

**FEATURE LEARNING AND PERSONALIZED SCREENING TECHNIQUES IN
HEALTHCARE**

A Dissertation
Presented to
The Academic Faculty

By

Chen Feng

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

Georgia Institute of Technology

August 2019

Copyright © Chen Feng 2019

FEATURE LEARNING AND PERSONALIZED SCREENING TECHNIQUES IN HEALTHCARE

Approved by:

Dr. Yajun Mei, advisor
H. Milton Stewart School of Industrial and Systems Engineering
Georgia Institute of Technology

Dr. Brani Vidakovic, advisor
H. Milton Stewart School of Industrial and Systems Engineering
Georgia Institute of Technology

Dr. Xiaoming Huo
H. Milton Stewart School of Industrial and Systems Engineering
Georgia Institute of Technology

Dr. Paul Griffin
School of Industrial Engineering,
Purdue University
Adjunct Professor, H. Milton Stewart School of Industrial and Systems Engineering
Georgia Institute of Technology

Dr. Ofer Sadan
School of Medicine
Emory University

Date Approved: May 06, 2019

Truth is ever to be found in simplicity, and not in the multiplicity and confusion of things

Isaac Newton

To my beloved parents and grandparents

ACKNOWLEDGEMENTS

My four-year Ph.D. study at the Georgia Institute of Technology is the most memorable time in my life. I would like to express my sincere gratitude to my advisor Prof. Yajun Mei for his continuous support of my Ph.D. study and related research. I got to know Prof. Mei in 2014 spring when I was a junior student at Xi'an Jiaotong University and was exchanged to Georgia Tech for one semester. I approached Prof. Mei and expressed my willingness to do research with him at that time. I still remember his kind response to my email, his patience in guiding me to the wonderful research world, and his encouragement when I felt frustrated by the difficulties in doing the research. It was his passion in research that motivated me to pursue a Ph.D. in Statistics, and ever since then, I make up my mind to be a person like him, kind-hearted, diligent, and knowledgeable. During my Ph.D. study, he has been very supportive, and he encourages me to do the research that I am interested in. Thanks to his support, I am able to do many interesting practical healthcare related projects with collaborators from different institutes besides my own theoretical research. He not only helps me with detailed research problems but also inspires me to have a bigger view. I would say, the most valuable thing that I learn from him is not any specific knowledge, but the ability to formulate an interesting research question from the real-life problem, which will be beneficial to my whole career life.

My sincere gratitude also goes to my co-advisor, Prof. Brani Vidakovic, for his patience, motivation, and immense knowledge. I started to work with Prof. Vidakovic in the second year of my Ph.D. study. I remember taking his Biostatistics course in Spring 2017, and after finishing the course, I asked him to be my co-advisor since I have great interests in his bio-related research. I was stuck by my sepsis project (which turns out to be Chapter 3 in my thesis) at that time and was under the big pressure of making less progress in my research. It was Prof. Vidakovic who save me from that frustrating situation by introducing me to a new wonderful research field of wavelet analysis and signal processing. Besides,

Prof. Vidakovic generously provides me with many of his resources, including but are not limited to offering me the opportunity to be the student statistical consultant of *Georgia Clinical & Translational Science Alliance*. I respect him as a knowledgeable scholar, as an elegant gentleman, and as one of my favorite seniors.

I would also like to thank Prof. Paul Griffin, Prof. Xiaoming Huo, and Prof. Ofer Sadan for serving on my thesis committee. Special thanks go to Prof. Paul Griffin for supporting me as a graduate research assistant when I was in my first year Ph.D. study and providing helpful advice to my research. I am grateful to Prof. Xiaoming Huo for his generous help, valuable comments, and insightful suggestions on my thesis. Many thanks go to Prof. Ofer Sadan for providing me with the opportunity to participate in interesting and significant clinical projects and introducing me with much medical knowledge.

I want to show my gratitude to Dr. Jelena Velickovic for sharing the real clinical data and providing professional interpretations to my analysis results. I would also like to extend a special thank to Prof. Stephen Sprigle for guiding me to the applied disability research. My sincere thanks also go to Dr. Maria Terres, who provided me an opportunity to join The Climate Corporation as an intern, and who gave me guidance and fruitful discussions. All of their expertise has been greatly beneficial to my research and career development.

I am very thankful to my dear friends for the sleepless nights we were working together before deadlines, and for all the fun we have had in the last four years. They include but are not limited to Chuanping Yu, Junying He, and Xi He. In particular, I am grateful to my boyfriend Dr. Yang Cao for all the stimulating discussions, for his selfless love, and for his tolerance of my negative emotions whenever I feel stressed out.

Finally, I would like to express my deepest thanks to my beloved mother Yanhong Zhao, father Guodong Feng, grandmother Guilan Lv and grandfather Wannian Feng, for their continuous love and support, especially during my pursuit of this Ph.D. I love them very much, and this dissertation is dedicated to them.

TABLE OF CONTENTS

Acknowledgments	v
List of Tables	xi
List of Figures	xiii
Chapter 1: Wavelet-Based Robust Estimation of Hurst Exponent with Application in Visual Impairment Classification	1
1.1 Introduction	1
1.2 General Trimean Estimators	4
1.3 Robust Estimations of Hurst Exponent	6
1.3.1 General Trimean of the Mid-energy (GTME) Method	8
1.3.2 General Trimean of the Logarithm of Mid-energy (GTLME) Method	10
1.4 Simulation	13
1.5 Application	16
1.6 Proof of Theorems 1.3.1 and 1.3.2	19
1.7 Conclusions	25
Chapter 2: Mammogram Diagnostics Using Robust Wavelet-based Estimator of Hurst Exponent	27
2.1 Introduction	27

2.2	Background	30
2.2.1	Non-decimated wavelet transforms	30
2.2.2	The fBm: Wavelet coefficients and Spectra	32
2.3	General Trimean Estimators	33
2.3.1	Tukey’s Trimean Estimator	35
2.3.2	Gastwirth Estimator	36
2.4	Methods	37
2.4.1	General Trimean of the Mid-energy (GTME) Method	39
2.4.2	General Trimean of the Logarithm of Mid-energy (GTLME) Method	41
2.4.3	Special Cases: Tukey’s Trimean and Gastwirth Estimators	42
2.5	Simulation	45
2.6	Application	46
2.7	Proof of Theorems 2.4.1 and 2.4.2	50
2.8	Proof of Lemmas 2.4.1 and 2.4.2	55
2.9	Conclusions	59
Chapter 3: A Personalized Threshold Method via Boosting for Sepsis Screening		61
3.1	Introduction	61
3.2	Background	63
3.2.1	Sepsis and qSOFA	63
3.2.2	Existing Methods to Obtain the Constant Threshold	64
3.2.3	Boosting Method	66
3.3	Our Proposed Personalized Threshold Method	67

3.3.1	Optimization Problem	68
3.3.2	Gradient Descent Algorithm	69
3.4	The Data Set	72
3.4.1	Study Population	72
3.4.2	Observed Data	73
3.4.3	General Characteristics	74
3.4.4	Exploratory Analysis on qSOFA Variables	74
3.5	Application to Sepsis Screening	76
3.5.1	Tuning parameters	77
3.5.2	Comparison to qSOFA criteria with constant thresholds	79
3.5.3	Comparison with other machine learning techniques	81
3.5.4	Interpretation and Implementation of Personalized qSOFA	82
3.6	Proof of Proposition 3.3.1	83
3.7	Conclusions	84
	Chapter 4: Regularized zero-and-one inflated beta regression model with ap- plication in predicting postoperative complications	87
4.1	Introduction	87
4.2	Background	90
4.2.1	Varying dispersion beta regression model	90
4.2.2	Zero-or-one inflated beta regression model	92
4.2.3	Regularized generalized linear models	93
4.3	Our proposed regularized beta regression	95
4.3.1	Zero-and-one inflated beta regression model	96

4.3.2	Likelihood inference	98
4.3.3	Variable selection algorithm	100
4.3.4	Classification Rule	104
4.4	Simulation	105
4.5	Application	107
4.6	Derivation of Score functions and Information matrix	113
4.7	Conclusions	120
Chapter 5: Conclusions and Future Research		122
References		137

LIST OF TABLES

1.1	Simulation Results for $N = 2^{10}$ fBm using Haar wavelet	14
1.2	Simulation Results for $N = 2^{11}$ fBm using Haar wavelet	15
1.3	Simulation Results for $N = 2^{12}$ fBm using Haar wavelet	16
1.4	Group characterization summary	18
1.5	Descriptive Statistics Group Summary (original noisy data)	19
1.6	Descriptive Statistics Group Summary (cleaned data)	20
1.7	Classification error	20
2.1	Simulation Results for $2^{10} \times 2^{10}$ fBm using Haar wavelet (300 Replications)	47
2.2	Descriptive Statistics Group Summary	49
2.3	Results of classification by logistic regression	50
3.1	Variables and corresponding definitions.	74
3.2	Descriptive statistics of interested variables.	75
3.3	Overall accuracy	79
3.4	Sensitivity	80
3.5	Specificity	80
3.6	Comparison with other machine learning techniques	82
4.1	Simulation Prediction Performance, average over 400 repetitions	106

4.2	Simulation Estimation Performance, average over 400 repetitions	107
4.3	Variables and corresponding definitions.	109
4.4	Prediction performance, average across 5-fold CV on training data	111
4.5	Prediction performance, on validation data	111

LIST OF FIGURES

1.1	Examples of PRB data from three groups: Left (healthy, control), middle (AMD group I, mild case), right (AMD group II, severe case)	2
1.2	Plot of $h(\alpha, p)$ against p on the left; Plot of α against p on the right	13
2.1	(a) Four types of wavelet coefficients with their locations in the tessellation of a 2-D scale mixing NDWT of depth of 3 ($J = 3$), with each block the size of $N \times N$. Coefficients c represent the coarsest approximation, h and v are the mix of coarse and detail information, and d carry detail information only. (b) Detail coefficients d and its diagonal blocks corresponding to 3 ($J = 3$) levels. (c) Symmetric random sampling from level-1 ($j = 1$) diagonal block divided into 6×6 ($M = 6$) grids.	38
2.2	Plot of $f(\alpha, p)$ against p on the left; Plot of α against p on the right	43
2.3	Histograms and Theoretical Distributions of \hat{H}	45
2.4	An example of mammograms with breast cancer	48
2.5	Using GME method to estimate Hurst exponent, boxplots in cancer and non-cancer groups on the left; normal density curves fitted in cancer and non-cancer groups on the right	48
3.1	Flowchart of cohort selection	73
3.2	Histograms of qSOFA variables.	76
3.3	Scatter plot of systolic blood pressure against respiratory rate	76
3.4	The weighted exponential loss versus the number of iterations with different learning rate α	78

3.5	The accuracy (left), sensitivity and specificity (right) with different w_+ 's and fixed $w_- = 1$	78
3.6	The predicted personalized threshold for respiratory rate against age.	82
3.7	Left: Screening for non-sepsis patient. Right: Screening for sepsis patient.	83
4.1	Left: Histogram of rescaled CCI in training set; Right: Histogram of rescaled CCI in validation set	109
4.2	MSEs in different ranges. Left: 5-fold CV results on training data. Right: Results on validation set	112
4.3	Variable Selection Results	113

SUMMARY

Data science is playing an increasingly important role in improving public health. Data used for public health studies are in various types, including electronic health records, image data, administrative data, claims data, and patient disease registries. The data variety provides opportunities and challenges for statisticians to impact public health in many ways. The value of statisticians lies in finding patterns in collected data, summarizing and presenting these in an effort to best describe the target population, and developing the necessary mathematical tools to ascertain associations of risk factors with the disease.

This dissertation aims to develop data-driven, efficient statistical and machine learning techniques in some modern real-world applications. We consider four different contexts: (i) Visual impairment classification based on noisy high-frequency pupillary response behavior data collected from human-computer interaction, (ii) breast cancer diagnosis using image data from plain Xray, (iii) personalized screening for sepsis disease based on regularly measured longitudinal biomarkers, (iv) prediction on the overall burden of postoperative complications using laboratory measurements.

In Chapter 1, we study the robust estimation of Hurst exponent from one dimensional high-frequency, time series data. High-frequency data from various sources often possess hidden patterns that reveal the effects of underlying functional differences, but such patterns cannot be explained by basic descriptive statistics, traditional statistical models, or global trends. For those complex high-frequency data, Hurst exponent becomes a powerful tool to detect muted or irregular change patterns. Hurst exponent quantifies the long memory, regularity, self-similarity, and scaling in a time series, and the robust estimation of Hurst exponent has recently become a topic of interest. In this chapter, we propose robust estimators of Hurst exponent based on non-decimated wavelet transforms and apply our methods to Pupillary response behavior (PRB) data to extract the Hurst exponent and then use it as a predictor to classify individuals with different degrees of visual impairment. At a high

level, the basic idea of all wavelet-based methods to estimate Hurst exponent is to explore the fact that Hurst exponent is linearly correlated to wavelet coefficients on the log-scale. In this study, we propose a general trimean estimator that balances the tradeoff between the median and extreme values and applied it on wavelet coefficients before correlating with Hurst exponent. By doing this, we are able to lessen the effects of outliers, thus achieving the robust estimation of Hurst exponent.

In Chapter 2, we extend the robust estimation of Hurst exponent to two-dimensional images and then apply the proposed method to mammograms to diagnose breast cancer. In the literature, researchers have developed many statistical and machine learning methods to do image classification, but most of them are black-box methods, and hard to interpret. In this chapter, we propose to use fractional Brownian motion (fBm) to model mammogram image, develop a robust estimator of Hurst exponent from two-dimensional fBm models based on non-decimated wavelet transforms, and then predict breast cancer using the extracted Hurst exponent. This allows us to use the underlying degree of self-similarity as a discriminatory descriptor to classify mammograms to benign and malignant. In addition, as compared to the one-dimensional case, it is more complicated for the two-dimensional images because the within-level correlation of non-decimated wavelet coefficients is defined in two-dimensional space and violates the independence assumption. Our main idea is to consider a symmetric random sampling technique to solve for such a correlation issue. Unlike the hard-to-interpret machine learning methods, our method helps to summarize the common features from the cancerous images and mimics the way how physicians make decisions in practice. It has been shown to be efficient and accurate in the early detection of breast cancer through simulation and case study.

Chapter 3 studies the personalized screening for Sepsis disease. Sepsis is a life-threatening complication of the infection. In 2016, a scoring criterion called quick Sequential (sepsis-related) Organ Failure Assessment (qSOFA) was proposed by a group of experts as a screening criterion for sepsis. To be more concrete, if at least two of the following three

conditions are satisfied, then an alarm will be set for a patient and physicians will conduct laboratory tests to further assess sepsis: 1) systolic blood pressure is ≤ 100 mm Hg; 2) respiratory rate is ≥ 22 breaths/min; 3) alteration in mental status (GCS score is less than 15). However, qSOFA does not perform well in practice, with very low sensitivity. Part of the reasons is that qSOFA uses constant thresholds for the biomarkers in regardless of patients' baseline information. Hence, we aim to improve qSOFA by developing a knowledge-based machine-learning method to self-learn the personalized thresholds that depend on patients' baseline information. The main idea to model the personalized threshold as functions of those demographic variables, and then use a boosting-based weighted exponential loss function to learn the personalized thresholds for efficient screening of sepsis. Our method yields efficient personalized monitoring, appropriate subject-specific intervention in early stages of sepsis, and thus a significant reduction of the mortality rate.

In Chapter 4 of the dissertation, our motivating example is to predict the overall burden of postoperative complications based on a real data set consisting of 206 adult patients who stayed in the Clinic for Digestive Surgery, Clinical Center of Serbia in Belgrade between November 2016 and October 2017. Recently, a novel critical scale called Comprehensive Complication Index (CCI) has been developed to capture the overall burden of complications in the postoperative period. CCI is widely used in reality, but it has several disadvantages: 1) It is calculated from a complicated procedure that requires the physicians and nurses to make records of every detail during patients' hospitalization, and is not practical for everyday use; 2) it can be calculated only retrospectively, when the hospitalization is finished, it can only reflect the results of perioperative treatment but cannot be used as a measure of patients' current status. In this chapter, we develop a zero-and-one inflated beta regression model to predict the CCI values based on patients' clinical covariates and also propose to estimate the unknown sparse coefficient vectors by maximizing the penalized log-likelihood function. Our proposed method not only can achieve a good prediction on CCI but also can select important clinical covariates leading to postoperative complications.

CHAPTER 1

**WAVELET-BASED ROBUST ESTIMATION OF HURST EXPONENT WITH
APPLICATION IN VISUAL IMPAIRMENT CLASSIFICATION**

1.1 Introduction

Visual impairment is defined as a functional limitation of the eyes or visual system. It can cause disabilities by significantly interfering with one's ability to function independently, to perform activities of daily living, and to travel safely through the environment, see [1, 2]. Many causes of severe visual impairment are hard to cure, however, there are conditions for which medical or surgical treatment will lessen the severity or progression of the vision loss, for example, recent advances in the treatment of age-related macular degeneration (AMD), see [3, 4, 5]. Precise classification of different degrees of visual impairment for AMD patients becomes increasingly important for the sake of early intervention. It has been suggested by [6] that the high-frequency pupillary response behavior (PRB) data can be useful in visual impairment classification. PRB refers to changes in pupil diameter in response to simple or complex stimuli, and in this study it was measured from older adults, including two groups diagnosed with AMD maintaining different ranges of visual acuities, and one visually healthy control group. Examples of PRB data from those three groups are shown in Figure 1.1. [6] indicated that there may be underlying unique patterns hidden within complex PRB data, and these patterns reveal the intrinsic individual differences in cognitive, sensory and motor functions. However, the proper description and interpretation of PRB is not straightforward, since it is affected by a variety of factors, including the ambient light, fatigue, and medication use. In fact, high-frequency, time series data from various sources often possess hidden patterns that reveal the effects of underlying functional differences, but such patterns cannot be explained by basic descriptive statistics, traditional

statistical models, or global trends. Thus powerful analytical tools are needed to detect these muted or irregular change patterns for those complex high-frequency data, like PRB.

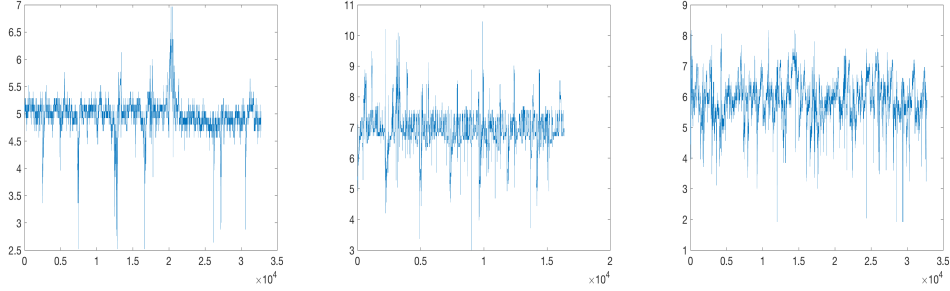


Figure 1.1: Examples of PRB data from three groups: Left (healthy, control), middle (AMD group I, mild case), right (AMD group II, severe case)

One powerful tool is the Hurst exponent, denoted by H in the sequel. It quantifies the long memory, regularity, self-similarity, and scaling in a time series, and has been used as an important feature in many applications, see [7, 8, 9, 10, 11, 12]. To be more concrete, a stochastic process, $\{X(t), t \in \mathbb{R}\}$ is self-similar with Hurst exponent H if $X(t) \stackrel{d}{=} \lambda^{-H} X(\lambda t)$, for any $\lambda \in \mathbb{R}^+$. Here the notation $\stackrel{d}{=}$ means the equality in all finite-dimensional distributions. Hurst exponent describes the rate at which autocorrelations decrease as the lag between two realizations in a time series increases. A value H in the range 0-0.5 indicates a zig-zagging intermittent time series with long-term switching between high and low values in adjacent pairs. A value H in the range 0.5 to 1 indicates a time series with long-term positive autocorrelations, which preserves trends on a longer time horizon and gives a time series more regular appearance.

One widely used example of self-similar Gaussian process is the fractional Brownian motion (fBm), which was first described by [13] and formalized by [14], also see [15, 16]. The fBm is a continuous-time Gaussian process $X(t)$, which starts at zero, has expectation zero for all t , and has the following covariance function:

$$E[X(t)X(s)] = \frac{1}{2}(|t|^{2H} + |s|^{2H} - |t - s|^{2H}).$$

One popular method to estimate Hurst exponent from fBm is the multiresolution analysis through wavelet transformations, see [17, 18, 19]. The idea is to explore the fact that H is linearly correlated to wavelet coefficients d_j 's at level j on the log-scale, and the following two estimation methods of H have been proposed: 1) Veitch & Abry (VA) method in [17] by weighted least square regression on the level-wise $\log_2(\bar{d}_j^2)$; 2) Soltani, Simard, & Boichu (SSB) method in [20] by first defining a mid-energy as $D_{j,k} = (d_{j,k}^2 + d_{j,k+N_j/2}^2)/2$, then taking the mean of the logarithm of mid-energies, and last applying weighted least square regression. Later [21] demonstrated that the SSB method yields more accurate estimators than the VA method since it takes the logarithm first and then averages. Unfortunately, both methods are sensitive to outlier coefficients and outlier multiresolution levels, inter and within level dependences, and distributional contaminations, and thus, it is important to robustify them.

The robust estimation of Hurst exponent H has recently become a topic of interest, see [22, 23, 21, 24]. Three robust estimation methods have been developed. The first one is the Theil-type regression (TT) method in [25], which modified the VA method by replacing the weighted least square regression with the Theil-type weighted regression in [26] to make it less sensitive to outlier levels. The second and third methods are MEDL and MEDLA proposed in [27], where the median, not mean, was used for level-wise wavelet coefficients. The difference between these two methods is that MEDL takes logarithm on squared wavelet coefficients, while MEDLA is similar to concept in SSB method that paired and averaged wavelet coefficients prior to taking logarithm. Although median is outlier-resistant, it can behave unexpectedly as a result of its non-smooth character. The fact that the median is not “universally the best outlier-resistant estimator” provides a practical motivation for examining alternatives that are intermediate in behavior between the very smooth but outlier-sensitive mean and the very outlier-insensitive but non-smooth median.

In this chapter, we propose to robustly estimate the Hurst exponent from the fBm model, where the mean or median of the wavelet coefficients are replaced by a general trimean

estimator that is inspired by [28] and [29]. Here the general trimean estimator is defined as a weighted average of the median and two quantiles symmetric about the median, which balances the tradeoff between median value and extreme values. It turns out that this will yield a robust estimator of the Hurst exponent from PRB data, which in turn allows us to efficiently classify PRB into groups with different degrees of visual impairment.

The remaining structure is as follows. Section 1.2 introduces the general trimean estimators; Section 1.3 describes estimation of Hurst exponent using the general trimean estimators and derives the asymptotic distributions of the proposed estimators. Section 1.4 provides the simulation results and compares the performance of the proposed methods to other standardly used, wavelet-based methods. The proposed methods are illustrated using the real PRB data for visual impairment classification in Section 1.5. The proofs of Theorems are provided in Section 1.6. The chapter is concluded with a summary and discussion in Section 1.7.

1.2 General Trimean Estimators

In this section, we propose a general trimean estimator under the point estimation context, which will be used later for robust estimation of Hurst exponent.

Let X_1, \dots, X_n be i.i.d. continuous random variables with pdf $f(x)$ and cdf $F(x)$ with mean μ . For $0 < p < 1$, let $Y_p = X_{[np]:n}$ denote a sample p th quantile, where $[np]$ denotes the greatest integer that is less than or equal to np . In the context of outliers, a remarkably efficient robust estimator of population mean μ is Tukey's trimean estimator in [28], defined as

$$\hat{\mu}_T = \frac{1}{4} Y_{1/4} + \frac{1}{2} Y_{1/2} + \frac{1}{4} Y_{3/4}. \quad (1.1)$$

Another robust estimator is the Gastwirth's estimator in [29],

$$\hat{\mu}_G = 0.3 Y_{1/3} + 0.4 Y_{1/2} + 0.3 Y_{2/3}. \quad (1.2)$$

Here we propose a general trimean estimator, which is defined as a weighted average of the distribution's median and its two quantiles Y_p and Y_{1-p} , for $p \in (0, 1/2)$:

$$\hat{\mu} = \frac{\alpha}{2} Y_p + (1 - \alpha) Y_{1/2} + \frac{\alpha}{2} Y_{1-p}. \quad (1.3)$$

The weights for the two quantiles are the same for Y_p and Y_{1-p} , and $\alpha \in [0, 1]$. This is equivalent to the weighted sum of the median and the average of Y_p and Y_{1-p} with weights $1 - \alpha$ and α :

$$\hat{\mu} = (1 - \alpha) Y_{1/2} + \alpha \left(\frac{Y_p + Y_{1-p}}{2} \right).$$

This general trimean estimator is more robust than mean but smoother than the median. It turns out that Tuckey's trimean estimator and Gastwirth's estimator are two special cases. To be specific, $\alpha = 1/2$, $p = 1/4$ in Tuckey's trimean estimator, and $\alpha = 0.6$, $p = 1/3$ in Gastwirth's estimator.

To derive its asymptotic distribution, we need to first define some notations for the population distribution. Let $0 < p < 1$ and ξ_p denotes the p th quantile of F , so that $\xi_p = \inf\{x | F(x) \geq p\}$. If F is monotone, the p th quantile is simply defined as $F(\xi_p) = p$. Moreover, define

$$A = \begin{bmatrix} \frac{\alpha}{2} & 1 - \alpha & \frac{\alpha}{2} \end{bmatrix}, \quad (1.4)$$

$$\boldsymbol{\xi} = [\xi_p \quad \xi_{1/2} \quad \xi_{1-p}]^T, \quad (1.5)$$

and the asymptotic covariance matrix of $\mathbf{y} = [Y_p \quad Y_{1/2} \quad Y_{1-p}]^T$ is

$$\Sigma = (\sigma_{ij})_{r \times r}, \quad \text{with} \quad \sigma_{ij} = \frac{p_i (1 - p_j)}{f(x_{p_i}) f(x_{p_j})}, \quad i \leq j, \quad (1.6)$$

see [30].

Now we are ready to present the asymptotic properties of our proposed general trimean estimator $\hat{\mu}$ in (1.3).

Lemma 1.2.1. *As $n \rightarrow \infty$, for $\hat{\mu}$ in (1.3),*

$$\sqrt{n}(\hat{\mu} - A \cdot \xi) \stackrel{approx}{\sim} \mathcal{N}(0, A\Sigma A^{-1}) \quad (1.7)$$

The proof of Lemma 1.2.1 follows directly from the asymptotic joint distribution of the order statistics, and thus are omitted.

1.3 Robust Estimations of Hurst Exponent

In this section, we propose two different robust methods for estimating Hurst exponent H under the fBm model through wavelet transformations.

For that purpose, let us first provide a brief background on non-decimated wavelet transforms (NDWT), also see [31] for more information. In a J -depth decomposition of a fBm of size N , a NDWT generates J detail levels and one smooth level, therefore containing $N \times (J + 1)$ wavelet coefficients, N in each level. In the Hurst exponent estimation literature, most research was based on the standard orthogonal discrete wavelet transforms (DWT), but NDWT turns out to have several advantages when employed for Hurst exponent estimation: 1) Input signals and images of arbitrary size can be processed due to the absence of decimation; 2) as a redundant transform, the NDWT increases the accuracy of the scaling estimation; 3) least square regression can be fitted to estimate H instead of weighted least square regression since the variances of the level-wise derived distributions based on logged NDWT coefficients do not depend on level; 4) local scaling can be assessed due to the time-invariance property. As we will discuss later, the price we pay is that the dependence of coefficients in NDWT is more profound than in DWT.

At high level, we propose to estimate Hurst exponent from NDWT as follows. At each detail level j , we generate $N/2$ mid-energies as $D_{j,k} = (d_{j,k}^2 + d_{j,k+N/2}^2)/2$, for $k = 1, 2, \dots, N/2$. Then we have two different approaches to robustly estimate Hurst exponent: One is based on mid-energies $D_{j,k}$ themselves, and the other is based on the logarithm of

mid-energies $\log D_{j,k}$. In each approach, we first calculate the general trimean estimator on $D_{j,k}$ or $(\log D_{j,k})$, and then derive its asymptotic distribution, which depends on Hurst exponent H and allows us to provide a robust estimation of H .

Note that the asymptotic distribution is conducted under the independent assumption between mid-energies. Unfortunately, for a fixed detail level j , these mid-energies or the logarithm versions are generally dependent. The good news is that their autocorrelations decay exponentially as their distance increases. Thus for the practice purpose, we will be able to reduce such dependency by increasing the distance between two consecutive points.

To be specific, we sample every M points from the original $N/2$ mid-energies or their logarithm versions, resulting in M groups in each level j . Note that at each level j , the M groups are generated by switching the starting point from $D_{j,1}$ ($\log D_{j,1}$) to $D_{j,M}$ ($\log D_{j,M}$). Since the distances are large, the $(N/2)/M$ sampled values within each group can be thought of as independent for the practice purpose. The general trimean estimators are then calculated from each of the M groups. Note that M must be divisible by $N/2$.

$$\begin{aligned} \text{Group 1: } & \{D_{j,1}, D_{j,1+M}, D_{j,1+2M}, \dots, D_{j,(N/2-M+1)}\} \\ & (\{\log(D_{j,1}), \log(D_{j,1+M}), \dots, \log(D_{j,(N/2-M+1)})\}) \\ \text{Group 2: } & \{D_{j,2}, D_{j,2+M}, D_{j,2+2M}, \dots, D_{j,(N/2-M+2)}\} \\ & (\{\log(D_{j,2}), \log(D_{j,2+M}), \dots, \log(D_{j,(N/2-M+2)})\}) \\ & \vdots \end{aligned}$$

$$\begin{aligned} \text{Group M: } & \{D_{j,M}, D_{j,2M}, D_{j,3M}, \dots, D_{j,N/2}\} \\ & (\{\log(D_{j,M}), \log(D_{j,2M}), \dots, \log(D_{j,N/2})\}) \end{aligned}$$

Below we will present our proposed two methods in two subsections. Section 1.3.1

introduces the general trimean of the mid-energy (GTME) method, and Section 1.3.2 discusses the general trimean of the logarithm of mid-energy (GTLME) method. These two methods are closely related, except switching the order of logarithm and general trimean estimator.

1.3.1 General Trimean of the Mid-energy (GTME) Method

Our proposed GTME method involves the following three steps:

- 1) Compute the general trimean estimators $\hat{\mu}_{j,i}$ on

$$\{D_{j,i}, D_{j,i+M}, D_{j,i+2M}, \dots, D_{j,(N/2-M+i)}\} := D(j, i),$$

where $D(j, i)$, for $1 \leq j \leq J$ and $1 \leq i \leq M$, is the i th group of mid-energies at level j in a J -level NDWT of a fBm of size N with Hurst exponent H .

- 2) For each $i = 1, 2, \dots, M$, calculate the regression slope $\hat{\beta}_i$ in the least square linear regression on pairs $(j, \log_2(\hat{\mu}_{j,i}))$, for $J_1 \leq j \leq J_2$.

- 3) Estimate the Hurst exponent by

$$\hat{H}_1 = \frac{1}{M} \sum_{i=1}^M \left(-\frac{\hat{\beta}_i}{2} - \frac{1}{2} \right) = -\frac{\bar{\beta}}{2} - \frac{1}{2}, \quad (1.8)$$

where $\bar{\beta} = \frac{1}{M} \sum_{i=1}^M \hat{\beta}_i$ is the average of regression slopes over the M groups for $i = 1, 2, \dots, M$.

The motivation of GTME method is based on the asymptotic distribution of $\hat{\mu}_{j,i}$ from Lemma 1.2.1:

$$\sqrt{N} (\hat{\mu}_{j,i} - c(\alpha, p) \lambda_j) \overset{\text{approx}}{\sim} \mathcal{N}(0, 2M f(\alpha, p) \lambda_j^2), \quad (1.9)$$

where

$$\begin{aligned}
c(\alpha, p) &= \frac{\alpha}{2} \log \left(\frac{1}{p(1-p)} \right) + (1-\alpha) \log 2, \\
f(\alpha, p) &= \frac{\alpha(1-2p)(\alpha-4p)}{4p(1-p)} + 1, \text{ and} \\
\lambda_j &= \sigma^2 \cdot 2^{-(2H+1)j}.
\end{aligned} \tag{1.10}$$

Here σ is the standard deviation of wavelet coefficients from level 0. Hence, $\log_2(\hat{\mu}_{j,i})$ is linearly related to $(2H+1)j$, which allows us to use the slopes $\hat{\beta}_i$ to estimate $2H+1$ and leads to the proposed estimator in (1.8).

For our proposed GTME method in (1.8), its asymptotic properties are established in the following theorem, whose proof is postponed in Section 1.6:

Theorem 1.3.1. *The estimator \hat{H}_1 follows the asymptotic normal distribution*

$$\sqrt{N}(\hat{H}_1 - H) \overset{\text{approx}}{\sim} \mathcal{N}(0, V_1). \tag{1.11}$$

The asymptotic variance V_1 is a constant number,

$$\begin{aligned}
V_1 &= \frac{6f(\alpha, p)}{(\log 2)^2 (c(\alpha, p))^2 q(J_1, J_2)}, \\
q(J_1, J_2) &= (J_2 - J_1)(J_2 - J_1 + 1)(J_2 - J_1 + 2),
\end{aligned} \tag{1.12}$$

where $c(\alpha, p)$ and $f(\alpha, p)$ are given in (1.10).

There are different ways to determine the tuning parameters α and p in general trimean estimator. One could use the settings in Tukey's trimean estimator ($\alpha = 1/2, p = 1/4$) or in Gastwirth's estimator ($\alpha = 0.6, p = 1/3$). Alternatively, we could find the optimal α and p in the sense of minimizing the asymptotic variance of general trimean estimators $\hat{\mu}_{j,i}$ in (1.9). To see this, we take partial derivatives of $f(\alpha, p)$ with respect to α and p and set

them to 0. The optimal α and p can be obtained by solving

$$\begin{aligned}\frac{\partial f(\alpha, p)}{\partial \alpha} &= -\frac{2p-1}{2p(1-p)}\alpha + \frac{1+p}{2(1-p)} - \frac{3}{2} = 0, \\ \frac{\partial f(\alpha, p)}{\partial p} &= \frac{\alpha(2-\alpha)}{2(1-p)^2} + \frac{\alpha^2(2p-1)}{4p^2(1-p)^2} = 0.\end{aligned}\tag{1.13}$$

Since $p \in (0, 1/2)$, and $\alpha \in [0, 1]$, we get the unique solution $p = 1 - \sqrt{2}/2 \approx 0.3$ and $\alpha = 2p \approx 0.6$. The Hessian matrix of $f(\alpha, p)$ is

$$\begin{bmatrix} \frac{\partial^2 f(\alpha, p)}{\partial \alpha^2} & \frac{\partial^2 f(\alpha, p)}{\partial \alpha \partial p} \\ \frac{\partial^2 f(\alpha, p)}{\partial \alpha \partial p} & \frac{\partial^2 f(\alpha, p)}{\partial p^2} \end{bmatrix} = \begin{bmatrix} -\frac{2p-1}{2p(1-p)} & \frac{2p^2-2\alpha p^2+\alpha(2p-1)}{2p^2(1-p)^2} \\ \frac{2p^2-2\alpha p^2+\alpha(2p-1)}{2p^2(1-p)^2} & \frac{2p^3\alpha(2-\alpha)+\alpha^2 p(1-p)+\alpha^2(2p-1)^2}{2p^3(1-p)^3} \end{bmatrix}.$$

Since $-\frac{2p-1}{2p(1-p)} > 0$ and the determinant is $5.66 > 0$ when $p = 1 - \sqrt{2}/2 \approx 0.3$ and $\alpha = 2p \approx 0.6$, the above Hessian matrix is positive definite. Therefore, $p = 1 - \sqrt{2}/2$ and $\alpha = 2 - \sqrt{2}$ provide the global minima of $f(\alpha, p)$, minimizing the asymptotic variance of $\hat{\mu}_{j,i}$. In comparing these optimal $\alpha \approx 0.6$ and $p \approx 0.3$ with $\alpha = 0.6$ and $p = 1/3$ from the Gastwirth estimator, curiously, we find that the calculated optimal general trimean estimator is very close to the Gastwirth estimator.

1.3.2 General Trimean of the Logarithm of Mid-energy (GTLME) Method

In this section, we propose our second method, the general trimean of the logarithm of mid-energy (GTLME) method, which takes logarithm first and then calculate the general trimean estimators. The GTLME method involves the following three steps:

- 1) Calculate the general trimean estimators $\hat{\mu}'_{j,i}$ on

$$\{\log(D_{j,i}), \log(D_{j,i+M}), \dots, \log(D_{j,(N/2-M+i)})\} := L(j, i),$$

where $L(j, i)$ is the i th group of logarithm of mid-energies at level j in a J -level NDWT of a fBm of size N with Hurst exponent H , $1 \leq i \leq M$ and $1 \leq j \leq J$.

2) Obtain the regression slope $\hat{\beta}'_i$ in the least square linear regressions on pairs $(j, \hat{\mu}'_{j,i})$ for $J_1 \leq j \leq J_2$.

3) Estimate the Hurst exponent by

$$\hat{H}_2 = \frac{1}{M} \sum_{i=1}^M \left(-\frac{\hat{\beta}'_i}{2 \log 2} - \frac{1}{2} \right) = -\frac{1}{2 \log 2} \bar{\beta}' - \frac{1}{2}, \quad (1.14)$$

where $\bar{\beta}' = \frac{1}{M} \sum_{i=1}^M \hat{\beta}'_i$ is the average of regression slopes over the M groups, $i = 1, 2, \dots, M$.

The motivation of GTLME method is from the asymptotic distribution of general trimean estimator $\hat{\mu}'_{j,i}$ that is derived from Lemma 1.2.1:

$$\sqrt{N}(\hat{\mu}'_{j,i} - (g(\alpha, p) + \log(\lambda_j))) \overset{\text{approx}}{\sim} \mathcal{N}(0, 2Mh(\alpha, p)), \quad (1.15)$$

where

$$\begin{aligned} g(\alpha, p) &= \frac{\alpha}{2} \log \left(\log \frac{1}{1-p} \cdot \log \frac{1}{p} \right) + (1-\alpha) \log(\log 2), \\ h(\alpha, p) &= \frac{\alpha^2}{4h_1(p)} + \frac{\alpha(1-\alpha)}{2h_2(p)} + \frac{(1-\alpha)^2}{(\log 2)^2}. \end{aligned} \quad (1.16)$$

The $h_1(p)$ and $h_2(p)$ are two functions of p that are provided in Section 1.6, $\lambda_j = \sigma^2 \cdot 2^{-(2H+1)j}$, and σ^2 is the variance of wavelet coefficients from level 0.

It is interesting to compare our two proposed methods, GTME and GTLME. The main difference is whether to calculate general trimean estimators before or after taking the logarithm. In GTLME method, the general trimean estimator $\hat{\mu}'_{j,i}$, not the $\log_2(\hat{\mu}'_{j,i})$, is linearly related to $(2H+1)j$, and we use the slopes $\bar{\beta}'_i$ to estimate $2H+1$, therefore leading to the proposed estimator in (1.14). Based on our extensive experience, the GTME seems more efficient in terms of smaller variance, whereas the GTLME method is more robust to outliers.

The asymptotic distribution of Hurst exponent estimator \hat{H}_2 in GTLME method is pro-

vided in the following theorem, whose proof is postponed in Section 1.6.

Theorem 1.3.2. *The estimator \hat{H}_2 follows the asymptotic normal distribution*

$$\sqrt{N}(\hat{H}_2 - H) \overset{approx}{\sim} \mathcal{N}(0, V_2). \quad (1.17)$$

The asymptotic variance V_2 is a constant number,

$$V_2 = \frac{6h(\alpha, p)}{(\log 2)^2 q(J_1, J_2)}, \quad (1.18)$$

where $q(J_1, J_2)$ is given in (1.12) and $h(\alpha, p)$ is in (1.16).

Now we want to determine the optimal tuning parameters α and p . Again, we can set $\alpha = 1/2$ and $p = 1/4$ from Tukey's trimean estimator or $\alpha = 0.6$ and $p = 1/3$ from Gastwirth's estimator. Here, we will find the optimal α and p by minimizing the asymptotic variance of general trimean estimator $\hat{\mu}'_{j,i}$ in (1.15), and the corresponding results also lead to the smallest asymptotic variance V_2 in (1.18). They can be obtained by solving

$$\frac{\partial h(\alpha, p)}{\partial \alpha} = 0, \text{ and } \frac{\partial h(\alpha, p)}{\partial p} = 0. \quad (1.19)$$

From the first equation in (1.19) it can be derived that

$$\alpha = \frac{\frac{2}{\log(2)^2} - \frac{1}{2}h_2(p)}{\frac{1}{2}h_1(p) - h_2(p) + \frac{2}{(\log 2)^2}}. \quad (1.20)$$

The second equation in (1.19) cannot be simplified to a finite form. As an illustration, we plot $h(\alpha, p)$ with p ranging from 0 to 0.5, and α is a function of p in (1.20). The plot of α against p is also shown in Figure 2.2. Numerical calculation gives $\alpha = 0.5965$ and $p = 0.24$. These optimal parameters are close to $\alpha = 0.5$ and $p = 0.25$ in the Tukey's trimean estimator, but put some more weight on the median.

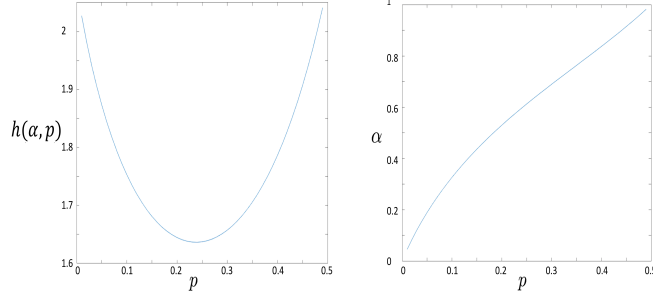


Figure 1.2: Plot of $h(\alpha, p)$ against p on the left; Plot of α against p on the right

1.4 Simulation

In this section, we will illustrate our proposed methods via simulation. We simulate one dimensional fBm signals of sizes $N = 2^{10}$, $N = 2^{11}$, and $N = 2^{12}$ with Hurst exponent $H = 0.3, 0.5, 0.7, 0.8, 0.9$, respectively. NDWT of depth $J = 10$ using Pollen wavelets with angles $\pi/6$ (Daubechies 2), $\pi/4$, $\pi/3$, and $\pi/2$ (Haar) are performed on each simulated signal to obtain wavelet coefficients. Pollen wavelets with different angles generates a family possessing continuum many wavelet bases of various degrees of regularity, see [32]. Special cases of Pollen's representation for $\pi/6$ and $\pi/2$ give Daubechies 2 filter and Haar filter, respectively.

