WAVELET METHODS AND STATISTICAL APPLICATIONS: NETWORK SECURITY AND BIOINFORMATICS

A Dissertation

by

DEUKWOO KWON

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

August 2005

Major Subject: Statistics

WAVELET METHODS AND STATISTICAL APPLICATIONS: NETWORK SECURITY AND BIOINFORMATICS

A Dissertation

by

DEUKWOO KWON

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Approved by:

Chair of Committee,	Marina Vannucci
Committee Members,	Michael T. Longnecker
	Johan Lim
	A.L. Narasimha Reddy
Head of Department,	Simon J. Sheather

August 2005

Major Subject: Statistics

ABSTRACT

Wavelet Methods and Statistical Applications:
Network Security and Bioinformatics. (August 2005)

Deukwoo Kwon, B.A., Yonsei University, Korea;
M.B.A., Korea Advanced Institute of Science and Technology;
M.S., Texas A&M University

Chair of Advisory Committee: Dr. Marina Vannucci

Wavelet methods possess versatile properties for statistical applications. We would like to explore the advantages of using wavelets in the analyses in two different research areas. First of all, we develop an integrated tool for online detection of network anomalies. We consider statistical change point detection algorithms, for both local changes in the variance and for jumps detection, and propose modified versions of these algorithms based on moving window techniques. We investigate performances on simulated data and on network traffic data with several superimposed attacks. All detection methods are based on wavelet packets transformations.

We also propose a Bayesian model for the analysis of high-throughput data where the outcome of interest has a natural ordering. The method provides a unified approach for identifying relevant markers and predicting class memberships. This is accomplished by building a stochastic search variable selection method into an ordinal model. We apply the methodology to the analysis of proteomic studies in prostate cancer. We explore wavelet-based techniques to remove noise from the protein mass spectra. The goal is to identify protein markers associated with prostate-specific antigen (PSA) level, an ordinal diagnostic measure currently used to stratify patients into

different risk groups.

To my parents

ACKNOWLEDGEMENTS

I would like to express special thanks to my advisor Dr. Marina Vannucci. Without her guidance and support this dissertation would not have existed. Her kindness and encouragement made the entire journey of this research very smooth and delightful. I am also grateful for Drs. Michael T. Longnecker, Johan Lim and A.L. Narasimha Reddy as committee members and Dr. Mahlet Tadesse. Their critical readings and helpful comments have made the dissertation richer.

I thank my parents. They always support whatever I decide. I thank my two sisters, Mikyeong and Ohkyeong, and their families. I give special thanks to Seonghee Lim. She gave me a pivotal moment in my life. She made me pursue my study. I wish she find her own way and succeed in her study. I also thank my friends, Jinyeong Park, Eunkyeong Park, Byeola Kim, Hojin Lee, Jeesun Jung, Joonjin Song, Kyongryun Kim, and Sinae Kim, who helped and encouraged me in many ways.

TABLE OF CONTENTS

CHAPTER Page

LIST OF FIGURES

LIST OF TABLES

CHAPTER I

INTRODUCTION

The main objective for this dissertation is to develop statistical methodologies for network security and for bioinformatics. In the dissertation we focus on applications of wavelet methods to the above two fields. Wavelet methods have been introduced to the statistical community in the last few years. Wavelets have versatile features, such as the ability to compress and/or denoise signals, they allow multi-scale decompositions, and possess time-frequency localization properties.

We treat the detection of network anomaly in network traffics and the problem of cancer classification with proteomic data in bioinformatics. While multi-scale decompositions and whitening property of wavelets are beneficial tools in detecting change points in the topic of network security, denoising procedure is the crucial tool for classification problem in proteomic data along with Bayesian methodology. Before we discuss two main parts for statistical applications we need to summarize theoretical foundations for wavelet methods in order to grasp how to use thes wavelet methods in the following applications.

In the first part of dissertation we propose a novel approach to detect network anomalies, which are particulary malicious attacks against a large scale network such as university network systems or commercial websites. We mainly focus on the online detection algorithms (or real-time detector). We define the performance of these algorithms in terms of detection delay time and number of false alarms. We prefer

The format and style follow that of *Journal of the American Statistical Association*.

short delay time and less false alarms. These two factors are related to each other reprociprocally. It means we cannot minimize two at the same time. The performance of on-line detection algorithms matters since the impact of network anomaly is enormous so that the quick reaction to the attacks is integral in the management of a large scale network.

The second part of dissertation deals with classification problem where the response variable is naturally ordered. Here we use prostate cancer proteomic mass spectra data with help of wavelet thresholding. This part comprises preprocessing procedure for prostate data and Bayesian ordinal probit analysis with variable selection. The purpose of this part is to find biomarkers for the prostate cancer mass spectra data. Bayesian variable selection plays an important role in identifying biomarkers.

CHAPTER II

WAVELETS

2.1 Introduction

In this chapter we describe the versatile features of wavelets in statistical analyses. Although we deal only with discrete wavelet transforms (DWT) we begin with the exposition of continuous wavelet transform (CWT). We provide the description of the standard discrete wavelet transform with its variants such as the maximal overlap discrete wavelet transform (MODWT), discrete wavelet packet transform (DWPT), and the combination of the two transforms, maximal overlap discrete wavelet packet transform (MODWPT).

Wavelet methods is one of orthogonal transformation which transforms data from original domain (typically time domain) to wavelet domain. This transformation enables us to give analytic tools in various fields . Analytic tools comprise denoising, nonlinear approximation through thresholding in signal processing, nonparametric function estimation, data compression in image processing, time-scale decomposition for time series analysis, and approximate decorrelation. Furthermore, we can enjoy efficient computations when using wavelet methods due to pyramidal algorithm in multiresolution analysis by Mallat (1989), which is faster than fast Fourier transform (FFT). Ogden (1997) and Vidakovic (1999) provide good references for wavelet methods in statistical analyses.

2.2 Basic concepts in wavelets

Wavelets can be considered in two ways: function approximation approach (or projection approach) and filtering approach (or signal processing approach).

In function approximation any function in $L_2(\mathbb{R})$ space can be written as a linear combinations of wavelet functions as follows:

$$\begin{aligned} f(t) &= \sum_j \sum_k a_{j,k}\phi_{j,k}(t), \\ &= \sum_k a_{j_0,k}\phi_{j_0,k}(t) + \sum_{j \geq j_0} \sum_k b_{j,k}\psi_{j,k}(t) \end{aligned}$$

where $a_{j,k} = \langle f(t), \phi_{j,k}(t) \rangle$, $b_{j,k} = \langle f(t), \psi_{j,k}(t) \rangle$ and $\phi_{j,k}, \psi_{j,k}$ are father and mother function respectively.

Scaling functions $\phi_{j,k}$ and wavelet functions $\psi_{j,k}$ ought to satisfy the following conditions:

$$\begin{aligned} \int \phi_{j,k}(t)\phi_{j,k'}(t)dt &= \delta_{k,k'} \\ \int \phi_{j,k}(t)\psi_{j',k'}(t)dt &= 0 \\ \int \psi_{j,k}(t)\psi_{j',k'}(t)dt &= \delta_{j,j'}\delta_{k,k'} \end{aligned}$$

where

$$\delta_{i,j} = \begin{cases} 1 & \text{if} \quad i = j \\ 0 & \text{if} \quad i \neq j \end{cases}$$

Smoothness, compact support, and asymmetry of scaling and wavelet functions distiguish from various functions such as Haar, Daubechies, and so on. See Daubechies (1992).

We can consider function approximation as projection with multiresolution analysis which connects to filtering approach in signal processing. First of all we consider spanned spaces by functions ϕ and ψ. Function spaces spanned by functions ϕ has a nested structure as follows.

Let $\mathcal{V}_j$ is a closed subspace spanned by $\phi_{j,k}, k \subset \mathbb{Z}$. The sequence of subspaces has the following properties:

$$\ldots \in \mathcal{V}_{-1} \in \mathcal{V}_0 \in \mathcal{V}_1 \in \mathcal{V}_2 \in \ldots$$
$$\cap_{j\in\mathbb{Z}}\mathcal{V}_j = \varnothing, \ \ \overline{\cup_{j\in\mathbb{Z}}\mathcal{V}_j} = L_2(\mathbb{R})$$

Wavelet spaces $\mathcal{W}_j$ are spanned by functions ψ. These spaces is related to the above spaces such as:

$$\begin{aligned} \mathcal{V}_{j+1} &= \mathcal{V}_j \bigoplus \mathcal{W}_j \\ L_2(\mathbb{R}) &= \cdots \bigoplus \mathcal{W}_{j-1} \bigoplus \mathcal{W}_j \bigoplus \mathcal{W}_{j+1} \bigoplus \cdots \end{aligned}$$

Hence by the multiresolution analysis we can rewrite function as follows.

$$f(t) = P^{j_0} f(t) + \sum_{j>j_0}\sum_k b_{j,k}\psi_{j,k}(t)$$

The first term in the above equatin is the projection of function $f(t)$ on V_{j_0}. The second sum is also rewritten as projections of function $f(t)$ on wavelet spaces $\mathcal{W}_l$ $l \geq j_0, l \in \mathbb{Z}$. At given j, space $\mathcal{V}_j$ and $\mathcal{W}_j$ are orthogonal to each other.

So far we focus on continous wavelet transform, we turn to the discrete wavelet transform. We can define functions $\phi_{j,k}$ and $\psi_{j,k}$ as follows:

$$\begin{aligned} \phi_{j,k}(t) &= 2^j\phi(2^jt - k) \\ \psi_{j,k}(t) &= 2^j\psi(2^jt - k) \end{aligned}$$

From the MRA we have the relations as follows. For any function in $\mathcal{V}_0$ can be written as a linear combination of the basis function $\phi_{1,k} = \sqrt{2}\phi(2t - k)$. Hence we have

$$\phi(t) = \sum_k h(k)\phi_{1,k}.$$

Due to the orthogonal complement $\mathcal{W}_j$ of $\mathcal{V}_j$ to $\mathcal{V}_{j+1}$ we define function ψ as follows.

$$\psi(t) = \sqrt{2}\sum_k (-1)^k h(-k+1)\phi(2t-k) = \sqrt{2}\sum_k g(k)\phi(2t-k).$$

The sequences $\{h(k), k \in \mathbb{Z}\}$ and $\{gh(k), k \in \mathbb{Z}\}$ are quadrature mirror filters in signal processing, that is, $g(k) = (-1)^k h(1-k)$. These filters are called as low-pass and high-pass filters, respectively.

2.3 Discrete wavelet transforms

Now we deal with a finite sequence of time series. Let $X = (x_0, \dots, x_{T-1})$ be a vector of observations from a stochastic process. The DWT is an orthogonal transformation of the data that operates via recursive filters according to the pyramidal algorithm proposed by Mallat (1989). If $T = 2^J$ the algorithm produces scaling coefficients at a coarsest level J, describing global features of the data, and wavelet coefficients at a number of finer scales $1, \dots, J$ describing local features. We denote with h and g the wavelet and scaling filter, respectively, and with L the width of the filters. At the first level, $j = 1$, wavelet coefficients $w_{1,t}$ and scaling coefficients $v_{1,t}$ are defined as

$$w_{1,t} = \sum_{l=0}^{L-1} h_l x_{2t+1-l \bmod N}, \quad v_{1,t} = \sum_{l=0}^{L-1} g_l x_{2t+1-l \bmod N}$$

The wavelet coefficients $w_{2,t}$ and scaling coefficients $v_{2,t}$ at level 2 are computed from the scaling coefficients at level 1 as follows

$$w_{2,t} = \sum_{l=0}^{L-1} h_l v_{1,2t+1-l \bmod N}, \quad v_{2,t} = \sum_{l=0}^{L-1} g_l v_{1,2t+1-l \bmod N}$$

Similary, at levels $j = 3, \dots, J$ the wavelet and scaling coefficients are obtained as

$$w_{j,t} = \sum_{l=0}^{L-1} h_l v_{j-1,2t+1-l \bmod N}, \quad v_{j,t} = \sum_{l=0}^{L-1} g_l v_{j-1,2t+1-l \bmod N}$$

Due to the decimating operator, at level j we have $\frac{T}{2^j}$ scaling and wavelet coefficients.

