Wavelet-based Data Reduction and Mining for Multiple Functional Data

A Thesis
Presented to
The Academic Faculty

by

Uk Jung

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

School of Industrial and Systems Engineering
Georgia Institute of Technology
July, 2004

Wavelet-based Data Reduction and Mining for Multiple Functional Data

Approved by:

Professor Jye-Chyi Lu, Committee Chair

Professor Xiaoming Huo

Professor Chen Zhou

Professor Roshan Vengazhiyil

Professor Farrokh Mistree
(Department of Mechanical Engineering)

Date Approved: 1 July 2004

With all my love

to my parents, brother, and sister

for their love, prayer, and encouragement

during this challenging journey

ACKNOWLEDGEMENTS

A wise man once said, "No man stands alone." I am not different in this endeavor. This work would not have been possible without the support of numerous people.

First, I would like to express my gratitude to my advisor, Dr. Jye-Chyi Lu, for his priceless advice, guidance, and financial support during this study. His constant encouragement and valuable advices are essential for the completion of this thesis. I will always be inspired by the high standards he sets for himself and his sharp focus and deep devotion to whatever he works on.

My parents, Yong-su Jung and Jung-ja Park, have served as role models of endless and unconditional love since I was born. They worked tirelessly to send my brother, sister, and me to college, even abroad. They showed me through their lives, the importance and power of education.

I would like to thank my thesis committee members: Dr. Chen Zhou, Dr. Farrokh Mistree, Dr. Xiaoming Huo, and Dr. Roshan Vengazhiyil for taking time to serve on my committee and for their questions and suggestions. My special thanks also go to Dr. Anthony Hayter and Dr. Brani Vidakovic for their effort and advice on my Ph.D. qualifying exam, and for the excellent text books they wrote, which benefit me a lot for my research work.

I also would like to thank all of members of the ISYE department, past and present, who have helped me overcome my mental blocks through numerous intellectual discussions.

Last but not least, all of my friends have been an unshakeable support to me for years and especially through the final and most difficult stage of my doctoral program.

The only thing I can make with all these people after this journey is a promise: a promise that my work is not finished yet, a promise that I will never stop developing my mind further, and a promise that I will do my best to live up to the reputation of Georgia Tech Ph.D. program.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

Advanced technology such as various types of automatic data acquisitions, management, and networking systems has created a tremendous capability for managers to access valuable production information to improve their operation quality and efficiency. Especially, due to the development of sensing and computer technology, on-line measurements of many process variables are available in current manufacturing processes. The functional data curve refers to an analog or digital signal measured during each operation cycle of a manufacturing process. In many manufacturing processes today, large volumes of functional data are being generated at an ever increasing pace. A set of Functional data is a class of very important in-process measurement, which contains rich information about the process condition and product quality for product design, process troubleshooting, quality/efficiency improvement and resource allocation decisions. It is highly desired to fully utilize the functional data curves for process monitoring and diagnosis. In this situation, signal processing and data mining techniques are more popular than ever in many fields, including intelligent manufacturing. As data sets increase in size, their exploration, manipulation, and analysis become more complicated and resource consuming. A major obstacle in those intelligent manufacturing systems is that tools for processing a large volume of information coming from numerous stages of manufacturing operations are not available. Thus, the underlying theme of this thesis is to reduce the size of data in a mathematically rigorous framework, and apply existing or new procedures to the reduced-size data for various decision-making purposes.

For the above purpose, wavelet transform is used in this research. Wavelet transforms model irregular data patterns such as cusps and lobes in a single curve better than the Fourier transform and standard statistical procedures. Most wavelets research in statistics focused on "data denoising" (also called "data shrinkage"), which screens out smaller sizes

of wavelet coefficients for removing data noises to obtain a smoother representation of the original data. Although they are proven to have quite good performance in a single curve shrinkage, a serious problem is often encountered when they are applied in data mining techniques for multiple curves, such as cluster analysis and classification. For example, when the existing shrinkage methods are applied to several curves which need to be analyzed together, different sets of selected wavelet coefficients for each curve are produced.

Most of the wavelet procedures are developed for a single data curve. The traditional typical wavelet model has a noise error component at each wavelet atom position to describe the narrow fluctuations at each time positions. This thesis, first, proposes *Wavelet-based Random-effect Model* which can generate multiple functional data signals which have wide fluctuations(between-signal variations) in the time domain. Also, the random-effect wavelet atom position in the model has *locally focused impact* which can be distinguished from other traditional random-effect models in biological field.

For the data-size reduction, in order to deal with heterogeneously selected wavelet coefficients for different single curves, this thesis introduces the newly-defined *Wavelet Vertical Energy* metric of multiple curves and utilizes it for the efficient data reduction method. As a result, if the vertical energy metric at certain wavelet atom position is large, it means that the wavelet atom position includes many important wavelet coefficients across all multiple curves, which represent most jumps or dips of each curve. The newly proposed method in this thesis will select important positions for the whole set of multiple curves by comparison between every vertical energy metrics and a threshold (*Vertical Energy Threshold; VET*) which will be optimally decided based on an objective function. The objective function balances the reconstruction error against a data reduction ratio. Also, the moment estimate of optimal threshold and its asymptotic properties are provided.

Based on class membership information of each signal obtained, this thesis proposes the *Vertical Group-Wise Threshold* method to increase the discriminative capability of the reduced-size data so that the reduced data set retains salient differences between classes as much as possible. The selection problem of class-wise thresholding scheme (intersection, union, and voting) is also briefly addressed. A new thresholding function using intersection

and a class-separability measure are proposed for finding the optimal threshold. A two-stage procedure based on these tools successfully increases the class separability with reasonably small loss of data reduction efficiency. Also, investigations on how several different situations can impact the performance of reconstruction accuracy, data reduction ratio, and class separability in the reduced-size data, are carried out using Monte-carlo simulations. A real-life example (Tonnage data) shows our proposed method is promising.

The following summarize the contributions of this thesis.

- A widely fluctuated multiple functional data signal set is well modelled using random-effects in wavelet domain. This has *locally focused impact*, which can be distinguished from other traditional random-effect models in biological field.

- A wavelet-based data reduction method for multiple curves is developed to deal with heterogeneously selected wavelet coefficients for different single curves, and its analytical properties are provided.

- The proposed data reduction method balances the reconstruction error against data reduction efficiency so that it is effective in capturing the key patterns in the multiple data signals. It also improves the time efficiency of clustering analysis.

- Based on class membership information of each signal obtained, the proposed method increases the discriminative capability of the reduced-size data. Consequently, the reduced data set retains salient differences between classes as much as possible.

CHAPTER I

INTRODUCTION

Advanced technology such as various types of automatic data acquisitions, management, and networking systems has created a tremendous capability for managers to access valuable production information to improve their operation quality and efficiency. Signal processing and data mining techniques are more popular than ever in many fields including intelligent manufacturing. As data sets increase in size, exploration, manipulation, and analysis become more complicated and resource consuming. Timely synthesized information is needed for product design, process trouble-shooting, quality/efficiency improvement and resource allocation decisions. A major obstacle in those intelligent manufacturing systems is that tools for processing a large volume of information coming from numerous stages of manufacturing operations are not available. Thus, the underlying theme of this thesis is to reduce the size of data in a mathematically rigorous framework, and apply existing and new procedures to the reduced-size data for various decision-making purposes.

There are many types of large size data requiring different data reduction techniques. The data studied in this thesis do not have many attributes (e.g., data from grocery sales) for "dimension reduction" typically performed in PCA(Principle Component Analysis) and other multivariate analysis. Many manufacturing practices indicated difficulties in handling complicated 'functional data' with nonstationary, correlated and dynamically changing patterns contributed from potential process faults, which are difficult to handle for standard data modelling techniques such as Fourier transform, polynomial regression, time series and neural network. Figure 1(a) shows an example taken from Nortel's antenna manufacturing process which has several nonstationary sharp-changes characterizing process behaviors or product characteristics. For example, the peaks and valleys in the center three main lobes of the data shown in Figure 1(b),which is an azimuth cut of Figure 1 (a), are very important for antenna signal quality. See Gardner *et al.* (1997), Bakshi (1998), Jin and Shi (1999),

Ganesan, Das, Sikder and Kumar (2002), Lada, Lu and Wilson (2002) for other studies of complicated functional data in semiconductor fabrication, chemical manufacturing and sheet-metal stamping applications.

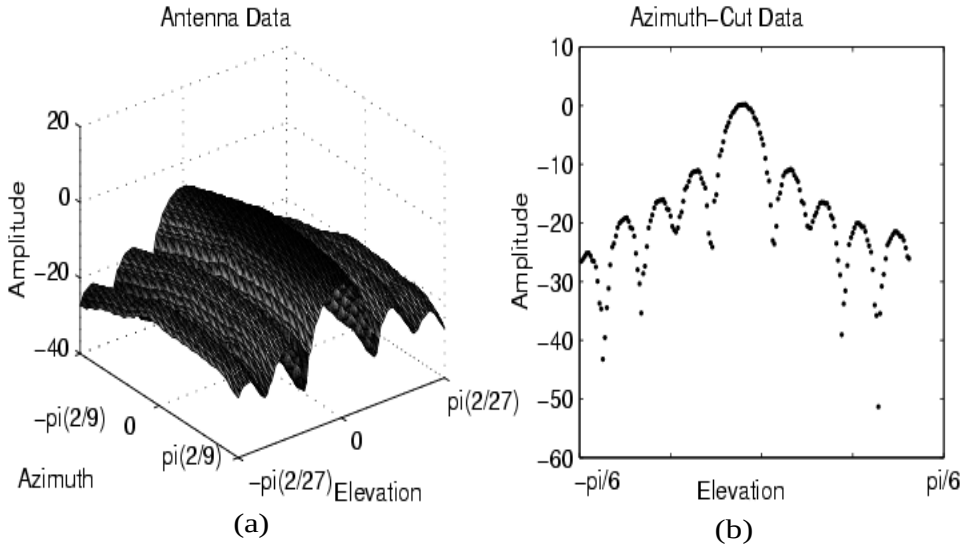
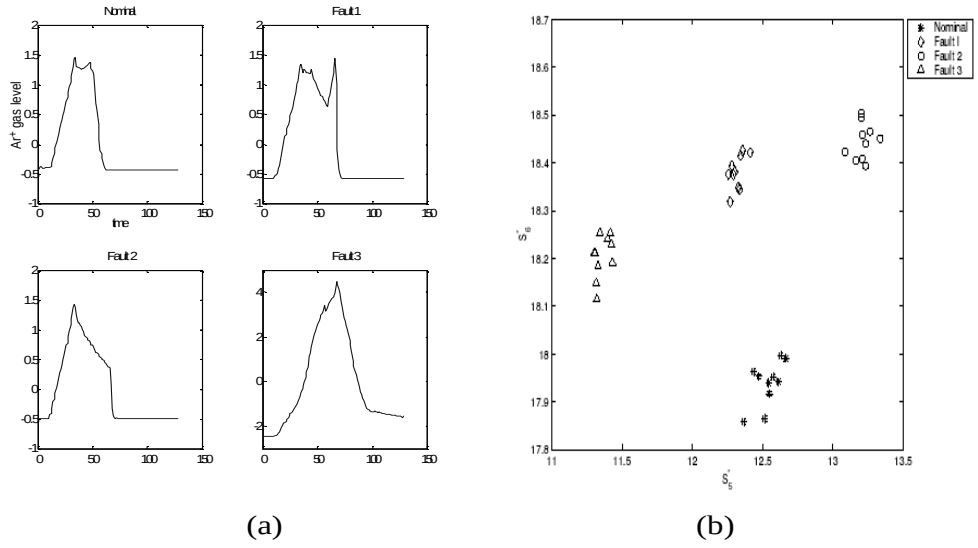


Figure 1: Data Signals from Nortel's Antenna Manufacturing Process

Most of the models and analysis of complicated functional data involve only one sequence of data. Examples are in Rao and Bopardikar (1998; Chapter 5), Jeong, Lu, Huo, Vidakovic and Chen (2003) and references therein. Many applications require multiple sets of complicated functional data. For instance, to estimate model parameters in a Phase-I study for establishing control limits in statistical process control (SPC) procedures (see Woodall (2000) for its definition), 18 sets of the antenna data are randomly collected under a baseline process condition. The two-dimensional antenna data presented in Figure 1(a) can be also viewed as multiple sets of (shifted in parallel) one-dimension functional data shown in Figure 1(b). The following briefly presents a motivating example for analyzing multiple sets of functional data.

Example 1. In a process of developing a pattern recognition procedure for understanding process faults in a semiconductor thin-film deposition experiment (see Rying, Ozturk, Bilbro and Lu (2003) for details), several sets of functional data were collected

from various process conditions. Figure 2 shows the separation of these curves into distinct classes based on the two most representative energy metrics, which are sums of squares of wavelet coefficients at different resolution levels. These energy metrics are called scalogram (Vidakovic 1999, page 289) in the signal processing field.



(a) (b)

Figure 2: Four Types of Signals from a Semiconductor Manufacturing Process.

Ganesan *et al.* (2002) and Lada *et al.* (2002) are other examples of engineering studies for understanding process behaviors and fault patterns with multiple sets of functional data. In biological and medical studies, many sets of repeated measurements collected at successive time points are analyzed for examining treatment effects and patient characteristics. See Zhang, Lin, Ras and Sowers (1998) for an example.

Due to the advance of information technology, larger amounts of data are now available for improving process quality and operation efficiency. There are many different purposes of using the reduced-size(RS) data. If RS data are constructed to detect specific types of known faults, a reduction procedure could be derived to minimize Type-one and /or Type-two errors in testing occurrence of fault. However, this data set might not be suitable for other purposes of analysis(e.g., failure prediction, design of experiment(DOE) for quality/efficiency improvement) or for handling cases when the fault patterns were changed.

Thus, the aim of our data reduction is to produce a small set of representative data suitable for many kinds of decisions of analysis for a set of functional curves. Moreover, if it is necessary, an accurate 'approximation' of the original data could be obtained for many types of analysis, (i.e., our procedure has the 'data compression' properties.) Ideally, these RS data could be combined with other data in the original or RS formats collected from many systems at different process stages and possibly located at various sites for process characterization, optimization and strategic planning purposes. As reviewed in Jeong *et al.* (2003), there are limited studies on analyzing potentially large-size *complicated* functional data. Some approaches do exist for analyzing data with smooth patterns. These include Functional Principle Component Analysis (PFCA; Ramsay and Silverman, 1997) and related procedures, e.g., Hall, Poskitt and Presnell's (2001) proposal of coordinates for representing their data curves. In dealing with complicated data patterns, engineering knowledge is commonly used to guide data preserving or feature extraction methods (Jin and Shi, 1999) for selecting representative data in smaller size for subsequent analyses. Many studies (e.g., Jin and Shi, 2001) used wavelet-based data denoising techniques (Donoho and Johnstone, 1994 and 1995) for data-reduction purposes. More examples are given in the data reduction paper by Jeong *et al.* (2003). Wavelet-based procedures are popular in these publications due to their ability to model sharp-changes in data patterns and the multi-scale data compression property. Thus, our studies will focus on wavelets.

Wavelet transforms model irregular data patterns such as cusps and lobes in Figure 1 (b) better than the standard statistical procedures mentioned above, and provide a *multi-resolution approximation* to the data (Mallat, 1998). Wavelet transforms have been demonstrated very useful in image and audio compression practices(e.g., Rao and Bopardikar, 1998; Chapter 5) and many data-denosing studies (e.g., Donoho and Johnstone, 1994) in various applications. However, the existing methods using wavelet are mostly about selecting representative wavelet coefficients for only a single curve. Then, for a system with large number of curves, the wavelet atom positions of selected wavelet coefficients are mostly different for different curves so that those methods unavoidably choose many unnecessary coefficients through a union concept(choose all coefficients at a wavelet atom position if

at least one of them is selected) for the use of data mining techniques such as clustering and classification. It may bring a serious inefficiency of data reduction for the whole set of curves. Thus, this thesis will focus on wavelet-based data reduction procedures for complicated multiple functional data, which can achieve high efficiency against the heterogeneously selected wavelet positions.

Chapter 2 reviews relevant literatures on the topics of this research, such as functional data analysis, data reduction and mining, and wavelet transformation, and compares them to this research. Chapter 3 introduces and explains the details about newly- defined *Wavelet-based Random-effect Model*. Most of the wavelet procedures are developed for a single data curve. The traditional typical wavelet model has a noise error component at each wavelet atom position to describe the narrow fluctuations at each time positions. This thesis, first, proposes *Wavelet-based Random-effect Model* which can generate multiple functional data signals which have wide fluctuations(between-signal variations) in time domain. Also, the random-effect wavelet atom position in the model has *locally focused impact* which can be distinguished from other traditional random-effect models in biological field.

For the data-size reduction, in order to deal with heterogeneously selected wavelet coefficients for different single curves, Chapter 4 introduces the *Wavelet Vertical Energy* metric, that is newly defined, of multiple curves and utilize it for the efficient data reduction method. If the wavelet vertical energy metric at certain wavelet atom position is large, the wavelet atom position includes many important wavelet coefficients across all multiple curves, which represent most jumps or dips of each curves. The newly proposed method in this chapter will select important positions for the whole set of multiple curves by comparison between every vertical energy metrics and a threshold (*Vertical Energy Threshold; VET*) which will be optimally decided based on an objective function. The objective function balances the reconstruction error against a data reduction ratio. Also the moment estimate of optimal threshold and its asymptotic properties are provided.

As far as the case that class membership information of each functional curve is available is concerned, another thresholding scheme, *Vertical Group-wise Threshold*, is explored in Chapter 5 with several data selection strategies (Union, Intersection, and Voting). This

chapter also combines the class-separability concept with the key components in the objective function of VET method. This combination in the *Vertical Group-wise Threshold(VGWT)* method was motivated by the idea that the reduced size data can maximize the ability to retain salient differences between classes. The class separability term given by the between-class variability using the class mean at each wavelet position is defined. The absolute value of each class means at each wavelet atom positions are taken into account to compare with a common threshold(VGWT). In order to achieve several purposes such as high signal reconstruction accuracy, efficient data reduction and retaining the class separability as much as possible in the reduced-size data, the guideline to get the optimal threshold is proposed.

Finally, Chapter 6 states the summary of results in the thesis and possible future research problems.

CHAPTER II

LITERATURE REVIEWS

The topics of this research can be categorized as the following three factors.

• Domain : Multiple functional data

• Purpose : Data reduction and mining

• Methodology : Wavelet transformation

In this section, some details to key references in each topic will be provided pointing out some that have been overlooked and not yet been explored.

2.1 Literature on Functional Data Analysis

Most statistical analysis involves one or more observations taken on each of a number of individuals in a sample, with the aim of making inferences about the general population from which the sample is drawn. In an increasing number of fields, these observations are curves or images. Curves and images are examples of functions, since an observed intensity ia available at each point on a line segment, a portion of a plane, or a volume. For this reason, we call observed curves and images 'functional data', and statistical methods for analyzing such data are described by the term 'functional data analysis' (FDA), coined by Ramsay and Dalzell (1991).

The goals of functional data analysis are essentially the same as for other branches of statistics, and include the following: (a) to represent and transform the data in ways that aid further analysis, (b) to display the data so as to highlight various characteristics, (c) to study important sources of pattern and variation among the data, and (d) to explain variation in an outcome or dependent variable by using input or independent variable information. Ramsay and Silverman (1997) illustrated the nature of functional data, these goals, and FDA tools through a series of examples, such as (a) Human Growth Data: Looking at Velocity and Acceleration, (b) The Mean Function and the Registration Problem,

(c) The Nondurable Goods Index and More Derivatives, (d) Functional Principle Components Analysis, (e) A Functional Linear Model and Regularization, and (f) Modeling with Derivatives: A Central Theme, etc.

Some research work has been reported on fully utilizing the functional data for the process monitoring and diagnosis purposes in manufacturing system. Jin and Shi(1999) proposed a "feature preserving" procedure to extract patterns in the waveform signal and link them to corresponding faults in stamping process. Pittner and Kamarthi(1999) proposed a wavelet-based procedure for feature extraction of waveform signals. They transform the waveform signals into the wavelet domain and then select the wavelet coefficients based on the magnitude of the coefficients. Lada, et al(2002) proposed a wavelet coefficient selection procedure not only based on the magnitude of the coefficients , but also based on an additional term that penalizes the number of selected coefficients. The purpose is to keep the number of the wavelet coefficients small to simplify further analysis.

