property of neural output may be related to the higher nervous centers responsible for control of walking rhythm.



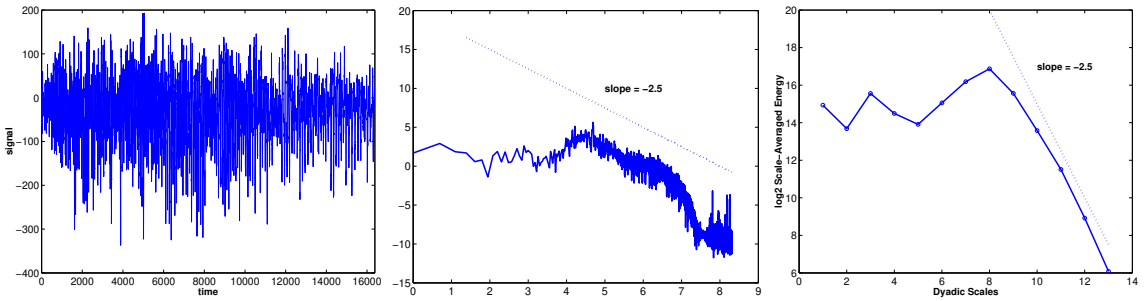
(a) (b) (c)

Figure 1.4: (a) Gait timing for Slow, Normal and Fast Walk;(b) scaling behavior in the Fourier domain, and (c) in the wavelet domain.

Participants in this experiment had no history of any neuromuscular, respiratory or cardiovascular disorders, and were taking no medications. Mean age was 21.7 years (range: 18-29 years). Height was 1.77 $\pm$ 0.08 meters (mean $\pm$ S.D.) and weight was 71.8 $\pm$ 10.7 kg. Subjects walked continuously on level ground around an obstacle free, long (either 225 or 400 meters), approximately oval path and the stride interval was measured using ultra-thin, force sensitive switches taped inside one shoe. Figure 1.4 shows 2048 data points for one subject. Slow and fast stride intervals have slopes of -0.91 and -0.97 respectively, and stride intervals for normal walk show scaling with -0.74 slope.

1.1.1.3 EEG Data

This data set gives fluctuations of measured electrical potential (in μV) derived from brain activity of a patient during an epileptic seizure. It was recorded in the ECT Lab at Duke University Medical Center (Courtesy of Dr. B. Krystal). A patient undergoing ECT therapy had measuring electrodes in his scalp and this particular time series is one of several "channels".



(a) (b) (c)

Figure 1.5: (a) EEG signal at seizure time; (b) scaling behavior in the Fourier domain, and (c) in the wavelet domain.

Outstanding problems for this kind analysis include the prediction, classification, and space-time localization of seizures, see Benedetto and Colella (1995) for wavelet based diagnostic methodology. The original data set covers a 104-second span at a frequency of 256 observations per second, but for our analysis we took a mid-segment of length 2^{14}. A power law with slope of -2.5 was found only at the end of spectrum (several "binomial decades").

1.1.2 Formal Definition of Self-similar Process

Definition 1.1. A random process $X(t), t > 0$ is called *self-similar* if for any $a > 0$, there exists $b > 0$ such that

$$X(at) \stackrel{d}{=} bX(t) \tag{1.1}$$

Lamperti (1962) proved the following result that connected the self-similar processes to the Hurst exponent.

Theorem 1.1. If random process $X(t), t \geq 0$ is nontrivial, stochastically continuous at 0, and self-similar, then there exists unique $H \geq 0$ such that $b = a^H$. If $X(0) = 0$ almost surely (*a.s.*), then $H > 0$.

Standard definition of self-similar processes that involves the self-similarity index is given below.

Definition 1.2. Process $X(t),\ t \geq 0$ is self-similar, with self-similarity index H (H-ss) if and only if there exists $H > 0$ such that for any $a > 0$, $X(at) \stackrel{d}{=} a^H X(t)$.

Uniqueness of H is not obvious from this definition, although, H is unique by the Lamperti's theorem. An example of a self-similar process is Standard Brownian Motion $B(t)$ in which $H = 1/2$, i.e, it is 1/2-ss. Indeed, the process $W(t) = 1/\sqrt{a}B(at)$ is standard Brownian motion, as well.

For 1-D data, there exist many estimation methods for self-similar processes including re-scaled range calculation (R/S), Fourier-Spectra, variance plots, quadrature variations, zero-level crossing, etc. For a comprehensive description, see Beran (1994), Doukhan et al. (2003), and Abry et al. (2013). Wavelet transforms are especially suitable for modeling self-similar phenomena, as is reflected in the literature. An overview is provided in Abry et al. (2000).

If self-similar processes have a stochastic structure, the scaling exponent becomes a parameter in a well-defined statistical model and can be estimated as such. For example, Fractional Brownian Motion (fBm) is an important and well understood model for data that scale. Their importance follows from the fact that they are unique Gaussian processes with stationary increments that are self-similar. We discuss this important process in the next subsection.

1.2 Fractional Brownian Motion (fBm)

Consider a Brownian motion process denoted as $B(t)$. $B_H(t)$ is defined as *fractional Brownian motion* with Hurst exponent H, $(0 < H < 1)$ as in Mandelbrot and Ness. (1968) and is represented by:

$$B_H(t) = \frac{1}{\Gamma(H+1/2)}\Big[\int_{-\infty}^{0}\big(|t-s|^{H-1/2} - |s|^{H-1/2}\big)dB(s) + \int_{0}^{t}|t-s|^{H-1/2}dB(s)\Big]. \tag{1.2}$$

The case when $0 < H < 0.5$ indicates a negatively correlated process, or anti-persistent process; the case when $0.5 < H < 1$ indicates that it is a positively correlated process and the process exhibits long-range dependence (LRD); the case when $H = 0.5$ indicates that the process is almost not a correlated process, it means the process is in fact a Brownian motion (Beran, 1994).

The process $B_H(t)$ is unique, in the sense that class of all fractional Brownian motions with exponent H coincides with the class of all Gaussian H-ss processes. However, a Gaussian process is H-ss with independent increments, if and only if $H = 1/2$, i.e., if it is a Brownian motion.

Alternatively, fractional Brownian motion with the Hurst exponent H could be defined as a random process that satisfies the following properties:

1. $B_H(t)$ has stationary increments, $B_H(t) - B_H(s) \stackrel{d}{=} B_H(t-s)$
2. $B_H(0) = 0$;
3. $E\big(B_H(t)\big) = 0 \quad \forall t$;
4. $E\big(B_H^2(t)\big) = \sigma^2|t|^{2H}, \quad \forall t$ and $\sigma^2 = var(B_H(1))$;
5. $B_H(t)$ is a continuous Gaussian process;

6. $B_H(t)$ is self-similar process;

7. $B_H(t)$ has auto-covariance function:

$$E\big(B_H(t)B_H(s)\big) = \frac{\sigma^2}{2}\big[|t|^{2H} + |s|^{2H} - |t-s|^{2H}\big]. \tag{1.3}$$

where $E|B_H(1)|^2 = \dfrac{\Gamma(2-2H)cos(\pi H)}{\pi H(1-2H)}$.

The fractional Brownian motion (fBm) is arguably among the most popular statistical models in signal and image processing when the process under consideration exhibits some scale-invariance properties.

1.3 Wavelets Basics

The first theoretical results in wavelets are connected with continuous wavelet decompositions of L^2 functions and go back to the early 1980s. Papers of Morlet et al. (1982) and Grossmann and Morlet (1985) 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 $\psi(x) \in L^2(\mathbb{R})$,

$$\psi_{a,b}(x) = \frac{1}{\sqrt{|a|}}\psi\left(\frac{x-b}{a}\right). \tag{1.4}$$

Normalization by $\frac{1}{\sqrt{|a|}}$ ensures that $||\psi_{a,b}(x)||$ is independent of a and b. The function ψ (called *the wavelet function* or *the mother wavelet*) is assumed to satisfy the *admissibility condition,*

$$C_\psi = \int_R \frac{|\Psi(\omega)|^2}{|\omega|}d\omega < \infty, \tag{1.5}$$

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

condition implies

$$0 = \Psi(0) = \int \psi(x)dx. \tag{1.6}$$

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 L^2 function $f(x)$, the continuous wavelet transformation is defined as a function of two variables

$$CWT_f(a, b) = \langle f, \psi_{a,b} \rangle = \int f(x)\overline{\psi_{a,b}(x)}dx. \tag{1.7}$$

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,*

$$f(x) = \frac{1}{C_{\psi}} \int_{R^2} CWT_f(a, b)\psi_{a,b}(x)\frac{da\ db}{a^2}. \tag{1.8}$$

If a is restricted to R^+, which is natural since a can be interpreted as a reciprocal of frequency, becomes

$$C_{\psi} = \int_0^{\infty} \frac{|\Psi(\omega)|^2}{\omega} d\omega < \infty, \tag{1.9}$$

and the *resolution of identity* relation in (1.8) takes the form:

$$f(x) = \frac{1}{C_{\psi}} \int_{-\infty}^{\infty} \int_0^{\infty} CWT_f(a, b)\psi_{a,b}(x)\frac{1}{a^2} da\ db. \tag{1.10}$$

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

Shifting Property. If $f(x)$ has a continuous wavelet transformation $CWT_f(a,b)$, then $g(x) = f(x-\beta)$ has the continuous wavelet transformation $CWT_g(a,b) = CWT_f(a, b-\beta)$.

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

Both the shifting property and the scaling property are simple consequences of changing variables under the integral sign.

Energy Conservation. From (1.10),

$$\int_{-\infty}^{\infty} |f(x)|^2 dx = \frac{1}{C_\psi} \int_{-\infty}^{\infty} \int_{0}^{\infty} |CWT_f(a,b)|^2 \frac{1}{a^2} da\, db. \tag{1.11}$$

Localization. Let $f(x) = \delta(x - x_0)$ be the Dirac pulse at the point x_0. Then, $CWT_f(a,b) = \frac{1}{\sqrt{a}} \psi(\frac{x_0 - b}{a})$.

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

$$F(u,v) = \frac{1}{C_\psi} \int_{-\infty}^{\infty} \int_{0}^{\infty} K(u,v;a,b) F(a,b) \frac{1}{a^2} da\, db, \tag{1.12}$$

i.e., K is a reproducing kernel. The associated reproducing kernel Hilbert space (RKHS) is defined as a 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

$f \in C^{\alpha}$(Hölder space with exponent α), then

$$|CWT_f(a,b)| \leq C|a|^{\alpha+1/2}. \tag{1.13}$$

Conversely, if a continuous and bounded function f satisfies the above condition, then $f \in C^{\alpha.}$

Example 1. Mexican hat or Marr's wavelet. The function

$$\psi(x) = \frac{d^2}{dx^2}[-e^{-x^2/2}] = (1-x^2)e^{-x^2/2} \tag{1.14}$$

is a wavelet known as the "Mexican hat" or Marr's wavelet.
By direct calculation one may obtain $C_\psi = 2\pi$.

Example 2. Poisson wavelet. The function $\psi(x) = -(1+\frac{d}{dx})\frac{1}{\pi}\,\frac{1}{1+x^2}$ is a wavelet known as the Poisson wavelet. The analysis of functions with respect to this wavelet is related to the boundary value problem of the Laplace operator.

The continuous wavelet transformation of a function of one variable is a function of two variables. Clearly, the transformation is redundant. To "minimize" the transformation one can 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 (Fig. 1.6) defined by

$$a = 2^{-j},\ b = k2^{-j},\ j,k \in Z, \tag{1.15}$$

will produce the minimal basis. Any coarser sampling will not give a unique inverse transformation; that is, the original function will not be uniquely recoverable. Moreover under mild conditions on the wavelet function ψ, such sampling produces an orthogonal basis $\{\psi_{jk}(x) = 2^{j/2}\psi(2^j x - k),\ j,k \in Z\}$.

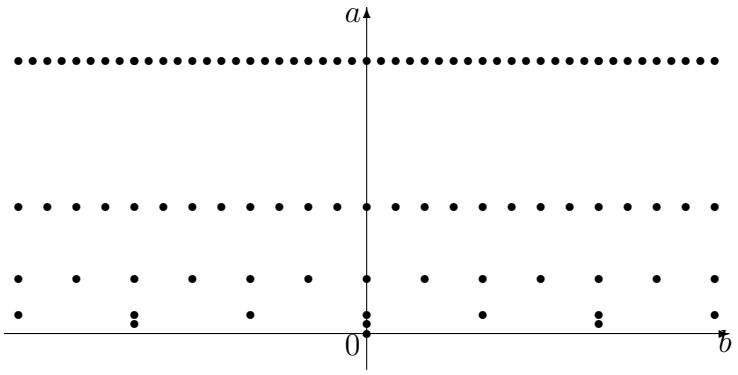


Figure 1.6: Critical Sampling in $R \times R^+$ half-plane ($a = 2^{-j}$ and $b = k\ 2^{-j}$).

There are other discretization choices. For example, selecting $a = 2^{-j},\ b = k$ will lead to non-decimated (or stationary) wavelets. For more general sampling, given by

$$a = a_0^{-j},\ b = k\, b_0\, a_0^{-j},\ j, k \in Z,\ a_0 > 1, b_0 > 0, \tag{1.16}$$

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

$$\psi_{jk}(x) = a_0^{j/2}\, \psi\left(\frac{x - k\, b_0\, a_0^{-j}}{a_0^{-j}}\right) = a_0^{j/2} \psi(a_0^j x - k\, b_0), \tag{1.17}$$

is (1.4) evaluated at (1.16).

Next, we consider wavelet transformations (wavelet series expansions) for values of a and b given by (1.15). An elegant theoretical framework for critically sampled wavelet transformation is *Mallat's Multiresolution Analysis* (Mallat, 1987, 1989a,b, 1998).

As an example, consider the signal Blocks from Wavelab toolbox. The wavelet coefficients at lower scales identify the discontinuities of the signal (see Fig. 1.7).

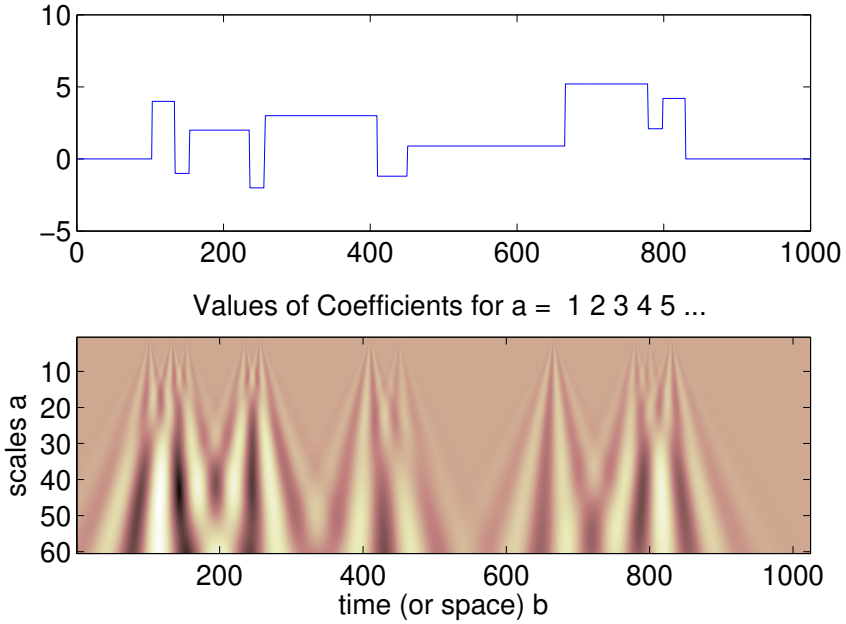


Figure 1.7: Signal Blocks (top) and the CWT (bottom)

1.3.1 Multiresolution Analysis

A multiresolution analysis (MRA) is a sequence of closed subspaces $V_n, n \in Z$ in $L^2(\mathbb{R})$ such that they lie in a containment hierarchy

$$\cdots \subset V_{-2} \subset V_{-1} \subset V_0 \subset V_1 \subset V_2 \subset \cdots . \tag{1.18}$$

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

$$\cap_n V_j = \{\mathbf{0}\}, \quad \overline{\cup_j V_j} = L^2(\mathbb{R}). \tag{1.19}$$

[With $\overline{A}$ we denoted the closure of a set A]. The hierarchy (1.18) is constructed such that (i) V-spaces are self-similar,

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

and (ii) there exists *a scaling function* $\phi \in V_0$ whose integer-translates span the space V_0,

$$V_0 = \left\{ f \in L_2(R) |\ \ f(x) = \sum_k c_k \phi(x-k) \right\},$$

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

Mild technical conditions on ϕ are necessary for future developments. It can be assumed that $\int \phi(x)dx \geq 0$. Later, we will prove 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, i.e.,

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

for some coefficients $h_k,\ k \in Z$. This equation is called the *scaling equation* (or two-scale equation) and it is fundamental in constructing, exploring, and utilizing wavelets.

Whenever a sequence of subspaces satisfies MRA properties, there exists (though not unique) an orthonormal basis for $L^2(\mathbb{R})$,

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

such that $\{\psi_{jk}(x),\ j\text{-fixed},\ k \in Z\}$ is an orthonormal basis of the "difference space" $W_j = V_{j+1} \ominus V_j$. The function $\psi(x) = \psi_{00}(x)$ is called a *wavelet function* or informally *the mother wavelet*.

Next, we detail the derivation of a wavelet function from the scaling function. Since $\psi(x) \in V_1$ (because of the containment $W_0 \subset V_1$), it can be represented as

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

for some coefficients $g_k,\ k \in Z$.

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

1.3.2 Haar Wavelets

In addition to their simplicity and formidable applicability, Haar wavelets have tremendous educational value. Here we illustrate some of the relations discussed earlier using the Haar wavelet. We start with $\phi(x) = \mathbb{I}(0 \leq x \leq 1)$ and pretend that everything else is unknown.

The scaling equation is very simple for the Haar case. By inspection of simple graphs of two scaled Haar wavelets $\phi(2x)$ and $\phi(2x+1)$ stuck to each other, we conclude that the scaling equation is

$$\phi(x) = \phi(2x) + \phi(2x-1) = \frac{1}{\sqrt{2}}\sqrt{2}\phi(2x) + \frac{1}{\sqrt{2}}\sqrt{2}\phi(2x-1),$$

which yields the wavelet filter coefficients:

$$h_0 = h_1 = \frac{1}{\sqrt{2}}.$$

Haar's wavelet has *linear phase,* i.e., the scaling function is symmetric in the time domain. By applying the inverse Fourier transformation we obtain

$$\psi(x) = \phi(2x) - \phi(2x-1)$$

in the time-domain. Therefore we "discovered" the Haar wavelet function ψ.

The Haar basis is not an appropriate basis for all applications for several reasons. The building blocks in Haar's decomposition are discontinuous functions that obviously are not effective in approximating smooth functions. Although the Haar wavelets are well localized in the time domain, in the frequency domain they decay at the slow rate of $O(\frac{1}{n})$.

1.4 Discrete Wavelet Transformations

Discrete wavelet transformations (DWT) are applied to the discrete data sets to produce discrete outputs. Transforming signals and data vectors by DWT is a process that resem-

bles the fast Fourier transformation (FFT), the Fourier method applied to a set of discrete measurements.

Table 1.1: The analogy between Fourier and wavelet methods

Fourier Methods	Fourier Integrals	Fourier Series	Discrete Fourier Transformations
Wavelet Methods	Continuous Wavelet Transformations	Wavelet Series	Discrete Wavelet Transformations

Discrete wavelet transformations map data from the time domain (the original or input data, signal vector) to the wavelet domain. The result is a vector of the same size. Wavelet transformations are linear and they can be defined by matrices of dimension $n \times n$ if they are applied to inputs of size n. Depending on boundary conditions, such matrices can be either orthogonal or "close" to orthogonal. When the matrix is orthogonal, the corresponding transformation is a rotation in R^n space in which the signal vectors represent coordinates of a single point. The coordinates of the point in the new, rotated space comprise the discrete wavelet transformation of the original coordinates.

Example. Let the vector be $\{1, 2\}$ and let $M(1, 2)$ be the point in R^2 with coordinates given by the data vector. The rotation of the coordinate axes by an angle of $\frac{\pi}{4}$ can be interpreted as a DWT in the Haar wavelet basis. The rotation matrix is

$$W = \begin{pmatrix} \cos\frac{\pi}{4} & \sin\frac{\pi}{4} \\ \cos\frac{\pi}{4} & -\sin\frac{\pi}{4} \end{pmatrix} = \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{pmatrix},$$

and the discrete wavelet transformation of $(1, 2)'$ is $W \cdot (1, 2)' = (\frac{3}{\sqrt{2}}, -\frac{1}{\sqrt{2}})'$. Notice that *the energy* (squared distance of the point from the origin) is preserved, $1^2 + 2^2 = (\frac{1}{2})^2 + (\frac{\sqrt{3}}{2})^2$, since W is a rotation.

Example. Let $\mathbf{y} = (1, 0, -3, 2, 1, 0, 1, 2)$. If Haar wavelet is used, the values $f(n) = y_n,\ n = 0, 1, \ldots, 7$ are interpolated by the father wavelet, the vector represent the sampled piecewise constant function. It is obvious that such defined f belongs to Haar's multireso-

lution space V_0.