Our proposed methods, GTME and GTLME, are then applied on the NDWT coefficients to estimate Hurst exponent H . We select different combinations of parameters α and p in each method, leading to the following 6 variations:

$$\text{I:} \quad \text{GTME with } \alpha = \frac{1}{2}, p = \frac{1}{4}; \quad (1.21)$$

$$\text{II:} \quad \text{GTLME with } \alpha = \frac{1}{2}, p = \frac{1}{4}; \quad (1.22)$$

$$\text{III:} \quad \text{GTME with } \alpha = 0.6, p = \frac{1}{3}; \quad (1.23)$$

$$\text{IV:} \quad \text{GTLME with } \alpha = 0.6, p = \frac{1}{3}; \quad (1.24)$$

$$\text{V:} \quad \text{GTME with } \alpha = 2 - \sqrt{2}, p = 1 - \frac{\sqrt{2}}{2}; \quad (1.25)$$

$$\text{VI:} \quad \text{GTLME with } \alpha = 0.5965, p = 0.24. \quad (1.26)$$

Table 1.1: Simulation Results for $N = 2^{10}$ fBm using Haar wavelet

H	Existing Methods					Proposed Methods					
	VA	SSB	MEDL	MEDLA	TT	I	II	III	IV	V	VI
\hat{H}											
0.3	0.2488	0.2490	0.2424	0.2421	0.2130	0.2461	0.2457	0.2456	0.2453	0.2460	0.2460
0.5	0.4947	0.5007	0.4860	0.4860	0.4971	0.4926	0.4914	0.4908	0.4905	0.4916	0.4920
0.7	0.7018	0.7251	0.6994	0.6974	0.7815	0.7078	0.7057	0.7051	0.7043	0.7065	0.7066
0.8	0.8007	0.8341	0.8036	0.8020	0.9299	0.8096	0.8082	0.8069	0.8064	0.8083	0.8091
0.9	0.9139	0.9502	0.9072	0.9002	1.0000	0.9075	0.9071	0.9058	0.9054	0.9069	0.9083
Variances											
0.3	0.0024	0.0021	0.0027	0.0021	0.0016	0.0020	0.0021	0.0020	0.0021	0.0020	0.0020
0.5	0.0044	0.0034	0.0034	0.0028	0.0061	0.0028	0.0029	0.0028	0.0029	0.0028	0.0029
0.7	0.0063	0.0054	0.0049	0.0038	0.0160	0.0037	0.0039	0.0037	0.0039	0.0037	0.0039
0.8	0.0091	0.0077	0.0064	0.0050	0.0255	0.0048	0.0053	0.0051	0.0054	0.0049	0.0053
0.9	0.0106	0.0081	0.0068	0.0046	0.0304	0.0040	0.0048	0.0044	0.0048	0.0042	0.0048
MSEs											
0.3	0.0050	0.0047	0.0060	0.0054	0.0092	0.0049	0.0050	0.0050	0.0051	0.0049	0.0050
0.5	0.0044	0.0034	0.0036	0.0030	0.0061	0.0028	0.0029	0.0029	0.0030	0.0028	0.0029
0.7	0.0063	0.0060	0.0049	0.0038	0.0226	0.0037	0.0039	0.0038	0.0039	0.0037	0.0039
0.8	0.0091	0.0089	0.0063	0.0050	0.0423	0.0048	0.0054	0.0051	0.0054	0.0050	0.0054
0.9	0.0108	0.0106	0.0068	0.0046	0.0760	0.0041	0.0048	0.0044	0.0048	0.0042	0.0049

Note. Different variations of our proposed methods can be found in (1.21)-(1.26).

Variations I and II are based on Tuckey's trimean estimator in (1.1), and variations III and III use Gastwirth's estimator in (1.2). The α and p in variations V and VI are the optimal values obtained in Section 3 to minimize the corresponding asymptotic variance of general trimean estimator.

Wavelet coefficients on each level are divided into eight groups ($M = 8$), and we use wavelet coefficients from levels 4 ($J_1 = 4$) to 10 ($J_2 = 10$) for the least square linear regression. The estimation performance of the proposed methods are compared to five other existing methods: VA method, SSB method, MEDL method, MEDLA method, and TT method. Those methods have been discussed in Introduction. Estimation performance is reported in terms of mean, variance, and mean square error (MSE) based on 300 repetitions for each case.

Table 1.2: Simulation Results for $N = 2^{11}$ fBm using Haar wavelet

H	Existing Methods					Proposed Methods					
	VA	SSB	MEDL	MEDLA	TT	I	II	III	IV	V	VI
\hat{H}											
0.3	0.2504	0.2438	0.2393	0.2392	0.2128	0.2412	0.2412	0.2407	0.2407	0.2408	0.2415
0.5	0.4846	0.4784	0.4688	0.4686	0.4812	0.4717	0.4712	0.4707	0.4705	0.4712	0.4715
0.7	0.7260	0.7124	0.6898	0.6909	0.8080	0.6968	0.6953	0.6956	0.6949	0.6962	0.6957
0.8	0.8262	0.8147	0.7893	0.7895	0.9576	0.7941	0.7932	0.7930	0.7926	0.7934	0.7936
0.9	0.9610	0.9345	0.9028	0.8960	1.0000	0.8984	0.9003	0.8980	0.8989	0.8981	0.9010
Variances											
0.3	0.0014	0.0009	0.0011	0.0009	0.0010	0.0008	0.0009	0.0009	0.0009	0.0009	0.0009
0.5	0.0025	0.0013	0.0014	0.0011	0.0037	0.0011	0.0012	0.0012	0.0012	0.0012	0.0012
0.7	0.0071	0.0027	0.0025	0.0019	0.0200	0.0019	0.0021	0.0020	0.0021	0.0019	0.0020
0.8	0.0099	0.0036	0.0031	0.0024	0.0307	0.0022	0.0024	0.0023	0.0024	0.0022	0.0024
0.9	0.0125	0.0049	0.0041	0.0028	0.0365	0.0026	0.0030	0.0027	0.0029	0.0026	0.0030
MSEs											
0.3	0.0038	0.0041	0.0048	0.0046	0.0086	0.0043	0.0043	0.0044	0.0044	0.0044	0.0043
0.5	0.0027	0.0018	0.0024	0.0021	0.0041	0.0019	0.0020	0.0020	0.0021	0.0020	0.0020
0.7	0.0078	0.0028	0.0026	0.0020	0.0316	0.0019	0.0021	0.0020	0.0021	0.0019	0.0020
0.8	0.0106	0.0038	0.0032	0.0025	0.0554	0.0022	0.0024	0.0023	0.0024	0.0022	0.0024
0.9	0.0162	0.0060	0.0041	0.0028	0.1134	0.0026	0.0030	0.0027	0.0029	0.0026	0.0030

Note. Different variations of our proposed methods can be found in (1.21)-(1.26).

The proposed methods perform the best using Haar wavelet (Pollen wavelets with angle $\pi/2$), and the simulation results are shown in Table 1.1 to Table 1.3 for fBm of sizes $N = 2^{10}$, $N = 2^{11}$, and $N = 2^{12}$, respectively. Similar results are obtained for other wavelets. For each H (corresponding to each row in the table), the smallest variances and MSEs are highlighted in bold. From simulation results, at least one of our 6 variations outperforms MEDL and MEDLA for all H and fBm of all three sizes. Compared with VA, SSB, and TT methods, our methods yield significantly smaller variances and MSEs when $H > 0.5$ for fBm of all three sizes. When $H = 0.3$ and 0.5 , our methods are still comparable to VA, SSB, and TT. The performances of our 6 variations are very similar regarding to variances and MSEs, in particular, variation I based on Tukey's trimean estimator of the mid-energy has the best performance among all of them.

Table 1.3: Simulation Results for $N = 2^{12}$ fBm using Haar wavelet

H	Existing Methods					Proposed Methods					
	VA	SSB	MEDL	MEDLA	TT	I	II	III	IV	V	VI
\hat{H}											
0.3	0.2446	0.2356	0.2324	0.2328	0.2066	0.2341	0.2333	0.2334	0.2331	0.2337	0.2335
0.5	0.4969	0.4746	0.4662	0.4662	0.4936	0.4688	0.4678	0.4680	0.4675	0.4684	0.4680
0.7	0.7375	0.6992	0.6840	0.6843	0.8214	0.6876	0.6865	0.6869	0.6862	0.6874	0.6867
0.8	0.8696	0.8113	0.7944	0.7919	1.0238	0.7948	0.7949	0.7944	0.7946	0.7945	0.7950
0.9	0.9931	0.9156	0.8928	0.8887	1.0000	0.8913	0.8928	0.8915	0.8923	0.8912	0.8932
Variances											
0.3	0.0007	0.0005	0.0007	0.0006	0.0005	0.0005	0.0006	0.0005	0.0006	0.0005	0.0006
0.5	0.0025	0.0007	0.0008	0.0006	0.0045	0.0006	0.0006	0.0006	0.0006	0.0006	0.0006
0.7	0.0077	0.0013	0.0014	0.0011	0.0224	0.0010	0.0011	0.0011	0.0011	0.0010	0.0011
0.8	0.0124	0.0019	0.0017	0.0013	0.0375	0.0013	0.0014	0.0014	0.0014	0.0013	0.0014
0.9	0.0169	0.0030	0.0029	0.0020	0.0507	0.0018	0.0021	0.0019	0.0021	0.0018	0.0021
MSEs											
0.3	0.0038	0.0047	0.0053	0.0051	0.0093	0.0049	0.0050	0.0050	0.0050	0.0049	0.0050
0.5	0.0025	0.0013	0.0019	0.0018	0.0046	0.0015	0.0016	0.0016	0.0017	0.0016	0.0016
0.7	0.0091	0.0013	0.0017	0.0013	0.0370	0.0011	0.0013	0.0012	0.0013	0.0012	0.0013
0.8	0.0172	0.0020	0.0018	0.0014	0.0875	0.0013	0.0014	0.0014	0.0014	0.0013	0.0014
0.9	0.0255	0.0033	0.0030	0.0021	0.1514	0.0018	0.0021	0.0020	0.0021	0.0019	0.0021

Note. Different variations of our proposed methods can be found in (1.21)-(1.26).

1.5 Application

In this section, we apply the proposed methods to PRB data in order to classify individuals according to their visual impairment. Participants in this study consists of 24 older adults, solicited from the patient pool of the Bascom Palmer Eye Institute of the University of Miami School of Medicine. Participants were selected on the basis of having either no ocular disease or only Age-related Macular Degeneration (AMD), as assessed by patient history and clinical testing.

Participants were assigned to three groups: one control group, and two experimental groups I and II. The control group is a set of individuals with healthy, unaffected vision and no evidence of any ocular disease or trauma. Individuals in two experimental groups had

varying visual acuity and were diagnosed with AMD. Patients in group II had more severe visual impairment than those in group I. The number of participants is 6 in control group, 8 in group I, and 10 in group II. In Introduction part, we have already shown in Figure 1.1 the examples of raw PRB data of three different individuals from control group, group I, and group II, respectively.

Researchers have utilized simple statistical methods for analyzing PRB, for example, comparing the relative mean or variance of pupil size deviation in response to stimuli; some sophisticated techniques have also been utilized, like power, frequency and spectral analysis using mathematical tools. However, they failed to characterize the underlying patterns within time series PRB data. Wavelet analysis to estimate the Hurst exponent of the high-frequency, time series physiological data is a useful tool for detecting these hidden patterns and differentiating individuals based on these unique patterns in their physiological behavior.

Like in many human-subject studies, the limited number of participants is a major disadvantage, but in PRB data set, each subject has enough measurements to segment into multiple pieces with a length of 2048 observations. Although this induces dependence between the data, we will use hierarchical models to accommodate for the subject induced dependence later. In order to illustrate the robustness of our methods, the proposed methods and other existing methods will be applied to both the original, noisy data and the cleaned data with blink and equipment artifacts removed. Here we use the same 6 variations of our proposed methods in Simulation section, and the parameter settings of those variations can be found in (1.21)-(1.26). The number of 2048 length original data and cleaned data within each group are shown in Table 1.4.

Table 1.5 and Table 1.6 provide descriptive statistics of the estimated Hurst exponent \hat{H} from original data and cleaned data, respectively. As can be seen from Table 1.6, for the cleaned data, the control group exhibited the smallest value for \hat{H} in both the mean and median. In fact, signals with smaller Hurst exponent H tend to be more disordered and

Table 1.4: Group characterization summary

Group	N	Visual Acuity	AMD	Number of original data	Number of cleaned data
Control	6	20/20-20/40	No	60	49
I	8	20/20-20/50	Yes	100	92
II	10	20/60-20/100	Yes	96	262

Note. N represents the number of individuals in the group; Visual Acuity signifies the range of Snellen acuity scores for the individuals in the given group; AMD indicates whether the individuals were diagnosed with age-related macular degeneration or not; Number of original data and Number of cleaned data show the number of 2048 length original data and cleaned data, respectively.

unsystematic, therefore individuals without visual impairment tend to have more disordered pupil diameter signals. However, for the original data, control group did not exhibit the smallest \hat{H} due to the noise caused by blinks and equipment artifacts, which can be seen from Table 1.5.

The objective is to classify the visual impairment groups based on the estimated Hurst exponent for a given 2048 length pupil diameter data. Before doing the classification, we need to first deal with subject induced dependence through the following hierarchical model. If we denote i to be the group index where the piece of observations is from, with $i = 0$ for control group, $i = 1$ for group I, $i = 2$ for group II, and n_j as the number of pieces generated from subject j ($j=1,2,\dots,24$), the estimated Hurst exponent \hat{H}_{ijk} for the k th piece of subject j nested in group i can be expressed in the following model:

$$\hat{H}_{ijk} = \mu + \alpha_i + \beta_{j(i)} + \epsilon_{ijk}, \quad (1.27)$$

where μ is the overall mean, α_i is the effect for i th group, $\beta_{j(i)}$ is the effect for j th participant within i th group, and ϵ_{ijk} is the random error. In avoid of dependency between data due to subject effects, the estimated $\hat{\beta}_{j(i)}$ is first subtracted from \hat{H}_{ijk} , and then multinomial logistic regression model is fitted on the data $\left\{ \left(\hat{H}_{ijk} - \hat{\beta}_{j(i)}, i \right), i = 0, 1, 2, j = 1, \dots, 24, k = 1, \dots, n_j \right\}$. To test the model performance, we randomly choose 80% of the data points to form a training set, and the remaining 20% forms the testing set. Model is developed on the training set and applied on the testing set.

Table 1.5: Descriptive Statistics Group Summary (original noisy data)

H	Existing Methods					Proposed Methods					
	VA	SSB	MEDL	MEDLA	TT	I	II	III	IV	V	VI
Mean of \hat{H}											
Control	0.2206	0.4242	0.3583	0.3660	0.2740	0.3524	0.3602	0.3602	0.3615	0.3567	0.3590
I	0.2781	0.5698	0.4195	0.4201	0.3346	0.4229	0.4255	0.4261	0.4262	0.4254	0.4246
II	0.1949	0.4391	0.3522	0.3297	0.2306	0.3099	0.3273	0.3214	0.3292	0.3163	0.3260
Median of \hat{H}											
Control	0.2336	0.4511	0.3416	0.3597	0.2795	0.3734	0.3681	0.3443	0.3540	0.3479	0.3719
I	0.2537	0.5696	0.4036	0.4227	0.3301	0.4219	0.4255	0.4197	0.4283	0.4174	0.4262
II	0.2107	0.4322	0.3686	0.3396	0.2544	0.3248	0.3370	0.3369	0.3406	0.3319	0.3377
Variance of \hat{H}											
Control	0.0197	0.0358	0.0349	0.0414	0.0191	0.0327	0.0333	0.0356	0.0353	0.0340	0.0317
I	0.0190	0.0344	0.0153	0.0158	0.0177	0.0186	0.0166	0.0174	0.0168	0.0181	0.0167
II	0.0225	0.0381	0.0190	0.0164	0.0293	0.0185	0.0167	0.0173	0.0167	0.0178	0.0167

Note. Different variations of our proposed methods can be found in (1.21)-(1.26).

Misclassification rates are reported in Table 1.7. When doing the classification on the original noised data, our robust methods performed the best, with the minimal misclassification error 37.21%. On the blinks-removed data, our methods outperformed or were comparable to other methods. In general, our methods provide a robust tool to classify different degrees of visual impairment for AMD patients.

1.6 Proof of Theorems 1.3.1 and 1.3.2

In this section, we provide the detailed proofs of Theorem 1.3.1 and Theorem 1.3.2.

Proof of Theorem 1.3.1. A single wavelet coefficient in a non-decimated wavelet transform of a fBm of size N with Hurst exponent H is normally distributed, with variance depending on its level j , therefore, each pair $d_{j,k}$ and $d_{j,k+N/2}$ in mid-energy $D_{j,k}$ are assumed to be independent and follow the same normal distribution.

$$d_{j,k}, d_{j,k+N/2} \sim \mathcal{N}(0, 2^{-(2H+1)j}\sigma^2).$$

Table 1.6: Descriptive Statistics Group Summary (cleaned data)

H	Existing Methods					Proposed Methods					
	VA	SSB	MEDL	MEDLA	TT	I	II	III	IV	V	VI
Mean of \hat{H}											
Control	0.1811	0.3650	0.3025	0.2804	0.2466	0.2722	0.2826	0.2806	0.2851	0.2772	0.2821
I	0.2751	0.5601	0.4048	0.4092	0.3311	0.3926	0.4040	0.4013	0.4076	0.3975	0.4031
II	0.2088	0.4356	0.3494	0.3246	0.2489	0.3028	0.3198	0.3140	0.3219	0.3085	0.3186
Median of \hat{H}											
Control	0.1775	0.3801	0.3311	0.3010	0.2387	0.3012	0.3065	0.3043	0.3070	0.3004	0.3040
I	0.2729	0.5210	0.4168	0.4117	0.3354	0.3959	0.4122	0.4033	0.4113	0.4011	0.4105
II	0.2301	0.4227	0.3580	0.3380	0.2865	0.3121	0.3329	0.3237	0.3345	0.3171	0.3312
Variance of \hat{H}											
Control	0.0094	0.0238	0.0126	0.0116	0.0091	0.0122	0.0122	0.0123	0.0122	0.0124	0.0121
I	0.0077	0.0310	0.0097	0.0131	0.0057	0.0121	0.0124	0.0124	0.0126	0.0122	0.0123
II	0.0149	0.0390	0.0182	0.0153	0.0195	0.0148	0.0152	0.0149	0.0152	0.0147	0.0152

Note. Different variations of our proposed methods can be found in (1.21)-(1.26).

Table 1.7: Classification error

	Existing Methods					Proposed Methods					
	VA	SSB	MEDL	MEDLA	TT	I	II	III	IV	V	VI
Blinks removed	0.4568	0.4074	0.4691	0.3704	0.4444	0.3951	0.3827	0.3827	0.3827	0.3951	0.3827
Original data	0.4651	0.3953	0.4535	0.3837	0.4419	0.3721	0.3837	0.3721	0.3721	0.3837	0.3721

Note. Different variations of our proposed methods can be found in (1.21)-(1.26).

Then the mid-energy is defined as

$$D_{j,k} = \frac{\left(d_{j,k}^2 + d_{j,k+N/2}^2\right)}{2}, \quad j = 1, \dots, J, \text{ and } k = 1, \dots, N/2,$$

and it can be readily shown that $D_{j,k}$ has exponential distribution with scale parameter $\lambda_j = \sigma^2 \cdot 2^{-(2H+1)j}$, i.e.,

$$f(D_{j,k}) = \lambda_j^{-1} e^{-\lambda_j^{-1} D_{j,k}}, \text{ for any } k = 1, \dots, N/2.$$

Therefore the i th subgroup $\{D_{j,i}, D_{j,i+M}, D_{j,i+2M}, \dots, D_{j,(N/2-M+i)}\}$ are i.i.d. $\exp(\lambda_j^{-1})$, and when applying general trimean estimator $\hat{\mu}_{j,i}$ on it, following the derivation in Section

1.2, we have

$$\boldsymbol{\xi} = \left[\log \left(\frac{1}{1-p} \right) \lambda_j \quad \log(2) \lambda_j \quad \log \left(\frac{1}{p} \right) \lambda_j \right]^T,$$

and

$$\Sigma = \begin{bmatrix} \frac{p}{(1-p)} \lambda_j^2 & \frac{p}{(1-p)} \lambda_j^2 & \frac{p}{(1-p)} \lambda_j^2 \\ \frac{p}{(1-p)} \lambda_j^2 & \lambda_j^2 & \lambda_j^2 \\ \frac{p}{(1-p)} \lambda_j^2 & \lambda_j^2 & \frac{1-p}{p} \lambda_j^2 \end{bmatrix}_{3 \times 3},$$

therefore, the asymptotic distribution of $\hat{\mu}_{j,i}$ is normal with mean

$$\begin{aligned} \mathbb{E}(\hat{\mu}_{j,i}) &= A \cdot \boldsymbol{x} \\ &= \left(\frac{\alpha}{2} \log \left(\frac{1}{p(1-p)} \right) + (1-\alpha) \log 2 \right) \lambda_j \\ &\triangleq c(\alpha, p) \lambda_j, \end{aligned}$$

and variance

$$\begin{aligned} \text{Var}(\hat{\mu}_{j,i}) &= \frac{2M}{N} A \Sigma A^T \\ &= \frac{2M}{N} \left(\frac{\alpha(1-2p)(\alpha-4p)}{4p(1-p)} + 1 \right) \lambda_j^2 \\ &\triangleq \frac{2M}{N} f(\alpha, p) \lambda_j^2. \end{aligned}$$

Since the Hurst exponent can be estimated as

$$\hat{H}_1 = -\frac{\bar{\beta}}{2} - \frac{1}{2}, \quad (1.28)$$

where $\bar{\beta} = 1/M \sum_{i=1}^M \hat{\beta}_i$ is the average regression slope in the least square linear regression on pairs $(j, \log_2(\hat{\mu}_{j,i}))$ from level J_1 to J_2 , $J_1 \leq j \leq J_2$. It can be easily derived that each $\hat{\beta}_i$ is a linear combination of $\log_2(\hat{\mu}_{j,i})$,

$$\hat{\beta}_i = \sum_{j=J_1}^{J_2} a_j \log_2(\hat{\mu}_{j,i}), \quad a_j = \frac{j - (J_1 + J_2)/2}{\sum_{j=J_1}^{J_2} (j - (J_1 + J_2)/2)^2}.$$

We can check that $\sum_{j=J_1}^{J_2} a_j = 0$ and $\sum_{j=J_1}^{J_2} a_j j = 1$. Also, if $X \sim \mathcal{N}(\mu, \sigma^2)$, the approxi-

mate expectation and variance of $g(X)$ are

$$\mathbb{E}(g(X)) = g(\mu) + \frac{g''(\mu)\sigma^2}{2}, \text{ and } \text{Var}(g(X)) = (g'(\mu))^2 \sigma^2,$$

based on which we calculate

$$\begin{aligned} \mathbb{E}(\log_2(\hat{\mu}_{j,i})) &= -(2H + 1)j + \text{Constant}, \text{ and} \\ \text{Var}(\log_2(\hat{\mu}_{j,i})) &= \frac{\frac{2M}{N} f(\alpha, p)}{(\log 2)^2 c^2(\alpha, p)}. \end{aligned}$$

Therefore

$$\begin{aligned} \mathbb{E}(\hat{\beta}_i) &= \sum_{j=J_1}^{J_2} a_j \mathbb{E}(\log_2(\hat{\mu}_{j,i})) = -(2H + 1), \text{ and} \\ \text{Var}(\hat{\beta}_i) &= \sum_{j=J_1}^{J_2} a_j^2 \text{Var}(\log_2(\hat{\mu}_{j,i})) := 4M \times V_1, \end{aligned}$$

and

$$\mathbb{E}(\hat{H}_1) = H, \text{ and } \text{Var}(\hat{H}_1) = \frac{1}{N} \cdot V_1, \quad (1.29)$$

where the asymptotic variance V_1 is a constant number independent of group number M and level j ,

$$V_1 = \frac{6f(\alpha, p)}{(\log 2)^2 (c(\alpha, p))^2 q(J_1, J_2)},$$

and

$$q(J_1, J_2) = (J_2 - J_1)(J_2 - J_1 + 1)(J_2 - J_1 + 2).$$

□

Proof of Theorem 1.3.2. We have stated that $D_{j,k} \sim \text{Exp}(\lambda_j^{-1})$ with scale parameter $\lambda_j = \sigma^2 \cdot 2^{-(2H+1)j}$, so that

$$f(D_{j,k}) = \lambda_j^{-1} e^{-\lambda_j^{-1} D_{j,k}}, \text{ for any } k = 1, \dots, N/2.$$

Let $y_{j,k} = \log(D_{j,k})$ for any $j = 1, \dots, J$ and $k = 1, \dots, N/2$. The pdf and cdf of $y_{j,k}$ are

$$f(y_{j,k}) = \lambda_j^{-1} e^{-\lambda_j^{-1} e^{y_{j,k}}} e^{y_{j,k}},$$

and

$$F(y_{j,k}) = 1 - e^{-\lambda_j^{-1} e^{y_{j,k}}}.$$

The p -quantile can be obtained by solving $F(y_p) = 1 - e^{-\lambda_j^{-1} e^{y_p}} = p$, and $y_p = \log(-\lambda_j \log(1-p))$.

Then it can be shown that $f(y_p) = -(1-p) \log(1-p)$. When applying the general trimean estimator $\hat{\mu}'_{j,i}$ on

$$\{\log(D_{j,i}), \log(D_{j,i+M}), \dots, \log(D_{j,(N/2-M+i)})\},$$

following the derivation in Section 1.2, we get

$$\boldsymbol{\xi} = \begin{bmatrix} \log\left(\log\left(\frac{1}{1-p}\right)\right) + \log(\lambda_j) \\ \log(\log 2) + \log(\lambda_j) \\ \log\left(\log\left(\frac{1}{p}\right)\right) + \log(\lambda_j) \end{bmatrix},$$

and

$$\boldsymbol{\Sigma} = \begin{bmatrix} \frac{p}{(1-p)(\log(1-p))^2} & \frac{p}{(1-p)\log(1-p)\log(\frac{1}{2})} & \frac{p}{(1-p)\log(1-p)\log p} \\ \frac{p}{(1-p)\log(1-p)\log(\frac{1}{2})} & \frac{1}{(\log 2)^2} & \frac{1}{\log(\frac{1}{2})\log p} \\ \frac{p}{(1-p)\log(1-p)\log p} & \frac{1}{\log(\frac{1}{2})\log p} & \frac{1-p}{p(\log p)^2} \end{bmatrix},$$

thus, the asymptotic distribution of $\hat{\mu}'_{j,i}$ is normal with mean

$$\begin{aligned} \mathbb{E}(\hat{\mu}'_{j,i}) &= A \cdot \boldsymbol{\xi} \\ &= \frac{\alpha}{2} \log\left(\log \frac{1}{1-p} \cdot \log \frac{1}{p}\right) + (1-\alpha) \log(\log 2) + \log(\lambda_j) \\ &\triangleq g(\alpha, p) + \log(\lambda_j), \end{aligned}$$

and variance

$$\begin{aligned}\text{Var}(\hat{\mu}'_{j,i}) &= \frac{1}{N/16} A \Sigma A^T \\ &= \frac{2M}{N} \left(\frac{\alpha^2}{4} h_1(p) + \frac{\alpha(1-\alpha)}{2} h_2(p) + \frac{(1-\alpha)^2}{(\log 2)^2} \right) \\ &\triangleq \frac{2M}{N} h(\alpha, p),\end{aligned}$$

where

$$\begin{aligned}h_1(p) &= \frac{p}{(1-p)(\log(1-p))^2} + \\ &\quad \frac{1-p}{p(\log p)^2} + \frac{2p}{(1-p)\log(1-p)\log p},\end{aligned}$$

and

$$h_2(p) = \frac{2p}{(1-p)\log(1-p)\log \frac{1}{2}} + \frac{2}{\log \frac{1}{2} \log p}.$$

Since the Hurst exponent can be estimated as

$$\hat{H}_2 = -\frac{1}{2 \log 2} \bar{\beta}' - \frac{1}{2}, \quad (1.30)$$

where $\bar{\beta}' = 1/M \sum_{i=1}^M \hat{\beta}'_i$ is the average regression slope in the least square linear regressions on pairs $(j, \hat{\mu}'_{j,i})$ from level J_1 to J_2 , $J_1 \leq j \leq J_2$. It can be easily derived that $\hat{\beta}'_i$ is a linear combination of $\hat{\mu}'_{j,i}$,

$$\hat{\beta}'_i = \sum_{j=J_1}^{J_2} a_j \hat{\mu}'_{j,i}, \quad a_j = \frac{j - (J_1 + J_2)/2}{\sum_{j=J_1}^{J_2} (j - (J_1 + J_2)/2)^2}.$$

Again, we can check that $\sum_{j=J_1}^{J_2} a_j = 0$ and $\sum_{j=J_1}^{J_2} a_j j = 1$. Therefore

$$\begin{aligned}\mathbb{E}(\hat{\beta}'_i) &= \sum_{j=J_1}^{J_2} a_j \mathbb{E}(\hat{\mu}'_{j,i}) = -(2H+1) \log 2, \text{ and} \\ \text{Var}(\hat{\beta}'_i) &= \sum_{j=J_1}^{J_2} a_j^2 \text{Var}(\hat{\mu}'_{j,i}) := 4(\log 2)^2 M \times V_2,\end{aligned}$$

and

$$\mathbb{E}(\hat{H}_2) = H, \text{ and } \text{Var}(\hat{H}_2) = \frac{1}{N} \cdot V_2, \quad (1.31)$$

where the asymptotic variance V_2 is a constant number independent of group number M and level j ,

$$V_2 = \frac{6f(\alpha, p)}{(\log 2)^2 q(J_1, J_2)},$$

and

$$q(J_1, J_2) = (J_2 - J_1)(J_2 - J_1 + 1)(J_2 - J_1 + 2).$$

□

1.7 Conclusions

In this chapter, we proposed two methods, GTME and GTLME, to improve the robust estimation of Hurst exponent from the fractional Brownian motion through wavelet transformations. The three key ideas in our proposed methods are: 1) We define a general trimean estimator that is a weighted average of median and two quantiles, and it turns out that the well known Tukey's trimean estimator and Gastwirth estimator are two special cases under this framework; 2) When utilizing non-decimated wavelet transforms (NDWT) wavelet coefficients to obtain Hurst exponent estimators, we reduce the dependency of NDWT wavelet coefficients by rearranging each level coefficients into groups, so that the distance between any two points within the same group are large enough; 3) Instead of using mean or median, we apply the general trimean estimator to wavelet coefficients (GTME) or the logarithm of wavelet coefficients (GTLME), and then derive its asymptotic distribution, which depends on Hurst exponent H and leads to the robust estimation of H .

The estimation performance of the proposed methods were compared to five other existing methods: Veitch & Abry (VA) method, Soltani, Simard, & Boichu (SSB) method, MEDL method, MEDLA method, and Theil-type regression (TT) method. Simulation re-

sults indicated our proposed methods yielded smaller variance and MSEs when estimating Hurst exponent H , in particular for large H 's. For fBm with small to moderate Hurst exponent, for example $H = 0.3$ or 0.5 , our methods still outperformed MEDL and MEDLA, and were comparable to VA, SSB, and TT.

Our proposed two methods have been applied to a real pupillary response behavior (PRB) data set for visual impairment classification. The unique pattern of PRB data cannot be efficiently represented by the trends or traditional statistical summaries of the signal, and our proposed methods helped to detect those unique patterns by estimating the Hurst exponent from the data. The estimated Hurst exponent was then used as a predictor in the multinomial logistic regression model to classify individuals with different degrees of visual impairment. It turns out that our robust methods yielded the smallest three-class misclassification rate 37.21% on the noisy PRB data. Besides, we noticed that the healthy group exhibited the smallest value for estimated Hurst exponent H , which indicated individuals without visual impairment had more disordered signals. This is common for many other biometric signals: EEG, EKG, high frequency protein mass-spectra, high resolution medical images of tissue, to list a few.

CHAPTER 2
MAMMOGRAM DIAGNOSTICS USING ROBUST WAVELET-BASED
ESTIMATOR OF HURST EXPONENT

The materials in this chapter were published in *New Frontiers of Biostatistics and Bioinformatics*, 2018

2.1 Introduction

Breast cancer is one of the major health concerns among women. It has been estimated by the National Cancer Institute that 1 in 8 women will be diagnosed with breast cancer during their lifetime. Early detection is proven to be the best strategy for improving prognosis. Most of the references dealing with automated breast cancer detection are based on microcalcifications [33, 34, 35, 36, 37]. Recently, predicting disease using image data becomes an active research area in statistics and machine learning [38, 39, 40, 41]. For example, Reiss and Ogden (2010) proposed a functional generalized linear regression model with images as predictors [38]. However, predicting breast cancer based on the tissue images directly is like a black-box. Physicians will have a hard time to summarize the common features from the cancerous images, and the prediction results are not easily interpreted. In this chapter, we study the scaling information from the tissue image and then predict breast cancer based on the estimated scaling parameter. It has been found in literatures that the scaling information is efficient and accurate in early detection of breast cancer [25, 42, 43, 44]. In fact, regular scaling is a common phenomenon in high-frequency signals and high-resolution digital images collected in real life. Examples can be found in a variety of fields including economics, telecommunications, physics, geosciences, as well as in biology and medicine [45, 46, 47, 48, 49, 50, 51].

The standard measure of regular scaling is the Hurst exponent, denoted by H in the

sequel. Recall that a stochastic process $\{X(\mathbf{t}), \mathbf{t} \in \mathbb{R}^d\}$ is self-similar with Hurst exponent H if, for any $\lambda \in \mathbb{R}^+$, $X(\mathbf{t}) \stackrel{d}{=} \lambda^{-H} X(\lambda \mathbf{t})$. Here the notation $\stackrel{d}{=}$ means the equality in all finite-dimensional distributions. The Hurst exponent quantifies the self-similarity and describes the rate at which autocorrelations decrease as the lag between two realizations in a time series increases. A value H in the range 0-0.5 indicates a zig-zagging intermittent time series with long-term switching between high and low values in adjacent pairs. A value H in the range 0.5 to 1 indicates a time series with long-term positive autocorrelations, which preserves trends on a longer time horizon and gives a time series more regular appearance.

Multiresolution analysis is one of the many methods to estimate the Hurst exponent. An overview can be found in [17, 18, 19]. In particular, the non-decimated wavelet transforms (NDWT) [31, 52, 53] has several potential advantages when employed for Hurst exponent estimation. Input signals and images of arbitrary size can be transformed in a straightforward manner due to the absence of decimation. As a redundant transform, the NDWT can decrease variance in the scaling estimation [27]. Least square regression can be fitted to estimate H instead of weighted least square regression since the variances of the level-wise derived distributions based on \log NDWT coefficients do not depend on level. Local scaling can be assessed due to the time-invariance property. Of course, the dependence of coefficients in NDWT is much more pronounced. Similar to Soltani et al [20], we will control this dependence by systematic sampling of coefficients on which the estimator is based.

Different wavelet-based methods for estimation of H have been proposed in the literature for the one-dimensional case. Veitch and Abry [17] suggested the estimation of H by weighted least square regression using the level-wise $\log_2(\overline{d_j^2})$. In addition, the authors corrected for the bias caused by the order of taking the logarithm and the average in $\log_2(\overline{d_j^2})$, where d_j indicates any detail coefficient at level j . We use $d_{j,k}$ to denote the k th coefficient at level j in the sequel. Soltani et al [20] defined a mid-energy as $D_{j,k} = (d_{j,k}^2 + d_{j,k+N_j/2}^2)/2$, and showed that the level-wise averages of $\log_2 D_{j,k}$ are

asymptotically normal and more stable, which is used to estimate H by regression. The estimators in Soltani et al [20] consistently outperform the estimators in Veitch and Abry [17]. Shen et al [21] showed that the method of Soltani et al [20] yields more accurate estimators since it takes the logarithm of the mid-energy first and then averages.

The robust estimation of H has recently become a topic of interest due to the presence of outlier coefficients and outlier multiresolution levels, inter and within level dependences, and distributional contaminations [22, 23, 21, 24]. Hamilton et al (2011) [25] came up with a robust approach based on Theil-type weighted regression [26], a method for robust linear regression that selects the weighted average of all slopes defined by different pairs of regression points. Like the VA method, they regress the level-wise $\log_2(\overline{d_j^2})$ against the level indices, but instead of weighted least square regression, they use the Theil-type weighted regression to make it less sensitive to outlier levels. Kang and Vidakovic [27] proposed MEDL and MEDLA methods based on non-decimated wavelets to estimate H . MEDL estimates H by regressing the medians of $\log d_j^2$ on level j , while MEDLA uses the level-wise medians of $\log((d_{j,k_1}^2 + d_{j,k_2}^2)/2)$ to estimate H , where k_1 and k_2 are properly selected locations at level j to approximate the independence.

Both MEDL and MEDLA use the median of the derived distribution instead of the mean, because the medians are more robust to potential outliers that can occur when logarithmic transform of a squared wavelet coefficient is taken and the magnitude of coefficient is close to zero. Although median is outlier-resistant, it can behave unexpectedly as a result of its non-smooth character. The fact that the median is not “universally the best outlier-resistant estimator” motivates us to develop the general trimean estimators of the level-wise derived distributions to estimate H , where the general trimean estimator was derived as a weighted average of the distribution’s median and two quantiles symmetric about the median, combining the median’s emphasis on center values with the quantiles’ attention to the tails. Tukey’s trimean estimator [28, 54] and Gastwirth estimator [55, 29, 56] are two special cases under such general framework.

In this chapter, we are concerned with the robust estimation of Hurst exponent in self-similar signals. Here, the focus is on images, but the methodology applies to multiscale context of arbitrary dimension. The properties of the proposed Hurst exponent estimators are studied both theoretically and numerically. The performance of the robust approach is compared with other standard wavelet-based methods (Veitch & Abry (VA) method, Soltani, Simard, & Boichu (SSB) method, median based estimators MEDL and MEDLA, and Theil-type (TT) weighted regression method).

The rest of the chapter consists of the following additional sections. Section 2.2 discusses background of non-decimated wavelet transforms and wavelet-based spectrum in the context of estimating the Hurst exponent for fractional Brownian motion (fBm). Section 2.3 introduces the general trimean estimators and discusses two special estimators following that general framework; Section 2.4 describes estimation of Hurst exponent using the general trimean estimators, presents distributional results on which the proposed methods are based, and derives optimal weights that minimize the variances of the estimators. Section 2.5 provides the simulation results and compares the performance of the proposed methods to other standardly used, wavelet-based methods. The proposed methods are applied to classify the digitized mammogram images as cancerous or non-cancerous in Section 2.6. The proofs of Theorems and Lemmas are included in Section 2.7 and Section 2.8, respectively. The chapter is concluded with a summary and discussion in Section 2.9.

2.2 Background

2.2.1 Non-decimated wavelet transforms

The non-decimated wavelet transforms (NDWT) [31, 52, 53] are redundant transforms because they are performed by repeated filtering with a minimal shift, or a maximal sampling rate, at all dyadic scales. Subsequently, the transformed signal contains the same number of coefficients as the original signal at each multiresolution level. We start by describing algorithmic procedure of 1-D NDWT and then expand to 2-D NDWT. Traditionally,

we perform a wavelet transformation as a convolution of an input data with wavelet and scaling filters. A principal difference between NDWT and DWT is the sampling rate.

Any square integrable function $f(x) \in \mathbf{L}_2(\mathbb{R})$ can be expressed in the wavelet domain as

$$f(x) = \sum_k c_{J_0,k} \phi_{J_0,k}(x) + \sum_{j \geq J_0} \sum_k d_{j,k} \psi_{j,k}(x),$$

where $c_{J_0,k}$ denote coarse coefficients, $d_{j,k}$ indicate detail coefficients, $\phi_{J_0,k}(x)$ represent scaling functions, and $\psi_{j,k}(x)$ signify wavelet functions. For specific choices of scaling and wavelet functions, the basis for NDWT can be formed from the atoms

$$\phi_{J_0,k}(x) = 2^{J_0/2} \phi(2^{J_0}(x - k)) \text{ and}$$

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

where $x \in \mathbb{R}$, j is a resolution level, J_0 is the coarsest level, and k is the location of an atom. Notice that atoms for NDWT have the constant location shift k at all levels, yielding the finest sampling rate on any level. The coarse coefficients $c_{J_0,k}$ and detail coefficients $d_{j,k}$ can be obtained via

$$c_{J_0,k} = \int f(x) \phi_{J_0,k}(x) dx \text{ and } d_{j,k} = \int f(x) \psi_{j,k}(x) dx. \quad (2.1)$$

In a J -level decomposition of an 1-D input signal of size N , a NDWT will yield $N \times (J+1)$ wavelet coefficients, including $N \times 1$ coarse coefficients and $N \times J$ detail coefficients.

Expanding on the 1-D definitions, we could easily describe 2-D NDWT of $f(x, y)$ with $(x, y) \in \mathbb{R}^2$. Several versions of 2-D NDWT exist, but we only focus on the scale-mixing version based on which our methods are proposed. For the scale-mixing 2-D NDWT, the wavelet atoms are

$$\phi_{J_{01}, J_{02}; \mathbf{k}}(x, y) = 2^{(J_{01} + J_{02})/2} \phi(2^{J_{01}}(x - k_1)) \phi(2^{J_{02}}(y - k_2)),$$

$$\psi_{J_{01}, J_{02}; \mathbf{k}}(x, y) = 2^{(J_{01}+J_{02})/2} \phi(2^{J_{01}}(x - k_1)) \psi(2^{J_{02}}(y - k_2)),$$

$$\psi_{j_1, J_{02}; \mathbf{k}}(x, y) = 2^{(j_1+J_{02})/2} \psi(2^{j_1}(x - k_1)) \phi(2^{J_{02}}(y - k_2)),$$

$$\psi_{j_1, j_2; \mathbf{k}}(x, y) = 2^{(j_1+j_2)/2} \psi(2^{j_1}(x - k_1)) \psi(2^{j_2}(y - k_2)),$$

where $\mathbf{k} = (k_1, k_2)$ is the location index, J_{01} and J_{02} are coarsest levels, $j_1 > J_{01}$, and $j_2 > J_{02}$. The wavelet coefficients for $f(x, y)$ after the scale-mixing NDWT can be obtained as

$$\begin{aligned} c_{J_{01}, J_{02}; \mathbf{k}} &= \iint f(x, y) \phi_{J_{01}, J_{02}; \mathbf{k}}(x, y) dx dy, \\ h_{J_{01}, j_2; \mathbf{k}} &= \iint f(x, y) \psi_{J_{01}, j_2; \mathbf{k}}(x, y) dx dy, \\ v_{j_1, J_{02}; \mathbf{k}} &= \iint f(x, y) \psi_{j_1, J_{02}; \mathbf{k}}(x, y) dx dy, \\ d_{j_1, j_2; \mathbf{k}} &= \iint f(x, y) \psi_{j_1, j_2; \mathbf{k}}(x, y) dx dy. \end{aligned} \tag{2.2}$$

Note that $c_{J_{01}, J_{02}; \mathbf{k}}$ are coarse coefficients and represent the coarsest approximation, $h_{J_{01}, j_2; \mathbf{k}}$ and $v_{j_1, J_{02}}$ represent the mix of coarse and detail information, and $d_{j_1, j_2; \mathbf{k}}$ carry information about details only. In our methods, only detail coefficients $d_{j_1, j_2; \mathbf{k}}$ are used to estimate H .

2.2.2 The fBm: Wavelet coefficients and Spectra

Among models having been proposed for analyzing the self-similar phenomena, arguably the most popular is the fractional Brownian motion (fBm) first described by Kolmogorov [13] and formalized by Mandelbrot and Van Ness [14].