From above review, the available FDA in manufacturing domain either focus on (a) representing, transforming, and displaying the data so as to highlight various characteristic or (b) extract features in data to study important sources of pattern and variation among the data. Very few effort have been made on the concept of efficient data reduction for the whole set of functional data. Jeong and Lu(2003) proposed the method of wavelet-based data reduction techniques for process fault detection. The proposed method minimized the objective functions to balance the tradeoff between data reduction and modelling accuracy for a single curve. For a system with large number of curves, the selected wavelet coefficients are mostly different for each curves. Then this method unavoidably choose many unnecessary coefficients through union concept for the use of data mining techniques such as CART so that the data reduction ratio of the whole set of curves become considerably low. In order to achieve high efficiency of data reduction, new approach on modelling and analyzing a set of multiple data curves are required. In this thesis, the wavelet based method to analyze a set of functional data curves will be proposed for the purpose.

2.2 Literature on Data Reduction and Mining

2.2.1 Data Mining

Data mining, the objective of which is to make predictions or discoveries involving a large amount of data, is an exciting field for both researchers and practitioners. However, *data mining* means different things to different people. For example the requirements and expectations for data mining in business and science-oriented applications may be quite different. Nevertheless, in parallel to diverging trends in various application, important common themes have also emerged from various application.

Articles and books on data mining are abundant. Data mining has been a favorite topic by academic researchers as well as business practitioners and is often discussed from very different perspectives. There are numerous books on data mining for practitioners, addressing practical concerns. For example, Berry and Linoff(1997) presented a wide range of data mining methods, and Westphal and Blaxton(1998) described how to use existing commercial tools to conduct data mining. On the other hand, within academia, issues related to data mining have been studied from different perspectives such as statistics, pattern recognition, database management system, and artificial intelligence(AI). For exmaple, the book by Fayyad et al.(1996) is a well-known volume on some research progress up to the year 1995. Kennedy et al.(1997) discussed pattern recognition techniques for data mining, and Cios et al(1998) promoted the use of computational intelligence techniques (such as rough theory, fuzzy logic, artificial neural networks) for data mining. More recently, Han and Kamber(2000) presented an excellent discussion on data mining mainly from a database perspective.

Note that different application may have very different focuses. For example, scientific data mining seems to focus mostly on finding explanations for the most variable elements of the data set(i.e., to find and explain the outliers). For example, one may want to understand the purchasing habits of most of our customers (skillicorn 1999). Applications of data mining include

1. Medicine, such as diagnosis and prognosis

2. Public administration

3. Marketing and finance

4. Scientific database

5. Fraud detection

6. Engineering, such as diagnostics of mechanisms and process

7. Data mining on the Web in text and heterogeneous data

Some illustrate the strong need of data mining research in manufacturing process, e.g., chemical manufacturing(Bakshi, 1998), nanomachining (e.g., see http://www.eng.usf.edu/das for the work of T.K.Das and his colleagues in chemical mechanical planarization(CMP) processes), semiconductor fabrication. However, the attention in this direction is far less than what "business operation" have received. To our knowledge, most of successful data mining applications with large size data are in grocery- for fashion-goods-sale studies, customer relationship management, telecommunication fraud analysis, etc. The recent book edited by Braha (2001) on *Data Mining for Design and Manufacturing* made attempts to bring engineers' attention in this important research area.

2.2.2 Data Reduction

Data can be reduced to a simpler form; for example, continuous variables can be discretized to get range variables. If we push this view point a little further, we can claim that *the task of data mining as a whole, can be viewed as a reduction process.* After all, the result of data mining is a more concise description of the original data themselves. The role of data mining is to discovery general patterns that describe the data. Theses patterns may have the form of rules, or some model. Each generated pattern represents a subset of the raw data. Knowledge extraction can be achieved through data reduction. Chen (2001) addressed in his book that some basic data mining techniques such as clustering, sampling, along with some other methods such as visualization(such as the use of histograms), singular value decomposition, wavelets, regression, loglinear models, can viewed as a kind of reduction method.

Lu(2001) summarized many data reduction procedures into three categories: Sampling Approaches, Modelling and Transformation Techniques, and Data Splitting Methods. The well known systematic, stratified and segmentation sampling methods and feature extraction procedures(e.g., Mallat and Hwang, 1992; Jin and Shi, 1999) are examples of sampling approaches. The classical regression, principle components analysis (PCA), Fourier and wavelet transformations and simple summary statistics (e.g., mean, variance) are examples of the second category. Similar (but different) to sampling approaches, data splitting methods such as kd-trees and c-means clustering are very useful in reducing data sizes.

As mentioned before, this thesis will focus on wavelet-based data reduction procedures for complicated functional data. In many engineering application (e.g., Lada, et al.(2002)) of the data de-nosing and the AMDL methods, we found that many coefficients were used to achieve a very small signal reconstruction error. By experimenting various numbers of coefficients used in the nonlinear signal approximation methods, we found that many sets of reconstructed signals using a fewer number of coefficients provide a very reasonable approximation to the original data. See Jeong and Lu (2002). More importantly, the selected wavelet coefficients were rather representative in most of the data analysis, e.g., chi-square test for process fault detection (e.g., Lada, et al. (2002)) or decision tree analysis for process fault classification (Jeong and Lu, 2002). This motivate us to search for a more aggressive "data reduction" method for multiple functional data curves.

Aggressive data reduction method for multiple functional data has a great potential at the age of high technology. The advance of computer, networking systems and automatic data acquisition instruments facilitates the growth of information in a form of functional data curves. Functional data curves represent a class of analog or digital signals over time, which normally can be measured using in-process sensors in a manufacturing process. It has broad potential applications, such as tonnage signals in stamping, torque signals in tapping, and force signals in welding, etc(See Jin and Shi(2001)). While those in-process measurements contain rich information about process condition and product quality, the massive amount of measurement data are a major obstacle to achieve quality and efficiency improvement in relatively short time. Also, if the data size is not very large, one can use

visualization techniques to examine potential systematic data patterns. However, when the data size become larger, visualization of large size data become more difficult.

2.3 Literature on Wavelet Transformation

2.3.1 Introduction to Discrete Wavelet Transformation(DWT)

In order to introduce the new thresholding method for representing well the whole set of multiple curves and keep the high data reduction ratio, the Wavelet transformation is briefly reviewed below.

A wavelet is a function $\psi(t) \in L^2(\mathbb{R})$ with the following basic properties

$$\int_{\mathbb{R}} \psi(t)\, dt = 0 \quad \text{and} \quad \int_{\mathbb{R}} \psi^2(t)\, dt = 1,$$

where $L^2(\mathbb{R})$ is the space of square integrable real functions defined on the real line $\mathbb{R}$ Wavelets can be used to create a family of time-frequency atoms, $\psi_{s,u}(t) = s^{1/2}\psi(st-u)$, via the dilation factor s and the translation u. We also require a scaling function $\phi(t) \in L^2(\mathbb{R})$ that satisfies

$$\int_{\mathbb{R}} \phi(t)\, dt \neq 0 \quad \text{and} \quad \int_{\mathbb{R}} \phi^2(t)\, dt = 1.$$

Selecting the scaling and wavelet functions as $\{\phi_{L,k}(t) = 2^{L/2}\phi(2^L t - k); k \in \mathbb{Z}\}$, $\{\psi_{j,k}(t) = 2^{j/2}\psi(2^j t - k); j \geq L, k \in \mathbb{Z}\}$, respectively, one can form an orthonormal basis to represent a signal function $f(t) \in L^2(\mathbb{R})$ as follows.

$$f(t) = \sum_{k\in\mathbb{Z}} c_{L,k}\phi_{L,k}(t) + \sum_{j\geq L}\sum_{k\in\mathbb{Z}} d_{j,k}\psi_{j,k}(t)$$

where $\mathbb{Z}$ denote the set of all integers $\{0, \pm 1, \pm 2, \ldots\}$, and the coefficients $c_{L,k} = \int_{\mathbb{R}} f(t)\phi_{L,k}(t)\, dt$ are considered to be the coarser-level coefficients characterizing smoother data patterns, and $d_{j,k} = \int_{\mathbb{R}} f(t)\psi_{j,k}(t)\, dt$ are viewed as the finer-level coefficients describing (local) details of data patterns. In practice, the following finite version of the wavelet series approximation is used:

$$\tilde{f}(t) = \sum_{k\in\mathbb{Z}} c_{L,k}\phi_{L,k}(t) + \sum_{j=L}^{J}\sum_{k\in\mathbb{Z}} d_{j,k}\psi_{j,k}(t)\,, \tag{1}$$

here $J > L$ and L correspond to the coarsest resolution level. Consider a sequence of data $\mathbf{y} = (y(t_1), \cdots, y(t_N))'$ taken from $f(t)$ or obtained as a realization of $y(t) = f(t) + \epsilon_t$ at

equally spaced discrete time points $t = t_i$'s, where ϵ_{t_i}'s are independent and identically distributed (i.i.d.) noises. The discrete wavelet transform (DWT) of $\mathbf{y}$ is defined as

$$\mathbf{d} = \boldsymbol{W}\mathbf{y}$$

where $\boldsymbol{W}$ is the orthonormal $N \times N$ DWT-matrix. From (1), we can write $\mathbf{d} = (\mathbf{c}_L, \mathbf{d}_L, \mathbf{d}_{L+1}, \cdots, \mathbf{d}_J)$, where $\mathbf{c}_L = (c_{L,0}, \cdots, c_{L,2^L-1}), \mathbf{d}_L = (d_{L,0}, \cdots, d_{L,2^L-1}), \cdots, \mathbf{d}_J = (d_{J,0}, \cdots, d_{J,2^J-1})$ are called scales or subbands. Using the inverse DWT, the $N \times 1$ vector $\mathbf{y}$ of the original signal curve can be "reconstructed" as $\mathbf{y} = \boldsymbol{W}'\mathbf{d}$. The process of transforming a data set via the DWT closely resembles the process of computing the Fast Fourier Transformation (FFT) of that data set. By applying the DWT to the data y_i's, $\mathbf{d} = \boldsymbol{W}\boldsymbol{y}$, we obtain the following model in the wavelet domain: $d_{j,k} = \theta_{j,k} + \eta_{j,k}$, for $j = L, \cdots, J$, $k = 0, 1, \cdots, 2^j - 1$, and $c_{L,k} = \theta_{L,k} + \eta_{L,k}$, for $k = 0, 1, \cdots, 2^L - 1$, where $J = \log_2 N - 1$. The model can be represented in the vector format as follows.

$$\mathbf{d} = \boldsymbol{\theta} + \boldsymbol{\eta} \tag{2}$$

where $\mathbf{d}, \boldsymbol{\theta}$ and $\boldsymbol{\eta}$ represent the collection of all coefficients, parameters and errors, respectively. Since $\boldsymbol{W}$ is an orthonormal transform, $\eta_{j,k}$'s are still i.i.d. $N(0, \sigma^2)$ (Vidakovic 1999, page 169). To simplify the notation, we use $\boldsymbol{d} = (d_1, d_2, \ldots, d_N)^\top$ instead of using c_{Lk}, d_{jk} for the components of $\boldsymbol{d}$ without any confusing.

2.3.2 Variety of Applications

Strictly speaking, wavelets are topic of pure mathematics, however in only a few years of existence as a theory of their own, they have shown great potential and applicability in many fields. Wavelet analysis is shown to be useful in handling irregular data patterns and in effectively reducing the data into a smaller number of representative wavelet coefficients. At present, statistical applications of wavelets predominately concentrate on curve estimation(Donoho and Jonston 1994 and 1995), time series analysis(Moulin 1994), survival analysis(Antoniadis 1999), classification(Learned and Willsky 1995) and Bayesian analysis(Vidakovic 1998). With increasing usage of automatic data collection tools, wavelet

models can be very important in the intelligent manufacturing research, where a huge amount of data must be analyzed in real-time for process control and improvement.

Some research work has been reported on applying wavelet transformation to functional data for the process monitoring and diagnosis purpose in manufacturing system. To detect faults in a time-dependent process, Lada, et al(2002) applied a discrete wavelet transform(DWT) to several independently replicated data sets generated by that process. The DWT can capture irregular data patterns such as sharp "jumps" better than the Fourier transform and standard procedures without adding much computational complexity. Their wavelet coefficient selection method effectively balances model parsimony against data reconstruction error. The few selected wavelet coefficients serve as the "reduced size" data set to facilitate an efficient decision-making method in situation with potentially large-volume data sets. Jin and Shi (2001) proposed an automatic feature extraction methodology for fault diagnostics purposes. In particular, for the monitoring of the normal process condition, they applied the Hotelling(1947) T^2 statistic to construct the SPC(statistical process control) limits. The data used in their T^2 statitic is the "denoised" wavelet coefficients from the Visu method developed in Donoho and Johnstone (1994). Their application data are from stamping processes. Koh, et al(1999) introduced an uniformly most powerful test on individual coefficients of the Haar transformation (one of wavelet families) of the cycle-based waveform signal. Based on this test, a monitoring system that is similar to Shewhart control chart is proposed to distinguish normal and abnormal conditions of the process based on cycle-based signals.

2.3.3 Threshold Strategy

One method often used to fit a single curve data using wavelets is to compute a set of multi-resolution approximation (Mallot,1998). This method involves first constructing an approximation to the data using the coarsest-scale signal and then adding increasingly finer levels of resolution. As more levels of resolution are used, the approximation to the target data set improves. At the finest level of resolution, the total number of estimated wavelet coefficients equals the size of the single curve data set so that the data set is exactly

reconstructed. While easy to use, this type of "linear" multi-resolution approximation tends to over-smooth the data. In order to avoid this drawback, Donoho and Johnston(1995) developed several wavelet based "shrinkage" techniques which is nonlinear approximation method to accurately represent small jumps or dips in the data. Nonlinear methods that select "important" wavelet coefficients(usually the largest in magnitude) and set to zero the "unimportant" coefficients(usually those representing noise) are effective with fewer coefficients than an approach based on a straightforward multi-resolution approximation. In the shrinkage scheme, wavelet coefficients are set to zero if their absolute values are below a certain *Threshold level*, $\lambda > 0$. Since the small size of wavelet coefficients in magnitude are usually contributed from data noises, thresholding out these coefficients has an effect of "removing data noises" so that the shrinkage methods are called data de-noising methods.

In using any type of wavelet threshold, the main issue is how to choose the threshold value λ. Choosing a very large threshold will make it difficult for a coefficient to be included in the data signal reconstruction, consequently resulting in an over-smoothing of the data curve. On the other hand, choosing a very small threshold value will allow many coefficients to be included in the reconstruction, giving a result close to the original noisy signal. The proper choice of threshold involves a careful balance of these principles. Comprehensive overview for threshold selection is given in Antoniadis, Gijbels and Gregoire (1997).

There are many wavelet model selection procedures in the literature that are based on the idea of selecting "important" wavelet coefficients and setting to zero the "unimportant" coefficients. These methods attempt to find an optimal number of coefficients to accurately represent the data, thereby leading to a simplified and smoother (less noisy)data. The next paragraph will briefly review the following three methods without giving their technical details.

SURE(Stein's Unbiased Risk Estimate) method proposed by Donoho and Johnston(1995) introduced a scheme that uses the wavelet coefficients at each resolution level to choose a different threshold. Wavelet coefficients smaller than the level-dependent threshold are set to zero. This method is very popular in practice. The AMDL (Approximation Minimum Description Length) procedure is proposed by Saito(1994). It minimize the cost function

AMDL(C) where C is the number of wavelet coefficients selected to be nonzero. As addressed in Antoniadis et al(1997) the AMDL function is similar to the Akaike information quantity commonly used in many statistical model selection procedures, including linear regression models. Jeong and Lu(2002) developed RRE method to find the optimal threshold λ to minimize the objective function which balances the goals of decreasing errors in the signal reconstruction and increasing the data reduction ratio. Its threshold level depends on signal in terms of its energy, and does not require the estimation of variance of noise while other data shrinkage methods require it.

All the thresholding methods introduced in most literatures so far is based on a single curve shrinkage concept. They are proved to have a quite good performance in a single curve shrinkage concept itself. However, when they are utilized for some data mining techniques for multiple curves, such as cluster analysis and classification, a serious problem is often confronted. For example, when we apply the existing shrinkage methods to several curves which are analyzed together, we usually experience the different sets of selected wavelet coefficients for each curve. This means that there is sometimes no selected wavelet coefficient at a certain position on certain curve which is supposed to be corresponding to other selected ones for different curve. It means there is nothing compared to other curves' information. The usual way to tackle this problem in the previous literatures (Jeong et al(2002), Jin and Shi(2001)) is that they used the "Union" concept. This means, if there is a selected wavelet coefficient at a certain position, other curves' wavelet coefficients at the same position are also selected. This method may bring a serious inefficiency of data reduction for the whole set of curves since there must be some unnecessary wavelet coefficients to represent and contain important information of the original curves.

Clearly, the above thresholding rule incurs the reduced size data from original data since we only pay attention to relatively a few thresholded coefficients. Thus, in this research, those thresholded data are treated as reduced size data in process fault detection, classification and other decisions for improving process quality. As far as it is concerned to deal with multiple functional data curves with the problem of heterogeneously selected wavelet coefficients mentioned in the above paragraph, few thresholding ideas for whole multiple

cures were come up with in the literature. In this thesis, we develop new thresholding method considering all multiple curves for high efficiency of data reduction and model representation.

CHAPTER III

WAVELET-BASED RANDOM-EFFECT MODEL

3.1 Introduction

Most of the wavelet procedures are developed for a single data curve. A typical model assumed in the above studies is $y(t) = f(t) + z(t)$, where the mean function $f(t)$ for the data $y(t)$ collected at time point t is a sum of wavelet coefficients multiplied by their wavelet bases as shown in Eq. (3) in Section 3.2, and the errors $z(t)$ at all time points are independent and identically distributed as normal with mean zero and variance σ^2. Figure 3 shows that this type of model will not be able to generate the multiple sets of functional data such as the 18 sets of antenna data we collected. Although only single curves at different levels of noise are shown, one can imagine that when multiple sets of data curves are generated, the wider fluctuations in the left and right sides of the antenna curves presented in Figure 3(a) cannot be produced from the above model. Moreover, data-denoising or -reduction procedures developed for single data curves cannot capture the common characteristics among all curves. Similar to random-effect models advocated in the biological and medical studies of repeated measurements, this chapter explores a random-effect model in the wavelet domain for a type of data set like in Figure 3(a).

Random-effect models in the wavelet domain are quite different from the traditional models used in the biological studies, where usually an intercept or slope is considered as random and the impact of random changes of this effect is well understood. In general, the wavelet-based random-effect model will have *locally focused impact*. In particular, if the random-effect is placed on the coarser level of coefficients, random changes of a certain wavelet coefficient could have a wide-range impact on many data points in the time domain. If the random-effect is placed on the finer level of coefficients, the range of time-domain data affected will be much narrower. Moreover, the supports of the coefficients in the coarser levels are overlapped. This leads to a possible "compounding" effect from two adjacent

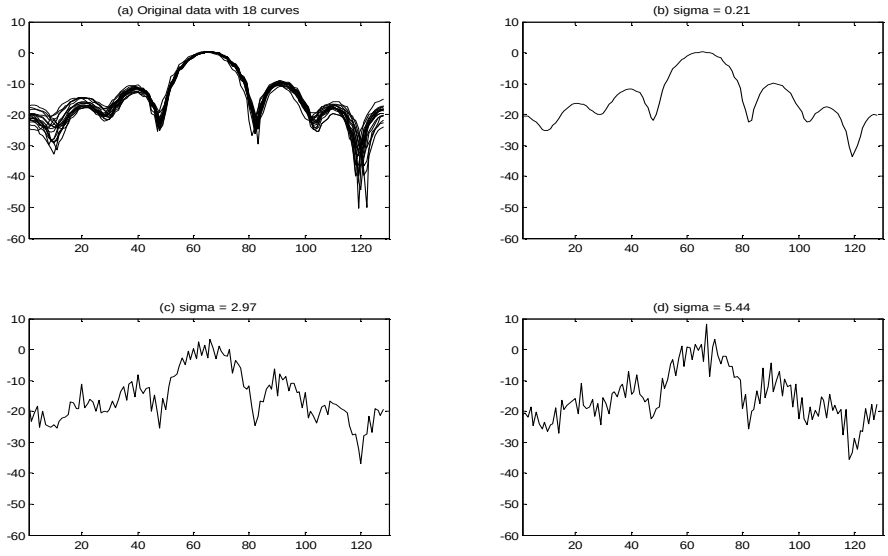


Figure 3: Problems with the Traditional Data Models

coefficients when they are both random. When there are multiple random-effects at various resolution levels of coefficients, the compounded impact is complicated. See section 3.2 for details of the proposed random-effect model and for studies of exploring these properties.