2.4 Maximal overlap wavelet transforms

In contrast to the DWT, the maximal overlap wavelet transform (MODWT) does not decimate the coefficients and therefore the number of scaling and wavelet coefficients at every level of the transfom is the same as the number of sample observations. Thus, the MODWT is also called undecimated DWT. The MODWT uses circular shifts of the scaling and wavelet filters. Although it loses the orthogonality and efficiency in computation, this transform obtains flexibility on the restriction on the sample size and invariance to circularly shifting the original data. Wavelet coefficients, $\tilde{w}_{j,t}$ and scaling coefficients, $\tilde{v}_{j,t}$ at levels $j, j = 1, \ldots, J$ are obtained as follows.

$$\tilde{w}_{1,t} = \sum_{l=0}^{L-1} \tilde{g}_l x_{t-l \bmod N}, \quad \tilde{v}_{1,t} = \sum_{l=0}^{L-1} \tilde{h}_l x_{t-l \bmod N}$$

$$\tilde{w}_{2,t} = \sum_{l=0}^{L-1} \tilde{g}_l \tilde{v}_{1,t-l \bmod N}, \quad \tilde{v}_{2,t} = \sum_{l=0}^{L-1} \tilde{h}_l \tilde{v}_{1,t-l \bmod N}$$

$$\tilde{w}_{j,t} = \sum_{l=0}^{L-1} \tilde{g}_l \tilde{v}_{j-1,t-l \bmod N}, \quad \tilde{v}_{j,t} = \sum_{l=0}^{L-1} \tilde{h}_l \tilde{v}_{j-1,t-l \bmod N}$$

The wavelet and scaling filters, $\tilde{g}_j,\ \tilde{h}_j$ are rescaled as $\tilde{g}_j = g_j/2^j,\ \tilde{h}_j = h_j/2^j$.

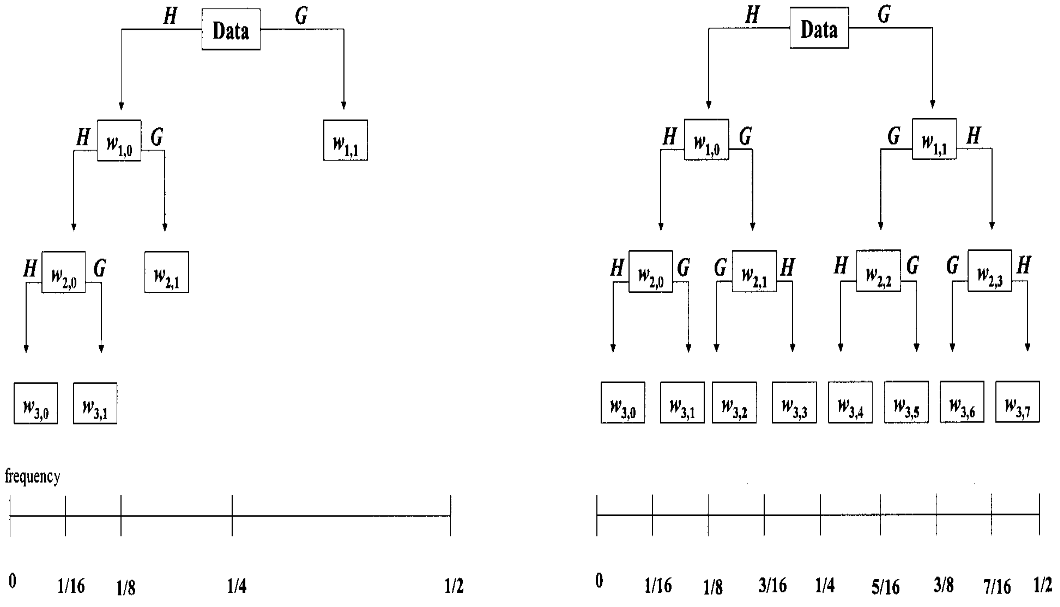


Figure 1: DWT and DWPT

2.5 Wavelet packet transforms

Wavelet packets, Wickerhauser (1994), induce a finer partition of the frequency space. We show this finer partitions of the freqency space in Figure 1. In contrast to the dyadic partitions of the traditional DWT in the left panel of Figure 1 wavelet packet transforms provide us equal-length frequency partitions. In the discrete wavelet packet transform (DWPT) or the undecimated version (MODWPT) both the scaling and wavelet coefficients are subject to the high-pass and low-pass filtering when computing the next level scaling and wavelet coefficients. With the standard transforms, scaling coefficients identify the frequency band $[0, 1/2^{J+1}]$, with J the coarsest level, while wavelet coefficients at level j describe the frequency band $[1/2^{j+1}, 1/2^{j}]$.

The discrete packet wavelet tranforms, DWPT and MODWPT, on the other hand, partition the whole frequency band, $[0, 1/2]$, into equal length frequency bands. For example, at a given level j, we have 2^j frequency partitions with equal length. This finer partition induced by the DWPT implies better decorrelation properties, as exploited in Percival et al. (2000), Whitcher (2001) and Gabbanini et al. (2004).
As a filtering of the original time series the MODWPT can be written as

$$\tilde{w}_{j,n,t} = \sum_{l=0}^{L-1} \tilde{f}_{j,n,l} x_{(t-l) \bmod T},$$

for $n = 0, \ldots, T-1$, where

$$\tilde{f}_{j,n,l} = \sum_{l=0}^{L-1} \tilde{f}_{n,k} \tilde{f}_{j-1,\lfloor n/2 \rfloor, l-2^{j-1}k}, \quad 0 \leq l \leq L-1$$

with

$$\tilde{f}_{n,l} = \begin{cases} \tilde{g}_l & \text{if } n \text{ mod 4=0 or 3} \\ \tilde{h}_l & \text{if } n \text{ mod 4=1 or 2} \end{cases}$$

and $\tilde{g}_l = (-1)^{l-1} \tilde{f}_{L-l-1}$, and such that $\{\tilde{f}_{1,0,l} = \tilde{g}_l,\ 0 \leq l \leq L-1\}$ and $\{\tilde{f}_{1,1,l} = \tilde{h}_l,\ 0 \leq l \leq L-1\}$.

2.6 Wavelet theresholding

Wavelet thresholding technique is one of good approach to remove noises in the data. More generally this wavelet thresholding is a particular case of shrinkage techniques. Hereafter we only deal with thresholding. Consider the standard univariate regression model:

$$y_i = f(x_i) + \sigma \epsilon_i, \quad \epsilon_i \sim N(0, 1),\ i = 1, \ldots, n$$

Our intention is that we want to extract true feature or signal f from the data by removing noises. We can reformulate the above problem with wavelet transforms. After the wavelet transform we get the following model in the wavelet domain:

$$d_i = \theta_i + \sigma \epsilon'_i,\ i = 1, \ldots, n,$$

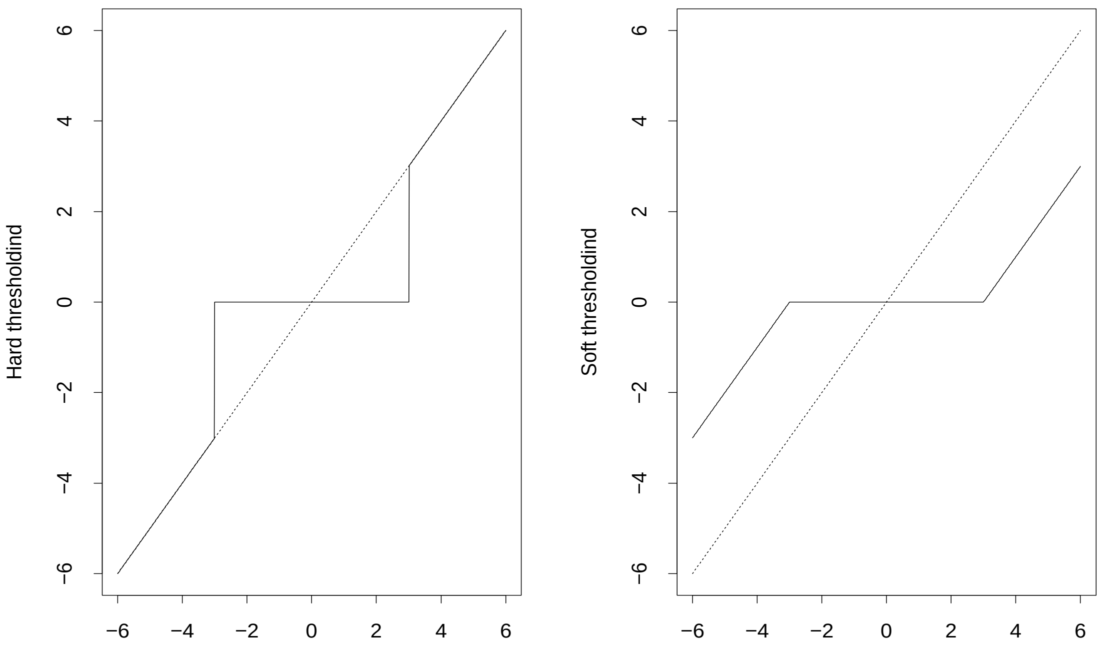


Figure 2: Hard and soft thresholding rule

where d_i is a empirical wavelet coefficient and θ_i true wavelet coefficient. Here due to the orthogonality of wavelet transform ϵ' is identical distribution to that of ϵ. We need to choose thresholding policies. There are two most common choices for the thresholding policies which are hard and soft thresholding rule. We show these two rules in Figure 2. We also give the mathematical expressions for these two rules as follows:

$$\begin{aligned}\delta^h(d,\lambda) &= d\mathbf{1}_{|d|>\lambda},\ \lambda \geq 0, d \in \mathbb{R} \\ \delta^s(d,\lambda) &= (d - \mathrm{sgn}(d)\lambda)d\mathbf{1}_{|d|>\lambda},\ \lambda \geq 0, d \in \mathbb{R}\end{aligned}$$

The key element in wavelet thresholding technique is to obtain an appropriate value for λ. The simplest one is the 'universal' threshold proposed by Donoho and Johnstone (1994). Alternative is for selecting a threshold value by minimizing Stein's unbiased estimator of risk (see Stein 1981) suggested by Donoho and Johnstone (1995).

CHAPTER III

APPLICATION FOR NETWORK SECURITY

3.1 Introduction

In this chapter we investigate the performances of an integrated tool for the detection of network anomalies with the goal of quickly identifying malicious attacks. Detection of network anomalies is a crucial task in network traffic management. Here we look at a network anomaly as a possible attack by a malicious user. Large scale network attacks cause huge costs and a waste of network resources. Early detection allows quick actions and minimizes network damage. In statistical terms, the detection of an anomaly can be considered as a change point problem. In this paper we consider two kinds of detection methods: Those that detect changes in the local variance of the data and those that detect jumps in the observed data. All statistical methods we consider are wavelet-based. Wavelet transformations have been proven to be a valid tool for the analysis of network traffic, mainly because of their locality and decorrelation properties, see for example Riedi et al. (1999), Gilbert et al. (1999), Gilbert (2001), Resnick et al. (2003) and Kim et al. (2004). We look at the implementation of the detection methods based on wavelet packet transformations. We explore performances on simulated data. We also analyze the trace data used in Kim et al. (2004), where the authors propose a novel definition of data correlation for the analysis of traffic packets and classify various types of network attacks as either variance changes or sharp jumps.

For detection we consider the iterated cumulative sums of squares (ICSS) algorithm and the Schwarz information criterion (SIC) algorithm, for the identification of multiple variance change points in sequence data, and the approach suggested by

Wang (1995) for the detection of sharp jumps and cusps in the data. We explore the implementation of these detection methods based on wavelet packets and assess performances in detecting network traffic attacks in real-time. The ICSS algorithm was originally proposed by Inclán and Tiao (1994) while Chen and Gupta (1996) suggested the use of the SIC algorithm for change detection. Whitcher et al. (2000) adapted the ICSS algorithm to discrete wavelet transforms (DWT) and to maximal overlap discrete wavelet transforms (MODWT), also knows as "non-decimated", "translation invariant" or "stationary". Their work is limited to the detection of variance change points for data that show long-range dependence (LRD). Gabbanini et al. (2004) extended the ICSS procedure to discrete wavelet packet transforms (DWPT) and maximal overlap discrete wavelet packet transforms (MODWPT). The use of wavelet packets allowed them to analyze a broader class of data than LRD.

Here we exploit the Gabbanini et al. (2004) method to see how effectively we can detect network traffic anomalies caused by malicious users' network attacks. While Gabbanini et al. (2004) used only the ICSS algorithm, we implement both the SIC and the ICSS algorithms based on wavelet packets. In addition, we extend the method of Wang to maximal overlap wavelet packets, i.e. MODWPT. In the sequel we will use the term "packet" with two different meanings. In network traffic terminology, data information is partitioned into small "chunks" called packets. The header of the packet contains useful information such as the addresses (source and destination) and the packet count. In wavelet theory terminology, the term packet indicates the particular frequency band at which the coefficients of a "packet" transform are associated.

3.2 Detection methods

In this section we describe two kinds of detection methods: Those that detect changes in the local variance of the data and those that detect jumps in the observed data. In the next section we will discuss our adaption of these methods to wavelet packets and related implementation issues.

3.2.1 Variance change points detection algorithms

We first summarize the ICSS and SIC detection algorithms for the detection of variance change points and describe a binary segmentation procedure that allows the adaption of these methods to the detection of multiple change points.