In this section, an overview of 1-D fBm and its extension to 2-D fBm is provided. Consider a stochastic process $\{X(t), t \in \mathbb{R}\}$ is self-similar with Hurst exponent H , then the 1-D detail coefficients defined in (2.1) satisfy

$$d_{jk} \stackrel{d}{=} 2^{-j(H+1/2)} d_{0k},$$

for a fixed level j [15]. If the process has stationary increments, i.e., $X(t+h) - X(t)$ is

independent of t , then $\mathbb{E}(d_{0k}) = 0$ and $\mathbb{E}(d_{0k}^2) = \mathbb{E}(d_{00}^2)$. We obtain

$$\mathbb{E}(d_{jk}^2) \propto 2^{-j(2H+1)}. \quad (2.3)$$

The Hurst exponent can be estimated by taking logarithms on both sides of equation (2.3). The wavelet spectrum is defined by the sequence $\{S(j) = \log \mathbb{E}(d_{jk}^2), j \in \mathbb{Z}\}$. Fractional Brownian motion (fBm), denoted as $B_H(\mathbf{t})$ is the unique Gaussian process with stationary increments that is self-similar [15, 16]. The definition of the one-dimensional fBm can be extended to the multivariate case. In particular, a two-dimensional fBm, $B_H(\mathbf{t})$, for $\mathbf{t} \in [0, 1] \times [0, 1]$ and $H \in (0, 1)$, is a Gaussian process with stationary zero-mean increments, satisfying

$$B_H(a\mathbf{t}) \stackrel{d}{=} a^H B_H(\mathbf{t}).$$

It can be shown that the detail coefficients $d_{j_1, j_2; k}$ defined in equation (2.2) satisfy

$$\log_2 \mathbb{E}(|d_{j_1, j_2; k}|^2) = -(2H + 2)j + C,$$

which defines the two-dimensional wavelet spectrum, from which the Hurst exponent can be estimated. Our proposed methods in next sections are based on but improve from this spectrum.

2.3 General Trimean Estimators

Let X_1, X_2, \dots, X_n be i.i.d. continuous random variables with pdf $f(x)$ and cdf $F(x)$. Let $0 < p < 1$, and let ξ_p denote the p th quantile of F , so that $\xi_p = \inf\{x | F(x) \geq p\}$. If F is monotone, the p th quantile is simply defined as $F(\xi_p) = p$.

Let $Y_p = X_{[np]:n}$ denote a sample p th quantile. Here $[np]$ denotes the greatest integer that is less than or equal to np . The general trimean estimator is defined as a weighted

average of the distribution's median and its two quantiles Y_p and Y_{1-p} , for $p \in (0, 1/2)$:

$$\hat{\mu} = \frac{\alpha}{2} Y_p + (1 - \alpha) Y_{1/2} + \frac{\alpha}{2} Y_{1-p}. \quad (2.4)$$

The weights for the two quantiles are the same for Y_p and Y_{1-p} , and $\alpha \in [0, 1]$. This is equivalent to the weighted sum of the median and the average of Y_p and Y_{1-p} with weights $1 - \alpha$ and α :

$$\hat{\mu} = (1 - \alpha) Y_{1/2} + \alpha \left(\frac{Y_p + Y_{1-p}}{2} \right).$$

This general trimean estimator turns out to be more robust than mean but smoother than the median. To derive its asymptotic distribution, the asymptotic joint distribution of sample quantiles is needed, as shown in Lemma 1; detailed proof can be found in [30].

Lemma 2.3.1. *Consider r sample quantiles, $Y_{p_1}, Y_{p_2}, \dots, Y_{p_r}$, where $1 \leq p_1 < p_2 < \dots < p_r \leq n$. If for any $1 \leq i \leq r$, $\sqrt{n}(\lfloor np_i \rfloor / n - p_i) \rightarrow 0$ is satisfied, then the asymptotic joint distribution of $Y_{p_1}, Y_{p_2}, \dots, Y_{p_r}$ is:*

$$\sqrt{n}((Y_{p_1}, Y_{p_2}, \dots, Y_{p_r}) - (\xi_{p_1}, \xi_{p_2}, \dots, \xi_{p_r})) \overset{approx}{\rightsquigarrow} \mathcal{MVN}(0, \Sigma),$$

where

$$\Sigma = (\sigma_{ij})_{r \times r},$$

and

$$\sigma_{ij} = \frac{p_i(1 - p_j)}{f(x_{p_i})f(x_{p_j})}, \quad i \leq j. \quad (2.5)$$

From Lemma 2.3.1, the asymptotic distribution of general trimean estimator will be normal as a linear combination of the components each with an asymptotic normal distribution. The general trimean estimator itself may be defined in terms of order statistics as

$$\hat{\mu} = A \cdot \mathbf{y},$$

where

$$A = \begin{bmatrix} \frac{\alpha}{2} & 1 - \alpha & \frac{\alpha}{2} \end{bmatrix}, \text{ and } \mathbf{y} = [Y_p \ Y_{1/2} \ Y_{1-p}]^T.$$

It can be easily verified that $\sqrt{n}(\lfloor pn \rfloor/n - p) \rightarrow 0$ for $p \in (0, 1/2]$. If we denote $\boldsymbol{\xi} = [\xi_p \ \xi_{1/2} \ \xi_{1-p}]^T$ the population quantiles, the asymptotic distribution of \mathbf{y} is

$$\sqrt{n}(\mathbf{y} - \boldsymbol{\xi}) \overset{\text{approx}}{\rightsquigarrow} \mathcal{MVN}(0, \Sigma),$$

where $\Sigma = (\sigma_{ij})_{3 \times 3}$, and σ_{ij} follows equation (2.5) for $p_1 = p$, $p_2 = 1/2$, and $p_3 = 1 - p$.

Therefore

$$\hat{\mu} \overset{\text{approx}}{\rightsquigarrow} \mathcal{N}(\mathbb{E}(\hat{\mu}), \text{Var}(\hat{\mu})),$$

with the theoretical expectation and variance being

$$\mathbb{E}(\hat{\mu}) = \mathbb{E}(A \cdot \mathbf{y}) = A \cdot \mathbb{E}(\mathbf{y}) = A \cdot \boldsymbol{\xi}, \quad (2.6)$$

and

$$\text{Var}(\hat{\mu}) = \text{Var}(A \cdot \mathbf{y}) = A \text{Var}(\mathbf{y}) A^T = \frac{1}{n} A \Sigma A^T. \quad (2.7)$$

2.3.1 Tukey's Trimean Estimator

Tukey's trimean estimator is a special case of the general trimean estimators, with $\alpha = 1/2$ and $p = 1/4$ in equation (2.4). To compute this estimator, we first sort the data in ascending order. Next, we take the values that are one-fourth of the way up this sequence (the first quartile), half way up the sequence (i.e., the median), and three-fourths of the way up the sequence (the third quartile). Given these three values, we then form the weighted average, giving the central (median) value a weight of $1/2$ and the two quartiles a weight of $1/4$ each.

If we denote Tukey's trimean estimator as $\hat{\mu}_T$, then

$$\hat{\mu}_T = \frac{1}{4} Y_{1/4} + \frac{1}{2} Y_{1/2} + \frac{1}{4} Y_{3/4}.$$

The asymptotic distribution is

$$\hat{\mu}_T \overset{\text{approx}}{\sim} \mathcal{N} \left(A_T \cdot \boldsymbol{\xi}_T, \frac{1}{n} A_T \Sigma_T A_T^T \right),$$

where $A_T = \begin{bmatrix} \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \end{bmatrix}$, $\boldsymbol{\xi}_T = [\xi_{1/4} \ \xi_{1/2} \ \xi_{3/4}]^T$, $\Sigma_T = (\sigma_{ij})_{3 \times 3}$ is the covariance matrix of the asymptotic multivariate normal distribution, and σ_{ij} follows equation (2.5) with $p_1 = 1/4$, $p_2 = 1/2$, and $p_3 = 3/4$.

2.3.2 Gastwirth Estimator

As Tukey's estimator, the Gastwirth estimator is another special case of the general trimean estimators, with $\alpha = 0.6$ and $p = 1/3$ in equation (2.4).

If we denote this estimator as $\hat{\mu}_G$, then

$$\hat{\mu}_G = 0.3 Y_{1/3} + 0.4 Y_{1/2} + 0.3 Y_{2/3}.$$

The asymptotic distribution can be derived as

$$\hat{\mu}_G \overset{\text{approx}}{\sim} \mathcal{N} \left(A_G \cdot \boldsymbol{\xi}_G, \frac{1}{n} A_G \Sigma_G A_G^T \right),$$

where $A_G = [0.3 \ 0.4 \ 0.3]$, $\boldsymbol{\xi}_G = [\xi_{1/3} \ \xi_{1/2} \ \xi_{2/3}]^T$, $\Sigma_G = (\sigma_{ij})_{3 \times 3}$, and σ_{ij} follows equation (2.5) with $p_1 = 1/3$, $p_2 = 1/2$, and $p_3 = 2/3$.

2.4 Methods

Our proposal for robust estimation of Hurst exponent H is based on non-decimated wavelet transforms (NDWT). In a J -depth decomposition of a 2-D fBm of size $N \times N$, a scale-mixing 2-D NDWT generates $(J + 1) \times (J + 1)$ blocks of coefficients, with each block the same size as original image, i.e., $N \times N$. The tessellation of coefficients of scale-mixing 2-D NDWT is shown in Figure 2.1(a). From the 2-D NDWT wavelets coefficients, our methods use the diagonal blocks ($j_1 = j_2 = j$) of the detail coefficients $d_{j_1, j_2; \mathbf{k}}$ to predict H , as is shown in Figure 2.1(b).

At each detail level j , the corresponding level- j diagonal block is of size $N \times N$, the same size as original image. Note that those coefficients $d_{j, j; \mathbf{k}}$ in level- j diagonal block are not independent, however, their autocorrelations decay exponentially, that is, they possess only the short memory. We reduce such within block dependency by dividing the block into $M \times M$ equal grids and then random sampling one coefficient from each grid, therefore increasing the distance between two consecutive coefficients. To improve the efficiency, here we apply symmetric sampling. To be specific, we partition the level- j diagonal block into four equal parts (top left, top right, bottom left, and bottom right), only sample from the $M^2/4$ grids at the top left, and then get the corresponding coefficients that have the same location in other parts, which is shown in Figure 2.1(c).

If assuming the coefficient $d_{j, j; (k_{i1}, k_{i2})}$ is randomly sampled from grid $i \in \{1, \dots, \frac{M^2}{4}\}$ at the top left part of level- j diagonal block, and $k_{i1}, k_{i2} \in \{1, 2, \dots, \frac{N}{2}\}$ being the corresponding location indexes, then we can extract corresponding coefficients $d_{j, j; (k_{i1}, k_{i2} + \frac{N}{2})}$, $d_{j, j; (k_{i1} + \frac{N}{2}, k_{i2})}$, and $d_{j, j; (k_{i1} + \frac{N}{2}, k_{i2} + \frac{N}{2})}$ from the top right, bottom left, and bottom right parts, respectively. From the set

$$\left\{ d_{j, j; (k_{i1}, k_{i2})}, d_{j, j; (k_{i1}, k_{i2} + \frac{N}{2})}, d_{j, j; (k_{i1} + \frac{N}{2}, k_{i2})}, d_{j, j; (k_{i1} + \frac{N}{2}, k_{i2} + \frac{N}{2})} \right\},$$

we could generate two mid-energies as

$$D_{i,j} = \frac{d_{j,j;(k_{i1},k_{i2})}^2 + d_{j,j;(k_{i1}+\frac{N}{2},k_{i2}+\frac{N}{2})}^2}{2} \quad (2.8)$$

$$D'_{i,j} = \frac{d_{j,j;(k_{i1},k_{i2}+\frac{N}{2})}^2 + d_{j,j;(k_{i1}+\frac{N}{2},k_{i2})}^2}{2}, \quad i \in \{1, \dots, \frac{M^2}{4}\},$$

where $D_{i,j}$ and $D'_{i,j}$ denote the two mid-energies corresponding to grid i at level j . If we denote D_j as the set of all mid-energies at level j , then

$$D_j = \{D_{1,j}, D'_{1,j}, D_{2,j}, D'_{2,j}, \dots, D_{\frac{M^2}{4},j}, D'_{\frac{M^2}{4},j}\}. \quad (2.9)$$

The $M^2/2$ mid-energies at each level j are treated as if they are independent. Note that M must be divisible by 2.

Our methods have two different versions, one is based on mid-energies D_j , while the other is using logged mid-energies $\log D_j$ (in bracket). First, the distribution of D_j ($\log D_j$) is derived under the independence approximation between $d_{j,j;(k_{i1},k_{i2})}$, $d_{j,j;(k_{i1},k_{i2}+\frac{N}{2})}$, $d_{j,j;(k_{i1}+\frac{N}{2},k_{i2})}$, and $d_{j,j;(k_{i1}+\frac{N}{2},k_{i2}+\frac{N}{2})}$. Next, we calculate the general trimean estimators from the level-wise derived distributions to estimate H .

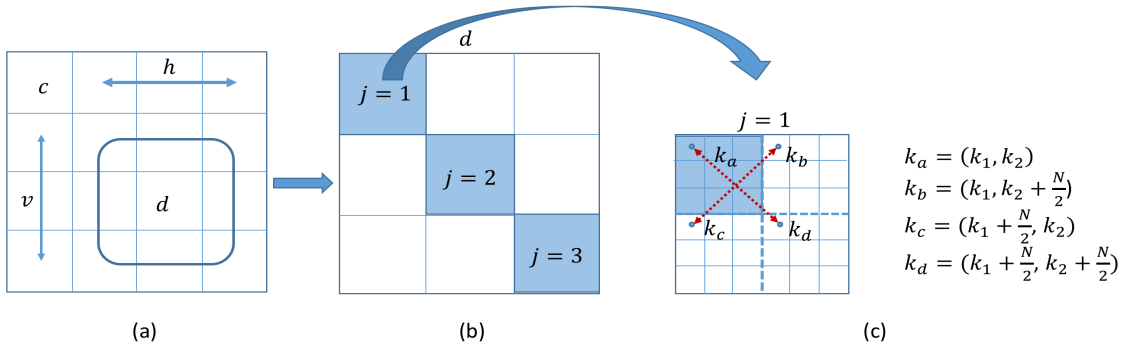


Figure 2.1: (a) Four types of wavelet coefficients with their locations in the tessellation of a 2-D scale mixing NDWT of depth of 3 ($J = 3$), with each block the size of $N \times N$. Coefficients c represent the coarsest approximation, h and v are the mix of coarse and detail information, and d carry detail information only. (b) Detail coefficients d and its diagonal blocks corresponding to 3 ($J = 3$) levels. (c) Symmetric random sampling from level-1 ($j = 1$) diagonal block divided into 6×6 ($M = 6$) grids.

2.4.1 General Trimean of the Mid-energy (GTME) Method

At each decomposition level j , the asymptotic distribution of the general trimean estimator on $M^2/2$ mid-energies in D_j is derived, from which we find the relationship between the general trimean estimators and H . The general trimean of the mid-energy (GTME) method is described in the following theorem:

Theorem 2.4.1. *Let $\hat{\mu}_j$ be the general trimean estimator based on the $M^2/2$ mid-energies in D_j defined by (2.9) at level j in a J -level NDWT of a 2-D fBm of size $N \times N$ with Hurst exponent H . Then, the asymptotic distribution of $\hat{\mu}_j$ is normal,*

$$\hat{\mu}_j \overset{\text{approx}}{\sim} \mathcal{N} \left(c(\alpha, p) \lambda_j, \frac{2}{M^2} f(\alpha, p) \lambda_j^2 \right), \quad (2.10)$$

where

$$c(\alpha, p) = \frac{\alpha}{2} \log \left(\frac{1}{p(1-p)} \right) + (1-\alpha) \log 2,$$

$$f(\alpha, p) = \frac{\alpha(1-2p)(\alpha-4p)}{4p(1-p)} + 1,$$

$$\lambda_j = \sigma^2 \cdot 2^{-(2H+2)j},$$

and σ^2 is the variance of wavelet coefficients from level 0, the Hurst exponent can be estimated as

$$\hat{H} = -\frac{\hat{\beta}}{2} - 1, \quad (2.11)$$

where $\hat{\beta}$ is the regression slope in the least square linear regression on pairs $(j, \log_2(\hat{\mu}_j))$ from level J_1 to J_2 , $J_1 \leq j \leq J_2$. The estimator \hat{H} follows the asymptotic normal distribution

$$\hat{H} \overset{\text{approx}}{\sim} \mathcal{N}(H, V_1), \quad (2.12)$$

where the asymptotic variance V_1 is a constant number independent of sample size N and

level j ,

$$V_1 = \frac{6f(\alpha, p)}{(\log 2)^2 M^2 c^2 (\alpha, p) q(J_1, J_2)},$$

and

$$q(J_1, J_2) = (J_2 - J_1)(J_2 - J_1 + 1)(J_2 - J_1 + 2). \quad (2.13)$$

The proof of Theorem 2.4.1 is deferred to section 2.7.

To find the optimal α and p by minimizing the asymptotic variance of $\hat{\mu}_j$, we take partial derivatives of $f(\alpha, p)$ with respect to α and p and set them to 0. The optimal $\hat{\alpha}$ and \hat{p} can be obtained by solving

$$\begin{aligned} \frac{\partial f(\alpha, p)}{\partial \alpha} &= -\frac{2p-1}{2p(1-p)}\alpha + \frac{1+p}{2(1-p)} - \frac{3}{2} = 0, \\ \frac{\partial f(\alpha, p)}{\partial p} &= \frac{\alpha(2-\alpha)}{2(1-p)^2} + \frac{\alpha^2(2p-1)}{4p^2(1-p)^2} = 0. \end{aligned} \quad (2.14)$$

Since $\alpha \in [0, 1]$ and $p \in (0, 1/2)$, we get the unique solution $\alpha = 2p \approx 0.6$ and $p = 1 - \sqrt{2}/2 \approx 0.3$. The Hessian matrix of $f(\alpha, p)$ is

$$\begin{aligned} &\begin{bmatrix} \frac{\partial^2 f(\alpha, p)}{\partial \alpha^2} & \frac{\partial^2 f(\alpha, p)}{\partial \alpha \partial p} \\ \frac{\partial^2 f(\alpha, p)}{\partial \alpha \partial p} & \frac{\partial^2 f(\alpha, p)}{\partial p^2} \end{bmatrix} = \\ &\begin{bmatrix} -\frac{2p-1}{2p(1-p)} & \frac{2p^2-2\alpha p^2+\alpha(2p-1)}{2p^2(1-p)^2} \\ \frac{2p^2-2\alpha p^2+\alpha(2p-1)}{2p^2(1-p)^2} & \frac{2p^3\alpha(2-\alpha)+\alpha^2 p(1-p)+\alpha^2(2p-1)^2}{2p^3(1-p)^3} \end{bmatrix}. \end{aligned}$$

Since $-\frac{2p-1}{2p(1-p)} > 0$ and the determinant is $5.66 > 0$ when $\alpha = 2p \approx 0.6$ and $p = 1 - \sqrt{2}/2 \approx 0.3$, the above Hessian matrix is positive definite. Therefore, $\hat{\alpha} = 2 - \sqrt{2}$ and $\hat{p} = 1 - \sqrt{2}/2$ provide the global minima of $f(\alpha, p)$, minimizing also the asymptotic variance of $\hat{\mu}_{j,i}$. In comparing these optimal $\hat{\alpha} \approx 0.6$ and $\hat{p} \approx 0.3$ with $\alpha = 0.6$ and $p = 1/3$ from the Gastwirth estimator, curiously, we find that the optimal general trimean

estimator is very close to the Gastwirth estimator.

2.4.2 General Trimean of the Logarithm of Mid-energy (GTLME) Method

Previously discussed the GTME method calculates the general trimean estimator of the mid-energy first and then takes the logarithm. In this section, we will calculate the general trimean estimator of the logged mid-energies at each level j . The following theorem describes the general trimean of the logarithm of mid-energy, the GTLME method.

Theorem 2.4.2. *Let $\hat{\mu}_j$ be the general trimean estimator based on $\log(D_j)$, which is the set of $M^2/2$ logged mid-energies at level j in a J -level NDWT of a 2-D fBm of size $N \times N$ with Hurst exponent H , and $1 \leq j \leq J$. Then, the asymptotic distribution of $\hat{\mu}_j$ is normal,*

$$\hat{\mu}_j \overset{\text{approx}}{\sim} \mathcal{N} \left(c(\alpha, p) + \log(\lambda_j), \frac{2}{M^2} f(\alpha, p) \right), \quad (2.15)$$

where

$$c(\alpha, p) = \frac{\alpha}{2} \log \left(\log \frac{1}{1-p} \cdot \log \frac{1}{p} \right) + (1-\alpha) \log(\log 2),$$

$$f(\alpha, p) = \frac{\alpha^2}{4g_1(p)} + \frac{\alpha(1-\alpha)}{2g_2(p)} + \frac{(1-\alpha)^2}{(\log 2)^2},$$

$g_1(p)$ and $g_2(p)$ are two functions of p given in the Appendix,

$$\lambda_j = \sigma^2 \cdot 2^{-(2H+2)j},$$

and σ^2 is the variance of wavelet coefficients from level 0. The Hurst exponent can be estimated as

$$\hat{H} = -\frac{1}{2 \log 2} \hat{\beta} - 1, \quad (2.16)$$

where $\hat{\beta}$ is the regression slope in the least square linear regressions on pairs $(j, \hat{\mu}_j)$ from

level J_1 to J_2 , $J_1 \leq j \leq J_2$. The estimator \hat{H} follows the asymptotic normal distribution

$$\hat{H} \overset{\text{approx}}{\sim} \mathcal{N}(H, V_2), \quad (2.17)$$

where the asymptotic variance V_2 is a constant number independent of sample size N and level j ,

$$V_2 = \frac{6f(\alpha, p)}{(\log 2)^2 M^2 q(J_1, J_2)},$$

and $q(J_1, J_2)$ is given in equation (2.13).

The proof of Theorem 2.4.2 is provided in section 2.7. Similarly, as for the GTME, the optimal α and p which minimize the asymptotic variance of $\hat{\mu}_j$ can be obtained by solving

$$\frac{\partial f(\alpha, p)}{\partial \alpha} = 0, \text{ and } \frac{\partial f(\alpha, p)}{\partial p} = 0. \quad (2.18)$$

From the first equation in (2.18) it can be derived that

$$\alpha = \frac{\frac{2}{\log(2)^2} - \frac{1}{2}g_2(p)}{\frac{1}{2}g_1(p) - g_2(p) + \frac{2}{(\log 2)^2}}.$$

The second equation in (2.18) cannot be simplified to a finite form. As an illustration, we plot the $f(\alpha, p)$ with p ranging from 0 to 0.5 and α being a function of p . The plot of α against p is also shown in Figure 2.2. Numerical computation gives $\hat{\alpha} = 0.5965$ and $\hat{p} = 0.24$. These optimal parameters are close to $\alpha = 0.5$ and $p = 0.25$ in the Tukey's trimean estimator, but put some more weight on the median.

2.4.3 Special Cases: Tukey's Trimean and Gastwirth Estimators

The Tukey's trimean of the mid-energy (TTME) method and Gastwirth of the mid-energy (GME) method are described in the following Lemma.

Lemma 2.4.1. *Let $\hat{\mu}_j^T$ and $\hat{\mu}_j^G$ be the Tukey's trimean and Gastwirth estimators based on*

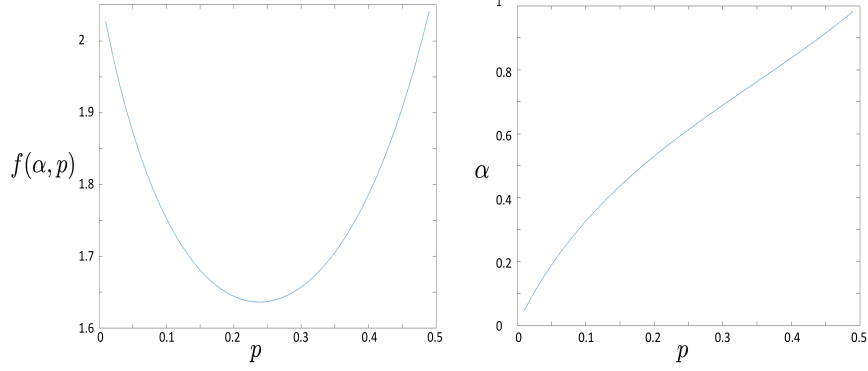


Figure 2.2: Plot of $f(\alpha, p)$ against p on the left; Plot of α against p on the right

D_j defined in (2.9). Then the asymptotic distributions of $\hat{\mu}_j^T$ and $\hat{\mu}_j^G$ are normal:

$$\hat{\mu}_j^T \overset{\text{approx}}{\sim} \mathcal{N} \left(c_1 \lambda_j, \frac{5}{3M^2} \lambda_j^2 \right), \quad (2.19)$$

$$\hat{\mu}_j^G \overset{\text{approx}}{\sim} \mathcal{N} \left(c_2 \lambda_j, \frac{1.67}{M^2} \lambda_j^2 \right), \quad (2.20)$$

where c_1 and c_2 are constant numbers and can be found in the appendix, $\lambda_j = \sigma^2 \cdot 2^{-(2H+2)j}$, and σ^2 is the variance of wavelet coefficients from level 0. The Hurst exponent can be estimated as

$$\hat{H}^T = -\frac{\hat{\beta}^T}{2} - 1, \text{ and } \hat{H}^G = -\frac{\hat{\beta}^G}{2} - 1, \quad (2.21)$$

where $\hat{\beta}^T$ and $\hat{\beta}^G$ are the regression slopes in the least square linear regression on pairs $(j, \log_2(\hat{\mu}_j^T))$ and pairs $(j, \log_2(\hat{\mu}_j^G))$ from level J_1 to J_2 , $J_1 \leq j \leq J_2$. The estimators \hat{H}^T and \hat{H}^G follow the asymptotic normal distributions

$$\hat{H}^T \overset{\text{approx}}{\sim} \mathcal{N}(H, V_1^T), \text{ and } \hat{H}^G \overset{\text{approx}}{\sim} \mathcal{N}(H, V_1^G), \quad (2.22)$$

where the asymptotic variances V_1^T and V_1^G are constant numbers,

$$V_1^T = \frac{5}{(\log 2)^2 M^2 c_1^2 q(J_1, J_2)},$$

$$V_1^G = \frac{5.01}{(\log 2)^2 M^2 c_2^2 q(J_1, J_2)}.$$

The function $q(J_1, J_2)$ is the same as equation (2.13) in Theorem 2.4.1.

The following Lemma describes the Tukey's trimean (TTLME) and Gastwirth (GLME) of the logarithm of mid-energy method.

Lemma 2.4.2. *Let $\hat{\mu}_j^T$ and $\hat{\mu}_j^G$ be the Tukey's trimean estimator and Gastwirth estimator based on $\log(D_j)$ defined in the Theorem 2.4.2. The asymptotic distributions of $\hat{\mu}_j^T$ and $\hat{\mu}_j^G$ are normal,*

$$\hat{\mu}_j^T \overset{\text{approx}}{\sim} \mathcal{N}(- (2H + 2) \log 2j + c_3, V_T), \quad (2.23)$$

$$\hat{\mu}_j^G \overset{\text{approx}}{\sim} \mathcal{N}(- (2H + 2) \log 2j + c_4, V_G), \quad (2.24)$$

where c_3, V_T, c_4 and V_G are constant numbers and can be found in the appendix. The Hurst exponent can be estimated as

$$\hat{H}^T = -\frac{\hat{\beta}^T}{2 \log 2} - 1, \text{ and } \hat{H}^G = -\frac{\hat{\beta}^G}{2 \log 2} - 1, \quad (2.25)$$

where $\hat{\beta}^T$ and $\hat{\beta}^G$ are the regression slopes in the least square linear regression on pairs $(j, \hat{\mu}_j^t)$ and pairs $(j, \hat{\mu}_j^g)$ from level J_1 to J_2 , $J_1 \leq j \leq J_2$. The estimators \hat{H}^T and \hat{H}^G follow the asymptotic normal distributions

$$\hat{H}^T \overset{\text{approx}}{\sim} \mathcal{N}(H, V_2^T), \text{ and } \hat{H}^G \overset{\text{approx}}{\sim} \mathcal{N}(H, V_2^G), \quad (2.26)$$

where the asymptotic variances V_2^T and V_2^G are constant numbers,

$$V_2^T = \frac{3V_T}{(\log 2)^2 q(J_1, J_2)},$$

$$V_2^G = \frac{3V_G}{(\log 2)^2 q(J_1, J_2)}.$$

The function $q(J_1, J_2)$ is provided in equation (2.13).

The proofs of Lemma 2.4.1 and Lemma 2.4.2 are shown in section 2.8. To verify the asymptotic normal distributions of predictors in Lemma 2.4.1 and Lemma 2.4.2, we perform an NDWT of depth 10 on 300 simulated fBm's with $H = 0.3$. We use resulting wavelet coefficients from levels 4 to 10 inclusive to estimate H. Figure 2.3 shows the histograms and theoretical distributions of \hat{H} using TTME, TTLME, GME, and GLME methods, respectively.

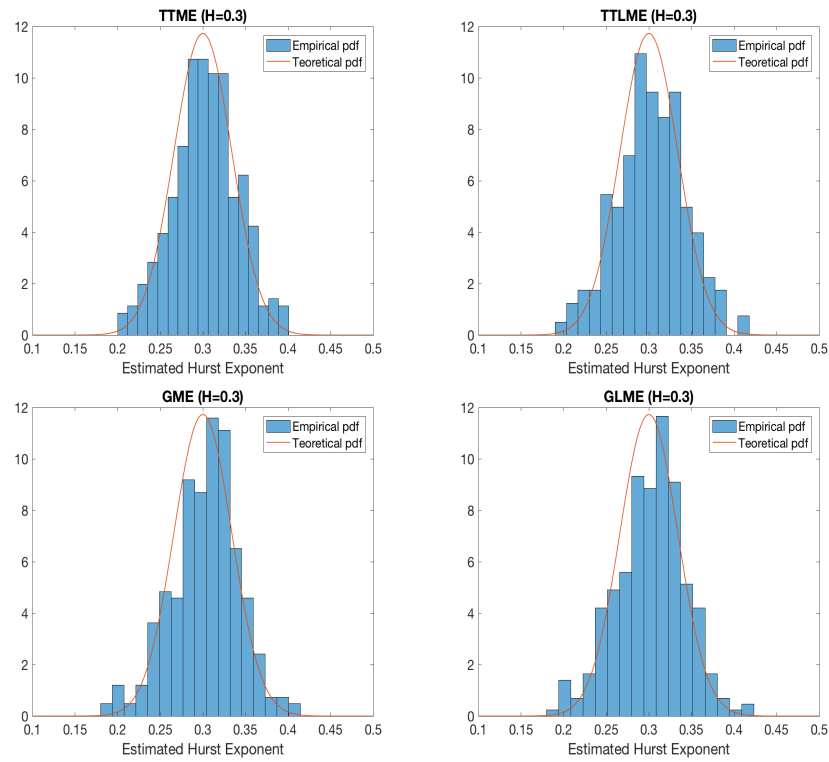


Figure 2.3: Histograms and Theoretical Distributions of \hat{H}

2.5 Simulation

We simulate 2-D fBm of sizes $2^{10} \times 2^{10}$ ($N = 2^{10}$) with Hurst exponent $H = 0.3, 0.5, 0.7, 0.8, 0.9$, respectively. NDWT of depth $J = 10$ using Haar wavelet is performed on the simulated signal to obtain wavelet coefficients. The two-dimensional fBm signals were simulated based on the method of Wood and Chan [57].

The proposed methods (with 6 variations) are applied on the NDWT detail coefficients to estimate Hurst exponent H . Each level diagonal block is divided into 16×16 grids ($M = 16$) for all proposed methods, and we use wavelet coefficients from levels 4 to 10 for the least square linear regression. The estimation performance of the proposed methods are compared to five other existing methods: Veitch & Abry (VA) method, Soltani, Simard, & Boichu (SSB) method, MEDL method, MEDLA method, and Theil-type regression (TT) method. The GTME and GTLME methods are based on the optimal parameters which minimize the variances. Estimation performance is reported in terms of mean, variance, and mean square error (MSE) based on 300 repetitions for each case.

The simulation results are shown in Table 2.1. For each H (corresponding to each row in the table), the smallest variances and MSEs are highlighted in bold. From simulations results, all our 6 variations outperform SSB, MEDL, MEDLA, and TT methods for all H 's regarding to variances and MSEs. Compared with VA method, our methods yield significantly smaller variances and MSEs when $H > 0.5$. When $H = 0.3$, our methods are still comparable to VA. Although the performances of our 6 variations are very similar regarding to variances and MSEs, the TTME method based on Tukey's trimean estimator of the mid-energy has the best performance among all of them. The variances of GTME based on the optimal parameters are very close or equal to those of GME and TTME methods in most cases. Besides, in most cases the optimized GTLME method is superior to other logged mid-energy methods TTLME and GLME with respect to variances; however, such superiority is not significant, since the variances are close to each other.

2.6 Application

In this section, we apply the proposed methodology to classification of digitized mammogram images. The digitized mammograms were obtained from the University of South Florida's Digital Database for Screening Mammography (DDSM) [58]. All cases examined had biopsy results which served as ground truth. Researchers used the HOWTEK

Table 2.1: Simulation Results for $2^{10} \times 2^{10}$ fBm using Haar wavelet (300 Replications)

H	Existing Methods					Proposed Methods					
	VA	SSB	MEDL	MEDLA	TT	TTME	TTLME	GME	GLME	GTME	GTLME
\hat{H}											
0.3	0.3103	0.3055	0.3018	0.3031	0.3054	0.3032	0.3028	0.3032	0.3034	0.3028	0.3030
0.5	0.5220	0.5132	0.5095	0.5102	0.5151	0.5126	0.5111	0.5108	0.5100	0.5118	0.5116
0.7	0.7382	0.7235	0.7175	0.7165	0.7326	0.7193	0.7179	0.7193	0.7184	0.7199	0.7181
0.8	0.8458	0.8261	0.8200	0.8204	0.8398	0.8222	0.8214	0.8208	0.8206	0.8212	0.8221
0.9	0.9593	0.9328	0.9241	0.9274	0.9641	0.9303	0.9282	0.9287	0.9278	0.9295	0.9287
Variances											
0.3	0.0014	0.0016	0.0026	0.0020	0.0017	0.0015	0.0016	0.0016	0.0016	0.0015	0.0016
0.5	0.0020	0.0017	0.0027	0.0018	0.0034	0.0013	0.0016	0.0014	0.0016	0.0014	0.0016
0.7	0.0037	0.0019	0.0030	0.0026	0.0086	0.0018	0.0021	0.0020	0.0021	0.0019	0.0020
0.8	0.0050	0.0021	0.0027	0.0023	0.0095	0.0018	0.0020	0.0020	0.0021	0.0019	0.0020
0.9	0.0073	0.0021	0.0028	0.0022	0.0168	0.0018	0.0019	0.0019	0.0020	0.0018	0.0019
MSEs											
0.3	0.0015	0.0016	0.0026	0.0020	0.0017	0.0015	0.0016	0.0016	0.0016	0.0015	0.0016
0.5	0.0025	0.0019	0.0027	0.0019	0.0037	0.0015	0.0017	0.0016	0.0017	0.0015	0.0017
0.7	0.0052	0.0025	0.0033	0.0028	0.0097	0.0022	0.0024	0.0024	0.0025	0.0023	0.0024
0.8	0.0070	0.0027	0.0031	0.0028	0.0110	0.0023	0.0024	0.0024	0.0025	0.0023	0.0025
0.9	0.0108	0.0032	0.0033	0.0030	0.0208	0.0027	0.0027	0.0027	0.0028	0.0027	0.0027

scanner at the full 43.5-micron per pixel spatial resolution to scan 45 mammograms from patients with normal studies (control group) and 79 from patients with confirmed breast cancer (study group). Figure 2.4 shows an example of mammograms from study group, and it is almost impossible for physicians to distinguish a cancerous mammogram with a non-cancerous mammogram just by eyes. Each subject contains two mammograms from a screening exam, one craniocaudal projection for each side breast. We only keep one projection for each subject, either right side or left side breast image. A sub-image of size 1024×1024 was taken manually from each mammogram.

Our methods were then applied on each sub-image to estimate the Hurst exponent parameter for each subject. To be specific, the NDWT of depth $J = 10$ using Haar wavelet was performed on each sub-image to obtain wavelet coefficients. The proposed methods (with 6 variations) are applied on the NDWT detail coefficients to estimate Hurst exponent H . Each level diagonal block is divided into 16×16 grids ($M = 16$) for all proposed meth-

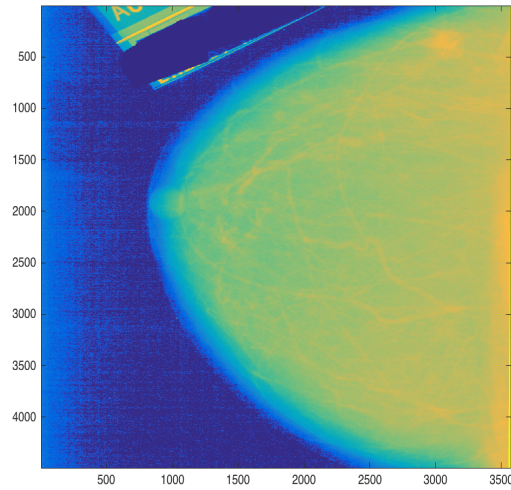


Figure 2.4: An example of mammograms with breast cancer

ods, and we use levels 4 to 10 for the least square linear regression. Veitch & Abry (VA) method, Soltani, Simard, & Boichu (SSB) method, MEDL method, MEDLA method, and Theil-type regression (TT) method were applied, as well, to compare with our methods.

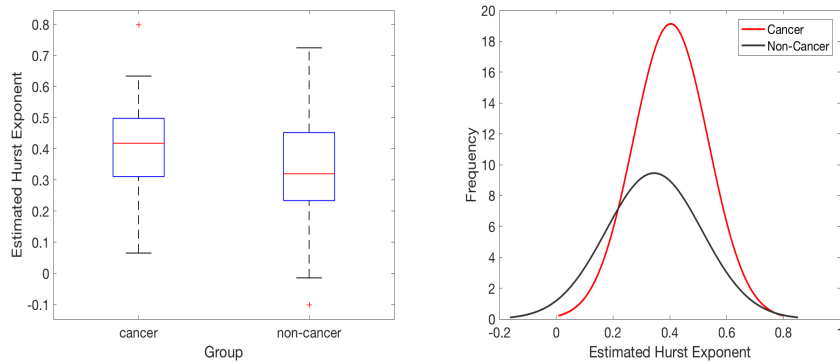


Figure 2.5: Using GME method to estimate Hurst exponent, boxplots in cancer and non-cancer groups on the left; normal density curves fitted in cancer and non-cancer groups on the right

Table 2.2 provides descriptive statistics of the estimated Hurst exponent \hat{H} in each group using our proposed methods and other standard methods to compare with. To visualize the difference in \hat{H} across cancer and non-cancer groups, we present in Figure 2.5 the boxplots of estimated H and fitted normal density curves in two groups based on proposed

GME method. As can be seen, the non-cancer group exhibited a smaller value for \hat{H} in both the mean and median, and the variance of \hat{H} is slightly larger. In fact, images with smaller Hurst exponent tend to be more disordered and unsystematic, therefore healthy individuals tend to have more rough breast tissue images.

Table 2.2: Descriptive Statistics Group Summary

Group	Existing Methods					Proposed Methods					
	VA	SSB	MEDL	MEDLA	TT	TTME	TTLME	GME	GLME	GTME	GTLME
Mean of \hat{H}											
Control	0.3570	0.3457	0.3323	0.3403	0.3716	0.3454	0.3422	0.3444	0.3420	0.3450	0.3430
Study	0.4310	0.4038	0.3935	0.4023	0.4203	0.4061	0.4026	0.4031	0.4019	0.4053	0.4027
Median of \hat{H}											
Control	0.3368	0.3339	0.3326	0.3140	0.3871	0.3248	0.3188	0.3198	0.3240	0.3263	0.3278
Study	0.4286	0.4147	0.3865	0.4165	0.4204	0.4194	0.4211	0.4178	0.4150	0.4168	0.4209
Variance of \hat{H}											
Control	0.0267	0.0270	0.0268	0.0298	0.0305	0.0284	0.0279	0.0285	0.0279	0.0281	0.0277
Study	0.0159	0.0172	0.0198	0.0175	0.0128	0.0169	0.0173	0.0174	0.0175	0.0175	0.0174

For subject i , we generated the data $\{Y_i, H_i\}$, where H_i represents the estimated Hurst exponent, and Y_i is the indicator of the disease status with 1 and 0 signifying cancer and non-cancer, respectively. The subjects were classified using a logistic regression model by treating H_i as the predictor and Y_i as the response. The overall classification accuracy, true positive rate (sensitivity), and true negative rate (specificity) were obtained by using a four fold cross validation. Instead of the constant 0.5 threshold, we used a training-data-determined adaptive threshold, i.e., each time the threshold of the logistic regression was first chosen to maximize Youden index on the training set and then applied to the testing set to classify.

Table 2.3 summarizes the results of the classification for each estimation method. The best classification rate (0.6538) and sensitivity (0.7217) were both achieved using GME estimator, and the best specificity (0.5530) was achieved using TT or TTME estimator (highlighted in bold). In general, the six variations of our robust method performed better as compared to other methods in classification of breast cancers using mammograms.

Real-world images like mammograms may be characterized by non-stationary conditions such as extreme values, causing outlier coefficients in multiresolution levels after NDWT. VA method estimates H by weighted least square regression using the level-wise $\log_2 \left(\overline{d_{j,j}^2} \right)$, and SSB method uses $\overline{\log_2 D_j}$, with D_j defined in (2.9), they are easily affected by those within level outliers, in that they both use mean of derived distributions on level-wise detail coefficients to estimate H . Besides, potential outliers can also occur when logarithmic transform is taken and the magnitude of coefficient is close to zero. Like the VA method, TT method regress the level-wise $\log_2 \left(\overline{d_{j,j}^2} \right)$ against the level indices, but instead of weighted least square regression, they use the Theil-type weighted regression, the weighted average of all slopes between different pairs of regression points, to make it less sensitive to outlier levels. However, it is still not robust to within level outlier coefficients. MEDL and MEDLA use the median of the derived distribution instead of the mean. Although median is outlier-resistant, it can behave unexpectedly as a result of its non-smooth character. To improve, our methods (6 derivations) use the general trimean estimator on non-decimated wavelet detail coefficients of the transformed data, combining the median's emphasis on central values with the quantiles' attention to the extremes. Besides, in the context of our scenario, Theil-type regression is equivalent to least square regression, since the variance of our pair-wise slop is independent of levels and sample size. Those explain why our robust methods performed the best in classification of mammograms.

Table 2.3: Results of classification by logistic regression

	Existing Methods					Proposed Methods					
	VA	SSB	MEDL	MEDLA	TT	TTME	TTLME	GME	GLME	GTME	GTLME
Overall accuracy	0.629	0.597	0.645	0.589	0.547	0.622	0.613	0.654	0.605	0.628	0.645
Sensitivity	0.695	0.659	0.709	0.623	0.543	0.659	0.659	0.722	0.647	0.684	0.708
Specificity	0.511	0.491	0.532	0.534	0.553	0.553	0.534	0.536	0.532	0.528	0.534

2.7 Proof of Theorems 2.4.1 and 2.4.2

In this section, we will provide the technical proofs of Theorem 2.4.1 and Theorem 2.4.2.