3.2 Model Formulation

Denote by $\boldsymbol{y}_i = [y_{i1}, y_{i2}, \cdots, y_{iN}]^T$ a vector of N equally-spaced data points from a signal curve where $N = 2^J$ with some positive integer J and $i = 1, 2, \cdots, M$. The superscript T represents the transpose operator. Let $\boldsymbol{Y} = [\boldsymbol{y}_1^T, \boldsymbol{y}_2^T, \cdots, \boldsymbol{y}_M^T]^T$ be the collection of M multiple sets of functional data. When a discrete wavelet transform (DWT) $\boldsymbol{W}$ is applied to a data set, the matrix of wavelet coefficients obtained from this transformation is $\boldsymbol{D} = \boldsymbol{YW}$, where $\boldsymbol{D} = [\boldsymbol{d}_1^T, \boldsymbol{d}_2^T, \cdots, \boldsymbol{d}_M^T]^T$, $\boldsymbol{d}_i = [d_{i1}, d_{i2}, \cdots, d_{iN}]^T$, and d_{im} is the wavelet coefficient at the mth wavelet-position for the ith data curve. See Mallat (1998; Chapter IV) for details of the discrete wavelet transform. When $\boldsymbol{W}$ is orthonormal, the original observations $\boldsymbol{Y}$ can be recovered using the inverse DWT. That is, through $\boldsymbol{Y} = \boldsymbol{DW}^T$ the original data can be expressed as a linear sum of products of wavelet coefficients ($c_{L,k}$ and $d_{j,k}$) and their

corresponding wavelet-basis functions ($\phi_{L,k}(t)$ and $\psi_{j,k}(t)$) as follows:

$$\tilde{f}(t) = \sum_{k\in\mathbb{Z}} c_{L,k}\phi_{L,k}(t) + \sum_{j=L}^{J}\sum_{k\in\mathbb{Z}} d_{j,k}\psi_{j,k}(t), \tag{3}$$

where the resolution (or scale) level J is greater than the coarsest level L, and $\mathbb{Z}$ is the set of positive integers. Note that if there are N data points, there will be N wavelet coefficients. The data reduction procedure rests on the idea that many of the coefficients are set as zeros such that only a smaller number of bases are used to "reconstruct" the original data curve. See Vidokovic (1999) for more details of the wavelet approximation models.

In the literature of random-effect models (Zhang *et al.*, 1998), which variable is a random-effect and which is a fixed-effect are determined based on "external information" (e.g., covariate or prior knowledge). This research will follow that tradition. However, we propose a simple guideline to find the random-effect variable. See Example 2 for the use of the normal probability plot to identify the seven random effects from all wavelet coefficients for modelling antenna data curves. The following examples explains the role of the random-effect in the wavelet models.

Example 2. Consider a simple example with the functional data shown in Figure 4(a) (in solid line) from the nominal process of a semiconductor thin-film deposition experiment. Assume that there is only one random coefficient $c_{4,2}$. The dotted line in Figure 4(a) represents the data-curve reconstructed with only fixed-effect wavelet coefficients. The solid line in Figure 4(b) presents the reconstructed data curve with both fixed- and random-effect curves added together, where the dashed line is a realization of the random-effect $c_{4,2}$ multiplied by its wavelet basis. Note that only the data in the support area around 50 to 85 are affected by this random effect. Figure 4(c) presents another set of curves similar to Figure 4(b), but the random-effect $c_{4,2}$ is generated far away from its mean. Thus, the shape of the dashed line is quite different from what we see in Figure 4(b), and then the sum of the fixed- and random-effects will be different from the original curve. Most difference occurs around the peak (around the time point 70). See Example 4 for more explorations with the real-life antenna data.

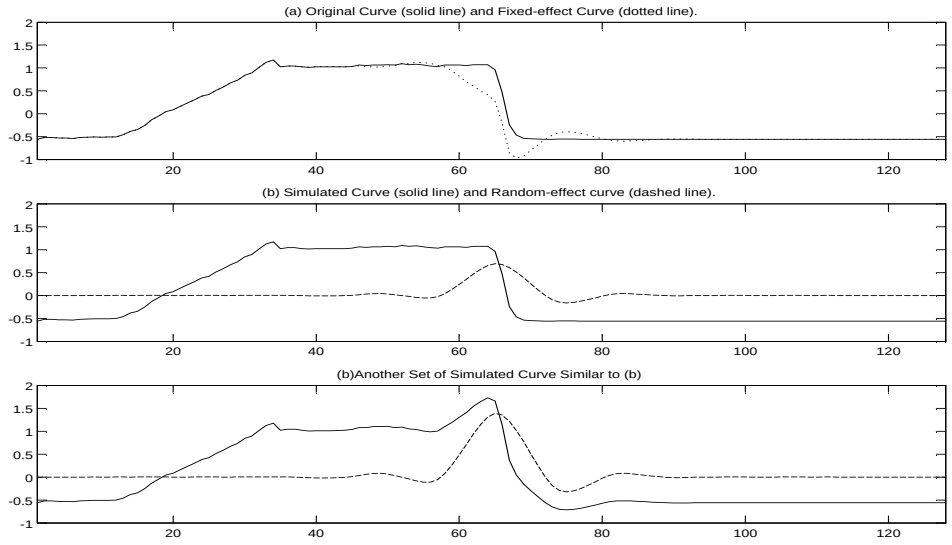


Figure 4: Impact of Random-effects

Define u_m as one if the wavelet coefficients at the mth wavelet-position is a random-effect; zero, otherwise, where $m = 1, 2, \ldots, N$. Here, m can be at any resolution level described in Eq. (1). Denote by $\boldsymbol{V} = [u_1, u_2, ..., u_N]^T$ a column vector. Define $\boldsymbol{U} = [\boldsymbol{V}^T, \ldots, \boldsymbol{V}^T]^T$ with $\boldsymbol{V}$ repeated M times as a column of $M \times N$ indicator variables for locating the coefficients that are random.

The distributions of the random coefficients R_m's are assumed to be independent normal with mean zero and variance ${\tau_m}^2$. Figure 5 in Example 3 shows an example that the assumption of normality is reasonable in the analysis of antenna data. Besides the "between-curve" variation from these random effects, the "within-curve" variation of wavelet coefficients is characterized by the common process variance σ^2 for all curves. Other than these two sources of variations, these M replicated curves have a common mean structure for all wavelet coefficients. Using the similar notation of $\boldsymbol{U}$, the mean column vector is defined as $\boldsymbol{\theta} = [{\boldsymbol{\theta}_{vN}}^T, \ldots, {\boldsymbol{\theta}_{vN}}^T]^T$, where ${\boldsymbol{\theta}_{vN}}^T$ is repeated M times and $\boldsymbol{\theta}_{vN} = [\theta_1, \theta_2, \ldots, \theta_N]^T$. Thus, the model of the wavelet coefficients $\boldsymbol{D}$ from M curves is as follows:

$$\boldsymbol{D} = \boldsymbol{\theta} + \boldsymbol{R} + \boldsymbol{Z}, \tag{4}$$

where $\boldsymbol{R} = [\boldsymbol{R}_{vN}{}^T, \ldots, \boldsymbol{R}_{vN}{}^T]^T$, $\boldsymbol{R}_{vN} = [R_1 \times u_1, \ldots, R_N \times u_N]^T$ and $\boldsymbol{Z}$ is a column of $M \times N$ random errors with the normal distribution $N(0, \sigma^2)$. Note that the indicators of the random-effects u_m's are involved in the vector $\boldsymbol{R}$. Based on this model, Figure 7 shows the comparison of the simulated multiple curves to the original data. Overall, the model captures key characteristics of variations in multiple curves. See the next example for details of deciding which coefficients are random.

Example 3. The formal research of deciding which coefficient is random in the wavelet-thresholding content (for data reduction and denoising purposes) is complicated and thus deferred to a future research. The following presents a simple idea of using the popular normal quantile-quantile plot (see Figure 5) to identify the random-effects, which exhibit excessive variance compared to the process variance σ^2 from random noises.

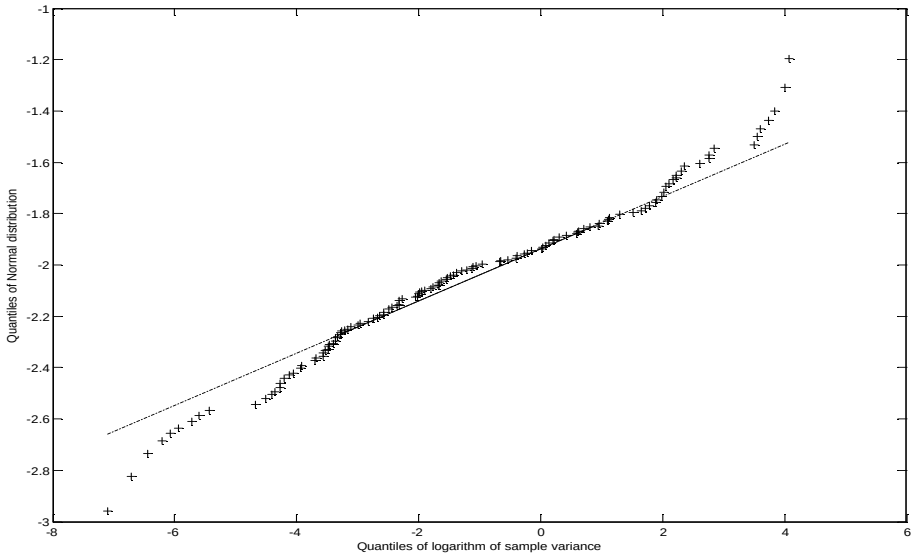


Figure 5: Normal Quantile-Quantile Plot of $\ln s_m^2$.

Our decision of random effects for $\boldsymbol{D}$ is rested on an implicit hypothesis of so called "factor sparsity" in the analysis of active effects without replicates (see Lenth (1989) and Box and Meyer (1986) for details and examples). That is, there will be only a few random-effects with a "significantly" large size of variance in the wavelet coefficients from replicated

data curves. Specifically, note that

$$(M-1)s_m^2 = \sum_{i=1}^{M}(d_{im} - \overline{d}_m)^2 \sim {\sigma_*}^2 \chi^2_{M-1}, \quad \text{and} \quad {\sigma_*}^2 = \sigma^2 + u_m {\tau_m}^2, \tag{5}$$

where χ^2_a is the chi-square distribution with a degree of freedom. Taking a logarithm to this sample variance yields $\ln s_m^2 = \ln {\sigma_*}^2 + \ln[\chi^2_{M-1}/(M-1)]$. The distribution of $\ln s_m^2$ can be approximated by $N(\ln {\sigma_*}^2, 2(M-1)^{-1})$. When the wavelet coefficient is a random effect, i.e., $u_m = 1$, the mean of s_m^2 will become $\sigma^2 + \tau_m^2$ which should be significantly larger than σ^2. Thus, when plot the quantiles of $\ln s_m^2$ against the quantiles of $N(\ln \sigma^2, 2(M-1)^{-1})$ in a normal probability plot, the random-effect coefficients will not be in a straight line. The variance σ^2 can be estimated by a pooled-variance using Donoho and Johnstone's (1994) robust estimate:

$$\hat{\sigma}^2 = M^{-1} \sum_{i=1}^{M} 0.6745^{-1} median(\mid d_{im} \mid : N/2+1 \le m \le N). \tag{6}$$

Figure 5 shows that there are several significantly larger-size variance terms. For example, the first seven in the upper-right corner of the plot show a clear departure of the straight line. The top 11 (or even 24) coefficients (in terms of their between-curve variance) could also be considered as random effects.

3.3 Antenna Data Example

Example 4. Consider only the first seven effects in the upper-right corner of Figure 5 as random wavelet coefficients. That is, the coefficients $c_{4,8}$, $c_{4,9}$, $c_{4,10}$, $c_{4,11}$, $c_{4,12}$, $d_{4,16}$, and $d_{5,32}$ are random. See Figure 7(a) for their wavelet bases. Most of them have overlapped support areas, and the coefficient $c_{L,k}$'s with father wavelets $\phi_{L,k}(t)$ have wider support than $d_{j,k}$'s with mother wavelets $\psi_{j,k}(t)$. Because all the support areas from these random-effects are only on the very left and very right sides of the antenna data, Figure 6(b) shows that when generating several realizations of random-effects, only the very left and very right sides of the antenna data curve have significant random fluctuations. Moreover, as experienced in Figure 3, this type of fluctuations is very different from the data noises. (See in the bottom two figures of Figure 3).

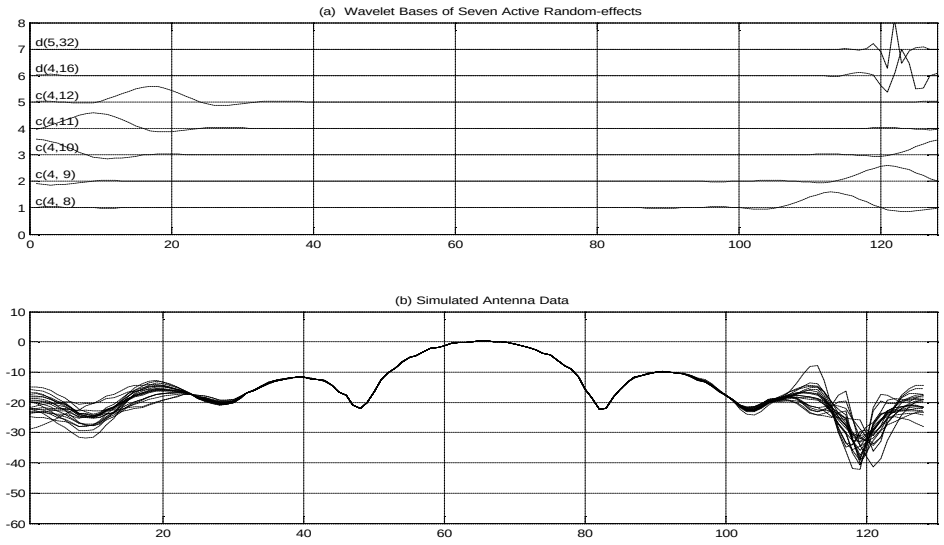


Figure 6: Wavelet Bases of Random-effects

Figure 7 shows different cases of simulated antenna data from the random-effect model with three different numbers (7, 11 and 24) of random-effects. Because the results look alike, we will use the case with seven random-effects in further studies for simplicity. Note that the simulated data curves based on the normality assumption of random-effects resemble the original data curves. Thus, we will stay with the normality assumption without any transformation. This assumption also makes the development of data-reduction procedures easier.

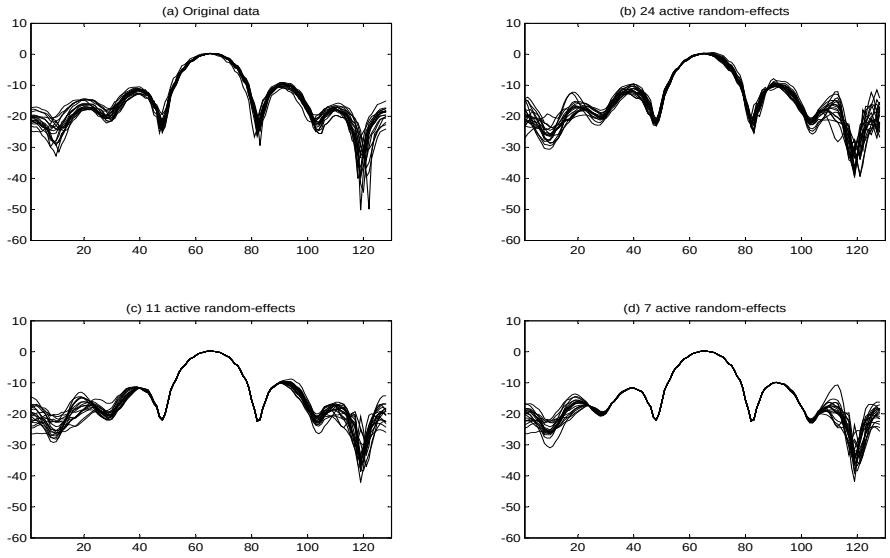


Figure 7: Simulated Multiple Sets of Data Curves

CHAPTER IV

VERTICAL ENERGY THRESHOLD(VET) WITHOUT CLASS INFORMATION

4.1 Introduction

Most of the work in the wavelet literature for handling multiple data curves are to apply data denoising (or reduction) procedures to each curve one at a time. Then, they use the union or intersection of wavelet coefficients selected from individual data curves to construct a "combined" representative set of coefficients for approximating the original data curves. See Lada *et al.* (2002) for such an example. Thus, their selections are based on the optimality criteria designed for a single data curve, not for multiple data curves. This chapter presents a criteria designed for multiple curves and its optimization details.

According to the previous literatures in data shrinkage methods, it was found that the reconstructed signals using a fewer number of wavelet coefficients provided a very reasonable approximation to the original data. In other words, the selected wavelet coefficients are rather representative in most of the data analysis. The newly proposed method in this proposal will also follow the principle that large magnitude wavelet coefficients (in their absolute value) will better characterize each signal patterns and retain more information. This principle will be expanded to that the large magnitude wavelet vertical energy will better characterize the whole set of signals and retain more information against the problem of heterogeneously selected wavelet coefficients for different single curves.

Figure 1 represents a kind of data in a form of curves. This data is from a project sponsored by Nortel and the NSF. The goal of the project was to develop timely product testing procedures to monitor antenna production quality, and to help trouble-shoot process imperfections. Nortel built a product functionality testing chamber to collect antenna signal patterns similar to data plotted in Figure 1. Although more detailed data could be obtained,

for the convenience of our data processing, for each of the 28 antennae studied we used at 128 elevation*150 azimuth grid to collect 19200 signal-amplitudes. The cusps and lobes of the azimuth cut of antenna data presented are difficult to handle by standard techniques such as polynomial regression or Fourier transform. Looking into the antenna data closely, we found that there is a certain systematic pattern useful in further data analysis. For example, although the data in azimuth cuts have many humps, data in the elevation cuts are rather smooth. It seems that there are several curves which are just vertical shifted from a curve with a certain pattern. However, unfortunately, even though almost all curves have a very similar systematic pattern, sets of thresholded wavelet coefficients for different curve are quite different each other after discrete wavelet transformation is applied to each curve.

In order to deal with heterogeneously selected wavelet coefficients for different single curves, we come up with several ideas. First, one can use the sum of wavelet coefficients at same position across all curves. However, in the case that many coefficients have different sign and same magnitude, the simple sum can not measure the importance of the wavelet positions. Secondly, the sum of absolute value of wavelet coefficients can be considered. Although this idea can successively measure the importance of a few wavelet positions, absolute sign causes the difficulty of deriving the distributional characteristics of meta-data(data of data). Thus, in this research, we introduce the wavelet vertical energy metric of multiple curves and utilize it for the efficient data reduction method.

4.2 Overall Relative Reconstruction Error(ORRE)

Inspired by the popularity of the scalogram (see Vidakovic (1999), page 289 for details), we develop the following vertical-energy based thresholding (VET) procedure. When a wavelet-position is selected, coefficients from all curves at this position will be selected. Learning from the hard-thresholding idea used in the procedures for single-curve based data reduction (see Jeong *et al.*, 2003), we propose the following minimization criteria,*Overall Relative Reconstruction Error;ORRE,*:

$$ORRE(\lambda) = \frac{\sum_{m=1}^{N} E[\|\boldsymbol{d}_{vm}(1 - I(\|\boldsymbol{d}_{vm}\|^2 > \lambda))\|^2]}{\sum_{m=1}^{N} E[\|\boldsymbol{d}_{vm}\|^2]} + \xi \cdot \frac{\sum_{m=1}^{N} E[I(\|\boldsymbol{d}_{vm}\|^2 > \lambda)]}{N}. \quad (7)$$

Note that the thresholding procedure described in the indicator function $I(\|\boldsymbol{d}_{vm}\|^2 > \lambda)$ is based on a "vertical energy" metric,

$$\|\boldsymbol{d}_{vm}\|^2 = {d_{1m}}^2 + {d_{2m}}^2 + \cdots + {d_{Mm}}^2, \quad m = 1, 2, \ldots, N, \tag{8}$$

which is the sum of all wavelet coefficients at the mth wavelet-position. This metric is similar to scalogram. However, in the scalogram applications, the coefficients are from the same wavelet-resolution level based on data in a single curve, not from different data curves. See Jeong, Chen and Lu (2003) for more details of the scalogram and its applications. Other metric such as the sum of absolute value of the coefficients could be used. However, after trying several choices we found that our vertical-energy is easier for deriving the optimum, the estimate of the thresholding parameter λ and its distribution properties. Thus, this article will not explore other methods.