The iterated cumulative sums of squares (ICSS) algorithm aims at testing and identifying multiple variance changes in a sequence of independent observations. Null and alternative hypotheses are specified as

$$H_0 : \sigma_1^2 = \sigma_2^2 = \ldots = \sigma_T^2 \quad \text{versus} \quad H_a : \sigma_1^2 = \cdots = \sigma_k^2 \neq \sigma_{k+1}^2 = \cdots = \sigma_T^2.$$

We denote with $C_k = \sum_{t=1}^{k} x_t^2$ the cumulative sum of squares of a series of uncorrelated random variables $\{x_t\}$ with mean 0 and variances $\sigma_t^2,\ t = 1, \ldots, T$. The test statistic is $D = \max(D^+, D^-)$ where

$$\begin{aligned} D^+ &= \max_{1 \leq k \leq T-1} \left(\frac{k+1}{T} - P_k \right) \\ D^- &= \max_{1 \leq k \leq T-1} \left(P_k - \frac{k}{T} \right) \\ P_k &= \frac{C_k}{C_T},\ k = 1, \ldots, T. \end{aligned}$$

Variance change points are located by looking at $k^* = \text{argmax}_k D$. When the maximum absolute value of D exceeds a certain predetermined value, then we take the point k^* as the change point estimate. Whitcher et al. (2000) obtained predetermined values for D under the null hypothesis by using Monte Carlo simulation. Inclán and Tiao (2004) showed that when the random variables $\{x_t\}$ are independent distributed the asymptotic distribution of D is that one of a Brownian bridge. Whitcher et al. (2000) suggested to use at least $T = 128$ sample size to conform with this asymptotic approximation.

The Schwarz information criterion (SIC) was suggested by Schwarz (1978) and is one of the modifications of Akaike information criterion (AIC) introduced by Akaike (1974). These criteria are useful tools for model selection. Let $\{x_t\}$ be a sequence of independent and identically distributed random variables with probability density function $f(\cdot|\theta)$, where f is a model with K parameters, that is,

$$\begin{aligned}\text{Model}(k) &= \{f(\cdot|\theta) : \theta = (\theta_1, \theta_2, \ldots, \theta_K), \theta \in \boldsymbol{\Theta}_k\}\\ \text{where } \boldsymbol{\Theta}_k &= \{\boldsymbol{\Theta}_k : \theta_{k+1} = \theta_{k+2} = \cdots = \theta_K\}, \quad k = 1, \ldots, K-1.\end{aligned}$$

The SIC is defined as $-2\log L(\bar{\theta_k}) + p\log T$, where $L(\bar{\theta_k})$ is the maximum likelihood function for the model(k), p is the number of parameters in the model, and n is the total number of samples. We specify the form of $SIC(T)$ and $SIC(k)$ as follows

$$\begin{aligned}SIC(T) &= T\log 2\pi + T\log\hat{\sigma}^2 + T + \log T\\ SIC(k) &= T\log 2\pi + k\log\hat{\sigma}_1^2 + (T-k)\log\hat{\sigma}_T^2 + T + 2\log T\end{aligned}$$

where

$$\hat{\sigma}^2 = \frac{1}{T}\sum_{i=1}^{T}(x_i - \bar{x})^2, \quad \hat{\sigma}_1^2 = \frac{1}{k}\sum_{i=1}^{k}(x_i - \bar{x})^2, \quad \text{and} \quad \hat{\sigma}_T^2 = \frac{1}{(T-k)}\sum_{i=k+1}^{T}(x_i - \bar{x})^2.$$

Under the same null and alternative hypotheses described above for the case of the ICSS algorithm, the null hypothesis is now rejected based on the principle of minimum information criterion, that is, we reject if $SIC(T) \geq \min_{2\leq k\leq T-2} SIC(k)$ and estimate the change point as $\hat{k}$ such that

$$SIC(\hat{k}) = \min_{2\leq k\leq T-2} SIC(k).$$

Notice that we can only detect change points that occur between the second and $(T-2)^{th}$ point.

The SIC algorithm does not require knowledge of the distribution of the test statistic. A modification of the method, more robust to data fluctuation, introduces a significant level α and its corresponding critical value C_α so that the null hypothesis is rejected if $SIC(T) \geq \min_{2\leq k\leq T-2} SIC(k) + C_\alpha$. The value C_α can be determined such that

$$1-\alpha = P\left[SIC(T) < \min_{2\leq k\leq T-2} SIC(k) + C_\alpha | H_0\right],$$

see Chen and Gupta (1996).

3.2.2 The binary segmentation procedure

Methods described above were designed for location of single change points. In the application section we will use the binary segmentation procedure to test and locate multiple change points. At the first stage of the procedure we test the null hypothesis for the whole data. If we do not reject H_0 we declare that there is no change point in the whole sequence, otherwise we divide the data into two sub-sequences as determined by the change point located. At the second stage we test the two sub-sequences and repeat the above procedure until we do not find any further change point. Several candidate change points may result from this procedure. At the third stage we check these points as follows. For a given possible change point we determine the sub-sequence between the previous possible change point and the next change point and repeat the test. If we still reject H_0 we keep this point as a change point, otherwise we remove it from the list of candidates. This confirmatory step helps to reduce masking effect and to get more reliable change point estimates. Inclán and Tiao (1994) describe this procedure in detail.

3.2.3 Multiple jumps detection: The Wang's method

Wang's algorithm enables us to detect sudden jumps and sharp cusps in a time series by using discrete wavelet transforms. The idea is simple to understand: A sudden jump affects the magnitudes of wavelet coefficients, thus one can set a threshold level to identify the location at which the jump occurs. Wang suggested to apply the DWT to the data and use the universal threshold of Donoho and Johnstone (1994),

$$\text{Universal threshold} \quad \lambda = \hat{\sigma}\sqrt{2\log n}$$

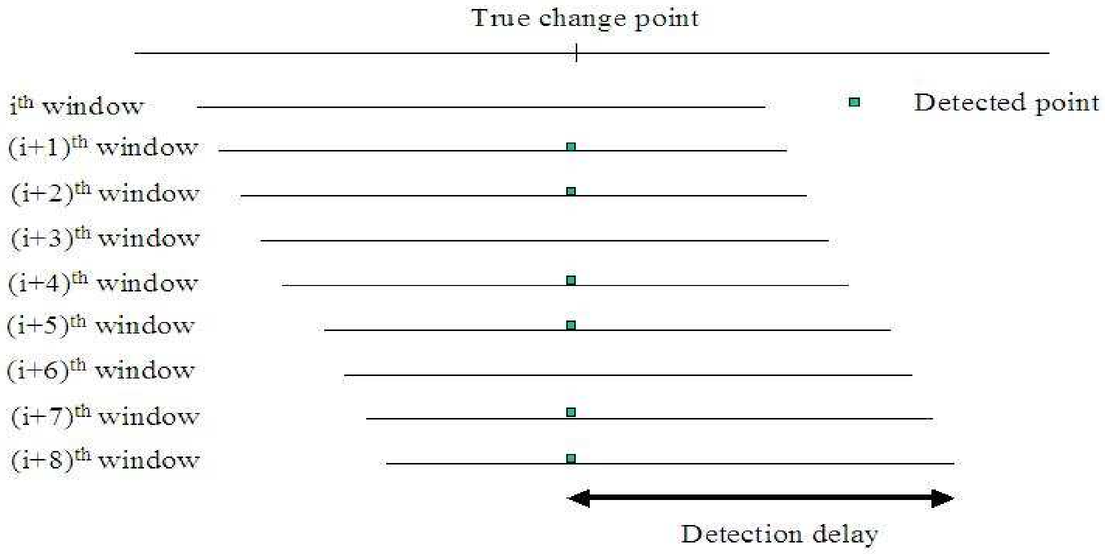
$$\hat{\sigma} = 1.4826 \cdot \text{MEDIAN}[|d^{J-1} - \text{MEDIAN}(d^{J-1})|]$$

where d^{J-1} is the vector of the finest wavelet coefficients of the wavelet transform and $\hat{\sigma}$ is the MAD estimate. Points above the threshold in absolute value are declared jump points.

3.3 Detection schemes

We implement the detection methods previously described using wavelet packet transformations. We use a moving window approach so that the methods can be used for online detection. We indicate these modified procedures as MWICSS (moving window ICSS), MWSIC (moving window SIC) and MWWJ (moving window Wang's jump detection). Whitcher et al. suggested that the sample size for the ICSS algorithm be at least 128 for better approximation. In the next section we investigate performances for several different window sizes. We use the same window lengths for the MWSIC, for a better comparison. For the MWWJ algorithm we also try smaller sizes.

Having chosen the length of the window, the data sequence is examined for change points by sliding the window along the data one point at the time and recording all change points detected. For all detection tests we use a 0.05 significance level. Detected points indicate network anomalies. We declare an anomaly to be a potential attack if it is detected by our procedures in a number of consecutive windows. In other words, we look at the detection frequency as the number of times the anomaly is detected and declare an attack if this exceeds a preselected threshold value. Our moving window procedure and the calculation of the detection frequency is explained in Figure 3, where we use a square symbol to indicate whether the point is detected in a particular window. With a preselected threshold of 6 or higher the point in the figure would be declared an attack. The choice of the threshold implies a trade-off between fast detection and false alarms. Specifically, we want to detect changes as fast as possible after they occur but also want to avoid false alarms. As the threshold value increases we are able to avoid more and more false alarms but with an increase in the detection delay. In the analyses reported here we aimed at decreasing the detection delay for a given false alarm level and look at the mean delay as a performance



Decision delay: the time of which detection frequency is equal to 6 (if *k* is set to 6) – true change point

Figure 3: Schematic representation of the moving window and detection frequency procedures

measure for online detection.

We now give step-by-step descriptions of the implementations of the detection procedures we propose.

3.4 Procedure for variance change detection

In a generic window of size m we test for variance change points as follows.

- Step I: We apply the DWPT and MODWPT. The maximum level of the transforms depends on the length of window. Whitcher et. al. recommend to use at least 128 data points to implement the variance change test. Moreover, we want to apply to the coefficients the Ljung-Box test for autocorrelation with maximum lag 10 (see step II). We therefore compute wavelet transforms up to level 4.

- Step II: The application of the MWICSS and MWSIC algorithms to test for variance changes requires uncorrelated data. We therefore choose the DWPT packet with highest P-value among those packets of the tree for which the null hypothesis of the Ljung-Box test for autocorrelation is not rejected. The statistic for this test is defined as

  $$Q = m(m+2)\sum_{k=1}^{l} \frac{\hat{\rho}^2(k)}{m-k},$$

  where $\hat{\rho}^2(k)$ is a squared correlation coefficient at lag k and l is arbitrary chosen (see Ljung and Box, 1978). Here we use a lag of 10, since we use at most 150 data points at a time.

- Step III: We test for variance changes (with either the ICSS or the SIC algorithm) using the coefficients of the DWPT packet selected from Step II. If the null hypothesis that no variance change occurs is rejected then we identify the location of the change point using now the non-decimated wavelet packet coefficients of the packet selected in Step II.

- Step IV: Using the binary segmentation procedure we repeat Steps I-III with subsequent subseries until no further variance change point is found. In the case of the ICSS procedure we also perform the additional confirmatory step on all identified potential change points by using subseries of data between adjacent points, as suggested by Inclán and Tiao (1994).

- Step V: We record information of the type (t_j, f_j) where t_j is a time location and f_j is its frequency of detection, i.e. how many times a change at that point has been detected by the method up to the window under consideration. We declare a certain time point to be a variance change if its frequency of detection is greater than or equal to a predetermined threshold k. A smaller k implies faster detection but also a larger number of false alarms.

3.5 Procedure for jump detection

For jump detection we adapt the procedure suggested by Wang to wavelet packets, specifically to MODWPT coefficients. This allows us to locate the jump points more precisely since the MODWPT is not subsampled.

In a generic window of size m we test for jumps in the data as follows.

- Step I: We apply the MODWPT up to level J.

- Step II: We compute a threshold value λ using the finest wavelet coefficients of the MODWPT (the wavelet coefficients of packet $[1, 1]$) according to the formula given in Section 3.2.2 with slight modification (see Vidakovic, 1999).

- Step III: We check wavelet coefficients and find those that exceed the threshold value. In general terms, resolution level j identifies the dyadic interval with width proportional to 2^{j-1}. Wang pointed out that jumps are better detected using relatively narrow widths. In our simulation study we found best detection performances when using the wavelet coefficients at levels 5 and 4. Among all packets at a given level, better performances were obtained at lower frequencies. Results we report here were obtained by considering the locations of the wavelet coefficient of packet [5,1] of the MODWPT for which the absolute value is larger than the threshold value λ. In case we have multiple points as jump points within a given window we choose the closest point to end point of the window. We declare a new jump point if the detected point is at least 20 points away from the jump detected in the previous window.

3.6 Simulation study

3.6.1 Purpose of the study

We performed a simulation study to better understand the relative performances of the iterated cumulative sum of squares (ICSS) and the Schwarz Information Criterion (SIC) algorithms. We simulated data and computed mean delays under several different settings. The aim of the study was to assess how two different factors, the window size and the variance ratio, affect the performance of the MWICSS and MWSIC algorithms. We also looked at the robustness of the distributional and model assumptions on the data.