Proof of Theorem 2.4.1. A single wavelet coefficient in a non-decimated wavelet transform of a 2-D fBm of size $N \times N$ with Hurst exponent H is normally distributed, with variance depending on its level j . The four coefficients in each set

$$\left\{ d_{j,j;(k_{i1},k_{i2})}, d_{j,j;(k_{i1},k_{i2}+\frac{N}{2})}, d_{j,j;(k_{i1}+\frac{N}{2},k_{i2})}, d_{j,j;(k_{i1}+\frac{N}{2},k_{i2}+\frac{N}{2})} \right\}$$

are assumed to be independent and follow the same normal distribution.

$$d_{j,j;(k_{i1},k_{i2})}, d_{j,j;(k_{i1},k_{i2}+\frac{N}{2})}, d_{j,j;(k_{i1}+\frac{N}{2},k_{i2})}, d_{j,j;(k_{i1}+\frac{N}{2},k_{i2}+\frac{N}{2})} \sim \mathcal{N}\left(0, 2^{-(2H+2)j} \sigma^2\right).$$

Then the mid-energies in D_j defined in (2.9) and (2.8) can be readily shown to have exponential distribution with scale parameter $\lambda_j = \sigma^2 \cdot 2^{-(2H+2)j}$. Therefore at each detail level j , the mid-energies in D_j are i.i.d. $\mathcal{Exp}(\lambda_j^{-1})$, and when applying general trimean estimator $\hat{\mu}_j$ on D_j , following the derivation in Section 2.3, we have

$$\boldsymbol{\xi} = \left[\log\left(\frac{1}{1-p}\right) \lambda_j \quad \log(2) \lambda_j \quad \log\left(\frac{1}{p}\right) \lambda_j \right]^T,$$

and

$$\Sigma = \begin{bmatrix} \frac{p}{(1-p)} \lambda_j^2 & \frac{p}{(1-p)} \lambda_j^2 & \frac{p}{(1-p)} \lambda_j^2 \\ \frac{p}{(1-p)} \lambda_j^2 & \lambda_j^2 & \lambda_j^2 \\ \frac{p}{(1-p)} \lambda_j^2 & \lambda_j^2 & \frac{1-p}{p} \lambda_j^2 \end{bmatrix}_{3 \times 3},$$

therefore, the asymptotic distribution of $\hat{\mu}_{j,i}$ is normal with mean

$$\begin{aligned} \mathbb{E}(\hat{\mu}_{j,i}) &= A \cdot \boldsymbol{x} \\ &= \left(\frac{\alpha}{2} \log\left(\frac{1}{p(1-p)}\right) + (1-\alpha) \log 2 \right) \lambda_j \\ &\triangleq c(\alpha, p) \lambda_j, \end{aligned}$$

and variance

$$\begin{aligned}\text{Var}(\hat{\mu}_{j,i}) &= \frac{2}{M^2} A \Sigma A^T \\ &= \frac{2}{M^2} \left(\frac{\alpha(1-2p)(\alpha-4p)}{4p(1-p)} + 1 \right) \lambda_j^2 \\ &\triangleq \frac{2}{M^2} f(\alpha, p) \lambda_j^2.\end{aligned}$$

Since the Hurst exponent can be estimated as

$$\hat{H} = -\frac{\hat{\beta}}{2} - 1, \quad (2.27)$$

where $\hat{\beta}$ is the regression slope in the least square linear regression on pairs $(j, \log_2(\hat{\mu}_j))$ from level J_1 to J_2 , $J_1 \leq j \leq J_2$. It can be easily derived that $\hat{\beta}$ is a linear combination of $\log_2(\hat{\mu}_j)$,

$$\hat{\beta} = \sum_{j=J_1}^{J_2} a_j \log_2(\hat{\mu}_j), \quad a_j = \frac{j - (J_1 + J_2)/2}{\sum_{j=J_1}^{J_2} (j - (J_1 + J_2)/2)^2}.$$

We can check that $\sum_{j=J_1}^{J_2} a_j = 0$ and $\sum_{j=J_1}^{J_2} a_j j = 1$. Also, if $X \sim \mathcal{N}(\mu, \sigma^2)$, the approximate expectation and variance of $g(X)$ are

$$\mathbb{E}(g(X)) = g(\mu) + \frac{g''(\mu)\sigma^2}{2}, \quad \text{and} \quad \text{Var}(g(X)) = (g'(\mu))^2 \sigma^2,$$

based on which we calculate

$$\mathbb{E}(\log_2(\hat{\mu}_j)) = -(2H+2)j + \text{Constant}, \quad \text{and} \quad \text{Var}(\log_2(\hat{\mu}_j)) = \frac{\frac{2}{M^2} f(\alpha, p)}{(\log 2)^2 c^2(\alpha, p)}.$$

Therefore

$$\mathbb{E}(\hat{\beta}) = \sum_{j=J_1}^{J_2} a_j \mathbb{E}(\log_2(\hat{\mu}_j)) = -(2H+2), \quad \text{and} \quad \text{Var}(\hat{\beta}) = \sum_{j=J_1}^{J_2} a_j^2 \text{Var}(\log_2(\hat{\mu}_j)) := 4V1,$$

and

$$\mathbb{E}(\hat{H}) = H, \quad \text{and} \quad \text{Var}(\hat{H}) = V1, \quad (2.28)$$

where the asymptotic variance V_1 is a constant number independent of sample size N and level j ,

$$V_1 = \frac{6f(\alpha, p)}{(\log 2)^2 M^2 c^2(\alpha, p) q(J_1, J_2)},$$

and

$$q(J_1, J_2) = (J_2 - J_1)(J_2 - J_1 + 1)(J_2 - J_1 + 2).$$

□

Proof of Theorem 2.4.2. We have stated that each mid-energy in D_j follows $\mathcal{Exp}(\lambda_j^{-1})$ with scale parameter $\lambda_j = \sigma^2 \cdot 2^{-(2H+2)j}$. If we denote the k th element in $\log(D_j)$ as $y_{j,k}$ for $k = 1, \dots, \frac{M^2}{2}$ and $j = 1, \dots, J$, the pdf and cdf of $y_{j,k}$ are

$$f(y_{j,k}) = \lambda_j^{-1} e^{-\lambda_j^{-1} e^{y_{j,k}}} e^{y_{j,k}},$$

and

$$F(y_{j,k}) = 1 - e^{-\lambda_j^{-1} e^{y_{j,k}}}.$$

The p -quantile can be obtained by solving $F(y_p) = 1 - e^{-\lambda_j^{-1} e^{y_p}} = p$, and $y_p = \log(-\lambda_j \log(1-p))$.

Then it can be shown that $f(y_p) = -(1-p) \log(1-p)$. When applying the general trimean estimator $\hat{\mu}_j$ on $\log(D_j)$, following the derivation in Section 2.3, we get

$$\boldsymbol{\xi} = \begin{bmatrix} \log\left(\log\left(\frac{1}{1-p}\right)\right) + \log(\lambda_j) \\ \log(\log 2) + \log(\lambda_j) \\ \log\left(\log\left(\frac{1}{p}\right)\right) + \log(\lambda_j) \end{bmatrix},$$

and

$$\boldsymbol{\Sigma} = \begin{bmatrix} \frac{p}{(1-p)(\log(1-p))^2} & \frac{p}{(1-p)\log(1-p)\log\left(\frac{1}{2}\right)} & \frac{p}{(1-p)\log(1-p)\log p} \\ \frac{p}{(1-p)\log(1-p)\log\left(\frac{1}{2}\right)} & \frac{1}{(\log 2)^2} & \frac{1}{\log\left(\frac{1}{2}\right)\log p} \\ \frac{p}{(1-p)\log(1-p)\log p} & \frac{1}{\log\left(\frac{1}{2}\right)\log p} & \frac{1-p}{p(\log p)^2} \end{bmatrix},$$

thus, the asymptotic distribution of $\hat{\mu}_{j,i}$ is normal with mean

$$\begin{aligned}\mathbb{E}(\hat{\mu}_{j,i}) &= A \cdot \boldsymbol{\xi} \\ &= \frac{\alpha}{2} \log \left(\log \frac{1}{1-p} \cdot \log \frac{1}{p} \right) + (1-\alpha) \log(\log 2) + \log(\lambda_j) \\ &\triangleq c(\alpha, p) + \log(\lambda_j),\end{aligned}$$

and variance

$$\begin{aligned}\text{Var}(\hat{\mu}_{j,i}) &= \frac{2}{M^2} A \Sigma A^T \\ &= \frac{2}{M^2} \left(\frac{\alpha^2}{4} g_1(p) + \frac{\alpha(1-\alpha)}{2} g_2(p) + \frac{(1-\alpha)^2}{(\log 2)^2} \right) \\ &\triangleq \frac{2}{M^2} f(\alpha, p),\end{aligned}$$

where

$$\begin{aligned}g_1(p) &= \frac{p}{(1-p)(\log(1-p))^2} + \\ &\quad \frac{1-p}{p(\log p)^2} + \frac{2p}{(1-p)\log(1-p)\log p},\end{aligned}$$

and

$$g_2(p) = \frac{2p}{(1-p)\log(1-p)\log \frac{1}{2}} + \frac{2}{\log \frac{1}{2} \log p}.$$

Since the Hurst exponent can be estimated as

$$\hat{H} = -\frac{1}{2 \log 2} \hat{\beta} - 1, \quad (2.29)$$

where $\hat{\beta}$ is the regression slope in the least square linear regressions on pairs $(j, \hat{\mu}_j)$ from level J_1 to J_2 , $J_1 \leq j \leq J_2$. It can be easily derived that $\hat{\beta}$ is a linear combination of $\hat{\mu}_j$,

$$\hat{\beta} = \sum_{j=J_1}^{J_2} a_j \hat{\mu}_j, \quad a_j = \frac{j - (J_1 + J_2)/2}{\sum_{j=J_1}^{J_2} (j - (J_1 + J_2)/2)^2}.$$

Again, we can check that $\sum_{j=J_1}^{J_2} a_j = 0$ and $\sum_{j=J_1}^{J_2} a_j j = 1$. Therefore

$$\mathbb{E}(\hat{\beta}) = \sum_{j=J_1}^{J_2} a_j \mathbb{E}(\hat{\mu}_{j,i}) = -(2H+2) \log 2, \text{ and } \text{Var}(\hat{\beta}) = \sum_{j=J_1}^{J_2} a_j^2 \text{Var}(\hat{\mu}_{j,i}) := 4(\log 2)^2 V_2,$$

and

$$\mathbb{E}(\hat{H}) = H, \text{ and } \text{Var}(\hat{H}) = V_2, \quad (2.30)$$

where the asymptotic variance V_2 is a constant number independent of simple size N and level j ,

$$V_2 = \frac{6f(\alpha, p)}{(\log 2)^2 M^2 q(J_1, J_2)},$$

and $q(J_1, J_2)$ is given in equation (2.13). □

2.8 Proof of Lemmas 2.4.1 and 2.4.2

The proofs of Lemma 2.4.1 and Lemma 2.4.2 will be discussed in this section.

Proof of Lemma 2.4.1. When applying Tukey's trimean estimator $\hat{\mu}_j^T$ on D_j , following the derivation in Section 2.3.1, we have

$$\boldsymbol{\xi}_T = \begin{bmatrix} \log\left(\frac{4}{3}\right) \lambda_j \\ \log(2) \lambda_j \\ \log(4) \lambda_j \end{bmatrix},$$

and

$$\Sigma_T = \begin{bmatrix} \frac{1}{3} \lambda_j^2 & \frac{1}{3} \lambda_j^2 & \frac{1}{3} \lambda_j^2 \\ \frac{1}{3} \lambda_j^2 & \lambda_j^2 & \lambda_j^2 \\ \frac{1}{3} \lambda_j^2 & \lambda_j^2 & \frac{1}{3} \lambda_j^2 \end{bmatrix}_{3 \times 3},$$

therefore, the asymptotic distribution of $\hat{\mu}_j^T$ is normal with mean

$$\mathbb{E}(\hat{\mu}_j^T) = A_T \cdot \boldsymbol{\xi}_T = \frac{1}{4} \log\left(\frac{64}{3}\right) \lambda_j \triangleq c_1 \lambda_j,$$

and variance

$$\text{Var}(\hat{\mu}_{j,i}^T) = \frac{2}{M^2} A_T \Sigma_T A_T^T = \frac{5}{3M^2} \lambda_j^2.$$

When applying Gastwirth estimator $\hat{\mu}_j^G$ on D_j , following the derivation in Section 2.3.2, we have

$$\boldsymbol{\xi}_G = \begin{bmatrix} \log\left(\frac{3}{2}\right) \lambda_j \\ \log(2) \lambda_j \\ \log(3) \lambda_j \end{bmatrix},$$

and

$$\Sigma_G = \begin{bmatrix} \frac{1}{2} \lambda_j^2 & \frac{1}{2} \lambda_j^2 & \frac{1}{2} \lambda_j^2 \\ \frac{1}{2} \lambda_j^2 & \lambda_j^2 & \lambda_j^2 \\ \frac{1}{2} \lambda_j^2 & \lambda_j^2 & \frac{1}{2} \lambda_j^2 \end{bmatrix},$$

therefore, the asymptotic distribution of $\hat{\mu}_j^G$ is normal with mean

$$\begin{aligned} \mathbb{E}(\hat{\mu}_{j,i}^G) &= A_G \cdot \boldsymbol{\xi}_G \\ &= \left(0.3 \times \log\left(\frac{9}{2}\right) + 0.4 \times \log(2)\right) \lambda_j \\ &\triangleq c_2 \lambda_j, \end{aligned}$$

and variance

$$\text{Var}(\hat{\mu}_{j,i}^G) = \frac{2}{M^2} A_G \Sigma_G A_G^T = \frac{1.67}{M^2} \lambda_j^2.$$

Based on equation (2.28), we have

$$\hat{H}^T \overset{\text{approx}}{\sim} \mathcal{N}(H, V_1^T), \text{ and } \hat{H}^G \overset{\text{approx}}{\sim} \mathcal{N}(H, V_1^G), \quad (2.31)$$

where the asymptotic variances V_1^T and V_1^G are constant numbers,

$$V_1^T = \frac{5}{(\log 2)^2 M^2 c_1^2 q(J_1, J_2)},$$

$$V_1^G = \frac{5.01}{(\log 2)^2 M^2 c_2^2 q(J_1, J_2)}.$$

The function $q(J_1, J_2)$ is the same as equation (2.13) in Theorem 2.4.1. \square

Proof of Lemma 2.4.2. When applying Tukey's trimean estimator $\hat{\mu}_j^T$ on $\log(D_j)$, following the derivation in Section 2.3.1, we have

$$\boldsymbol{\xi}_T = \begin{bmatrix} \log(\log(\frac{4}{3})) + \log(\lambda_j) \\ \log(\log 2) + \log(\lambda_j) \\ \log(\log 4) + \log(\lambda_j) \end{bmatrix},$$

and

$$\Sigma_T = \begin{bmatrix} \frac{1}{3(\log(\frac{3}{4}))^2} & \frac{1}{3\log(\frac{3}{4})\log(\frac{1}{2})} & \frac{1}{3\log(\frac{3}{4})\log(\frac{1}{4})} \\ \frac{1}{3\log(\frac{3}{4})\log(\frac{1}{2})} & \frac{1}{(\log 2)^2} & \frac{1}{\log(\frac{1}{2})\log(\frac{1}{4})} \\ \frac{1}{3\log(\frac{3}{4})\log(\frac{1}{4})} & \frac{1}{\log(\frac{1}{2})\log(\frac{1}{4})} & \frac{3}{(\log 4)^2} \end{bmatrix},$$

therefore, the asymptotic distribution of $\hat{\mu}_j^T$ is normal with mean

$$\begin{aligned} \mathbb{E}(\hat{\mu}_{j,i}^T) &= A_T \cdot \boldsymbol{\xi}_T \\ &= -(2H + 2) \log 2 \cdot j + \log \sigma^2 + \\ &\quad \frac{1}{4} \log \left(\log \left(\frac{4}{3} \right) \cdot \log 4 \right) + \frac{1}{2} \log(\log 2) \\ &\triangleq -(2H + 2) \log 2 \cdot j + c_3 \end{aligned}$$

and variance

$$\begin{aligned}
\text{Var}(\hat{\mu}_{j,i}^T) &= \frac{2}{M^2} A_T \Sigma_T A_T^T \\
&= \frac{2}{M^2} \left(\frac{1}{48 (\log \frac{3}{4})^2} + \frac{1}{12 \log \frac{3}{4} \log \frac{1}{2}} + \frac{1}{24 \log \frac{3}{4} \log \frac{1}{4}} + \right. \\
&\quad \left. \frac{1}{4 (\log 2)^2} + \frac{1}{4 \log \frac{1}{2} \log \frac{1}{4}} + \frac{3}{16 (\log \frac{1}{4})^2} \right) \\
&\triangleq V_T.
\end{aligned}$$

When applying Gastwirth estimator $\hat{\mu}_j^G$ on $\log(D_{j,i})$, following the derivation in Section 2.3.2, we have

$$\boldsymbol{\xi}_G = \begin{bmatrix} \log(\log(\frac{3}{2})) + \log(\lambda_j) \\ \log(\log 2) + \log(\lambda_j) \\ \log(\log 3) + \log(\lambda_j) \end{bmatrix},$$

and

$$\Sigma_G = \begin{bmatrix} \frac{1}{2(\log \frac{2}{3})^2} & \frac{1}{2 \log(\frac{2}{3}) \log(\frac{1}{2})} & \frac{1}{2 \log(\frac{1}{3}) \log(\frac{2}{3})} \\ \frac{1}{2 \log(\frac{2}{3}) \log(\frac{1}{2})} & \frac{1}{(\log 2)^2} & \frac{1}{\log(\frac{1}{2}) \log(\frac{1}{3})} \\ \frac{1}{2 \log(\frac{1}{3}) \log(\frac{2}{3})} & \frac{1}{\log(\frac{1}{2}) \log(\frac{1}{3})} & \frac{2}{(\log 3)^2} \end{bmatrix},$$

therefore, the asymptotic distribution of $\hat{\mu}_j^G$ is normal with mean

$$\begin{aligned}
\mathbb{E}(\hat{\mu}_{j,i}^G) &= A_g \cdot \boldsymbol{\xi}_G \\
&= -(2H + 2) \log 2 \cdot j + \log \sigma^2 + \\
&\quad 0.3 \times \log \left(\log \left(\frac{3}{2} \right) \cdot \log 3 \right) + 0.4 \times \log(\log 2) \\
&\triangleq -(2H + 2) \log 2 \cdot j + c_4
\end{aligned}$$

and variance

$$\begin{aligned}
\text{Var}(\hat{\mu}_{j,i}^G) &= \frac{2}{M^2} A_G \Sigma_G A_G^T \\
&= \frac{2}{M^2} \left(\frac{0.09}{2 \left(\log \frac{2}{3}\right)^2} + \frac{0.12}{\log \frac{2}{3} \log \frac{1}{2}} + \frac{0.09}{\log \frac{1}{3} \log \frac{2}{3}} + \right. \\
&\quad \left. \frac{0.16}{\left(\log \frac{1}{2}\right)^2} + \frac{0.24}{\log \frac{1}{2} \log \frac{1}{3}} + \frac{0.18}{\left(\log \frac{1}{3}\right)^2} \right) \\
&\triangleq V_G.
\end{aligned}$$

Based on quation (2.30), we can easily derive

$$\hat{H}^T \overset{\text{approx}}{\sim} \mathcal{N}(H, V_2^T), \text{ and } \hat{H}^G \overset{\text{approx}}{\sim} \mathcal{N}(H, V_2^G), \quad (2.32)$$

where the asymptotic variances V_2^T and V_2^G are constant numbers,

$$V_2^T = \frac{3V_T}{(\log 2)^2 q(J_1, J_2)},$$

$$V_2^G = \frac{3V_G}{(\log 2)^2 q(J_1, J_2)}.$$

The function $q(J_1, J_2)$ is provided in equation (2.13). □

2.9 Conclusions

The work in this chapter has been published in the book *New Frontiers of Biostatistics and Bioinformatics*, see [59]. In this study, we proposed methodologies and derived 6 variations to improve the robustness of estimation of Hurst exponent H in two-dimensional setting. Non-decimated wavelet transforms (NDWT) are utilized for its redundancy and time-invariance. Instead of using mean or median of the derived distribution on level-wise wavelet coefficients, we defined the general trimean estimators that combine the median's emphasis on center values with the quantiles' attention to the extremes and used them on

the level-wise derived distributions to estimate H .

The proposed variations were: 1) Tukey's trimean of the mid-energy (TTME) method; 2) Tukey's trimean of the logged mid-energy (TTLME) method; 3) Gastwirth of the mid-energy (GME) method; 4) Gastwirth of the logged mid-energy (GLME) method; 4) general trimean of the mid-energy (GTME) method; 6) general trimean of the logarithm of mid-energy (GTLME) method. The GTME and GTLME methods are based on the derived optimal parameters in general trimean estimators to minimize the asymptotic variances. Tukey's trimean and Gastwirth estimators are two special cases following the general trimean estimators' framework. These estimators are applied on both mid-energy (as defined by Soltani et al.,[20]) and logarithm of the mid-energy at each NDWT level detail coefficient diagonal block. The estimation performance of the proposed methods are compared to five other existing methods: Veitch & Abry (VA) method, Soltani, Simard, & Boichu (SSB) method, MEDL method, MEDLA method, and Theil-type regression (TT) method.

Simulation results indicate all our 6 variations outperform SSB, MEDL , MEDLA, and TT methods for all H 's regarding to variances and MSEs. Compared with VA method, our methods yield significantly smaller variances and MSEs when $H > 0.5$. When $H = 0.3$, our methods are still comparable to VA. Although the performances of our 6 variations are very similar regarding to variances and MSEs, the TTME method based on Tukey's trimean estimator of the mid-energy has the best performance among all of them.

The proposed methods have been applied to digitized mammograms to classify patients with and without breast cancer. Our methods helped to differentiate individuals based on the estimated Hurst parameters \hat{H} . Higher values for \hat{H} have been found in cancer group, and individuals with breast cancer have smoother breast tissue images. This increase of regularity with increase of the degree of pathology is common for many other biometric signals: EEG, EKG, high frequency protein mass-spectra, high resolution medical images of tissue, to list a few.

CHAPTER 3

A PERSONALIZED THRESHOLD METHOD VIA BOOSTING FOR SEPSIS SCREENING

3.1 Introduction

Sepsis is life-threatening organ dysfunction caused by a dysregulated host response to infection, and it is one of the biggest patient safety risks in healthcare settings [60, 61, 62, 63]. Nearly half of patients who die in hospitals are septic, and the natural mortality rate for sepsis is between 25% and 50% [60]. Recently, screening as a decision support mechanism for early detection of sepsis has been widely advocated, since early identification of sepsis and the timely medical intervention could significantly decrease sepsis-related mortality and are cost-effective [64, 65, 66, 67, 68, 69, 70, 71]. In 2016, a task force committee [72] recommended patient screening for sepsis by a the scoring criterion termed quick Sequential (sepsis related) Organ Failure Assessment (qSOFA), and conducting laboratory tests to further assess sepsis if needed. The qSOFA score is essentially a constant thresholding technique regardless of patients' baseline information, and it uses three easy-to-measure biomarkers: systolic blood pressure, respiratory rate, and Glasgow Coma Scale score (a score for mental status).

The main statistical approach in qSOFA criteria is to dichotomize each biomarker X , and raise a screening warning if $X \geq c$ (or $X < c$) for some constant threshold c . The constant thresholds in qSOFA are determined based on physiology, clinical experience, and statistical analysis such as Receiver Operating Characteristic (ROC) curve [73, 74]. Indeed, there has been extensive research in the statistical literature to find the optimal threshold. The five most popular methods are: 1) the minP approach [75], 2) the Youden index [76], 3) the closest-to-(0, 1) criteria [77], 4) the concordance probability method [78], and 5) the

index of union [79].

Unfortunately, there are some drawbacks to the qSOFA score. It demonstrates very low specificity in ICU populations [80, 81]. At emergency department (ED) triage, qSOFA scores poorly in identifying sepsis, and is likewise poor in both pre-hospital and ED triage for predicting intensive care unit (ICU) stays of three or more days [82, 83, 84, 85]. One possible explanation is that the current qSOFA uses thresholds that are constant for all patients, regardless of patients' demographic differences, leading to the low prediction accuracy.

In this chapter, we propose to improve qSOFA by self-learning suitable personalized thresholds for different sepsis patients. Our key idea is to use boosting [86, 87, 88], a popular machine learning technique, to obtain the personalized thresholds based on patient-related information. As a result, our method can better identify sepsis patients who could benefit from time critical interventions. In addition, our method is similar to original qSOFA, and it is easy to interpret and can be used for real-time monitoring.

Besides screening, another concept in sepsis is assessing. Machine learning methods have been developed for sepsis assessing, see [89, 90, 91, 92, 93, 94, 95, 96, 97]. For instance, Giannini et al. [98] applied the Random Forest classifier to predict patients at elevated risk of developing severe sepsis and/or septic shock by using Electronic Health Record (EHR) data; Shimabukuro et al. [99] studied a machine learning-based severe sepsis prediction system for reductions in average length of stay and in-hospital mortality rate; Nemati et al. [100] developed and validated an Artificial Intelligence Sepsis Expert algorithm for early prediction of sepsis. However, these black-box machine learning methods are not as appropriate a method for screening due to lack of interpretation and implementation difficulties for real-time monitoring. Here, we essentially develop a knowledge based machine learning method that is suitable for sepsis screening.

The organization of the remaining sections are as follows. Section 3.2 introduces the necessary background knowledge, including the details of sepsis and qSOFA score, existing

methods to obtain the optimal constant threshold and boosting machine learning method. Section 3.3 describes the proposed boosting based method to estimate the personalized threshold, in which the exponential loss and gradient descent algorithm were applied. Section 3.4 explores the data we use to demonstrate the method. Section 3.5 presents the application sepsis screening and compares the performance of the proposed method to the original qSOFA criteria, other standards used, constant threshold approaches, and machine learning methods. The proof of the proposition is provided in Section 3.6. We provide conclusions in Section 3.7.

3.2 Background

In this section, we introduce the background information of sepsis and qSOFA score, the existing approaches to obtain constant thresholds, and the boosting method in three subsections, respectively.

3.2.1 Sepsis and qSOFA

Sepsis is not a specific illness but rather a syndrome producing a similar innate immune response as infection. It is differentiated from infection in a dysregulated host response and the presence of organ dysfunction [72]. Considerable changes have been made on how it is defined based on a better understanding of the underlying pathobiology.

The initial definition of sepsis was developed in 1991, and it was assessed by the Systemic Inflammatory Response Syndrome (SIRS) to infection, which includes a patient's temperature, heart rate, respiratory rate, and white blood cell count [101, 102]. In this definition, sepsis is defined as infection with at least 2 of the 4 SIRS criteria satisfied. However, SIRS criteria do not necessarily indicate organ dysfunction or failure. In 2016, therefore, a task force developed the current definition of sepsis, and recommended two stages of monitoring sepsis: screening and assessing [72]. Instead of the widely used SIRS criteria, the 2016 definitions recommended clinically characterizing a sepsis patient by the Sequential

Organ Failure Assessment (SOFA) score, which is used to identify organ dysfunction. Moreover, the 2016 task force committee recommended a new score criterion termed quickSOFA (qSOFA), a bedside screening measurement identifying patients with suspected infection.

The qSOFA score relies on three important variables for sepsis: respiratory rate, systolic blood pressure, and Glasgow Coma Scale (GCS) score, and the current qSOFA guideline is to check whether these three observed variables are normal or abnormal as compared to their respective constant critical threshold values. A screening alarm is raised if two out of three variables are abnormal. To be specific, in the qSOFA criterion, one is monitoring 1) whether the alteration in mental status (GCS score is less than 15), 2) whether systolic blood pressure is ≤ 100 mm Hg, or 3) whether respiratory rate is ≥ 22 breaths/min. These constant critical thresholds are derived from the two group comparisons of sepsis versus non-sepsis patients, and do not take into account the patient’s baseline demographic characteristics such as age, sex, admission location, admission type, ethnicity, insurance, and marital status.

3.2.2 Existing Methods to Obtain the Constant Threshold

The most common approaches to determine the threshold for a biomarker are via Receiver Operating Characteristic (ROC) curve analysis. The ROC curve is a mapping of the sensitivity versus 1-specificity for all possible thresholds. Thresholds dichotomize the test values, and therefore provide the diagnosis whether the subject is diseased or not. A threshold is referred to as optimal when it classifies most of the individuals correctly. Let X be a continuous biomarker that is assumed to be predictive of disease Y ($Y = 1$ for diseased and $Y = -1$ for not diseased). For any possible cut-point c of X , the data can be formed into a 2×2 table,

	$Y = 1$	$Y = -1$
$X \geq c$	$s = s(c)$	$r = r(c)$
$X < c$	$u = u(c)$	$v = v(c)$

Sensitivity(Se) and specificity(Sp) are defined as follows,

$$\text{Se}(c) = P(X \geq c|Y = 1) = \frac{s}{s + u}, \quad \text{Sp}(c) = P(X < c|Y = -1) = \frac{v}{r + v}.$$

Various criteria for the optimal threshold value c have been proposed [75, 76, 77, 78, 79] based on above 2-by-2 table. We briefly describe them here.

minP Approach [75]: The optimal threshold c is selected so as to maximize the standard chi square statistic

$$\chi_1^2(c) = \frac{(s + r + u + v)(sv - ur)^2}{(s + r)(u + v)(s + u)(r + v)}.$$

Youden Index [76]: Youden index, $\text{Se}(c) + \text{Sp}(c) - 1$, combines sensitivity and specificity into a single measure. Maximizing the Youden index, one is able to find the cut point that has the largest value in the sum of sensitivity and specificity or in the difference between sensitivity and the false positive rate.

Closest-to-(0,1) Criterion [77]: Ideal point (0, 1) on ROC curve represents zero false positives and perfect sensitivity. The “optimal” threshold is defined as the point on the ROC curve closest to (0,1), i.e., find c to minimize $\sqrt{(1 - \text{Se}(c))^2 + (1 - \text{Sp}(c))^2}$.

Concordance Probability [78]: The concordance probability method defines the optimal threshold as the point c maximizing the product of sensitivity and specificity $\text{Se}(c)\text{Sp}(c)$. The concordance probability of dichotomized measure at threshold c , can be expressed as the area of a rectangle associated with the ROC curve. The threshold c maximizing $\text{Se}(c)\text{Sp}(c)$ actually maximizes the area of the rectangle.

Index of Union [79]. The method defines the optimal cut-point value c as the point minimizing the summation of absolute values of the differences between AUC and sensitivity and AUC and specificity, $|\text{Se}(c) - \text{AUC}| + |\text{Sp}(c) - \text{AUC}|$.

All these methods are based on the biomarker itself and the outcome variable alone, and do not take into account any individual-specific information. In real clinical practice, patient-related variables are usually available besides the biomarker, such as the easily accessible demographics. How to utilize the extra information to determine a personalized threshold for diagnostic or screening purposes has rarely been considered.

3.2.3 Boosting Method

Boosting is one of the new machine learning techniques for classification. One well-known boosting algorithm is called "AdaBoost.M1." proposed by Freund and Schapire (1997) [103]. The three key ideas in boosting method include: 1) Using a sign function as the classification rule for binary outcome; 2) Optimizing a novel exponential loss function that approximates the error rate; 3) Assigning different weights on observations based on their influences.

Given a binary outcome $Y \in \{-1, 1\}$ and a vector of predictor variables \mathbf{X} , a classifier $f(\mathbf{X})$ is a sign function of some statistics, producing a prediction taking one of the two values $\{-1, 1\}$,

$$f(\mathbf{X}) = \text{sign}(g(\mathbf{X})), \text{ where } g(\mathbf{X}) = \sum_{m=1}^M \alpha_m f_m(\mathbf{X}), \quad (3.1)$$

and $f_m(x)$, $m = 1, 2, \dots, M$, are weak learners and $g(\mathbf{X})$ combines them together to make a better classification. One of the key ideas in boosting method is to replace the discrete 0-1 classification error by the exponential loss. To be more concrete, in classification with a -1/1 response, the error rate of training data is

$$\bar{\text{err}} = \frac{1}{N} \sum_{i=1}^N I(y_i \neq \text{sign}(g(\mathbf{X}_i))) = \frac{1}{N} \sum_{i=1}^N I(y_i \cdot \text{sign}(g(\mathbf{X}_i)) < 0). \quad (3.2)$$

Observations with positive $y_i \cdot \text{sign}(g(\mathbf{X}_i)) > 0$ are correctly classified, while those with negative $y_i \cdot \text{sign}(g(\mathbf{X}_i)) < 0$ are misclassified. The error rate is not a smoothing function,

and therefore it is not a favorable loss function for classification. Boosting method replaces the term $I(y_i \cdot \text{sign}(g(\mathbf{X}) < 0))$ by a novel exponential loss function

$$L(y, g(\mathbf{X})) = \exp(-y \cdot g(\mathbf{X})). \quad (3.3)$$

It has been proven that the exponential loss is a monotone continuous approximations to misclassification loss. In the training process, the exponential criterion concentrates much more influence on observations with negative $y \cdot g(\mathbf{X})$.

Boosting method essentially is a gradient descent algorithm that finds the parameters α'_m s to minimize the exponential loss function on the training data. It turns out that it can also be thought of as applying different weights to training observations (\mathbf{X}_i, y_i) , $i = 1, 2, \dots, N$. Those observations that were difficult to predict (misclassified by weak learners) will have larger weights, whereas less weights will be assigned to those easy-to-predict observations that were classified correctly by weak learners.

3.3 Our Proposed Personalized Threshold Method

In this section, we propose a boosting alike method to obtain the personalized threshold. Suppose we have a training data of the form $(Y_i, X_i, u_{i1}, u_{i2}, \dots, u_{iq})$, for $i = 1, \dots, N$, where $Y_i \in \{-1, 1\}$ is the binary outcome, X_i is the biomarker whose threshold needs to be determined, $u_{i1}, u_{i2}, \dots, u_{iq}$ are q baseline characteristics, and we define $\mathbf{u}_i = (1, u_{i1}, u_{i2}, \dots, u_{iq})^T \in \mathbb{R}^{q+1}$, where 1 corresponds to the intercept in proposed model. The biomarker X_i could be respiratory rate, systolic blood pressure, etc., and the binary Y_i is predicted by comparing X_i with threshold $c(\mathbf{u}_i)$. Here the threshold $c(\mathbf{u}_i)$ is a function of baseline characteristics \mathbf{u}_i , while the existing methods estimate the optimal constant threshold without using the extra information, such as a patient's gender, age, weight, etc. We define the classification

rule as:

$$\hat{Y}_i = \begin{cases} 1, & \text{if } X_i \geq c(\mathbf{u}_i) \\ -1, & \text{otherwise} \end{cases} = \text{sign}(X_i - c(\mathbf{u}_i)), \quad (3.4)$$

and we assume

$$c(\mathbf{u}_i) = \beta_0 + u_{i1}\beta_1 + \cdots + u_{iq}\beta_q = \mathbf{u}_i^T \boldsymbol{\beta}, \quad (3.5)$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \dots, \beta_q)^T \in \mathbb{R}^{q+1}$. Note that for the case where $\hat{Y}_i = 1$ if $X_i \leq c(\mathbf{u}_i)$, we could let $X'_i = -X_i$, and then fit X'_i in model (3.4).

Below we will present our proposed method to obtain the threshold $c(\mathbf{u}_i)$ in two subsections. Section 3.1 formulates an optimization problem, and section 3.2 introduces the algorithm to solve the problem.

3.3.1 Optimization Problem

The personalized threshold can be obtained by solving an optimization problem. To be more concrete, the $(q + 1)$ -dimensional unknown parameters $\boldsymbol{\beta}$ in (3.5) can be estimated by minimizing the misclassification rate. The remaining multiple challenges include: First, the function $\text{sign}(X_i - c(\mathbf{u}_i))$ in (3.4) is not continuous; second, the 0-1 loss function $I\{\hat{Y}_i \neq \text{sign}(X_i - c(\mathbf{u}_i))\}$ is non-smoothing; third, the consequences of misclassifying sepsis and non-sepsis patients are different, due to the high mortality rate of the sepsis.

In order to address those challenges, we modify the boosting method by minimizing a smooth surrogate weighted exponential loss function,

$$J(\boldsymbol{\beta}) = \frac{1}{N} \sum_{i=1}^N \left(w_{Y_i} \cdot e^{-Y_i(X_i - \boldsymbol{\beta}^T \mathbf{u}_i)} \right), \quad (3.6)$$

where the weight w_{Y_i} depends on outcome Y_i . We define $w_{Y_i} = w_+$ when $Y_i = +1$, and $w_{Y_i} = w_-$ when $Y_i = -1$. Equivalently we can write $w_{Y_i} = (w_+ \cdot (Y_i + 1) + w_- \cdot (1 - Y_i)) / 2$. The values of w_+ and w_- are user specified.

Then the unknown coefficients β can be estimated as below,

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^{(q+1)}}{\operatorname{argmin}} J(\beta). \quad (3.7)$$

This provides a personalized threshold $\hat{c}_i = \mathbf{u}_i^T \hat{\beta}$ in (3.5) for the i -th subject for $i = 1, \dots, N$.

The motivation of the proposed weighted exponential loss function (3.6) is based on the fact that our classification rule in (3.4) is very similar to that in boosting method (3.1). To see this, we define $f_i \triangleq X_i - c(\mathbf{u}_i)$, then the prediction on the outcome Y_i is

$$\hat{Y}_i = \operatorname{sign}(f_i). \quad (3.8)$$

Therefore we borrow the idea from boosting and replace the 0 – 1 loss function $I(Y_i \neq \operatorname{sign}(f_i))$ by the exponential loss $\exp(-Y_i f_i)$. Besides, boosting method assigns different weights to observations, and inspired by this, we introduce two different weights, w_+ and w_- , depending on whether $Y_i = +1$ or -1 , in order to take into account the different consequences of misclassification. These lead us to consider the loss function

$$L(Y_i, f_i) = w_+ \cdot e^{-Y_i f_i} \cdot I(Y_i = 1) + w_- \cdot e^{-Y_i f_i} \cdot I(Y_i = -1).$$

The weighted exponential loss function (3.6) is then derived from $J(\beta) = \frac{1}{N} \sum_{i=1}^N L(Y_i, f_i)$.

3.3.2 Gradient Descent Algorithm

We apply the gradient descent algorithm to solve the parameter $\hat{\beta}$ in (3.7). The procedure can be divided into forward propagation and backward propagation steps. The forward propagation step constructs the cost function vector J in (3.6). If the observed data is $\{Y_i, X_i, u_{i1}, u_{i2}, \dots, u_{iq}\}$, we denote $\mathbf{X} = (X_1, X_2, \dots, X_N)^T$, $\mathbf{Y} = (Y_1, Y_2, \dots, Y_N)^T$, $\mathbf{U} = (\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_N)$, and $\mathbf{u}_i = (1, u_{i1}, u_{i2}, \dots, u_{iq})^T$. The unknown parameter is $\beta =$

$(\beta_0, \beta_1, \beta_2, \dots, \beta_q)^T$. The weights in (3.6) are denoted by $\mathbf{W} = (w_{Y_1}, w_{Y_2}, \dots, w_{Y_N})^T$, where $w_{Y_i} \triangleq (w_+ \cdot (Y_i + 1) + w_- \cdot (1 - Y_i))/2$, with w_{Y_i} being w_+ if $Y_i = 1$ and w_- if $Y_i = -1$.

In order to minimize the average cost, we use backwards propagation by finding the derivative of J with respect to β using the chain rule, and then moving in the direction to reduce total cost. This is repeated until convergence. If we denote $d\beta$ as the derivatives of J with respect to β in current iteration, then the value of β will be updated by $\beta \leftarrow \beta - \alpha d\beta$. The learning rate α is a given small number. After T iterations, we obtain $\hat{\beta}$ as the final estimation of unknown $(q + 1)$ -dim parameter β , then the estimated thresholds can be calculated as $\hat{c} = \mathbf{U}^T \hat{\beta}$ in (3.5).

In summary, our proposed algorithm for personalized threshold $c(\mathbf{u}_i)$ in (3.5) through solving the optimization problem in (3.7) can be presented as follows.

Algorithm 1 Our Proposed Algorithm for Personalized Threshold

Require: $\mathbf{Y}, \mathbf{X}, \mathbf{U}, N, w_- > 0, w_+ > 0, \alpha, T$

- 1: **Initialization:** $\beta_i \leftarrow 0 \forall i \in \{0, 1, 2, \dots, q\}$, $\mathbf{W} = (w_+(\mathbf{Y} + 1) + w_-(1 - \mathbf{Y}))/2$
- 2: **for all** $t = 1, 2, \dots, T$ **do**
- 3: $\mathbf{f} = \mathbf{X} - \mathbf{U}^T \beta$
- 4: $\mathbf{L} = \exp(-\mathbf{Y} * \mathbf{f})$ {*: Element-wise product}
- 5: $\mathbf{J} = \frac{1}{N} \mathbf{L}^T \mathbf{W}$
- 6: $d\mathbf{f} = -\frac{1}{N} \mathbf{W} * \mathbf{Y} * \exp(-\mathbf{Y} * \mathbf{f})$
- 7: $d\beta = -\mathbf{U} d\mathbf{f}$
- 8: $\beta \leftarrow \beta - \alpha d\beta$
- 9: **end for**
- 10: $\hat{c} = \mathbf{U}^T \beta$

} Forward Propagation
 } Backward Propagation

The following proposition shows that $\hat{c} = \mathbf{U}^T \hat{\beta}$ always exists and is well-defined as a point estimate of the optimal threshold.

Proposition 3.3.1. *The weighted exponential loss function $J(\beta)$ in (3.6) is a convex function with respect to β , and thus the gradient descent algorithm converges if we choose a small enough learning rate and long enough optimization steps.*

The proof of Proposition 3.3.1 is postponed to Section 3.6. By formulating the parameter estimation as a convex optimization problem that has a unique solution, we simplify

both the numerical computation and the tuning process for the algorithm parameters. It is now straight forward to evaluate the prediction errors of the classifier in (3.4) through the testing data and cross-validation.

Remark 1. We propose a knowledgeable-based machine learning method that keeps the simple thresholding idea, but at the same time, mimics the idea of boosting. Besides, it is easy to notice that our method is close to the logistic regression model with the coefficient of biomarker fixed at 1. Although classical machine learning methods are becoming increasingly popular in improving health care, they are usually a black box for physicians and nurses and are not suitable for screening due to interpretation and implementation difficulties, which is the reason why constant thresholding method is still commonly adopted in real life. As a special case of logistic regression model, our personalized thresholding method is a surrogate that combines the easily implemented thresholding method with machine learning technique, improving the predictive accuracy without adding too much computational burden or complexity. Later, we will compare our proposed method with logistic regression and AdaBoosting in the context of sepsis screening.

Remark 2. Our proposed method can be extended to multiple biomarkers. When applying our proposed method to sepsis screening, we need to estimate the personalized thresholds for both respiratory rate and systolic blood pressure in qSOFA criteria, but for GCS score, the constant threshold of 15 is kept in that obtaining the personalized threshold for discrete biomarker is out of scope of this chapter. We define c_{i1} and c_{i2} to be the personalized thresholds of respiratory rate and systolic blood pressure, respectively, depending on variables u'_{ij} 's that are subject's baseline characteristics. We assume $c_{ik} = \beta_{0k} + u_{i1}\beta_{1k} + \dots + u_{iq}\beta_{qk}$, for some unknown parameters β_{jk} 's, with $k = 1, 2$. Ideally, we want to find suitable choices of the $2(q + 1)$ parameters β_{jk} 's from the training data, so that the qSOFA criteria could has good predictive performance for the testing data. However, it is non-trivial to jointly estimate them simultaneously from the training data. We therefore decompose the $2(q + 1)$ -dimensional estimation problem into 2 different $(q + 1)$ -dimensional estimation problems,

and estimate the $(q + 1)$ -dimensional vector $(\beta_{0k}, \dots, \beta_{qk})$ in the personalized thresholds c_{ik} recursively one at a time for each $k = 1, 2$. In general this might lose statistical efficiency since we ignore the intercorrelation between the biomarkers, but it will gain computational efficiency. Moreover, it is a reasonable approach for the sepsis screening context, as the three biomarkers (respiratory rate, systolic blood pressure, and Glasgow Coma Scale (GCS) score) characterize different physical and mental aspects of sepsis patients. In particular, the classifier is considered to have good properties only if each biomarker yields a good prediction of binary outcome, and the constant thresholds of the current qSOFA guideline are also based on the component-to-component optimization.