The motivation of the criteria ORRE (overall relative reconstruction error) is from a simple idea of balancing the reconstruction error and the data-reduction ratio. The use of "normalizing constants" to make the two balancing terms compatible is critical. Note that all the data-denoising procedures (Donoho and Johnstone, 1994, 1995) do not have this type of normalization factors. Jeong *et al.* (2003) used empirical studies (in the single-curve situation) to show that without the normalization factors, the procedures studied were not effective for data-reduction purpose. This normalization idea is also motivated from many engineering applications (e.g., Mallat, 1998, pages 378-391), where the reconstruction error

$$RE = ||\boldsymbol{f} - \hat{\boldsymbol{f}}||/||\boldsymbol{f}||, \quad \text{where} \quad ||\boldsymbol{f}|| = (\sum_{i=1}^{N} f(t_i)^2)^{1/2},$$

is commonly used in comparing signal approximation quality. It characterizes the accuracy of the approximation to the original data. Thus, the first component of the objective function Eq. (7) represents a "normalized" reconstruction error from the approximated wavelet model $\boldsymbol{Y} = \boldsymbol{D}^*\boldsymbol{W}^T$. This article utilizes a thresholding parameter λ to decide which wavelet-domain data to keep and which to discard in decision-making analysis. Ideally, only a small portion of the data satisfying $\boldsymbol{d}_{vm} \cdot I(\|\boldsymbol{d}_{vm}\|^2 > \lambda))$ should be selected for meeting the data-reduction goal.

The second component of Eq. (7) is the normalized number of coefficients used. Note that there shall be an M factor in both numerator and denominator for the total number of coefficients considered from all M data curves. However, they cancel each other. In order to keep the representation or approximation model simple, this term acts as a penalty for avoiding the use of excessive number of coefficients. Similar penalty ideas have been used in ridge regression (Hastie *et al.*, 2001, page 59) and neural network (Hastie *et al.*, 2001, page 356). For example, the ridge regression finds the optimal choice of estimate for the regression coefficients by minimizing the following objective function:

$$\sum_{i=1}^{N}(y_i - \beta_0 - \beta_1 x_{i1} - ... - \beta_p x_{ip})^2 + \omega \sum_{j=1}^{p} {\beta_j}^2,$$

where ω is a weighting parameter like ξ in Eq. (7). Note that this objective function is not normalized as done in Eq. (7).

The weighting parameter ξ is user-selected or provided by methods such as generalized cross-validation (GCV) (Weyrich and Warhola, 1998). However, as experienced from Weyrich and Warhola (1998) further studies are needed for developing the GCV-like selection of ξ in our problem and understanding its properties. For simplicity, this article will use $\xi = 1$, which places equal weights on both components in the follow-up studies.

In the "follow-up analyses" (see Section 4.4 for examples), the selected coefficients are treated as the "reduced-size" data for various types of decision-making. Thus, we will deal with a small size of wavelet-domain data instead of the large size of the original time-domain data for saving computing time and storage space. The following section presents the analytical properties of the VET method. Section 4 conducts simulation comparisons of VET and single-curve based methods.

Remark: [1] In the objective function $ORRE(\lambda)$, the thresholding parameter λ is applied to all curves. When the traditionally used single-curve based data-denoising or -reduction method is used on multiple curves, the thresholding parameters could be different for distinct data curves. Even the idea of union or intersection could be used to select a common set of wavelet positions across all curves to reconstruct the data, but in that case the property of the thresholding parameters is changed. Thus, the properties of the resulted

ad hoc procedures are unknown, and analytical study of their properties is needed in future work.

[2] The main purpose of the ORRE-based data reduction method is to rigorously construct a representative "reduced-size" data in many types of further analysis. The following are a few examples. In developing statistical process control tools for monitoring complicated function data, multiple sets of data curves are needed in "Type-I" studies for establishing model parameters and so on. See Jeong and Lu (2004) for details. Sections 4.3 and 4.4 present classification and data visualization studies for several classes of multiple curves. Analysis of variance for complicated and large-size functional data could be conducted on the VET-selected wavelet coefficients, where their wavelet bases for all data curves are the same. Because our procedure is developed for general-purpose use, it can be improved for specific decision-making analysis. For example, one can add some kind of "class-separation" measures to ORRE for distinguishing curves in different classes. This will possibly improve the accuracy of classification for data classes. Further investigations are needed, and they are out of the scope of the studies in this thesis.

4.3 Optimal λ and Its Estimator

Solving the optimization problem with the objective function $ORRE(\lambda)$ requires proof of existence and uniqueness of the optimal solution. See Theorem 1 for details and Theorem 2 for an estimate of the thresholding parameter and its large-sample distribution properties. Proofs are left in Appendices.

Theorem 1. *Consider the model stated as $\boldsymbol{D} = \boldsymbol{\theta} + \boldsymbol{R} + \boldsymbol{Z}$. The objective function $ORRE(\lambda)$ is minimized uniquely at $\lambda = \lambda_{NM}$, where*

$$\lambda_{NM} = \sum_{m=1}^{N} E(\|\boldsymbol{d}_{vm}\|^2)/N = \sum_{m=1}^{N} M(\sigma^2 + \theta_m^2 + u_m\tau_m^2)/N = M\sigma^2 + M(\sum_{m=1}^{N} \theta_m^2 + u_m\tau_m^2)/N.$$

Since each wavelet coefficient is independent and distributed as normal, thus each wavelet-position's vertical energy follows a non-central chi-square distribution. Using this result and following some calculus derivations, the theorem is proved.

Proof of Theorem 1. Since $\|\boldsymbol{d}_{vm}\|^2$ is equal to $\sum_{i=1}^{M} d_{im}^2$, where d_{im}'s are independent and distributed as $N(\theta_m, u_m\tau_m^2+\sigma^2)$, it follows that $Y_m = \|\boldsymbol{d}_{vm}\|^2$ has a non-central chi-square distribution. That is, $Y_m \sim (\sigma^2+u_m\tau_m^2)\ \ \chi^2\left(M, M[\theta_m^2/(\sigma^2+u_m\tau_m^2)]\right)$, where $E[Y_m] = M(\sigma^2+\theta_m^2+u_m\tau_m^2)$ and $Var[Y_m] = 2M(\sigma^2+u_m\tau_m^2)(\sigma^2+u_m\tau_m^2+2\theta_m^2)$.

Denote by

$$\begin{aligned}\Lambda_m(\lambda) &= E[\|I(\|\boldsymbol{d}_{vm}\|^2<\lambda)\boldsymbol{d}_{vm}\|^2] = E[I(\|\boldsymbol{d}_{vm}\|^2<\lambda)\|\boldsymbol{d}_{vm}\|^2] \\ &= E[I(Y_m<\lambda)y_m] = \int_0^{\lambda} y_m f_m(y_m)dy_m,\end{aligned}$$

and

$$\begin{aligned}\Psi_m(\lambda) &= E[I(\|\boldsymbol{d}_{vm}\|^2>\lambda)] = Pr(\|\boldsymbol{d}_{vm}\|^2>\lambda) \\ &= Pr(Y_m>\lambda) = 1-\int_0^{\lambda} f_m(y_m)dy_m,\end{aligned}$$

where $f_m(y_m)$ is a noncentral chi-square density of Y_m. It follows that

$$\sum_{m=1}^{N} E[\|\boldsymbol{d}_{vm}(1-I(\|\boldsymbol{d}_{vm}\|^2>\lambda))\|^2] = \sum_{m=1}^{N} E[\|\boldsymbol{d}_{vm}\|^2 I(\|\boldsymbol{d}_{vm}\|^2<\lambda)] = \sum_{m=1}^{N}\Lambda_m(\lambda).$$

Then, $ORRE(\lambda)$ can be written as

$$\begin{aligned}ORRE(\lambda) &= \frac{\sum_{m=1}^{N} E[\|\boldsymbol{d}_{vm}(1-I(\|\boldsymbol{d}_{vm}\|^2>\lambda))\|^2]}{\sum_{m=1}^{N} E[\|\boldsymbol{d}_{vm}\|^2]} + \frac{\sum_{m=1}^{N} E[I(\|\boldsymbol{d}_{vm}\|^2>\lambda)]}{N} \\ &= \frac{\sum_{m=1}^{N}\Lambda_m(\lambda)}{\sum_{m=1}^{N} E[\|\boldsymbol{d}_{vm}\|^2]} + \frac{\sum_{m=1}^{N}\Psi_m(\lambda)}{N}.\end{aligned}$$

Because

$$\partial\Psi_m(\lambda)/\partial\lambda = \partial\left(1-\int_0^{\lambda} f_m(y_m)dy_m\right)/\partial\lambda = -f_m(y_m) < 0$$

and

$$\partial\Lambda_m(\lambda)/\partial\lambda = \partial\left(\int_0^{\lambda} y_m f_m(y_m)dy_m\right)/\partial\lambda = \lambda f_m(\lambda) = -\lambda\partial\Psi_m(\lambda)/\partial\lambda,$$

then

$$\begin{aligned}\frac{\partial ORRE(\lambda)}{\partial\lambda} &= -\lambda\left(\sum_{m=1}^{N}\frac{\partial\Psi_m(\lambda)}{\partial\lambda}\right)\left(\frac{1}{\sum_{m=1}^{N} E[\|\boldsymbol{d}_{vm}\|^2]}\right) + \frac{1}{N}\left(\sum_{m=1}^{N}\frac{\partial\Psi_m(\lambda)}{\partial\lambda}\right) \\ &= \left(-\frac{\lambda}{\sum_{m=1}^{N} E[\|\boldsymbol{d}_{vm}\|^2]} + \frac{1}{N}\right)\left(\sum_{m=1}^{N}\frac{\partial\Psi_m(\lambda)}{\partial\lambda}\right) = 0\end{aligned}$$

if and only if $\lambda = \sum_{m=1}^{N} E(\|\boldsymbol{d}_{vm}\|^2)/N$.

Theorem 2. *A simple and closed-form estimate of the thresholding parameter λ_{NM} is the following moment estimate,*

$$\widehat{\lambda}_{NM} = \sum_{m=1}^{N}\sum_{i=1}^{M} d_{im}^2/N.$$

Then, $\widehat{\lambda}_{NM}$ is a strongly consistent estimate of λ_{NM}, and its asymptotic distribution is $\sqrt{N}(\widehat{\lambda}_{NM}-\lambda_{NM})/\sigma_{NM}^* \xrightarrow{d} N(0,1)$, where $(\sigma_{NM}^*)^2 = 2M\sum_{m=1}^{N}(\sigma^2+u_m\tau_m^2)(\sigma^2+u_m\tau_m^2+2\theta_m^2)/N$.

The reason that we did not present the maximum likelihood estimate is due to the use of robust estimating procedure in Eq. (6) for the σ^2, instead of its MLE.

Proof of Theorem 2. Since DWT is orthonomal, $|\theta_m|,\ m = 1, 2, \ldots, N$, are uniformly bounded as $N \longrightarrow \infty$. Without loss of generality, we assume that $|\theta_m| < C_1,\ 0 < \tau_m^2 < C_2,\ m = 1, 2, \ldots, N$, where C_1 and C_2 constants. Therefore,

$$\begin{aligned}\sum_{m=1}^{\infty} Var(\sum_{i=1}^{M} d_{im}^2)/m^2 &= \sum_{m=1}^{\infty} Var(y_m)/m^2 = \sum_{m=1}^{\infty} 2M(\sigma^2+u_m\tau_m^2)(\sigma^2+u_m\tau_m^2+2\theta_m^2)/m^2 \\ &< \sum_{m=1}^{\infty} 2M(\sigma^2+C_2)(\sigma^2+C_2+2C_1)/m^2 < \infty, \quad \text{where } M < \infty.\end{aligned}$$

Thus, according to the Kolmogorov Theorem (Serfling, 1980, page 27), we conclude that

$$\left(\frac{1}{N}\sum_{m=1}^{N}(\sum_{i=1}^{M} d_{im}^2) - \frac{1}{N}\sum_{m=1}^{N} E(\|\boldsymbol{d}_{vm}\|^2)\right) \xrightarrow{w.p.1} 0.$$

That is, $\left(\widehat{\lambda}_{NM} - \lambda_{NM}\right) \xrightarrow{w.p.1} 0$

It is sufficient to verify the Lindeberg-Feller theorem (Serfling, 1980; page 29) for showing the asymptotic normality of $\widehat{\lambda}_{NM}$. Let Y_m be independent with means μ_m, finite variance σ_m^2 and distribution function F_m. Suppose that $B_N^2 = \sum_{m=1}^{N}\sigma_m^2$ satisfies $\sigma_N^2/B_N^2 \longrightarrow 0$, $B_N^2 \longrightarrow \infty$ as $N \longrightarrow \infty$. Then, according to the Lindeberg-Feller theorem,

$$\frac{1}{N}\sum_{m=1}^{N} Y_i \ \ is \ \ AN\left(\frac{1}{N}\sum_{m=1}^{N}\mu_m,\ \frac{1}{N^2}B_N^2\right)$$

if and only the Lindeberg condition,

$$\sum_{m=1}^{N}\frac{1}{B_N^2}\int_{|t-\mu_m|>\epsilon B_N}(t-\mu_m)^2 dF_m(t) \longrightarrow 0,\ N \longrightarrow \infty,\ \epsilon > 0,$$

is satisfied. In our problem, the notation can be interpreted as follow:

$$\widehat{\lambda}_{NM} = \sum_{m=1}^{N} Y_m/N, \quad Y_m \sim (\sigma^2 + u_m\tau_m^2)\ \chi^2(M, M\theta_m^2/(\sigma^2 + u_m\tau_m^2))$$

$$\lambda_{NM} = \sum_{m=1}^{N} \mu_m/N, \quad \mu_m = E(Y_m) = M(\sigma^2 + \theta_m^2 + u_m\tau_m^2)$$

$$(\sigma_{NM}^*)^2 = B_N^2/N = \sum_{m=1}^{N} \sigma_m^2/N, \quad \sigma_m^2 = Var(Y_m) = 2M(\sigma^2 + u_m\tau_m^2)(\sigma^2 + u_m\tau_m^2 + 2\theta_m^2).$$

Note that $B_n^2 = \sum_{m=1}^{N} \sigma_m^2 = N \cdot C_B$, where C_B is a finite constant. Then, the Lindeberg condition in our problem can be satisfied as follows:

$$\sum_{m=1}^{N} \frac{1}{N \cdot C_B} \int_{|t-\mu_m|>\epsilon\sqrt{N\cdot C_B}} (t-\mu_m)^2 dF_m(t) = \sum_{m=1}^{N} \frac{1}{N \cdot C_B}\ O\left(\int_{t>\mu_m+\epsilon\sqrt{N\cdot C_B}} (t-\mu_m)^2 dF_m(t)\right)$$

$$= \sum_{m=1}^{N} \frac{1}{N \cdot C_B}\ O\left(\int_{t\epsilon\sqrt{N\cdot C_B}} t^2 f_m(t)\ dt\right) = \sum_{m=1}^{N} \frac{1}{N \cdot C_B}\ O\left(\epsilon^2 \cdot N \cdot C_B \cdot f_m(\epsilon\sqrt{N \cdot C_B})\right) \longrightarrow 0.$$

Note that it is true that $f_m(\epsilon\sqrt{N \cdot C_B})$ converge to zero as N goes to infinity using the fact that $f_m(t)$ is a noncentral chi-square density function.

4.4 Illustrative Example

The popular testing data curve in Mallat (1998, page 378) is used for our simulation studies. Three models are considered. Model 1 has only one random effect set at the coarsest wavelet resolution level. Model 2 entertains two random effects, one at the coarsest level and the other at the second coarsest level. Model 3's random effects includes all coefficients at the coarsest level (denoted as "Case 1"; see Figure 8(a) for data), the second coarsest level (denoted as "Case 2"; see Figure 8(c) for data), or both of them (denoted as "Case 3"; see Figure 8(e) for data). The standard deviations (s.d.'s) τ_m of the random effects for Model 3 are all set at 4.5 and the s.d. σ for the noise error is set at 1. Other s.d.'s are explored as well. The conclusion is similar and thus skipped here. Each curve in Figure 8(b), (d), and (f) is an individual-curve example of each 10 simulated curves of (a), (c), and (e), respectively.

Although the data structures in Model 3 are much more complicated than the other two models, and the supports of these random effects are complicatedly overlapped, the

observations learned from these three models are alike. Thus, only the results for the most complicated model, Model 3, are presented here. Three methods are applied to the simulated data: VET and VisuShrink with union and VisuShrink with intersection, where VisuShrink is a commonly used single-curve based data denoising procedure developed in Donoho and Johnstorne (1994). Table 1 presents comparison results using the following measures: (1) Relative Error: RE= $\sum_i \|\boldsymbol{f}_i - \widehat{\boldsymbol{f}}_i\| / \sum_i \|\boldsymbol{f}_i\|$; (2) Reduction ratio (%): RR=(1 - Number of selected positions/N)$\times$ 100; (3) Overall Relative Reconstruction Error: $ORRE$ = (1) + (2). After 400 simulation runs, the average of each measure in all these three cases is reported. The standard deviations are all less than 0.0001 so that they are not presented in the Table 1. Results in Table 1 show that VET has about 95% data reduction ratio,

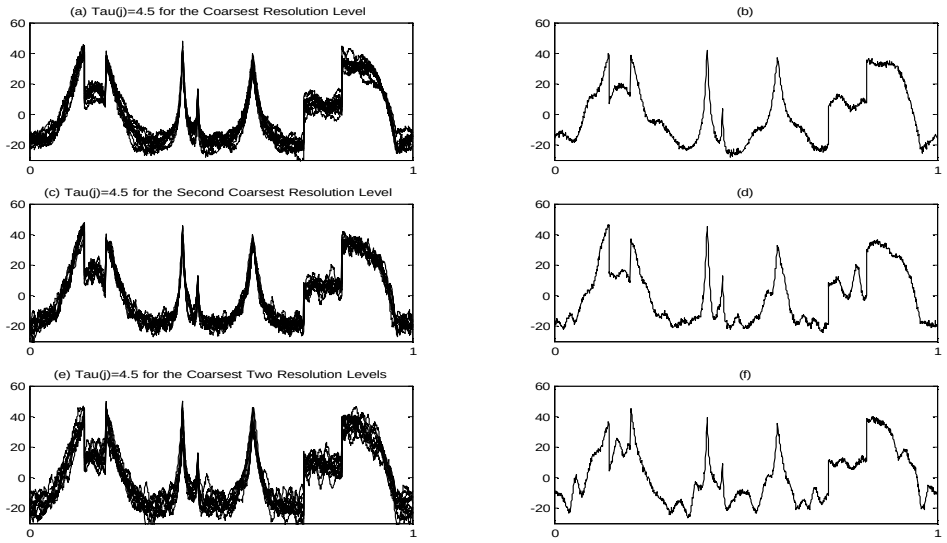


Figure 8: Simulated Data Curves from Model 3

which is quite aggressive. However, as seen from Figure 9, when using only 5% of the data, the original data curves can be constructed satisfactorily. Compared to the reconstruction curves from the other two methods, although VET has larger relative errors than the other methods, visually, the errors are reasonably small as seen in Figure 9. Compared in terms of ORRE, VET performs the best for all cases considered. VET has about 18.91 % (Case 3) to 28.36% (Case 1) smaller ORRE than the VisuShrink-intersection method, and 30.65 %

Table 1: Comparison of Data-reduction Methods

Method		Model 3		
		Case 1	Case 2	Case 3
VisuShrink (Union)	RE	0.0061	0.0064	0.0061
	RR	88.02%	86.9%	87.07%
	ORRE	0.1258	0.1374	0.1354
VisuShrink (Intersection)	RE	0.0064	0.0340	0.0366
	RR	91.44%	91.93%	92.08%
	ORRE	0.0920	0.1147	0.1158
VET	RE	0.0234	0.0375	0.0437
	RR	95.76%	94.76%	94.98%
	ORRE	0.0659	0.0899	0.0939

(Case 3) to 47.62 % (Case 1) smaller ORRE than the VisuShrink-union method. A simpler case such as Case 1 has larger difference. The VET method is applied to the real-life

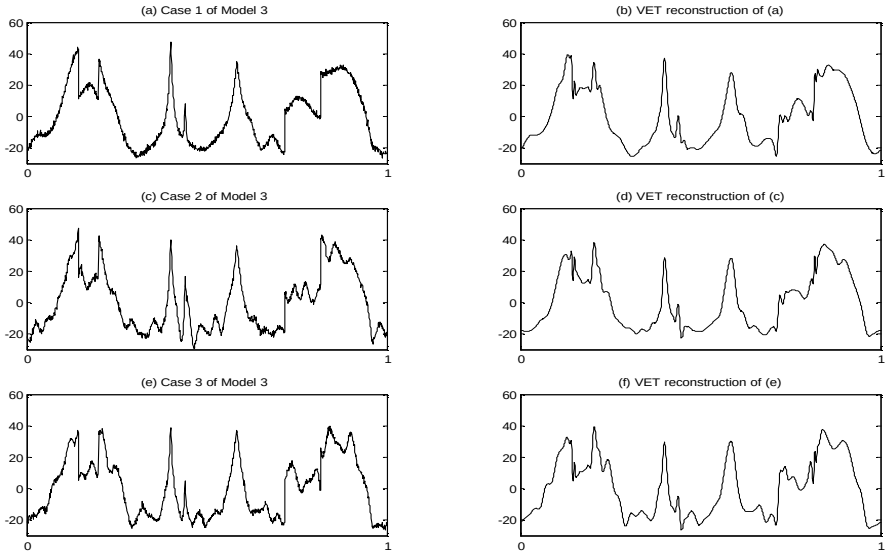


Figure 9: Reconstructed Data Curves

antenna data set consisting of 18 curves. Figure 10 presents the comparisons of the original antenna data and the reconstruction from the VET method. In total, 39 wavelet positions (the wavelet coefficients in the coarsest resolution level and the thresholded coefficients in other resolution levels) were used out of 128. The data reduction ratio is 69.53% and the

reconstruction looks very reasonable and captures the patterns in peaks and valleys. See Figure 10 for details.