3.6.2 Simulation scheme

We simulated normal random sequences of length 250 with one change point in the variance located at point 201. For convenience we set the mean of the data to zero. We used four different variance ratios, one vs. four, four vs. one, one vs. sixteen, and sixteen vs. one. For each variance ratio we replicated the experiment 200 times. We adopted the same detection scheme that we used in the previous section. We looked at three different window sizes, 128, 140, and 150. For window size 128 we used windows sliding from point 74 to 114, from point 62 to 102 for window size 140, and from point 52 to 92 for widow size 150. We set the threshold level to 2, that is we recorded end points of windows where change points were detected for the second time. We measured detection delays as differences between the actual change point (the 201 data point) and the end points. We repeated this scheme for the different variance ratios under investigation. We looked at the mean delays and their standard errors from the 200 experiments as criteria for performance comparison.

3.6.3 Results for simulations

Results on normal data are reported in Table 1. We repeated the entire simulation with data from a Laplace distribution, see Table 2, and from an AR(1) process with normal errors, see Table 3. **Variance ratio**: For increasing variance ratios (1 vs. 4 and 1 vs. 16 variance ratio), both MWICSS and MWSIC can capture change points with mean delay around 17 and 7 points, respectively, away from the end point of the analyzing window. Performances in the case of a one vs. sixteen ratio appear to be better than those for the case of one vs. four ratio. This is an obvious result since a bigger variance change should be easier to detect. In these cases the absolute mean delays are in general quite small. However, when the variance changes from large to small, for example from four to one or from sixteen to one, both algorithms show worse performances, with mean delays almost doubled. A variance change from large to small may take more time to be detected because of the bigger oscillations of the signal in the first part that tend to dominate over the latter part.

Table 1: Summary of four variance ratios for MWICSS and MWSIC for normal distribution

variance ratio	method		window size		
			128	140	150
1 vs. 4	MWICSS	mean	16.21	16.72	18.45
		std. err.	0.64	0.62	0.63
	MWSIC	mean	17.86	17.62	18.88
		std. err.	0.71	0.67	0.65
4 vs. 1	MWICSS	mean	32.83	31.76	32.92
		std. err.	0.45	0.39	0.27
	MWSIC	mean	31.48	31.26	31.24
		std. err.	0.46	0.48	0.47
1 vs. 16	MWICSS	mean	6.02	6.24	7.65
		std. err.	0.28	0.29	0.28
	MWSIC	mean	6.06	5.92	7.61
		std. err.	0.26	0.24	0.24
16 vs. 1	MWICSS	mean	35.05	34.88	34.97
		std. err.	0.32	0.34	0.27
	MWSIC	mean	22.24	21.96	23.71
		std. err.	0.43	0.42	0.41

Table 2: Summary of four variance ratios for MWICSS and MWSIC for Laplace distribution

variance ratio	method		window size		
			128	140	150
1 vs. 4	MWICSS	mean	16.57	17.06	19.71
		std. err.	0.64	0.63	0.67
	MWSIC	mean	19.38	18.60	19.76
		std. err.	0.74	0.67	0.65
4 vs. 1	MWICSS	mean	31.29	29.57	31.78
		std. err.	0.68	0.86	0.72
	MWSIC	mean	28.46	27.12	29.41
		std. err.	0.63	0.64	0.58
1 vs. 16	MWICSS	mean	6.69	6.85	8.46
		std. err.	0.32	0.30	0.32
	MWSIC	mean	7.23	7.05	8.85
		std. err.	0.38	0.34	0.36
16 vs. 1	MWICSS	mean	35.6	35.69	35.58
		std. err.	0.32	0.33	0.33
	MWSIC	mean	21.78	22.10	23.64
		std. err.	0.48	0.49	0.46

Table 3: Summary of four variance ratios for MWICSS and MWSIC for AR(1) with normal errors ($\phi = -0.1$)

variance ratio	method		window size		
			128	140	150
1 vs. 4	MWICSS	mean	17.00	18.02	19.63
		std. err.	0.62	0.66	0.64
	MWSIC	mean	19.16	19.47	21.33
		std. err.	0.71	0.68	0.66
4 vs. 1	MWICSS	mean	36.29	30.00	31.25
		std. err.	0.26	0.54	0.38
	MWSIC	mean	30.62	30.86	32.31
		std. err.	0.51	0.59	0.49
1 vs. 16	MWICSS	mean	5.90	6.11	7.47
		std. err.	0.23	0.25	0.23
	MWSIC	mean	6.26	6.10	7.89
		std. err.	0.28	0.28	0.27
16 vs. 1	MWICSS	mean	34.78	34.57	35.28
		std. err.	0.38	0.39	0.32
	MWSIC	mean	22.82	23.23	24.68
		std. err.	0.46	0.46	0.46

Window size: From all three tables we conclude that different window sizes do not affect the detection performance since the variations in detection delays are quite small. Given the reduction in computation time and in cost we suggest to use small window sizes.

MWICSS vs. MWSIC: Both methods show reasonably good performance in the increasing variance ratio cases for both normal and Laplace distributions. In the decreasing variance ratio cases, i.e. four vs. one and sixteen vs. one, we notice that the MWSIC performs better than the MWICSS for the case of a large difference between the two variance values (16 vs 1). The MWSIC algorithm showed large differences in detection performance according to whether we used the additional checking procedure or not. Results here reported were obtained without this procedure. Similar comments apply to results obtained by generating data from an AR(1) process with normal errors. Here, in addition, we notice an improvement in the standard errors for both methods for the cases four vs. one and one vs. sixteen.

Mean delay: An another goal of the simulation study was to investigate how much we can reduce the detection delay. In the case of increasing variance ratios the best detections were 6-8 data points away from the end of the window. That is, we have to endure a 6-8 delay.

3.7 Analysis of network data

3.7.1 Network trace data

Kim et al. (2004) suggest a new data structure for network anomaly detection. Their data structure is based on the concept of correlation between adjacent sampling periods. They use IP addresses and their packet counts from the packet header data. Their computation procedure intends to convert discrete type information into a continuous signal. Within a given sampling period (e.g. one minute) IP addresses and their packet counts are stored for all traffic flows. An IP address has four fields with word-size of 256 locations, that is, a total of 1024 words. For a given traffic flow its packet count is recorded at the number of each field of IP address. In order to obtain a signal, correlation numbers are computed for the four fields at a given sampling point as follows:

$$C_i(t) = \frac{\sum_{j=0}^{255}[\text{packet count}_j(t-1) \times \text{packet count}_j(t)]}{\sqrt{\sum_{j=0}^{255}(\text{packet count}_j(t))^2}} \quad \text{where } i = 1, \cdots, 4.$$

The correlation signal is defined as:

$$S(t) = \alpha_0 + \alpha_1(\sum_{i=1}^{4} w_i C_i(t)), \quad \text{where } \sum_{i=1}^{4} w_i = 1.$$

This linear transformation ensures that the signal lies in the range between zero and one hundred. We illustrate this procedure in Figure 2 with a simple example.

Kim et al. (2004) analyze internet traffic traces from NLANR (National Laboratory for Applied Network Research). They apply the following sampling scheme:

They sampled one minute of traffic to compute their correlation signal and then paused for one minute. The resulting correlation signal consists of 4,302 data points for a 3-day trace. These data were considered as an ambient trace, that is, without noticeable attacks against the network. They then simulated nine kinds of attacks with various behaviors, as motivated by recent SQL Slammer and Code Red attacks. The nine attacks were classified as follows:

(1) **Duration:** The first 6 attacks last for 2 hours, the remaining 3 attacks for 1 hour.

(2) **Persistance:** The first 3 attacks send malicious packets for 3 minutes and pause for 3 minutes. Such pattern is repeated through the attack duration. While the filtering may mitigate the overhead of the attacker's continuing scan traffic, a more sophisticated attacker might have stopped scanning and it may be possible to conceal attacker's intentions through repeating attack and pause periods. The other remaining attacks continue to assault throughout the attack period.

(3) **IP address:** The first attack among every 3 attacks targets a single destination IP address. In a hypothetical situation, the attackers target a famous site such as the White House, CNN or Yahoo, etc. This target may be really one host in case of 32-bit prefix, occasionally aggregated neighboring hosts in case of x-bit prefix. The 2nd attack style imitates from the IP address generation scheme of the notorious Code Red II worm. That is to say, a portion of addresses preserve the class-A and a partition of addresses preserve class-B for the infiltration efficiency. The 3rd type is a randomly generated address that was used for the Code Red I and SQL Slammer worm.

(4) **Protocol:** The 3 major protocols, ICMP, TCP, and UDP, are used in turn.

(5) **Port:** The second port among every 3 attacks targets randomly generated destination ports. It is useful to detect portscan that is used to probe a loosely

Table 4: Description of nine simulated attacks

	1	2	3	4	5	6	7	8	9
Duration	2h	2h	2h	2h	2h	2h	1h	1h	1h
Persistency	intermittence	intermittence	intermittence	persistency	persistency	persistency	persistency	persistency	persistency
IP	single	semi-random	random random	single single	semi-random	random random	single single	semi-random	random random
Protocol	ICMP	TCP	UDP	ICMP	TCP	UDP	ICMP	TCP	UDP
Port	#80	random	#1434	#80	random	#1434	#80	random	#1434
Size	random	4KB	404B	random	4KB	404B	random	4KB	404B

defensive port. The first port is a representative #80 that stands for the reserved port for well-known services. The third port is a #1434 that acts for the ephemeral client port, which is used in SQL Slammer worm.

(6) **Size:** There are three different byte counts of packets. The three denominations are random size, 4K bytes and 404 bytes.

The attacks can be described by a 3-tuple (duration, persistency, and IP address). These attacks were superimposed to the ambient traces from NLANR. The ratio of attack and normal traffic is 1:2 in packet counts. The resulting correlation signal is shown in Figure 4. We summarize the features of the nine attacks in Table 4. The first three attacks exhibit variance changes, while the other 6 show also sudden up and down jumps.

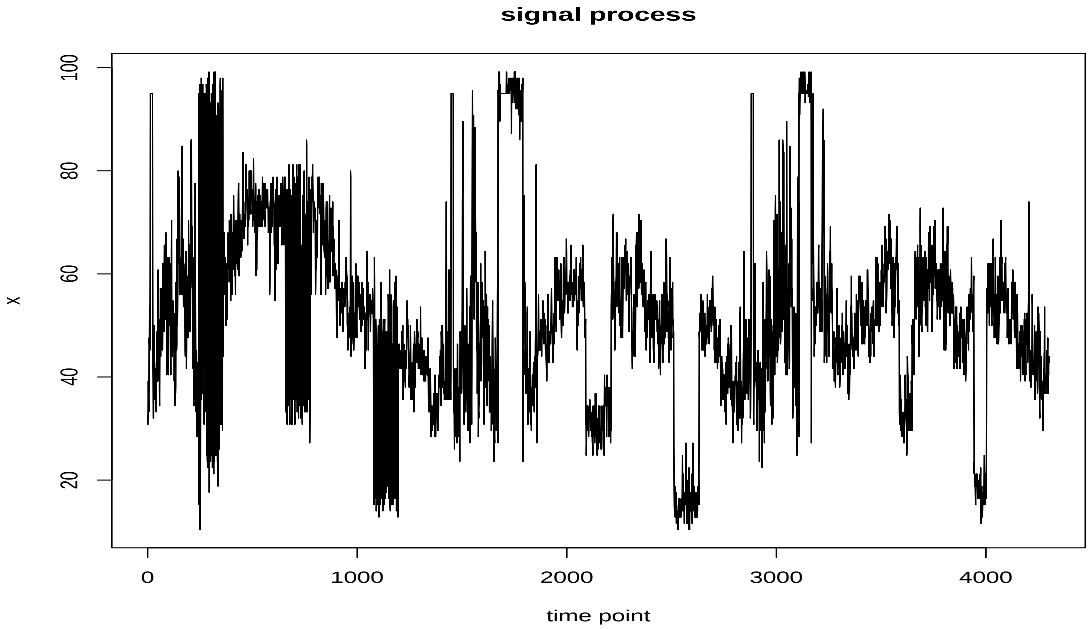


Figure 4: Correlation signal

Figure 5 shows the sub-sequence of the data corresponding to the second attack. In the same figure we also report autocorrelation functions of the data, of the DWT wavelet coefficients at levels 2 and 3 and of two DWPT packets. This figure clearly shows the additional flexibility of the DWPT versus the DWT at decorrelating data.

3.7.2 Results

We examined several different combinations of the window size and the wavelet family. We used three different wavelet families, the Haar wavelets, Daubechies wavelets with 2 vanishing moments, and the least asymmetric wavelets with 4 vanishing moments (Daubechies, 1992). In order to reduce the number of false alarms we used the threshold approach as previously described, that is, we considered change points those for which the detection numbers are equal or greater than the threshold value.