3.4 The Data Set

We use the Medical Information Mart for Intensive Care III (MIMIC-III) database (version 1.4) [104, 105], a large and freely-available database comprised of de-identified health-related data associated with over forty thousand patients who stayed in critical care units of the Beth Israel Deaconess Medical Center between 2001 and 2012. There were a total of 46,520 patients in the data set. The International Classification of Diseases, Ninth Revision (ICD-9) coding was used to identify sepsis and non-septic patients. ICD-9 is a list of codes intended for the classification of diseases and a wide variety of signs, symptoms, abnormal findings, complaints, social circumstances, and external causes of injury or disease.

Below we will present the details in four subsections. In section 3.4.1, we introduce how the study group and control group are selected. Section 3.4.2 shows the interested variables. We discuss the summary statistics of those variables in Section 3.4.3 and conduct exploratory analysis on qSOFA biomarkers in Section 3.4.4.

3.4.1 Study Population

The details of cohort selection from the Mimic III data is provided in Figure 3.1. There are 36,543 adult patients (aged 18 years or older), 4,233 of which have sepsis-related ICD-

9 codes (995.91 for sepsis, 995.92 for severe sepsis, and 785.52 for septic shock). In some cases, a patient is assigned more than one ICD-9 code, and if any of the ICD-9 codes is sepsis related, we consider them as being diagnosed with sepsis. We retrieve comprehensive clinical data, including patient demographic and clinical measurements for qSOFA biomarkers. After excluding those sepsis patients who have no observations in the qSOFA variables within the first 24 hours after admission, we generate a study group of 3,771 adult patients with sepsis.

We form a control group by randomly sampling 4,000 adult non-sepsis patients from the MIMIC III database after excluding those patients with sepsis related ICD-9 codes, having infection plus meeting SIRS criteria, and with missing observations in the qSOFA variables.

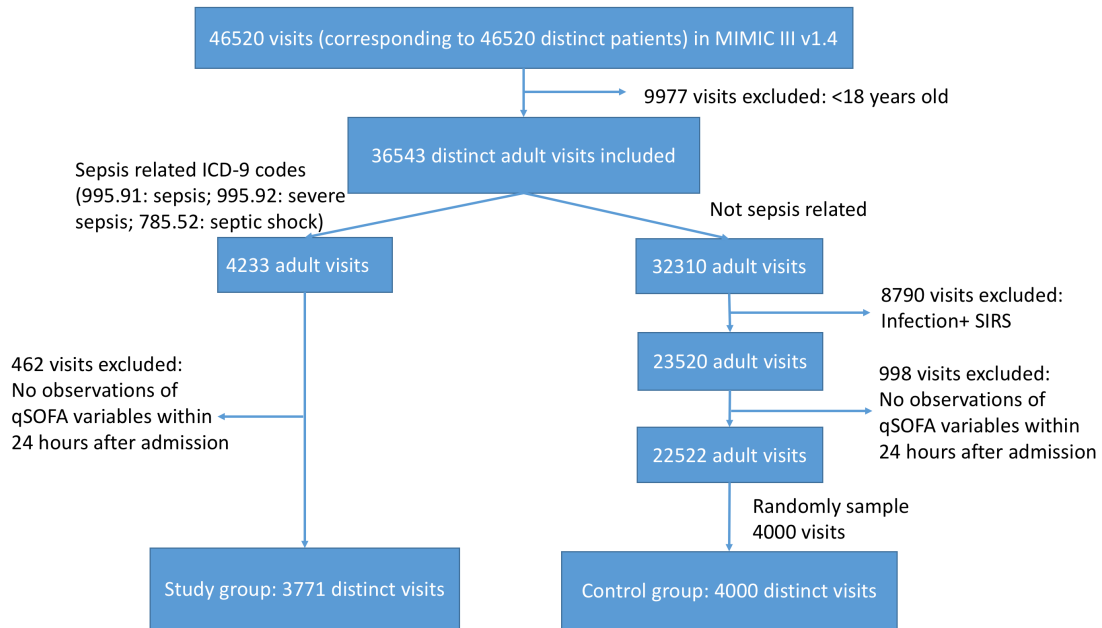


Figure 3.1: Flowchart of cohort selection

3.4.2 Observed Data

The observed data can be written in the form $\{Y_i, X_{i1}, X_{i2}, X_{i3}, u_{i1}, u_{i2}, \dots, u_{i7}\}$ for $i = 1, 2, \dots, N$, where $Y_i = -1$ or 1 indicates whether the i -th subject is diagnosed without or with sepsis. The triplet (X_{i1}, X_{i2}, X_{i3}) denotes the qSOFA variables of respiratory rate,

Table 3.1: Variables and corresponding definitions.

Variables	Definitions
Y_i	Sepsis indicator for patient i
X_{i1}	Maximum respiratory rate within first 24 hours for patient i
X_{i2}	Minimum systolic blood pressure within first 24 hours for patient i
X_{i3}	Minimum Glasgow Coma Scale (GCS) score within first 24 hours for patient i
u_{i1}	Age for patient i
u_{i2}	Sex for patient i
u_{i3}	Admission location for patient i (1=Emergency room admit; 0=Others)
u_{i4}	Admission type for patient i (1=Emergency and Urgent; 0=Others)
u_{i5}	Ethnicity for patient i (1=White; 2=Black; 3=Hispanic; 4=Others)
u_{i6}	Insurance type for patient i (1=Medicaid; 0= Self pay)
u_{i7}	Marital status for patient i (1=Married; 0=Others)

systolic blood pressure, and GCS scores observed for the i -th subject within the first 24 hours after ICU admission. The $(u_{i1}, u_{i2}, \dots, u_{i7})$ variables represent the demographic variables of age, gender, admission location, admission type, ethnicity, insurance, and marital status. The total number of patients N is 7,771, including 3,771 with sepsis and 4,000 without sepsis. Variable definitions are provided in Table 3.1.

3.4.3 General Characteristics

Descriptive statistics are calculated for all variables of interest, and shown in Table 3.2. The mean and standard deviation (SD) are compared across sepsis and non-sepsis groups using two-sample t -tests for continuous data. Categorical data are presented as counts and percentages, and they are compared between two groups by Fisher's exact test or Chi-square test. All of the selected variables are significantly correlated with the sepsis outcome ($p < 0.05$).

3.4.4 Exploratory Analysis on qSOFA Variables

Histograms for the qSOFA variables are shown in Figure 3.2. Sepsis patients tend to have lower GCS score, higher respiratory rate, and lower systolic blood pressure as compared

Table 3.2: Descriptive statistics of interested variables.

Variables		Sepsis (<i>N</i> = 3771)	Non-Sepsis (<i>N</i> = 4000)	<i>p</i> -Values
Max Respiratory Rate, mean (SD)		30.6 (8.9)	26.5 (7.4)	7.6×10^{-106}
Min Systolic Blood Pressure, mean (SD)		82.3 (17.4)	93.5 (18.1)	7.2×10^{-163}
Altered mental status, count (%) (Min GCS < 15)		1580 (41.9)	1407 (35.2)	1.2×10^{-9}
Age, mean (SD)		65.3 (15.6)	61.6 (16.6)	9.9×10^{-25}
Sex, count (%)	Male	2148 (57.0)	2397 (59.9)	8.1×10^{-3}
	Female	1623 (43.0)	1603 (40.1)	
Admission location, count (%)	Emergency Room	1767 (46.9)	1528 (38.2)	1.3×10^{-14}
	Others	2004 (53.1)	2472 (61.8)	
Admission type, count (%)	Emergency and Urgent	3635 (96.4)	3110 (77.8)	0
	Others	136 (3.6)	890 (22.3)	
Ethnicity, count (%)	White	2728 (72.3)	2830 (70.8)	3.6×10^{-6}
	Black	349 (9.3)	280 (7)	
	Hispanic	125 (3.3)	143 (3.6)	
	Others	569 (15.1)	747 (18.7)	
Insurance, count (%)	Medicaid	2703 (71.7)	2302 (57.6)	0
	Self pay	1068 (28.3)	1698 (42.5)	
Marital status, count (%)	Married	1719 (45.6)	2075 (51.9)	3.0×10^{-8}
	Others	2052 (54.4)	1925 (48.1)	

to non-sepsis patients. However, the difference of GSC scores among the two cohorts is not strong, and most of the patients have GCS score 14 or 15 (conscious mental status). In our study, we focus on obtaining the personalized thresholds for respiratory rate and systolic blood pressure, and keep the constant cutoff 15 for GCS score, since GCS score is a discrete variable. The scatter plot of systolic blood pressure against respiratory rate for the two groups is presented in Figure 3.3. Most of the sepsis cohort lie on the lower right (high respiratory rate and low systolic blood pressure), while most of the non-sepsis cohort lie on the upper left (low respiratory rate and high systolic blood pressure). qSOFA criteria with constant thresholds may classify those patients well, however, the two cohorts overlap in the middle of the plot, and hence constant thresholds may lose power in identifying sepsis patients among them.

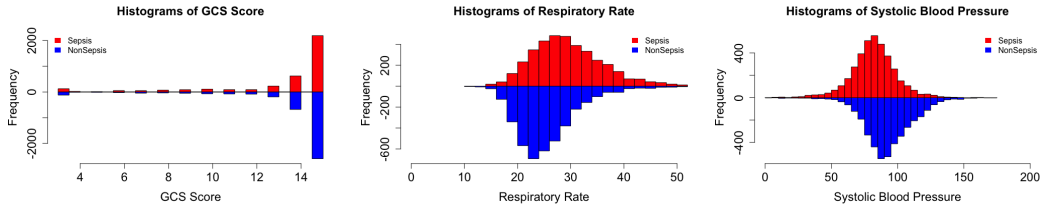


Figure 3.2: Histograms of qSOFA variables.

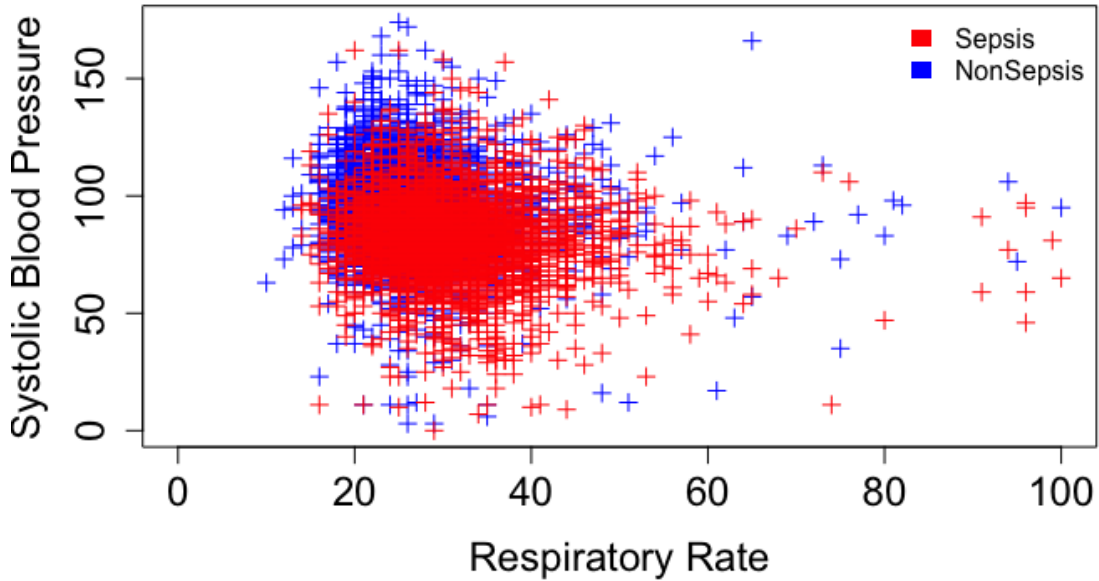


Figure 3.3: Scatter plot of systolic blood pressure against respiratory rate

3.5 Application to Sepsis Screening

In this section, we apply our proposed personalized threshold method to MIMIC III data set for sepsis screening. For the purpose of comparison, we consider 6 baseline methods, including the original qSOFA criteria and five other standard methods, minP, Youden Index, Closest-to-(0,1), Concordance Probability, and Index of Union. Since our focus is on the prediction and classification, we use the random cross-validation to evaluate performances of all methods. Specifically, for each iteration, we randomly divide the real data into training (80%) and testing (20%) sets, and apply our method and the six baseline methods on

the training set to obtain the thresholds for respiratory rate and systolic blood pressure, respectively, and then the obtained thresholds are used to classify subjects as septic or non-septic in the testing set to calculate the classification accuracy, sensitivity, and specificity. We then repeat this process 100 times, and report the averaged testing error statistics.

In our proposed personalized threshold method, we replace the constant thresholds for respiratory rate and systolic blood pressure by the personalized thresholds in the qSOFA criteria and keep the constant threshold of 15 for GCS score, i.e., the altered mental status occurs when GCS score is less than 15. The parameters $T = 30000$, $\alpha = 0.001$, $w_+ = 1$, and $w_- = 1$ are selected based on a grid search to maximize the prediction accuracy.

For better presentations, we split this section into four subsection. Section 3.5.1 discusses the tuning parameters in our proposed method. Section 3.5.2 and Section 3.5.3 compare our method with other existing constant threshold methods and machine learning techniques, respectively. In Section 3.5.4, we provide the interpretation of our personalized threshold method and illustrate how to implement it in practice.

3.5.1 Tuning parameters

In this subsection, we illustrate how the learning rate α and the total number of iterations T influence the speed of convergence in our gradient descent algorithm 1. In addition, we discuss the tradeoff between sensitivity and specificity from tuning the weights w_+ and w_- in the proposed cost function (3.6).

Figure 3.4 shows how the learning rate influenced the loss function (w_+ and w_- were set to be 1) over all of the data. The learning rates of 0.05, 0.1, 0.15, and 0.2 were chosen to ensure the algorithm reached the global optimal. Indeed, larger learning rates tend to converge faster.

In our application, the parameters $T = 30000$ and $\alpha = 0.001$ are selected by grid search to maximize the averaged testing accuracy in a five-fold cross validation. Notice that the algorithm with $T = 30000$ and $\alpha = 0.001$ might not converge on the training data, but it

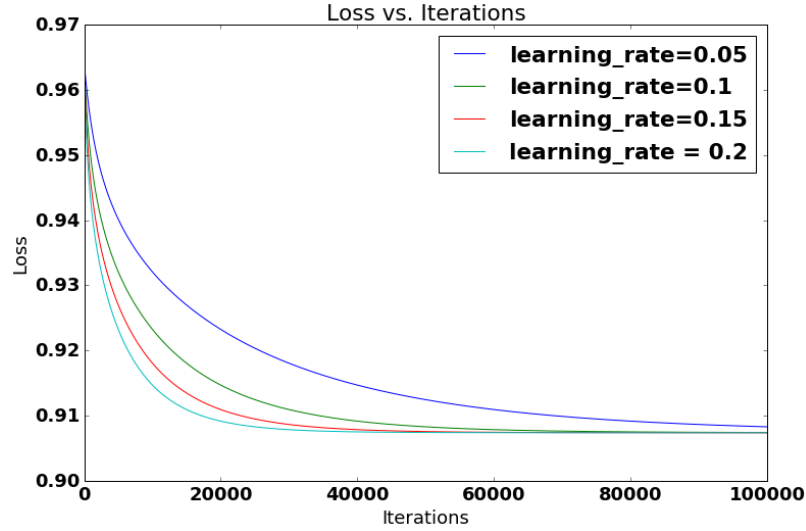


Figure 3.4: The weighted exponential loss versus the number of iterations with different learning rate α

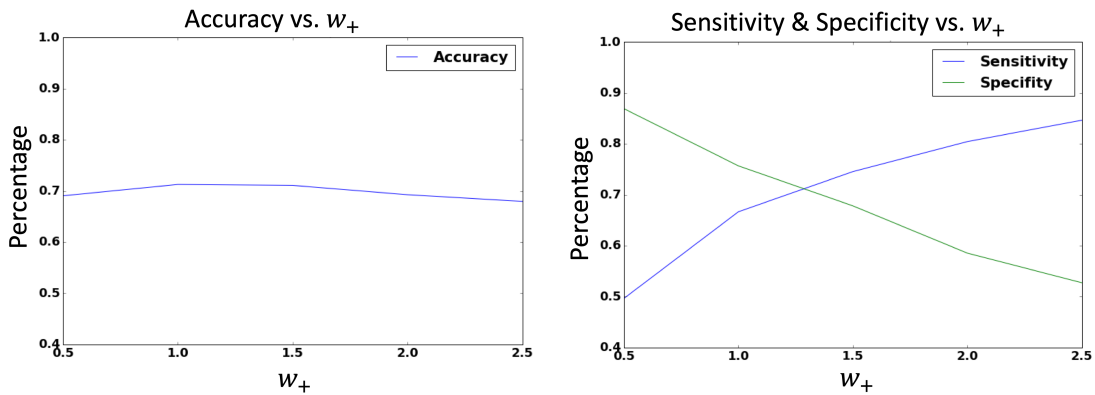


Figure 3.5: The accuracy (left), sensitivity and specificity (right) with different w_+ 's and fixed $w_- = 1$.

led to the highest classification accuracy. This observation is in line with “early stopping” concept in machine learning, which is to avoid overfitting when training a learner with an iterative method, such as gradient descent [106, 107, 108, 109, 110].

Figure 3.5 shows the averaged accuracy, sensitivity, and specificity on the testing data in a five-fold cross validation for fixed $w_- = 1$ and different values for w_+ . The learning rate and number of iterations are set as $\alpha = 0.001$ and $T = 30000$. The overall accuracy is essentially constant when we change the values of w_+ . There is an obvious increasing trend of sensitivity as we increase w_+ , since it penalizes more for the misclassification of

Table 3.3: Overall accuracy

Methods	Individual Biomarker		Combined in qSOFA
	Respiratory Rate	Systolic Blood Pressure	
Personalized threshold	0.6781	0.6889	0.6850
qSOFA threshold	0.5464	0.5899	0.5853
minP	0.6456	0.6359	0.6497
Youden	0.6466	0.6371	0.6524
Closest-to-(0,1)	0.6467	0.6379	0.6536
Concordance Probability	0.6466	0.6380	0.6534
Index of Union	0.6468	0.6377	0.6528

patients with sepsis. This demonstrates the flexibility of Algorithm 1

3.5.2 Comparison to qSOFA criteria with constant thresholds

We compare the personalized threshold method with the original qSOFA criteria and five other standard methods, minP, Youden Index, Closest-to-(0,1), Concordance Probability, and Index of Union. The averaged classification accuracy, sensitivity, and specificity of different methods based on each individual biomarker and their combination in qSOFA criteria are compared as shown in Tables 3.3, 3.4, and 3.5, respectively. When using respiratory rate alone to classify, we identify a patient as having sepsis if the measurement is greater than the obtained threshold, while for systolic blood pressure, a sepsis patient is identified if it is less than the optimal threshold. When combining them in the qSOFA criteria, a patient is identified to be of high risk of developing sepsis if two of the following three criteria are satisfied: respiratory rate is greater than the obtained threshold, systolic blood pressure is less than its corresponding threshold, and GCS score is less than 15.

The personalized threshold method yields the largest overall accuracies. The accuracies for constant thresholds in the qSOFA criteria are the lowest. The five standard methods, minP, Youden Index, Closest-to-(0,1), Concordance Probability, and Index of Union, have similar classification accuracies, which are higher than those for qSOFA with constant cutoffs but lower than our personalized method.

Table 3.4: Sensitivity

Methods	Individual Biomarker		
	Respiratory Rate	Systolic Blood Pressure	Combined in qSOFA
Personalized threshold	0.6509	0.6924	0.6354
qSOFA threshold	0.9294	0.8841	0.8966
minP	0.6179	0.5816	0.5647
Youden	0.6240	0.6156	0.5858
Closest-to-(0,1)	0.6238	0.6328	0.5947
Concordance Probability	0.6241	0.6296	0.5931
Index of Union	0.6240	0.6210	0.5886

Table 3.5: Specificity

Methods	Individual Biomarker		
	Respiratory Rate	Systolic Blood Pressure	Combined in qSOFA
Personalized threshold	0.7038	0.6857	0.7318
qSOFA threshold	0.1851	0.3125	0.2917
minP	0.6718	0.6871	0.7301
Youden	0.6680	0.6574	0.7152
Closest-to-(0,1)	0.6680	0.6426	0.7092
Concordance Probability	0.6681	0.6458	0.7103
Index of Union	0.6679	0.6534	0.7135

In table 3.4, qSOFA with constant thresholds corresponds to the highest sensitivities: 92.94% for respiratory rate, 88.41% for systolic blood pressure, and 89.66% for their combination, while using our personalized threshold method, the sensitivities are 65.09%, 69.24%, and 63.54%, respectively. Other standard methods have the sensitivities ranging from 61.79% to 62.41% for respiratory rate, from 58.16% to 63.28% for systolic blood pressure, and from 56.47% to 59.47% for them combined.

The classification specificities are detailed in Table 3.5. qSOFA with constant thresholds has the lowest specificities (18.51% for respiratory rate, 31.25% for systolic blood pressure, and 29.17% for their combination in qSOFA), while using the personalized thresholds, the specificities are increased to 70.38% for respiratory rate, 68.57% for systolic blood pressure, and 73.18% for their combination in qSOFA. The specificities of other standard

methods range from 66.79% to 67.18% for respiratory rate, from 64.26% to 68.71% for systolic blood pressure, and from 70.92% to 73.01% for them combined in qSOFA.

In general, the personalized threshold method yields the largest prediction accuracy and the best balance of sensitivity and specificity. Note that in this application, we choose the parameters $w_+ = w_- = 1$ in our cost function, however, we can adjust the balance between sensitivity and specificity by choosing different values of w_+ and w_- .

3.5.3 Comparison with other machine learning techniques

We also apply logistic regression and AdaBoosting to predict sepsis using the qSOFA variables respiratory rate, systolic blood pressure, and GCS score, together with baseline demographic variables age, sex, admission location, admission type, ethnicity, insurance, and marital status. As described previously, the models are built on the training set (80%) and then applied on the testing set (20%) to classify the subjects with 100 repetitions of randomly splitting. The comparisons between personalized qSOFA with logistic regression and AdaBoosting are presented in Table 3.6.

The averaged classification accuracy of the personalized qSOFA is close to those obtained from logistic regression and AdaBoosting. The averaged sensitivity of the personalized qSOFA is less than those from the machine learning methods, but the specificity of the personalized qSOFA is the largest. In general, the personalized qSOFA is comparable to the more difficult-to-interpret machine learning methods.

Compared to machine learning methods, personalized qSOFA can be easily manipulated and interpreted by physicians and nurses, and requires no statistical training. Using the personalized qSOFA for sepsis screening is also efficient, since there is only a one time calculation of the personalized thresholds for each patient. Machine learning methods, on the other hand, require an update each time there is a new observation, which is not computationally efficient.

Table 3.6: Comparison with other machine learning techniques

Methods	Overall Accuracy	Sensitivity	Specificity
Personalized qSOFA	0.6850	0.6354	0.7318
Logistic Regression	0.7182	0.7277	0.7094
AdaBoosting	0.7171	0.7146	0.7196



Figure 3.6: The predicted personalized threshold for respiratory rate against age.

3.5.4 Interpretation and Implementation of Personalized qSOFA

After applying the proposed model to the data, we obtain the estimated personalized thresholds for each individual. Here, we are going to explore how the estimated thresholds are related to the patients' baseline information. Figure 3.6 plots the predicted personalized threshold for respiratory rate against age. The personalized cutoffs show a decreasing trend as age increases. This suggests that for older patients with lower respiratory rate, the threshold should be set lower than that set for their younger counterparts. Patients with the same age may differ in other baseline characteristics, however, which would lead to different recommended personalized thresholds to improve overall classification accuracy.

An advantage of our approach is that we only use easily accessible demographic variables to estimate the personalized thresholds. Therefore, the threshold can be calculated

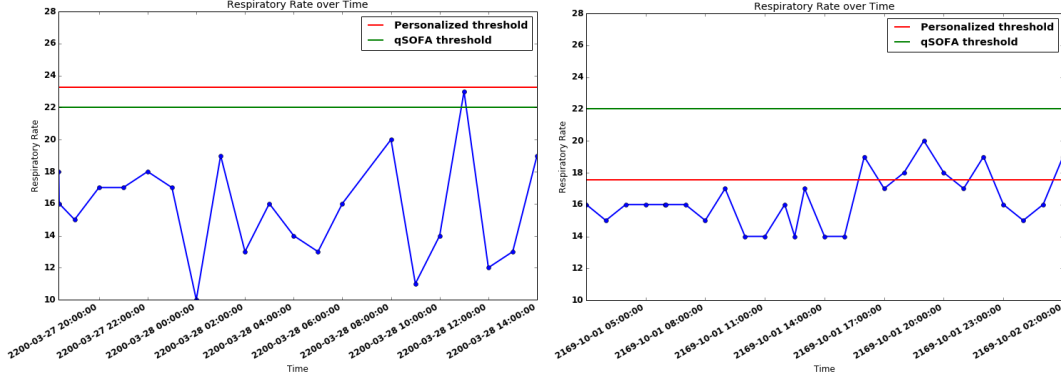


Figure 3.7: Left: Screening for non-sepsis patient. Right: Screening for sepsis patient.

and fixed as soon as the patient is admitted. The personalized threshold can be treated and manipulated in exactly the same way as the constant threshold in sepsis screening. Figure 3.7 illustrates this. It shows two selected examples of screening for respiratory rate using the personalized threshold and the constant threshold 22 in qSOFA criteria: one for a non-sepsis patient, and the other for a sepsis patient. As can be seen from the plots, the personalized threshold for each patient is a fixed line just as the constant qSOFA threshold, but their values are different for different subjects.

3.6 Proof of Proposition 3.3.1

It suffices to show that the weighted exponential loss function $J(\beta)$ in (3.6) is a convex function with respect to β . Recall that $\beta = [\beta_0, \dots, \beta_q]^T$ and without loss of generality, we can rewrite that $\mathbf{u}_i = [u_{i0}, \dots, u_{iq}]^T$ with $u_{i0} = 1$, for any $i \in \{1, \dots, N\}$. It is sufficient to prove that for any $\mathbf{z} \in \mathbb{R}^{q+1}$, $\mathbf{z}^T [\nabla^2 J(\beta)] \mathbf{z} \geq 0$, where $\nabla^2 J(\beta)$ is the Hessian matrix of $J(\beta)$. First, we compute the first order derivative of $J(\beta)$ with respect to β . For any $j \in \{0, \dots, q\}$, we have

$$\frac{\partial J(\beta)}{\partial \beta_j} = \frac{1}{N} \sum_{i=1}^N (w_{Y_i} \cdot e^{-Y_i f_i} \cdot Y_i u_{ij})$$

Second, we compute the second order derivatives. For any $j \in \{0, \dots, q\}$ and any $k \in \{0, \dots, q\}$,

$$\begin{aligned}\frac{\partial^2 J(\boldsymbol{\beta})}{\partial \beta_j^2} &= \frac{1}{N} \sum_{i=1}^N (w_{Y_i} \cdot e^{-Y_i f_i} \cdot (Y_i u_{ij})^2), \\ \frac{\partial^2 J(\boldsymbol{\beta})}{\partial \beta_j \partial \beta_k} &= \frac{1}{N} \sum_{i=1}^N (w_{Y_i} \cdot e^{-Y_i f_i} \cdot Y_i^2 u_{ij} u_{ik}).\end{aligned}$$

Therefore, for any $\mathbf{z} = [z_0, \dots, z_q]^T$, we have

$$\begin{aligned}\mathbf{z}^T [\nabla^2 J(\boldsymbol{\beta})] \mathbf{z} &= \frac{1}{N} \sum_{j=0}^q \sum_{k=0}^q \sum_{i=1}^N (w_{Y_i} Y_i^2 \cdot e^{-Y_i f_i} \cdot z_j z_k u_{ij} u_{ik}) \\ &= \frac{1}{N} \sum_{i=1}^N w_{Y_i} Y_i^2 \cdot e^{-Y_i f_i} \sum_{j=0}^q \sum_{k=0}^q (z_j z_k u_{ij} u_{ik}) \\ &= \frac{1}{N} \sum_{i=1}^N w_{Y_i} Y_i^2 \cdot e^{-Y_i f_i} (\mathbf{z}^T \mathbf{u}_i)^2.\end{aligned}\tag{3.9}$$

Since $w_{Y_i} \geq 0$ for all $i \in \{1, \dots, N\}$, we have that $\mathbf{z}^T [\nabla^2 J(\boldsymbol{\beta})] \mathbf{z} \geq 0$ for all \mathbf{z} . \square

3.7 Conclusions

Sepsis is difficult to identify and diagnose, and unfortunately, there is not a validated standard diagnostic test for sepsis at present. The idea of “screening” plus “assessing” as recommended by the 2016 Task Force [72] is attractive, but there are some drawbacks in using the qSOFA score. The most important limitation is that it has rather low sensitivity in identifying patients at high risk of developing sepsis.

We developed a personalized threshold method that is able to adjust the thresholds in the qSOFA criteria based on the subject’s baseline characteristics, age, sex, admission location, admission type, ethnicity, insurance, and marital status. We assumed the personalized thresholds were a linear function of those demographic variables and developed a boosting-based method to obtain the personalized thresholds for efficient screening of sepsis. The gradient descent algorithm was applied to obtain the unknown parameters in the linear

function to calculate the personalized thresholds. The method yielded efficient personalized monitoring, appropriate subject-specific intervention in early stages of sepsis, and thus a significant reduction of the mortality rate.

Our method was applied to the MIMIC III data (ICU populations) to find the optimal personalized thresholds for the qSOFA variables of respiratory rate and systolic blood pressure. The constant thresholds in qSOFA were replaced by those obtained from our method for classifying patients as septic or non-septic. We compared personalized qSOFA with the original qSOFA criteria and five other standard methods to obtain the optimal constant threshold for a single biomarker (minP, Youden Index, Closest-to-(0,1), Concordance Probability, and Index of Union). Our method yielded the largest overall testing accuracy for identifying sepsis patients. The constant qSOFA had a high sensitivity but a very low specificity in ICU populations, while our personalized qSOFA yielded a better balance. In general, the five standard methods performed better than constant qSOFA but worse than the personalized qSOFA.

The personalized qSOFA has comparable performance to logistic regression and AdaBoosintg, but has the benefit of being easily used and interpreted by physicians and nurses. Our personalized qSOFA method only requires a one time calculation of the personalized threshold for each patient, whereas machine learning methods require an update each time there is a new observation, which is not computationally efficient. In addition, the balance of sensitivity and specificity can be easily adjusted in our method by tuning the weighting parameters.

There are several limitations that should be mentioned. When obtaining the personalized thresholds for respiratory rate and systolic blood pressure in qSOFA criteria, we did not jointly estimate them, and therefore ignored the possible correlation between them. Although we did preliminary analysis and found that the baseline characteristic variables age, sex, admission location, admission type, ethnicity, insurance, and marital status were significantly correlated with sepsis outcome, we did not consider variable selection when

putting them in the model to estimate the thresholds. In future studies, more clinical variables may become available, and therefore performing variable selection will likely be necessary. Finally, we only focused on the scenario where the threshold is a linear function of those baseline characteristics, which might not always be the case.

It is worth noting that although we demonstrated the use of personalized thresholds for sepsis screening, the general approach can be applied to other clinical screening applications. Examples include personalizing the HAVOC score, a clinical score for predicting atrial fibrillation in patients with cyptogenic stroke or transient ischemic attack [111] and the Fong clinical risk score for predicting colorectal cancer recurrence [112].

CHAPTER 4

REGULARIZED ZERO-AND-ONE INFLATED BETA REGRESSION MODEL WITH APPLICATION IN PREDICTING POSTOPERATIVE COMPLICATIONS

4.1 Introduction

Postoperative complications occur in up to 50% of patients undergoing major, elective surgery [113, 114]. The postoperative complications are the leading cause of after-surgery morbidity and mortality [115], and have a deeper impact on long-term survival after surgery [116, 117]. Early and precise prediction on postoperative complications can help to modify the postoperative course and guarantee timely intervention, thereby increasing the survival rate and improving the quality of life. However, the lack of standardization in definitions and quantification of complications continues to be a major obstacle. It is very common that authors inconsistently quantify complications as "minor" and "major", or report only those personally judged to be significant [118].

In 2013, Slankamenac et al. developed the Comprehensive Complication Index (CCI), a novel scale to measure surgical morbidity [119, 120]. The CCI was developed taking into account the severity of multiple complications as evaluated by different stakeholders, including patients, surgeons, perioperative physicians, and nurses. It is designed to capture the overall burden of complications in the postoperative period. The CCI is a linear scale ranging from 0 (no complications) to 100 (death). As such, it represents a convenient method for quantifying postoperative morbidity that sums up the relative contribution of each individual complication.

However, the CCI is calculated from a complicated procedure, and is not practical for everyday use [119]. Besides, it can be calculated only retrospectively, when the hospitalization is finished, it can only reflect the results of perioperative treatment but cannot

be used as a measure of patients' current status. In order to simplify the calculation of CCI and make it prospective, we aim to develop a machine learning method to self-learn the CCI values based on patients' clinical covariates, such as heart rate, diabetes status, etc. The major challenges include that CCI is within a limited range, besides it exhibits heteroskedasticity, and the simple linear regression that assumes the response is random variable taking any real number is not suitable here.

Statistical modeling of limited range data has received close attention in the past decades. Different strategies have been proposed, among which beta regression models are of particular interests [121, 122, 123, 124, 125, 126, 127, 128]. In beta regression models, the response variable, usually in $(0, 1)$, is assumed to be beta distributed. As is well known, beta distribution is very flexible in the sense that its density can have quite different shapes depending on the values of the two parameters that index the distribution. The beta density is

$$\pi(y; \mu, \phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)} y^{\mu\phi-1} (1-y)^{(1-\mu)\phi-1}, \quad y \in (0, 1), \quad (4.1)$$

where $0 < \mu < 1$ is the mean, $\phi > 0$ is a precision (or dispersion) parameter, and $\Gamma(\cdot)$ is the gamma function. The mean and variance are $\mathbb{E}(y) = \mu$ and $\text{var}(y) = \mu(1-\mu)/(\phi+1)$. The mean μ and precision parameter ϕ of the beta distributed response is then modeled as a function of a set of exogenous variables. Compared to usual regression models, such as normal linear or nonlinear regression models, beta regression models are more suitable to model variables in open interval $(0, 1)$, for example the rates, percentages and proportions. However, they cannot be applied when response includes a non negligible number of zeros and ones, a situation often found in empirical research.

The inflated beta regression models have been proposed to model the cases where zeros or ones appear [129, 130, 131, 132]. The zero-or-one inflated beta regression model is one of the most famous approaches proposed by Ospina in 2012 [130], and is developed for continuous proportions when the data contains zeros or ones. It models response variable using a mixture of two distributions: a beta distribution and a degenerate distribution in a

known value c , where c equals zero or one, depending on the case. Under this scheme, the probability density function of the response variable y is given by

$$f_c(y; \alpha, \mu, \phi) = \begin{cases} \alpha, & \text{if } y = c \\ (1 - \alpha)\pi(y; \mu, \phi), & \text{if } y \in (0, 1) \end{cases}, \quad (4.2)$$

where $\pi(y; \mu, \phi)$ is the beta density in (4.1), and $0 < \alpha < 1$ is the probability mass at c and represents the probability of observing zero or one. The mean and variance are $\mathbb{E}(y) = \alpha c + (1 - \alpha)\mu$ and $\text{var}(y) = \alpha(1 - \alpha)(c - \mu)^2 + (1 - \alpha)\mu(1 - \mu)/(\phi + 1)$. The parameters of the mixture distribution are then modeled as functions of other interested variables. Unfortunately, this zero-or-one inflated beta regression model is not suitable to model the cases where both zeros and ones exist.

In this chapter, we propose a zero-and-one inflated beta regression model for continuous response variables in the range of $[0, 1]$. We assume a mixture of a beta distribution and a Bernoulli distribution for response variables [133], which has the density function

$$f(y; \alpha, \gamma, \mu, \phi) = \begin{cases} \alpha(1 - \gamma), & \text{if } y = 0 \\ (1 - \alpha)\pi(y; \mu, \phi), & \text{if } y \in (0, 1) \\ \alpha\gamma, & \text{if } y = 1 \end{cases}, \quad (4.3)$$

with $0 < \alpha, \gamma, \mu < 1$ and $\phi > 0$. Then we will relate the density parameters to linear predictors through link functions. In addition, we will conduct variable selection for zero-and-one inflated beta regression model.

Similar to the linear model, variable selection is an important issue for beta regression models, and two different variable selection approaches have been proposed: 1) Zhao et al. proposed variable selection for beta regression model based on a penalized likelihood method [134]; 2) Bayer and Cribari-Neto proposed a fast two step model selection scheme to select variables using some efficient model selection criteria [135]. However, both of the above methods are for simple beta regression models without considering the cases when

zeros or/and ones appear.

We borrow the idea from the regularized generalized linear regression models, and develop an efficient penalized likelihood-based method that uses the elastic net penalty to select important explanatory variables for the proposed zero-and-one inflated beta regression model. We also propose a fast algorithm to solve for the penalized maximum log-likelihood function, mimicking the ideas of proximal Newton-type methods for convex optimization.

The remainder of this chapter is organized as follows. In Section 4.2, we introduce background information, including the beta regression model, zero-or-one inflated beta regression model, and variable selection for generalized linear models. Section 4.3 describes our proposed regularized zero-and-one inflated beta regression model. Simulation results are reported in Section 4.4, and we will apply the proposed model to predict CCI in Section 4.5. The derivations of score function and fisher information matrix are shown in Section 4.6. We provide conclusions in Section 4.7.

4.2 Background

We will introduce the background information in three subsections, including the well known varying dispersion beta regression model (VBRM), the zero-or-one inflated beta regression model, and the algorithms for estimation of generalized linear models with convex penalties.

4.2.1 Varying dispersion beta regression model

Beta regression model was first proposed by Ferrari [121], and was built for beta distributed random variables. The density of the beta distribution is given in (4.1), where it is indexed by μ and ϕ . the parameter μ is the mean of the response variable, and ϕ can be interpreted as a precision parameter in the sense that, for fixed μ , the larger the value of ϕ , the smaller the variance of the variable [121, 124, 123]. Note that the parameterization of the beta density

is not unique, and another common expression is indexed by $p = \mu\phi$ and $q = (1 - \mu)\phi$,

$$\pi(y; p, q) = \frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)} y^{p-1} (1-y)^{q-1}, 0 < y < 1.$$

It is easy to check that $p > 0$, $q > 0$, and the mean and variance are

$$\mathbb{E}(y) = \frac{p}{p+q} \quad \text{and} \quad \text{var}(y) = \frac{pq}{(p+q)^2(p+q+1)}.$$

However, here we use the beta density in (4.1), since it is typically more useful to model the mean of the response and the precision parameter.

It is well known that the densities of beta distribution can display different shapes depending on the values of the two parameters μ and ϕ . To be more concrete, when $\mu = 1/2$ it tends to be symmetric, while for the cases with $\mu \neq 1/2$, the shape is asymmetric. Additionally, when $\mu = 1/2$ and $\phi = 2$, the density reduces to that of a standard uniform distribution.

Let y_1, \dots, y_n be independent random variables, and each y_t , $t = 1, \dots, n$, follows the beta density in (4.1) with mean μ_t and precision ϕ_t . In the standard formulation of the beta regression model [121], it is assumed that the precision parameter $\phi_t = \phi$ is constant across all y_t 's. However, such assumptions cannot hold in many practical situations. The varying dispersion beta regression model (VBRM) allows for the dispersion to vary across observations [136, 137]. Therefore, in VBRM, y_t is assumed to follow the beta density with mean μ_t and precision ϕ_t ,

$$\pi(y_t; \mu_t, \phi_t) = \frac{\Gamma(\phi_t)}{\Gamma(\mu_t\phi_t)\Gamma((1-\mu_t)\phi_t)} y_t^{\mu_t\phi_t-1} (1-y_t)^{(1-\mu_t)\phi_t-1}, \quad y_t \in (0, 1), \quad t = 1, \dots, n. \quad (4.4)$$

The model is obtained by assuming the mean μ_t and precision ϕ_t can be written as

$$h_1(\mu_t) = \mathbf{Z}_t^T \boldsymbol{\sigma} = \eta_{1t}, \quad (4.5)$$

$$h_2(\phi_t) = \mathbf{Q}_t^T \boldsymbol{\tau} = \eta_{2t}, \quad (4.6)$$

where $\mathbf{Z}_t = (1, Z_{t1}, \dots, Z_{tm})^T \in \mathbb{R}^{m+1}$ and $\mathbf{Q}_t = (1, Q_{t1}, \dots, Q_{tl})^T \in \mathbb{R}^{l+1}$ are covariates which can either be the same or different depending on cases, $\boldsymbol{\sigma} = (\sigma_0, \sigma_1, \dots, \sigma_m)^T$ and $\boldsymbol{\tau} = (\tau_0, \tau_1, \dots, \tau_l)^T$ are vectors of unknown coefficients, $\boldsymbol{\eta}_1 = (\eta_{1t}, \dots, \eta_{1n})^T$ and $\boldsymbol{\eta}_2 = (\eta_{2t}, \dots, \eta_{2n})^T$ are called the linear predictors, and $h_1(\cdot)$, $h_2(\cdot)$ are strictly monotonic and twice differentiable link functions, which often are chosen as logistic function and logarithm function in practice, respectively. The VBRM refers to (4.4) - (4.6).

The unknown coefficients $\boldsymbol{\sigma}$ and $\boldsymbol{\tau}$ can be estimated by maximum likelihood methods, and the log-likelihood function and Fisher information matrix for the VBRM have been derived by Ferrari and Cribari-Neto [121] and Cribari-Neto and Zeilis [138]. More details of VBRM can be found in the literatures [126, 127, 139, 136, 137, 140]. We omit those details here.

4.2.2 Zero-or-one inflated beta regression model

Oftentimes, data include zeros or ones. When this is the case, the VBRM does not provide a satisfactory description of the data. The zero-or-one inflated beta regression model might be a better choice.

A general class of zero-or-one inflated beta regression models can be defined as follows. Let y_1, \dots, y_n be independent random variables such that each y_t , for $t = 1, \dots, n$, has probability density function in (4.2) with parameters α_t , μ_t , and ϕ_t ,

$$f_c(y_t; \alpha_t, \mu_t, \phi_t) = \begin{cases} \alpha_t, & \text{if } y_t = c \\ (1 - \alpha_t)\pi(y_t; \mu_t, \phi_t), & \text{if } y_t \in (0, 1) \end{cases}, \quad (4.7)$$

where c is either zero or one. We assume that α_t , μ_t , and ϕ_t are defined as

$$h_1(\mu_t) = \mathbf{Z}_t^T \boldsymbol{\sigma} = \eta_{1t}, \quad (4.8)$$

$$h_2(\phi_t) = \mathbf{Q}_t^T \boldsymbol{\tau} = \eta_{2t}, \quad (4.9)$$

$$h_3(\alpha_t) = \mathbf{V}_t^T \boldsymbol{\rho} = \eta_{3t}, \quad (4.10)$$

where $\mathbf{Z}_t = (1, Z_{t1}, \dots, Z_{tm})^T \in \mathbb{R}^{m+1}$, $\mathbf{Q}_t = (1, Q_{t1}, \dots, Q_{tl})^T \in \mathbb{R}^{l+1}$, and $\mathbf{V}_t = (1, V_{t1}, \dots, V_{tr})^T \in \mathbb{R}^{r+1}$ are observations on known covariates, $\boldsymbol{\sigma} = (\sigma_0, \sigma_1, \dots, \sigma_m)^T$, $\boldsymbol{\tau} = (\tau_0, \tau_1, \dots, \tau_l)^T$, and $\boldsymbol{\rho} = (\rho_0, \rho_1, \dots, \rho_r)^T$ are vectors of unknown coefficients, $\boldsymbol{\eta}_1 = (\eta_{11}, \dots, \eta_{1n})^T$, $\boldsymbol{\eta}_2 = (\eta_{21}, \dots, \eta_{2n})^T$, and $\boldsymbol{\eta}_3 = (\eta_{31}, \dots, \eta_{3n})^T$ are predictor vectors, and $h_1(\cdot)$, $h_2(\cdot)$, and $h_3(\cdot)$ are twice continuously differentiable link functions.