The following matrix equation gives the connection between $\mathbf{y}$ and the wavelet coefficients (data in the wavelet domain).

$$\begin{bmatrix} 1 \\ 0 \\ -3 \\ 2 \\ 1 \\ 0 \\ 1 \\ 2 \end{bmatrix} = \begin{bmatrix} \frac{1}{2\sqrt{2}} & \frac{1}{2\sqrt{2}} & \frac{1}{2} & 0 & \frac{1}{\sqrt{2}} & 0 & 0 & 0 \\ \frac{1}{2\sqrt{2}} & \frac{1}{2\sqrt{2}} & \frac{1}{2} & 0 & -\frac{1}{\sqrt{2}} & 0 & 0 & 0 \\ \frac{1}{2\sqrt{2}} & \frac{1}{2\sqrt{2}} & -\frac{1}{2} & 0 & 0 & \frac{1}{\sqrt{2}} & 0 & 0 \\ \frac{1}{2\sqrt{2}} & \frac{1}{2\sqrt{2}} & -\frac{1}{2} & 0 & 0 & -\frac{1}{\sqrt{2}} & 0 & 0 \\ \frac{1}{2\sqrt{2}} & -\frac{1}{2\sqrt{2}} & 0 & \frac{1}{2} & 0 & 0 & \frac{1}{\sqrt{2}} & 0 \\ \frac{1}{2\sqrt{2}} & -\frac{1}{2\sqrt{2}} & 0 & \frac{1}{2} & 0 & 0 & -\frac{1}{\sqrt{2}} & 0 \\ \frac{1}{2\sqrt{2}} & -\frac{1}{2\sqrt{2}} & 0 & -\frac{1}{2} & 0 & 0 & 0 & \frac{1}{\sqrt{2}} \\ \frac{1}{2\sqrt{2}} & -\frac{1}{2\sqrt{2}} & 0 & -\frac{1}{2} & 0 & 0 & 0 & -\frac{1}{\sqrt{2}} \end{bmatrix} \cdot \begin{bmatrix} c_{00} \\ d_{00} \\ d_{10} \\ d_{11} \\ d_{20} \\ d_{21} \\ d_{22} \\ d_{23} \end{bmatrix}.$$

The solution is

$$\begin{bmatrix} c_{00} \\ d_{00} \\ d_{10} \\ d_{11} \\ d_{20} \\ d_{21} \\ d_{22} \\ d_{23} \end{bmatrix} = \begin{bmatrix} \sqrt{2} \\ -\sqrt{2} \\ 1 \\ -1 \\ \frac{1}{\sqrt{2}} \\ -\frac{5}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} \end{bmatrix}.$$

Thus,

$$f = \sqrt{2}\phi_{-3,0} - \sqrt{2}\psi_{-3,0} + \psi_{-2,0} - \psi_{-2,1} + \frac{1}{\sqrt{2}}\psi_{-1,0} - \frac{5}{\sqrt{2}}\psi_{-1,1} + \frac{1}{\sqrt{2}}\psi_{-1,2} - \frac{1}{\sqrt{2}}\psi_{-1,3}.$$

The solution is easy to verify. For example, when $x \in [0, 1)$,

$$f(x) = \sqrt{2} \cdot \frac{1}{2\sqrt{2}} - \sqrt{2} \cdot \frac{1}{2\sqrt{2}} + 1 \cdot \frac{1}{2} + \frac{1}{\sqrt{2}} \cdot \frac{1}{\sqrt{2}} = 1/2 + 1/2 = 1 \ (= y_0).$$

Performing wavelet transformations by multiplying the input vector with an appropriate orthogonal matrix is conceptually straightforward, but of limited practical value. Storing and manipulating transformation matrices when inputs are long (> 2000) may not even be feasible.

In the context of image processing, Burt and Adelson (1983a,b) developed orthogonal and biorthogonal pyramid algorithms. Pyramid or cascade procedures process an image at different scales, ranging from fine to coarse, in a tree-like algorithm. The images can be denoised, enhanced or compressed by appropriate scale-wise treatments.

Mallat (1989a,b) was the first to link wavelets, multiresolution analyses and cascade algorithms in a formal way. Mallat's cascade algorithm gives a constructive and efficient recipe for performing the discrete wavelet transformation. It relates the wavelet coefficients from different levels in the transformation by filtering with $\mathbf{h}$ and $\mathbf{g}$. Mallat's algorithm can be viewed as a wavelet counterpart of Danielson-Lanczos algorithm in fast Fourier transformations.

It is convenient to link the original signal with the space coefficients from the space V_J, for some J. Such link is exact for interpolating wavelets (Haar, Shannon, some biorthogonal and halfband-filter wavelets) and close to exact for other wavelets, notably coiflets. Then, coarser smooth and complementing detail spaces are (V_{J-1}, W_{J-1}), (V_{J-2}, W_{J-2}), etc. Decreasing the index in V-spaces is equivalent to coarsening the approximation to the data.

By a straightforward substitution of indices in the scaling equations, one obtains

$$\phi_{j-1,l}(x) = \sum_{k \in Z} h_{k-2l} \phi_{jk}(x) \ \text{ and } \ \psi_{j-1,l}(x) = \sum_{k \in Z} g_{k-2l} \phi_{jk}(x). \tag{1.20}$$

The relations in (1.20) are fundamental in developing the cascade algorithm.

Consider 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 represented uniquely 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 customary to denote the coefficients associated with $\phi_{jk}(x)$ and $\psi_{jk}(x)$ by c_{jk} and d_{jk}, respectively.

Thus,

$$\begin{aligned} v_j(x) &= \sum_k c_{j,k}\phi_{j,k}(x) \\ &= \sum_l c_{j-1,l}\phi_{j-1,l}(x) + \sum_l d_{j-1,l}\psi_{j-1,l}(x) \\ &= v_{j-1}(x) + w_{j-1}(x). \end{aligned} \tag{1.21}$$

By using the general scaling equations (1.20), orthogonality of $w_{j-1}(x)$ and $\phi_{j-1,l}(x)$ for any j and l, and additivity of inner products, we obtain

$$c_{j-1,l} = \langle v_j, \phi_{j-1,l}\rangle = \langle v_j, \sum_k h_{k-2l}\phi_{j,k}\rangle = \sum_k h_{k-2l}\langle v_j, \phi_{j,k}\rangle = \sum_k h_{k-2l}c_{j,k}.$$

Similarly, $d_{j-1,l} = \sum_k g_{k-2l}c_{j,k}$.

The cascade algorithm works in the reverse direction as well. 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 describes a single step in the reconstruction algorithm.

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

For DAUB2, the scaling equation at integers is

$$\phi(n) = \sum_{k=0}^{3} h_k \sqrt{2}\phi(2n-k). \tag{1.22}$$

Recall that $\mathbf{h} = \{h_0, h_1, h_2, h_3\} = \{\frac{1+\sqrt{3}}{4\sqrt{2}}, \frac{3-\sqrt{3}}{4\sqrt{2}}, \frac{3+\sqrt{3}}{4\sqrt{2}}, \frac{1-\sqrt{3}}{4\sqrt{2}}\}$.

Since $\phi(0) = \sqrt{2}h_0\phi(0)$ and $\sqrt{2}h_0 \neq 1$, it follows that $\phi(0) = 0$. Also, $\phi(3) = 0$. For $\phi(1)$ and $\phi(2)$ we obtain the system

$$\begin{bmatrix} \phi(1) \\ \phi(2) \end{bmatrix} = \sqrt{2} \cdot \begin{bmatrix} h_1 & h_0 \\ h_3 & h_2 \end{bmatrix} \cdot \begin{bmatrix} \phi(1) \\ \phi(2) \end{bmatrix}.$$

From $\sum_k \phi(x-k) = 1$ it follows that $\phi(1) + \phi(2) = 1$. Solving for $\phi(1)$ and $\phi(2)$ we obtain

$$\phi(1) = \frac{1+\sqrt{3}}{2} \text{ and } \phi(2) = \frac{1-\sqrt{3}}{2}.$$

Now, one can refine ϕ,

$$\phi\left(\frac{1}{2}\right) = \sum_k h_k \sqrt{2}\phi(1-k) = h_0\sqrt{2}\phi(1) = \frac{2+\sqrt{3}}{4},$$

$$\phi\left(\frac{3}{2}\right) = \sum_k h_k \sqrt{2}\phi(3-k) = h_1\sqrt{2}\phi(2) + h_2\sqrt{2}\phi(1) = \frac{3+\sqrt{3}}{4}\cdot\frac{1-\sqrt{3}}{2} + \frac{3-\sqrt{3}}{4}\cdot\frac{1+\sqrt{3}}{2} = 0,$$

$$\phi\left(\frac{5}{2}\right) = \sum_k h_k \sqrt{2}\phi(5-k) = h_3\sqrt{2}\phi(2) = \frac{2-\sqrt{3}}{4},$$

or ψ,

$$\psi(-1) = \psi(2) = 0,$$

$$\psi\left(-\frac{1}{2}\right) = \sum_k g_k \sqrt{2}\phi(-1-k) = h_1\sqrt{2}\phi(1) = -\frac{1}{4}, \quad [g_n = (-1)^n h_{1-n}]$$

$$\psi(0) = \sum_k g_k \sqrt{2}\phi(0-k) = g_{-2}\sqrt{2}\phi(2) + g_{-1}\sqrt{2}\phi(1) = -h_2\sqrt{2}\phi(1) = -\frac{\sqrt{3}}{4}, \text{etc.}$$

In its general form, wavelet basis is an orthonormal basis in $L^2(\mathbb{R})$, formed by:

$$\psi_{j,k}(x) = 2^{j/2}\psi(2^j x - k) \tag{1.23}$$

$$\phi_{j,k}(x) = 2^{j/2}\phi(2^j x - k) \tag{1.24}$$

as dilation and translation of a wavelet function $\psi(x)$ and scaling function $\phi(x)$. The family $\{\psi_{j,k}\}_{1\leq j\leq J, 0\leq k\leq 2^{n-j}-1}$ and $\{\phi_{J,k}\}_{0\leq k\leq 2^{n-J}-1}$ consists of orthonormal basis.

Decomposition of a function $f(x) \in L^2(\mathbb{R})$ in wavelet domain is given by

$$f(x) = \sum_{j=1}^{J}\sum_{k=0}^{2^{n-j}-1} d_{j,k}\psi_{j,k}(x) + \sum_{k=0}^{2^{n-J}-1} s_{J,k}\phi_{J,k}(x) \tag{1.25}$$

where $d_{j,k} = \int f(x)\psi_{j,k}(x)dx$, and $s_{j,k} = \int f(x)\phi_{j,k}(x)dx$. We refer to the set of coefficients $\{d_{j,k}\}_{1\leq j\leq J, 0\leq k\leq 2^{n-j}-1}$ and $\{s_{J,k}\}_{0\leq k\leq 2^{n-J}-1}$ as detail and scaling coefficients, respectively. Here, n_0 indicates the coarsest scale or lowest resolution level of the transform, and larger values of j correspond to higher resolutions.

The norm of the function f is defined as:

$$\|f\|^2 = \sum_{j=1}^{J}\sum_{k=0}^{2^{n-j}-1} (d_{j,k})^2 + \sum_{k=0}^{2^{n-J}-1} (s_{J,k})^2 \tag{1.26}$$

1.4.1 Wavelet Analysis of Self-similar Signals

In this subsection, we discuss some properties of self-similar signals in the wavelet domain.

Lets consider $f(x)$ as a fractional Brownian motion with Hurst exponent H. As shown by Remenyi and Vidakovic (2013), there is relationship between the expected energy of wavelet coefficients is a linear function of the Hurst exponent. We know that the detail

coefficients in Wavelet decomposition is:

$$d_{j,k} = \int f(x) 2^{j/2} \psi(2^j x - k) dx \tag{1.27}$$

Therefore:

$$\begin{aligned} E(d_{j,k}^2) &= 2^j \int\int E\big[f(u)f(v)\big]\psi(2^j u - k)\psi(2^j v - k) du dv \\ &= 2^j \int\int \frac{\sigma^2}{2}\big(|u|^{2H} + |v|^{2H} - |u-v|^{2H}\big)\psi(2^j u - k)\psi(2^j v - k) du dv \\ &= \frac{\sigma^2}{2} 2^j \Big[\int |u|^{2H} \psi(2^j u - k) \Big(\int \psi(2^j v - k) dv \Big) du \\ &\qquad + \int |v|^{2H} \psi(2^j v - k) \Big(\int \psi(2^j u - k) du \Big) dv \\ &\qquad - \int\int |u-v|^{2H} \psi(2^j u - k)\psi(2^j v - k) du dv \Big] \end{aligned} \tag{1.28}$$

Since $\int \psi(2^j v - k) dv = \int \psi(2^j u - k) du = 0$ the two first integrals vanish. By considering $p = 2^j(u - v)$ and $q = 2^j v - k$ we have:

$$\begin{aligned} E(d_{j,k}^2) &= -\frac{\sigma^2}{2} 2^j \int\int |2^{-j} p|^{2H} \psi(p+q)\psi(q) 2^{-j} dp 2^{-j} dq \\ &= 2^{-2Hj} 2^{-j} \Big(-\frac{\sigma^2}{2} \int\int |p|^{2H} \psi(p+q)\psi(q) dp dq \Big) \\ &= 2^{-(2H+1)j} \big(-\frac{\sigma^2}{2} \big) V_\psi \end{aligned} \tag{1.29}$$

where $V_\psi = \int\int |p|^{2H} \psi(p+q)\psi(q) dp dq$.

If we take log_2 from both side we have:

$$log_2 E(d_{j,k}^2) = -(2H+1)j + C \tag{1.30}$$

where $C = log_2(-\sigma^2 V_\psi / 2)$.

The wavelet spectra can be plotted the points $(j, log_2 E(d_{j,k}^2))$, and consequently the Hurst exponent can be estimated by computing the slope of this spectra.

CHAPTER 2

ASSESSMENT OF SCALING BY AUTO-CORRELATION SHELLS

2.1 Introduction

Analysis of multi-resolution signals and time-series data has wide applications in biology, medicine, engineering, etc. In many cases, the large-scale (low-frequency) features of a signal including basic descriptive statistics, trends, smoothed functional estimates, do not carry useful information about the phenomenon of interest. On the other hand, the study of small-scale (high-frequency) features that look like noise may be more informative even though extracting such informative features are not always straightforward. For example, the pupil diameter in humans fluctuates with high frequency (in hundreds of Hz), and prolonged monitoring of such a diameter leads to massive data sets. Researchers found that the fast dynamics of changes in pupil diameter is associated with eye pathologies (e.g., macular degeneration (Moloney et al., 2006), yet the low-frequency features like trend and mean of the data are clinically irrelevant since the magnitude of the diameter depends on the ambient light and not on the inherent eye pathology.

However, scaling and extracting such high-frequency features, by analyzing the data in the time domain is impossible. To perform scaling a variety of tools such as Structure Functions, Spectrograms, Logscale Diagrams, q-th order Logscale Diagrams have been effectively used. As an example, in Fourier Log-Spectrograms or Logscale Diagrams, if it is possible to fit a straight line with particular slope of $-\alpha$, over duration of several decades (octaves, "binary-decades"), then the scaling in the data is present.

Much of the literature in this area has focused on orthonormal bases because of their interesting properties including the simplicity of the implementation using numerical algo-

rithms, and the capability of precisely detecting edges of signals. Although the analysis of scale-to-scale growth or decay of the orthonormal wavelet coefficients makes the estimation of the local behavior of signals possible, these coefficients are not shift-invariant (Saito (1994), chapter 7). To address this issue we can use orthonormal shells (Abry et al., 2003). However, the representation of a signal in an orthonormal shell is not symmetric (Daubechies, 1988) due to the asymmetric shape of the compactly supported wavelets used in the shell. Moreover, there might be too many zero-crossings because of the rough shape of the wavelets (Mallat, 1991). These drawbacks are of critical in the the edge detection applications where the scale-to-scale analysis of the coefficients is necessary (Mallat and Zhong, 1992).

Additionally, choosing ranges in frequencies for Fourier tools or in scales for wavelet-based tools, and estimating the slope and its variation are important steps since in many situations estimation of the slope is non-robust. This robustness is influenced by several factors, including the quality of data, the closeness of the slope to zero, presence of a periodicity, or injection of energy at a particular scale. Especially the selection of lower scale is more critical. The high variability of spectra at lower scales is influenced by several factors, some of which have nothing to do with the nature of data. For instance, in the Logscale diagrams, points at low scales are obtained by averaging substantially less empirical values of energy (squared wavelet coefficients) than those at high scales. For example, if the scale $l = 10$ averages 1024 energies, the scale $l = 3$ averages only 8 values.

To address the foregoing issues, one can use Auto-Correlation (AC) Shell for signal analysis instead of wavelets or orthonormal shells. The AC shell is a redundant shift-invariant representation of the signal that is obtained by using Auto-Correlation function of compactly supported wavelets (Saito and Beylkin, 1993), and considering dilation and translation. AC Shell is exactly symmetric which allows to detect zero-crossings and compute the slopes at these points. It also simplifies the scale-to-scale analysis of the coefficients. Additionally, it solves the problem of the unbalanced number of coefficients in

computing the average of empirical energy values at different scales as the number of coefficients in each scale is the same and equal to the length of the original signal.

In this chapter, we utilize the AC Shell to propose a feature extraction method that can effectively capture small-scale information of a signal. We study the properties of the proposed method using extensive simulations. To show its effectiveness, we validate our feature extraction method in the context of classification to identify patients with ovarian cancer through the analysis of their MRI images.

The organization of this chapter is as follows. In Section 2, we discuss theoretical background of AC Shells. In Section 3, we present the proposed feature extraction method and discuss its theoretical properties. Section 4 provides simulations and comparative study with existing benchmarks including wavelets counterpart. In Section 5, a real-life application of the proposed methodology is presented, in which we utilize the proposed feature extraction method combined with Support Vector Machine (SVM) classification to identify patients with ovarian cancer by analyzing their MRI images. In Section 5 we provide conclusions and delineate some possible directions for future research.

2.2 Auto-correlation Shell Transform

In this section, we introduce a shift-invariant transform, known as the Auto-Correlation (AC) shells that have some advantages over the Wavelet shells. We begin with reviewing the wavelet decomposition using orthonormal basis in $L^2(\mathbb{R})$. The family $\{\psi_{j,k}\}_{1\leq j\leq J, 0\leq k\leq 2^{n-j}-1}$ and $\{\phi_{J,k}\}_{0\leq k\leq 2^{n-J}-1}$ consists of orthonormal basis. Decomposition of a function $f(x) \in L^2(\mathbb{R})$ in wavelet domain is given by:

$$f(x) = \sum_{j=1}^{J} \sum_{k=0}^{2^{n-j}-1} d_{j,k}\psi_{j,k}(x) + \sum_{k=0}^{2^{n-J}-1} s_{J,k}\phi_{J,k}(x) \tag{2.1}$$

where $d_{j,k} = \int f(x)\psi_{j,k}(x)dx$, and $s_{j,k} = \int f(x)\phi_{j,k}(x)dx$. We refer to the set of coefficients $\{d_{j,k}\}_{1\leq j\leq J, 0\leq k\leq 2^{n-j}-1}$ and $\{s_{J,k}\}_{0\leq k\leq 2^{n-J}-1}$ as detail and scaling coefficients,

respectively. Here, J indicates the coarsest scale or lowest resolution level of the transform, and larger values of j correspond to higher resolutions.

The norm of the function f is defined as

$$\|f\|^2 = \sum_{j=1}^{J} \sum_{k=0}^{2^{n-j}-1} (d_{j,k})^2 + \sum_{k=0}^{2^{n-J}-1} (s_{J,k})^2 \tag{2.2}$$

It is possible to study local behavior of a signal by scale-to-scale analyzing of orthonormal wavelet coefficients, but these coefficients are not shift-invariant. To add the shift-invariance property, we can define orthonormal shells. Consider the family of functions $\tilde{\psi}_{j,k}(x) = 2^{j/2}\psi(2^j(x-k))$ and $\tilde{\phi}_{j,k}(x) = 2^{J/2}\phi(2^J(x-k))$. We call these functions $\{\tilde{\psi}_{j,k}(x)\}_{1\le j\le J, 0\le k\le N-1}$ and $\{\tilde{\phi}_{j,k}(x)\}_{0\le k\le N-1}$ where $N = 2^{n-J}$, a *shell* of the orthonormal wavelets or *orthonormal shell* in short.

Decomposition of a function $f(x)$ in orthonormal shell is given by:

$$f(x) = \sum_{j=1}^{J} \sum_{k=0}^{N-1} \tilde{d}_{j,k}\tilde{\psi}_{j,k}(x) + \sum_{k=0}^{N-1} \tilde{s}_{J,k}\tilde{\phi}_{J,k}(x) \tag{2.3}$$

where $\tilde{d}_{j,k} = \int f(x)\tilde{\psi}_{j,k}(x)dx$ and $\tilde{s}_{J,k} = \int f(x)\tilde{\phi}_{J,k}(x)dx$. We refer to the set of coefficients $\{\tilde{d}_{j,k}\}_{1\le j\le J, 0\le k\le N-1}$ and $\{\tilde{s}_{J,k}\}_{0\le k\le N-1}$ as *orthonormal shell coefficients*.

The new family of functions defined by the above equations can also considered as basis in multi-resolution analysis. They are complete, but they are redundant and not orthonormal (Saito, 1994). Therefore, the decomposition of a function in these bases is not unique. The representation of signals using this family of bases are shift-invariant among different scales. The norm defined as

$$\|f(x)\|_{\mathcal{S}}^2 = \sum_{j=1}^{J} 2^j \sum_{k=0}^{N-1} \left(\tilde{d}_{j,k}\right)^2 + 2^J \sum_{k=0}^{N-1} \left(\tilde{s}_{J,k}\right)^2 \tag{2.4}$$

The factor 2^j in (2.4) is used to offset the redundancy of this presentation, since this presentation at the j-th scale is 2^j times more redundant than the orthonormal wavelet

representation. Therefore, it can be seen that $\|f\|^2 = \|f\|_{\mathcal{S}}^2$

The AC functions of compactly supported wavelets were introduced by Saito and Beylkin (1993). These functions have some interesting properties that among them, symmetry (which is not the necessary case for the corresponding wavelets) and smoothness, are critical for denoising purposes. Using these functions one can define a new shift-invariant transform in the AC shell.

2.2.1 Auto-Correlation Function of Compactly Supported Wavelets

Consider ψ a wavelet function and ϕ the corresponding scaling function. By definition, auto-correlation function are defined by:

$$\Phi(x) = \int_{-\infty}^{\infty} \phi(y)\phi(y-x)dy \tag{2.5}$$

$$\Psi(x) = \int_{-\infty}^{\infty} \psi(y)\psi(y-x)dy \tag{2.6}$$

Because of orthonormal bases $\{\phi(x-k)\}_{0\leq k\leq N-1}$ and $\{\psi(x-k)\}_{0\leq k\leq N-1}$, AC functions have 0 and 1 values at integer points, i.e., for $k \in \mathbb{Z}$

$$\Phi(k) = \delta_{0k} = \begin{cases} 1 & for \quad k = 0 \\ 0 & for \quad k \neq 0 \end{cases} \tag{2.7}$$

and

$$\Psi(k) = \delta_{0k}$$

where δ_{0k} is the Kronecker Delta. Besides, Φ and Ψ have vanishing moments given by:

$$\mathcal{M}_{\Psi}^{m} = \int_{-\infty}^{\infty} x^m \Psi(x)dx = 0 \quad for \quad 0 \leq m \leq L-1 \tag{2.8}$$

$$\mathcal{M}_{\Phi}^{m} = \int_{-\infty}^{\infty} x^m \Phi(x) dx = 0 \quad for \quad 1 \leq m \leq L-1 \tag{2.9}$$

where $\int_{-\infty}^{\infty} \Phi(x)dx = 1$. We can see that $m = 0$ in (2.8) implies $\int_{-\infty}^{\infty} \Psi(x)dx = 0$. As mentioned earlier, the AC functions $\Psi(x)$ and $\Phi(x)$ are *symmetric* which was not necessary the case for ϕ and ψ, and are *smoother* than the original functions, as can be seen in Figure 2.1.