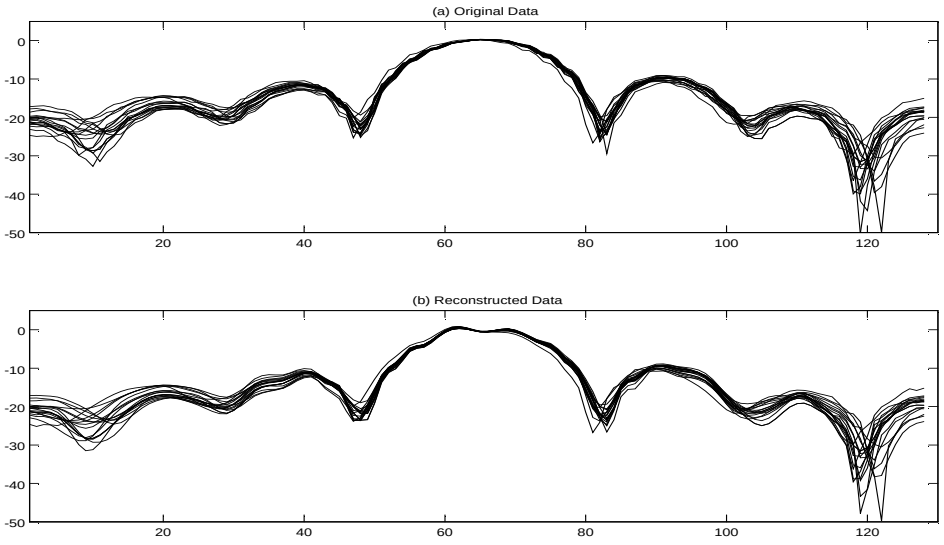


Figure 10: Reconstructed Antenna Data Curves

4.5 Decision based on the Reduced-size VET Data

When the data size is large, many standard software packages will face problems with limited working space and the data analysis time will be longer. Some of the iterative real-time data-exploration activities are not feasible in this situation. For example, in one of the studies in Jeong *et al.* (2003), typical commercial cluster analysis software cannot process 3,600 data curves (with 1,024 observations in each curve). This section uses examples to illustrate the potential of analyzing the reduced-size data obtained from the VET method in various decision-making applications.

Four different groups of process signals in Figure 11 are constructed based on certain modification of Mallat's data. Compared to the Nominal Class data given in Figures 8 and 9, Fault Class 1 has a different shape in the first peak around time point 180. Fault Class 2 has a only one jump level (instead of two) between time points 750 to 900. Fault Class 3 does not have the second peak in time points 380 to 450 and also has a smaller vertical

jump in time points 740 to 840. Each group has seven replicates (with 1,024 data points in each replicate), and there are 32 random effects assumed in the coarsest level of wavelet coefficients. The standard deviation τ of the random-effects is set at 20, and the standard deviation of process noises is 0.5.

Apply the VET method to the $7 \times 4 = 28$ data curves and select representative wavelet coefficients. Based on these selected coefficients, we will make simple plots to see if these four groups of data curves are very different. Moreover, we will conduct a hierarchical classification study to these coefficients for distinguishing curves in these four groups. Figure

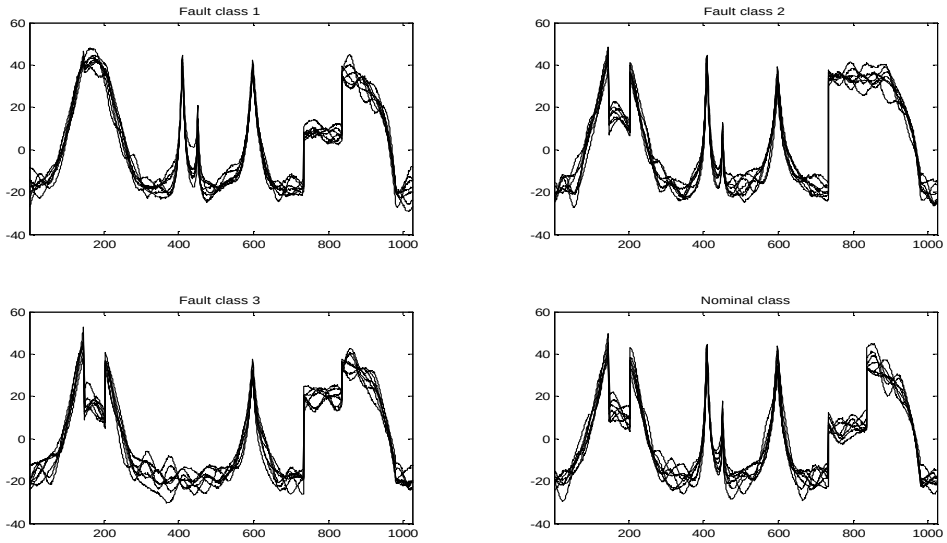


Figure 11: Four Groups of Simulated Data Curves

12 shows the selected wavelet-positions using the VET method for the "combined data set" from all four groups. The family of symmlet-8 is used in the discrete wavelet transformation for all data curves in this experiment. The vertical energy metrics for all 1,024 positions are displayed in six different panels representing wavelet's resolution levels. The thresholding value from VET is plotted in each panel, and the vertical energy larger than the threshold is marked in "triangle." It is found that all wavelet-positions (32 of them) in the coarsest level $c_{5,k}$ are selected. Then, seven out of 32 positions in the second coarsest level $d_{5,k}$, three out of 64 positions in the next level $d_{6,k}$, one out of 128 positions in $d_{7,k}$, and none

out of 256 positions in $d_{8,k}$ and none out of 512 positions in the finest level $d_{9,k}$ are selected, respectively. Thus, 43 out of 1,024 wavelet-positions are selected to reconstruct all four different groups of replicated data curves. This implies that the data reduction ratio is 95.8 %, i.e., only 4.2 % of the original wavelet domain data were used to reconstruct the original time domain data with a little loss of accuracy (see Figure 13 for its visual presentation). Importantly, the same wavelet-positions are used to reconstruct all four distinct groups of data curves even though the curves in different groups could be very different. This is important in further analysis such as wavelet-based functional analysis of variance studies. In such analysis, these curves will be compared using the same set of "reduced-size" data, which are the selected coefficients. To simplify the presentation and thus avoid many

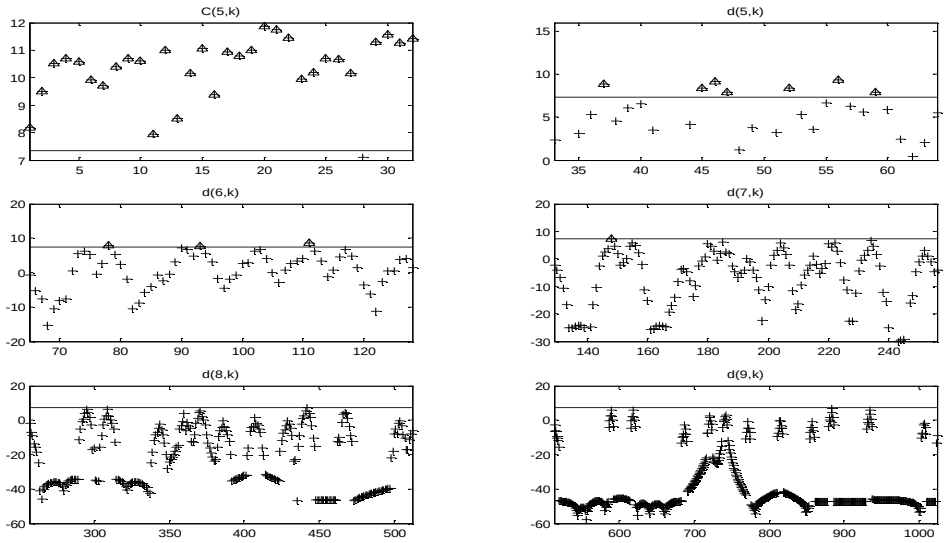


Figure 12: Vertical Energy at Each Resolution Level

replicated curves being placed in one plot, Figure 13 shows a pair of one representative original and reconstructed data curves for each group of curves in each row. Overall, these reconstructed curves capture the major differences among the four groups very well. In particular, the difference between Fault Class 2 and Nominal Class in panels (d) and (h) in the amount of vertical drop of the rectangle-shaped dip around 740 to 840 time points is captured well. The uniqueness of Fault Class 1 in panel (b) at the first dip around 150 to

Table 2: Elapsed Time for Hierarchical Clustering Analysis

	Elapsed Time (seconds)	
Number of Total Replicated Signals	VET Wavelet Domain	Original Time Domain
28	0.03	0.31
112	0.24	2.80
448	7.62	53.68
896	55.27	225.50

200 time points compared with the other groups is well depicted. Flat signal around 430 time point of Fault Class 3 in panel (f) is well distinguished from the other groups.

Many types of decision analysis can be applied to the VET-selected wavelet coefficients, here, we apply the VET-based reduced-size data for cluster analysis. *Hierarchical clustering* (see Duda, Hart, and Stork (2001) for its definition) is a way to investigate the grouping of a data set. The result is the construction of a hierarchy, which is a tree-like structure (*dendrogram*). In the structure, each data curve is treated as an *object* and is presented on the x-axis, and the other axis portrays the steps in the hierarchical procedure. Starting with each object represented as a separate class, the dendrogram graphically shows how the clusters are combined at each step of the procedure until all are contained in a single class. Figure 14 shows the details of its application to the above four groups of curves with 28 total replicates. The results from both wavelet-domain reduced-size data and the original time-domain data are exactly the same in the dendrogram plot. In fact, even as the number of replicates increases, e.g., 112, 448 or 896, the dendrograms from both wavelet and time domains are the same. Moreover, both achieve the perfect clustering: the identified class membership from the dendrogram is the same as the class assignment used to generate the data curves. Table 2 reports an experiment of elapse-time calculation for processing the wavelet-domain reduced-size data and the original time-domain data using the dendrogram. Matlab's commands, Tic(at starting) and Toc(at ending), provide a convenient way to calculate the elapsed time. In our experiment, the above random-effect model is used to generate different numbers of total replicates, e.g., 28, 112, 448, and 896, with equal size of replicates for each of the four groups. The experiments were implemented

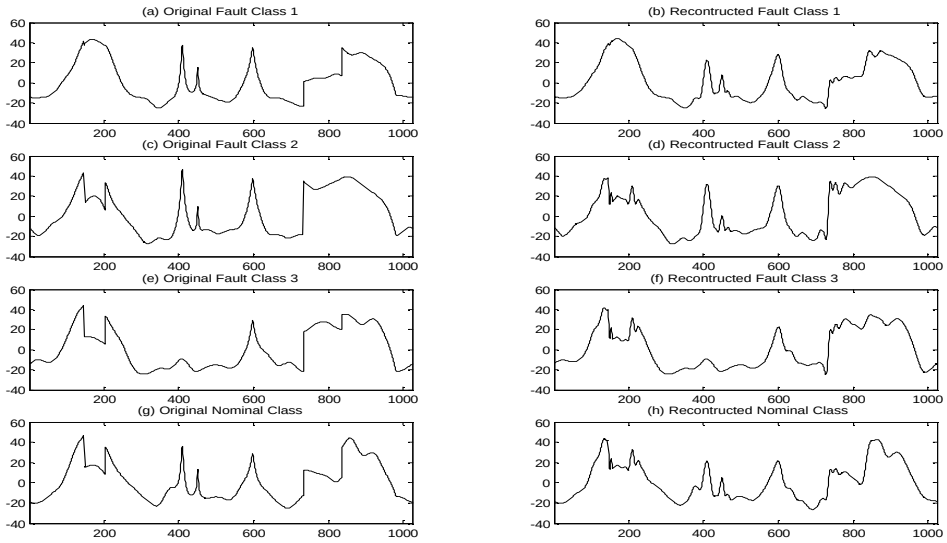


Figure 13: Reconstructed Data Curves

in MATLAB with Intel Pentium-III 996 MHz processor. When small number of replicates were used for clustering analysis, the difference of computing time was negligible. However, as the number of replicates increased, the differences becomes more significant. For the total of 896 replicates, the processing time in the original time-domain data was five times larger than the reduced-size data.

Figures 15 (a) and (b) plot a few selected wavelet coefficients pairwisely. Depending on the choice of the coefficients, the separation between the replicates in the four groups is different. Figure 15(c) plots the scalogram of all selected coefficients at the coarsest two resolution levels. The scalogram plot has the best separation among the three cases presented here. Since our current research focuses on one class of data curves, future research is needed to select a VET-like metric considering class separability for increasing the separation among groups of replications.

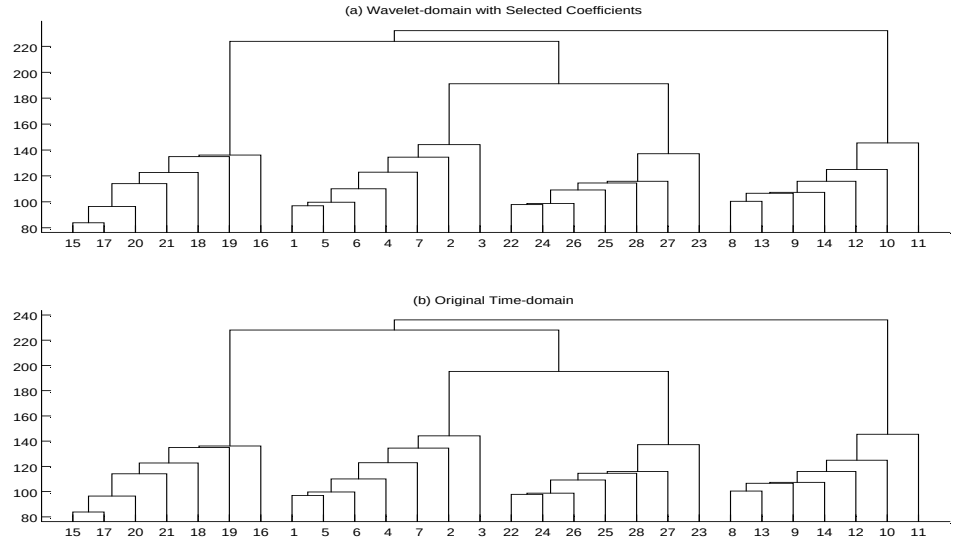


Figure 14: Hierarchical Clustering by the Dendrogram

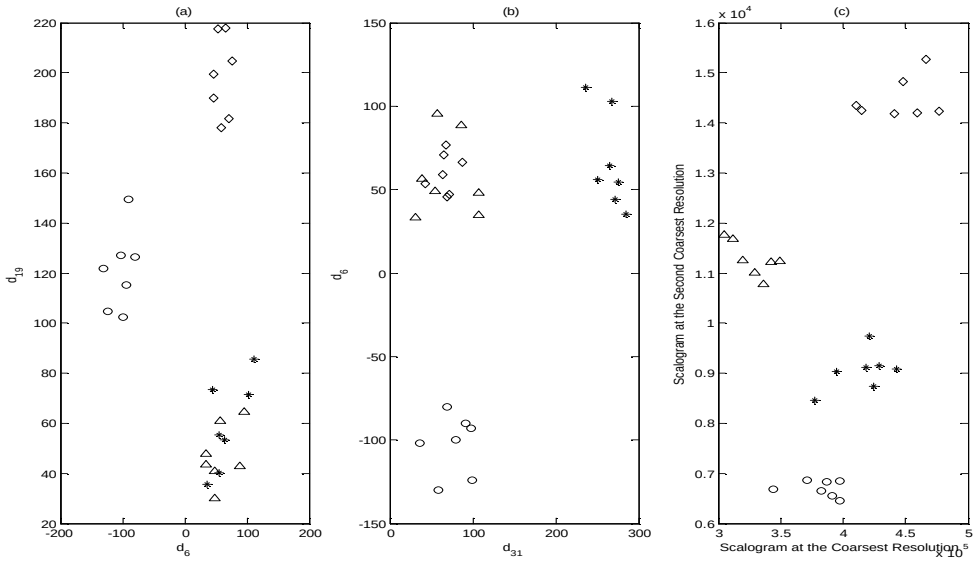


Figure 15: Clusters of Data Curves

CHAPTER V

VERTICAL GROUP-WISE THRESHOLD(VGWT) WITH CLASS INFORMATION

5.1 Introduction

It is conjectured that VET(Vertical Energy Threshold) method must be very efficient especially when a set of curves are homogeneous in their pattern. Cluster analysis provides us the class information of each curves which indicates that the curves in same class probably have same or similar pattern even though there are a little noise. Thus, as long as we have the class information of each curve, VET can be applied to each class successfully. In this section, we would like to explore some further problems of applying VET to the process containing multi-classes in its output.

As indicated in the previous section, the number of wavelet positions used to represent a class can be different for different classes from the same process. Moreover, even if the number of selected wavelet coefficients are same, the wavelet positions can be different. Our challenge is to decide on a set of wavelet positions to represent adequately the overall data structure of the process and perform further data mining analysis, there are a number of different selection strategies based on VET such as union, intersection, and voting strategy.

There have not been any research result about how these different strategies affect the efficiency of representing the data structure and how to measure the efficiency itself. The motivation of exploring this aspect is based on the assumption that there might be the most suitable selection strategy depending on different process in a sense of statistical inference. It can be said that the most suitable selection strategy can be regarded as an intrinsic parameter of the system. If we can clarify the most suitable strategy for different processes, it will enable us to perform more justified data reduction procedure in real practice. We will explore the mathematically rigorous best strategy selection scheme using objective

function-based comparison.

Also, an expansion of this research is directed to combine the class separability concept with our key components of data reduction goal. It was motivated that our reduced size data can maximize the ability to make different classes further distinguished (separated) each other in terms of distance of each class means. This combination, called *Vertical Group-Wise Threshold(VGWT)* method, focuses the ability to retain salient differences between classes. The absolute value of each class means at wavelet positions were taken into account to compare with a common threshold λ in order to reduce the size of data. Use of the absolute value of each class mean at wavelet positions was reasonable approach since class means at wavelet positions are good representation of different classes at each wavelet positions and the absolute value can successively measure the importance of wavelet positions. In order to achieve several purposes such as signal reconstruction accuracy, data reduction efficiency and retaining the class separability as much as possible in the reduced-size data, the guideline to get the optimal threshold is proposed.

5.2 VET for Data with Class Information

5.2.1 Selection Strategies (Union, Intersection and Voting)

For simplicity of notations, we newly define d_{ijk}, a wavelet coefficient at jth position of ith curve in kth class, and $\boldsymbol{d}_{vj}^{k} = (d_{1jk}, d_{2jk}, \ldots, d_{ijk}, \ldots, d_{M_k jk})^{\top}$ as *wavelet vertical vector of class* k where $k = 1, 2, \ldots, K$ is an index for classes, $j = 1, 2, \ldots, N$ for wavelet positions, and $i = 1, 2, \ldots, M_k$ for curves in class k. We use x_{jk} instead of $\|\boldsymbol{d}_{vj}^{k}\|^2$ without any confusion and $\boldsymbol{X}_j$ for a set of all x_{jk} at position j.

$$x_{jk} = \|\boldsymbol{d}_{vj}^{k}\|^2, \quad k = 1, 2, \ldots, K, \ j = 1, 2, \ldots, N$$

$$\boldsymbol{X}_j = \{\boldsymbol{x}_{j1}, \boldsymbol{x}_{j2}, \ldots, x_{jk}\} \quad j = 1, 2, \ldots, N$$

We will use an indicator variable of several selection strategies such as Union, Intersection, and Voting denoted by $\Lambda_{union}(\boldsymbol{X}_j)$, $\Lambda_{intersept}(\boldsymbol{X}_j)$, and $\Lambda_{voting}(\boldsymbol{X}_j)$ respectively. See Figure 16 and 17 to understand better the direct use of VET for several classes and choices of several selection strategies.