The maximum likelihood inference for $\boldsymbol{\sigma}$, $\boldsymbol{\tau}$, and $\boldsymbol{\rho}$ have been discussed by Ospina in [130], which is omitted here.

4.2.3 Regularized generalized linear models

In recent years, there has been an enormous amount of research activity devoted to regularization methods [141, 142, 143, 144, 145, 146, 147]. The Lasso [148] is a popular method for regression that uses an ℓ_1 penalty to achieve a sparse solution. The elastic net [149] uses a penalty that is part ℓ_1 , part ℓ_2 . It includes lasso as a special case, and is particularly useful in the situation where there are many correlated predictor variables.

Fast algorithms have been developed for estimation of generalized linear models with convex penalties [150, 151]. The basic ideas of the algorithms are: 1) Forming quadratic approximation to the log-likelihood function using Taylor expansion about current estimates; 2) reformatting the quadratic approximation into the form of weighted least squares; 3) updating the parameters by solving the penalized weighted least-squares problems.

Let us illustrate this procedure using the simple two-class logistic regression. Suppose that the response variables g_1, \dots, g_N are binary, taking values $\{0, 1\}$. The logistic regres-

sion model represents the probability of response being 1 through a linear function of the predictors:

$$P(g_i = 1|\mathbf{x}_i) = \frac{1}{1 + \exp(-\mathbf{x}_i^T \boldsymbol{\beta})},$$

where $\mathbf{x}_i = (1, x_{i1}, \dots, x_{is})^T \in \mathbb{R}^{s+1}$ is a vector of predictors, $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_s)^T \in \mathbb{R}^{s+1}$ is a vector of unknown coefficients that is sparse, and $i = 1, \dots, N$. It is not hard to check that the log-likelihood function can be written in the following explicit form,

$$\ell(\boldsymbol{\beta}) = \frac{1}{N} \sum_{i=1}^N \left\{ g_i \cdot \mathbf{x}_i^T \boldsymbol{\beta} - \log(1 + e^{\mathbf{x}_i^T \boldsymbol{\beta}}) \right\}. \quad (4.11)$$

We then fit this model by regularized maximum binomial likelihood, and the unknown $\boldsymbol{\beta}$ coefficients are estimated by maximizing the penalized log likelihood

$$\max_{\boldsymbol{\beta} \in \mathbb{R}^{s+1}} \{ \ell(\boldsymbol{\beta}) - \lambda \text{Pen}(\boldsymbol{\beta}) \}. \quad (4.12)$$

The term $\text{Pen}(\boldsymbol{\beta})$ is any convex penalty, such as lasso and elastic-net.

The approach in [150] solves (4.12) in the following three steps:

1) Forming a quadratic approximation to the log-likelihood function (4.11):

If the current estimation is $\tilde{\boldsymbol{\beta}} \in \mathbb{R}^{s+1}$, then based on Taylor expansion, the log-likelihood function (4.11) can be approximated by

$$\ell_Q(\boldsymbol{\beta}) = \ell(\tilde{\boldsymbol{\beta}}) + (\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}})^T \left. \frac{\partial \ell}{\partial \boldsymbol{\beta}} \right|_{\tilde{\boldsymbol{\beta}}} + \frac{1}{2} (\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}})^T \left. \frac{\partial^2 \ell}{\partial \boldsymbol{\beta}^2} \right|_{\tilde{\boldsymbol{\beta}}} (\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}) + O(\|\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}\|_2^2). \quad (4.13)$$

It can be calculated that

$$\begin{aligned} \left. \frac{\partial \ell}{\partial \boldsymbol{\beta}} \right|_{\tilde{\boldsymbol{\beta}}} &= \frac{1}{N} X^T (\mathbf{g} - \tilde{\mathbf{p}}), \\ \left. \frac{\partial^2 \ell}{\partial \boldsymbol{\beta}^2} \right|_{\tilde{\boldsymbol{\beta}}} &= -\frac{1}{N} X^T W X, \end{aligned}$$

where $X = [\mathbf{x}_1, \dots, \mathbf{x}_N] \in \mathbb{R}^{(s+1) \times N}$, $\mathbf{g} = (g_1, \dots, g_N)^T$, $\tilde{\mathbf{p}} = (\tilde{p}_1, \dots, \tilde{p}_N)^T$ with $\tilde{p}_i =$

$P(g_i = 1|\mathbf{x}_i, \tilde{\boldsymbol{\beta}})$, and $W = \text{diag}(w_1, \dots, w_N) \in \mathbb{R}^{N \times N}$ with $w_i = \tilde{p}_i(1 - \tilde{p}_i)$. Therefore the quadratic approximation (4.13) can be rewritten into:

$$\ell_Q(\boldsymbol{\beta}) = \ell(\tilde{\boldsymbol{\beta}}) + \frac{1}{N} \sum_{i=1}^N (g_i - \tilde{p}_i) \mathbf{x}_i^T (\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}) - \frac{1}{2N} \sum_{i=1}^N \left(\mathbf{x}_i^T (\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}) \right)^2 w_i + O(\|\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}\|_2^2). \quad (4.14)$$

2) Reformatting (4.14) into weighted least squares:

It is well known that the Newton algorithm for maximizing the unpenalized log-likelihood function (4.11) amounts to iteratively reweighted least squares. To see this, let us reformat (4.14) into the following weighted least squares,

$$\ell_Q(\boldsymbol{\beta}) = -\frac{1}{2N} \sum_{i=1}^N w_i \left(\mathbf{x}_i^T (\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}) + \frac{g_i - \tilde{p}_i}{w_i} \right)^2 + C(\tilde{\boldsymbol{\beta}}) \quad (4.15)$$

$$= -\frac{1}{2N} \sum_{i=1}^N w_i (\bar{g}_i - \mathbf{x}_i^T \boldsymbol{\beta})^2 + C(\tilde{\boldsymbol{\beta}}), \quad (4.16)$$

where $w_i = \tilde{p}_i(1 - \tilde{p}_i)$ is the weight and $\bar{g}_i = \mathbf{x}_i^T \tilde{\boldsymbol{\beta}} + \frac{g_i - \tilde{p}_i}{w_i}$ is the working response. The last term is constant. The Newton update is obtained by minimizing ℓ_Q .

3) Updating the parameters by solving the penalized weighted least-squares problems:

Now, let us solve problem (4.12). For each value of λ , we compute (4.16) about the current estimates $\tilde{\boldsymbol{\beta}}$, then solve the penalized weighted least-squares problem:

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^{s+1}} \left\{ \frac{1}{2N} \sum_{i=1}^N w_i (\bar{g}_i - \mathbf{x}_i^T \boldsymbol{\beta})^2 + \lambda \text{Pen}(\boldsymbol{\beta}) \right\}.$$

Following the above ideas, we will derive similar algorithm to solve the regularized zero-and-one inflated beta regression in next Section.

4.3 Our proposed regularized beta regression

In previous Section, we introduced the varying dispersion beta regression model and zero-and-one inflated beta regression model, however, they are not suitable for modeling data that

contains both zeros and ones. For this situation, we propose a zero-and-one inflated beta regression model in Section 4.3.1, provide its likelihood inference in Section 4.3.2, consider variable selection for the proposed model in Section 4.3.3, and define the classification rule in Section 4.3.4.

4.3.1 Zero-and-one inflated beta regression model

The zero-and-one inflated beta regression model is defined as follows. Suppose we have a training data of the form $\{y_t, \mathbf{Z}_t, \mathbf{Q}_t, \mathbf{V}_t, \mathbf{X}_t\}$, for $t = 1, \dots, n$, where $y_t \in [0, 1]$ is the response variable, $\mathbf{Z}_t = (1, Z_{t1}, \dots, Z_{tm})^T \in \mathbb{R}^{m+1}$, $\mathbf{Q}_t = (1, Q_{t1}, \dots, Q_{tl})^T \in \mathbb{R}^{l+1}$, $\mathbf{V}_t = (1, V_{t1}, \dots, V_{tr})^T \in \mathbb{R}^{r+1}$, and $\mathbf{X}_t = (1, X_{t1}, \dots, X_{tk})^T \in \mathbb{R}^{k+1}$ are known covariates where 1 corresponds to the intercept term. The zero-and-one inflated beta regression model can be defined as follows. Assume that each y_t , for $t = 1, \dots, n$ has the probability density function in (4.46) with parameters $0 < \mu_t < 1$, $\phi_t > 0$, $0 < \alpha_t < 1$, and $0 < \gamma_t < 1$,

$$f(y_t; \mu_t, \phi_t, \alpha_t, \gamma_t) = \begin{cases} \alpha_t(1 - \gamma_t), & \text{if } y_t = 0 \\ (1 - \alpha_t)\pi(y_t; \mu_t, \phi_t), & \text{if } y_t \in (0, 1) \\ \alpha_t\gamma_t, & \text{if } y_t = 1 \end{cases}, \quad (4.17)$$

which is a mixture between a beta distribution and a Bernoulli distribution. To be more concrete, the cumulative distribution of y_t can be written as

$$F(y_t; \mu_t, \phi_t, \alpha_t, \gamma_t) = \alpha_t \mathbf{Ber}(y_t; \gamma_t) + (1 - \alpha_t) \Pi(y_t; \mu_t, \phi_t),$$

where $\mathbf{Ber}(\cdot; \gamma)$ represents the cumulative distribution of a Bernoulli random variable with parameter γ , and $\Pi(\cdot; \mu, \phi)$ is the cumulative distribution of beta random variable. Such mixture distribution is named as zero-and-one inflated beta distribution by Ospina [133], based on which we call our proposed model the zero-and-one inflated beta regression.

After some algebra, the mean and variance of y_t , for $t = 1, \dots, n$ are

$$\begin{aligned}\mathbb{E}(y_t) &= \alpha_t \gamma_t + (1 - \alpha_t) \mu_t, \\ \text{var}(y_t) &= \alpha_t \gamma_t (1 - \gamma_t) + (1 - \alpha_t) \frac{\mu_t (1 - \mu_t)}{\phi_t + 1} + \alpha_t (1 - \alpha_t) (\gamma_t - \mu_t)^2.\end{aligned}$$

Note that $\mathbb{E}(y_t)$ is the weighted average of the mean of Bernoulli distribution and mean of beta distribution with weights α_t and $1 - \alpha_t$. Besides, $\mathbb{E}(y_t | y_t \in (0, 1)) = \mu_t$ and $\text{var}(y_t) = \mu_t (1 - \mu_t) (\phi_t + 1)$. Other properties of zero-and-one inflated beta distribution can be found in [133].

We then assume the unknown parameters can be expressed as linear combinations of covariates through link functions,

$$h_1(\mu_t) = \mathbf{Z}_t^T \boldsymbol{\sigma} = \eta_{1t}, \quad (4.18)$$

$$h_2(\phi_t) = \mathbf{Q}_t^T \boldsymbol{\tau} = \eta_{2t}, \quad (4.19)$$

$$h_3(\alpha_t) = \mathbf{V}_t^T \boldsymbol{\rho} = \eta_{3t}, \quad (4.20)$$

$$h_4(\gamma_t) = \mathbf{X}_t^T \boldsymbol{\beta} = \eta_{4t}, \quad (4.21)$$

where $\boldsymbol{\sigma} = (\sigma_0, \sigma_1, \dots, \sigma_m)^T$, $\boldsymbol{\tau} = (\tau_0, \tau_1, \dots, \tau_l)^T$, $\boldsymbol{\rho} = (\rho_0, \rho_1, \dots, \rho_r)^T$, and $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_k)^T$ are vectors of unknown coefficients, $\boldsymbol{\eta}_1 = (\eta_{11}, \dots, \eta_{1n})^T$, $\boldsymbol{\eta}_2 = (\eta_{21}, \dots, \eta_{2n})^T$, $\boldsymbol{\eta}_3 = (\eta_{31}, \dots, \eta_{3n})^T$, and $\boldsymbol{\eta}_4 = (\eta_{41}, \dots, \eta_{4n})^T$ are predictor vectors, $h_1 : (0, 1) \rightarrow \mathbb{R}$, $h_2 : (0, \infty) \rightarrow \mathbb{R}$, $h_3 : (0, 1) \rightarrow \mathbb{R}$, and $h_4 : (0, 1) \rightarrow \mathbb{R}$ are twice continuously differentiable link functions. Various different links can be used, for μ_t , α_t , and γ_t , we may choose logit link $\log(x/1-x)$, probit link $\Phi^{-1}(x)$, complementary log-log link $\log(-\log(1-x))$, and log-log link $-\log(-\log(x))$. Possible links for ϕ_t are log link $\log(x)$ and square-root link \sqrt{x} .

Notice that the role of covariates \mathbf{Z}_t , \mathbf{Q}_t , \mathbf{V}_t , and \mathbf{X}_t and their corresponding coefficients $\boldsymbol{\sigma}$, $\boldsymbol{\tau}$, $\boldsymbol{\rho}$, and $\boldsymbol{\beta}$ are clear in model (4.17) - (4.21). \mathbf{Z}_t and $\boldsymbol{\sigma}$ control $\mathbb{E}(y_t | y_t \in (0, 1))$,

\mathbf{Q} and $\boldsymbol{\tau}$ influence the precision of the distribution of y_t given that $y_t \in (0, 1)$, \mathbf{V}_t and $\boldsymbol{\rho}$ affect the probability of $y_t \in \{0, 1\}$, and \mathbf{X}_t and $\boldsymbol{\beta}$ determines the probability of y_t being 1 given that $y_t \in \{0, 1\}$. Here we assume the covariates \mathbf{Z}_t , \mathbf{Q}_t , \mathbf{V}_t , and \mathbf{X}_t are different, however, they can be the same set of covariates in practice.

Throughout this chapter, we assume that the response is constrained to the standard unit interval $[0, 1]$. The model, however, is still useful for situations where the response is restricted to the interval $[a, b]$, where a and b are known scalars, $a < b$. In this case, one would model $(y - a)/(b - a)$ instead of modeling y directly.

4.3.2 Likelihood inference

In this section, we will derive the score function and fisher information matrix for unknown parameters $\boldsymbol{\sigma}$, $\boldsymbol{\tau}$, $\boldsymbol{\rho}$, and $\boldsymbol{\beta}$. The results will then be used in the next section to develop the variable selection algorithm.

The likelihood function for $\boldsymbol{\sigma} \in \mathbb{R}^{(m+1)}$, $\boldsymbol{\tau} \in \mathbb{R}^{(l+1)}$, $\boldsymbol{\rho} \in \mathbb{R}^{(r+1)}$, and $\boldsymbol{\beta} \in \mathbb{R}^{(k+1)}$ given y_1, \dots, y_n is

$$\begin{aligned} L(\boldsymbol{\sigma}, \boldsymbol{\tau}, \boldsymbol{\rho}, \boldsymbol{\beta}) &= \prod_{t=1}^n f(y_t; \mu_t, \phi_t, \alpha_t, \gamma_t) \\ &= L_1(\boldsymbol{\rho})L_2(\boldsymbol{\beta})L_3(\boldsymbol{\sigma}, \boldsymbol{\tau}), \end{aligned}$$

with

$$\begin{aligned} L_1(\boldsymbol{\rho}) &= \prod_{t=1}^n \alpha_t^{\mathbb{1}_{\{0,1\}}(y_t)} (1 - \alpha_t)^{1 - \mathbb{1}_{\{0,1\}}(y_t)}, \\ L_2(\boldsymbol{\beta}) &= \prod_{t: y_t \in \{0,1\}} \gamma_t^{y_t} (1 - \gamma_t)^{1 - y_t}, \\ L_3(\boldsymbol{\sigma}, \boldsymbol{\tau}) &= \prod_{t: y_t \in (0,1)} \pi(y_t; \mu_t, \phi_t). \end{aligned}$$

The notation $\mathbb{1}_{\{0,1\}}(y_t)$ is an indicator function that equals to 1 if $y_t \in \{0, 1\}$ and 0 otherwise. The function $\pi(y_t; \mu_t, \phi_t)$ is the density function of beta distribution given in (4.4).

The parameters $\mu_t = h_1^{-1}(\eta_{1t})$, $\phi_t = h_2^{-1}(\eta_{2t})$, $\alpha_t = h_3^{-1}(\eta_{3t})$, and $\gamma_t = h_4^{-1}(\eta_{4t})$, as defined in (4.18) - (4.21), are functions of $\boldsymbol{\sigma}, \boldsymbol{\tau}, \boldsymbol{\rho}, \boldsymbol{\beta}$, respectively.

The likelihood function above have three terms: The first term only depends on $\boldsymbol{\rho}$ regarding to the discrete components 0 and 1; the second term only contains $\boldsymbol{\beta}$ corresponding to the probability of being 1 for the discrete components; the third term is based on $\boldsymbol{\sigma}$ and $\boldsymbol{\tau}$ only, associating with continuous component. Since the parameters are separable, the maximum likelihood inferences can be performed separately, as well. More details of how the likelihood function for zero-and-one beta distribution is derived can be found in [133].

The log-likelihood function can be written as

$$\ell(\boldsymbol{\sigma}, \boldsymbol{\tau}, \boldsymbol{\rho}, \boldsymbol{\beta}) = \frac{1}{n} (\ell_1(\boldsymbol{\rho}) + \ell_2(\boldsymbol{\beta}) + \ell_3(\boldsymbol{\sigma}, \boldsymbol{\tau})), \quad (4.22)$$

with

$$\begin{aligned} \ell_1(\boldsymbol{\rho}) &= \sum_{t=1}^n (\mathbb{1}_{\{0,1\}}(y_t) \log(\alpha_t) + (1 - \mathbb{1}_{\{0,1\}}(y_t)) \log(1 - \alpha_t)) \triangleq \sum_{t=1}^n \ell_{1t}(\alpha_t), \\ \ell_2(\boldsymbol{\beta}) &= \sum_{t:y_t \in \{0,1\}} (y_t \log(\gamma_t) + (1 - y_t) \log(1 - \gamma_t)) \triangleq \sum_{t:y_t \in \{0,1\}} \ell_{2t}(\gamma_t), \\ \ell_3(\boldsymbol{\sigma}, \boldsymbol{\tau}) &= \sum_{t:y_t \in (0,1)} \left(\log \Gamma(\phi_t) - \log \Gamma(\phi_t \mu_t) - \log \Gamma(\phi_t(1 - \mu_t)) + (\phi_t \mu_t - 1) \log y_t + \right. \\ &\quad \left. (\phi_t(1 - \mu_t) - 1) \log(1 - y_t) \right) \triangleq \sum_{t:y_t \in (0,1)} \ell_{3t}(\mu_t, \phi_t). \end{aligned}$$

The score functions are obtained by differentiating the log-likelihood function with respect to each of the unknown parameters $\boldsymbol{\sigma}, \boldsymbol{\tau}, \boldsymbol{\rho}, \boldsymbol{\beta}$, respectively. The details are postponed in Section 4.6.

$$\begin{aligned} U_{\boldsymbol{\sigma}}(\boldsymbol{\sigma}, \boldsymbol{\tau}) &= \frac{1}{n} \mathbf{Z}^T P_{\boldsymbol{\sigma}}, & U_{\boldsymbol{\tau}}(\boldsymbol{\sigma}, \boldsymbol{\tau}) &= \frac{1}{n} \mathbf{Q}^T P_{\boldsymbol{\tau}}, \\ U_{\boldsymbol{\rho}}(\boldsymbol{\rho}) &= \frac{1}{n} \mathbf{V}^T P_{\boldsymbol{\rho}}, & U_{\boldsymbol{\beta}}(\boldsymbol{\beta}) &= \frac{1}{n} \mathbf{X}^T P_{\boldsymbol{\beta}}, \end{aligned} \quad (4.23)$$

where $\mathbf{Z}^T = (\mathbf{Z}_1, \dots, \mathbf{Z}_n) \in \mathbb{R}^{(m+1) \times n}$, $\mathbf{Q}^T = (\mathbf{Q}_1, \dots, \mathbf{Q}_n) \in \mathbb{R}^{(l+1) \times n}$, $\mathbf{V}^T = (\mathbf{V}_1, \dots, \mathbf{V}_n) \in \mathbb{R}^{(r+1) \times n}$, and $\mathbf{X}^T = (\mathbf{X}_1, \dots, \mathbf{X}_n) \in \mathbb{R}^{(k+1) \times n}$ are known covariates, and $P_{\boldsymbol{\sigma}} \in \mathbb{R}^n$,

$P_\tau \in \mathbb{R}^n$, $P_\rho \in \mathbb{R}^n$, and $P_\beta \in \mathbb{R}^n$ are vectors depending on unknown parameters, which are provided in (4.51) - (4.54) in Section 4.6. The maximum likelihood estimators (MLEs) of σ , τ , ρ , β can be obtained as the solutions of $U_\sigma(\sigma, \tau) = 0$, $U_\tau(\sigma, \tau) = 0$, $U_\rho(\rho) = 0$, and $U_\beta(\beta) = 0$. No closed form expressions for these estimators exist, and their computations can be performed numerically using a nonlinear optimization algorithm, such as Newton-Raphson, Fisher's scoring. More details can be found in [130].

The Fisher-information matrix is useful for computing asymptotic standard errors of MLEs, see more details in [133]. It is provided as follows,

$$\begin{bmatrix} K_{\sigma\sigma} & K_{\sigma\tau} & 0 & 0 \\ K_{\tau\sigma} & K_{\tau\tau} & 0 & 0 \\ 0 & 0 & K_{\rho\rho} & 0 \\ 0 & 0 & 0 & K_{\beta\beta} \end{bmatrix},$$

where the elements are $K_{\sigma\sigma} = \frac{1}{n} \mathbf{Z}^T W_1 \mathbf{Z}$, $K_{\tau\tau} = \frac{1}{n} \mathbf{Q}^T W_2 \mathbf{Q}$, $K_{\sigma\tau} = K_{\tau\sigma}^T = \frac{1}{n} \mathbf{Z}^T W_{12} \mathbf{Q}$, $K_{\rho\rho} = \frac{1}{n} \mathbf{V}^T W_3 \mathbf{V}$, and $K_{\beta\beta} = \frac{1}{n} \mathbf{X}^T W_4 \mathbf{X}$. The W_1 , W_2 , W_{12} , W_3 , and W_4 are $n \times n$ matrixes that depend on unknown parameters and are provided in (4.66) - (4.70) in Section 4.6.

4.3.3 Variable selection algorithm

Variable selection is important for regression models, since it can improve the prediction accuracy and enhance model interpretability with parsimonious representations. In Section 4.2.3, we introduced the variable selection methods for generalized linear models. Following that idea, in this section, we develop a penalized likelihood-based method to select important explanatory variables for our proposed zero-and-one inflated beta regression model.

We fit our proposed model in (4.17)- (4.21) by regularized maximum likelihood, and

maximize the penalized log likelihood

$$\max_{\boldsymbol{\sigma}, \boldsymbol{\tau}, \boldsymbol{\rho}, \boldsymbol{\beta}} \{ \ell(\boldsymbol{\sigma}, \boldsymbol{\tau}, \boldsymbol{\rho}, \boldsymbol{\beta}) - \lambda_{\boldsymbol{\sigma}} \text{Pen}(\boldsymbol{\sigma}) - \lambda_{\boldsymbol{\tau}} \text{Pen}(\boldsymbol{\tau}) - \lambda_{\boldsymbol{\rho}} \text{Pen}(\boldsymbol{\rho}) - \lambda_{\boldsymbol{\beta}} \text{Pen}(\boldsymbol{\beta}) \}, \quad (4.24)$$

where $\ell(\boldsymbol{\sigma}, \boldsymbol{\tau}, \boldsymbol{\rho}, \boldsymbol{\beta})$ is the log-likelihood function in (4.22), $\boldsymbol{\lambda} = (\lambda_{\boldsymbol{\sigma}}, \lambda_{\boldsymbol{\tau}}, \lambda_{\boldsymbol{\rho}}, \lambda_{\boldsymbol{\beta}})$ are regularization parameters which can be chosen by a data-driven criterion such as cross validation and BIC criterion [152], and $\text{Pen}(\cdot)$ is a convex penalty function. Note that we assume the penalty function is the same for all parameters, however, the methodology we propose can also be applied to the case where the penalty functions are different. To be more concrete, we consider the following two different penalties:

$$\text{Pen}(\boldsymbol{\beta}) = \|\boldsymbol{\beta}\|_{\ell_1}, \quad (\text{lasso}) \quad (4.25)$$

$$\text{Pen}(\boldsymbol{\beta}) = \frac{1}{4} \|\boldsymbol{\beta}\|_{\ell_2}^2 + \frac{1}{2} \|\boldsymbol{\beta}\|_{\ell_1}. \quad (\text{elastic-net}) \quad (4.26)$$

The elastic-net penalty is a weighted average of the ridge-regression penalty and the lasso penalty, and here we set the weight to be 1/2.

To solve the optimization problem in (4.24), we borrow the ideas from generalized linear models with convex penalty in [150], and solve the problem in the following three steps:

- 1) Forming partial quadratic approximations to the log-likelihood function (4.22):

Unlike the simple two-class logistic regression model that we illustrated in Section 4.2.3, here the form of quadratic approximation to the log-likelihood can be tedious, due to the fact that we have four different unknown parameters: $\boldsymbol{\sigma}$, $\boldsymbol{\tau}$, $\boldsymbol{\rho}$, and $\boldsymbol{\beta}$. However, we can avoid these complexities by forming a partial quadratic approximation to the log-likelihood (4.22), allowing only one parameter to vary at a time. In fact, such idea has been applied to solve regularized multinomial regression, see [150].

If the current estimations are $\tilde{\boldsymbol{\sigma}} = (\tilde{\sigma}_0, \dots, \tilde{\sigma}_m)^T$, $\tilde{\boldsymbol{\tau}} = (\tilde{\tau}_0, \dots, \tilde{\tau}_l)^T$, $\tilde{\boldsymbol{\rho}} = (\tilde{\rho}_0, \dots, \tilde{\rho}_r)^T$, and $\tilde{\boldsymbol{\beta}} = (\tilde{\beta}_0, \dots, \tilde{\beta}_k)^T$, then we can calculate $\tilde{\boldsymbol{\mu}} = (\tilde{\mu}_1, \dots, \tilde{\mu}_n)^T$, $\tilde{\boldsymbol{\phi}} = (\tilde{\phi}_1, \dots, \tilde{\phi}_n)^T$, $\tilde{\boldsymbol{\alpha}} =$

$(\tilde{\alpha}_1, \dots, \tilde{\alpha}_n)^T$, $\tilde{\gamma} = (\tilde{\gamma}_1, \dots, \tilde{\gamma}_n)^T$ in (4.18) - (4.21). Based on Taylor expansion, the four partial quadratic approximations to the log-likelihood function (4.22) are:

$$\begin{aligned}
l_{\sigma} &= \ell(\tilde{\sigma}, \tilde{\tau}, \tilde{\rho}, \tilde{\beta}) + (\sigma - \tilde{\sigma})^T \frac{\partial \ell}{\partial \sigma} \Big|_{\tilde{\sigma}, \tilde{\tau}, \tilde{\rho}, \tilde{\beta}} + \frac{1}{2} (\sigma - \tilde{\sigma})^T \frac{\partial^2 \ell}{\partial \sigma^2} \Big|_{\tilde{\sigma}, \tilde{\tau}, \tilde{\rho}, \tilde{\beta}} (\sigma - \tilde{\sigma}) + O(\|\sigma - \tilde{\sigma}\|_2^2), \\
l_{\tau} &= \ell(\tilde{\sigma}, \tilde{\tau}, \tilde{\rho}, \tilde{\beta}) + (\tau - \tilde{\tau})^T \frac{\partial \ell}{\partial \tau} \Big|_{\tilde{\sigma}, \tilde{\tau}, \tilde{\rho}, \tilde{\beta}} + \frac{1}{2} (\tau - \tilde{\tau})^T \frac{\partial^2 \ell}{\partial \tau^2} \Big|_{\tilde{\sigma}, \tilde{\tau}, \tilde{\rho}, \tilde{\beta}} (\tau - \tilde{\tau}) + O(\|\tau - \tilde{\tau}\|_2^2), \\
l_{\rho} &= \ell(\tilde{\sigma}, \tilde{\tau}, \tilde{\rho}, \tilde{\beta}) + (\rho - \tilde{\rho})^T \frac{\partial \ell}{\partial \rho} \Big|_{\tilde{\sigma}, \tilde{\tau}, \tilde{\rho}, \tilde{\beta}} + \frac{1}{2} (\rho - \tilde{\rho})^T \frac{\partial^2 \ell}{\partial \rho^2} \Big|_{\tilde{\sigma}, \tilde{\tau}, \tilde{\rho}, \tilde{\beta}} (\rho - \tilde{\rho}) + O(\|\rho - \tilde{\rho}\|_2^2), \\
l_{\beta} &= \ell(\tilde{\sigma}, \tilde{\tau}, \tilde{\rho}, \tilde{\beta}) + (\beta - \tilde{\beta})^T \frac{\partial \ell}{\partial \beta} \Big|_{\tilde{\sigma}, \tilde{\tau}, \tilde{\rho}, \tilde{\beta}} + \frac{1}{2} (\beta - \tilde{\beta})^T \frac{\partial^2 \ell}{\partial \beta^2} \Big|_{\tilde{\sigma}, \tilde{\tau}, \tilde{\rho}, \tilde{\beta}} (\beta - \tilde{\beta}) + O(\|\beta - \tilde{\beta}\|_2^2),
\end{aligned}$$

The first derivatives of log-likelihood function with respect to those four parameters are provided in (4.47) - (4.50), and the secondary derivatives are given in (4.55) - (4.55). Plugging in the calculated first and secondary derivatives, we can rewrite the partial quadratic approximations into:

$$\begin{aligned}
l_{\sigma} &= \ell(\tilde{\sigma}, \tilde{\tau}, \tilde{\rho}, \tilde{\beta}) + \frac{1}{n} \sum_{t=1}^n \tilde{p}_{\sigma t} \mathbf{Z}_t^T (\sigma - \tilde{\sigma}) - \frac{1}{2n} \sum_{t=1}^n (\mathbf{Z}_t^T (\sigma - \tilde{\sigma}))^2 \tilde{w}_{\sigma t} + O(\|\sigma - \tilde{\sigma}\|_2^2), \\
l_{\tau} &= \ell(\tilde{\sigma}, \tilde{\tau}, \tilde{\rho}, \tilde{\beta}) + \frac{1}{n} \sum_{t=1}^n \tilde{p}_{\tau t} \mathbf{Q}_t^T (\tau - \tilde{\tau}) - \frac{1}{2n} \sum_{t=1}^n (\mathbf{Q}_t^T (\tau - \tilde{\tau}))^2 \tilde{w}_{\tau t} + O(\|\tau - \tilde{\tau}\|_2^2), \\
l_{\rho} &= \ell(\tilde{\sigma}, \tilde{\tau}, \tilde{\rho}, \tilde{\beta}) + \frac{1}{n} \sum_{t=1}^n \tilde{p}_{\rho t} \mathbf{V}_t^T (\rho - \tilde{\rho}) - \frac{1}{2n} \sum_{t=1}^n (\mathbf{V}_t^T (\rho - \tilde{\rho}))^2 \tilde{w}_{\rho t} + O(\|\rho - \tilde{\rho}\|_2^2), \\
l_{\beta} &= \ell(\tilde{\sigma}, \tilde{\tau}, \tilde{\rho}, \tilde{\beta}) + \frac{1}{n} \sum_{t=1}^n \tilde{p}_{\beta t} \mathbf{X}_t^T (\beta - \tilde{\beta}) - \frac{1}{2n} \sum_{t=1}^n (\mathbf{X}_t^T (\beta - \tilde{\beta}))^2 \tilde{w}_{\beta t} + O(\|\beta - \tilde{\beta}\|_2^2),
\end{aligned}$$

where $\tilde{p}_{\sigma t} = p_{\sigma t}(\tilde{\mu}_t, \tilde{\phi}_t)$, $\tilde{p}_{\tau t} = p_{\tau t}(\tilde{\mu}_t, \tilde{\phi}_t)$, $\tilde{p}_{\rho t} = p_{\rho t}(\tilde{\alpha}_t)$, and $\tilde{p}_{\beta t} = p_{\beta t}(\tilde{\gamma}_t)$ can be found in (4.51) - (4.54), and $\tilde{w}_{\sigma t} = w_{\sigma t}(\tilde{\mu}_t, \tilde{\phi}_t)$, $\tilde{w}_{\tau t} = w_{\tau t}(\tilde{\mu}_t, \tilde{\phi}_t)$, $\tilde{w}_{\rho t} = w_{\rho t}(\tilde{\alpha}_t)$, and $\tilde{w}_{\beta t} = w_{\beta t}(\tilde{\gamma}_t)$ are given in (4.61) - (4.65).

2) Reformatting partial quadratic approximations into weighted least squares:

It is not hard to show that those partial quadratic approximations can be reformatted into

the following weighted least squares,

$$\ell_{\sigma} = -\frac{1}{2n} \sum_{t=1}^n \tilde{w}_{\sigma t} (\bar{y}_{t\sigma} - \mathbf{Z}_i^T \boldsymbol{\sigma})^2 + C_1(\tilde{\boldsymbol{\sigma}}, \tilde{\boldsymbol{\tau}}, \tilde{\boldsymbol{\rho}}, \tilde{\boldsymbol{\beta}}), \quad \bar{y}_{t\sigma} = \mathbf{Z}_i^T \tilde{\boldsymbol{\sigma}} + \frac{\tilde{p}_{\sigma t}}{\tilde{w}_{\sigma t}}, \quad (4.27)$$

$$\ell_{\tau} = -\frac{1}{2n} \sum_{t=1}^n \tilde{w}_{\tau t} (\bar{y}_{t\tau} - \mathbf{Q}_i^T \boldsymbol{\tau})^2 + C_2(\tilde{\boldsymbol{\sigma}}, \tilde{\boldsymbol{\tau}}, \tilde{\boldsymbol{\rho}}, \tilde{\boldsymbol{\beta}}), \quad \bar{y}_{t\tau} = \mathbf{Q}_i^T \tilde{\boldsymbol{\tau}} + \frac{\tilde{p}_{\tau t}}{\tilde{w}_{\tau t}}, \quad (4.28)$$

$$\ell_{\rho} = -\frac{1}{2n} \sum_{t=1}^n \tilde{w}_{\rho t} (\bar{y}_{t\rho} - \mathbf{V}_i^T \boldsymbol{\rho})^2 + C_3(\tilde{\boldsymbol{\sigma}}, \tilde{\boldsymbol{\tau}}, \tilde{\boldsymbol{\rho}}, \tilde{\boldsymbol{\beta}}), \quad \bar{y}_{t\rho} = \mathbf{V}_i^T \tilde{\boldsymbol{\rho}} + \frac{\tilde{p}_{\rho t}}{\tilde{w}_{\rho t}}, \quad (4.29)$$

$$\ell_{\beta} = -\frac{1}{2n} \sum_{t=1}^n \tilde{w}_{\beta t} (\bar{y}_{t\beta} - \mathbf{X}_i^T \boldsymbol{\beta})^2 + C_4(\tilde{\boldsymbol{\sigma}}, \tilde{\boldsymbol{\tau}}, \tilde{\boldsymbol{\rho}}, \tilde{\boldsymbol{\beta}}), \quad \bar{y}_{t\beta} = \mathbf{X}_i^T \tilde{\boldsymbol{\beta}} + \frac{\tilde{p}_{\beta t}}{\tilde{w}_{\beta t}}, \quad (4.30)$$

where $\bar{y}_{t\sigma}$, $\bar{y}_{t\tau}$, $\bar{y}_{t\rho}$, and $\bar{y}_{t\beta}$ are the working responses, and the last terms $C_1 - C_4$ are constants. The Newton updates are obtained by minimizing ℓ_{σ} , ℓ_{τ} , ℓ_{ρ} , and ℓ_{β} in (4.27) - (4.30).

3) Updating the parameters by solving the penalized weighted least-squares problems:

Now, we are ready to solve problem (4.24). For given values of λ_{σ} , λ_{τ} , λ_{ρ} , λ_{β} , we compute (4.27) - (4.30) about the current estimates $\tilde{\boldsymbol{\sigma}}$, $\tilde{\boldsymbol{\tau}}$, $\tilde{\boldsymbol{\rho}}$, and $\tilde{\boldsymbol{\beta}}$, then use any existing convex programming solver, such as CVX in Matlab [153, 154], to solve the penalized weighted least-squares problem:

$$\min_{\boldsymbol{\sigma} \in \mathbb{R}^{m+1}} \left\{ \frac{1}{2n} \sum_{t=1}^n \tilde{w}_{\sigma t} (\bar{y}_{t\sigma} - \mathbf{Z}_i^T \boldsymbol{\sigma})^2 + \lambda_{\sigma} \text{Pen}(\boldsymbol{\sigma}) \right\}, \quad (4.31)$$

$$\min_{\boldsymbol{\tau} \in \mathbb{R}^{l+1}} \left\{ \frac{1}{2n} \sum_{t=1}^n \tilde{w}_{\tau t} (\bar{y}_{t\tau} - \mathbf{Q}_i^T \boldsymbol{\tau})^2 + \lambda_{\tau} \text{Pen}(\boldsymbol{\tau}) \right\}, \quad (4.32)$$

$$\min_{\boldsymbol{\rho} \in \mathbb{R}^{r+1}} \left\{ \frac{1}{2n} \sum_{t=1}^n \tilde{w}_{\rho t} (\bar{y}_{t\rho} - \mathbf{V}_i^T \boldsymbol{\rho})^2 + \lambda_{\rho} \text{Pen}(\boldsymbol{\rho}) \right\}, \quad (4.33)$$

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^{k+1}} \left\{ \frac{1}{2n} \sum_{t=1}^n \tilde{w}_{\beta t} (\bar{y}_{t\beta} - \mathbf{X}_i^T \boldsymbol{\beta})^2 + \lambda_{\beta} \text{Pen}(\boldsymbol{\beta}) \right\}. \quad (4.34)$$

To summarize, the procedure to obtain the unknown parameters through solving (4.24) can be presented in the following nested loops:

Outer loop: Cycle over different combinations of λ_{σ} , λ_{τ} , λ_{ρ} , and λ_{β} .

Middle loop: Update the partial quadratic approximations $\ell_\sigma, \ell_\tau, \ell_\rho, \ell_\beta$ in (4.27) - (4.30) using the current estimations $\tilde{\sigma}, \tilde{\tau}, \tilde{\rho},$ and $\tilde{\beta}$.

Inner loop: Run the CVX on the penalized weighted least squares problems (4.31) - (4.34).

As a result, we could obtain the estimated parameters as a function of $\boldsymbol{\lambda} = (\lambda_\sigma, \lambda_\tau, \lambda_\rho, \lambda_\beta)$, which can be denoted as $\hat{\boldsymbol{\sigma}}(\boldsymbol{\lambda}), \hat{\boldsymbol{\tau}}(\boldsymbol{\lambda}), \hat{\boldsymbol{\rho}}(\boldsymbol{\lambda}),$ and $\hat{\boldsymbol{\beta}}(\boldsymbol{\lambda})$. The optimal $\boldsymbol{\lambda}$ can be obtained to minimize the BIC,

$$\begin{aligned} (\lambda_\sigma^{\text{opt}}, \lambda_\tau^{\text{opt}}, \lambda_\rho^{\text{opt}}, \lambda_\beta^{\text{opt}}) &= \min_{\boldsymbol{\lambda}} \text{BIC}(\boldsymbol{\lambda}), \quad \text{and} \\ \text{BIC}(\boldsymbol{\lambda}) &= -\frac{2}{n} \ell(\hat{\boldsymbol{\sigma}}(\boldsymbol{\lambda}), \hat{\boldsymbol{\tau}}(\boldsymbol{\lambda}), \hat{\boldsymbol{\rho}}(\boldsymbol{\lambda}), \hat{\boldsymbol{\beta}}(\boldsymbol{\lambda})) + df_\lambda \times \frac{\log(n)}{n}, \end{aligned} \quad (4.35)$$

where df_λ is the number of nonzero coefficients of $\hat{\boldsymbol{\sigma}}(\boldsymbol{\lambda}), \hat{\boldsymbol{\tau}}(\boldsymbol{\lambda}), \hat{\boldsymbol{\rho}}(\boldsymbol{\lambda}), \hat{\boldsymbol{\beta}}(\boldsymbol{\lambda})$. In practice, we can use grid search to find the optimal tuning λ 's.

4.3.4 Classification Rule

Assume $\hat{\boldsymbol{\sigma}}, \hat{\boldsymbol{\tau}}, \hat{\boldsymbol{\rho}}, \hat{\boldsymbol{\beta}}$ are the final estimated parameters that are solved through (4.24). Then we can calculate $\hat{\mu}_t = h_1^{-1}(\mathbf{Z}_t^T \hat{\boldsymbol{\sigma}})$, $\hat{\phi}_t = h_2^{-1}(\mathbf{Q}_t^T \hat{\boldsymbol{\tau}})$, $\hat{\alpha}_t = h_3^{-1}(\mathbf{V}_t^T \hat{\boldsymbol{\rho}})$, and $\hat{\gamma}_t = h_4^{-1}(\mathbf{X}_t^T \hat{\boldsymbol{\beta}})$, for $t = 1, \dots, n$. We define the classification rule as:

$$\hat{y}_t = \begin{cases} 1, & \text{if } \hat{\alpha}_t \geq c_1, \quad \hat{\gamma}_t \geq c_2 \\ 0, & \text{if } \hat{\alpha}_t \geq c_1, \quad \hat{\gamma}_t < c_2 \\ \hat{\mu}_t, & \text{if } \hat{\alpha}_t < c_1 \end{cases}, \quad (4.36)$$

where c_1 and c_2 are the cutoffs that can be pre-defined or can be determined through cross validation to minimize the prediction errors.

It is now straight forward to evaluate the prediction errors of the classifier in (4.36) through the testing data and cross-validation.

4.4 Simulation

In this section, we exhibit numerical performances of the regularized zero-and-one inflated beta regression model through simulation studies. The numerical performances include the prediction performance on response variable and the estimation performance on unknown parameters.