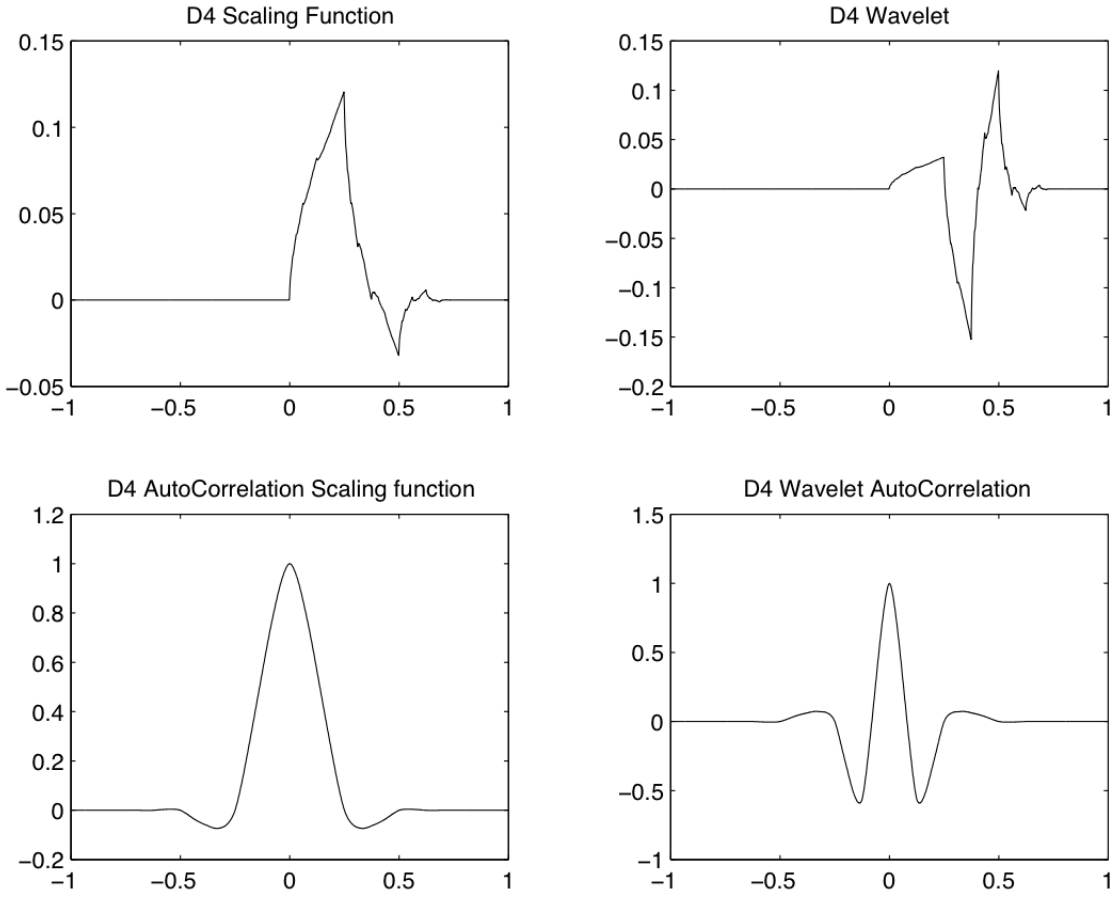


Figure 2.1: Wavelet and Scaling functions of Daubechies 4 wavelet with their Auto-Correlation functions

2.2.2 Auto-Correlation Shell

To have the shift-invariant property, we consider the following family of functions,

$$\left\{\tilde{\Psi}_{j,k}(x)\right\}_{1\leq j\leq J, 0\leq k\leq N-1} \quad and \quad \left\{\tilde{\Phi}_{J,k}(x)\right\}_{0\leq k\leq N-1}$$

where

$$\tilde{\Psi}_{j,k}(x) = 2^{j/2}\Psi(2^j(x-k)) \tag{2.10}$$

$$\tilde{\Phi}_{j,k}(x) = 2^{J/2}\Phi(2^J(x-k)) \tag{2.11}$$

As mentioned earlier, the decomposition in the Auto-Correlation Shell is a shift-invariant transformation. But this decomposition has a significant redundancy compared to the orthonormal shell decomposition. Specifically, for a signal $f(x)$ of length N, we get $N \times (L+1)$ coefficients from its decomposition in the AC shell at level L.

The other interesting property of this representation is its relation to the orthonormal shell of the corresponding compactly supported wavelet. On each scale, the AC Shell coefficients, $D_{j,k}$ and orthonormal shell coefficients, $\tilde{d}_{j,k}$ are convertible to each other independent of other scales (Saito and Beylkin, 1993). It is known that (Rayana, 1998):

$$\int f_d^j(y)2^j\psi(2^j(y-x))dy = \sum_{k=0}^{N-1} D_{j,k}\Psi(x-k) \quad \forall x \tag{2.12}$$

$$\int f_s^j(y)2^j\phi(2^j(y-x))dy = \sum_{k=0}^{N-1} S_{j,k}\Phi(x-k) \quad \forall x \tag{2.13}$$

where

$$f_d^j(y) = \sum_{k=0}^{N-1} \tilde{d}_{j,k}\phi(y-k) \tag{2.14}$$

$$f_s^j(y) = \sum_{k=0}^{N-1} \tilde{s}_{j,k}\phi(y-k) \tag{2.15}$$

and $\tilde{s}_{j,k}$ and $\tilde{d}_{j,k}$ are the orthonormal shell coefficients, respectively.

For integer k we have:

$$D_{j,k} = \int f_d^j(y) 2^j \psi(2^j(y-k))dy \tag{2.16}$$

$$S_{j,k} = \int f_s^j(y) 2^j \phi(2^j(y-k))dy \tag{2.17}$$

2.3 Feature extraction using AC Shell Spectra

Recall that a stochastic process $\{X(\boldsymbol{t}), \boldsymbol{t} \in \mathbb{R}^d\}$ is self-similar with scaling exponent (or *Hurst exponent*) H if, for any $\lambda \in \mathbb{R}^+$, $X(\lambda \boldsymbol{t}) \stackrel{d}{=} \lambda^H X(\boldsymbol{t})$, where "$\stackrel{d}{=}$" denotes the equality in all finite dimensional distributions. The scaling exponent possesses important information about the stochastic process structure and how it scales. Hence, it can be used as a distinguishing feature among different stochastic processes (e.g., case vs control, normal vs. anomalous).

Wavelet spectra that shows the energy level of wavelet coefficients at different scales (decomposition level), is a capable tool in capturing the self-similarity in the signals and estimating the scaling exponent. Some important pioneering work in this area was done by Flandrin and his collaborators (Abry et al., 1993; Flandrin, 1989, 1992a; Flandrin and Goncalves, 1992). It can be shown (Vidakovi, 1999) that the expected value of the energy of wavelet coefficients at each scale can be represented by a linear function of the scale whose slope linearly depends on the Hurst exponent. Specifically,

$$log_2 E(d_{j,k}^2) = -(2H+1)j + C, \tag{2.18}$$

where $d^2_{j,k}$ is the energy at scale j, H is the Hurst exponent, and $C = log_2(-\sigma^2 V_\psi/2)$.

Therefore, as can be seen in Figure 2.2 (b), by plotting the points $(j, log_2 E(d^2_{j,k}))$ we will get a spectra that can be used to estimate the slope and consequently the Hurst (scaling) exponent.

One problem with the estimate of the Hurst exponent obtained from the wavelet spectra is the high variance of the such estimate. As mentioned earlier, these estimates can be used as features for classification. Therefore, the high variability may impact the accuracy of the classification or anomaly detection procedure. To address this issue, in the next subsection, we propose a new method for estimation of the Hurst exponent, using AC Shell spectra. AC Shell due to symmetry, smoothness and redundancy properties will result in estimates with less variations, therefore, useful for classification.

2.3.1 Hurst Exponent Estimation using AC-Shell Spectra

Consider the AC Shell coefficients of signal S as $(S_{J,k}, D_{1,k}, D_{2,k}, ..., D_{J,k})$ where J is a fixed level smaller than $log_2(N) - 1$. $S_{J,k}$ are scaling AC Shell coefficients and $D_{j,k}$ are detail AC Shell coefficients for $k = 0, ..., N-1$, given by $D_{j,k} = \int f^j_d(y) 2^{-j} \psi\big(2^{-j}(y - k)\big) dy$, where $f^j_d(y) = \sum_{k=0}^{N-1} \tilde{d}_{j,k} \phi(y - k)$.

Proposition 2.1. The expected energy of AC Shell coefficients is given by

$$\mathbb{E}(D^2_{j,k}) = 2^{-(2H+1)j} \frac{\sigma^2_H}{2} V_\psi Q_{\psi,j}, \tag{2.19}$$

where V_ψ and $Q_{\psi,j}$ depend on wavelet functions.

Proof.

$$\begin{aligned} \mathbb{E}(D^2_{j,k}) &= \mathbb{E}\Big\{ \Big(\int f^j_d(y) 2^j \psi\big(2^j(y-k)\big) dy \Big) \Big(\int f^j_d(z) 2^j \psi\big(2^j(z-k)\big) dz \Big) \Big\} \\ &= 2^{2j} \int \int \mathbb{E}\Big(f^j_d(y) f^j_d(z) \Big) \psi\big(2^j(y-k)\big) \psi\big(2^j(z-k)\big) dy dz \end{aligned} \tag{2.20}$$

Now, we calculate $\mathbb{E}\Big(f_d^j(y)f_d^j(z)\Big)$ and then we plug it in the above equation.

$$\begin{aligned}
\mathbb{E}\Big(f_d^j(y)f_d^j(z)\Big) &= \mathbb{E}\Big(\sum_{n=0}^{N-1}\tilde{d}_{j,n}\phi(y-n)\sum_{m=0}^{N-1}\tilde{d}_{j,m}\phi(z-m)\Big)\\
&= \sum_{n=0}^{N-1}\sum_{m=0}^{N-1}\mathbb{E}(\tilde{d}_{j,n}\tilde{d}_{j,m})\phi(y-n)\phi(z-m)\\
&= \sum_{n=0}^{N-1}\sum_{m=0}^{N-1}\mathbb{E}\Big[\int f(u)\tilde{\psi}_{j,n}(u)du\int f(v)\tilde{\psi}_{j,m}(v)dv\Big]\phi(y-n)\phi(z-m)\\
&= \sum_{n=0}^{N-1}\sum_{m=0}^{N-1}\phi(y-n)\phi(z-m)\Big[\int\int\mathbb{E}\big(f(u)f(v)\big)\tilde{\psi}_{j,n}(u)\tilde{\psi}_{j,m}(v)dudv\Big]\\
&= \sum_{n=0}^{N-1}\sum_{m=0}^{N-1}\phi(y-n)\phi(z-m)\Big[\int\int\mathbb{E}\big(f(u)f(v)\big)\\
&\qquad\qquad 2^{j/2}\psi\big(2^j(u-n)\big)2^{j/2}\psi\big(2^j(v-m)\big)dudv\Big]
\end{aligned} \tag{2.21}$$

By using the form of auto-covariance function of fBm, we have:

$$\mathbb{E}\big(f(u)f(v)\big) = \frac{\sigma_H^2}{2}(|u|^{2H}+|v|^{2H}-|u-v|^{2H}) \tag{2.22}$$

By plugging in this in the main equation, we can see

$$\begin{aligned}
\mathbb{E}\Big(f_d^j(y)f_d^j(z)\Big) = 2^j\sum_{n=0}^{N-1}\sum_{m=0}^{N-1}\phi(y-n)\phi(z-m)\frac{\sigma_H^2}{2}\Big[&\\
&\int|u|^{2H}\psi\big(2^j(u-n)\big)\Big(\int\psi\big(2^j(v-m)\big)dv\Big)du\\
&+\int|v|^{2H}\psi\big(2^j(v-n)\big)\Big(\int\psi\big(2^j(u-m)\big)du\Big)dv\\
&-\int\int|u-v|^{2H}\psi\big(2^j(u-n)\big)\psi\big(2^j(v-m)\big)dudv\Big].
\end{aligned} \tag{2.23}$$

The first two integrals inside the above brackets are zero, as $\int\psi(x)dx=0$, for the third doubled integral we use change of variables in the form of $p=2^j(u-n)-2^j(v-m)=$

$2^j(u-v+m-n)$ and $q=2^j(v-m)$.

Consequently, we have $u-v=2^{-j}p+n-m$ and $2^j(u-n)=p+q$. So,

$$\mathbb{E}\Big(f_d^j(y)f_d^j(z)\Big)=2^j\frac{\sigma_H^2}{2}\sum_{n=0}^{N-1}\sum_{m=0}^{N-1}\phi(y-n)\phi(z-m)\Big[\int\int|2^{-j}p+n-m|^{2H}\psi(p+q)\psi(q)(2^{-j}dp)(2^{-j}dq)\Big] \tag{2.24}$$

for choice of $n=m=k$, we get,

$$\begin{aligned}\mathbb{E}\Big(f_d^j(y)f_d^j(z)\Big)&=\big(2^j\big)\big(2^{-2j}\big)\big(2^{-2Hj}\big)\frac{\sigma_H^2}{2}\sum_{k=0}^{N-1}\phi(y-k)\phi(z-k)\Big[\\ &\int\int|p|^{2H}\psi(p+q)\psi(q)dpdq\Big]\\ &=2^{-(2H+1)j}\frac{\sigma_H^2}{2}V_\psi\sum_{k=0}^{N-1}\phi(y-k)\phi(z-k),\end{aligned} \tag{2.25}$$

where $V_\psi=\int\int|p|^{2H}\psi(p+q)\psi(q)dpdq$ does not depend on j, but just on H and ψ.
We finally have the $\mathbb{E}\Big(f_d^j(y)f_d^j(z)\Big)$ to plug into the main equation. That is,

$$\begin{aligned}\mathbb{E}(D_{j,k}^2)&=\int\int\mathbb{E}\Big(f_d^j(y)f_d^j(z)\Big)2^j\psi\big(2^j(y-k)\big)2^j\psi\big(2^j(z-k)\big)dydz\\ &=2^{-(2H+1)j}\frac{\sigma_H^2}{2}V_\psi\\ &\sum_{k=0}^{N-1}\bigg(\int\phi(y-k)\psi\big(2^j(y-k)\big)2^jdy\bigg)\bigg(\int\phi(z-k)\psi\big(2^j(z-k)\big)2^jdz\bigg)\end{aligned} \tag{2.26}$$

The Last summation depends on the wavelet function ψ and j, so, we call it $Q_{\psi,j}$. To summarize:

$$\mathbb{E}(D_{j,k}^2)=2^{-(2H+1)j}\frac{\sigma_H^2}{2}V_\psi Q_{\psi,j} \tag{2.27}$$

□

Based on the results of the foregoing proposition, we can see the Hurst exponent of

a fraction Brownian motion can be estimated from the slope in the equation given in the proposition. The empirical counterpart of this equation is a regression model defined on the pairs of $\big(j, log_2(\overline{D^2}_{j,k})\big)$, where $\overline{D^2}_{j,k})$ (average of squared detail AC Shell coefficients at level j) is an empirical counterpart of $E(D^2_{j,k})$. The sample mean in can be replaced by sample median or any other location estimation to produce more robust estimators of the spectra. Also, the regression should be weighted since the variances in the levels are not equal anymore.

The slope of this linear equation is an estimator for Hurst Exponent of fBm.

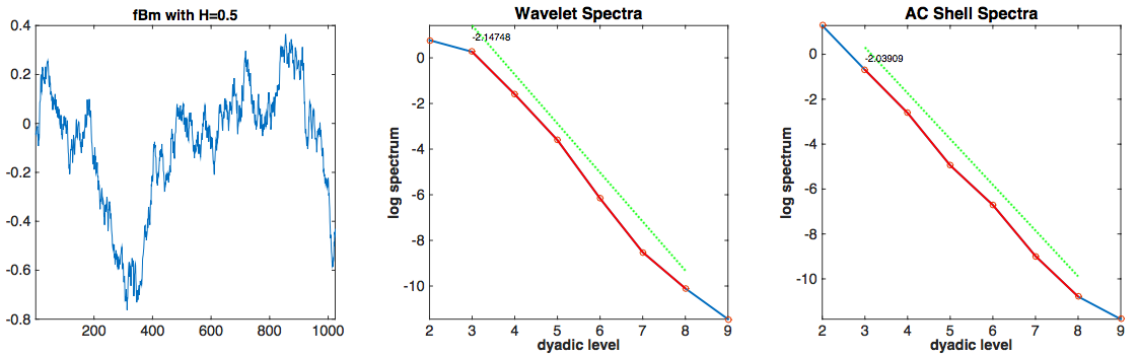
$$Slope \simeq -(2H + 1)$$

therefore a biased AC Shell based estimator for Hurst exponent can be introduced by:

$$\hat{H} \simeq -(Slope + 1)/2$$

where the slopes can be computed by the above AC Shell methods.

Figure 2.2 (c) shows AC Shell spectra for an fBm with $H = 0.5$ by length 1024.



(a) (b) (c)

Figure 2.2: (a) fBm with $H = 0.5$ (we expect $slope = -2$), (b) Wavelet spectra with $slope = -2.14748$, (c) AC Shell spectra with $slope = -2.03909$

2.4 Evaluation of the Proposed Method using Simulations

In this section, we evaluate the performance of the proposed method for estimation of the Hurst exponent using AC Shell spectra. For this purpose, we generate a variety of the fBm time series of length 1024 with the different Hurst exponents of $H = 0.3, 0.4, 0.5$. For each signal, we compute the slope of spectra by using Wavelet, AC Shell spectra. This procedure is repeated 1000 times to capture the variability of estimates. We use two wavelet basis functions, namely, Daubechies 4 and Symmelet 4. Table 2.1 contains the mean and (variance) of slopes and Table 2.2 shows the estimated Hurst exponent and (Mean Square Error) of the estimated Hurst exponent.

Table 2.1: Mean and variance of computed slopes with Wavelet and AC Shell methods by Daubechies4 and Symmlet4 wavelets. In each cell we have mean and (variance) of 1000 times computed slope for a fBm with Hurst exponent H

	Slope	Daubechies		Symmlet	
H	$-(2H+1)$	Wavelet	AC Shell	Wavelet	AC Shell
0.3	-1.6	-1.6096 (0.0223)	-1.6123 (0.0172)	-1.5635 (0.0196)	-1.5537 (0.0143)
0.4	-1.8	-1.8239 (0.0179)	-1.8237 (0.0131)	-1.8067 (0.0197)	-1.7839 (0.0134)
0.5	-2.0	-1.9587 (0.0116)	-1.9516 (0.0076)	-01.9978 (0.0167)	-1.9523 (0.0111)

As can be seen from Table 2.1, both Wavelets and AC Shell perform similarly in estimating the slope, but the variances of slopes in AC Shell method is smaller than the one with Wavelet method. For example, for $H = 0.5$, the variance of the estimated slopes (standard error) using Wavelet is 0.0116, while this number for AC Shell is 0.0076, when Daubechies4 basis function is used. Also, this smaller variability does not depend on the type of the basis function, and may help us to use this feature for creating accurate classifiers. Additionally, the MSE results reported in Table 2.2, which is the sum of the squared bias and variance, confirms the observations made from the slope estimates.

Table 2.2: Hurst exponent estimation with Wavelet and AC Shell methods by Daubechies and Symmlet wavelet. The number in parenthesis shows MSE of each estimation based on 1000 iteration

	Daubechies		Symmlet	
H	Wavelet	AC Shell	Wavelet	AC Shell
0.3	0.3048 (0.0056)	0.3062 (0.0043)	0.2818 (0.0052)	0.2769 (0.0041)
0.4	0.4120 (0.0046)	0.4118 (0.0034)	0.4033 (0.0049)	0.3919 (0.0034)
0.5	0.4793 (0.0033)	0.4758 (0.0025)	0.4989 (0.0042)	0.4761 (0.0033)

Furthermore, Figure 2.3 show the distribution of the estimated slopes obtained in different replications using boxplots. These boxplots again confirm that the variability of the slope estimates obtained by AC Shell is less than those of Wavelets. This difference in variability is more profound when $H = 0.5$

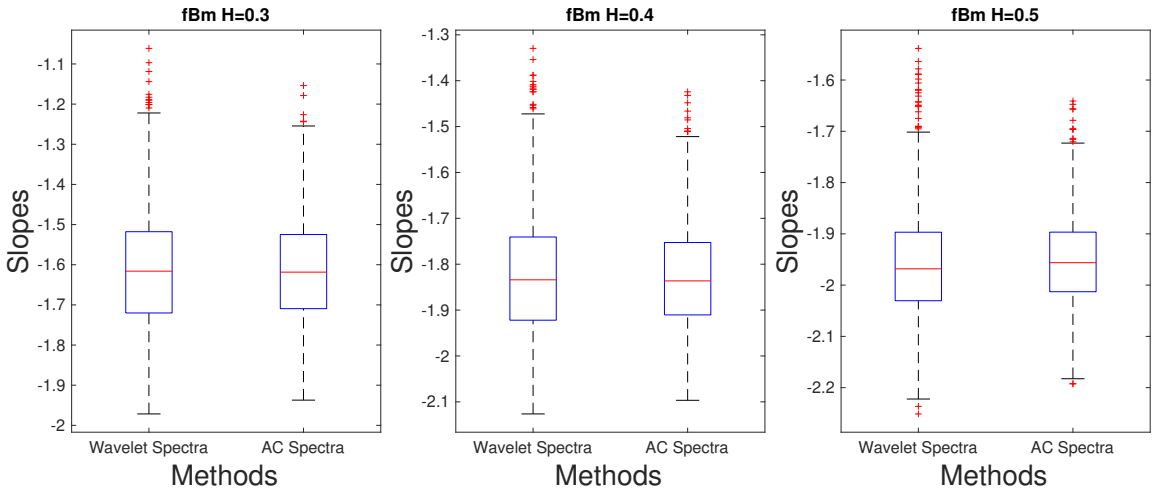


Figure 2.3: Boxplot of estimated slopes based on Wavelet Spectra and AC Shell Spectra for fBm with Hurst exponent (a) $H = 0.3$, (b) $H = 0.4$, (c) $H = 0.5$

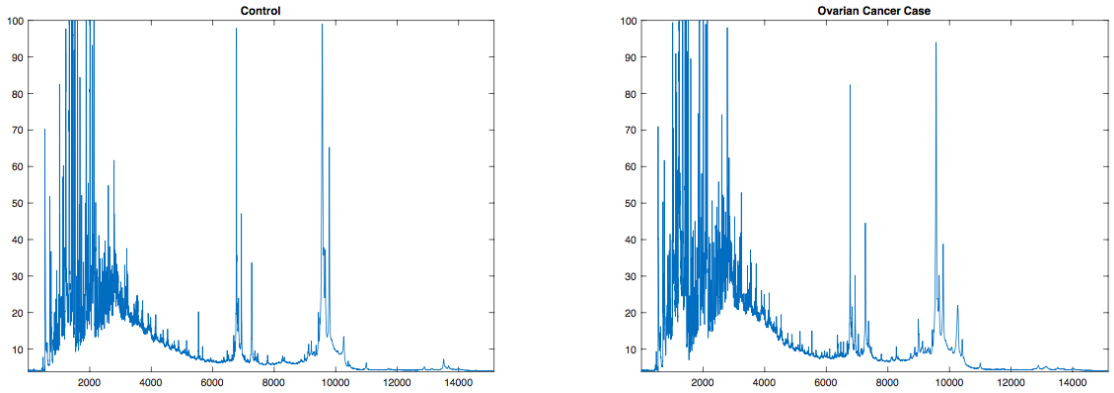
In short, the simulation study conducted in this section showed that the slope estimation of the AC Shell spectra leads to less variable and therefore more robust features than the Wavelet counterpart.