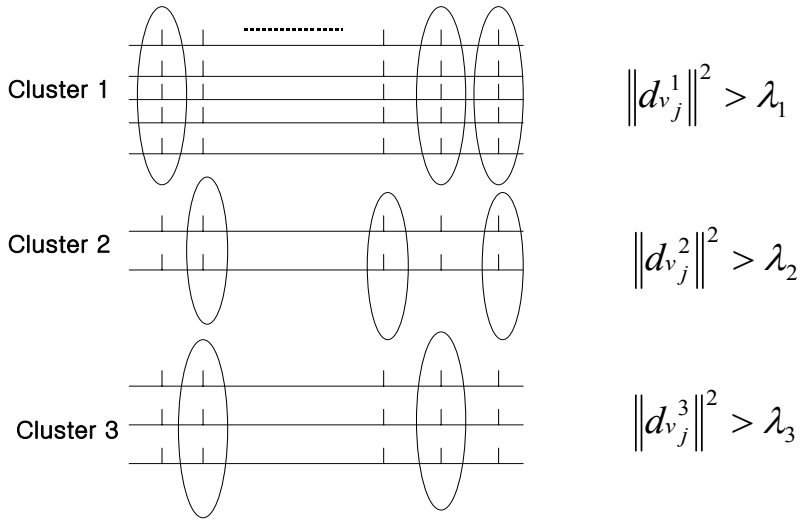


Figure 16: Direct use of Vertical energy threshold.

One can use the union set of all wavelet positions selected by VET method for the all class data sets. This approach gives a comprehensive selection of the representative coefficients that covers many of the data fluctuations across all classes and captures the most important features of each classes. However the number of wavelet positions in the union set can be very large against efficient data reduction. The indicator variable for Union strategy will be obtained like below.

$$\begin{aligned}
\Lambda_{union}(\boldsymbol{X}_j) &= max(I(\boldsymbol{x}_{j1} > \lambda_1), I(\boldsymbol{x}_{j2} > \lambda_2), \cdots, I(x_{jk} > \lambda_K)) \\
&= \coprod_{k=1}^{K} I(x_{jk} > \lambda_k) \\
&= 1 - (1 - I(\boldsymbol{x}_{j1} > \lambda_1))(1 - I(\boldsymbol{x}_{j2} > \lambda_2)) \cdots (1 - I(x_{jk} > \lambda_K))
\end{aligned}$$

An alternative approach is to select the intersection set of all wavelet positions that are selected by the VET method on every class. This strategy will keep the number of selected wavelet positions smaller than any other strategies. The weakness of this method is that the more detailed data patterns for each class can be ignored so that it will make the data

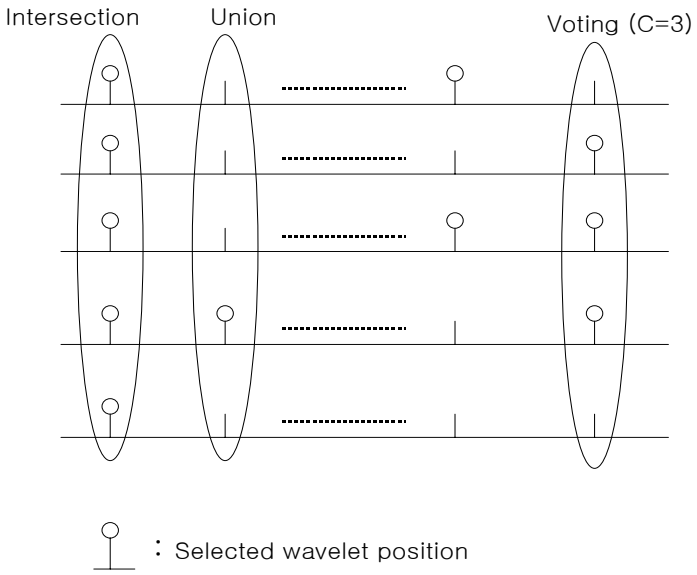


Figure 17: Example of several selection strategies.

model approximation overly smoothed. The indicator variable for Intersection strategy will be obtained like below.

$$\begin{aligned}
\Lambda_{intersection}(\boldsymbol{X}_j) &= min(I(\boldsymbol{x}_{j1} > \lambda_1), I(\boldsymbol{x}_{j2} > \lambda_2), \cdots, I(x_{jk} > \lambda_K)) \\
&= \prod_{k=1}^{K} I(x_{jk} > \lambda_k) \\
&= I(\boldsymbol{x}_{j1} > \lambda_1) I(\boldsymbol{x}_{j2} > \lambda_2) \cdots I(x_{jk} > \lambda_K)
\end{aligned}$$

The above two strategies can be generalized to the other strategy, which is called Voting strategy. This approach is to select the set of all wavelet positions that are selected by the VET method on at least a certain number of classes, say C classes. If C equals one, then it means Union strategy, and C equals K (the number of all classes), it means Intersection strategy. This idea could lead to a theoretically justified procedure for fining optimal value for C. However, it is still under development and its performance remains to be studied.

The indicator variable for Voting strategy with constant C will be obtained like below.

$$\Lambda_{voting}(\boldsymbol{X}_j) = \begin{cases} 1 & \text{if } \sum_{k=1}^{K} I(x_{jk} > \lambda_k) > C \\ 0 & \text{otherwise.} \end{cases} \tag{9}$$

In the next section, The objective function for best selection strategy problem will be formulated based on the indicator variables introduced in this section. It turns out that the least voting strategy constant C is the key of strategy selection and an optimal C for a certain system should be numerically obtained from previous data.

5.2.2 Direct Use of VET to Individual Class

In Chapter 4, we obtained the optimal threshold value for each class such as λ_k where k is an index for classes. The objective function-based best strategy selection problem addressed in the previous section can be simplified to the problem of finding optimal C of the following objective function

$$\begin{aligned} ORRE_{voting} &= \frac{E\left[\sum_{j=1}^{N}\sum_{k=1}^{K}\|\boldsymbol{d}_{vj}^{k}(1-\Lambda_{voting}(\boldsymbol{X}_j))\|^2\right]}{E\left[\sum_{j=1}^{N}\sum_{k=1}^{K}\|\boldsymbol{d}_{vj}^{k}\|^2\right]} \\ &\quad + \frac{E\left[\sum_{j=1}^{N}\Lambda_{voting}(\boldsymbol{X}_j)\right]}{N} \\ &= \frac{E\left[\sum_{j=1}^{N}\left\{\sum_{k=1}^{K}\|\boldsymbol{d}_{vj}^{k}\|^2\right\}(1-\Lambda_{voting}(\boldsymbol{X}_j))\right]}{E\left[\sum_{j=1}^{N}\sum_{k=1}^{K}\|\boldsymbol{d}_{vj}^{k}\|^2\right]} \\ &\quad + \frac{E\left[\sum_{j=1}^{N}\Lambda_{voting}(\boldsymbol{X}_j)\right]}{N} \\ &= \frac{E\left[\sum_{j=1}^{N}\left\{\sum_{k=1}^{K}x_{jk}\right\}(1-I(\sum_{k=1}^{K}I(x_{jk}>\lambda_k)>C))\right]}{E\left[\sum_{j=1}^{N}\sum_{k=1}^{K}x_{jk}\right]} \\ &\quad + \frac{E\left[\sum_{j=1}^{N}I(\sum_{k=1}^{K}I(x_{jk}>\lambda_k)>C)\right]}{N} \end{aligned}$$

The first component of above objective function represents a "normalized" reconstruction error from the approximated wavelet model structured by selected wavelet coefficients using VET and the least voting strategy constant C. The term $\boldsymbol{d}_{vj}^{k} \cdot \Lambda_{voting}(\boldsymbol{X}_j)$ is the

shrunken vertical vector of wavelet coefficients for class k by hard thresholding. The second component is the normalized number of coefficients used. A constant multiplier α that can be added to the second term may be considered to control the trade-off between the two terms as mentioned before.

Particularly, when C is equal to 1 or K, the above objective function equals to the followings respectively.

When $C = 1$

$$ORRE_{union} = \frac{E\left[\sum_{j=1}^{N}\left\{\sum_{k=1}^{K} x_{jk}\right\}\prod_{k=1}^{K}(1 - I(x_{jk} > \lambda_k))\right]}{E\left[\sum_{j=1}^{N}\sum_{k=1}^{K} x_{jk}\right]} + \frac{E\left[\sum_{j=1}^{N}\prod_{k=1}^{K}(1 - I(x_{jk} > \lambda_k))\right]}{N}$$

When $C = K$

$$ORRE_{intersection} = \frac{E\left[\sum_{j=1}^{N}\left\{\sum_{k=1}^{K} x_{jk}\right\}(1 - \prod_{k=1}^{K} I(x_{jk} > \lambda_k))\right]}{E\left[\sum_{j=1}^{N}\sum_{k=1}^{K} x_{jk}\right]} + \frac{E\left[\sum_{j=1}^{N}\prod_{k=1}^{K} I(x_{jk} > \lambda_k)\right]}{N}$$

In the voting strategy, optimal selection of constant C might be possible through numerical search technique since closed form solution does not exist.

REMARKS 5.2.1. *:*

1. The λ_k we obtained is optimal for class k itself. However, when we deal with several classes all together with objective function, we can not assure that $\{\lambda_1, \cdots, \lambda_K\}$ is optimal for objective function with several classes. There might be the optimal set of thresholds $\{\lambda_1^*, \cdots, \lambda_K^*\}$ for several classes.

2. In order to find the optimal set of thresholds $\{\lambda_1^*, \cdots, \lambda_K^*\}$, we might need to use some numerical search methods such as Newton-Raphson method using $\{\lambda_1, \cdots, \lambda_K\}$ as the starting point. However, we don't hereby dig out the solution for this problem, but clarify the questions we confront.

5.3 Vertical Group-Wise Threshold(VGWT) with Between-Class Separability

5.3.1 Class Separability with Threshold Rule

Generally speaking, classification performance depends on four factors : class separability, the training sample size, dimensionality, and classifier type(or discriminant function). To improve classification performance, attention is often focused on seeking improvements on the factors other than class separability because class separability is usually considered inherent and predetermined. The objective of this section is to call attention to the fact that class separability can be increased by careful selection of reduced sized data as compensation of possible loss of data reduction efficiency.

The most important property of a classification system is its ability to find the most informative features describing the objects that are classified, because it guarantees as compact decision rules as possible. In order to design a simple and efficient classification and segmentation scheme one has to select features that are most effective in showing the salient differences between the signals. This selection may or may not be appropriate for other tasks such that approximation or compression. In other words the selection must give the best minimal set of features in terms of the separability of signal classes in the feature space. Examples of quantitative measures of class separability are Bayes error, Bhattacharya distance, divergence based or variational distribution distances and scatter matrix based measures.

Unlike Mean Square Error(MSE) which is the most widely used criterion for signal representation, class separability measures are typically invariant under any non-singular, linear or non-linear, transformation. However any non-singular mapping used for dimensionality reduction results in losing some of classification information. Our objective is to find the mapping that provides the maximum class-separability for a given range of acceptable reduction in space dimension satisfying comparatively accurate signal approximation as well. In other words we are searching among all possible transformations for the best subspace which preserves class-separations as much as possible in the lowest possible dimensional space fulfilling good signal reconstruction .

A simplified and yet elegant way of formulating criteria of class separability is based on within and between class scatter matrices which are used widely in discriminant analysis of statistics. Usually within-class, between-class, and mixture scatter matrices are used to formulate the criteria of class separability. A within-class scatter matrix of the input vectors X for L classes is expressed by

$$SSW_X = \sum_{i=1}^{L} P_i E[(X - m_i)(X - m_i)^T | w_i] = \sum_{i=1}^{L} P_i \Sigma_i$$

where $P_i (Pr(w = w_i))$ means the prior probability of class i, m_i is the conditional mean vector and Σ_i is the conditional covariance matrix. A between class scatter matrix is expressed as

$$SSB_X = \sum_{i=1}^{L} P_i (m_i - m_0)(m_i - m_0)^T = \sum_{i=1}^{L-1} \sum_{j=i+1}^{L} P_i P_j (m_i - m_j)(m_i - m_j)^T$$

where m_0 is the overall mean vector. The optimal features are determined by optimizing the Fisher criteria given by

$$tr(SSW_X^{-1} SSB_X) \quad or \quad tr(SSB_X)/tr(SSW_X)$$

However, this criteria will be calculated by a single set of samples from a certain distribution of signal. So the more nature-oriented criteria of class-separability would be

$$J_X = E[tr(SSW_X^{-1} SSB_X)] \quad or \quad E[tr(SSB_X)]/E[tr(SSW_X)]$$

. So we are seeking a transformation A from R^n to R^m, $X \subset R^n \longrightarrow^A Y \subset R_m$, with $m < n$ such that A optimizes J_Y, i.e. minimizes the drop in cost $J_X - J_Y$ incurred by the maximum reduction in the feature space dimensionality satisfying the high accuracy of signal representation using Y.

In order to define the class separability measure in our context, we define d_{ijk}, a wavelet coefficient at jth position of ith curve in kth class where $k = 1, 2, \ldots, K$, $j = 1, 2, \ldots, N$, and $i = 1, 2, \ldots, n_k$. Denoted by $\overline{d}_{.jk}$ the average of d_{ijk}'s over all $i = 1, \cdots, n_k$ curves in the kth class, and $\overline{d}_{.j.}$ the average over all curves from all classes at jth wavelet atom position. Let us take a given input $\boldsymbol{d}_{h11}, \cdots, \boldsymbol{d}_{hn_K K}$ where $\boldsymbol{d}_{hik} = \{d_{i1k}, d_{i2k}, \cdots, d_{iNk}\}$ $(i = 1, \cdots, n_k$ and $k = 1, \cdots, K)$ is N-dimensional row vector (horizontal vector), i.e. N is the total number

of wavelet atom positions. Let the input matrix $\boldsymbol{D}_{n\times N} = \{\boldsymbol{d}_{h11}, \cdots, \boldsymbol{d}_{hn_K K}\}^T$ be formed by the row vectors $\boldsymbol{d}_{hik}$ where $n = n_1 + \cdots + n_K$. The inputs should be classified into classes k $(k = 1, \cdots, K)$ which possess a priori probabilities P_k and the cardinality of the classes is n_k. P_k can be estimated by $\frac{n_k}{n}$. Let $\boldsymbol{D'}_{n\times N'} = \{\boldsymbol{d'}_{h11}, \cdots, \boldsymbol{d'}_{hn_K K}\}^T$ be generated by a thresholding rule (feature selection technique) from $\boldsymbol{D}$, where $\boldsymbol{d'}_{hik}$ are N'-dimensional row vector. $(N' < N)$. $\boldsymbol{D'}$ is generated by deleting some $N - N'$ columns of $\boldsymbol{D}$. A criterion to measure the class separability is defined as $J(\boldsymbol{D'}) = \frac{E[tr(SSB_{\boldsymbol{D'}})]}{E[tr(SSW_{\boldsymbol{D'}})]}$ where $SSB_{\boldsymbol{D'}}$ is the between class scatter matrix and $SSW_{\boldsymbol{D'}}$ is the within class scatter matrix of $\boldsymbol{D'}$ which is transformed by a certain thresholding rule such as $\boldsymbol{D'} = \Lambda(\boldsymbol{D}, \lambda)$.

At this stage, as we discussed in the previous section, there are several options of the thresholding function to decide $N - N'$ columns of $\boldsymbol{D}$(e.g. $N - N'$ numbers of j wavelet atom positions out of N) to be deleted for the reduction of feature space. Here we decide to use the intersection concept using class mean $\overline{d}_{.jk}$. Thus, when $\boldsymbol{d}_{vj} = \{d_{1j1}, \cdots, d_{n_K jK}\}^T$ is defined as n-dimensional column vector (vertical vector) in the input matrix $\boldsymbol{D}$, the decision whether jth wavelet atom position in the input matrix $\boldsymbol{D}$ should be deleted or not can be made by a thresholding function such as

$$\boldsymbol{d'}_{vj} := \Lambda(\boldsymbol{d}_{vj}, \lambda) = \begin{cases} \boldsymbol{d}_{vj} & \text{if} \quad \Lambda_{intersection}(\boldsymbol{d}_{vj}, \lambda) = 1 \\ \text{zero vector} & \text{o/w} \end{cases}$$

using the indicator function $\Lambda_{intersection}(\boldsymbol{d}_{vj}, \lambda)$ at jth position such as

$$\begin{aligned} \Lambda_{intersection}(\boldsymbol{d}_{vj}, \lambda) &= min(I(|\overline{d}_{.j1}| > \lambda), I(|\overline{d}_{.j2}| > \lambda), \cdots, I(|\overline{d}_{.jK}| > \lambda)) \\ &= \prod_{k=1}^{K} I(|\overline{d}_{.jk}| > \lambda) \\ &= I(|\overline{d}_{.j1}| > \lambda) I(|\overline{d}_{.j2}| > \lambda) \cdots I(|\overline{d}_{.jK}| > \lambda) \end{aligned}$$

where

$$I(|\overline{d}_{.jk}| > \lambda) := \begin{cases} 1 & \text{if} \quad |\overline{d}_{.jk}| > \lambda \\ 0 & \text{o/w} \end{cases}$$

Then the class separbility measure $J(\boldsymbol{D}')$ in our context becomes

$$\begin{aligned} J(\boldsymbol{D}') &= J(\boldsymbol{D}, \lambda) \\ &= \frac{E[tr(SSB_{\boldsymbol{D}'})]}{E[tr(SSW_{\boldsymbol{D}'})]} \\ &= \frac{E[tr(SSB_{\boldsymbol{D},\lambda})]}{E[tr(SSW_{\boldsymbol{D},\lambda})]} \\ &= \frac{E\left[\sum_{j=1}^{N} \quad (\quad \sum_{k=1}^{K} n_k(\overline{d}_{.jk} - \overline{d}_{.j.})^2 \cdot \prod_{k=1}^{K} I(|\overline{d}_{.jk}| > \lambda) \quad) \quad \right]}{E\left[\sum_{j=1}^{N} \quad (\quad \sum_{k=1}^{K} (\sum_{i=1}^{n_k} (d_{ijk} - \overline{d}_{.jk})^2) \cdot \prod_{k=1}^{K} I(|\overline{d}_{.jk}| > \lambda) \quad) \quad \right]} \end{aligned}$$

The use of the class mean $\overline{d}_{.jk}$ in the indicator function was considered making sense that it can be well used for several reasons. First, one may consider some assumptions for ease of computation of $J(\boldsymbol{D}')$ since $E[tr(SSB_{\boldsymbol{D}'})]$ is composed of linear combination of $\overline{d}_{.jk}$. Second, since we have several classes of curves, it is our main interest to study how the each class mean curve looks like. Then, the class mean can represent well each class so that 'reconstruction error' term can be replaced with 'class mean reconstruction error' for modelling accuracy concept. The practical implementation of this idea will be conducted in the following section.