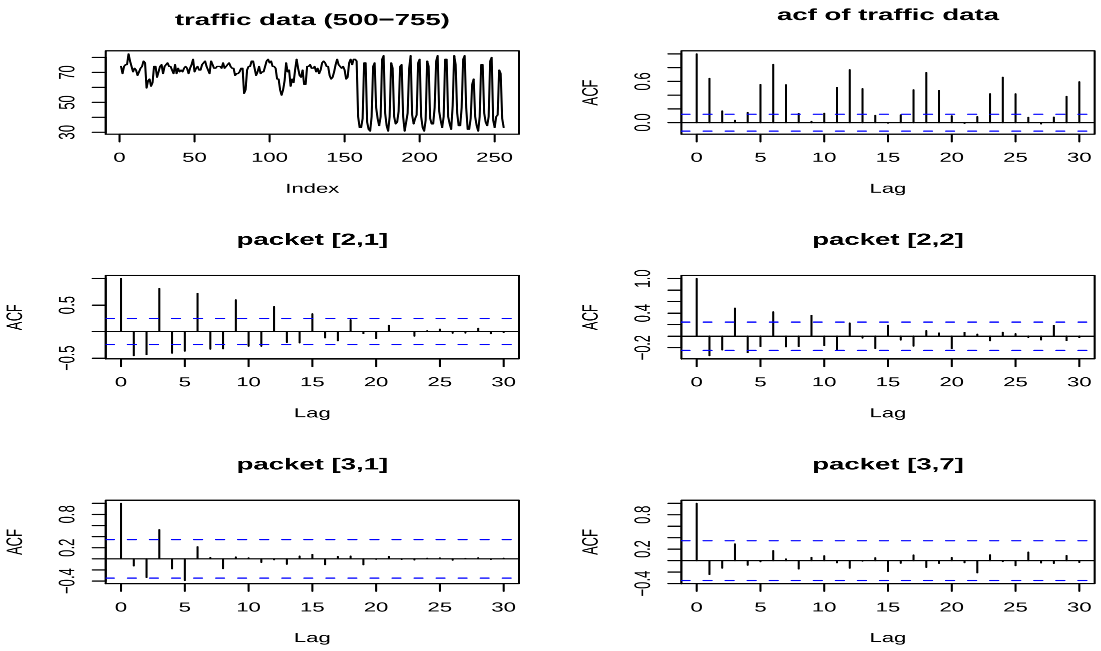


Figure 5: Attack n.2 with autocorrelation functions of the data, of the DWT wavelet coefficients at levels 2 and 3 and of two DWPT packets.

When computing detection delays we considered a change point successfully detected if a point that falls within 10 time points from the actual change point was detected by our procedure.

We report here results we obtained with Haar wavelets, which showed the best performances. We considered only 8 attacks, that is, 16 change points, among the 9 simulated. We ignored the last attack because of the moving window and threshold approach we adopted. Table 5 reports detection delays for 4 threshold values between 3 and 15. We measured the detection delay as the time difference between the actual change point and the earliest point detected by our procedure. Numbers in the first column of Table 5 indicate the 16 change points (numbered from 1 to 16) that define starting and ending of the first to the eighth attack.

Table 5: Detection delays for MWICSS and MWSIC

threshold		3						6					
window		128		140		150		128		140		150	
method		ICSS	SIC	ICSS	SIC	ICSS	SIC	ICSS	SIC	ICSS	SIC	ICSS	SIC
	1	12	12	12	16	16	18	16	16	16	-	23	-
	2	65	53	77	113	79	87	71	77	85	101	87	101
	3	8	8	8	8	10	10	14	14	14	14	14	14
	4	58	62	58	74	68	68	70	74	74	104	84	114
	5	4	5	5	7	4	9	7	8	8	19	8	19
change	6	65	70	70	64	78	74	79	80	77	74	82	92
	7	12	11	16	11	14	13	20	14	24	14	26	16
	8	29	35	29	51	23	27	35	47	35	71	33	61
	9	45	105	49	33	52	75	55	-	57	77	63	127
point	10	1	-	1	-	2	-	5	-	12	-	6	-
	11	38	16	20	28	18	102	52	88	50	104	52	111
	12	8	51	4	81	2	53	13	73	9	93	14	87
	13	-	9	-	7	-	9	-	41	-	45	-	47
	14	20	14	24	22	20	26	30	30	30	28	30	34
	15	64	-	110	-	116	-	-	-	-	-	-	-
	16	42	-	77	132	-	-	-	-	-	-	-	-
mean delay		31.40	34.69	37.33	46.22	35.86	43.92	35.93	65.64	37.77	74.45	40.31	90.17
total points detected		107	141	106	137	104	136	55	63	58	66	60	64

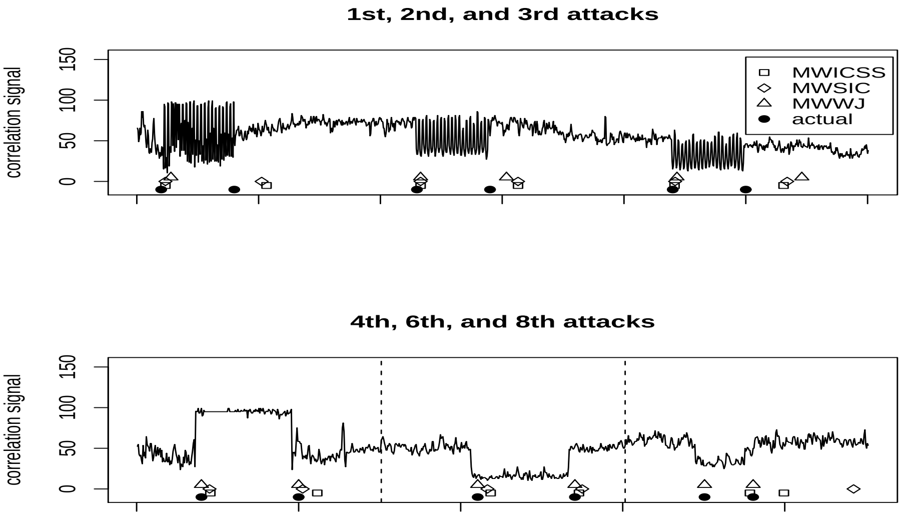
threshold		9						15					
window		128		140		150		128		140		150	
method		ICSS	SIC	ICSS	SIC	ICSS	SIC	ICSS	SIC	ICSS	SIC	ICSS	SIC
	1	27	31	27	-	33	-	-	-	-	-	-	-
	2	77	97	97	122	93	111	116	118	132	-	138	-
	3	18	18	18	18	20	20	31	-	-	-	-	-
	4	80	118	82	118	92	126	94	-	110	-	118	-
	5	11	22	12	27	12	25	19	-	18	-	18	-
change	6	108	94	88	92	94	100	121	-	126	-	131	-
	7	26	17	30	17	33	23	-	32	-	28	-	28
	8	64	51	60	77	58	75	89	-	103	-	109	-
	9	61	-	63	-	69	136	-	-	-	-	-	-
point	10	22	-	42	-	20	-	-	-	-	-	-	-
	11	62	101	63	120	19	127	-	-	-	-	-	-
	12	20	85	16	101	47	99	-	-	-	-	-	-
	13	-	52	-	64	-	99	-	-	-	-	-	-
	14	44	78	36	37	-	38	52	108	105	-	107	126
	15	-	-	-	-	-	-	-	-	-	-	-	-
	16	-	-	-	-	-	-	-	-	-	-	-	-
mean delay		47.69	60.30	48.77	62.60	49.17	78.08	71.15	86.00	99.00	28.00	108.17	77.00
total points detected		54	61	57	66	58	64	32	39	28	12	29	13

In Table 6 we report detection delays for the MWWJ algorithm with four different widow sizes. The detection criterion for MWWJ is as follows. We set to 20 the gap size value to decide whether a jump occurs. For a given window size we find all locations at which the absolute value of the MODWPT coefficients exceeds λ (computed using the MODWPT coefficients of the finest level). Then we record the closest location to the end point of the window. We compare this location with the one of the previous window. When the difference between two points is equal to or greater than the predetermined gap size, this new point is declared as a jump point. As expected, performances of the three different detection methods vary according to the attack type. MWWJ detects all 12 jump-type change points without delay while it shows worse performances in capturing variance change points (first three

Table 6: Detection delays for MWWJ

window		100	128	140	150
change point	1	-	16	16	16
	2	-	-	-	-
	3	6	6	6	6
	4	24	27	30	33
	5	7	7	7	7
	6	63	92	97	100
	7	0	0	0	0
	8	0	0	0	0
	9	0	0	0	0
	10	1	1	1	1
	11	0	0	0	0
	12	0	0	0	0
	13	0	0	0	0
	14	0	0	0	0
	15	0	0	0	0
	16	0	0	0	0
mean delay		7.14	10.53	11	11.36
total points detected		27	28	28	23

attacks, see Table 6), particularly for the first attack. Note that the 2nd and 3rd attacks are not “pure” variance change points. Indeed, they contain both a jump in the mean level as well as a variance change. As for the MWICSS and the MWSIC, performances are different for the single attacks. For the 1st, 2nd, and 3rd attacks the two methods show comparable behaviour, with a slight better performance of the MWICSS. For the 4th and 7th attacks the MWSIC does a better job at capturing the starting point while the MWICSS performs better in detecting the end point of the attack. MWSIC shows bad behaviour for the 5th attack, by missing it in most cases, and performs worse than MWICSS in the detection of the 6th attack. The 8th attack is a very difficult case to detect, although MWICSS with a small threshold does a decent job, even if with a considerable detection delay. As a general result, MWICSS may be preferable to MWSIC since it shows smaller mean detection delays. Here MWSIC was performed without the confirmatory step as additional checking procedure previously described because we noticed that including such additional checking would worsen the performances of the MWSIC method. On the contrary, when used with the MWICSS algorithm the confirmatory step was beneficial. Plots of Figure 6 give a graphical representation of the performances of the three detection methods. There, each of the two subplots contains a different portion of the signal, displaying 1st, 2nd, 3rd attacks and 4th, 6th and 8th attacks respectively, as representatives of the two different kind of change point, in mean and in variance. Results for MWICSS and MWSIC are for a threshold level 2 and window size 128 (see Table 5), those for MWWJ are for window size 128 (see Table 6). In these plots, the solid circles indicate the real change points, the square rectangles the points detected by the MWICSS, the diamonds those detected by the MWSIC, and the triangles those detected by the MWWJ. Notice how the MWICSS and MWSIC algorithms do a better job at detecting attacks of the first type, that show variance changes. However, there appears to be an asymmetric aspect in the detection of these two methods, in that both the MWICSS and the MWSIC detect the start of the attacks but show a relative large


1st, 2nd, and 3rd attacks

4th, 6th, and 8th attacks

Figure 6: Performances of the three algorithms

delay in detecting the ending points. In other words, these algorithms seem to be sensitive to the location of the change points and to the variance ratio, as already suggested by the simulation study of the previous section.

For online network attack detection, our results suggest that a simultaneous use of both MWICSS (or MWSIC) and the MWWJ algorithms give best results, allowing the detection of attacks of different types. Indeed, the average detection delay for all methods is 10.63 minutes. In addition, if we consider the starting points of the attacks only, as points of primary interest in network attack detection, the mean detection delay is 1.06 minutes, with a threshold level 2.

CHAPTER IV

APPLICATION FOR BIOINFORMATICS

4.1 Introduction

We propose a Bayesian model for the analysis of high-throughput data where the outcome of interest has a natural ordering. The method provides a unified approach for identifying relevant markers and predicting class memberships. This is accomplished by building a stochastic search variable selection method into an ordinal model. We apply the methodology to the analysis of proteomic studies in prostate cancer. We also explore wavelet-based techniques to remove noise from the protein mass spectra. The goal is to identify protein biomarkers associated with prostate-specific antigen (PSA) level, this ordinal diagnostic measure currently used to stratify patients into different risk groups.

Recently, there has also been interest in using protein mass spectroscopy to detect discriminating molecular markers in Petricoin et al. (2002a). The diagnostic categories often consist of tumor versus normal tissues, different types of malignancies, and subtypes of a specific cancer. Several variable selection methods have been developed to address this problem in Brown et al. (1998a, 1998b, 2002) and Sha et al. (2003). These procedures are tailored towards classification into nominal categories. In some cases, however, the outcome of interest may have an ordered scale. Examples of variables with a natural ordering include the stage of a tumor and quantitative clinical factors such as white blood cell counts. Applying methods designed for nominal variables to such problems is not optimal since the information about the ordering will be ignored.

In this chapter, we propose a Bayesian variable selection method for classification into ordinal categories. In this method, the ordered outcomes are related to the PSA levels using a data augmentation approach. The variable selection procedure proposed by George and McCulloch (1997) is built into the model through a latent binary inclusion/exclusion vector. Markov chain Monte Carlo (MCMC) stochastic search techniques are used to update this latent vector and explore the prohibitively large space of variable subsets for promising models. For posterior inference, Bayesian model averaging techniques proposed by Madigan and Raftery (1994)are used to identify discriminating variables and predict the ordered group membership of a sample. This allows us to account for the uncertainty inherent in the model selection process. In addition, in the proteomic application, we propose wavelet based techniques to remove noise from the mass spectra as part of the preprocessing steps required before analysis.