2.5 Case Study: Classification of Ovarian Cancer Spectrum Data

The development of tools for the early cancer diagnosis is a major open problem, and clinicians have investigated a variety of diagnosis techniques. Recently, they have discovered that cancer may affect the blood mass spectrum, and studied diagnosis methods based on the analysis of mass-spectrum data, which provide information about proteins and their fragments (Bakhtiar and Nelson, 2001; Bakhtiar and Tse, 2000; Yates, 2000). The blood mass spectrum, as shown in Figure (2.4), is a curve, where the x- axis shows the ratio of the weight of a specific molecule to its electric charge, and the y-axis is the signal intensity for the same molecule. The mass-spectrum analysis is a fast inexpensive procedure based on a sample of a patient's blood, and it may potentially allow cancer screening with little discomfort to a patient.

The dataset we used in this study includes the mass spectra of 162 patients with ovarian cancer and 91 healthy people. Each mass-spectrum curve consists of 15,154 points. The dataset is available at *http://clinicalproteomics.steem.com* (Tang et al., 2004). The main goal of this case study is to extract robust features that can help distinguish between the case and control instances. For this purpose, we apply both Wavelet-based and AC Shell-based spectra to extract features and use them to train a support vector machine (SVM) classifier. The results are compared in terms of Sensitivity, Specificity, Precision, and Accuracy.

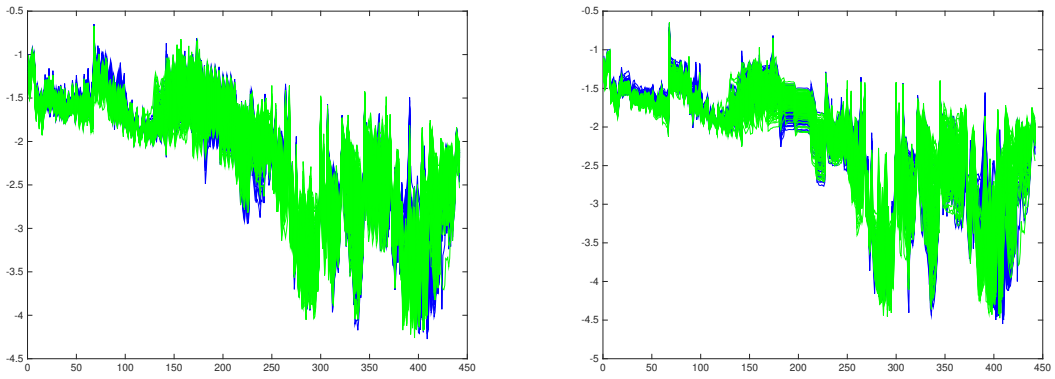
We begin with some preprocessing step to prepare the signals for analysis. We first use an overlapping window with the size of 2^{10} and shift it along the entire signal with step size 2^5 to create 442 sub-signals. Then, using the wavelet "Symmlet 4", we compute the slope of the spectra for all sub-signals with wavelet and AC methods. The mean of slopes in Wavelet method over all lags and people in the control group is -1.9748, which shows that Hurst exponent H is close to 0.5. As discussed in the simulation section, in this case, the AC Shell spectra results in features (slope estimates) with smaller variation, which can lead to a better separation between control and cancer. Figure (2.5) shows the 442 slopes



(a) (b)

Figure 2.4: A sample of blood mass spectrum for (a) a control and (b) a cancer case person

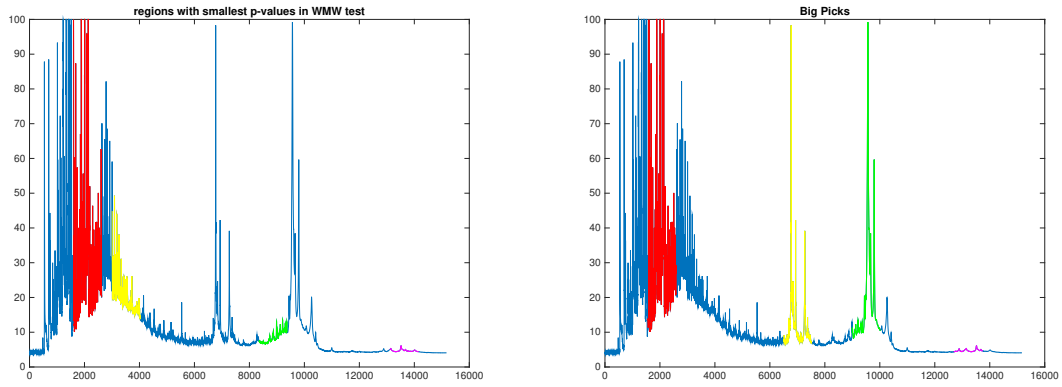
for all control and cancer cases in (a) Wavelet method and (b) AC method.



(a) (b)

Figure 2.5: Slopes of spectra for 442 sub-signals with (a) wavelet method and (b) AC method. Blue used for control and green for cancer cases

In order to find the most informative sub-signals, we use Wilcoxon-Mann-Withny (WMW) test to compare the mean of slopes for the control and cancer group for each region (sub-signal). From 442 regions, 339 and 347 regions have smaller p-values than 0.05 for wavelet and AC methods respectively. We pick 4 regions with the smallest p-values to train the classification model. Figure (2.6) (a) shows those 4 regions with smallest p-value for on of the control patients.



(a) (b)

Figure 2.6: 4 regions in mass spectrometry of serum proteins of a control case as an example.(a) regions with smallest p-values in WMW test (b) regions with biggest picks in practice

In addition, we consider another scenario, where we pick 4 regions with the largest peaks and regions of interest. These regions are shown in Figure (2.6) (b). After feature extraction (estimation of the slope in spectra), we train an SVM classifier for each scenario (i.e., largest peaks and smallest p-values). We randomly split the data into a training set (169 observations, two third of the data) and a testing set (84 observations, a third of the data) with a similar proportion of cancer cases. Then, the slopes of 4 regions are used as features to train the SVM model with Gaussian or Radial Basis Function (rbf) kernels. The outcome of the SVM classifier on test data is used to compute sensitivity, specificity, precision, and accuracy. This procedure is repeated 1000 times and the average assessment metrics for each scenario and feature extraction method is reported in Tables 2.3 - 2.4.

Table 2.3: SVM results based on slopes of 4 regions with smallest p-values of WMW test in mass spectrometry of serum proteins for both Wavelet and AC methods

Method	Sensitivity	Specificity	Precision	Accuracy
Wavelet	0.9656	0.9255	0.9587	0.9510
AC	0.9822	0.9665	0.9813	0.9765

As can be seen from the tables, for both region scenarios, the SVM classifier trained by using the AC Shell spectra slopes as features outperform the wavelet counterpart, in

Table 2.4: SVM results based on slopes of 4 regions with biggest picks in mass spectrometry of serum proteins for both Wavelet and AC methods

Method	Sensitivity	Specificity	Precision	Accuracy
Wavelet	0.9291	0.6942	0.8432	0.8438
AC	0.9670	0.8632	0.9265	0.9293

terms of all assessment measures including sensitivity, specificity, precision, and accuracy. For example, for the four-peak scenario, the specificity of the AC-Shell SVM classifier is around .86, meaning that 86% of control people are correctly classified. This specificity is 17% higher than that of the Wavelet SVM classifier which is around 0.69. The 4% difference in the sensitivity between the two methods, indicates that AC-Shell SVM classifier can outperform its Wavelet counterpart in detecting the cancer cases. The overall accuracy and precision of the AC-Shell SVM classifier are both around 0.93, which are 8% higher than that of Wavelet's. The main reason for this significant difference between the two methods is that AC Shell can generate more robust features with smaller variations that Wavelet.

2.6 Conclusions

In this chapter, we proposed a new method for robust feature extraction from signal with scaling property. We suggested the use of slope estimate of the AC Shell spectra (i.e., the energy of AC Shell coefficients at different scales) as a feature and showed that it has less variation than the slope estimate of the Wavelet spectra. This makes AC Shell-based feature more robust and hence it results in a better classification performance. We validated the proposed method using simulations by generating random realizations of fBm with different Hurst exponents and computing the bias, variance and MSE for both Wavelet and AC Shell spectra slopes. The results confirmed our hypothesis that AC shell features have smaller variance. Furthermore, we applied our proposed method in analyzing blood mass spectra to detect cancer cases and distinguish them from healthy people. We trained two SVM classifiers with features obtained from Wavelets and AC Shell spectra slopes and

measured the performance using sensitivity, specificity, precision, and accuracy measures. The results again indicated that the classifier that was trained by using AC Shell features is superior.

CHAPTER 3

BAYESIAN BINARY REGRESSIONS IN WAVELET-BASED FUNCTION ESTIMATION

3.1 Introduction

Wavelet shrinkage has been widely used in nonparametric statistics and signal processing for a variety of purposes including denoising noisy signals an images, dimension reduction, and variable/feature selection. Wavelet shrinkage follows a three-step procedure: 1) transformation of the original signals into the wavelet domain and obtaining wavelet coefficients; 2) shrinkage of the coefficients using a thresholding function; and 3) Transformation of the shrunk coefficients back to the original domain, or utilization of of the low dimensional thresholded coefficients in building a regression/classification model. Examples of shrinkage methods include universal soft and hard thresholding (Donoho, 1995; Donoho and Johnstone, 1994a), Stein's Unbiased Risk Estimate (SURE)-based shrinkage (Donoho and Johnstone, 1994b), and Bayes' shrink (Chang et al., 2000).

Although the traditional wavelet shrinkage methods are effective and popular, they have one major drawback. In these methods the shrinkage process only relies on the information of the coefficient being thresholded and the information contained in the neighboring coefficients is ignored. To address this issue (Remenyi and Vidakovic, 2013) proposed a Bayesian wavelet-based denoising methodology based on the total energy of a neighboring pair of coefficients at the same decomposition level plus their parental coefficient. Their proposed shrinkage model is based on a Bayesian hierarchical model using a contaminated exponential prior on the total mean energy in a neighborhood of wavelet coefficients. The hyperparameters of their model are estimated empirically. In this approach the shrinkage

is performed based on based on the ratio of the estimated and observed energy. They validated their method and showed its superior performance through simulations and a case study from inductance plethysmography.

Auto-Correlation (AC) Shell has some advantage over the wavelets by providing a redundant shift-invariant representation of the signal that is obtained by using Auto-Correlation function of compactly supported wavelets [5] Saito and Beylkin (1992) and considering dilation and translation. AC Shell is exactly symmetric which allows to detect zero-crossings and compute the slopes at these points. It also simplifies the scale-to-scale analysis of the coefficients. Additionally, it solves the problem of the unbalanced number of coefficients in computing the average of empirical energy values at different scales as the number of coefficients in each scale is the same and equal to the length of the original signal.

In function estimation and denoising applications, the standard AC Shell methods shrink the empirical coefficients independently, by comparing their magnitudes with a threshold value. The information of other coefficients has no influence on behavior of a particular coefficients. However, due to redundant representation of signals and coefficients obtained by AC Shells, the dependency of neighboring coefficients and the amount of shared information between them increases. Therefore, it would be vital to propose a new thresholding approach for AC Shells coefficients that considers the information of neighboring coefficients.

For this purpose, we develop a new Bayesian denoising for AC Shell coefficients approach that integrates logistic regression, universal thresholding, and Bayesian inference. We validate the proposed method using extensive simulations and a case study of denoising Atomic Force Microscopy (AFM) signals measuring the adhesion strength between two materials at the nano-newton scale.

3.2 Overview of Proposed Methodology

In this section, we provide an overview of the neighborhood AC shell denoising. It has been inspired by block shrinkage (De Canditiis and Vidakovic, 2004) and Λ-Shrinkage (Remenyi and Vidakovic, 2013).

In the subsequent sections, we discuss neighborhood AC shell denoising details. The proposed method can be summarized through the following steps:

- Step 1. Decompose the original signal by using AC Shell basis functions and obtain the corresponding coefficients.

- Step 2. Apply the universal hard thresholding function to the coefficients individually and indicate the unthresholded coefficients by 1 and the rest by zero. This information is used as prior in the Bayesian setting.

- Step 3. For each coefficient, consider the two immediate left and right neighbors. The information provided by the energy of these neighbor coefficients is useful for determining the significance of the coefficient, hence, considered as as the likelihood.

- Step 4. The posterior, defined as the probability that a coefficient is unthresholded given the energy of the neighboring coefficients, are computed by fitting a logistic regression. The posterior probabilities are thresholded to identify the significant coefficients.

- Step 5. The final list of significant coefficients is formed by finding the union of unthresholded priors and posteriors. The remainder of the coefficients are thresholded and set to zero. $D_{j,k}$ for $j = 1, ..., L$ and $k = 0, ..., N-1$ where N is signal length and L is the level of decomposition

3.3 Neighboring AC Shell Denoising

In this section we elaborate the proposed Bayesian method for denoising a signal using AC Shell decomposition.

3.3.1 AC Shell Decomposition and Thresholding

The proposed method begins with decomposing the signal using the AC Shell basis. Consider a noisy observed signal Y, defined by

$$Y_i = f(x_i) + \epsilon_i; i = 1, 2, ...N, \tag{3.1}$$

where $f(x_i)$ is the true function and ϵ_i is the white noise. The goal is to denoise the observed signal and obtain the true function $f(x)$.

To obtain the signal representation in the AC Shell domain, we consider the following family of functions,

$$\left\{\tilde{\Psi}_{j,k}(x)\right\}_{1\leq j\leq J, 0\leq k\leq N-1} \quad and \quad \left\{\tilde{\Phi}_{J,k}(x)\right\}_{0\leq k\leq N-1}, \tag{3.2}$$

where $\tilde{\Psi}_{j,k}(x) = 2^{j/2}\Psi(2^j(x-k))$ and $\tilde{\Phi}_{j,k}(x) = 2^{J/2}\Phi(2^J(x-k))$ and the coefficients $\{D_{j,k}\}_{1\leq j\leq J, 0\leq k\leq N-1}$ and $\{S_{J,k}\}_{0\leq k\leq N-1}$ are defined as in equations (2.16) and (2.17).

In practice since the true function is unknown the noisy coefficients are computed by replacing $f(x)$ with the noisy signal Y. Note that similar to wavelet denoising the thresholding is only performed on detail coefficients, $D_{j,k}$

To perform denoising according to the proposed approach, we first threshold the noisy coefficients using the universal hard thresholding (Donoho, 1995):

$$b_{j,k} = \mathbb{I}\{|D_{j,k}| > \lambda\}; j = 1, ..., L; k = 0, ..., N-1 \tag{3.3}$$

where $\mathbb{I}$ is an indicator function, and $\lambda = \sqrt{2 \log N}\hat{\sigma}$, where $\hat{\sigma}$ is an estimator of standard deviation of noise present, and N is the size of the original signal. Given the redundancy of the transform, we estimate $\hat{\sigma}$ by averaging two estimators, which are sample standard deviations of wavelet coefficients at every odd and even locations, respectively, within the finest level of detail. Basically, if the magnitude of the coefficient is larger than the threshold, the function returns 1 and zero, otherwise. This thresholding function is applied on all the detail coefficients and the binary values, $b_{j,k}$, are computed and considered as the prior information for our Bayesian thresholding.

3.3.2 Computing Posteriors using Logistic Regression Model

As mentioned earlier, due to redundancy resulting from the AC Shell decomposition, the neighboring coefficients contain information about the significance of for each AC Shell coefficient. Therefore, this information should be incorporated in the denoising process.

Consider two immediate neighbor coefficients on the left and right sides of the coefficient of interest, $D_{j,k}$, as shown in Figure 3.1. That is, $D_{j,k-2}, D_{j,k-1}, D_{j,k+1}$, and $D_{j,k+2}$. The energy of these coefficients, i.e., the squared coefficients, are considered as the likeli-

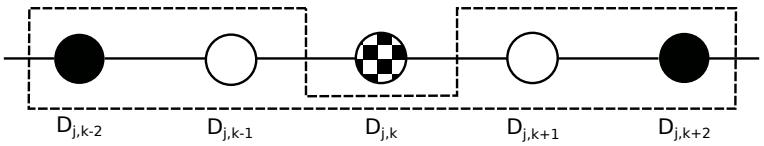


Figure 3.1: 2-steps coefficient neighbors in AC Shell decomposition

hood information that are integrated with the prior information obtained from thresholding the AC Shell coefficients, i.e., $b_{j,k}$. The posterior probabilities are defined by:

$$\begin{aligned} \Pr(\text{significant } D_{j,k} \mid D^2_{j,k-2}, D^2_{j,k-1}, D^2_{j,k+1}, D^2_{j,k+2}) \approx \\ l(D^2_{j,k-2}, D^2_{j,k-1}, D^2_{j,k+1}, D^2_{j,k+2}) \ \times \ \Pr(\text{significant } D_{j,k}). \end{aligned} \tag{3.4}$$

To compute the posterior probabilities, we fit a logistic regression model, in which the response data is the $b_{j,k}$ and the input is the energy of the neighboring coefficients. Specif-

ically, the logit of the posterior is given by:

$$\log\left(\frac{p_{j,k}}{1-p_{j,k}}\right) = \beta_0 + \beta_1 D^2_{j,k-2} + \beta_2 D^2_{j,k-1} + \beta_3 D^2_{j,k+1} + \beta_3 D^2_{j,k+2}, \tag{3.5}$$

where $p_{j,k} = P(b_{j,k}|D^2_{j,k-2}, D^2_{j,k-1}, D^2_{j,k+1}, D^2_{j,k+2})$, is the posterior probability for coefficient $D_{j,k}$. The coefficients β_0, β_1, β_2, β_3 are estimated by maximizing likelihood function (Myers et al., 2002).

The posterior probabilities are then thresholded by applying a universal hard thresholding function given by

$$c_{j,k} = \mathbb{I}\{p_{j,k} > z\}, \quad z \in (0,1). \tag{3.6}$$

The threshold z is determined by leave-one-out cross-validation such that the mean square error of signal reconstruction is minimized.

The final denoising step is performed by finding the union set of unthresholded prior and posteriors, i.e., $b_{j,k}$ and $c_{j,k}$, and setting the remainder coefficients to zero. That is:

$$D^{Neighbor}_{j,k} = \begin{cases} D_{j,k} & \text{if} \quad |D_{j,k}| \geq \lambda \quad or \quad p_{j,k} \geq z \\ 0 & otherwise \end{cases} \tag{3.7}$$

Lastly, the denoised AC Shell coefficients are transformed back to the original domain and the denoised signal, $\hat{f}(x)$ will be estimated.