We can take advantage of computation with further assumption which can leads us to mainly focus on the between class separability of the thresholded data. When all curves are classified to several class, small difference of curves in magnitude is usually experienced to decide that those curves are in same class. That is, we may assume that 'within class variance' is quite small compared to 'between class variance'. (exceptional case may be possible though.) From the above assumption, the denominator of class separability term can also be modified. Under the assumption of very small 'within class variance', applying the thresholding rule to 'within class variance' of the original data does not make big change in magnitude. Thus, we can get rid of the indicator variable for the denominator of $J(\boldsymbol{D}')$, then 'within class variance' term becomes a constant which is nothing to do with the thresholding rule. Finally, we can only pay attention to 'between class variance' term to increase the class separability measure. Once we have reduced the size of data, it is important to realize how much portion of the original between class separability can be reflected in the reduced size data's between class separability. So the ratio of $E[tr(SSB_{\boldsymbol{D}'})]$

to the $E[tr(SSB_{\boldsymbol{D}})]$ will be main interest and can be used as the measure of the efficiency in terms of the class separability. The new measure of class separability of the reduced-size data, called *BCSR(Between Class Separability Ratio)*, is that

$$\begin{aligned} BCSR(\lambda) &= \frac{E[tr(SSB_{\boldsymbol{D}'})]}{E[tr(SSB_{\boldsymbol{D}})]} \\ &= \frac{E[tr(SSB_{\boldsymbol{D}}(\lambda))]}{E[tr(SSB_{\boldsymbol{D}})]} \\ &= \frac{E\left[\sum_{j=1}^{N} \quad \{ \quad \sum_{k=1}^{K} n_k(\overline{d}_{.jk} - \overline{d}_{.j.})^2 \cdot \prod_{k=1}^{K} I(|\overline{d}_{.jk}| > \lambda) \quad \} \quad \right]}{E\left[\sum_{j=1}^{N} \sum_{k=1}^{K} n_k(\overline{d}_{.jk} - \overline{d}_{.j.})^2 \quad \right]} \end{aligned}$$

5.3.2 Two-Stage Procedure

We propose a "two-stage procedure" to achieve our several purposes. As mentioned before, the informative features can be different according to different tasks so that the optimal feature selection method also varies for different purposes as well. If one just wants to maximize the ability to show the salient differences between signals, the number of informative features(important wavelet atom positions in our context) could be very small. However, we here still want to keep the important features for high level of accuracy in signal approximation.

So we apply ORRE (Overall Relative Reconstruction Error) concept, like in the previous chapter, in our new circumstance (data knowledge with class membership information, and different thresholding function) for balancing the accuracy of the signal class mean reconstruction and high performance in data reduction in the first stage. More specifically, our strategy here is that we first find optimal λ_0 in the thresholding function for balancing low level of the class-mean-reconstruction-error(CMRE) and large reduction in data size that can be measured as the used-data-ratio(UDR). In order to balance these two objectives CMRE and UDR, the criterion ORRE,which is a weighted sum of these two measure, will be used like in the previous chapter.

Then, in the second stage, we decide a certain upper limit of ORRE according to the engineering knowledge and historical experience so that we can get a certain range of possible λ's. It will guarantee that any λ in that range satisfies the acceptable level of ORRE. In the

range of λ's, we study the behavior of our class separability measure(BCSR; Between Class Separability Ratio), then we find the optimal λ^* which maximize the BCSR in the range of λ's. In certain case, it would be hard to find the upper limit of acceptable ORRE when one does not have the engineering knowledge and historical experience. Even in this case, we proposed a guideline of how to decide the upper limit(U) of ORRE. It will be discussed in the chapter later.

5.3.3 ORRE-driven Optimal λ_0

Similar to the optimization problem in the Chapter 4.2, we here still use the criterion ORRE which is the weighted sum of two components; "signal reconstruction error" and "data reduction efficiency".

Since we have several classes of multiple curves, we become more interested in the signal class mean so that we use the signal-class-mean-reconstruction-error-ratio (CMRE) for signal reconstruction error component.

$$CMRE(\lambda) = \frac{E\left[\sum_{j=1}^{N}\sum_{k=1}^{K} n_k(\overline{d}_{.jk} - \overline{d}_{.jk}\cdot\prod_{k=1}^{K} I(|\overline{d}_{.jk}| > \lambda))^2\right]}{E\left[\sum_{j=1}^{N}\sum_{k=1}^{K} n_k\overline{d}_{.jk}^2\right]}$$

The nominator represents the signal-class-mean-reconstruction error which is the mean square of difference between before and after thresholding. The denominator is "normalizing constant" which characterizes the accuracy of the approximation to the original signal class mean data, which is same to the nominator in the case that none of j wavelet atom positions is deleted. Thus, CMRE represents a "normalized" reconstruction error from the approximated signal class mean data after the thresholding rule is applied with a certain λ. Ideally, a small value of this component explains good approximation performance.

The second component of ORRE is the used-data-ratio(UDR).

$$UDR(\lambda) = E\left[\frac{\sum_{j=1}^{N}\prod_{k=1}^{K} I(|\overline{d}_{.jk}| > \lambda)}{N}\right] \quad = \quad \frac{1}{N}\sum_{j=1}^{N} E\left[\prod_{k=1}^{K} I(|\overline{d}_{.jk}| > \lambda)\right]$$

It is the normalized number of coefficients used. Like the ORRE in the previous chapter, note that there shall be an $n = n_1 + n_2 + \cdots + n_K$ factor in both numerator and denominator for the total number of coefficients considered from all n data curves. However, they cancel each other.

In order to estimate the expectation term(E) of both CMRE and UDR in the maximum likelihood manner, considering the invariance property of MLE, we can use the meta-function $R_{jk}(\lambda)$ with the assumption of normal distribution of $\overline{d}_{.jk}$ with mean μ_{jk} and variance σ^2_{jk}, such as

$$\begin{aligned} R_{jk}(\lambda) &= E[I(|\bar{d}_{.jk}| > \lambda)] \\ &= 1 - \Phi\left(\frac{\lambda_{jk} - \mu_{jk}}{\sigma_{jk}}\right) + \Phi\left(\frac{-\lambda_{jk} - \mu_{jk}}{\sigma_{jk}}\right) \end{aligned}$$

where two common functions $\phi(y)$ and $\Phi(y)$ are defined as

$$\begin{aligned} \phi(y) &= \frac{1}{\sqrt{2\pi}} \exp(-0.5y^2) \\ \Phi(y) &= \int_{-\infty}^{y} \phi(z)dz \end{aligned}$$

Details are explained and proved in the appendix 1.

Using the meta-function above, the CMRE and UDR can be expressed as follows

$$CMRE(\lambda) = \left(\sum_{j=1}^{N}\{1 - \prod_{k=1}^{K} R_{jk}(\lambda)\}\sum_{k=1}^{K} n_k(\mu^2_{jk} + \sigma^2_{jk})\right) / \left(\sum_{j=1}^{N}\sum_{k=1}^{K} n_k(\mu^2_{jk} + \sigma^2_{jk})\right)$$

$$UDR(\lambda) = \sum_{j=1}^{N}\prod_{k=1}^{K} R_{jk}(\lambda)/N$$

Details for this derivation is proved in the appendix 3.

Due to the complexity of CMRE and UDR functions, it is impossible to get the closed form solution of the optimal λ_0 which minimize the $ORRE(\lambda)$. From the derivation of CMRE and UDR, the $ORRE(\lambda) = CMRE(\lambda) + UDR(\lambda)$ can be computable in the most computational software, such as Matlab, so that we can search the optimal λ_0 which minimize the $ORRE(\lambda)$ in iterative procedure such as Golden-section search. More brute force method to find the root of the function is a simple piece-wise plotting using acceptably small interval.

5.3.4 Optimal λ^* with Known U

Here we consider the class separability term with respect to λ. The between-class-separability measure with respect to λ, $E[tr(SSB(\lambda))]$, is

$$E[tr(SSB(\lambda))] = E\left[\sum_{j=1}^{N}\sum_{k=1}^{K} n_k(\overline{d}_{.jk} - \overline{d}_{.j.})^2 \cdot \prod_{k=1}^{K} I(|\overline{d}_{.jk}| > \lambda)\right]$$

and the original between-class-separability measure is

$$E[tr(SSB)] = E\left[\sum_{j=1}^{N}\sum_{k=1}^{K} n_k(\overline{d}_{.jk} - \overline{d}_{.j.})^2\right]$$

Like the previous section, in order to estimate the expectation term(E) of both $E[tr(SSB(\lambda))]$ and $E[tr(SSB)]$ in the maximum likelihood manner, considering the invariance property of MLE, we can use the two other meta-functions $P_{jk}(\lambda)$, and $Q_{jk}(\lambda)$ with the assumption of normal distribution of $\overline{d}_{.jk}$ with mean μ_{jk} and variance σ^2_{jk}, such as

$$\begin{aligned}
P_{jk}(\lambda) &= E[\bar{d}^2_{.jk} I(|\bar{d}_{.jk}| > \lambda)] \\
&= \left\{1 - \Phi\left(\frac{\lambda_{jk} - \mu_{jk}}{\sigma_{jk}}\right) + \Phi\left(\frac{-\lambda_{jk} - \mu_{jk}}{\sigma_{jk}}\right)\right\}(\sigma^2_{jk} + \mu^2_{jk}) \\
&\quad -2\sigma_{jk}\left\{\phi\left(\frac{-\lambda_{jk} - \mu_{jk}}{\sigma_{jk}}\right) - \phi\left(\frac{\lambda_{jk} - \mu_{jk}}{\sigma_{jk}}\right)\right\} \\
&\quad -\sigma^2_{jk}\left\{\left(\frac{-\lambda_{jk} - \mu_{jk}}{\sigma_{jk}}\right)\cdot\phi\left(\frac{-\lambda_{jk} - \mu_{jk}}{\sigma_{jk}}\right) - \left(\frac{\lambda_{jk} - \mu_{jk}}{\sigma_{jk}}\right)\cdot\phi\left(\frac{\lambda_{jk} - \mu_{jk}}{\sigma_{jk}}\right)\right\}
\end{aligned}$$

and

$$\begin{aligned}
Q_{jk}(\lambda) &= E[\bar{d}_{.jk} I(|\bar{d}_{.jk}| > \lambda)] \\
&= \left\{1 - \Phi\left(\frac{\lambda_{jk} - \mu_{jk}}{\sigma_{jk}}\right) + \Phi\left(\frac{-\lambda_{jk} - \mu_{jk}}{\sigma_{jk}}\right)\right\}\cdot\mu_{jk} \\
&\quad -\sigma_{jk}\left\{\phi\left(\frac{-\lambda_{jk} - \mu_{jk}}{\sigma_{jk}}\right) - \phi\left(\frac{\lambda_{jk} - \mu_{jk}}{\sigma_{jk}}\right)\right\}
\end{aligned}$$

Then $E[tr(SSB(\lambda))]$ and $E[tr(SSB)]$ can be computed as below.

$$E[tr(SSB(\lambda))] = \sum_{j=1}^{N}\sum_{k=1}^{K} n_k(1 - \frac{n_k}{n_T})P_{jk}(\lambda)\prod_{r\neq k} R_{jk}(\lambda) - \frac{2}{n_T}\sum_{j=1}^{N}\sum_{a\neq b} Q_{ja}(\lambda)Q_{jb}(\lambda)\prod_{r\neq a,b} R_{jr}(\lambda)$$

$$E[tr(SSB)] = \sum_{j=1}^{N}\sum_{k=1}^{K} n_k(1 - \frac{n_k}{n_T})(\sigma^2_{jk} + \mu^2_{jk}) - \frac{2}{n_T}(\sum_{j=1}^{N}\sum_{a\neq b} n_a n_b \mu_{ja}\mu_{jb})$$

Details are explained in the appendix 1 and 2. Then our class separability measure $BCSR(\lambda)$ will be defined as

$$BCSR(\lambda) = \frac{E[tr(SSB(\lambda))]}{E[tr(SSB)]}$$

Then, when we have the engineering knowledge and historical experience so that a certain upper limit of ORRE, U, can be decided, the problem of finding optimal λ^* in the range of possible λ's can be formulated as

$$\begin{aligned} \max_{\lambda} \quad & E[tr(SSB(\lambda))] \\ s.t. \quad & ORRE(\lambda) \leq U \end{aligned}$$

The objective function is $E[tr(SSB(\lambda))]$ instead of $BCSR(\lambda)$ because the only nominator $E[tr(SSB(\lambda))]$ of $BCSR(\lambda)$ is related to λ. So it is actually same problem of maximizing $BCSR(\lambda)$ with respect to λ.

The case that we do not have the engineering knowledge and historical experience for deciding U will be covered in Section 5.3.5.

The solution of the constraint optimization problem above is

$$\begin{aligned} \text{If} \quad & U < \min ORRE(\lambda), && \text{No feasible solution} \\ \text{else if} \quad & U = \min ORRE(\lambda), && \lambda^* = ORRE^{-1}(U) \\ \text{else if} \quad & U > \min ORRE(\lambda), && \lambda^* = \{\lambda; \quad \lambda = ORRE^{-1}(U), \quad \lambda < \lambda_0\} \\ & && \quad = \min ORRE^{-1}(U) \end{aligned}$$

The constraint $ORRE(\lambda) \leq U$ results in the range of λ such as $\{\lambda; min\ ORRE^{-1}(U) \leq \lambda \leq max\ ORRE^{-1}(U), U \geq min\ ORRE(\lambda)\}$. Also $E[tr(SSB(\lambda))]$ is a monotonically decreasing function of λ (*see Lemma 5.3.1.*). So the optimal λ^* will be the smallest value in the range of all possible λ's. Then the proof of the optimal solution set is done.

LEMMA 5.3.1. $E[tr(SSB(\lambda))]$ *is a monotonically decreasing function of* λ

For the proof of Lemma 5.3.1, see the appendix 4.

Maximum Likelihood Estimator of Optimal λ

The theoretically optimal solution of λ^* is proved as $min\ ORRE^{-1}(U; \mu_{jk}, \sigma_{jk})$. Then the MLE of the optimal λ can be obtained , according to the Invariance Property (*see appendix 5*), as

$$\hat{\lambda} = min\ ORRE^{-1}(U; \hat{\mu}_{jk}, \hat{\sigma}_{jk})$$

where, with the fact that the distribution of $\bar{d}_{.jk}$ is assumed as $N(\mu_{jk}, \sigma^2_{jk})$, $\hat{\mu}_{jk}$ and $\hat{\sigma}_{jk}$ are the MLE of μ_{jk} and σ_{jk} respectively.

$$\begin{aligned} \hat{\mu}_{jk} &= \sum_{i=1}^{n_k} d_{ijk}/n_k \\ \hat{\sigma}^2_{jk} &= \sum_{i=1}^{n_k} (d_{ijk} - \hat{\mu}_{jk})^2/n_k^2 \end{aligned}$$

Then we need to solve the equation $ORRE(\hat{\lambda}) = U$. Like most practical application, one can not simply solve the equation in closed form. Instead one has to use iterative methods and one of the most famous one is Newton-Rapson. We wish to apply Taylors Approximation in solving $h(\lambda) = ORRE(\lambda) - U = 0$. Suppose λ_0 is a "good guess" of the solution to $h(\lambda) = 0$. Then

$$h(\hat{\lambda}) = 0 \approx h(\lambda_0) + h'(\lambda_0)(\hat{\lambda} - \lambda_0)$$

Solving the last "equation" gives

$$\hat{\lambda} \approx \lambda_0 - \frac{h(\lambda_0)}{h'(\lambda_0)}$$

In this special practice we do have a "good guess" λ_0 with 0 because we are looking for the minimum λ which make the equation equals zero and it tends to be a lot smaller than the possible largest λ. The formula suggests the following iterative scheme which is known as the Newton-Rapson algorithm. Start with some initial value λ_0 and then calculate successively

$$\begin{aligned} \lambda_1 &\approx \lambda_0 - \frac{h(\lambda_0)}{h'(\lambda_0)} \\ \lambda_2 &\approx \lambda_1 - \frac{h(\lambda_1)}{h'(\lambda_1)} \\ \lambda_3 &\approx \cdots \end{aligned}$$

In most (but certainly not all) problems arising in practice there will be very little change in the λ-values after a few iterations. One can then stop and the final value is then taken to be the maximum likelihood estimate.

Differentiating the $h(\lambda)$ would not give a simple analytical form due to the complexity of function itself such as

$$\begin{aligned} h(\lambda;\hat{\mu}_{jk},\hat{\sigma}_{jk}) &= ORRE(\lambda) - U \\ &= CMRE(\lambda) + DRR(\lambda) - U \\ &= \left(\sum_{j=1}^{N}\{1-\prod_{k=1}^{K}\hat{R}_{jk}(\lambda)\}\sum_{k=1}^{K} n_k(\hat{\mu}_{jk}^2+\hat{\sigma}_{jk}^2)\right) / \left(\sum_{j=1}^{N}\sum_{k=1}^{K} n_k(\hat{\mu}_{jk}^2+\hat{\sigma}_{jk}^2)\right) \\ &\quad + \sum_{j=1}^{N}\prod_{k=1}^{K}\hat{R}_{jk}(\lambda)/N - U \end{aligned}$$

where

$$\hat{R}_{jk}(\lambda) = 1 - \Phi\left(\frac{\lambda-\hat{\mu}_{jk}}{\hat{\sigma}_{jk}}\right) + \Phi\left(\frac{-\lambda-\hat{\mu}_{jk}}{\hat{\sigma}_{jk}}\right)$$

However, it is computationally possible to differentiate $h(\lambda)$ using Secant method(see appendix 6).

Asymptotic Property of Maximum Likelihood Estimator of Optimal λ

Let $\boldsymbol{\theta}$ be a $2nK \times 1$ parameter vector,

$$\begin{aligned} \boldsymbol{\theta} &= (\mu_{11},\cdots,\mu_{nK},\sigma_{11}^2,\cdots,\sigma_{nK}^2)^T \\ &= (\theta_1,\theta_2,\cdots,\theta_{2nK})^T \end{aligned}$$

, with maximum likelihood estimator $\hat{\boldsymbol{\theta}}$ such that

$$\hat{\boldsymbol{\theta}} \sim N(\boldsymbol{\theta}, Var(\hat{\boldsymbol{\theta}}))$$

where $Var(\hat{\boldsymbol{\theta}})$ is a covariance matrix of $\hat{\boldsymbol{\theta}}$ and its element can, due to the assumption of independence among wavelet atom positions and among classes, be understood as

$$\begin{aligned} Var(\hat{\mu}_{jk}) &= \sigma_{jk}^2/n_k, \quad Var(\hat{\sigma}_{jk}^2) = 2(n_k-1)\sigma_{jk}^4/n_k^2, \\ Cov(\hat{\mu}_{jk},\hat{\mu}_{j'k'}) &= Cov(\hat{\sigma}_{jk}^2,\hat{\sigma}_{j'k'}^2) = Cov(\hat{\mu}_{jk},\hat{\sigma}_{jk}) = Cov(\hat{\mu}_{jk},\hat{\sigma}_{j'k'}) = 0 \end{aligned}$$

where $j \neq j'$ and $k \neq k'$. Estimating a nonlinear function $g(U;\boldsymbol{\theta})$ and its asymptotic distribution can be obtained using a general method called the Delta method explained below.