We apply the methods to the analysis of prostate cancer studies conducted using protein mass spectroscopy technologies. Prostate cancer is the most frequently diagnosed and the second leading cause of cancer death in men in the United States (see Jemal et al., 2003). Despite these high rates, it is often an indolent disease and patients can remain asymptomatic for years. Currently, patients prognostic and treatment assignment are based on clinical stage, serum PSA levels. PSA is a glycoprotein produced primarily by the epithelial cells that line the acini and ducts of the prostate gland, and is concentrated in prostatic tissue. Serum PSA levels are normally low and tend to increase proportionally to the pathological stage of the tumor (see Stamey and Kabalin, 1989). Protein mass spectroscopy experiments have also been used to detect markers that distinguish men with different PSA levels (Petricoin et al., 2002b). We will focus on the analysis of the protein spectra from Petricoin et al.. Our goal is to identify relevant markers, both at the transcriptional and post-translational levels, related to the state of tumor as measured by the PSA levels.

4.2 Bayesian ordinal probit model

4.2.1 Probit model for ordinal outcomes

Let (Z, X) denote the observed data, where $Z_{n\times 1}$ is the vector of ordered categorical outcomes and $X_{n\times p}$ is the matrix of covariates. The responses Z_i take one of J values, $0, \dots, J-1$. Each outcome Z_i is associated with a vector $(p_{i,0}, \dots, p_{i,J-1})$, where $p_{i,j} = P(Z_i = j)$ is the probability that subject i falls in the class ordered j. The probabilities $p_{i,j}$ can be related to the linear predictor $x_i\beta$ by adopting a data augmentation approach in Albert and Chib (1993). We assume that there exists a latent continuous random variable Y_i, such that

$$Y_i = \alpha + x_i'\beta + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2), \quad i = 1, \dots, n, \tag{4.2.1}$$

where β is a $p \times 1$ vector of regression coeffifcients and σ^2 is set to 1 to make the model identifiable. The correspondence between the observed outcome Z_i and the latent variable Y_i is defined by

$$Z_i = j \ \text{ if } \ \delta_j < Y_i \le \delta_{j+1}, \qquad j = 0, \dots, J-1, \tag{4.2.2}$$

where the boundaries δ_j are unknown and $-\infty = \delta_0 < \delta_1 < \ldots < \delta_{J-1} < \delta_J = \infty$.

4.2.2 Incorporating variable selection into model

Without loss of generality, we assume in the sequel that X has been centered, so that its columns sum to zero. Thus, $\text{rank}(X) \le \min(n-1, p)$.

We deal with high dimensional data sets where the number p of covariates is substantially larger than the sample size n and most of the variables provide no information about the outcome of interest. The method we propose here for variable selection can be viewed as a univariate version of the approach presented by Sha et al. for multinomial probit models. In this context, however, the correspondence between Z_i and Y_i uses different boundaries that account for the natural ordering of the outcome. In order to identify the predictive variables, we introduce a latent binary inclusion/exclusion vector γ that induces a mixture prior on the regression coefficients. We specify conjugate priors for the intercept $\alpha \sim \mathcal{N}(\alpha_0, h)$ and the regression coefficients of the included variables $\beta_\gamma \sim \mathcal{N}(\beta_{0_\gamma}, H_\gamma)$. The simplest form for the prior of γ is to assume its elements to be independently and identically distributed Bernoulli random variables, $\pi(\gamma) = w^{p_\gamma}(1-w)^{p-p_\gamma}$, where w is the proportion of variables expected a priori to be related to the outcome and p_γ is the number of included variables. This prior can be relaxed and more uncertainty can be introduced by assuming a further beta prior on w.

4.2.3 Hyperparameter settings

A vague prior can be specified on the intercept parameter α by setting h large, so that the value ascribed to the prior mean becomes irrelevant. We set $\alpha_0 = 0$ and $\beta_0 = 0$. For a given γ, the prior on β depends on the matrix H_γ. Brown et al.(1998b) discuss relative merits and drawbacks of different specifications. Here we use $H = cI$, which is easier to calibrate. The parameter c regulates the amount of shrinkage in the model. In general, we want to avoid very small values of c which cause too much regularization and large values that can induce nonlinear shrinkage as a result of Lindley's paradox (see Lindley, 1957). For the boundary parameters, we need to impose one constraint to ensure identifiability; without loss of generality we take $\delta_1 = 0$. We assign diffuse priors for the remaining parameters.

4.2.4 Model fitting

The prior beliefs are then updated with information from the data. We do this using Markov chain Monte Carlo (MCMC) techniques. The model fitting can be made more efficient by integrating out the parameters α and β. The MCMC sampler starts from a set of arbitrary parameters and the following steps are iterated:

- Step I:

  Update the latent vector Y from its posterior distribution given (γ, δ, X, Z), which is a truncated normal density under the constraints defined in equation (4.2.2)

  $$Y|(\gamma, \delta, X, Z) \sim \mathcal{N}_\delta(1\alpha_0 + X_\gamma\beta_{0\gamma}, P_\gamma), \tag{4.2.3}$$

  where $P_\gamma = I_n + h11' + X_\gamma H_\gamma X'_\gamma$, 1 is an $n \times 1$ vector of ones, I_n is an $n \times n$ identity matrix.

- Step II:

  Update the latent variable selection vector γ from its conditional posterior distribution

  $$\pi(\gamma|Y, \delta, X, Z) \propto \pi(\gamma) \cdot \pi(Y|\gamma, \delta, X, Z). \tag{4.2.4}$$

  This is accomplished using a Metropolis algorithm as in Sha et al.. In this approach, the sampler visits a sequence of models that differ successively in one or two variables. At a generic step, a candidate model, γ^{new}, is generated by randomly choosing among a set of transition moves. These moves consist of adding or deleting a variable by choosing one of the γ_k's ($k = 1, \ldots, p$) and changing its value, or swapping the status of two variables by choosing independently and at random a 0 and a 1 and exchanging their values. The proposed γ^{new} is accepted with a probability that depends on the ratio of the relative posterior probabilities of the new vector versus the one visited at the previous iteration.

- Step III:

  Update the boundary parameters from their posterior densities given $(\gamma, X, Z, \delta_{(-j)})$, where $\delta_{(-j)}$ is the vector δ without the j-th element. These conditional distributions are uniform $[\max\{Y_i : Z_i = j-1\} \wedge \delta_{j-1}, \min\{Z_i : Y_i = j\} \wedge \delta_{j+1}]$, as described in Albert and Chib (1993).

4.2.5 Posterior inference

The MCMC procedure results in a list of visited variable subsets, γ, as well as sampled δ and Y vectors with their corresponding relative posterior probabilities. In order to draw posterior inference, we first need to impute the latent vector Y, which can be viewed as missing data. Let $\widehat{Y}$ and $\widehat{\delta}$ be the estimates obtained by averaging respectively over the sampled Y and δ vectors. The normalized conditional probabilities $\pi(\gamma|\widehat{Y}, \widehat{\delta}, X, Z)$, which identify promising variable subsets, can be computed for all distinct vectors γ visited by the MCMC sampler. The marginal posterior probabilities of inclusion for single variables, $\pi(\gamma_k = 1|\widehat{Y}, \widehat{\delta}, X, Z),\ k = 1, \ldots, p$, can also be derived from these posterior probabilities.

Inference on class prediction can be done in various ways. If a further set of observations is available for validation, least squares prediction based on a single "best" model can be computed:

$$\widehat{Y}_f = \tilde{\alpha} + X_{f(\gamma)} \tilde{\beta}_{(\gamma)}, \qquad (4.2.5)$$

where γ is the vector with highest posterior probability, X_γ consists of the covariates selected by γ, $\tilde{\alpha} = \bar{\widehat{Y}}$, $\tilde{\beta}_(\gamma) = (X'_\gamma X_\gamma + H_\gamma^{-1})^{-1} X'_\gamma \widehat{Y}$. Alternatively, we can use

Bayesian model averaging over a set of a posteriori likely models to estimate Y_f:

$$\widehat{Y}_f = \sum_{\gamma} \left(\tilde{\alpha} + X_{f(\gamma)} \tilde{\beta}_{(\gamma)} \right) \pi(\gamma | \widehat{Y}, \widehat{\delta}, X, Z). \tag{4.2.6}$$

The ordered categorical outcomes can then be predicted using the correspondence

$$\widehat{Z}_{f,i} = j \quad \text{if } \widehat{\delta}_j < \widehat{Y}_{f,i} \leq \widehat{\delta}_{j+1}. \tag{4.2.7}$$

In situations where the sample size is limited dividing the data into a training and a validation set may not be possible. In such cases, one can resort to sampling-based methods for cross-validation prediction (see Gelfand,1996). A cross-validation predictive distribution for sample i can be calculated using $\pi(\gamma, Y, \delta | X, Z)$ as importance sampling density for $\pi(\gamma, Y, \delta | X, Z_{(i)})$, where $Z_{(i)}$ is the outcome vector Z without the i-th element:

$$\begin{aligned} P(Z_i = j | X, Z) &= \int_{\gamma} \int_{Y} \int_{\delta} P(Z_i = j | X, Z_{(i)}, \gamma, Y, \delta) \cdot \pi(\gamma, Y, \delta | X, Z_{(i)}) \; d\delta \; dY \; d\gamma \\ &\propto \frac{1}{M} \sum_{t=1}^{M} P(\delta_j^{(t)} < Y_i \leq \delta_{j+1}^{(t)} | X, Z_{(i)}, \gamma^{(t)}, Y^{(t)}) \qquad (4.2.8) \\ &= \frac{1}{M} \sum_{t=1}^{M} \Phi \left(\delta_{j+1}^{(t)} - \mu_{Y_i}^{(t)} \right) - \Phi \left(\delta_j^{(t)} - \mu_{Y_i}^{(t)} \right), \end{aligned}$$

where t indexes the MCMC iterations, $\Phi(.)$ is the normal cumulative density function, $\mu_{Y_i}^{(t)} = \tilde{\alpha}^{(t)} + x_{i,\gamma^{(t)}} \tilde{\beta}_{\gamma}^{(t)}$ with $\tilde{\alpha} = \bar{Y}^{(t)}$, $\tilde{\beta}_{\gamma}^{(t)} = ({X'}_{\gamma^{(t)}} X_{\gamma^{(t)}} + H_{\gamma}^{-1})^{-1} {X'}_{\gamma^{(t)}} Y^{(t)}$, and $x_{i,\gamma^{(t)}}$ are sample i's measurements for the variables selected by $\gamma^{(t)}$. The class membership of sample i can then be predicted by the mode of the predictive distribution:

$$\widehat{Z}_i =_{0 \leq j \leq J-1} P(Z_i = j | X, Z). \tag{4.2.9}$$

4.3 Preprocessing of mass spectrometry profiles

Protein mass spectra are inherently noisy and require substantial preprocessing before analysis. A mass spectrum can be represented as a curve where the x-axis indicates the ratio of a particular molecule's weight to its electrical charge (m/z) and the y-axis represents a signal intensity corresponding to the abundance of the molecule in the sample. Most peaks in the spectrum are associated with proteins or peptides and constitute important features. The goal of the analysis is often to identify peaks related to specific outcomes, such as different malignancies or clinical responses. Before proceeding to the data analysis, a number of preprocessing steps, such as removal of baseline and noise, normalization and calibration of samples, are needed. The procedures to achieve these are still experimental and no standard has yet been established.

4.3.1 Baseline correction

This step is required to remove the ion overload and chemical noise that are usually higher at smaller m/z values. There is no general solution to this problem because baseline characteristics vary from one experiment to another and each spectrum has to be assessed individually. For the data considered in this paper, the baseline was already subtracted by the original investigators. Indeed, we can see from the spectra plot in Figure 7 that there is no evident baseline artifact.