3.4 Validation of the neighboring AC Shell Denoising using Simulations

In this section, we evaluate the performance of the proposed denoising methodology and compare it with benchmark methods using simulations. We consider four functions with different shapes and degrees of smoothness, namely, Doppler, Bumps, HeaviSine, and Blocks (Donoho and Johnstone, 1994a). Each simulated signal is of length $N = 2048$. A sample of each signal type is shown in Figure 3.2. For each signal type, 100 instances

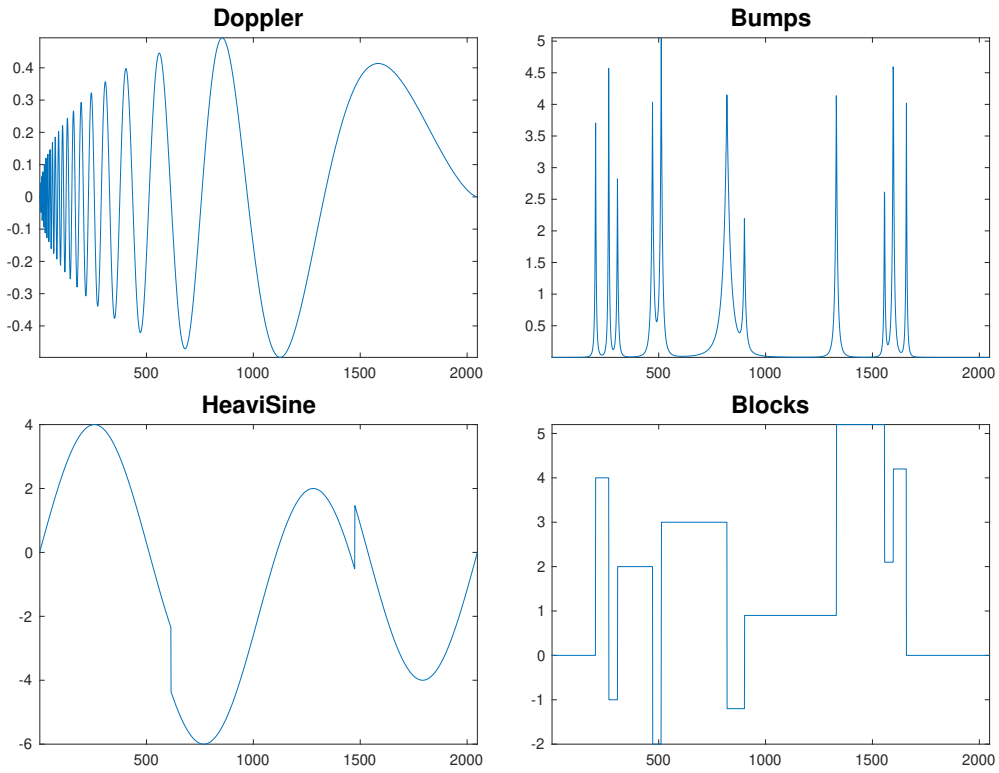
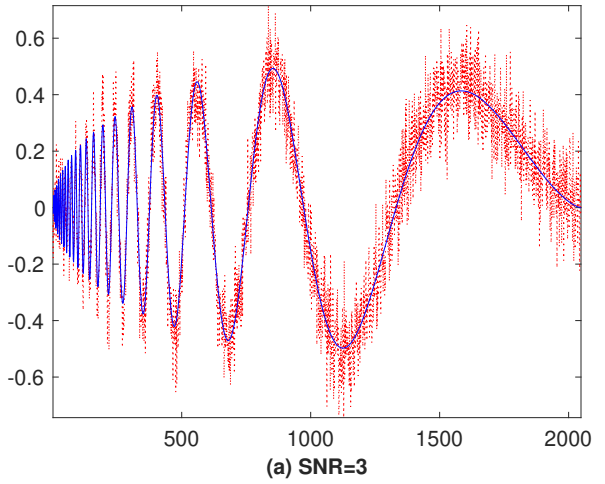


Figure 3.2: 4 different signals to check the performance of proposed method

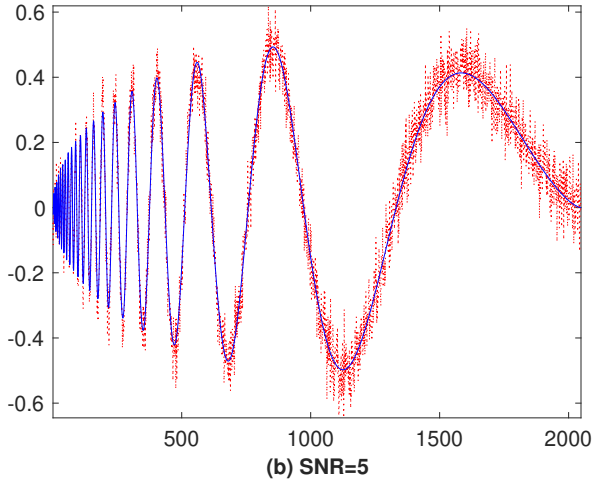
with three signal-to-noise ratios (SNR) of 3, 5, and 7 are generated. A sample of noisy Doppler signal with different SNR values are shown in Figure 3.3.

We analyze the simulated data using four methods and compare the results. Specifically, we apply 1) Wavelet denoising, 2) Neighboring Wavelet denoising, 3)AC Shell denoising, and 4) Neighboring AC Shell denoising to the simulated data with different SNRs and compute the mean squared error of reconstructed signal (i.e., $\|f(\hat{x}) - f(x)\|^2/N$). We use Daubechies 4 basis and 5 levels of decomposition. As an example, we show the denoised signals using each method against the true function for the Doppler signal with the SNR of 3 in Figure 3.4.

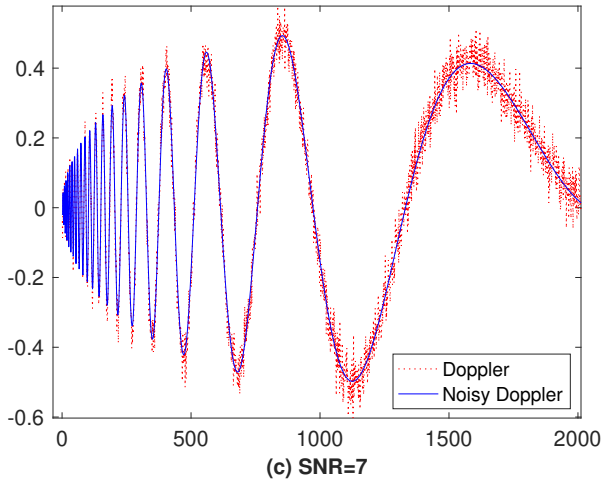
The MSE of each smoothing/denoising method for different signal type - SNR combinations are given in Figures **??**. As can be seen from these figures for the Doppler and HeaviSine where the true signals are smooth, AC Shell and Neighboring AC Shell significantly outperform the Wavelet-based benchmarks. For example, for the Doppler signal the



(a) SNR=3



(b) SNR=5



(c) SNR=7

Figure 3.3: Doppler Signal (blue line) with different amount of noises (red line) with (a) Signal-to-Noise SNR=3, (b) SNR=5 and (c) SNR=7

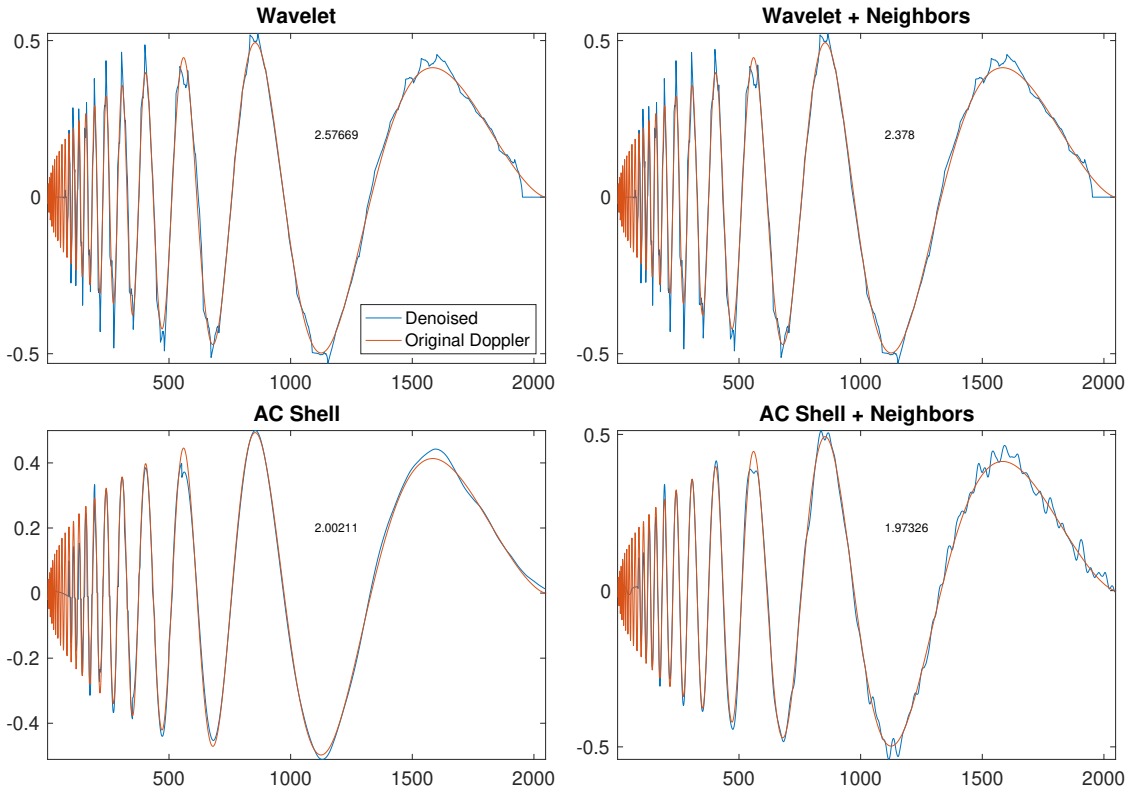


Figure 3.4: Compare denoised signal with the original Doppler in 4 different methods based on MSE

median MSE for Neighboring AC is around 2, while this value for Neighboring Wavelet is around 2.75. This can be attributed to the redundancy obtained by the AC Shell basis. However, for non-smooth signals like Bumps and Blocks, Wavelet-based methods have lower MSE values.

Moreover, by comparing the neighboring methods with the regular AC Shell and Wavelets denoising, we can see the value that neighbor coefficients brings to the denoising process. In all signal types neighboring methods have lowers MSEs than their non-neighboring counterparts. For example, for the Bumps signal the median MSE for Neighboring AC is around 19, while this value for AC Shell is more than 20. This indicates the value of including the neighboring information in the smoothing and denoising process.

Finally, we compare the performance of different methods under different SNR values. The foregoing observations are mostly consistent across different SNR values.

3.5 Case Study

In this section, we validate the proposed denoising method using real data. Specifically, we denoise a signal captured by an atomic force microscope. The atomic force microscopy (AFM) is a type of scanned proximity probe microscopy that measures the adhesion strength between two materials at the nano-newton scale. The AFM data from the adhesion measurements between carbohydrate and the cell adhesion molecule (CAM) E-Selection was collected by Bryan Marshall from the Department of Biomedical Engineering at Georgia Institute of Technology. The technical description and details provided in Marshall et al. (2001) is illustrated in Figure 3.10. A sample of the AFM signal is shown in Figure 3.9.

In AFM, a cantilever beam is adjusted until it bonds with the surface of a sample, and then, the force required to separate the beam and sample is measured from the beam deflection. Additionally, researchers are interested in the shape of the signal in the first segment (i.e., the first 350 observations), prior to cantilever detachment. Hence, identifying the drop point is an important part of the process. However, beam vibration can be caused by external factors such as thermal energy of the surrounding air or even the footsteps of someone outside the laboratory. In Figure 3.10 the vibration of a beam shows that the noise can mask the deflection signal and drop point. Therefore, it is important to first denoise the signal and then, identify the deflection point.

To denoise the AFM signal, we decomposed it of size 3,000 into 10 decomposition levels using the DWT with a 6-tab Daubechies wavelet (3 vanishing moments) and applied hard thresholding on wavelet coefficients. The threshold for this process is set as $\sqrt{2\log N}\hat{\sigma}$, where $\hat{\sigma}$ is an estimator of standard deviation of noise, and N is the size of the original signal. Given the redundancy of the transform, we estimate $\hat{\sigma}$ by averaging two estimators, $\hat{\sigma}$ and $\hat{e}$, which are sample standard deviations of wavelet coefficients at every odd and even locations, respectively, within the finest level of detail.

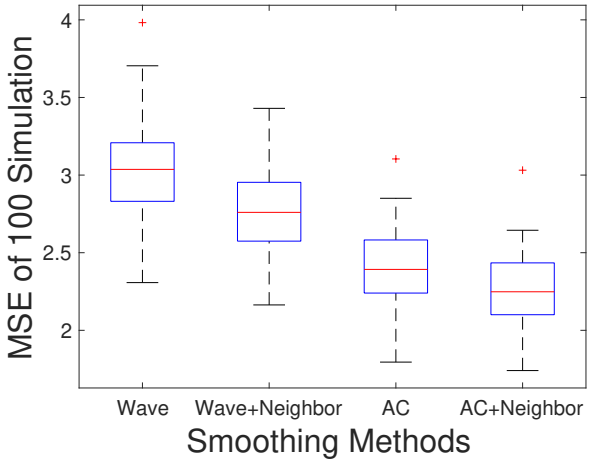
We apply four denoising methods discussed in the previous section and compare the

results. The original and denoised signals obtained from applying each method are plotted in Figure 3.11. As can be seen from the results, the AC Shell methods outperform the Wavelet-based methods in smoothing the signal. Additionally, the neighboring methods have slightly better performance compared with the original AC Shell and Wavelets denoising methods.

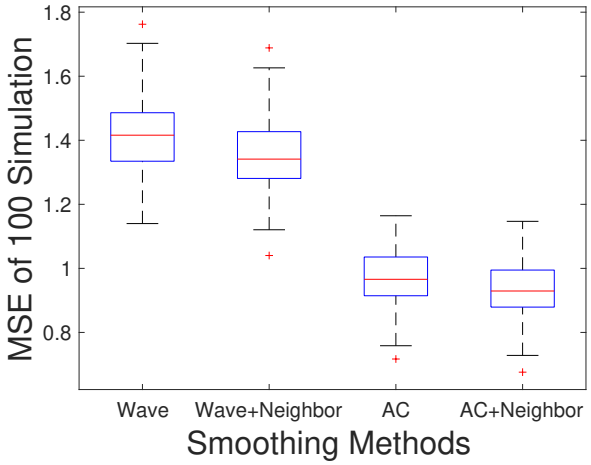
In short, both the simulations and case study show the superiority of the proposed method over the existing benchmarks and underline the importance of both information redundancy resulting from AC Shell basis as well as the neighboring information in denoising signals.

3.6 Conclusions

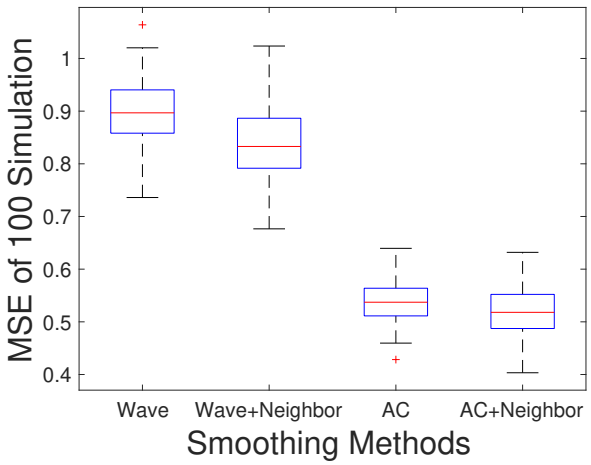
In this chapter, we utilized the redundant information property obtained by applying the AC Shells on noisy signals to devise a novel denoising/smoothing method. For this purpose, we proposed to incorporate the information of its neighboring AC Shell coefficients in identifying whether a coefficient is significant or should be removed. A Bayesian framework was proposed that combines the thresholded coefficients as the prior, with the energy of neighboring coefficients as the likelihood information to obtain the posterior probability of thresholding a coefficient. We used both simulations and a case study of analyzing and smoothing the atomic force microscopy signals to evaluate the performance of the proposed denoising method. The results indicated that the proposed method outperforms existing denoising methods for different signal-to-noise ratios. The results showed that if the signal is not smooth the proposed Neighboring AC Shell method performs poorly. The extension of this method to be applicable to non-smooth signal is an important topic that requires further research.



(a) Doppler, SNR=3

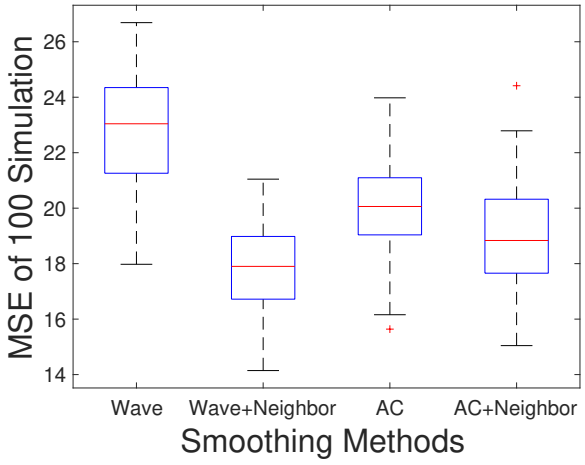


(b) Doppler, SNR=5

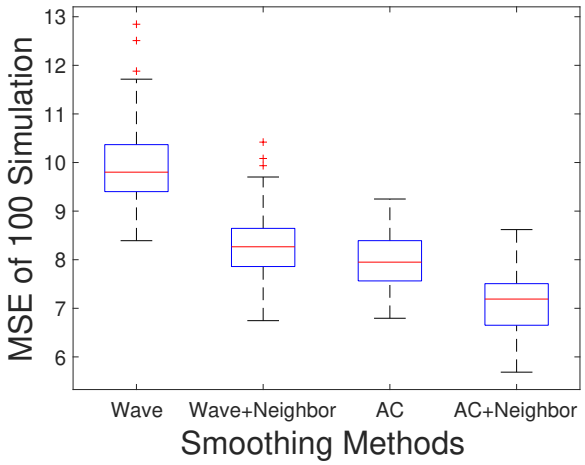


(c) Doppler, SNR=7

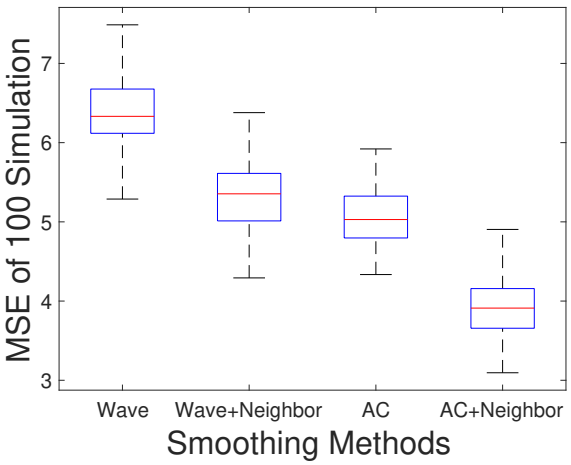
Figure 3.5: Boxplot of Mean Squared Error (MSE) for 4 different denoising methods (smoothing) for noisy Doppler signal with (a) $SNR = 3$, (b) $SNR = 5$ and (c) $SNR = 7$



(a) Bumps, SNR=3

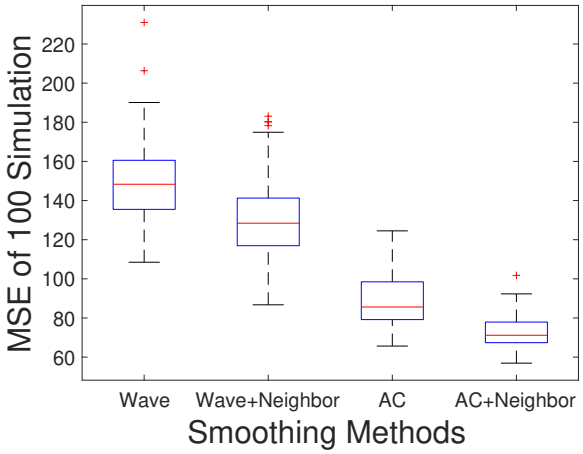


(b) Bumps, SNR=5

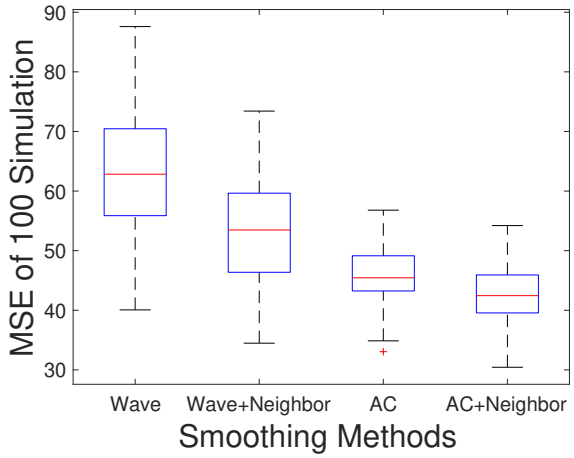


(c) Bumps, SNR=7

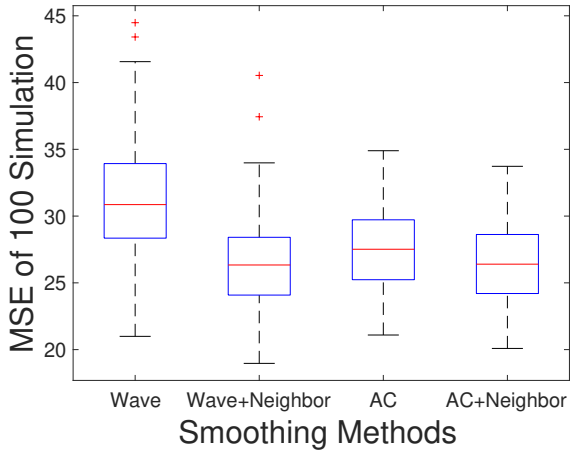
Figure 3.6: Boxplot of Mean Squared Error (MSE) for 4 different denoising methods (smoothing) for noisy Bumps signal with (a) $SNR = 3$, (b) $SNR = 5$ and (c) $SNR = 7$



(a) HeaviSine, SNR=3

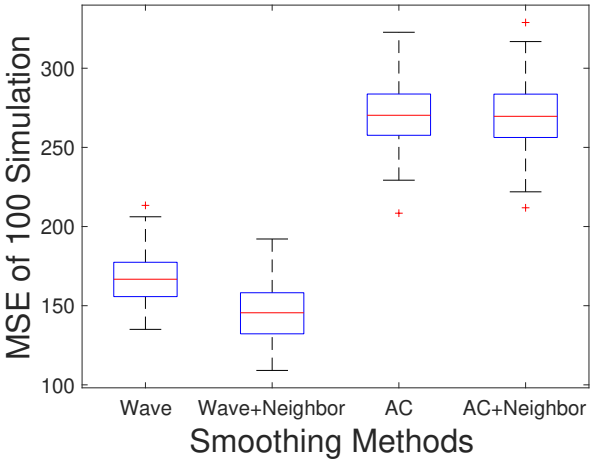


(b) HeaviSine, SNR=5

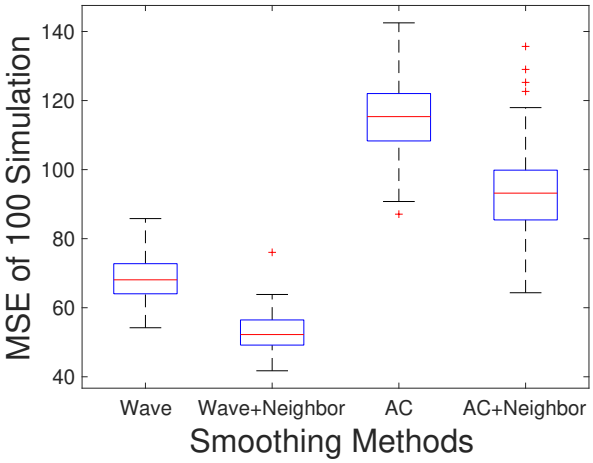


(c) HeaviSine, SNR=7

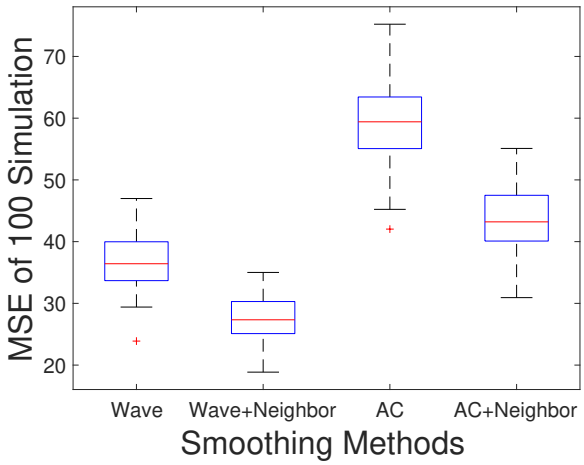
Figure 3.7: Boxplot of Mean Squared Error (MSE) for 4 different denoising methods (smoothing) for noisy HeaviSine signal with (a) $SNR = 3$, (b) $SNR = 5$ and (c) $SNR = 7$