The training and testing data are generated from the following model, for $t = 1, \dots, n$,

$$\begin{aligned}
 \log\left(\frac{\mu_t}{1 - \mu_t}\right) &= \mathbf{Z}_t^T \boldsymbol{\sigma}, \quad \mathbf{Z}_t = (1, Z_{t1}, \dots, Z_{tm})^T, \quad \boldsymbol{\sigma} = (\sigma_0, \sigma_1, \dots, \sigma_m)^T, \\
 \log(\phi_t) &= \mathbf{Q}_t^T \boldsymbol{\tau}, \quad \mathbf{Q}_t = (1, Q_{t1}, \dots, Q_{tl})^T, \quad \boldsymbol{\tau} = (\tau_0, \tau_1, \dots, \tau_l)^T, \\
 \log\left(\frac{\alpha_t}{1 - \alpha_t}\right) &= \mathbf{V}_t^T \boldsymbol{\rho}, \quad \mathbf{V}_t = (1, V_{t1}, \dots, V_{tr})^T, \quad \boldsymbol{\rho} = (\rho_0, \rho_1, \dots, \rho_r)^T, \quad (4.37) \\
 \log\left(\frac{\beta_t}{1 - \gamma_t}\right) &= \mathbf{X}_t^T \boldsymbol{\beta}, \quad \mathbf{X}_t = (1, X_{t1}, \dots, X_{tk})^T, \quad \boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_k)^T, \\
 y_t &: \text{Generated using classifier in (4.36), with } c_1 = c_2 = 0.5.
 \end{aligned}$$

We set $m = l = r = k = 8$, sample size $n = 1000$, and the parameters as follows:

$$\begin{aligned}
 \boldsymbol{\sigma} &= (1, 1, 1, 1, 0, 0, 0, 0)^T, \quad \boldsymbol{\tau} = (1, 1, 1, 1, 0, 0, 0, 0)^T, \\
 \boldsymbol{\rho} &= (0, 0.1, 0.1, 0.1, 0, 0, 0, 0)^T, \quad \boldsymbol{\beta} = (0, 0.1, 0.1, 0.1, 0, 0, 0, 0)^T.
 \end{aligned}$$

We consider lasso penalty in (4.25) and elastic-net penalty in (4.26). For each penalty, we first generate the training data $(Z_{t1}, \dots, Z_{t8}), (Q_{t1}, \dots, Q_{t8}), (V_{t1}, \dots, V_{t8}),$ and (X_{t1}, \dots, X_{t8}) independently from $\text{unif}(-1, 1)$, and obtain the response y_t based on model in (4.37), for $t = 1, \dots, 1000$. The testing data with sample size $n = 1000$ are generated in the same manner in order to validate the model prediction performance. We then apply our proposed regularized zero-and-one beta regression method on training set $\{y_t, \mathbf{Z}_t, \mathbf{Q}_t, \mathbf{V}_t, \mathbf{X}_t\}$, for $t = 1, \dots, 1000$, to calculate the estimated parameters $\hat{\boldsymbol{\sigma}}, \hat{\boldsymbol{\tau}}, \hat{\boldsymbol{\rho}}, \hat{\boldsymbol{\beta}}$ and use them to predict \hat{y}_t 's from the testing set. We repeat this process 400 times.

The main purpose of our proposed method is to provide accurate predictions on re-

Table 4.1: Simulation Prediction Performance, average over 400 repetitions

Penalty	MSE _{pred} in (4.38) \pm SD	MAE _{pred} in (4.39) \pm SD	PCP ₁ in (4.40) \pm SD	PCP ₀ in (4.41) \pm SD
lasso	0.0785 (0.0304)	0.1398 (0.0357)	0.8373 (0.0759)	0.8390 (0.0738)
elastic-net	0.0300 (0.0103)	0.0774 (0.0138)	0.9467 (0.0254)	0.9436 (0.0269)

sponses that are in the range of $[0, 1]$. In each repetition, the model prediction performance on testing data is evaluated by mean square error (MSE_{pred}), mean absolute error (MAE_{pred}), percentage of correct prediction on 1's (PCP₁), and percentage of correct prediction on 0's (PCP₀),

$$\text{MSE}_{\text{pred}} = \frac{1}{n} \sum_{t=1}^n (\hat{y}_t - y_t)^2, \quad (4.38)$$

$$\text{MAE}_{\text{pred}} = \frac{1}{n} \sum_{t=1}^n |\hat{y}_t - y_t|, \quad (4.39)$$

$$\text{PCP}_1 = \frac{\text{number of correct predictions on 1's}}{\text{number of 1's in } y_t\text{'s}}, \quad (4.40)$$

$$\text{PCP}_0 = \frac{\text{number of correct predictions on 0's}}{\text{number of 0's in } y_t\text{'s}}. \quad (4.41)$$

Then, the mean MSE_{pred} \pm standard deviation, mean MAE_{pred} \pm standard deviation, mean PCP₁ \pm standard deviation, and mean PCP₀ \pm standard deviation of the 400 repetitions are reported in Table 4.1. The MSE_{pred} and MAE_{pred} represent the prediction errors and are the smaller the better, while the PCP₁ and PCP₀ are the percentages of correct predictions on 1's and 0's, thus the more close to 1 the better.

The model estimation performance on unknown parameters is measured by the overall mean square error (MSE_{est}), the percentage of zero coefficients correctly estimated to be zero (C), and the percentage of non-zero coefficients correctly estimated to be non-zero

Table 4.2: Simulation Estimation Performance, average over 400 repetitions

Penalty	MSE _{est} in (4.42) \pm SD	C in (4.43) \pm SD	IC in (4.44) \pm SD
lasso	0.1514 (0.0254)	0.9886 (0.0290)	0.7607 (0.0581)
elastic-net	0.1990 (0.0138)	0.9614 (0.0537)	0.7857 (0.0000)

(IC),

$$\text{MSE}_{\text{est}} = \frac{1}{N} (\|\hat{\boldsymbol{\sigma}} - \boldsymbol{\sigma}\|_{\ell_2}^2 + \|\hat{\boldsymbol{\tau}} - \boldsymbol{\tau}\|_{\ell_2}^2 + \|\hat{\boldsymbol{\rho}} - \boldsymbol{\rho}\|_{\ell_2}^2 + \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_{\ell_2}^2), \quad (4.42)$$

$$\text{C} = \frac{\text{number of correctly estimated zeros}}{\text{number of true zeros}}, \quad (4.43)$$

$$\text{IC} = \frac{\text{number of correctly estimated non-zeros}}{\text{number of true non-zeros}}, \quad (4.44)$$

where $N = m + l + r + k + 4 = 36$ is the total number of parameters, and the number of zeros and non-zeros in the true parameters are 22 and 14, respectively. The mean $\text{MSE}_{\text{est}} \pm$ standard deviation, mean $\text{C} \pm$ standard deviation, and mean $\text{IC} \pm$ standard deviation of the 400 repetitions are then reported in Table 4.2. The MSE_{est} is expected to be close to 0, while C and IC are expected to be close to 1.

From Table 4.1 - 4.2, we can make the following observations: 1) Overall, our proposed model performs well in predicting both the continuous part within the range $(0, 1)$ and the discrete part $\{0, 1\}$; 2) the lasso penalty performs better regarding to estimating the unknown parameters, and can correctly exclude almost all the insignificant predictors with zero coefficients, however, its prediction performance is worse than elastic-net penalty; 3) As a compromise of lasso and ridge, the elastic-net penalty can make good predictions on the response, and at the same time can efficiently select the subset of important predictors.

4.5 Application

In this section, we will apply our proposed regularized zero-and-one inflated beta regression model to predict the Comprehensive Complication Index (CCI) for patients undergoing major abdominal surgery. We consider lasso penalty in (4.25) and elastic-net penalty in

(4.26) in our proposed model, and refer to them as "Lasso" and "Elastic-net" methods, respectively, in the sequel. We use five-fold cross validation on the training data to choose the regularization parameters $\lambda = (\lambda_\sigma, \lambda_\tau, \lambda_\rho, \lambda_\beta)$ in (4.24) and the cutoffs c_1 and c_2 in classification rule (4.36). It turns out that the selected parameters are the same for two penalties. Specifically, $\lambda_\sigma^{\text{opt}} = 0.003$, $\lambda_\tau^{\text{opt}} = 0.005$, $\lambda_\rho^{\text{opt}} = 0.005$, $\lambda_\beta^{\text{opt}} = 0.01$ are obtained by minimizing the BIC criteria in (4.35) and the cutoffs $c_1^{\text{opt}} = 0.5$ and $c_2^{\text{opt}} = 0.5$ are selected by minimizing the prediction mean square error. For the purpose of comparison, we also consider the varying dispersion beta regression model as the baseline method, which is referred to as "Baseline method" in the sequel. Our proposed model with chosen tuning parameters and the beta regression model are fitted on the whole training data, and then applied on the validation set to predict CCI response. The model prediction performance is evaluated through the mean square error (MSE), mean absolute error (MAE), percentages of correct prediction on 1's and 0's (PCP₁ and PCP₀) that are calculated by comparing the predicted values with true responses in validation set.

The training set in this study consists of 206 adult patients who stayed in the Clinic for Digestive Surgery, Clinical Center of Serbia in Belgrade between November 2016 and October 2017. Participants were selected on the basis of: 1) having an elective major or major + abdominal surgery lasting for more than 2 hours; 2) having limited physiologic reserve of one or more organs as assessed by patient history and clinical testing or aging over 70 years; 3) having a planned postoperative ICU stay of at least 24 hours. The validation set consists of 60 patients who stayed in the same clinic during 2018, and is collected for validation purpose only, which means those patients are not used in the modeling process.

The CCI is a linear scale ranging from 0 (no complications) to 100 (death), and we need to first divide it by 100 to rescale it in the range $[0, 1]$. The proposed model is then applied to predict the rescaled CCI. Figure 4.1 shows the rescaled CCI in the training set (on left) and validation set (on right). As can be seen, the number of 1's (N=14) is relatively small compared to that of 0's (N=81) in the training data. If we train the model on the original

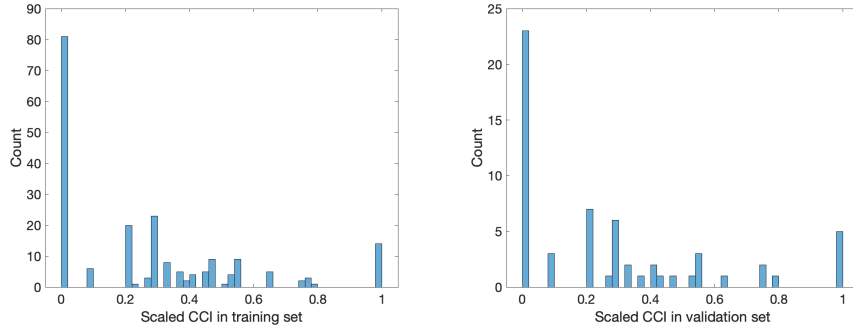


Figure 4.1: Left: Histogram of rescaled CCI in training set; Right: Histogram of rescaled CCI in validation set

Table 4.3: Variables and corresponding definitions.

Variables	Definitions
y_t	Response in $[0, 1]$
X_{t1}	Serum concentration of albumin (g/l) on the first postoperative day
X_{t2}	Base excess on admission to the ICU (mEq/l).
X_{t3}	History of diabetes (0=absent; 1=present)
X_{t4}	Heart rate on admission to ICU (beats/minute)
X_{t5}	Heart rate 12 hours after ICU admission (beats/minute)
X_{t6}	Surgery involving more than one organ due to the local invasiveness of disease (0=no; 1= yes)
X_{t7}	Administration of at least one unit of allogenic packed red blood cells within 24 hours (0=no; 1=yes)

training set, the classifier will tend to classify death as no complications. To avoid the prediction bias caused by such imbalance, and inspired by the upsampling idea in Machine Learning context, we randomly sample 67 (81-14) subjects who have response 1's with replacement from training data and then add them back to make the numbers of 0's and 1's the same. The observed training data can then be written in the form $\{y_t, X_{t1}, \dots, X_{t7}\}$ for $t = 1, \dots, n$, where $Y_t \in [0, 1]$ is the rescaled CCI, the (X_{t1}, \dots, X_{t7}) variables are Albumin2, BE, Diabetes, HR, HR.12hr, Multiorgan.resection, and Transfusion. Variable definitions are provided in Table 4.3. The total number of patients n is 273, including 81 with response 1's and response 0's, respectively, and 111 patients whose response is in the range of $(0, 1)$.

Before applying our proposed regularized zero-and-one inflated beta regression model, we first considered a simple linear regression model fitted by ordinary least squares, how-

ever, the model exhibits heteroskedasticity according to the studentized Breusch and Pagan test [155, 156]. Hence, linear regression model is not suitable here to predict CCI response. In our proposed methods, we assume the response y_t follows the zero-and-one inflated beta distribution, and has the density function in (4.17) with parameters μ_t , ϕ_t , α_t , and γ_t . We then assume those parameters can be expressed as linear combinations of variables $\mathbf{X}_t = (X_{t1}, \dots, X_{t7})$ through link functions,

$$\begin{aligned}
\log\left(\frac{\mu_t}{1-\mu_t}\right) &= \mathbf{X}_t^T \boldsymbol{\sigma}, \quad \boldsymbol{\sigma} = (\sigma_0, \sigma_1, \dots, \sigma_7)^T, \\
\log(\phi_t) &= \mathbf{X}_t^T \boldsymbol{\tau}, \quad \boldsymbol{\tau} = (\tau_0, \tau_1, \dots, \tau_7)^T, \\
\log\left(\frac{\alpha_t}{1-\alpha_t}\right) &= \mathbf{X}_t^T \boldsymbol{\rho}, \quad \boldsymbol{\rho} = (\rho_0, \rho_1, \dots, \rho_7)^T, \\
\log\left(\frac{\beta_t}{1-\gamma_t}\right) &= \mathbf{X}_t^T \boldsymbol{\beta}, \quad \boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_7)^T,
\end{aligned} \tag{4.45}$$

where $\boldsymbol{\sigma}$, $\boldsymbol{\tau}$, $\boldsymbol{\rho}$, and $\boldsymbol{\beta}$ are the unknown coefficients, which can be sparse since the variables $\mathbf{X}_t = (X_{t1}, \dots, X_{t7})$ are not necessarily all important in predicting parameters μ_t , ϕ_t , α_t , and γ_t . We will apply our proposed variable selection algorithm on the training set to estimate those unknown coefficients, and then predict the responses y_t 's in the validation set using the classification rule in (4.36).

As a comparison, we will also consider the varying dispersion beta regression model. Since beta regression can only model the unit interval $(0, 1)$, hence we need to transform the response from $[0, 1]$ to $(0, 1)$ before fitting the baseline beta regression model. Let's consider the following transformations:

$$y'_t = \frac{y_t + \epsilon}{1 + 2 * \epsilon}, \quad \text{with } \epsilon = 0.1.$$

Then the beta regression model is fitted on $y'_t \in (0, 1)$ in training set, but the predicted value will be transformed back to the scale $[0, 1]$ through $\hat{y}_t = \hat{y}'_t \times (1 + 2 * \epsilon) - \epsilon$ before comparing to the true responses in the validation set.

Table 4.4: Prediction performance, average across 5-fold CV on training data

Method	MSE_{pred} in (4.38) \pm SD	MAE_{pred} in (4.39) \pm SD	PCP_1 in (4.40) \pm SD	PCP_0 in (4.41) \pm SD
Lasso	0.0591 (0.0168)	0.1391 (0.0213)	1 (0)	0.7279 (0.1057)
Elastic-net	0.0555 (0.0075)	0.1367 (0.0115)	1 (0)	0.7154 (0.1061)
Baseline method	0.0900 (0.0555)	0.2074 (0.0663)	0 (0)	0 (0)

Table 4.5: Prediction performance, on validation data

Method	MSE_{pred} in (4.38)	MAE_{pred} in (4.39)	PCP_1 in (4.40)	PCP_0 in (4.41)
Lasso	0.0955	0.2145	0.8000	0.7391
Elastic-net	0.0955	0.2144	0.8000	0.7391
Baseline method	0.1382	0.3122	0	0

The prediction performances of the baseline method and our proposed method with lasso and elastic-net penalties on training data through five-fold cross-validation are shown in Table 4.4, and the model performances on the validation set are shown in Table 4.5. We also calculate the MSEs and MAEs for responses in different ranges. Figure 4.2 shows the bar plots of mean MSEs of five-fold cross validation on training data (on left) and MSEs on validation data (on right).

As can be seen, our zero-and-one inflated beta regression model with lasso penalty and elastic-net penalty have very similar prediction performances and both are better than the baseline varying dispersion beta regression model. The two advantages of our method include: 1) We can predict the edge values 1 (death) and 0 (no complications) well, however, the baseline method is not able to provide exact predictions on the 1's and 0's; 2) Our methods can select the subset of important variables, while the baseline method is not able to do variable selection. Besides, from Figure 4.2, we notice that our methods with lasso and elastic-net penalties perform better than the baseline method in predicting the values close to 0 and 1, but as a compromise, they perform worse in predicting the values in between, although such difference is not large. In practice, accurate predictions on values close to 0 (no complications) and 1 (death) are more important, since they represent the two extreme cases and can be more informative than the values in the middle.

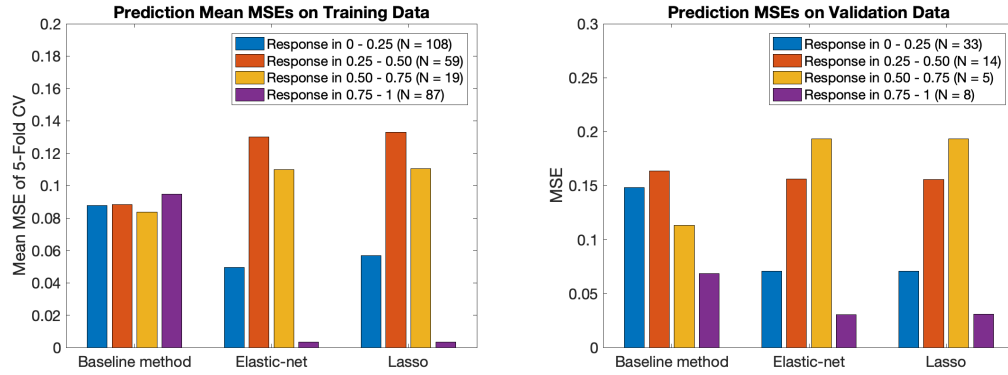


Figure 4.2: MSEs in different ranges. Left: 5-fold CV results on training data. Right: Results on validation set

The variable selection results are shown in Figure 4.3, where x-axis represents different variables and y-axis is the coefficient for each variable. In this application, the lasso and elastic net penalties select the same subset of variables in all four models, which explains why they have similar prediction performance as is shown in Tables 4.4 and 4.5. However, they have certain differences and are suitable to different situations. To be more concrete, lasso is known to be indifferent to very correlated predictors, and will tend to pick one and ignore the rest. The elastic-net penalty is a compromise between ridge and lasso. It performs much like the lasso, but removes any degeneracies and wild behavior caused by extreme correlations. We also notice that, interestingly, the important predictors are different in models for four different parameters, and the effects (signs) of the same predictor can also be different in different models, for example, the BE in the model for μ_t and the model for ϕ_t .

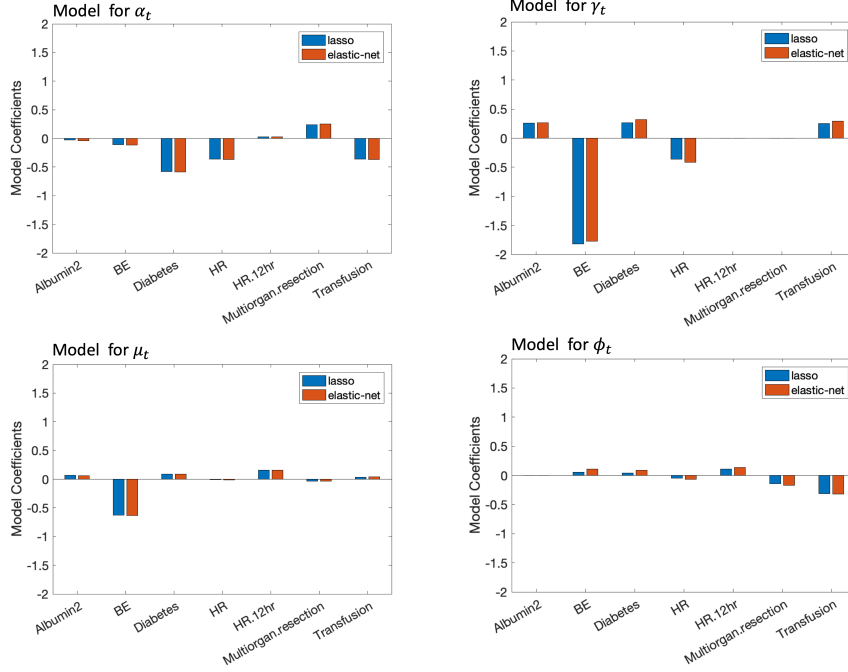


Figure 4.3: Variable Selection Results

4.6 Derivation of Score functions and Information matrix

In this section, we derive the score functions and fisher information matrix. Before that, we need to first define some notations,

$$y_t^* = \begin{cases} \log \frac{y_t}{1-y_t}, & \text{if } y \in (0, 1) \\ 0, & \text{otherwise} \end{cases}, y_t^\dagger = \begin{cases} \log(1 - y_t), & \text{if } y \in (0, 1) \\ 0, & \text{otherwise} \end{cases}, \quad (4.46)$$

and we have

$$\begin{aligned} \mu_t^* &= \mathbb{E}(y_t^* | y_t \in (0, 1)) = \psi(\mu_t \phi_t) - \psi((1 - \mu_t) \phi_t), \\ \mu_t^\dagger &= \mathbb{E}(y_t^\dagger | y_t \in (0, 1)) = \psi((1 - \mu_t) \phi_t) - \psi(\phi_t), \\ v_t^* &= \text{var}(y_t^* | y_t \in (0, 1)) = \psi'(\mu_t \phi_t) - \psi'((1 - \mu_t) \phi_t), \\ v_t^\dagger &= \text{var}(y_t^\dagger | y_t \in (0, 1)) = \psi'((1 - \mu_t) \phi_t) - \psi'(\phi_t), \\ c_t^{*\dagger} &= \text{cov}(y_t^*, y_t^\dagger | y_t \in (0, 1)) = -\psi'((1 - \mu_t) \phi_t), \end{aligned}$$

where $\psi(x) = d \log \Gamma(x)/dx$ is the digamma function.

Score functions. The elements of the score vector are

$$\begin{aligned}
U_{\sigma_i}(\boldsymbol{\sigma}, \boldsymbol{\tau}) &= \frac{\partial \ell(\boldsymbol{\sigma}, \boldsymbol{\tau}, \boldsymbol{\rho}, \boldsymbol{\beta})}{\partial \sigma_i} \\
&= \frac{\partial \ell_3(\boldsymbol{\sigma}, \boldsymbol{\tau})}{\partial \sigma_i} \\
&= \sum_{t: y_t \in (0,1)} \frac{\partial \ell_{3t}(\mu_t, \phi_t)}{\partial \sigma_i} \\
&= \sum_{t: y_t \in (0,1)} \frac{\partial \ell_{3t}(\mu_t, \phi_t)}{\partial \mu_i} \cdot \frac{\partial \mu_t}{\partial \eta_{1t}} \cdot \frac{\partial \eta_{1t}}{\partial \sigma_i} \\
&= \sum_{t: y_t \in (0,1)} \left(\frac{\phi_t(y_t^* - \mu_t^*)}{h_1'(\mu_t)} \cdot Z_{ti} \right), \quad \text{for } i = 0, \dots, m.
\end{aligned}$$

$$\begin{aligned}
U_{\tau_i}(\boldsymbol{\sigma}, \boldsymbol{\tau}) &= \frac{\partial \ell(\boldsymbol{\sigma}, \boldsymbol{\tau}, \boldsymbol{\rho}, \boldsymbol{\beta})}{\partial \tau_i} \\
&= \frac{\partial \ell_3(\boldsymbol{\sigma}, \boldsymbol{\tau})}{\partial \tau_i} \\
&= \sum_{t: y_t \in (0,1)} \frac{\partial \ell_{3t}(\mu_t, \phi_t)}{\partial \tau_i} \\
&= \sum_{t: y_t \in (0,1)} \frac{\partial \ell_{3t}(\mu_t, \phi_t)}{\partial \phi_i} \cdot \frac{\partial \phi_t}{\partial \eta_{2t}} \cdot \frac{\partial \eta_{2t}}{\partial \tau_i} \\
&= \sum_{t: y_t \in (0,1)} \left(\frac{\mu_t(y_t^* - \mu_t^*) + (y_t^\dagger - \mu_t^\dagger)}{h_2'(\phi_t)} \cdot Q_{ti} \right), \quad \text{for } i = 0, \dots, l.
\end{aligned}$$

$$\begin{aligned}
U_{\rho_i}(\boldsymbol{\rho}) &= \frac{\partial \ell(\boldsymbol{\mu}, \boldsymbol{\phi}, \boldsymbol{\rho}, \boldsymbol{\beta})}{\partial \rho_i} \\
&= \frac{\partial \ell_1(\boldsymbol{\rho})}{\partial \rho_i} \\
&= \sum_{t=1}^n \frac{\partial \ell_{1t}(\alpha_t)}{\partial \rho_i} \\
&= \sum_{t=1}^n \frac{\partial \ell_{1t}(\alpha_t)}{\partial \alpha_t} \cdot \frac{\partial \alpha_t}{\partial \eta_{3t}} \cdot \frac{\partial \eta_{3t}}{\partial \rho_i} \\
&= \sum_{t=1}^n \left(\frac{\mathbb{1}_{\{0,1\}}(y_t) - \alpha_t}{\alpha_t(1 - \alpha_t)h'_3(\alpha_t)} \cdot V_{ti} \right), \quad \text{for } i = 0, \dots, r.
\end{aligned}$$

$$\begin{aligned}
U_{\beta_i}(\boldsymbol{\beta}) &= \frac{\partial \ell(\boldsymbol{\mu}, \boldsymbol{\phi}, \boldsymbol{\rho}, \boldsymbol{\beta})}{\partial \beta_i} \\
&= \frac{\partial \ell_2(\boldsymbol{\beta})}{\partial \beta_i} \\
&= \sum_{y_t \in \{0,1\}} \frac{\partial \ell_{2t}(\gamma_t)}{\partial \beta_i} \\
&= \sum_{y_t \in \{0,1\}} \frac{\partial \ell_{2t}(\gamma_t)}{\partial \gamma_t} \cdot \frac{\partial \gamma_t}{\partial \eta_{4t}} \cdot \frac{\partial \eta_{4t}}{\partial \beta_i} \\
&= \sum_{t=1}^n \left(\frac{\mathbb{1}_{\{0,1\}}(y_t)(y_t - \gamma_t)}{\gamma_t(1 - \gamma_t)h'_4(\gamma_t)} \cdot X_{ti} \right), \quad \text{for } i = 0, \dots, k.
\end{aligned}$$

Therefore, the score functions are:

$$U_{\boldsymbol{\sigma}}(\boldsymbol{\sigma}, \boldsymbol{\tau}) = \frac{\partial \ell(\boldsymbol{\sigma}, \boldsymbol{\tau}, \boldsymbol{\rho}, \boldsymbol{\beta})}{\partial \boldsymbol{\sigma}} = \frac{1}{n} \mathbf{Z}^T P_{\boldsymbol{\sigma}}, \quad (4.47)$$

$$U_{\boldsymbol{\tau}}(\boldsymbol{\sigma}, \boldsymbol{\tau}) = \frac{\partial \ell(\boldsymbol{\sigma}, \boldsymbol{\tau}, \boldsymbol{\rho}, \boldsymbol{\beta})}{\partial \boldsymbol{\tau}} = \frac{1}{n} \mathbf{Q}^T P_{\boldsymbol{\tau}}, \quad (4.48)$$

$$U_{\boldsymbol{\rho}}(\boldsymbol{\rho}) = \frac{\partial \ell(\boldsymbol{\sigma}, \boldsymbol{\tau}, \boldsymbol{\rho}, \boldsymbol{\beta})}{\partial \boldsymbol{\rho}} = \frac{1}{n} \mathbf{V}^T P_{\boldsymbol{\rho}}, \quad (4.49)$$

$$U_{\boldsymbol{\beta}}(\boldsymbol{\beta}) = \frac{\partial \ell(\boldsymbol{\sigma}, \boldsymbol{\tau}, \boldsymbol{\rho}, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \frac{1}{n} \mathbf{X}^T P_{\boldsymbol{\beta}}, \quad (4.50)$$

where $\mathbf{Z}^T = (\mathbf{Z}_1, \dots, \mathbf{Z}_n) \in \mathbb{R}^{(m+1) \times n}$, $\mathbf{Q}^T = (\mathbf{Q}_1, \dots, \mathbf{Q}_n) \in \mathbb{R}^{(l+1) \times n}$, $\mathbf{V}^T = (\mathbf{V}_1, \dots, \mathbf{V}_n) \in$

$\mathbb{R}^{(r+1) \times n}$, and $\mathbf{X}^T = (\mathbf{X}_1, \dots, \mathbf{X}_n) \in \mathbb{R}^{(k+1) \times n}$ are known covariates, and

$$P_{\boldsymbol{\sigma}} = (p_{\sigma 1}, \dots, p_{\sigma n})^T, \quad p_{\sigma t} = \frac{\mathbb{1}_{(0,1)}(y_t) \phi_t (y_t^* - \mu_t^*)}{h_1'(\mu_t)} \triangleq p_{\sigma t}(\mu_t, \phi_t), \quad (4.51)$$

$$P_{\boldsymbol{\tau}} = (p_{\tau 1}, \dots, p_{\tau n})^T, \quad p_{\tau t} = \frac{\mathbb{1}_{(0,1)}(y_t) \left(\mu_t (y_t^* - \mu_t^*) + (y_t^\dagger - \mu_t^\dagger) \right)}{h_2'(\phi_t)} \triangleq p_{\tau t}(\mu_t, \phi_t), \quad (4.52)$$

$$P_{\boldsymbol{\rho}} = (p_{\rho 1}, \dots, p_{\rho n})^T, \quad p_{\rho t} = \frac{\mathbb{1}_{\{0,1\}}(y_t) - \alpha_t}{\alpha_t (1 - \alpha_t) h_3'(\alpha_t)} \triangleq p_{\rho t}(\alpha_t), \quad (4.53)$$

$$P_{\boldsymbol{\beta}} = (p_{\beta 1}, \dots, p_{\beta n})^T, \quad p_{\beta t} = \frac{(y_t - \gamma_t) \mathbb{1}_{\{0,1\}}(y_t)}{\gamma_t (1 - \gamma_t) h_4'(\gamma_t)} \triangleq p_{\beta t}(\beta_t). \quad (4.54)$$

□

Fisher information matrix. Let us first present the secondary derivatives of the log-likelihood function with respect to the unknown parameters, which will be used to obtain the fisher information matrix and the quadratic approximations to the log-likelihood function.

The elements of secondary derivatives are:

$$\begin{aligned} J_{\sigma_i, \sigma_{i'}}(\boldsymbol{\sigma}, \boldsymbol{\tau}) &= \frac{\partial^2 \ell(\boldsymbol{\sigma}, \boldsymbol{\tau}, \boldsymbol{\rho}, \boldsymbol{\beta})}{\partial \sigma_i \partial \sigma_{i'}} = \frac{\partial U_{\sigma_i}(\boldsymbol{\sigma}, \boldsymbol{\tau})}{\partial \sigma_{i'}}, \quad \text{for } i = 1, \dots, m; i' = 1, \dots, m, \\ &= -\frac{1}{n} \sum_{t: y_t \in (0,1)} \left(\frac{\phi_t^2 v_t^* - 2\phi_t^2 c_i^{*\dagger}}{(h_1'(\mu_t))^2} + \frac{\phi_t (y_t^* - \mu_t^*) h_1''(\mu_t)}{(h_1'(\mu_t))^3} \right) Z_{ti} Z_{ti'}. \end{aligned}$$

$$\begin{aligned} J_{\tau_i, \tau_{i'}}(\boldsymbol{\sigma}, \boldsymbol{\tau}) &= \frac{\partial^2 \ell(\boldsymbol{\sigma}, \boldsymbol{\tau}, \boldsymbol{\rho}, \boldsymbol{\beta})}{\partial \tau_i \partial \tau_{i'}} = \frac{\partial U_{\tau_i}(\boldsymbol{\sigma}, \boldsymbol{\tau})}{\partial \tau_{i'}}, \quad \text{for } i = 1, \dots, l; i' = 1, \dots, l, \\ &= -\frac{1}{n} \sum_{t=1}^n \mathbb{1}_{(0,1)}(y_t) \left(\frac{\mu_t^2 v_t^* + v_t^\dagger + 2\mu_t (1 - \mu_t) c_i^{*\dagger}}{(h_2'(\phi_t))^2} + \right. \\ &\quad \left. \frac{\left(\mu_t (y_t^* - \mu_t^*) + (y_t^\dagger - \mu_t^\dagger) \right) h_2''(\phi_t)}{(h_2'(\phi_t))^3} \right) Q_{ti} Q_{ti'}. \end{aligned}$$

$$\begin{aligned}
J_{\sigma_i, \tau_{i'}}(\boldsymbol{\sigma}, \boldsymbol{\tau}) &= \frac{\partial^2 \ell(\boldsymbol{\sigma}, \boldsymbol{\tau}, \boldsymbol{\rho}, \boldsymbol{\beta})}{\partial \sigma_i \partial \tau_{i'}} = \frac{\partial U_{\sigma_i}(\boldsymbol{\sigma}, \boldsymbol{\tau})}{\partial \tau_{i'}}, \quad \text{for } i = 1, \dots, m; i' = 1, \dots, l, \\
&= -\frac{1}{n} \sum_{t=1}^n \mathbb{1}_{(0,1)}(y_t) \left(\frac{\phi_t (\mu_t v_t^* + (1 - 2\mu_t) c_t^{*\dagger}) - (y_t^* - \mu_t^*)}{h_1'(\mu_t) h_2'(\phi_t)} \right) Z_{ti} Q_{ti'}.
\end{aligned}$$

$$\begin{aligned}
J_{\rho_i, \rho_{i'}}(\boldsymbol{\rho}) &= \frac{\partial^2 \ell(\boldsymbol{\sigma}, \boldsymbol{\tau}, \boldsymbol{\rho}, \boldsymbol{\beta})}{\partial \rho_i \partial \rho_{i'}} = \frac{\partial U_{\rho_i}(\boldsymbol{\rho})}{\partial \rho_{i'}}, \quad \text{for } i = 1, \dots, r; i' = 1, \dots, r, \\
&= -\frac{1}{n} \sum_{t=1}^n \left(\left(\frac{\mathbb{1}_{(0,1)}(y_t)}{(1 - \alpha_t)^2} + \frac{\mathbb{1}_{\{0,1\}}(y_t)}{\alpha_t^2} \right) \frac{1}{(h_3'(\alpha_t))^2} + \frac{(\mathbb{1}_{\{0,1\}}(y_t) - \alpha_t) h_3''(\alpha_t)}{\alpha_t (1 - \alpha_t) (h_3'(\alpha_t))^3} \right) V_{ti} V_{ti'}.
\end{aligned}$$

$$\begin{aligned}
J_{\beta_i, \beta_{i'}}(\boldsymbol{\beta}) &= \frac{\partial^2 \ell(\boldsymbol{\sigma}, \boldsymbol{\tau}, \boldsymbol{\rho}, \boldsymbol{\beta})}{\partial \beta_i \partial \beta_{i'}} = \frac{\partial U_{\beta_i}(\boldsymbol{\beta})}{\partial \beta_{i'}}, \quad \text{for } i = 1, \dots, k; i' = 1, \dots, k, \\
&= -\frac{1}{n} \sum_{t=1}^n \mathbb{1}_{\{0,1\}}(y_t) \left(\frac{y_t - 2y_t \gamma_t + \gamma_t^2}{\gamma_t^2 (1 - \gamma_t)^2 (h_4'(\gamma_t))^2} + \frac{(y_t - \gamma_t) h_4''(\gamma_t)}{\gamma_t (1 - \gamma_t) (h_4'(\gamma_t))^3} \right) X_{ti} X_{ti'}.
\end{aligned}$$

Therefore the secondary derivatives can be written as:

$$J_{\boldsymbol{\sigma}}(\boldsymbol{\sigma}, \boldsymbol{\tau}) = \frac{\partial^2 \ell(\boldsymbol{\sigma}, \boldsymbol{\tau}, \boldsymbol{\rho}, \boldsymbol{\beta})}{\partial \boldsymbol{\sigma}^2} = -\frac{1}{n} \mathbf{Z}^T W_{\boldsymbol{\sigma}} \mathbf{Z}, \quad (4.55)$$

$$J_{\boldsymbol{\tau}}(\boldsymbol{\sigma}, \boldsymbol{\tau}) = \frac{\partial^2 \ell(\boldsymbol{\sigma}, \boldsymbol{\tau}, \boldsymbol{\rho}, \boldsymbol{\beta})}{\partial \boldsymbol{\tau}^2} = -\frac{1}{n} \mathbf{Q}^T W_{\boldsymbol{\tau}} \mathbf{Q}, \quad (4.56)$$

$$J_{\boldsymbol{\sigma}\boldsymbol{\tau}}(\boldsymbol{\sigma}, \boldsymbol{\tau}) = \frac{\partial^2 \ell(\boldsymbol{\sigma}, \boldsymbol{\tau}, \boldsymbol{\rho}, \boldsymbol{\beta})}{\partial \boldsymbol{\sigma} \partial \boldsymbol{\tau}} = -\frac{1}{n} \mathbf{Z}^T W_{\boldsymbol{\sigma}\boldsymbol{\tau}} \mathbf{Q}, \quad (4.57)$$

$$J_{\boldsymbol{\tau}\boldsymbol{\sigma}}(\boldsymbol{\sigma}, \boldsymbol{\tau}) = \frac{\partial^2 \ell(\boldsymbol{\sigma}, \boldsymbol{\tau}, \boldsymbol{\rho}, \boldsymbol{\beta})}{\partial \boldsymbol{\tau} \partial \boldsymbol{\sigma}} = -\frac{1}{n} \mathbf{Q}^T W_{\boldsymbol{\sigma}\boldsymbol{\tau}} \mathbf{Z}, \quad (4.58)$$

$$J_{\boldsymbol{\rho}}(\boldsymbol{\rho}) = \frac{\partial^2 \ell(\boldsymbol{\sigma}, \boldsymbol{\tau}, \boldsymbol{\rho}, \boldsymbol{\beta})}{\partial \boldsymbol{\rho}^2} = -\frac{1}{n} \mathbf{V}^T W_{\boldsymbol{\rho}} \mathbf{V}, \quad (4.59)$$

$$J_{\boldsymbol{\beta}}(\boldsymbol{\beta}) = \frac{\partial^2 \ell(\boldsymbol{\sigma}, \boldsymbol{\tau}, \boldsymbol{\rho}, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}^2} = -\frac{1}{n} \mathbf{X}^T W_{\boldsymbol{\beta}} \mathbf{X}, \quad (4.60)$$

where $W_{\boldsymbol{\sigma}} = \text{diag}(w_{\sigma 1}, \dots, w_{\sigma n})$, $W_{\boldsymbol{\tau}} = \text{diag}(w_{\tau 1}, \dots, w_{\tau n})$, $W_{\boldsymbol{\sigma}\boldsymbol{\tau}} = \text{diag}(w_{\sigma\tau 1}, \dots, w_{\sigma\tau n})$,

$W_\rho = \text{diag}(w_{\rho 1}, \dots, w_{\rho n})$, and $W_\beta = \text{diag}(w_{\beta 1}, \dots, w_{\beta n})$, and for $t = 1, \dots, n$,

$$w_{\sigma t} = \mathbb{1}_{(0,1)}(y_t) \left(\frac{\phi_t^2 v_t^* - 2\phi_t^2 c_t^{*\dagger}}{(h'_1(\mu_t))^2} + \frac{\phi_t(y_t^* - \mu_t^*)h''_1(\mu_t)}{(h'_1(\mu_t))^3} \right), \quad (4.61)$$

$$w_{\tau t} = \mathbb{1}_{(0,1)}(y_t) \left(\frac{\mu_t^2 v_t^* + v_t^\dagger + 2\mu_t(1 - \mu_t)c_t^{*\dagger}}{(h'_2(\phi_t))^2} + \frac{(\mu_t(y_t^* - \mu_t^*) + (y_t^\dagger - \mu_t^\dagger))h''_2(\phi_t)}{(h'_2(\phi_t))^3} \right) \quad (4.62)$$

$$w_{\sigma\tau t} = \mathbb{1}_{(0,1)}(y_t) \left(\frac{\phi_t(\mu_t v_t^* + (1 - 2\mu_t)c_t^{*\dagger}) - (y_t^* - \mu_t^*)}{h'_1(\mu_t)h'_2(\phi_t)} \right), \quad (4.63)$$

$$w_{\rho t} = \left(\frac{\mathbb{1}_{(0,1)}(y_t)}{(1 - \alpha_t)^2} + \frac{\mathbb{1}_{\{0,1\}}(y_t)}{\alpha_t^2} \right) \frac{1}{(h'_3(\alpha_t))^2} + \frac{(\mathbb{1}_{\{0,1\}}(y_t) - \alpha_t)h''_3(\alpha_t)}{\alpha_t(1 - \alpha_t)(h'_3(\alpha_t))^3}, \quad (4.64)$$

$$w_{\beta t} = \mathbb{1}_{\{0,1\}}(y_t) \left(\frac{y_t - 2y_t\gamma_t + \gamma_t^2}{\gamma_t^2(1 - \gamma_t)^2(h'_4(\gamma_t))^2} + \frac{(y_t - \gamma_t)h''_4(\gamma_t)}{\gamma_t(1 - \gamma_t)(h'_4(\gamma_t))^3} \right). \quad (4.65)$$

The secondary derivatives are in complex forms, and to simplify, we will use $w_{\sigma t}(\mu_t, \phi_t)$, $w_{\tau t}(\mu_t, \phi_t)$, $w_{\sigma\tau t}(\mu_t, \phi_t)$, $w_{\rho t}(\alpha_t)$, and $w_{\beta t}(\gamma_t)$ to represent the expressions of $w_{\sigma t}$, $w_{\tau t}$, $w_{\sigma\tau t}$, $w_{\rho t}$, and $w_{\beta t}$ in (4.61) - (4.65).