4.3.2 Noise removal by wavelet methods

Wavelets are families of orthonormal bases that can be used to parsimoniously represent functions. Following the seminal work of Donoho and Johnstone (1994), wavelet thresholding has successfully been used in various applications to remove noise and recover the true signal intensities. This is accomplished by applying a wavelet transform to the data and mapping wavelet coefficients that fall below a threshold to 0 (hard thresholding) or shrinking all coefficients toward 0 (soft thresholding). One can also opt between a universal or an adaptive thresholding rule. The former applies

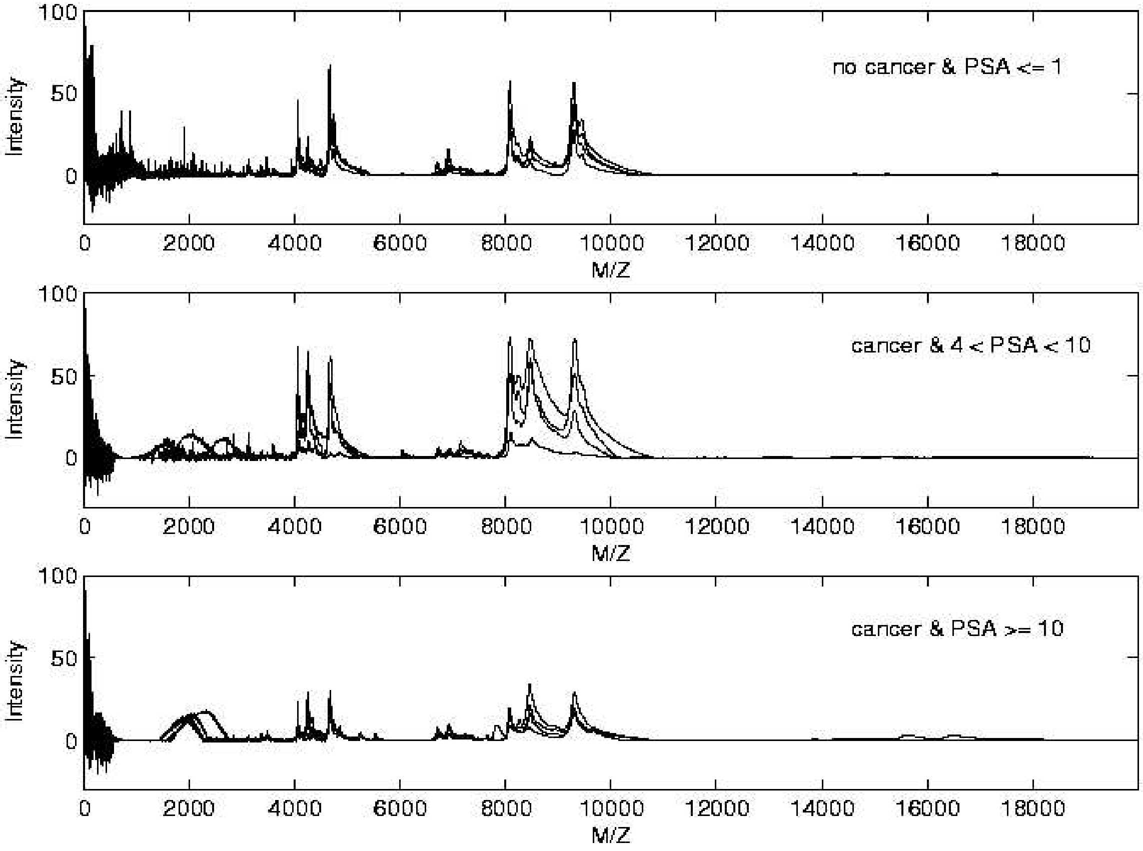


Figure 7: Profiles of four mass spectra from each class

the same threshold, i.e. identical cut-off value or same amount of shrinkage for all wavelet coefficients, whereas the latter uses a threshold that depends on the resolution level of the wavelet coefficients. An inverse wavelet transform is then applied to the thresholded coefficients leading to a smoothed estimate of the function.

We interpolated the mass spectra on a grid of equally spaced m/z values using piecewise cubic splines. We considered three different grids with 10,000, 12,000 and 15,000 equi-spaced points in the range of the data. We noticed better qualitative denoising with undecimated transforms over standard decimated discrete wavelet transforms (DWT). These transforms do not impose restrictions on the length of the signal and are shift-invariant, i.e., they are not affected by the starting position of the signal. We used the maximum overlap discrete wavelet transforms (MODWT) (see Percival and Walden, 2000). We explored different choices of wavelet basis (Haar wavelets, Daubechies wavelets with 4 vanishing moments, least symmetric Daubechies wavelets with 8 vanishing moments), different thresholding rules (hard versus soft and universal versus adaptive). See Daubechies (1992). In general, we noticed that the universal hard threshold removes lots of coefficients and the universal soft threshold tends to attenuate some of the distinctive peaks. The adaptive soft thresholding approach, on the other hand, does a better job at preserving the peaks. As a result of this investigation, we chose to interpolate the data on a grid of 15,000 points and used the MODWT with Daub(4) along with an adaptive soft thresholding rule.

4.3.3 Peak identification

A crucial step for the identification and quantification of proteins in mass spectra is to find m/z values that correspond to peak intensities. We used the peak detection methods implemented in the Bioconductor PROcess package (WWW.BIOCONDUCTOR.ORG). For each spectrum, peaks were identified as m/z values with signal intensities satisfying at least three of the following criteria: (1) the intensity exceeds a specified threshold value; (2) the intensity exceeds a constant times the median absolute deviation estimate of noise in a given window; (3) the intensity is a local maximum within a given window; (4) the ratio of the area under the peak, i.e., the sum of the intensities within a bandwidth, versus the maximum area among all peaks is greater than a pre-specified constant.

4.3.4 Normalization

Systematic variations often exist between spectra and need to be minimized. These are due to varying amounts of protein samples or differences in the detector sensitivity . We used a global normalization approach in which the signal intensities are scaled by a common factor. For a given peak in a spectrum, we defined the normalization constant as the ratio of the area under the peak to the median area of the corresponding peaks in all spectra. See Bolstad et al. (2003).

4.3.5 Alignment

Mass spectra exhibit shifts along the horizontal axis between replicate spectra. In general, the instruments have an accuracy of 0.1 to 0.3% on the m/z scale. Thus, detected peaks that have masses within the percentage accuracy are considered identical. We merged peaks that have m/z measurements within 0.2% of each other and assigned the new peak the average m/z values and the maximum intensity.

4.4 Results

We use the surface-enhanced laser desorption and ionization time-of-flight (SELDI-TOF) protein mass spectra from Petricoin et al. (2002b). The complete data set is available at HOME.CCR.CANCER.GOV/NCIFDAPROTEOMICS/PPATTERNS.ASP and consist of 63 samples with no evidence of disease and PSA ≤ 1 ng/ml, 190 benign samples with PSA levels ranging from less than 1 ng/ml to greater than 10 ng/ml, 26 cancer samples with $4 < \text{PSA} < 10$, and 43 cancer samples with PSA ≥ 10 ng/ml. Petricoin et al.(2002b) were interested in investigating the ability of serum protein profiles to discriminate between different prostate conditions based on their PSA levels. They considered 25 samples with no evidence of disease and PSA levels ≤ 1 ng/ml, and 31 prostate cancer samples with PSA ≥ 4 ng/ml in their training set. They used genetic algorithms and self-organizing maps to identify discriminating protein markers. They then used the selected markers in an independent test set to separate benign and tumor samples whose PSA levels spanned all possible ranges. Their approach does not specifically address the issue of validating the performance of the classifiers since the outcomes considered in the training and test sets are different.

In the sequel, we remove the benign samples from analysis because their PSA levels span all possible ranges. We focus on the other three groups (63 non-diseased with PSA ≤ 1, 26 cancer with $4 <$ PSA < 10, and 43 cancer with PSA ≥ 10). Figure 7 displays the mass spectra for four patients from each of the three groups. Each mass spectra represents the expression profile of 15,154 peptides defined by their m/z values. We note some clear differences between the different classes. We divided the data into a training set (70% of data, i.e. 92 spectra) and a validation set (40 mass spectra) to assess the prediction performance of the classifiers. We preprocessed the spectra as described in the previous section. After applying the wavelet thresholding for noise removal, the peak identification and alignment steps resulted in 53 peaks in the training set. We located these same 53 peaks in the test set in order to assess the selected classifiers.

We fitted the ordinal probit model with variable selection to identify protein markers that discriminate among the three groups. We used a Bernoulli prior with 10 variables expected to distinguish the classes. We ran four MCMC chains with 100,000 iterations each and discarded the first half as burn-in to eliminate dependence on the starting points. We used $c = 0.1$ for the covariance hyperparameter of the regression coefficients. Each chain visited about 20,000 distinct models after the burn-in period. The majority of the visited models contained 5 to 15 variables. The marginal probabilities of inclusion for single peaks are shown in Figure 8. Indices with high posterior probabilities correspond to important markers that discriminate between the different groups. There is a good concordance among the four plots and posterior inference was drawn on the pooled output from the four MCMC chains.

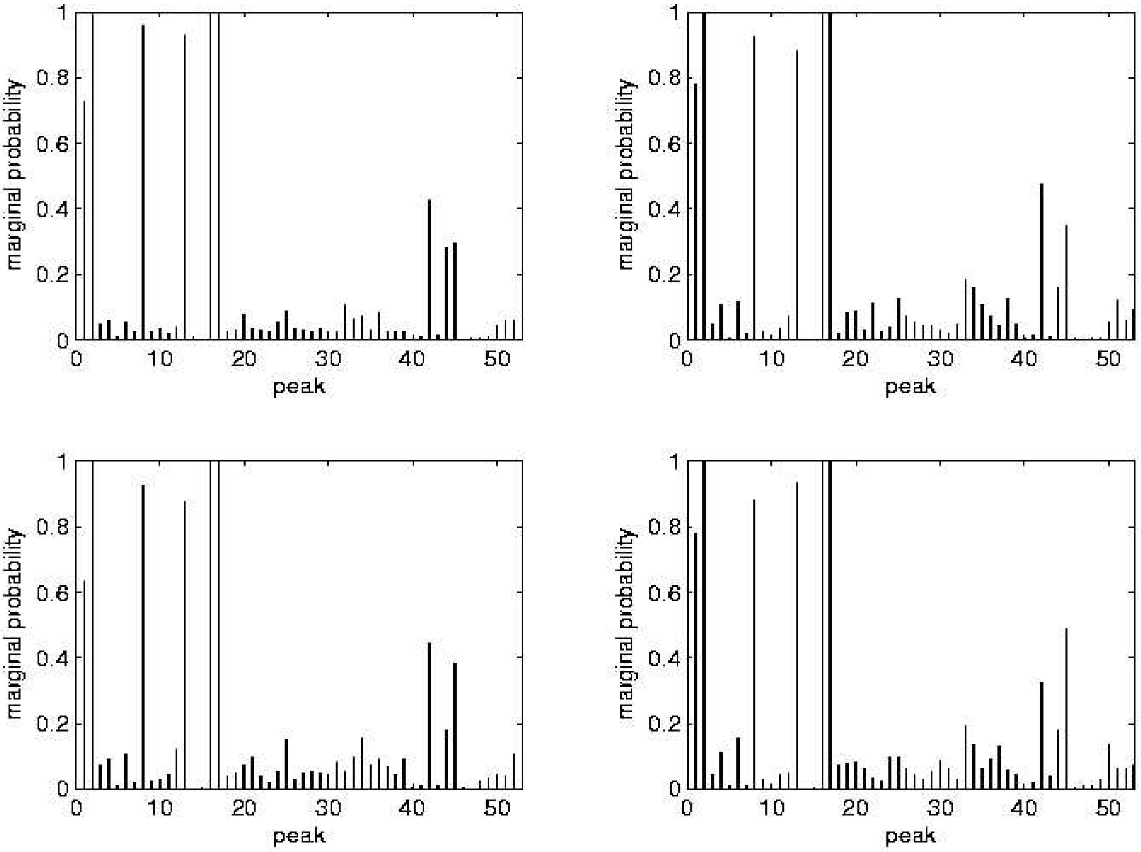


Figure 8: Marginal posterior probabilities of inclusion for single peaks in each of the four MCMC chains

We considered variables with large marginal posterior probabilities as well as markers included in the "best" models, i.e., γ vectors with high joint posterior probabilities. As we can see from the list of selected markers reported in Table 7, there is a good agreement between the selections based on the marginal and joint posterior probabilities. The majority of the selected peaks have m/z values lower than 7000 with only three peaks falling in a higher range. We note that among the seven peaks selected by Petricoin et al. (2002b) six had m/z values lower than 6000. Figure 9 displays surface representations of single spectra for the first 10,000 m/z values in each of the three groups plus the benign group that was not considered for analysis.

Table 7: List of selected markers

Criterion	selected markers (m/z values)			
Best γ	693.19	930.48	2011.6	3466
	4665.7	4739.1	7683.8	8983.5
$\pi(\gamma_k \| \mathbf{X}, \mathbf{Z}) > 0.5$	693.19	930.48	2011.6	3466
	4665.7	4739.1		
10 best γ vectors	693.19	930.48	2011.6	3466
	4665.7	4739.1	6312.1	6554.7
	7683.8	7793.1	7886.4	8327.7
	8603.6	8983.5	19495	

The arrows on top of the graph indicate peaks that appeared in the best model. We note that they clearly distinguish the different groups. The performance of the selected discriminants was assessed by predicting the class membership in the validation set. We considered a least squares prediction as well as a Bayesian model averaging (BMA) prediction based on the single best and on the 10 best models. Table 8 reports the misclassification error rates for each of these prediction approaches. They ranged between 22.5 and 27.5%. For comparison, we also looked at commonly used classification methods, such as k-nearest neighbor (KNN) and nonlinear support vector machines (SVM). For KNN we considered values of k ranging from 1 to 20 and we report the results for $k = 2$. We note that all the methods have fairly high error rates with LDA performing slightly better. However, we have to keep in mind that this approach does not provide selection of the actual discriminating markers. We also note that the largest misclassification errors were associated with the cancer group that has $4 < \text{PSA} < 10$. This is known to be a range where the PSA levels correspond to a rather heterogeneous group.