(a) Blocks, SNR=3



(b) Blocks, SNR=5



(c) Blocks, SNR=7

Figure 3.8: Boxplot of Mean Squared Error (MSE) for 4 different denoising methods (smoothing) for noisy Blocks signal with (a) $SNR = 3$, (b) $SNR = 5$ and (c) $SNR = 7$

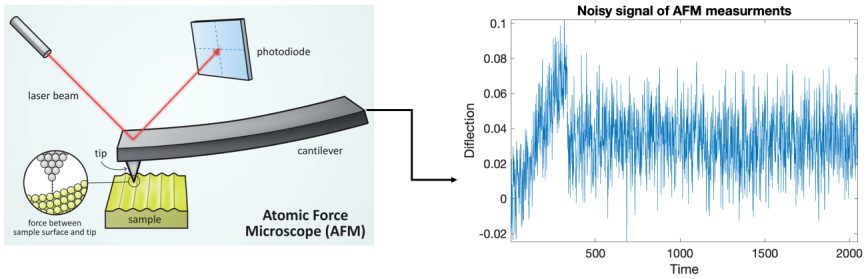


Figure 3.9: AFM illustration and a sample signal

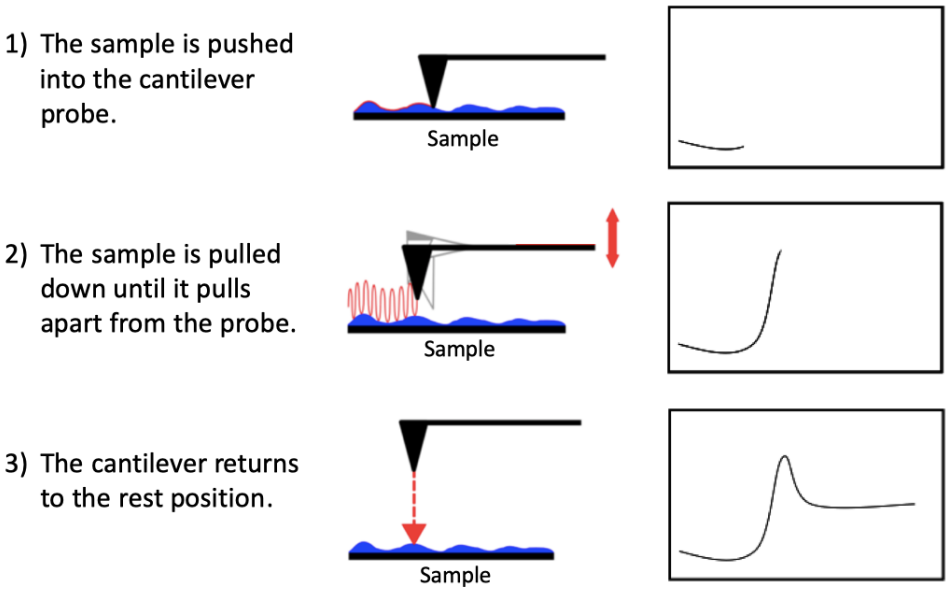


Figure 3.10: Steps of collecting data from Atomic Force Microscopy (AFM) and a sample signal

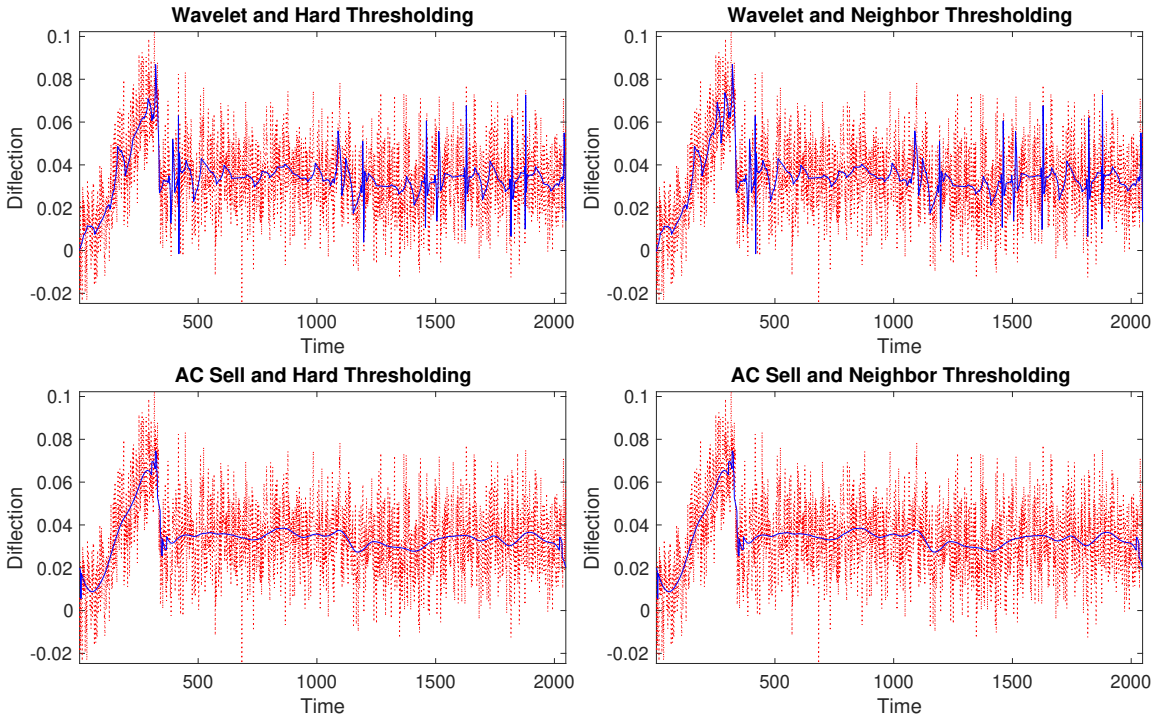


Figure 3.11: Denoising AFM measurements signal with both Hard and Neighbor thresholding for Wavelet and AC Shell decomposition with Daubechies 4 wavelet

CHAPTER 4

BAYESIAN METHOD IN COMBINING GENETIC AND HISTORICAL RECORDS OF TRANSATLANTIC SLAVE TRADE IN THE AMERICAS

4.1 Introduction

In the era between 1515 and 1865, more than 12 millions people were enslaved and forced to move from Africa to North and Latin America, which has had significant social, cultural, health and genetic impacts across the Americas. The shipping documents have recorded the origin and disembarkation of enslaved people. For example, the data show that more than 10 millions enslaved people disembarked in Central America, South America, and the Caribbean, and fewer than half a million disembarked in North America (Eltis, 2007).

However, over time due to slave trades they have been moved across North America. This makes identification of African American's origins particularly challenging. The genealogy study that focuses on tracing one's family ancestry and origins is an ancient human desire (Potter-Phillips, 1999). Traditionally, genealogy study has been done via the exploration of historical records, family tress and birth certificates. Due to recent advancements in the field of genetics, genealogy has been revolutionized and become more accurate by DNA marker-based methods (Aulicino, 2013; Fitzpatrick and Yeiser, 2005). DNA sequencing provides accurate, unbiased and sensitive markers measuring the relationships among family members in the family trees, and helps identify individual ancestral origins. The use of DNA-markers are more pronounced when there is a lack of historical ancestral records (Gates Jr., 2010).

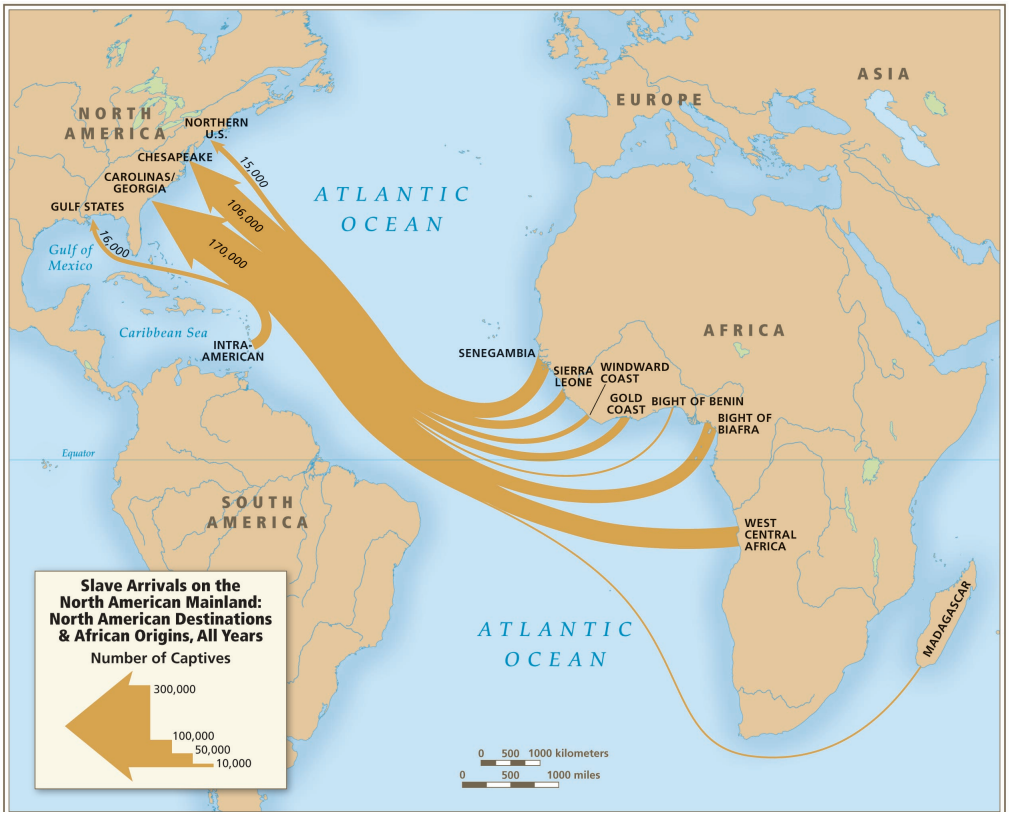


Figure 4.1: Slave arrivals on the North American mainland: North American destinations and African origins, all years. Attributed to *https:// tracingafricanroots .wordpress.com/*

Mitochondrial DNA (mtDNA) and Y-DNA sequences (haplotypes) are two most popular markers widely used for genetic genealogy (Cann et al., 1987; Pakendorf and Stoneking, 2005; Stumpf and Goldstein, 2001). mtDNA and Y-DNA are sex-specific markers that help with determining the female and male lineages, respectively. Additionally, they can show geographical differentiation that localizes the ancient ancestral origins. However, as discussed in the literature (Emery et al., 2015; Salas et al., 2005; Stefflova et al., 2011), mtDNA and Y-DNA genetic markers have some limitation for genealogical studies. They are only capable of capturing single unbroken ancestral lineages, hence, can be used for fully identifying an individual ancestry. Additionally, although these markers can provide continental differentiation, they have poor spatial resolution that makes it hard to localize ancestry assignment as these markers are distributed across different sub-continental regions.

To overcome the foregoing drawbacks, a hybrid approach has increasingly become more relevant that combines the genetic markers results with the historical records that show the transatlantic journeys of enslaved people. Addition of the journey data can provide with substantially increased resolution in ancestry. Historical information could also be combined with genetic information at the population level to increase confidence in genetic-based ancestry assignments. However, in the genealogy context, there is little research for the integration of historical and genetic data. Rishishwar et al. (2015) used a Bayesian approach for the combination of population-level historical records with genetic marker data for determining the ancestry of Afro-Colombian. Micheletti et al. (2020) used Bayesian modeling to compare the timing of genetic connections between African populations and individuals from the Americas with historical records of the transatlantic shipping of enslaved Africans.

In this chapter, we propose a Bayesian modeling framework to integrate genetic data, namely genome sequences as well as genotypes, and its geographic distribution in Africa with historical records of the transatlantic shipping of enslaved Africans to increase the spatial resolution of ancestry assignments for African-Americans. The proposed Bayesian framework uses the the voyage data from historical records available in the transatlantic slave trade database as prior probabilities and combine them with genetic markers of Afro-Americans, considered as the likelihood information to estimated the posterior (updated) probabilities of their ancestry assignments to geographical regions in Africa.

4.2 Dataset Description

4.2.1 Genome Sequence Data and Admixture Analysis:

A total of 427 whole genome sequences or genotypes, taken form the 1000 Genomes Project (1KGP) (McVean, 2010, 2012) and Human Genome Diversity Project (HGDP) . Specifically, the Mandenka population were taken from the HGDP, all the rest from 1KGP.

Number of cases from each region showed in Table 4.1

Table 4.1: Number of whole genome sequences from each region

Regions	GWD	MSL	ESN	Mandenka	YRI	Sum
Number of cases	113	85	99	22	108	427

Whole genome sequences for 60 admixed Afro-American from the 1000 Genomes Project. Whole genome sequence variant data from 1KGP individuals were merged with genome genotype data from HGDP individuals using the program PLINK version 1.9 (Purcell et al., 2007), keeping only those sites common to both datasets and correcting SNP strand orientations for consistency as needed. These processes were done separately for genome sequence and genotype data together and for genome sequence data alone.

Allele sharing distances between pairs of genomes were computed as the fraction of differences between SNP calls. The program ADMIXTURE v1.23 (Alexander et al., 2009) was used to estimate the admixture fractions of six putative ancestral populations among Afro-American genome sequences. ADMIXTURE was run with default settings and k = 6 ancestral populations. The program SupportMix (Ver. Jul 18, 2012) (Omberg et al., 2010) was used to characterize the regional (locus-specific) ancestry admixture fractions in the Afro-American genomes using default settings. A sample of these fractions is given in Table 4.2. Also, the references along with corresponding regions and the number of unrelated individuals are summarized on Table 4.3.

Table 4.2: Copying fractions from African American individuals (ASW), and reference African populations (ESN, GWD, LWK, Mandenka, MSL and YRI)

Individuals	GWD	MSL	ESN	Mandenka	YRI	LWK
ASW1	0.1559312	0.1467045	0.2455052	0.03247513	0.2478764	0.1715075
ASW2	0.1693882	0.1428647	0.2600602	0.02752310	0.2499406	0.1502233
ASW3	0.1820081	0.1438091	0.2217078	0.03044837	0.2353801	0.1866465
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
ASW60	0.1623608	0.1451124	0.2368917	0.03243721	0.2485918	0.1746061

Table 4.3: References and related regions plus the number of unrelated individuals

Reference	Region	Num. of Individuals
GWD	Gambian in Western Divisions or Senegambia	113
MSL	Sierra Leone	85
Mandenka	Windward Coast and Gold Coast	22
ESN	Esan in Nigeria	99
YRI	Yoruba in Ibadan, Nigeria	108
LWK	Luhya in Webuye, Kenya	

4.2.2 Historical Records and Transatlantic Voyages Data

Historical data on the African ancestral origins and voyages of enslaved people from those origins of the modern Afro-American population, compiled from records of trans-Atlantic slave voyages available on slave voyages website . The dataset covers voyages from 1626 to 1875 originated from 6 geographical regions in Africa, namely, Senegambia and off-shore Atlantic, Sierra Leone, Windward Coast and Gold Coast, Bight of Benin, Bight of Biafra and South-east Africa and Indian ocean islands. The total numbers of voyages from each in this 250 years time period are, 111822, 54339, 92947, 11456, 82726 and 10551 respectively. The detailed voyage statistics for 25 year periods are given in Table 4.4.

Table 4.4: Number of voyages from each region in the time period of 1626 till 1875

Year	GWD	MSL	Mandenka	ESN	YRI	LWK	Totals
1626-1650	0	0	0	0	0	0	0
1651-1675	2403	0	0	0	1627	0	4030
1676-1700	4884	0	726	573	5519	2604	14306
1701-1725	11571	735	10789	2875	17741	3342	47053
1726-1750	32508	3490	11147	1506	35799	527	84977
1751-1775	41135	21171	39337	4518	16027	381	122569
1776-1800	8505	10063	12834	510	395	0	32307
1801-1825	10816	18880	18114	1348	5513	3697	58368
1826-1850	0	0	0	0	105	0	105
1851-1875	0	0	0	126	0	0	126
Totals	111822	54339	92947	11456	82726	10551	363841

4.3 Ancestry Assignments to Geographical Regions using a Bayesian Approach

In this section, we describe our proposed Bayesian method for ancestry assignments of Afro-Americans to geographical regions in Africa. The Bayesian method incorporates the historical records of transatlantic voyages that provide prior knowledge about the origins, with the the regional ancestry fractions of individuals obtained from the number of chunks of genome matches with references, considered as the likelihood information to compute the posterior probabilities. For any given Afro-American individual with genome data, their African ancestral origin can be assigned by finding the posterior probability of coming from any of the six ancestral regions.

According to the Bayes' rule,

$$\Pr(Region \mid Regional\ ancestry\ fractions) = \frac{\Pr(Regional\ ancestry\ fractions \mid Region) \times \Pr(Region)}{\Pr(Regional\ ancestry\ fractions)}, \tag{4.1}$$

where:

$\Pr(Region \mid Regional\ ancestry\ fractions)$ is the posterior probability,

$\Pr(Regional\ ancestry\ fractions \mid Region)$ is the likelihood,

$\Pr(Region)$ is the prior probability, and

$\Pr(Regional\ ancestry\ fractions)$ is the marginal genome probability considered as a normalizing constant.

We assume that the likelihood function of the number of regional ancestry matches or chunks that copy region k follows a multinomial distribution. Specifically, let $f_{ik}; i = 1, 2, ..., 60, k = 1, 2, ..., 6$ denote the number of chunks that copies region k for the individ-

ual i. That is,

$$f_{i1}, ..., f_{i6} | (p_{i1}, ..., p_{i6}) \sim \mathrm{MN}\Big(N, (p_{i1}, ..., p_{i6})\Big), \tag{4.2}$$

where $N = \sum_{k=1}^{K} f_{ik}$, and p_{ik} is the probability of copying from region k for sample i. The probability mass function is given by

$$g\Big((f_{i1}, ..., f_{i6}) | (p_{i1}, ..., p_{i6})\Big) = \frac{N!}{f_{i1}! ... f_{i6}!} p_{i1}^{f_{i1}} p_{i2}^{f_{i2}} ... p_{iK}^{f_{iK}}. \tag{4.3}$$

We also consider the conjugate prior for multinomial distribution, namely Dirichlet, i.e., the probability of a voyage from area k has a Dirichlet distribution $(p_{i1}, ..., p_{iK}) \sim \mathrm{Drich}\Big(\alpha_{i1}, ..., \alpha_{iK}\Big)$. The probability density function of the prior is given by

$$h\big(p_{i1}, ..., p_{iK}\big) = \frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \prod_k p_{ik}^{\alpha_k - 1}, \tag{4.4}$$

where $p_{ik} \in (0, 1)$ and $\sum_k p_{ik} = 1$.

As Dirichlet is a conjugate prior for the Multinomial likelihood, the posterior distribution would have a close-form with Dirichlet distribution, i.e.,

$$\begin{aligned} \big(p_{i1}, ..., p_{iK}\big) \mid \big(f_{i1}, ..., f_{iK}, &\alpha_{i1}, ..., \alpha_{iK}\big) \\ &\sim \mathrm{Drich}\Big(\alpha_1 + f_{i1}, ..., \alpha_K + f_{iK}\Big) \end{aligned} \tag{4.5}$$

The posterior pdf is written by

$$l\big(p_{i1}, ..., p_{iK}\big) = \frac{\Gamma(\sum_k (\alpha_k + f_{ik}))}{\prod_k \Gamma(\alpha_k + f_{ik})} \prod_k p_{ik}^{\alpha_k + f_{ik} - 1}, \tag{4.6}$$

where $p_{ik} \in (0, 1)$ and $\sum_k p_{ik} = 1$.

4.3.1 Estimation of Hyperparameter of Prior Distribution

We deal with the unknown prior hyperparameters, α_k's, in two different ways: First, we use Markov Chain Monte Carlo (MCMC) simulations and generate α_k's from a non-informative prior, namely a wide uniform distribution. The generated α_k's are used to generate prior probabilities which are in turn utilized along with f_{ik}'s to draw from the posterior distributions, $p_1, ..., p_6$. The average posterior draws are used to estimate the posterior probabilities and consequently assign an individual's ancestry region.

Second, we use empirical Bayes method. Using the method of moment, and the voyage data, we can estimate the hyperparameters. Specifically, the following set of equations are used to estimate the hyperparameters, α_k.

$$\mathbb{E}(p_k) = \frac{\alpha_k}{\sum_k \alpha_k} \tag{4.7}$$

$$\mathrm{Var}(p_k) = \frac{\mathbb{E}(p_k)\big(1 - \mathbb{E}(p_k)\big)}{1 + \sum_k \alpha_k} \tag{4.8}$$

$$\log \sum_k \alpha_k = \frac{1}{K-1} \sum_{k=1}^{K-1} \log \left(\frac{\mathbb{E}(p_k)\big(1 - \mathbb{E}(p_k))\big)}{\mathrm{Var}(p_k)} - 1 \right) \tag{4.9}$$

$\mathbb{E}(p_k)$ and $\mathrm{Var}(p_k)$ are estimated from the voyage data. $\hat{\mathbb{E}}(p_k) = \frac{x_k}{n}$, where x_k is the number of the voyages from region k, and n is the total number of voyages. $\hat{\mathrm{Var}}(p_k)$ is estimated by the sample variance, in which estimated $\mathbb{E}(p_k)$ from each 25 year period is considered as an observation.