Besides, it is easy to verify the facts that

$$\begin{aligned} \mathbb{E}(\mathbb{1}_{(0,1)}(y_t)) &= 1 - \alpha_t, \\ \mathbb{E}(\mathbb{1}_{\{0,1\}}(y_t)) &= \alpha_t, \\ \mathbb{E}(y_t \mathbb{1}_{\{0,1\}}(y_t)) &= \alpha_t \gamma_t, \\ \mathbb{E}((1 - y_t) \mathbb{1}_{\{0,1\}}(y_t)) &= \alpha_t(1 - \gamma_t), \\ \mathbb{E}(y_t^* \mathbb{1}_{(0,1)}(y_t)) &= (1 - \alpha_t)\mu_t^*, \\ \mathbb{E}(y_t^\dagger \mathbb{1}_{(0,1)}(y_t)) &= (1 - \alpha_t)\mu_t^\dagger, \end{aligned}$$

Therefore, we have

$$\begin{aligned}\mathbb{E}(J_{\sigma_i, \sigma_{i'}}(\boldsymbol{\sigma}, \boldsymbol{\tau})) &= -\frac{1}{n} \sum_{i=1}^n \frac{(1 - \alpha_t) (\phi_t^2 v_t^* - 2\phi_t^2 c_i^{*\dagger})}{(h'_1(\mu_t))^2} Z_{ti} Z_{ti'}, \\ \mathbb{E}(J_{\tau_i, \tau_{i'}}(\boldsymbol{\sigma}, \boldsymbol{\tau})) &= -\frac{1}{n} \sum_{i=1}^n \frac{(1 - \alpha_t) (\mu_t^2 v_t^* + v_t^\dagger + 2\mu_t(1 - \mu_t)c_t^{*\dagger})}{(h'_2(\phi_t))^2} Q_{ti} Q_{ti'}, \\ \mathbb{E}(J_{\sigma_i, \tau_{i'}}(\boldsymbol{\sigma}, \boldsymbol{\tau})) &= -\frac{1}{n} \sum_{i=1}^n \frac{(1 - \alpha_t) \phi_t (\mu_t v_t^* + (1 - 2\mu_t)c_t^{*\dagger})}{(h'_1(\mu_t)h'_2(\phi_t))} Z_{ti} Q_{ti'}, \\ \mathbb{E}(J_{\rho_i, \rho_{i'}}(\boldsymbol{\rho})) &= -\frac{1}{n} \sum_{i=1}^n \frac{\frac{1}{1-\alpha_t} + \frac{\alpha_t}{\alpha_t}}{(h'_3(\alpha_t))^2} V_{ti} V_{ti'}, \\ \mathbb{E}(J_{\beta_i, \beta_{i'}}(\boldsymbol{\beta})) &= -\frac{1}{n} \sum_{i=1}^n \frac{\frac{\alpha_t}{1-\gamma_t} + \frac{\alpha_t}{\gamma_t}}{(h'_4(\gamma_t))^2} X_{ti} X_{ti'},\end{aligned}$$

The entries in fisher information matrix are then obtained by

$$\begin{aligned}K_{\boldsymbol{\sigma}\boldsymbol{\sigma}} &= \mathbb{E}(-J_{\boldsymbol{\sigma}}(\boldsymbol{\sigma}, \boldsymbol{\tau})) = \frac{1}{n} \mathbf{Z}^T W_1 \mathbf{Z}, \\ K_{\boldsymbol{\tau}\boldsymbol{\tau}} &= \mathbb{E}(-J_{\boldsymbol{\tau}}(\boldsymbol{\sigma}, \boldsymbol{\tau})) = \frac{1}{n} \mathbf{Q}^T W_2 \mathbf{Q}, \\ K_{\boldsymbol{\sigma}\boldsymbol{\tau}} &= K_{\boldsymbol{\tau}\boldsymbol{\sigma}}^T = \mathbb{E}(-J_{\boldsymbol{\sigma}\boldsymbol{\tau}}(\boldsymbol{\sigma}, \boldsymbol{\tau})) = \frac{1}{n} \mathbf{Z}^T W_{12} \mathbf{Q}, \\ K_{\boldsymbol{\rho}\boldsymbol{\rho}} &= \mathbb{E}(-J_{\boldsymbol{\rho}}(\boldsymbol{\rho})) = \frac{1}{n} \mathbf{V}^T W_3 \mathbf{V}, \\ K_{\boldsymbol{\beta}\boldsymbol{\beta}} &= \mathbb{E}(-J_{\boldsymbol{\beta}}(\boldsymbol{\beta})) = \frac{1}{n} \mathbf{X}^T W_4 \mathbf{X},\end{aligned}$$

where

$$W_1 = \text{diag}(w_{11}, \dots, w_{1n}), w_{1t} = \frac{(1 - \alpha_t) (\phi_t^2 v_t^* - 2\phi_t^2 c_t^{*\dagger})}{(h'_1(\mu_t))^2}, \quad (4.66)$$

$$W_2 = \text{diag}(w_{21}, \dots, w_{2n}), w_{2t} = \frac{(1 - \alpha_t) (\mu_t^2 v_t^* + v_t^\dagger + 2\mu_t(1 - \mu_t)c_t^{*\dagger})}{(h'_2(\phi_t))^2}, \quad (4.67)$$

$$W_{12} = \text{diag}(w_{12;1}, \dots, w_{12;n}), w_{12;t} = \frac{(1 - \alpha_t)\phi_t (\mu_t v_t^* + (1 - 2\mu_t)c_t^{*\dagger})}{(h'_1(\mu_t)h'_2(\phi_t))}, \quad (4.68)$$

$$W_3 = \text{diag}(w_{31}, \dots, w_{3n}), w_{3t} = \frac{\frac{1}{1-\alpha_t} + \frac{1}{\alpha_t}}{(h'_3(\alpha_t))^2}, \quad (4.69)$$

$$W_4 = \text{diag}(w_{41}, \dots, w_{4n}), w_{4t} = \frac{\frac{\alpha_t}{1-\gamma_t} + \frac{\alpha_t}{\gamma_t}}{(h'_4(\gamma_t))^2}. \quad (4.70)$$

4.7 Conclusions

An accurate and reproducible method for complication reporting is essential for a valid assessment of the outcomes of surgery. The Comprehensive Complication Index (CCI) is a novel linear scale ranging from 0 to 100, designed to capture the overall burden of complications after surgery. The most important limitation is that it is calculated by a complicated formula, and is not practical in everyday use. In this study, we develop a machine learning method to self-learn the limited-range CCI values based on patients' clinical covariates.

We propose a regularized zero-and-one inflated beta regression model for responses that are in the range of $[0, 1]$. One would model $(y - a)/(b - a)$ instead of modeling y directly for the cases where the response is restricted to the interval $[a, b]$, such as CCI where $a = 0$ and $b = 100$. To be more concrete, we assume the response follows a zero-and-one inflated beta distribution, which is a mixture of a beta distribution and a Bernoulli distribution, and then relate the density parameters to linear combinations of predictors through unknown coefficients and link functions. The unknown coefficients are obtained by maximizing a penalized log likelihood function on training data, and we propose to use the lasso penalty

and elastic net penalty to realize variable selection. The trained model can then be applied on the testing set to predict the responses in $[0, 1]$.

Our method has been validated through simulation study. It turns out lasso penalty performs worse in predicting the responses, but it is better regarding to parameter estimation, and can correctly select the subset of important predictors. As a compromise of lasso and ridge, the elastic-net penalty can make good predictions on the response, and at the same time can efficiently select the subset of important predictors.

We also apply our proposed method to a real data set to predict the CCI. The training data was collected from the Clinic for Digestive Surgery, Clinical Center of Serbia in Belgrade between November 2016 and October 2017, and the validation set was collected in the same clinic during 2018. Compared to the baseline varying dispersion beta regression method, our methods can predict the responses better, especially for the values that are close to 0 (no complications) and 1 (death). In addition, we can select the subset of important variables.

It is worth noting that although we demonstrated the use of regularized zero-and-one inflated beta regression model for estimating complication index CCI, the general approach can be applied to other applications. Examples include predicting the G.P.A which is in the range of $[0, 4]$.

CHAPTER 5

CONCLUSIONS AND FUTURE RESEARCH

This thesis develops feature learning and personalized screening methodologies in health-care, and it deals with the following four different contexts: (i) Visual impairment classification based on noisy high-frequency pupillary response behavior data collected from human-computer interaction, (ii) breast cancer diagnosis using image data from plain Xray, (iii) personalized screening for sepsis disease based on regularly measured longitudinal biomarkers, (iv) prediction on the overall burdens of postoperative complications using laboratory measurements.

Our research opens several further research directions in public health.

- **Feature extraction.** In the first two chapters of this thesis, we develop a robust estimation of the Hurst exponent from 1-D and 2-D fractional Brownian motions (fBm) based on wavelet spectra, but we only consider constant Hurst exponent. It will be interesting to consider time changing Hurst exponent for time series data. Moreover, it is also interesting to investigate other methods to estimate the Hurst exponent, for instance, the method based on convex rearrangements of properly filtered versions of the process.
- **Personalized screening.** Chapter 3 of this thesis opens several new directions in the context of screening: 1) How to jointly estimate the personalized thresholds for multiple biomarkers by taking into account the possible correlation structures between these biomarkers; 2) how to develop more complicated personalized threshold that can be non-linear functions of the baseline characteristics; 3) how to conduct both variable selection and personalized threshold estimation simultaneously.
- **Beta regression model.** In the last chapter, we study the zero and one inflated beta

regression model and consider variable selection for it. There are a couple of future directions: 1) It is useful to establish the consistency and asymptotic normality properties of the proposed parameter estimation method; 2) it is interesting to consider the non-convex penalty terms, such as the capped-L1 regularization. It is crucial to develop computationally efficient numerical algorithm to solve corresponding non-convex optimization problem.

REFERENCES

- [1] J. M. Geraci, C. M. Ashton, D. H. Kuykendall, M. L. Johnson, and L. Wu, “International classification of diseases, 9th revision, clinical modification codes in discharge abstracts are poor measures of complication occurrence in medical inpatients,” *Medical Care*, vol. 35, no. 6, pp. 589–602, 1997.
- [2] S. K. West, G. S. Rubin, A. T. Broman, B. Munoz, K. Bandeen-Roche, and K. Turano, “How does visual impairment affect performance on tasks of everyday life?: The see project,” *Archives of Ophthalmology*, vol. 120, no. 6, pp. 774–780, 2002.
- [3] P. J. Rosenfeld, D. M. Brown, J. S. Heier, D. S. Boyer, P. K. Kaiser, C. Y. Chung, and R. Y. Kim, “Ranibizumab for neovascular age-related macular degeneration,” *New England Journal of Medicine*, vol. 355, no. 14, pp. 1419–1431, 2006.
- [4] D. C. Fletcher and R. A. Schuchard, “Visual function in patients with choroidal neovascularization resulting from age-related macular degeneration: The importance of looking beyond visual acuity,” *Optometry and Vision Science*, vol. 83, no. 3, pp. 178–189, 2006.
- [5] R. L. Avery, D. J. Pieramici, M. D. Rabena, A. A. Castellarin, A. N. Maan, and M. J. Giust, “Intravitreal bevacizumab (avastin) for neovascular age-related macular degeneration,” *Ophthalmology*, vol. 113, no. 3, pp. 363–372, 2006.
- [6] K. P. Moloney, J. A. Jacko, B. Vidakovic, F. Sainfort, V. K. Leonard, and B. Shi, “Leveraging data complexity: Pupillary behavior of older adults with visual impairment during hci,” *ACM Transactions on Computer-Human Interaction (TOCHI)*, vol. 13, no. 3, pp. 376–402, 2006.
- [7] J. Engel Jr, A. Bragin, R. Staba, and I. Mody, “High-frequency oscillations: What is normal and what is not?” *Epilepsia*, vol. 50, no. 4, pp. 598–604, 2009.
- [8] G. G. Gregoriou, S. J. Gotts, H. Zhou, and R. Desimone, “High-frequency, long-range coupling between prefrontal and visual cortex during attention,” *Science*, vol. 324, no. 5931, pp. 1207–1210, 2009.
- [9] G. Katul, B. Vidakovic, and J. Albertson, “Estimating global and local scaling exponents in turbulent flows using discrete wavelet transformations,” *Physics of Fluids*, vol. 13, no. 1, pp. 241–250, 2001.

- [10] K. Park and W. Willinger, *Self-similar network traffic and performance evaluation*. Wiley Online Library, 2000.
- [11] T. Woods, T. Preeprem, K. Lee, W. Chang, and B. Vidakovic, “Characterizing exons and introns by regularity of nucleotide strings,” *Biology Direct*, vol. 11, no. 1, p. 6, 2016.
- [12] B. Zhou, “High-frequency data and volatility in foreign-exchange rates,” *Journal of Business & Economic Statistics*, vol. 14, no. 1, pp. 45–52, 1996.
- [13] A. Kolmogorov, “Wienersche spiralen und einige andere interessante kurven im hilbertschen raum,” *Acad Sci USSR (NS)*, vol. 26, pp. 115–118, 1940.
- [14] B. B. Mandelbrot and J. W. Van Ness, “Fractional Brownian motions, fractional noises and applications,” *SIAM Review*, vol. 10, no. 4, pp. 422–437, 1968.
- [15] P. Abry, P. Flandrin, M. S. Taqqu, and D. Veitch, “Self-similarity and long-range dependence through the wavelet lens,” *Theory and Applications of Long-Range Dependence*, pp. 527–556, 2003.
- [16] P. Abry, “Scaling and wavelets: An introductory walk,” *Processes with Long-Range Correlations*, pp. 34–60, 2003.
- [17] P. Abry, P. Flandrin, M. S. Taqqu, and D. Veitch, “Wavelets for the analysis, estimation, and synthesis of scaling data,” *Self-Similar Network Traffic and Performance Evaluation*, pp. 39–88, 2000.
- [18] P. Abry, P. Gonçalves, and P. Flandrin, “Wavelets, spectrum analysis and 1/f processes,” in *Wavelets and Statistics*, Springer, 1995, pp. 15–29.
- [19] P. Abry, P. Goncalves, and J. L. Véhel, *Scaling, Fractals and Wavelets*. John Wiley & Sons, 2013.
- [20] S. Soltani, P. Simard, and D. Boichu, “Estimation of the self-similarity parameter using the wavelet transform,” *Signal Processing*, vol. 84, no. 1, pp. 117–123, 2004.
- [21] H. Shen, Z. Zhu, and T. C. Lee, “Robust estimation of the self-similarity parameter in network traffic using wavelet transform,” *Signal Processing*, vol. 87, no. 9, pp. 2111–2124, 2007.
- [22] C. L. Franzke, T. Graves, N. W. Watkins, R. B. Gramacy, and C. Hughes, “Robustness of estimators of long-range dependence and self-similarity under non-gaussianity,” *Phil. Trans. R. Soc. A*, vol. 370, no. 1962, pp. 1250–1267, 2012.

- [23] J. Park and C. Park, “Robust estimation of the hurst parameter and selection of an onset scaling,” *Statistica Sinica*, pp. 1531–1555, 2009.
- [24] H. Sheng, Y. Chen, and T. Qiu, “On the robustness of hurst estimators,” *IET Signal Processing*, vol. 5, no. 2, pp. 209–225, 2011.
- [25] E. K. Hamilton, S. Jeon, P. R. Cobo, K. S. Lee, and B. Vidakovic, “Diagnostic classification of digital mammograms by wavelet-based spectral tools: A comparative study,” in *2011 IEEE International Conference on Bioinformatics and Biomedicine*, 2011, pp. 384–389.
- [26] H. Theil, “A rank-invariant method of linear and polynomial regression analysis,” in *Henri Theils Contributions to Economics and Econometrics*, Springer, 1992, pp. 345–381.
- [27] M. Kang and B. Vidakovic, “Medl and medla: methods for assessment of scaling by medians of log-squared nondecimated wavelet coefficients,” *ArXiv e-prints*, Mar. 2017. arXiv: 1703.04180 [stat.ME].
- [28] J. W. Tukey, *Exploratory Data Analysis*. Reading, Mass., 1977.
- [29] J. L. Gastwirth and M. L. Cohen, “Small sample behavior of some robust linear estimates of locations,” *Journal of the American Statistical Association*, vol. 65, p. 946, 1970.
- [30] A. DasGupta, *Asymptotic theory of statistics and probability*. Springer Science & Business Media, 2008.
- [31] G. P. Nason and B. W. Silverman, “The stationary wavelet transform and some statistical applications,” in *Wavelets and Statistics*, Springer, 1995, pp. 281–299.
- [32] B. Vidakovic, “Pollen bases and daubechies-lagarias algorithm in matlab,” *Jacket’s Wavelets Website*. <http://www.isye.gatech.edu/~brani/datasoft/DL.pdf>, 2002.
- [33] I. El-Naqa, Y. Yang, M. N. Wernick, N. P. Galatsanos, and R. M. Nishikawa, “A support vector machine approach for detection of microcalcifications,” *IEEE transactions on medical imaging*, vol. 21, no. 12, pp. 1552–1563, 2002.
- [34] P. Kestener, J. M. Lina, P. Saint-Jean, and A. Arneodo, “Wavelet-based multifractal formalism to assist in diagnosis in digitized mammograms,” *Image Analysis & Stereology*, vol. 20, no. 3, pp. 169–174, 2011.
- [35] B. K. Bala and S. Audithan, “Wavelet and curvelet analysis for the classification of microcalcification using mammogram images,” in *Second International Con-*

ference on Current Trends In Engineering and Technology - ICCTET 2014, 2014, pp. 517–521.

- [36] T. Netsch and H.-O. Peitgen, “Scale-space signatures for the detection of clustered microcalcifications in digital mammograms,” *IEEE Transactions on medical imaging*, vol. 18, no. 9, pp. 774–786, 1999.
- [37] T. C. Wang and N. B. Karayiannis, “Detection of microcalcifications in digital mammograms using wavelets,” *IEEE transactions on medical imaging*, vol. 17, no. 4, pp. 498–509, 1998.
- [38] P. T. Reiss and R. T. Ogden, “Functional generalized linear models with images as predictors,” *Biometrics*, vol. 66, no. 1, pp. 61–69, 2010.
- [39] H. Zhou, L. Li, and H. Zhu, “Tensor regression with applications in neuroimaging data analysis,” *Journal of the American Statistical Association*, vol. 108, no. 502, pp. 540–552, 2013.
- [40] V. Zippunikov, B. Caffo, D. M. Yousem, C. Davatzikos, B. S. Schwartz, and C. Crainiceanu, “Functional principal component model for high-dimensional brain imaging,” *NeuroImage*, vol. 58, no. 3, pp. 772–784, 2011.
- [41] P. T. Reiss, R. T. Ogden, J. J. Mann, and R. V. Parsey, “Functional logistic regression with pet imaging data: A voxel-level clinical diagnostic tool,” *Journal of Cerebral Blood Flow & Metabolism*, vol. 25, no. 1_suppl, S635–S635, 2005.
- [42] O. Nicolis, P. Ramírez-Cobo, and B. Vidakovic, “2d wavelet-based spectra with applications,” *Computational Statistics & Data Analysis*, vol. 55, no. 1, pp. 738–751, 2011.
- [43] P. Ramírez-Cobo and B. Vidakovic, “A 2d wavelet-based multiscale approach with applications to the analysis of digital mammograms,” *Computational Statistics & Data Analysis*, vol. 58, pp. 71–81, 2013.
- [44] S. Jeon, O. Nicolis, and B. Vidakovic, “Mammogram diagnostics via 2-d complex wavelet-based self-similarity measures,” *The São Paulo Journal of Mathematical Sciences*, vol. 8, no. 2, pp. 265–284, 2014.
- [45] C. Feng and B. Vidakovic, “Estimation of the hurst exponent using trimean estimators on nondecimated wavelet coefficients,” *ArXiv preprint arXiv:1709.08775*, 2017.
- [46] J. Engel Jr, A. Bragin, R. Staba, and I. Mody, “High-frequency oscillations: What is normal and what is not?” *Epilepsia*, vol. 50, no. 4, pp. 598–604, 2009.

- [47] G. G. Gregoriou, S. J. Gotts, H. Zhou, and R. Desimone, “High-frequency, long-range coupling between prefrontal and visual cortex during attention,” *Science*, vol. 324, no. 5931, pp. 1207–1210, 2009.
- [48] G. Katul, B. Vidakovic, and J. Albertson, “Estimating global and local scaling exponents in turbulent flows using discrete wavelet transformations,” *Physics of Fluids*, vol. 13, no. 1, pp. 241–250, 2001.
- [49] K. Park and W. Willinger, *Self-similar network traffic and performance evaluation*. Wiley Online Library, 2000.
- [50] T. Woods, T. Preeprem, K. Lee, W. Chang, and B. Vidakovic, “Characterizing exons and introns by regularity of nucleotide strings,” *Biology direct*, vol. 11, no. 1, p. 6, 2016.
- [51] B. Zhou, “High-frequency data and volatility in foreign-exchange rates,” *Journal of Business & Economic Statistics*, vol. 14, no. 1, pp. 45–52, 1996.
- [52] B. Vidakovic, *Statistical modeling by wavelets*. John Wiley & Sons, 2009, vol. 503.
- [53] D. B. Percival and A. T. Walden, *Wavelet methods for time series analysis*. Cambridge University Press, 2006, vol. 4.
- [54] D. F. Andrews and F. R. Hampel, *Robust estimates of location: Survey and advances*. Princeton University Press, 2015.
- [55] G. J. L., “On robust procedures,” *Journal of the American Statistical Association*, vol. 61, p. 929, 1966.
- [56] G. J. L. and R. H., “On robust linear estimators,” *Annals of Mathematical Statistics*, vol. 40, p. 24, 1969.
- [57] A. T. Wood and G. Chan, “Simulation of stationary gaussian processes in $[0, 1]^d$,” *Journal of Computational and Graphical Statistics*, vol. 3, no. 4, pp. 409–432, 1994.
- [58] M. Heath, K. Bowyer, D. Kopans, R. Moore, and W. P. Kegelmeyer, “The digital database for screening mammography,” in *Proceedings of the 5th international workshop on digital mammography*, Medical Physics Publishing, 2000, pp. 212–218.
- [59] C. Feng, Y. Mei, and B. Vidakovic, “Mammogram diagnostics using robust wavelet-based estimator of hurst exponent,” in *New Frontiers of Biostatistics and Bioinformatics*, Springer, 2018, pp. 109–140.

- [60] D. F. Gaieski, J. M. Edwards, M. J. Kallan, and B. G. Carr, "Benchmarking the incidence and mortality of severe sepsis in the united states," *Critical Care Medicine*, vol. 41, no. 5, pp. 1167–1174, 2013.
- [61] J.-L. Vincent, S. M. Opal, J. C. Marshall, and K. J. Tracey, "Sepsis definitions: Time for change," *Lancet*, vol. 381, no. 9868, p. 774, 2013.
- [62] D. Kojic, B. H. Siegler, F. Uhle, C. Lichtenstern, P. P. Nawroth, M. A. Weigand, S. Hofer, and T. Brenner, "Are there new approaches for diagnosis, therapy guidance and outcome prediction of sepsis?" *World Journal of Experimental Medicine*, vol. 5, no. 2, p. 50, 2015.
- [63] E. Rezende, J. M. Silva Junior, A. M. Isola, E. V. Campos, C. P. Amendola, and S. L. Almeida, "Epidemiology of severe sepsis in the emergency department and difficulties in the initial assistance," *Clinics*, vol. 63, no. 4, pp. 457–464, 2008.
- [64] A. F. Shorr, S. T. Micek, W. L. Jackson, and M. H. Kollef, "Economic implications of an evidence-based sepsis protocol: Can we improve outcomes and lower costs?" *Critical Care Medicine*, vol. 35, no. 5, pp. 1257–1262, 2007.
- [65] A. Castellanos-Ortega, B. Suberviola, L. A. García-Astudillo, F. Ortiz, J. Llorca, and M. Delgado-Rodríguez, "Late compliance with the sepsis resuscitation bundle: Impact on mortality," *Shock*, vol. 36, no. 6, pp. 542–547, 2011.
- [66] D. Talmor, D. Greenberg, M. D. Howell, A. Lisbon, V. Novack, and N. Shapiro, "The costs and cost-effectiveness of an integrated sepsis treatment protocol," *Critical Care Medicine*, vol. 36, no. 4, pp. 1168–1174, 2008.
- [67] E. Gyang, L. Shieh, L. Forsey, and P. Maggio, "A nurse-driven screening tool for the early identification of sepsis in an intermediate care unit setting," *Journal of Hospital Medicine*, vol. 10, no. 2, pp. 97–103, 2015.
- [68] L. J. Moore, S. L. Jones, L. A. Kreiner, B. McKinley, J. F. Sucher, S. R. Todd, K. L. Turner, A. Valdivia, and F. A. Moore, "Validation of a screening tool for the early identification of sepsis," *Journal of Trauma and Acute Care Surgery*, vol. 66, no. 6, pp. 1539–1547, 2009.
- [69] N. Villegas and L. J. Moore, "Sepsis screening: Current evidence and available tools," *Surgical Infections*, vol. 19, no. 2, pp. 126–130, 2018.
- [70] X. Zhou, Y. Ye, and G. Tang, "Sepsis screening tools in the era of sepsis 3.0," *Surgical Infections*, vol. 19, no. 5, pp. 553–553, 2018.

- [71] E. D. Krebs, T. E. Hassinger, C. A. Guidry, P. S. Berry, N. R. Elwood, and R. G. Sawyer, “Non-utility of sepsis scores for identifying infection in surgical intensive care unit patients,” *The American Journal of Surgery*, 2018.
- [72] M. Singer, C. S. Deutschman, C. W. Seymour, M. Shankar-Hari, D. Annane, M. Bauer, R. Bellomo, G. R. Bernard, J.-D. Chiche, C. M. Coopersmith, *et al.*, “The third international consensus definitions for sepsis and septic shock (sepsis-3),” *JAMA*, vol. 315, no. 8, pp. 801–810, 2016.
- [73] C. W. Seymour, V. X. Liu, T. J. Iwashyna, F. M. Brunkhorst, T. D. Rea, A. Scherag, G. Rubenfeld, J. M. Kahn, M. Shankar-Hari, M. Singer, *et al.*, “Assessment of clinical criteria for sepsis: For the third international consensus definitions for sepsis and septic shock (sepsis-3),” *JAMA*, vol. 315, no. 8, pp. 762–774, 2016.
- [74] E. P. Raith, A. A. Udy, M. Bailey, S. McGloughlin, C. MacIsaac, R. Bellomo, and D. V. Pilcher, “Prognostic accuracy of the sofa score, sirs criteria, and qsofa score for in-hospital mortality among adults with suspected infection admitted to the intensive care unit,” *JAMA*, vol. 317, no. 3, pp. 290–300, 2017.
- [75] R. Miller and D. Siegmund, “Maximally selected chi square statistics,” *Biometrics*, pp. 1011–1016, 1982.
- [76] W. J. Youden, “Index for rating diagnostic tests,” *Cancer*, vol. 3, no. 1, pp. 32–35, 1950.
- [77] J Perkins Neil and F Schisterman Enrique, “The inconsistency of optimal cut-points using two roc based criteria,” *American Journal of Epidemiology*, vol. 163, pp. 670–675, 2006.
- [78] X. Liu, “Classification accuracy and cut point selection,” *Statistics in Medicine*, vol. 31, no. 23, pp. 2676–2686, 2012.
- [79] I. Unal, “Defining an optimal cut-point value in roc analysis: An alternative approach,” *Computational and Mathematical Methods in Medicine*, vol. 2017, 2017.
- [80] S. M. Fernando, A. Tran, M. Taljaard, W. Cheng, B. Rochweg, A. J. Seely, and J. J. Perry, “Prognostic accuracy of the quick sequential organ failure assessment for mortality in patients with suspected infection,” *Ann Intern Med*, vol. 168, no. 4, pp. 266–75, 2018.
- [81] M. Singer and M. Shankar-Hari, “Qsofa, cue confusion,” *Annals of internal medicine*, vol. 168, no. 4, pp. 293–295, 2018.
- [82] S. Tusgul, P.-N. Carron, B. Yersin, T. Calandra, and F. Dami, “Low sensitivity of qsofa, sirs criteria and sepsis definition to identify infected patients at risk of com-

- plication in the prehospital setting and at the emergency department triage,” *Scandinavian Journal of Trauma, Resuscitation and Emergency Medicine*, vol. 25, no. 1, p. 108, 2017.
- [83] M. Brabrand, U. Havshøj, and C. A. Graham, “Validation of the qsofa score for identification of septic patients: A retrospective study,” *European Journal of Internal Medicine*, vol. 36, e35–e36, 2016.
- [84] E. J. Giamarellos-Bourboulis, T. Tsaganos, I Tsangaris, M Lada, C Routsis, D Sina- pidis, M Koupetori, M Bristianou, G Adamis, K Mandragos, *et al.*, “Validation of the new sepsis-3 definitions: Proposal for improvement in early risk identification,” *Clinical Microbiology and Infection*, vol. 23, no. 2, pp. 104–109, 2017.
- [85] M. Dorsett, M. Kroll, C. S. Smith, P. Asaro, S. Y. Liang, and H. P. Moy, “Qsofa has poor sensitivity for prehospital identification of severe sepsis and septic shock,” *Prehospital Emergency Care*, vol. 21, no. 4, pp. 489–497, 2017.
- [86] J. Friedman, T. Hastie, and R. Tibshirani, *The elements of statistical learning*. Springer Series in Statistics New York, 2001, vol. 1.
- [87] T. Hastie, S. Rosset, J. Zhu, and H. Zou, “Multi-class adaboost,” *Statistics and its Interface*, vol. 2, no. 3, pp. 349–360, 2009.
- [88] M. Collins, R. E. Schapire, and Y. Singer, “Logistic regression, adaboost and breg- man distances,” *Machine Learning*, vol. 48, no. 1-3, pp. 253–285, 2002.
- [89] E. R. Swenson, N. D. Bastian, and H. B. Nembhard, “Data analytics in health pro- motion: Health market segmentation and classification of total joint replacement surgery patients,” *Expert Systems with Applications*, vol. 60, pp. 118–129, 2016.
- [90] C. Tucker, Y. Han, H. Black Nembhard, W.-C. Lee, M. Lewis, N. Sterling, and X. Huang, “A data mining methodology for predicting early stage parkinson’s dis- ease using non-invasive, high-dimensional gait sensor data,” *IIE Transactions on Healthcare Systems Engineering*, vol. 5, no. 4, pp. 238–254, 2015.
- [91] C. S. Tucker, I. Behoora, H. B. Nembhard, M. Lewis, N. W. Sterling, and X. Huang, “Machine learning classification of medication adherence in patients with move- ment disorders using non-wearable sensors,” *Computers in Biology and Medicine*, vol. 66, pp. 120–134, 2015.
- [92] F. Jaimes, J. Farbiarz, D. Alvarez, and C. Martínez, “Comparison between logistic regression and neural networks to predict death in patients with suspected sepsis in the emergency room,” *Critical Care*, vol. 9, no. 2, R150, 2005.

- [93] S. W. Thiel, J. M. Rosini, W. Shannon, J. A. Doherty, S. T. Micek, and M. H. Kollef, "Early prediction of septic shock in hospitalized patients," *Journal of Hospital Medicine*, vol. 5, no. 1, pp. 19–25, 2010.
- [94] D. B. Knox, M. J. Lanspa, K. G. Kuttler, S. C. Brewer, and S. M. Brown, "Phenotypic clusters within sepsis-associated multiple organ dysfunction syndrome," *Intensive Care Medicine*, vol. 41, no. 5, pp. 814–822, 2015.
- [95] C. C. Polito, A. Isakov, A. H. Yancey, D. K. Wilson, B. A. Anderson, I. Bloom, G. S. Martin, and J. E. Sevransky, "Prehospital recognition of severe sepsis: Development and validation of a novel ems screening tool," *American Journal of Emergency Medicine*, vol. 33, no. 9, pp. 1119–1125, 2015.
- [96] K. E. Henry, D. N. Hager, P. J. Pronovost, and S. Saria, "A targeted real-time early warning score (trewscore) for septic shock," *Science Translational Medicine*, vol. 7, no. 299, 299ra122–299ra122, 2015.
- [97] S. Mani, A. Ozdas, C. Aliferis, H. A. Varol, Q. Chen, R. Carnevale, Y. Chen, J. Romano-Keeler, H. Nian, and J.-H. Weikamp, "Medical decision support using machine learning for early detection of late-onset neonatal sepsis," *Journal of the American Medical Informatics Association*, vol. 21, no. 2, pp. 326–336, 2014.
- [98] H. M. Giannini, C. Chivers, M. Draugelis, A. Hanish, B. Fuchs, P. Donnelly, M. Lynch, L. Meadows, S. J. Parker, W. D. Schweickert, *et al.*, "Development and implementation of a machine-learning algorithm for early identification of sepsis in a multi-hospital academic healthcare system," in *D15. Critical Care: Do We Have a Crystal Ball? Predicting Clinical Deterioration and Outcome in Critically Ill Patients*, American Thoracic Society, 2017, A7015–A7015.
- [99] D. W. Shimabukuro, C. W. Barton, M. D. Feldman, S. J. Mataraso, and R. Das, "Effect of a machine learning-based severe sepsis prediction algorithm on patient survival and hospital length of stay: A randomised clinical trial," *BMJ Open Respiratory Research*, vol. 4, no. 1, e000234, 2017.
- [100] S. Nemati, A. Holder, F. Razmi, M. D. Stanley, G. D. Clifford, and T. G. Buchman, "An interpretable machine learning model for accurate prediction of sepsis in the icu," *Critical Care Medicine*, vol. 46, no. 4, pp. 547–553, 2018.
- [101] R. C. Bone, R. A. Balk, F. B. Cerra, R. P. Dellinger, A. M. Fein, W. A. Knaus, R. M. Schein, and W. J. Sibbald, "Definitions for sepsis and organ failure and guidelines for the use of innovative therapies in sepsis," *Chest*, vol. 101, no. 6, pp. 1644–1655, 1992.
- [102] M. M. Levy, M. P. Fink, J. C. Marshall, E. Abraham, D. Angus, D. Cook, J. Cohen, S. M. Opal, J.-L. Vincent, G. Ramsay, *et al.*, "2001 sccm/esicm/accp/ats/sis

- international sepsis definitions conference,” *Intensive Care Medicine*, vol. 29, no. 4, pp. 530–538, 2003.
- [103] Y. Freund and R. E. Schapire, “A decision-theoretic generalization of on-line learning and an application to boosting,” *Journal of Computer and System Sciences*, vol. 55, no. 1, pp. 119–139, 1997.
- [104] A. E. Johnson, T. J. Pollard, L. Shen, H. L. Li-wei, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. A. Celi, and R. G. Mark, “Mimic-iii, a freely accessible critical care database,” *Scientific Data*, vol. 3, p. 160 035, 2016.
- [105] T. Pollard and A. Johnson III, *The mimic-iii clinical database*, 2017.
- [106] T. Zhang, B. Yu, *et al.*, “Boosting with early stopping: Convergence and consistency,” *The Annals of Statistics*, vol. 33, no. 4, pp. 1538–1579, 2005.
- [107] Y. Yao, L. Rosasco, and A. Caponnetto, “On early stopping in gradient descent learning,” *Constructive Approximation*, vol. 26, no. 2, pp. 289–315, 2007.
- [108] G. Raskutti, M. J. Wainwright, and B. Yu, “Early stopping and non-parametric regression: An optimal data-dependent stopping rule,” *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 335–366, 2014.
- [109] L. Prechelt, “Early stopping-but when?” In *Neural Networks: Tricks of the trade*, Springer, 1998, pp. 55–69.
- [110] F. Girosi, M. Jones, and T. Poggio, “Regularization theory and neural networks architectures,” *Neural Computation*, vol. 7, no. 2, pp. 219–269, 1995.
- [111] C. Kwong, A. Y. Ling, M. H. Crawford, S. X. Zhao, and N. H. Shah, “A clinical score for predicting atrial fibrillation in patients with cryptogenic stroke or transient ischemic attack,” *Cardiology*, vol. 138, pp. 133–140, 2017.
- [112] Y. Fong, J. Fortner, R. L. Sun, M. F. Brennan, and L. H. Bumgart, “Clinical score for predicting recurrence after hepatic resection for metastatic colorectal cancer,” *Annals of Surgery*, vol. 230, no. 3, p. 309, 1999.
- [113] A. Luna, P. Rebasas, S. Navarro, S. Montmany, D. Coroleu, J. Cabrol, and O. Colomer, “An evaluation of morbidity and mortality in oncologic gastric surgery with the application of possum, p-possum, and o-possum,” *World journal of surgery*, vol. 33, no. 9, pp. 1889–1894, 2009.
- [114] H. Wang, H. Wang, T. Chen, X. Liang, Y. Song, and J. Wang, “Evaluation of the possum, p-possum and e-pass scores in the surgical treatment of hilar cholangiocarcinoma,” *World journal of surgical oncology*, vol. 12, no. 1, p. 191, 2014.

- [115] R. M. Pearse, D. A. Harrison, P. James, D. Watson, C. Hinds, A. Rhodes, R. M. Grounds, and E. D. Bennett, "Identification and characterisation of the high-risk surgical population in the united kingdom," *Critical care*, vol. 10, no. 3, R81, 2006.
- [116] S. F. Khuri, W. G. Henderson, R. G. DePalma, C. Mosca, N. A. Healey, D. J. Kumbhani, *et al.*, "Determinants of long-term survival after major surgery and the adverse effect of postoperative complications," *Annals of surgery*, vol. 242, no. 3, p. 326, 2005.
- [117] S. Moonesinghe, S Harris, M. Mythen, K. Rowan, F. Haddad, M Emberton, and M. Grocott, "Survival after postoperative morbidity: A longitudinal observational cohort study," *British journal of anaesthesia*, vol. 113, no. 6, pp. 977–984, 2014.
- [118] M. A. Healy, A. J. Mullard, D. A. Campbell, and J. B. Dimick, "Hospital and payer costs associated with surgical complications," *JAMA surgery*, vol. 151, no. 9, pp. 823–830, 2016.
- [119] K. Slankamenac, R. Graf, J. Barkun, M. A. Puhan, and P.-A. Clavien, "The comprehensive complication index: A novel continuous scale to measure surgical morbidity," *Annals of surgery*, vol. 258, no. 1, pp. 1–7, 2013.
- [120] K. Slankamenac, N. Nederlof, P. Pessaux, J. De Jonge, B. P. Wijnhoven, S. Breitenstein, C. E. Oberkofler, R. Graf, M. A. Puhan, and P.-A. Clavien, "The comprehensive complication index: A novel and more sensitive endpoint for assessing outcome and reducing sample size in randomized controlled trials," *Annals of surgery*, vol. 260, no. 5, pp. 757–763, 2014.
- [121] S. Ferrari and F. Cribari-Neto, "Beta regression for modelling rates and proportions," *Journal of applied statistics*, vol. 31, no. 7, pp. 799–815, 2004.
- [122] R. Kieschnick and B. D. McCullough, "Regression analysis of variates observed on (0, 1): Percentages, proportions and fractions," *Statistical modelling*, vol. 3, no. 3, pp. 193–213, 2003.
- [123] P. L. Espinheira, S. L. Ferrari, and F. Cribari-Neto, "Influence diagnostics in beta regression," *Computational Statistics & Data Analysis*, vol. 52, no. 9, pp. 4417–4431, 2008.
- [124] ———, "On beta regression residuals," *Journal of Applied Statistics*, vol. 35, no. 4, pp. 407–419, 2008.
- [125] P. Paolino, "Maximum likelihood estimation of models with beta-distributed dependent variables," *Political Analysis*, vol. 9, no. 4, pp. 325–346, 2001.

- [126] M. Smithson and J. Verkuilen, “A better lemon squeezer? maximum-likelihood regression with beta-distributed dependent variables.,” *Psychological methods*, vol. 11, no. 1, p. 54, 2006.
- [127] A. B. Simas, W. Barreto-Souza, and A. V. Rocha, “Improved estimators for a general class of beta regression models,” *Computational Statistics & Data Analysis*, vol. 54, no. 2, pp. 348–366, 2010.
- [128] S. L. Ferrari and E. C. Pinheiro, “Improved likelihood inference in beta regression,” *Journal of Statistical Computation and Simulation*, vol. 81, no. 4, pp. 431–443, 2011.
- [129] C. J. Swearingen, M. M. Castro, and Z. Bursac, “Inflated beta regression: Zero, one and everything in between,” in *SAS Institute, SAS Global Forum*, 2012.
- [130] R. Ospina and S. L. Ferrari, “A general class of zero-or-one inflated beta regression models,” *Computational Statistics & Data Analysis*, vol. 56, no. 6, pp. 1609–1623, 2012.
- [131] L. Scott-Hayward, D. Borchers, M. Burt, S Barco, H. Haas, C. Sasso, and R. Smolowitz, “Use of zero-and one-inflated beta regression to model availability of loggerhead turtles off the east coast of the united states,” 2014.
- [132] I. Khan and S. Morris, “A non-linear beta-binomial regression model for mapping eortc qlq-c30 to the eq-5d-3l in lung cancer patients: A comparison with existing approaches,” *Health and quality of life outcomes*, vol. 12, no. 1, p. 163, 2014.
- [133] R. Ospina and S. L. Ferrari, “Inflated beta distributions,” *Statistical Papers*, vol. 51, no. 1, p. 111, 2010.
- [134] W. Zhao, R. Zhang, Y. Lv, and J. Liu, “Variable selection for varying dispersion beta regression model,” *Journal of Applied Statistics*, vol. 41, no. 1, pp. 95–108, 2014.
- [135] F. M. Bayer and F. Cribari-Neto, “Model selection criteria in beta regression with varying dispersion,” *Communications in Statistics-Simulation and Computation*, vol. 46, no. 1, pp. 729–746, 2017.
- [136] F. Cribari-Neto and T. C. Souza, “Testing inference in variable dispersion beta regressions,” *Journal of Statistical Computation and Simulation*, vol. 82, no. 12, pp. 1827–1843, 2012.
- [137] S. L. Ferrari, P. L. Espinheira, and F. Cribari-Neto, “Diagnostic tools in beta regression with varying dispersion,” *Statistica Neerlandica*, vol. 65, no. 3, pp. 337–351, 2011.

- [138] F. Cribari-Neto and A. Zeileis, “Beta regression in r,” 2009.
- [139] L.-C. Chien, “Diagnostic plots in beta-regression models,” *Journal of Applied Statistics*, vol. 38, no. 8, pp. 1607–1622, 2011.
- [140] A. V. Rocha and A. B. Simas, “Influence diagnostics in a general class of beta regression models,” *Test*, vol. 20, no. 1, pp. 95–119, 2011.
- [141] M. Yuan and Y. Lin, “Model selection and estimation in regression with grouped variables,” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 68, no. 1, pp. 49–67, 2006.
- [142] E. Candes, T. Tao, *et al.*, “The dantzig selector: Statistical estimation when p is much larger than n ,” *The annals of Statistics*, vol. 35, no. 6, pp. 2313–2351, 2007.
- [143] M. Y. Park and T. Hastie, “L1-regularization path algorithm for generalized linear models,” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 69, no. 4, pp. 659–677, 2007.
- [144] J. Fan and R. Li, “Variable selection via nonconcave penalized likelihood and its oracle properties,” *Journal of the American statistical Association*, vol. 96, no. 456, pp. 1348–1360, 2001.
- [145] J. H. Friedman, “Fast sparse regression and classification,” *International Journal of Forecasting*, vol. 28, no. 3, pp. 722–738, 2012.
- [146] T. Hastie, S. Rosset, R. Tibshirani, and J. Zhu, “The entire regularization path for the support vector machine,” *Journal of Machine Learning Research*, vol. 5, no. Oct, pp. 1391–1415, 2004.
- [147] J. Friedman, T. Hastie, and R. Tibshirani, “Sparse inverse covariance estimation with the graphical lasso,” *Biostatistics*, vol. 9, no. 3, pp. 432–441, 2008.
- [148] R. Tibshirani, “Regression shrinkage and selection via the lasso,” *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996.
- [149] H. Zou and T. Hastie, “Regularization and variable selection via the elastic net,” *Journal of the royal statistical society: Series B (statistical methodology)*, vol. 67, no. 2, pp. 301–320, 2005.
- [150] J. Friedman, T. Hastie, and R. Tibshirani, “Regularization paths for generalized linear models via coordinate descent,” *Journal of statistical software*, vol. 33, no. 1, p. 1, 2010.

- [151] J. D. Lee, Y. Sun, and M. Saunders, “Proximal newton-type methods for convex optimization,” in *Advances in Neural Information Processing Systems*, 2012, pp. 827–835.
- [152] H. Wang, R. Li, and C.-L. Tsai, “Tuning parameter selectors for the smoothly clipped absolute deviation method,” *Biometrika*, vol. 94, no. 3, pp. 553–568, 2007.
- [153] M. Grant and S. Boyd, *Cvx: Matlab software for disciplined convex programming, version 2.1*, <http://cvxr.com/cvx>, Mar. 2014.
- [154] ———, “Graph implementations for nonsmooth convex programs,” in *Recent Advances in Learning and Control*, ser. Lecture Notes in Control and Information Sciences, V. Blondel, S. Boyd, and H. Kimura, Eds., http://stanford.edu/~boyd/graph_dcp.html, Springer-Verlag Limited, 2008, pp. 95–110.
- [155] T. S. Breusch and A. R. Pagan, “A simple test for heteroscedasticity and random coefficient variation,” *Econometrica: Journal of the Econometric Society*, pp. 1287–1294, 1979.
- [156] R. Koenker, “A note on studentizing a test for heteroscedasticity,” *Journal of Econometrics*, vol. 17, no. 1, pp. 107–112, 1981.