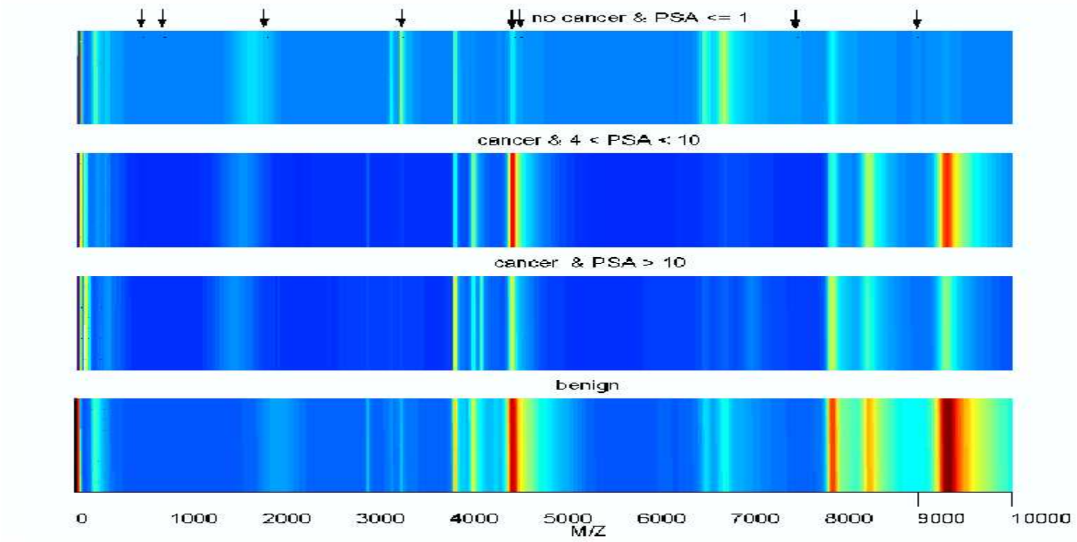


Figure 9: Surface representation of spectra from patients in four groups. Arrows at the top of the graph indicate peaks selected by our method

Table 8: Misclassification error rates

Prediction approach	overall error rate	PSA ≤ 1	$4 <$ PSA < 10	PSA ≥ 10
MCMC (pooled)				
Least squares	0.250	0/19	5/8	5/13
BMA (1 best)	0.225	0/19	4/8	5/13
BMA (10 best)	0.275	1/19	5/8	5/13
nonlinear SVM	0.250	1/19	3/8	5/13
kNN	0.225	1/19	5/8	3/13

CHAPTER V

SUMMARY AND FUTURE RESEARCH

5.1 Summary

We have witnessed versatile properties of wavelet methods for statistical applications in this dissertation. Mainly we use decorrelation property for Part I and denoising via wavelet thresholding technique for Part II.

The main goal of Part I is to develop an integrated tool for the detection of network anomalies and investigate performances using statistical analysis. We have proposed adaptations to wavelet packets of variance change detection methods and of a method for jump detection, and explored their implementation for online detection of network anomalies. These methods can capture several types of attacks against the network.

In Part II We have proposed a methodology for classification problems with ordinal outcomes. The method is well suited for the analysis of high-dimensional data and we have illustrated its applications using protein mass spectra from prostate cancer studies. The ordinal outcomes were defined in terms of PSA level. The prediction accuracies in both cases were between 70 and 80%. These error rates are partly due to the less than perfect specificity of these prognostic factors. Indeed, in order to identify reliable biomarkers, each outcome category must correspond to homogeneous groups.

We have also proposed wavelet-based techniques to remove noise from protein mass spectra. This procedure appears beneficial. We repeated the analysis without using the noise removal preprocessing step on the spectra. There were 93 markers identified by the peak detection and alignment procedures. We applied the ordinal probit model with variable selection to the data without wavelet thresholding procedure and obtained classification accuracies that were lower than those reported in previous chapter. This confirms the sensitivity of the results to noise in the data and the need for good preprocessing techniques. Here, we have used soft and adaptive wavelet thresholding to remove noise from the spectra. In future work, we will investigate alternative approaches, such as block shrinkage methods (see Cai, 1999). We conclude the second part by raising a couple of issues related to the analysis of SELDI-TOF mass spectra. It has been suggested that the current technology may be unreliable for low m/z values because of the effects of chemicals used to ionize the proteins. It may therefore be preferable to remove these low levels from the analysis. In addition, several criticisms have been raised on the use of SELDI-TOF technology for cancer detection (see Diamandis, 2004).

5.2 Future research

We focus on Bayesian ordinal probit model with Bayesian variable selection in the second part. Our approach for Bayesian ordinal probit is based on sampling from uniform distribution for cutoff point parameter, δ with Albert and Chib (1993). There are several advanced suggestions for sampling this parameter vecter in order to improve mixing ability such as Cowles (1996) and Nandram-Chen (1996) algorithms. We are going to investigate enhancement of our results with these above algorithms. Their works do not retain variable selection part. Our extension will be good endeavor in the Bayesian generalized linear models (GLMs).

We am going to consider other way by using a Bayesian clustering technique. Current work is based on the classification task with SELDI-TOF MS. Before we analyze data with Bayesian inference we do ‘curve clustering’ and find biomarkers.

REFERENCES

Akaike, H. (1974). “A new look at the statistical identification model.” *IEEE Transactions on Automatic Control*, 19, 716–723.

Albert, J. and Chib, S. (1993). “Bayesian analysis of binary and polychotomous response data.” *Journal of the American Statistical Association*, 88, 669–679.

Bolstad, B., Irizarry, R., Astrand, M., and Speed, T. (2003). “A comparison of normalization methods for high density oligonucleotide array data based on bias and variance.” *Bioinformatics*, 19, 185–193.

Brown, P., Vannucci, M., and Fearn, T. (1998a). “Bayesian wavelength selection in multicomponent analysis.” *Journal of Chemometrics*, 12, 173–182.

——— (1998b). “Multivariate Bayesian variable selection and prediction.” *Journal of Royal of the Statistical Society, Series B*, 60(3), 627–641.

——— (2002). “Bayes model averaging with selection of regressors.” *Journal of the Royal Statistical Society, Series B*, 64(3), 519–536.

Cai, T. (1999). “Adaptive wavelet estimation: A block thresholding and oracle inequality approach.” *Annanls of Statistics*, 27, 898–924.

Chen, J. and Gupta, A. (1997). “Testing and locating variance changepoints with application to stock prices.” *Journal of the American Statistical Association*, 92, 739–747.

Cowles, M. (1996). “Accelerating Markov chain Monte Carlo convergence for cumulative-link generalized linear models.” *Statistics and Computing*, 6, 883–904.

Daubechies, I. (1992). *Ten lectures on wavelets*. Philadelphia, SIAM.

Diamandis, E. (2004). “Analysis of serum proteomic patterns for early cancer diagnosis: Drawing attention to potential problems.” *Journal of the National Cancer Institute*, 96, 353–356.

Donoho, D. and Johnstone, I. (1994). “Ideal spatial adaptation by wavelet shrinkage.” *Biometrika*, 81(3), 425–455.

——— (1995). “Adapting to unknown smoothness via wavelet shrinkage.” *Journal of the American Statistical Association*, 90, 1200–1224.

Gabbanini, F., Vannucci, M., Bartoli, G., and Moro, A. (2004). “Wavelet packet methods for the analysis of variance of time series with application to crack widths on the Brunelleschi dome.” *Journal of Computational and Graphical Statistics*, 13(3), 639–658.

Gelfand, A. (1996). “Model determination using sampling-based methods.” In *Markov Chain Monte Carlo in Practice*, eds. W, Gilks, S, Richardson, and D, Spiegelhalter, 145–162. London, Chapman & Hall.

George, E. and McCulloch, R. (1997). “Approaches for Bayesian variable selection.” *Statistica Sinica*, 7, 339–373.

Gilbert, A. (2001). “Multiscale anaysis and data networks.” *Applied and Computational Harmonic Analysis*, 10(3), 185–202.

Gilbert, A., Willinger, W., and Feldman, A. (1999). “Scaling anlysis of random cascades, with applications to network traffic.” *IEEE Transactions on Information Theory*, 45(3), 971–991.

Inclán, C. and Tiao, G. (1994). "Use of cumulative sums of squares for retrospective detection of changes of variance." *Journal of the American Statistical Association*, 89, 913–923.

Jemal, A., Samuels, T., Ghafoor, A., Ward, E., and Thun, M. (2003). "Cancer statistics." *A Cancer Journal for Clinicians*, 53, 5–26.

Kim, S., Reddy, N., and Vannucci, M. (2004). "Detecting traffic anomalies through aggregate analysis of packet header data." *Proceedings of Networking 2004 (Athens, Greece, May)*, 3042, 1375–1384.

Lindley, D. (1957). "A statistical paradox." *Biometrika*, 44, 187–192.

Ljung, G. and Box, G. (1978). "On a measure of lack of fit in time series models." *Biometrika*, 65, 297–304.

Madigan, D. and Raftery, A. (1994). "Model selection and accounting for model uncertainty in graphical models using Occam's window." *Journal of the American Statistical Association*, 89, 1535–1546.

Mallat, S. (1989). "A theory of multiresolution signal decomposition: The wavelet representation." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11, 674–693.

Nandram, B. and Chen, M. (1996). "Reparameterizing the generalized linear model to accelerate Gibbs sampler convergence." *Journal of Statistical Computation and Simulation*, 54, 129–144.

Ogden, R. (1997). *Essential wavelets for statistical applications and data analysis*. Boston, Birkhuser.

Percival, D., Sardy, S., and Davison, A. (2000). "Wavestrapping time series: Adaptive wavelet-based bootstrapping." In *Nonlinear and Nonstationary Signal Processing*, eds. B. F, BJ, R, Smith, A, Walden, and P, Young, 442–470. Cambridge, UK, Cambridge University Press.

Percival, D. and Walden, A. (2000). *Wavelet methods for time series analysis*. London, Combridge University Press.

Petricoin, E., Ardekani, A., Hitt, B., Levine, P., Fusaro, V., Steinberg, S., Mills, G., Simone, C., Fishman, D., Kohn, E., and Liotta, L. (2002a). "Use of proteomic patterns in serum to identify ovarian cancer." *Lancet*, 359, 572–577.

Petricoin, E., Ornstein, D., Paweletz, C., Ardekani, A., Hackett, P., Hitt, B., Velassco, A., Trucco, C., Wiegand, L., Wood, K., Simone, C., Levine, P., Linehan, W., Emmert-Buck, M., Steinberg, S., Kohn, E., and Liotta, L. (2002b). "Serum proteomic pattern for detection of prostate cancer." *Journal of the National Cancer Institute*, 94, 1576–1578.

Resnick, S., G, G. S., Gilbert, A., and Willinger, W. (2003). "Wavelet analysis of conservative cascades." *Bernoulli*, 9, 97–135.

Riedi, R., Crouse, M., Ribeiro, V., and Baraniuk, R. (1999). "A multifractal wavelet model with application to network traffic." *IEEE transactions on Information Theory*, 45(3), 992–1018.

Schwarz, G. (1978). "Estimating the dimension of a model." *Annals of Statistics*, 6, 461–464.

Sha, N., Vannucci, M., Brown, P., Trower, M., Amphlett, G., and Falciani, F. (2003).

"Gene selection in arthritis classification with large-scale microarray expression profiles." *Comparative and Functional Genomics*, 4(2), 171–181.

Stamey, T. and Kabalin, J. (1989). "Prostate specific antigen in the diagnosis and treatment of adenocarcinoma of the prostate. I. untreated patients." *Journal of Urology*, 141, 1070–1075.

Stein, C. (1981). "Estimation of the mean of a multivariate normal distribution." *The Annals of Statistics*, 9, 1135–1151.

Vidakovic, B. (1999). *Statistical modelling by wavelets*. New York, Wiley.

Wang, Y. (1995). "Jump and sharp cusp detection by wavelets." *Biometrika*, 82(2), 385–397.

Whitcher, B. (2001). "Simulating gaussian stationary processes with unbounded spectra." *Journal of Computational and Graphical Statistics*, 10(1), 112–134.

Whitcher, B., Guttorp, P., and Percival, D. (2000). "Multiscale detection and location of multiple variance changes in the presence of long memory." *Journal of Statistical Computation and Simulation*, 68(1), 65–88.

Wickerhauser, M. (1994). *Adapted wavelet analysis from theory to software algorithms*. Massachusetts, A K Peters.

VITA

Deukwoo Kwon was born in Wonjoo, Korea. He received a Bachelor of Arts degree in economics from Yonsei University in Seoul, Korea in 1994. He received a Master of Business Administration degree in financial engineering from Korea Advanced Institute of Science and Technology, Korea in 2000 and Master of Science in statistics from Texas A&M University in College Station, Texas, under the direction of Dr. P. Fred Dahm in 2002. He continued his studies under the direction of Dr. Marina Vannucci, and received a Doctor of Philosophy degree from Texas A&M University in August 2005. His permanent address is 645-87 Bongcheon 1 dong, Kwanak-ku, Seoul, Korea.