Finally, using the estimated α_k's, and the likelihood observations, f_{ik}'s, the posterior means are estimated by

$$\mathbb{E}(p_{ik}) = \frac{\hat{\alpha}_k + f_{ik}}{\sum_k (\hat{\alpha}_k + f_{ik})}. \tag{4.10}$$

The region with the maximum mean a posterior is assigned to the corresponding individual.

4.4 Results and Discussions

In this section, we analyze the genome data combined with the voyage records using the proposed Bayesian methods, namely the Empirical Bayes and MCMC method, to localize ancestry assignment for individual Afro-Americans. Additionally, we perform a validation study on the genome data of the references sample to show the value of including prior information in the analysis.

4.4.1 Posterior Probabilities Results using Empirical Bayes

As discussed in the previous section, we use the method of moment to estimate the hyper-parameter α for each region using the the voyage data. We should note that since the voyage data are given for the periods of 25 years, we estimate the $\mathbb{E}(p_k)$ for each period and use these estimates to find the sample variance as the estimate for $\mathrm{Var}(p_k)$. The point estimations for prior probabilities, variances, and the hyper-parameters α for each region are given in Table 4.5.

Table 4.5: Estimation of $\mathbb{E}(p_k)$, $\mathrm{Var}(p_k)$ and the hyperparameters α

Estimate	ESN	GWD	LWK	Mandenka	MSL	YRI
$\mathbb{E}(p_k)$	0.0315	0.3073	0.0290	0.2555	0.1493	0.2274
$\mathrm{Var}(p_k)$	0.0002	0.0047	0.0018	0.0092	0.0135	0.0230
α	0.6846	6.6825	0.6305	5.5545	3.2473	4.9437

As the posterior distribution of the assignment probabilities is Dirichlet, its parameters are estimated by combining the number of matches with the estimated priors. That is, $\mathrm{Dirichlet}\Big(\alpha_{i1} + f_{i1}, ..., \alpha_{iK} + f_{iK}\Big)$. The f_{ik} is computed by multiplying the fraction matches, obtained by the "Supportmix" package, with $N = 400$. For each individual i the mean a posteriori values are considered as the posterior assignment probabilities, i.e., $\mathbb{E}(p_{ik}) = \dfrac{\hat{\alpha}_k + f_{ik}}{\sum_k(\hat{\alpha}_k + f_{ik})}$. The final assignment is done by finding the region with the maximum posterior probability, i.e., $R_i = argmax_k\Big(\mathbb{E}(p_{ik})\Big)$.

The posterior assignment probabilities for each individual along with the assignment probabilities obtained by only considering the genome data are illustrated in Figures 4.2, 4.3, and 4.4 using bar charts. The impact of the prior information on the assignment probabilities is clear from these figures. As can be seen for some regions inclduing the GWD and Mandenka these probabilities increase, while decreasing for some others such as LWK.

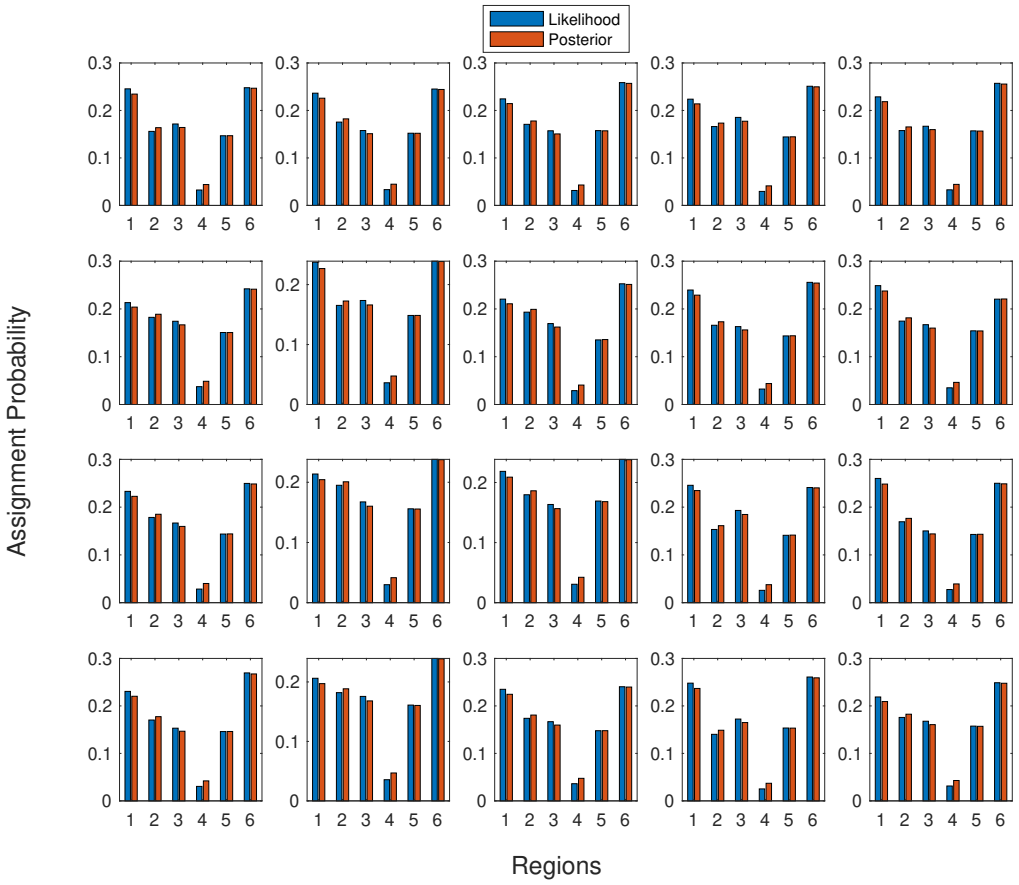


Figure 4.2: The probability of ancestry assignment for individuals 1 to 20, ordered from top left corner to bottom right

To have a clearer comparison between posterior probabilities and likelihood, we study the probability distributions across the individuals by plotting two sets of boxplots in Figure 4.9.

As can be seen from the figures, the prior information has increased the assignment

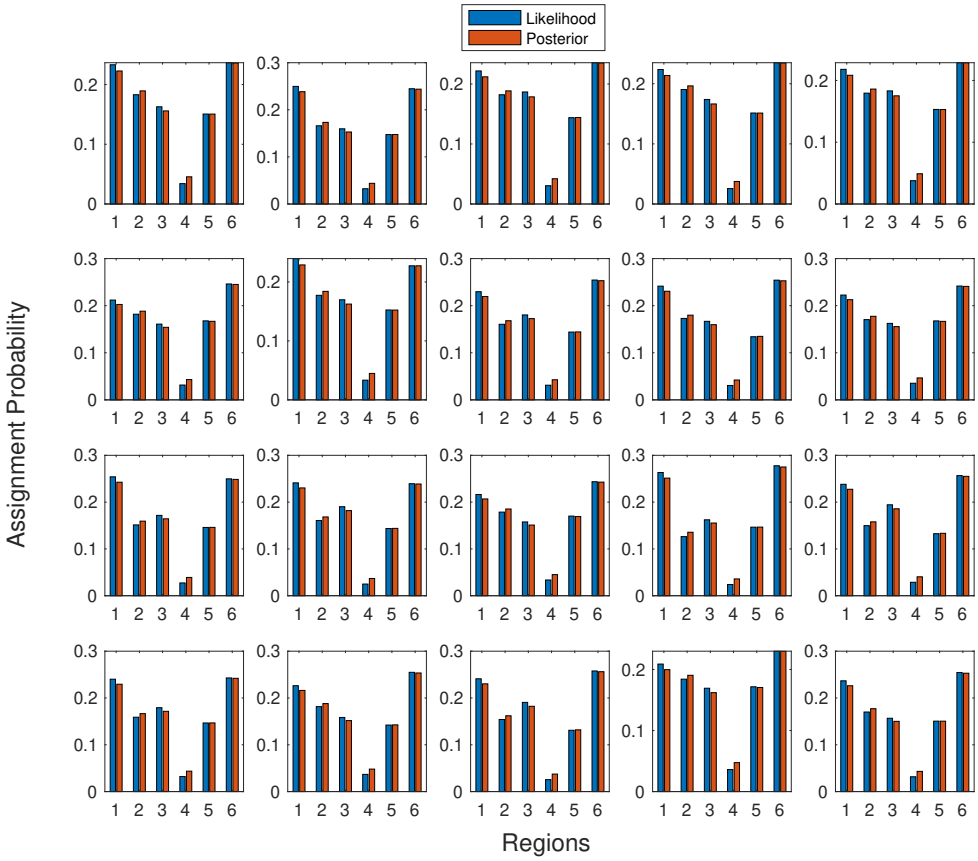


Figure 4.3: The probability of ancestry assignment for individuals 21 to 40, ordered from top left corner to bottom right

probabilities for some regions including GWD, and Mandenka. To further highlight the impact of the prior information, we plot the confusion matrix in Figure 4.6 that shows how many individuals' assignments have changed due to prior information. Based on the confusion matrix and the boxplots, it is clear that 11 individuals were assigned to the ESN region, 7 of whom were reassigned to the YIR region after incorporating the prior information. 49 individuals are assigned to YRI based on only likelihood information. However, due to a high prior probability (high voyage frequency) of YRI, the assignments based on posterior probability do not change.

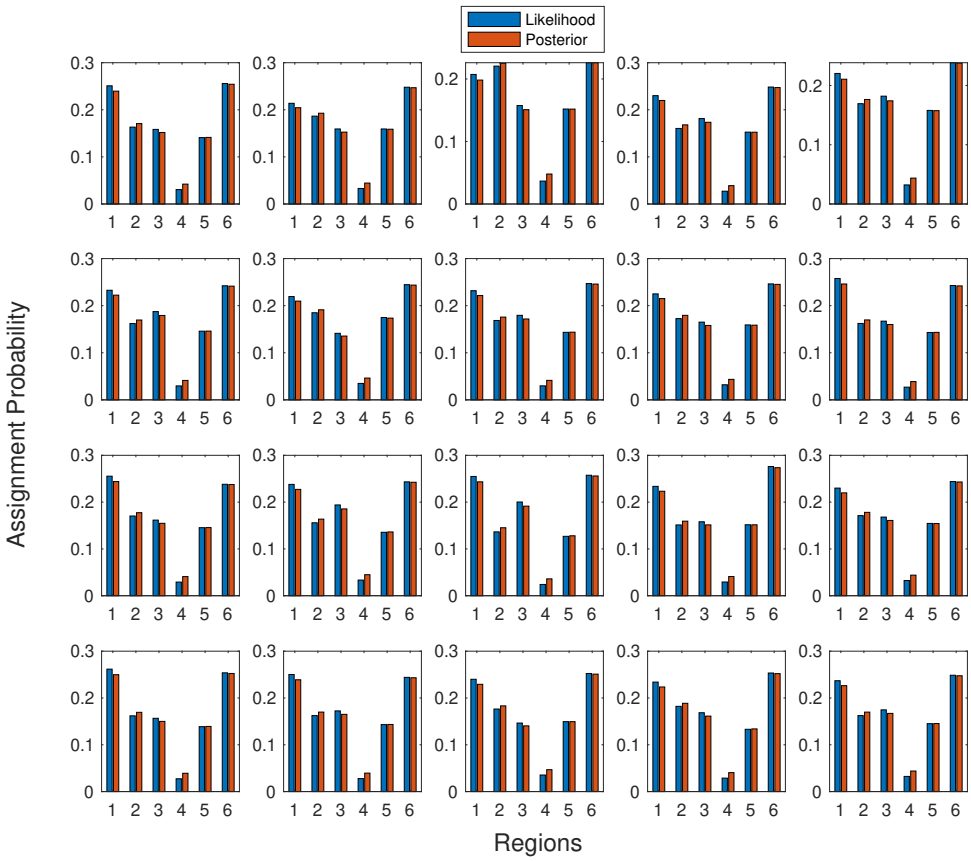


Figure 4.4: The probability of ancestry assignment for individuals 41 to 60, ordered from top left corner to bottom right

4.4.2 Validation study for Empirical Bayes

In this section, we use the reference individual genome data to validate the importance of including prior information in empirical Bayes approach using simulations. There are 427 reference individuals in this study, we use sampling with replacement to sample from reference individuals with the proportions defined by voyage data. We generate populations of 1000 individuals whose origins are known. Then, we apply the empirical Bayes discussed in the previous section to estimate the posterior probability and consequently determine the origin of simulated individuals. Next, the assigned origin is compared with the true origin to compute the accuracy. We repeat this procedure 1000 times.

The average confusion matrices for both likelihood-based method and posterior-based

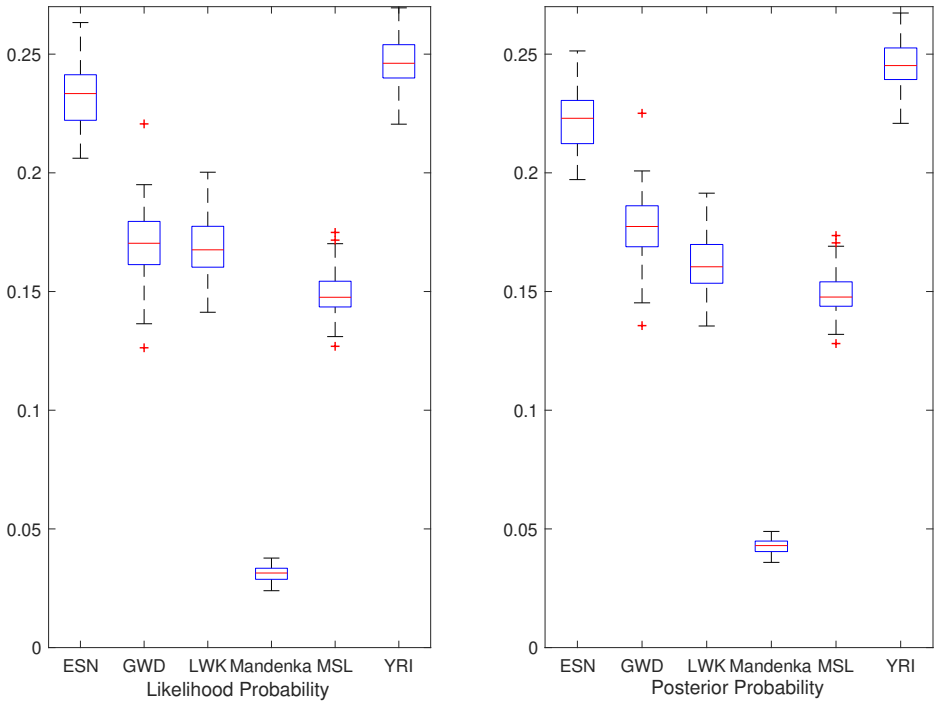


Figure 4.5: The probability of ancestry assignment for different regions, left panel: likelihood without priors, right panel: posterior probabilities

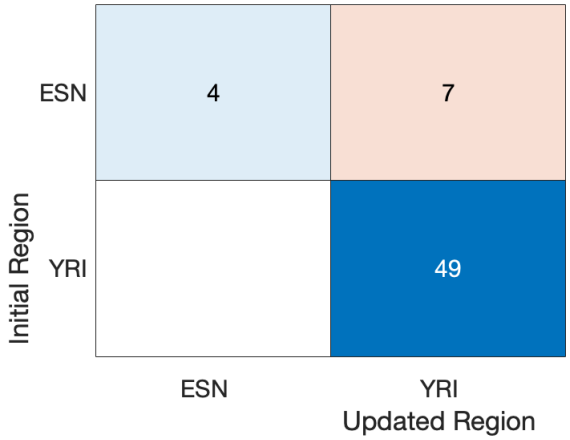


Figure 4.6: The confusion matrix of likelihood assignment vs posterior assignment

method are plotted in Figures 4.8 and 4.7, respectively. Comparing the two confusion matrices clearly shows the accuracy of the posterior-based assignment is more than that of the likelihood-based method. On average, 5 individuals that are mis-classified using the likelihoods are correctly classified using the posterior probabilities. This shows the

importance of the prior information in making more accurate determination of one's origin.

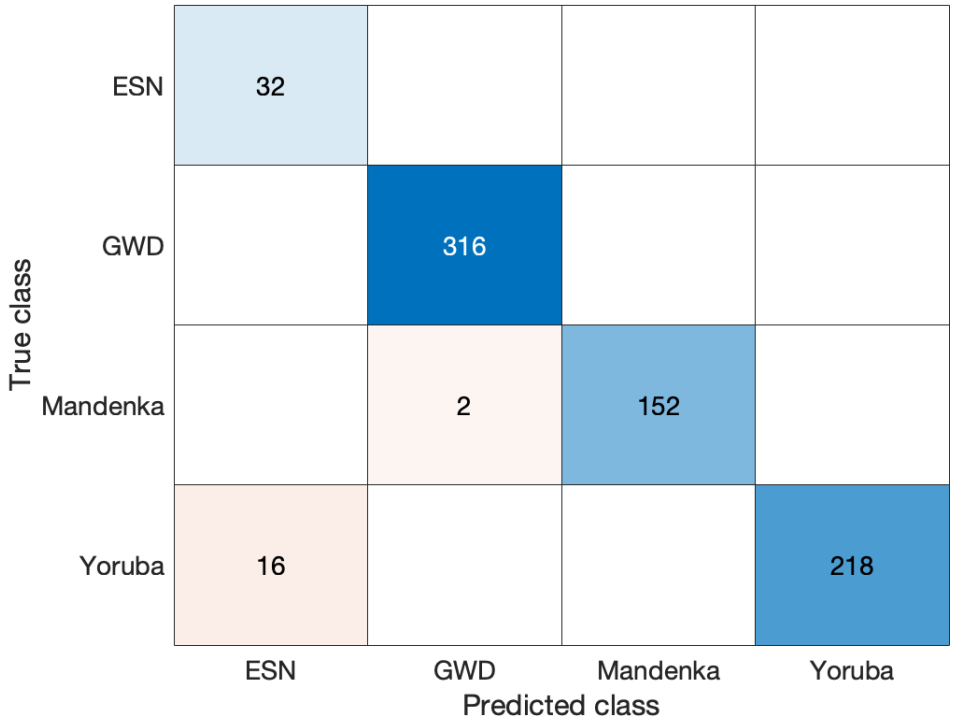


Figure 4.7: The confusion matrix of likelihood-based approach

To see the distribution of the errors for each method, we plot boxplots of 1000 error values in Figure 4.9. The boxplot clearly shows that the posterior-based method outperform the likelihood-based method. The median improvement is about 1%.

4.4.3 Posterior Probabilities Results using MCMC

In this approach, random draws are made from a non-informative uniform distribution for the prior hyper-parameters, i.e., $\alpha_k \sim \text{Uniform}(0.1, 5000)$. Then an MCMC sampling scheme is followed to obtain the posterior probabilities.

The resulting probability of ancestry assignment to a region obtained by using only the genome data and MCMC are shown in Figures 4.10 and 4.11, respectively. As can be seen from these figures the assignment probabilities for both methods are very similar. This is mainly because unlike the empirical Bayes approach that uses the voyage data to estimate

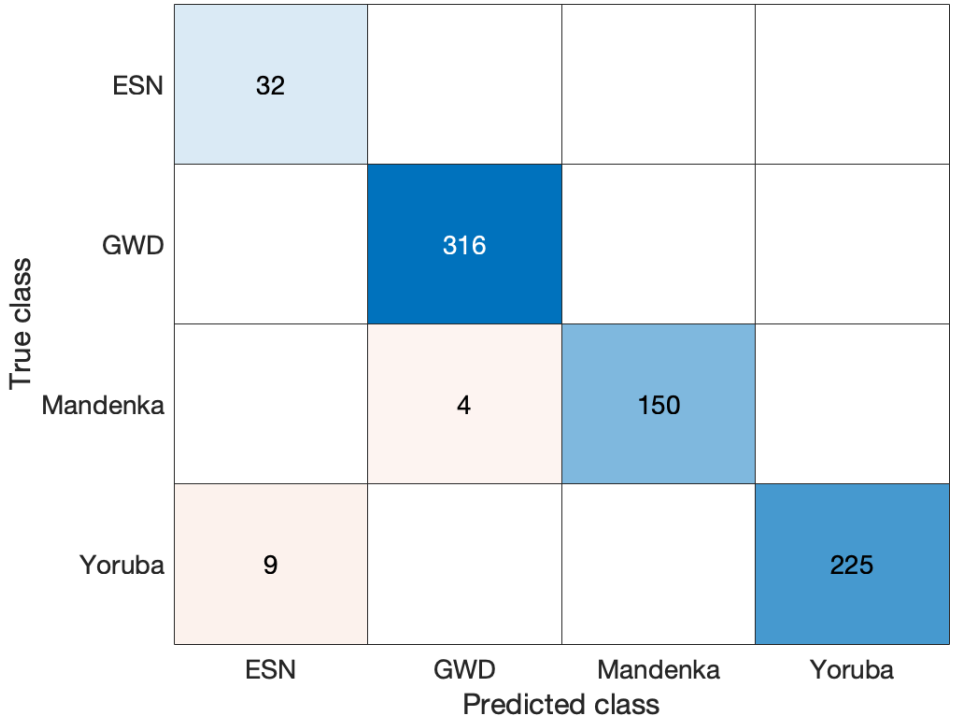


Figure 4.8: The confusion matrix of posterior-based approach

the hyper-parameters (informative priors), the non-informative flat priors chosen for sampling the Dirichlet's parameters add almost no information to the likelihood function and the genome data. This again emphasizes on the importance of the prior information.

4.5 Conclusions

Recently, due to advancements in the field of genetics, genealogy has been revolutionized and become more accurate by DNA marker-based methods. However, the poor spatial resolution of DNA marker-based methods makes it hard to localize ancestry assignment. To overcome the issue, in this chapter, we utilized Bayesian methodology to propose a hybrid approach that combines the genetic markers results with the historical records that show the transatlantic journeys of enslaved people.

The proposed methodology consists of two methods; the empirical Bayes, in which the hyper-parameters of the prior distribution are estimated using the data, and the MCMC

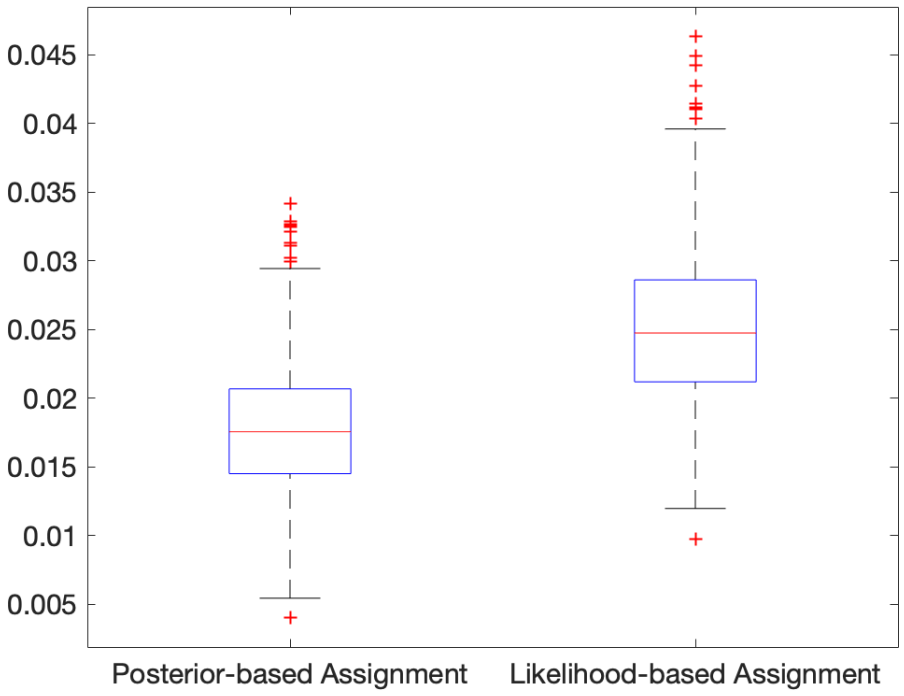


Figure 4.9: Boxplots of 1000 error values for each method (Likelihood versus Posterior)

method that assumes non-informative priors. We applied the proposed methodology to transatlantic voyage data and a sample of genome data from 60 Afro-American individuals. We showed the effectiveness of the proposed methodology and the importance of prior information in increasing the accuracy of ancestry assignment. The results showed that the empirical Bayes can improve ancestry assignment, while the MCMC that uses non-informative priors has little impact on the assignment and does not add much to the genome data.

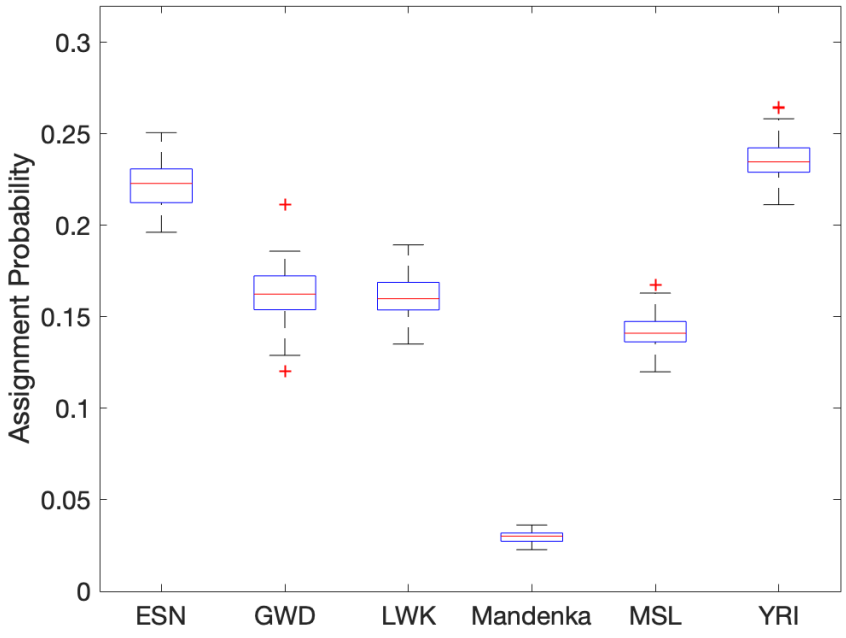


Figure 4.10: The probability of ancestry assignment to a region obtained by using only the genome data

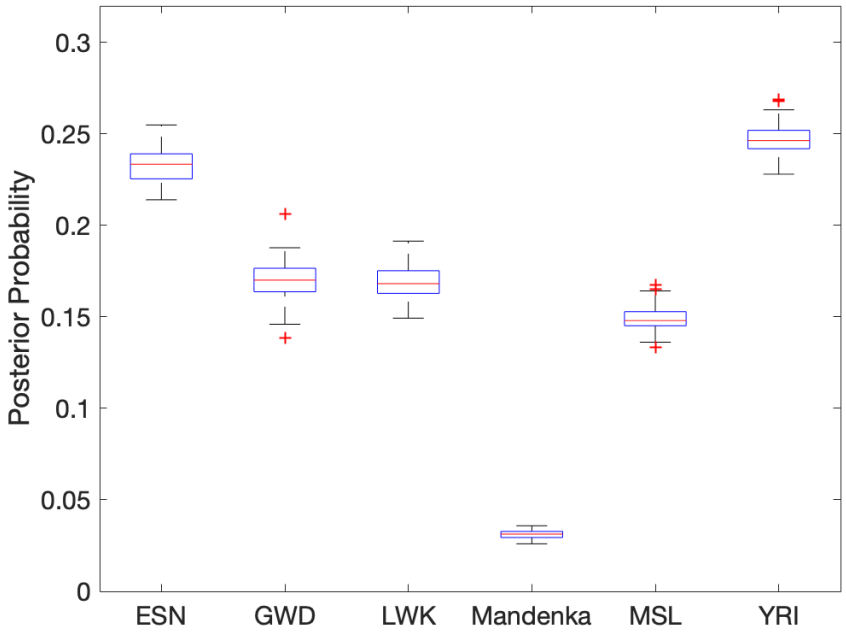


Figure 4.11: The probability of ancestry assignment to a region obtained by using MCMC

Appendices

APPENDIX A

AC SHELL DETAIL COEFFICIENTS ENERGY

$$\mathbb{E}(D_{j,k}^2) = \mathbb{E}\Big\{\Big(\int f_d^j(y)2^j\psi\big(2^j(y-k)\big)dy\Big)\Big(\int f_d^j(z)2^j\psi\big(2^j(z-k)\big)dz\Big)\Big\} \qquad (A.1)$$
$$= 2^{2j}\int\int \mathbb{E}\Big(f_d^j(y)f_d^j(z)\Big)\psi\big(2^j(y-k)\big)\psi\big(2^j(z-k)\big)dydz$$

Now, we calculate $\mathbb{E}\Big(f_d^j(y)f_d^j(z)\Big)$ and then we will replace it in (A.1).

$$\begin{aligned}
\mathbb{E}\Big(f_d^j(y)f_d^j(z)\Big) &= \mathbb{E}\Big(\sum_{n=0}^{N-1}\tilde{d}_{j,n}\phi(y-n)\sum_{m=0}^{N-1}\tilde{d}_{j,m}\phi(z-m)\Big)\\
&= \sum_{n=0}^{N-1}\sum_{m=0}^{N-1}\mathbb{E}(\tilde{d}_{j,n}\tilde{d}_{j,m})\phi(y-n)\phi(z-m)\\
&= \sum_{n=0}^{N-1}\sum_{m=0}^{N-1}\mathbb{E}\Big[\int f(u)\tilde{\psi}_{j,n}(u)du\int f(v)\tilde{\psi}_{j,m}(v)dv\Big]\phi(y-n)\phi(z-m)\\
&= \sum_{n=0}^{N-1}\sum_{m=0}^{N-1}\phi(y-n)\phi(z-m)\Big[\int\int \mathbb{E}\big(f(u)f(v)\big)\tilde{\psi}_{j,n}(u)\tilde{\psi}_{j,m}(v)dudv\Big]\\
&= \sum_{n=0}^{N-1}\sum_{m=0}^{N-1}\phi(y-n)\phi(z-m)\Big[\int\int \mathbb{E}\big(f(u)f(v)\big)\\
&\qquad\qquad 2^{j/2}\psi\big(2^j(u-n)\big)2^{j/2}\psi\big(2^j(v-m)\big)dudv\Big]
\end{aligned} \qquad (A.2)$$

By using the form of auto-covariance function of fBm, we have:

$$\mathbb{E}\big(f(u)f(v)\big) = \frac{\sigma_H^2}{2}(|u|^{2H}+|v|^{2H}-|u-v|^{2H})$$

replace this in (A.2),

$$\mathbb{E}\Big(f_d^j(y)f_d^j(z)\Big) = 2^j \sum_{n=0}^{N-1}\sum_{m=0}^{N-1}\phi(y-n)\phi(z-m)\frac{\sigma_H^2}{2}\Big[\int |u|^{2H}\psi\big(2^j(u-n)\big)\Big(\int \psi\big(2^j(v-m)\big)dv\Big)du + \int |v|^{2H}\psi\big(2^j(v-n)\big)\Big(\int \psi\big(2^j(u-m)\big)du\Big)dv - \int\int |u-v|^{2H}\psi\big(2^j(u-n)\big)\psi\big(2^j(v-m)\big)dudv\Big]$$

the first two integrals inside the above brackets are zero, as $\int \psi(x)dx = 0$, for the third doubled integral we consider some variable changes such as: $p = 2^j(u-n) - 2^j(v-m) = 2^j(u-v+m-n)$ and $q = 2^j(v-m)$.

By these variable changes, we will have $u - v = 2^{-j}p + n - m$ and $2^j(u-n) = p+q$. So,

$$\mathbb{E}\Big(f_d^j(y)f_d^j(z)\Big) = 2^j\frac{\sigma_H^2}{2}\sum_{n=0}^{N-1}\sum_{m=0}^{N-1}\phi(y-n)\phi(z-m)\Big[\int\int |2^{-j}p+n-m|^{2H}\psi(p+q)\psi(q)(2^{-j}dp)(2^{-j}dq)\Big]$$

for choice of $n = m = k$, we will get:

$$\begin{aligned}\mathbb{E}\Big(f_d^j(y)f_d^j(z)\Big) &= \big(2^j\big)\big(2^{-2j}\big)\big(2^{-2Hj}\big)\frac{\sigma_H^2}{2}\sum_{k=0}^{N-1}\phi(y-k)\phi(z-k)\Big[\int\int |p|^{2H}\psi(p+q)\psi(q)dpdq\Big] \\ &= 2^{-(2H+1)j}\frac{\sigma_H^2}{2}V_\psi\sum_{k=0}^{N-1}\phi(y-k)\phi(z-k)\end{aligned} \tag{A.3}$$

where $V_\psi = \int\int |p|^{2H}\psi(p+q)\psi(q)dpdq$ does not depend on j, but just on H and ψ.

We finally have the $\mathbb{E}\Big(f_d^j(y)f_d^j(z)\Big)$ to plug into (A.1) equation:

$$\mathbb{E}(D_{j,k}^2) = \int\int \mathbb{E}\Big(f_d^j(y) f_d^j(z)\Big) 2^j \psi\big(2^j(y-k)\big) 2^j \psi\big(2^j(z-k)\big) dy dz \tag{A.4}$$
$$= 2^{-(2H+1)j} \frac{\sigma_H^2}{2} V_\psi$$
$$\sum_{k=0}^{N-1} \bigg(\int \phi(y-k) \psi\big(2^j(y-k)\big) 2^j dy \bigg) \bigg(\int \phi(z-k) \psi\big(2^j(z-k)\big) 2^j dz \bigg)$$

The Last summation is dependent to the wavelet function ψ and j, so, we call it $Q_{\psi,j}$. To summarize:

$$\mathbb{E}(D_{j,k}^2) = 2^{-(2H+1)j} \frac{\sigma_H^2}{2} V_\psi Q_{\psi,j} \tag{A.5}$$

BIBLIOGRAPHY

Abry, P.; Flandrin, P.; Taqqu, M., and Veitch, D. Wavelets for the analysis, estimation, and synthesis of scaling data. *In K. Park and W. Willinger, editors, Self-Similar Network Traffic and Performance Evaluation,*, pages 39–88, Wiley, 2000.

Abry, P.; Flandrin, P.; Taqqu, M. S., and Veitch, D. Self-similarity and long-range dependence through the wavelet lens. *In Theory and applications of long-range dependence*, pages 527–556, 2003.

Abry, P; Goncalves, P, and Vehel, J L. *Scaling, Fractals and Wavelets*. Wiley-ISTE, 2013. ISBN 978-1-118-62290-2.

Abry, Patrice; Goncalves, Paulo, and Flandrin, Patrick. Wavelet-based spectral analysis of 1/f processes. *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 3:237–240, 1993.

Alexander, D. H.; Novembre, J., and Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome research*, 19:1655–1664, 2009.

Aulicino, E.D. *Genetic Genalogy: The Basics and Beyond*. AuthorHouse, Bloomington, 2013.

Bakhtiar, Ray and Nelson, Randall W. Mass spectrometry of the proteome. *Molecular Pharmacology*, 60:405–415, 2001.

Bakhtiar, Ray and Tse, F. L. S. Biological mass spectrometry: A primer. *Mutagenesis*, 15: 415–430, 2000.

Barnard, GA. Discussion of hurst. *Proceedings of the Institution of Civil Engineers*, 5: 551–553, 1956.

Benedetto, J J and Colella, D. De-noising using wavelets and cross validation. *Wavelet Applications in Signal and Image Processing III*, 2569 of NATO ASI Series C:512–521, 1995.

Beran, J. *Statistics for Long-memory Processes*. Chapman & Hall, 1994.

Burt, P. J. and Adelson, E. H. A multiresolution spline with applications to image mosaic. *ACM Trans. Graphics*, 2:674–693, 1983a.

Burt, P. J. and Adelson, E. H. The laplacian pyramid as a compact image code. *IEEE Trans. Comm*, COM-31:532–540, 1983b.

Cann, R.L.; Stoneking, M., and Wilson, A.C. Mitochondrial dna and human evolution. *Nature*, 325:31–36, 1987.

Chang, S. G.; Yu, B., and Vetterli, M. Adaptive wavelet thresholding for image denoising and compression. *IEEE Trans. Image Processing*, 9:1532–1546, 2000.

Daubechies, I. Orthonormal bases of compactly supported wavelets. *Comm. Pure and Appl. Math.*, 41, 1988.

De Canditiis, Daniela and Vidakovic, Brani. Wavelet bayesian block shrinkage via mixtures of normal-inverse gamma priors. *Journal of Computational and Graphical Statistics*, 13: 383–398, 2004.

Donoho, D. L. De-noising by soft-thresholding. *IEEE Transactions on Information Theory*, 42:613–627, 1995.

Donoho, D. L. and Johnstone, I. M. Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81:425–455, 1994a.

Donoho, D. L. and Johnstone, I. M. Adapting to unknown smoothness via wavelet shrinkage. *Journal of the American Statistical Association*, 90:1200–1224, 1994b.

Doukhan, P; Oppenheim, G, and Taqqu, M S. Theory and applications of long-range dependence. *Springer Science & Business Media*, 2003.

Eltis, D. A brief overview of the trans-atlantic slave trade. voyages. *Voyages:The Trans-Atlantic Slave Trade Database*, 2007. URL `http://www.slavevoyages.org/assessment/essay`.

Emery, L.S.; Magnaye, K.M.; Bigham, A.W.; Akey, J.M., and Bamshad, M.J. Estimates of continental ancestry vary widely among individuals with the same mtdna haplogroup. *Am. J. Hum. Genet*, 96:183–193, 2015.

Feller, W. The asymptotic distribution of the range of sums of independent random variables. *The Annals of Mathematical Statistics*, 22:427–432, 1951.

Fitzpatrick, C. and Yeiser, A. *DNA Genealogy*. Rice Book Press, Houston, 2005.

Flandrin, P. On the spectrum of fractional brownian motions. *IEEE Transaction on Information Theory*, 35:197–199, 1989.

Flandrin, P. Wavelet analysis and synthesis of fractional brownian motion. *IEEE Transaction on Information Theory*, 38:910–917, 1992a.

Flandrin, Patrick and Goncalves, Paulo. From wavelets to time-scale energy distributions. *Recent Advances in the Theory of wavelets, Academic Press*, pages 309–334, 1992.

Gates Jr., H.L. *Exploring Our Roots*. PBS, Boston, 2010.

Grossmann, A. and Morlet, J. Decomposition of functions into wavelets of constant shape and related transforms. *Mathematics and physics, lectures on recent results*, 1985.

HGDP, . Human genome diversity project. URL `https://www.hagsc.org/hgdp/`.

Lamperti, J. Semi-stable stochastic processes. *Transactions of the American Mathematical Society*, 104:62–78, 1962.

Mallat, S. G. A theor a theory for multir y for multiresolution signal decomposition: The w esolution signal decomposition: The wavelet representation. 1987.

Mallat, S. G. Multiresolution approximations and wavelet orthonormal bases of $l^2(r)$. *Transactions of the American Mathematical Society*, 315:69–87, 1989a.

Mallat, S. G. A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11, 1989b.

Mallat, S. G. Zero-crossings of a wavelet transform. *Information Theory, IEEE Transactions*, 37:1019–1033, 1991.

Mallat, S. G. *A wavelet tour of signal processing: the sparse way*. Academic Press, 1998.

Mallat, S. G. and Zhong, S. Characterization of signals from multiscale edges. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14:710–732, 1992.

Mandelbrot, B.B. and Ness., J. W. J. W. Van. Fractional brownian motions, fractional noises and applications. *SIAM rev.*, 10:422–437, 1968.

Marshall, Bryan; McEver, Rodger P., and Zhu, Cheng. Kinect rates and their force dependence of the p-selection/psgl-1 interaction measured by atomic force microscopy. *Bioengineering Conference ASME*, 50, 2001.

McVean, G. Corresponding Author The 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature*, 467:1061–1073, 2010.

McVean, G. Corresponding Author The 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491:56–65, 2012.

Micheletti, Steven J.; Bryc, Kasia; Esselmann, Samantha G. Ancona; Freyman, William A.; Moreno, Meghan E.; Poznik, G. David; Shastri, Anjali J.; 23andMe Research Team, ; Beleza, Sandra, and Mountain, Joanna L. Genetic consequences of the transatlantic slave trade in the americas. *American Journal of Humman Genetics*, 107:265–277, 2020.

Moloney, K. P.; Jacko, J. A.; Vidakovic, B.; Sainfort, F.; Leonard, V. K., and Shi, B. Leveraging data complexity: Pupillary behavior of older adults with visual impairment during hci. *Journal ACM Transactions on Computer-Human Interaction (TOCHI)*, 13: 376–402, 2006.

Morlet, J.; Arens, G.; Fourgeau, I., and Giard, D. Wave propagation and sampling theory. *Geophysics*, 47:203–236, 1982.

Myers, R. H.; Montgomery, D. C.; Vining, G. G., and Robinson, T. J. *Generalized Linear Models with Applications: in Engineering and the Sciences*. New York: John Wiley Sons, 2002.

Omberg, L.; Salit, J., and Hackett, N. et al. Inferring genome-wide patterns of admixture in qataris using fifty-five ancestral populations. *BMC genetics*, 13:49, 2010.

Pakendorf, B. and Stoneking, M. Mitochondrial dna and human evolution. *Annu. Rev. Genomics Hum. Genet*, 6:165–183, 2005.

Potter-Phillips, D. History of genealogy. family chronicles. *Voyages:The Trans-Atlantic Slave Trade Database*, 1999. URL `http://www.familychronicle.com/HistoryOfGenealogy.html`.

Purcell, S.; Neale, B.; K, Todd-Brown; Thomas, L.; Ferreira, M. A. R.; Bender, David; Maller, Julian; Sklar, Pamela; de Bakker, Paul I. W.; Daly, Mark J., and Sham, Pak C. Plink: a tool set for whole-genome association and population-based linkage analyses. *American journal of human genetics*, 81:559–575, 2007. URL `https://www.cog-genomics.org/plink/`.

Rayana, M. I. BEN. Denoising using decompositions in the auto-correlation shell. *University of California Davis, Department of Mathematics,AC Shell Matlab Package*, 1998.

Remenyi, Norbert and Vidakovic, Brani. $lambda$-neighborhood wavelet shrinkage. *Computational Statistics Data Analysis*, 57:404–416, 2013.

Rishishwar, Lavanya; Conley, Andrew B.; Vidakovic, Brani, and Jordan, I. King. A combined evidence bayesian method for human ancestry inference applied to afro-colombians. *Gene*, 574:345–351, 2015.

Saito, N. Local feature extraction and its applications using a library of bases. *Ph.D. thesis, Dept. of Mathematics, Yale University, New Haven, CT 06520 USA*, 41, 1994.

Saito, N. and Beylkin, G. Multiresolution representations using the autocorrelation functions of compactly supported wavelets. *IEEE Trans. Signal Process.*, 41:3584–3590, 1993.

Salas, A.; Carracedo, A.; Richards, M., and Macaulay, V. Charting the ancestry of african americans. *Am. J. Hum. Genet*, 77:676–680, 2005.

slave voyages website, . URL `http://www.slavevoyages.org/assessment/essay`.

Stefflova, K.; Dulik, M.C.; Barnholtz-Sloan, J.S.; Pai, A.A.; Walker, A.H., and Rebbeck, T.R. Dissecting the within-africa ancestry of populations of african descent in the americas. *PLoS One*, 6:e14495, 2011.

Stumpf, M.P. and Goldstein, D.B. Genealogical and evolutionary inference with the human y chromosome. *Science*, 291:1738–1742, 2001.

Tang, H.; Mukomel, Y., and Fink, E. Diagnosis of ovarian cancer based on mass spectra of blood samples. *IEEE International Conference on Systems, Man and Cybernetics*, 4: 3444–3450, 2004.

Vidakovi, Brani. *Statistical Modeling by Wavelets*. Wiley Series in Probability and Statistic, 1999.

Yates, John R. Mass spectrometry: From genomics to proteomics. *Trends in Genetics*, 16: 5–8, 2000.

VITA

Parisa Yousefi Zowj is a Ph.D. candidate in the School of Industrial Systems Engineering (ISyE) at Georgia Institute of Technology in the Bioinfaormatics program. She joined the Ph.D. program in 2017 and was taken under the advisement of Dr. Brani Vidakovic. Her research interests focus on developing multiresolution analytical tools including Wavelets and Auto-Correlation (AC) Shells for signal and image processing with applications in medical decision making. She has also been working on Bayesian methods for modeling genetic consequences of the transatlantic slave trade. Before beginning the Ph.D. program, she earned a Bachelor of Science degree and a Master of Art in Statistics from Iran in 2005 and 2008, respectively, as well as a Master's degree in Applied Statistics from ISyE, Georgia Tech in 2017.