Suppose $g(U;\boldsymbol{\theta})$ is a nonlinear continuous function of $\boldsymbol{\theta}$, in our context, $g(U;\boldsymbol{\theta}) = min\ ORRE^{-1}(U;\boldsymbol{\theta})$. Expending in Taylor series about the true value $\boldsymbol{\theta}$,

$$\begin{aligned} g(U;\hat{\boldsymbol{\theta}}) &= g(U;\boldsymbol{\theta}) + g'(U;\boldsymbol{\theta})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) + o(\parallel \hat{\boldsymbol{\theta}} - \boldsymbol{\theta} \parallel) \\ &= g(U;\boldsymbol{\theta}) + \sum_{j=1}^{2nK} \frac{\partial g}{\partial \theta_j}(\hat{\theta}_j - \theta_j) + o(\parallel \hat{\boldsymbol{\theta}} - \boldsymbol{\theta} \parallel) \end{aligned}$$

where we have defined

$$g'(U;\boldsymbol{\theta}) = \frac{\partial g}{\partial \boldsymbol{\theta}} = (\frac{\partial g}{\partial \theta_1}, \frac{\partial g}{\partial \theta_2}, \cdots, \frac{\partial g}{\partial \theta_{2nK}})^T$$

evaluated at $\boldsymbol{\theta}$. As the optimal λ is equal to $g(U;\boldsymbol{\theta})$,

$$\begin{aligned} \frac{\partial g(U;\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} &= \frac{\partial \lambda}{\partial \boldsymbol{\theta}} = \frac{\partial U}{\partial \boldsymbol{\theta}} \left(\frac{\partial U}{\partial \lambda}\right)^{-1} \\ &= \frac{\rho_1(\lambda,\boldsymbol{\theta})}{\rho_2(\lambda,\boldsymbol{\theta})} = \frac{\rho_1(g(U;\boldsymbol{\theta}),\boldsymbol{\theta})}{\rho_2(g(U;\boldsymbol{\theta}),\boldsymbol{\theta})} \end{aligned}$$

where $\rho_1(\lambda,\boldsymbol{\theta}) = \frac{\partial U}{\partial \boldsymbol{\theta}}$ and $\rho_2(\lambda,\boldsymbol{\theta}) = \frac{\partial U}{\partial \lambda}$. That is

$$g'(U;\boldsymbol{\theta}) = \frac{\rho_1(g(U;\boldsymbol{\theta}),\boldsymbol{\theta})}{\rho_2(g(U;\boldsymbol{\theta}),\boldsymbol{\theta})}$$

The vector g' has dimension $2nK \times 1$. Rearranging terms,

$$g(U;\hat{\boldsymbol{\theta}}) - g(U;\boldsymbol{\theta}) = g'(U;\boldsymbol{\theta})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) + o(\parallel \hat{\boldsymbol{\theta}} - \boldsymbol{\theta} \parallel)$$

We will estimate $g(U;\theta)$ by $g(U;\hat{\boldsymbol{\theta}})$. If $\hat{\boldsymbol{\theta}}$ is a maximum likelihood estimate, then $g(U;\hat{\boldsymbol{\theta}})$ is the MLE of $g(U;\boldsymbol{\theta})$. Taking the variance of both sides,

$$Var(g(U;\hat{\boldsymbol{\theta}}) - g(U;\boldsymbol{\theta})) = g'(U;\boldsymbol{\theta})^T Var(\hat{\boldsymbol{\theta}}) g'(U;\boldsymbol{\theta})$$

where $Var(\hat{\boldsymbol{\theta}})$ is defined above. This equation is the heart of the delta method, so one will write it out again as a scalar equation. Let g'_i be the i-th element of $g'(U;\boldsymbol{\theta})$,and let v_{ij} be the ij-element of the matrix $Var(\hat{\boldsymbol{\theta}})$. Then the variance of $g(U;\hat{\boldsymbol{\theta}})$ is

$$Var(g(U;\hat{\boldsymbol{\theta}})) = \sum_{i=1}^{2nK} \sum_{j=1}^{2nK} g'_i g'_j v_{ij}$$

In large samples and under regularity conditions, $g(U;\hat{\boldsymbol{\theta}})$ will converge to $g(U;\boldsymbol{\theta})$, and

$$g(U;\hat{\boldsymbol{\theta}}) \sim N\left(g(U;\boldsymbol{\theta}), \quad \left[\frac{\rho_1(g(U;\boldsymbol{\theta}),\boldsymbol{\theta})}{\rho_2(g(U;\boldsymbol{\theta}),\boldsymbol{\theta})}\right]^T Var(\hat{\boldsymbol{\theta}}) \left[\frac{\rho_1(g(U;\boldsymbol{\theta}),\boldsymbol{\theta})}{\rho_2(g(U;\boldsymbol{\theta}),\boldsymbol{\theta})}\right]\right)$$

Using λ notation,

$$\hat{\lambda} \sim N\left(\lambda, \quad \left[\frac{\rho_1(g(U;\boldsymbol{\theta}),\boldsymbol{\theta})}{\rho_2(g(U;\boldsymbol{\theta}),\boldsymbol{\theta})}\right]^T Var(\hat{\boldsymbol{\theta}}) \left[\frac{\rho_1(g(U;\boldsymbol{\theta}),\boldsymbol{\theta})}{\rho_2(g(U;\boldsymbol{\theta}),\boldsymbol{\theta})}\right]\right)$$

In practice, all derivatives and true $\boldsymbol{\theta}$ are evaluated at $\hat{\boldsymbol{\theta}}$. The implementation of all derivatives can be done using Secant method.

5.3.5 Guideline for the Selection of U

When we do not have the given U, the upper limit of acceptable ORRE, without the engineering knowledge and historical experience, we confront another problem of deciding reasonable value of U. That is, how to set up the acceptable upper limit of ORRE. We propose a guideline of how to set it up in this section. Referring Figure 18 will help one to understand every notation.

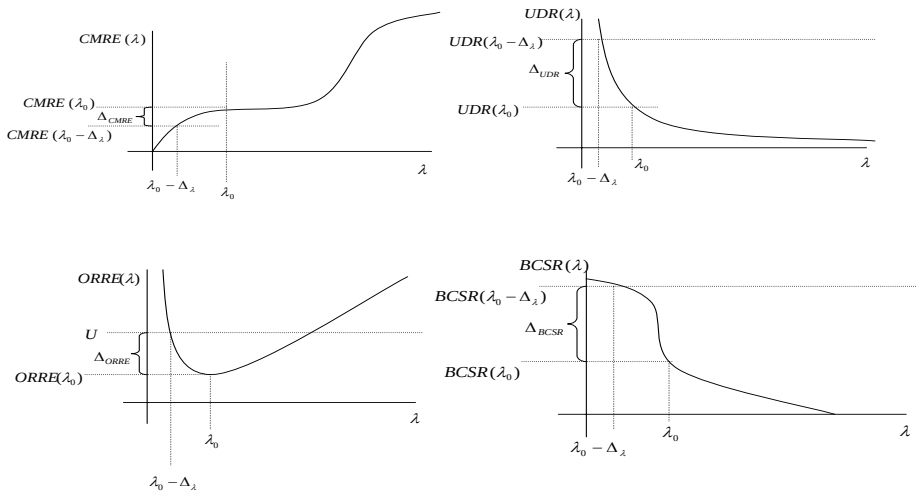


Figure 18: Notations

According to the notation illustrated in Figure 18 above, one can say

$$\begin{aligned} U &= ORRE(\lambda_0 - \triangle_\lambda) \\ &= ORRE(\lambda_0) + \triangle_{ORRE} \\ &= CMRE(\lambda_0) + UDR(\lambda_0) - \triangle_{CMRE} + \triangle_{UDR} \end{aligned}$$

Since we use the intersect concept in our thresholding function, λ_0 is usually located as comparatively small value(left side of the plot of ORRE) and $\triangle_{UDR}$ is much more sensitive to the $\triangle_\lambda$ than $\triangle_{CMRE}$ is. It can be seen in the plot of $CMRE(\lambda)$ and $UDR(\lambda)$. Than is, $\triangle_{CMRE} \ll \triangle_{UDR}$. then one would better focus on $\triangle_{UDR}$ more than $\triangle_{CMRE}$ in the problem of finding optimal U.

Since it is also true that $BCSR(\lambda_0 - \triangle_\lambda) = BCSR\lambda_0) + \triangle_{BCSR}$, the problem of finding U^* is actually same to the problem of finding $\triangle_\lambda^*$(optimal $\triangle_\lambda$) which can minimize $\triangle_{UDR}$ and maximize $\triangle_{BCSR}$, where

$$\begin{aligned} \triangle_{UDR} &= UDR(\lambda_0 - \triangle_\lambda) - UDR(\lambda_0), \\ \triangle_{BCSR} &= BCSR(\lambda_0 - \triangle_\lambda) - BCSR(\lambda_0) \end{aligned}$$

Using the wight (w) for the general purpose, like in our previous research in ORRE definition, we can define $\triangle_\lambda^*$ such as

$$\triangle_\lambda^* = \arg\max_{\triangle_\lambda} \{w \cdot \triangle_{BCSR} - (1-w) \cdot \triangle_{UDR}\}$$

Then the optimal λ^* and the optimal upper limit of ORRE, U^*, can be obtained as

$$\begin{aligned} \lambda^* &= \lambda_0 - \triangle_\lambda^* \\ U^* &= ORRE(\lambda_0 - \triangle_\lambda^*) \end{aligned}$$

In order to understand the impact of UDR and BCSR pattern to the optimal $\triangle_\lambda^*$, one may wish to use the coefficients of polynomial regression to characterize the pattern of UDR and BCSR. Let's assume that the third-order polynomial regression can approximate the $UDR(\lambda)$ and $BCSR(\lambda)$ $(0 < \lambda < \lambda_0)$ accurate enough. Since our interested range of λ is

$\{o < \lambda < \lambda_0\}$,where $BCSR(\lambda)$ can be increased, the input vector of λ for the regression will be discretized in the range.

$$UDR(\lambda) = f(\lambda) = \alpha_0 + \alpha_1\lambda + \alpha_2\lambda^2 + \alpha_3\lambda^3$$
$$BCSR(\lambda) = g(\lambda) = \beta_0 + \beta_1\lambda + \beta_2\lambda^2 + \beta_3\lambda^3$$

Then the $\triangle_{UDR}$ and $\triangle_{BCSR}$ become

$$\begin{aligned}
\triangle_{UDR} &= f(\lambda_0 - \triangle_\lambda) - f(\lambda_0) \\
&= \alpha_1(-\triangle_\lambda) + \alpha_2(-2\lambda_0\,\triangle_\lambda + \triangle_\lambda^2) + \alpha_3(-3\lambda_0^2\,\triangle_\lambda + 3\lambda_0\,\triangle_\lambda^2 - \triangle_\lambda^3) \\
\triangle_{BCSR} &= g(\lambda_0 - \triangle_\lambda) - g(\lambda_0) \\
&= \beta_1(-\triangle_\lambda) + \beta_2(-2\lambda_0\,\triangle_\lambda + \triangle_\lambda^2) + \beta_3(-3\lambda_0^2\,\triangle_\lambda + 3\lambda_0\,\triangle_\lambda^2 - \triangle_\lambda^3)
\end{aligned}$$

Then, when the weight(w) of objective function of $\triangle_\lambda$ is ignored (e.g. same weights), the objective function,$\triangle_{BCSR} - \triangle_{UDR}$, becomes

$$\begin{aligned}
\triangle_{BCSR} - \triangle_{UDR} &= -\triangle_\lambda^3\,(\beta_3 - \alpha_3) \\
&\quad + \triangle_\lambda^2\,\{3\lambda_0(\beta_3 - \alpha_3) + (\beta_2 - \alpha_2)\} \\
&\quad - \triangle_\lambda\,\{3\lambda_0^2(\beta_3 - \alpha_3) + 2\lambda_0(\beta_2 - \alpha_2) + (\beta_1 - \alpha_1)\}
\end{aligned}$$

In order to find optimal $\triangle_\lambda$ which maximize $\triangle_{BCSR} - \triangle_{UDR}$, we need to get the first derivative $\triangle_{BCSR} - \triangle_{UDR})'$.

$$\begin{aligned}
(\triangle_{BCSR} - \triangle_{UDR})' &= -3\,\triangle_\lambda^2\,(\beta_3 - \alpha_3) \\
&\quad + 2\,\triangle_\lambda\,\{3\lambda_0(\beta_3 - \alpha_3) + (\beta_2 - \alpha_2)\} \\
&\quad - \{3\lambda_0^2(\beta_3 - \alpha_3) + 2\lambda_0(\beta_2 - \alpha_2) + (\beta_1 - \alpha_1)\}
\end{aligned}$$

Thus the root of function $(\triangle_{BCSR} - \triangle_{UDR})'$ becomes

$$\triangle_\lambda = \frac{3\lambda_0(\beta_3 - \alpha_3) + (\beta_2 - \alpha_2) \pm \sqrt{2}\sqrt{(\beta_2 - \alpha_2)^2 - 3(\beta_3 - \alpha_3)(\beta_1 - \alpha_1)}}{3(\beta_3 - \alpha_3)}$$

Then the optimal $\triangle_\lambda^*$ would be one of the four candidates. That is, two from the root of

$(\triangle_{BCSR} - \triangle_{UDR})' = 0$, and two from the both end of range of $\triangle_\lambda$.

$$\begin{aligned}
\triangle_\lambda^{(0)} &= 0 \\
\triangle_\lambda^{(1)} &= \frac{3\lambda_0(\beta_3-\alpha_3) + (\beta_2-\alpha_2) - \sqrt{2}\sqrt{(\beta_2-\alpha_2)^2 - 3(\beta_3-\alpha_3)(\beta_1-\alpha_1)}}{3(\beta_3-\alpha_3)} \\
\triangle_\lambda^{(2)} &= \frac{3\lambda_0(\beta_3-\alpha_3) + (\beta_2-\alpha_2) + \sqrt{2}\sqrt{(\beta_2-\alpha_2)^2 - 3(\beta_3-\alpha_3)(\beta_1-\alpha_1)}}{3(\beta_3-\alpha_3)} \\
\triangle_\lambda^{(3)} &= \lambda_0
\end{aligned}$$

If the fifth-order polynomial regression is, for another example, applied, then

$$\begin{aligned}
\triangle_{BCSR} - \triangle_{UDR} &= -\triangle_\lambda^5 (\beta_5-\alpha_5) \\
&\quad + \triangle_\lambda^4 \{5\lambda_0(\beta_5-\alpha_5) + (\beta_4-\alpha_4)\} \\
&\quad - \triangle_\lambda^3 \{10\lambda_0^2(\beta_5-\alpha_5) + 4\lambda_0(\beta_4-\alpha_4) + (\beta_3-\alpha_3)\} \\
&\quad + \triangle_\lambda^2 \{10\lambda_0^3(\beta_5-\alpha_5) + 6\lambda_0^2(\beta_4-\alpha_4) + 3\lambda_0(\beta_3-\alpha_3) + (\beta_2-\alpha_2)\} \\
&\quad + \triangle_\lambda \{5\lambda_0^4(\beta_5-\alpha_5) + 4\lambda_0^3(\beta_4-\alpha_4) + 3\lambda_0^2(\beta_3-\alpha_3) + 2\lambda_0(\beta_2-\alpha_2) + (\beta_1-\alpha_1)\}
\end{aligned}$$

Then the optimal $\triangle_\lambda^*$ would be one of the roots (e.g. $\triangle_\lambda^{(1)}, \cdots, \triangle_\lambda^{(4)}$) of the $(\triangle_{BCSR} - \triangle_{UDR})'$, or 0(e.g. $\triangle_\lambda^{(0)}$) or λ_0(e.g. $\triangle_\lambda^{(5)}$) , which maximize the objective function $\triangle_{BCSR} - \triangle_{UDR}$.

Figure 19 is four types of Mallat data variation simulated with noise error variance $\sigma^2 = 1$ and random effect variance $\tau^2 = 3^2$. We applied several order of polynomial regression to the $BCSR(\lambda)$ and $UDR(\lambda)$ of this data set. The λ-axis of $CMRE(\lambda)$, $UDR(\lambda)$ and $BCSR(\lambda)$ is discretized with the unit of 0.87 and the ORRE-driven optimal λ_0 is obtained as 12.18 (e.g. 14th unit of λ in total 100 units; 100th units equals the maximum of $\overline{d}_{.jk}$). From Figure 20, the third-order polynomial regression would not be appropriate due to the inaccurate approximation of $BCSR(\lambda)$ and $UDR(\lambda)$ resulting inaccurate approximation of $\triangle_{BCSR} - \triangle_{UDR}$. That is, it leads to the far different optimal $\triangle_\lambda^*$ approximation. From Figure 22, the fifth-order polynomial regression would be accurate enough bringing us the quite reliable optimal $\triangle_\lambda^*$.

Table 3 shows that the result of VGWT without considering the increase of BCSR. This table illustrates that,with $\lambda_0 = 12.18$, only 26.56% of original data are remaining as non-zero

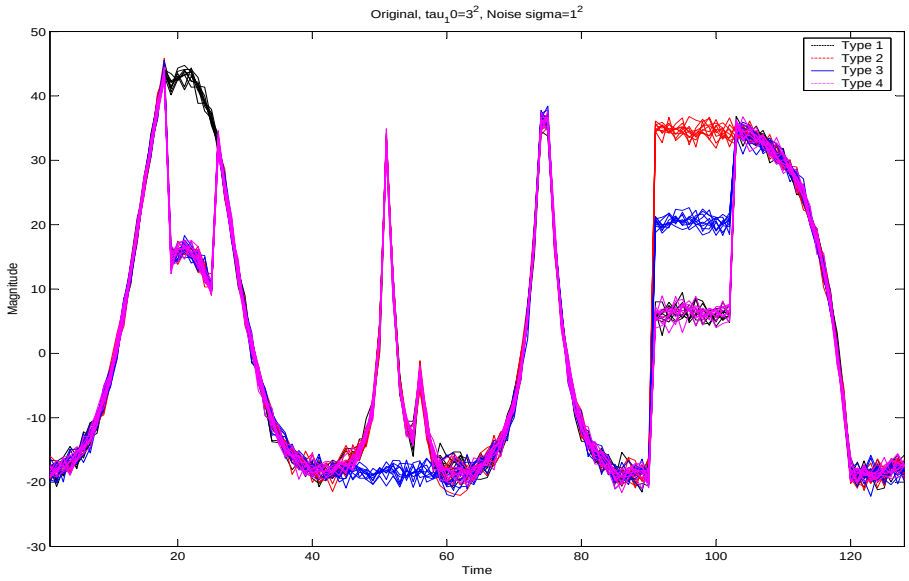


Figure 19: Four types of Mallat data

Table 3: Min-ORRE-based Statistic for Mallat data

σ^2	Min $ORRE(\lambda;\boldsymbol{\theta})$	$\hat{\lambda}$	$CMRE$	UDR	$BCSR$
1^2	0.3171	12.18	0.0515	0.2656	0.7054

and there is 5.15% of reconstruction error were measured. These two measures compose the minimum of ORRE(31.71%). At this time, 70.54% of original between-class variation is reflected in the reduced-size data(non-zero wavelet coefficients). Table 4 represents the result of the second stage with the upper limit of ORRE. It is shown that, as U is getting increased, UDR is getting worse and BCSR is getting better. Table 5 shows the result of deciding the optimal U (e.g. optimal $\triangle^*_\lambda$). As shown from the plot of $\triangle_{BCSR} - \triangle_{UDR}$ in Figure 22 (the optimal $\triangle^*_\lambda$ was the fourth unit), the optimal $\lambda^* = 8.70$ achieved the largest $\triangle_{BCSR} - \triangle_{UDR}$ (0.069). ($\lambda^* = \lambda_0 - \triangle^*_\lambda = 12.18 - 4 * (0.87) = 8.70$). As the result, 28.73% (2.17% up) of original data are remaining as non-zero and 79.61% (9.07% up) of original between-class variation is reflected in the non-zero wavelet coefficients.

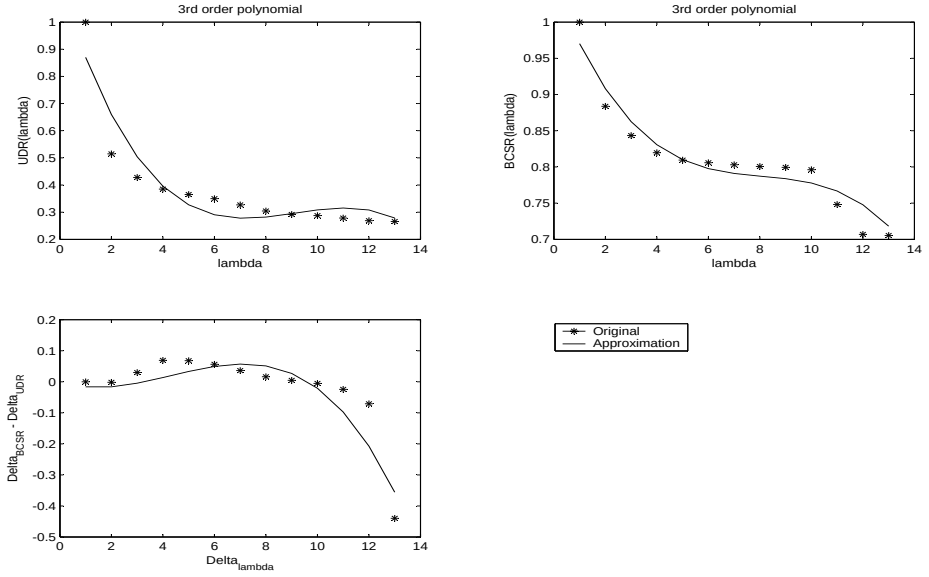


Figure 20: Third-order polynomial regression for $BCSR(\lambda)$ and $UDR(\lambda)$

Table 4: Upper-limit-ORRE(U)-based Statistic for Mallat data

σ^2	U	$\hat{\lambda}$	$CMRE$	UDR	$BCSR$	$\triangle_{BCSR} - \triangle_{UDR}$
1^2	0.32	9.5372	0.0511	0.2689	0.7071	-0.0016
				($\triangle_{UDR} = 0.0033$)	($\triangle_{BCSR} = 0.0017$)	
	0.35	5.9412	0.0422	0.3078	0.8009	0.0533
				($\triangle_{UDR} = 0.0422$)	($\triangle_{BCSR} = 0.0955$)	
	0.45	1.8610	0.0290	0.4210	0.8392	-0.0216
				($\triangle_{UDR} = 0.1554$)	($\triangle_{BCSR} = 0.1338$)	

Table 5: Optimal $\triangle_\lambda$-based Statistic for Mallat data

σ^2	Optimal U^*	λ^*	$CMRE$	UDR	$BCSR$	$\triangle_{BCSR} - \triangle_{UDR}$
1^2	0.3308	8.70	0.0435	0.2873	0.7961	0.0690
				($\triangle_{UDR} = 0.0217$)	($\triangle_{BCSR} = 0.0907$)	

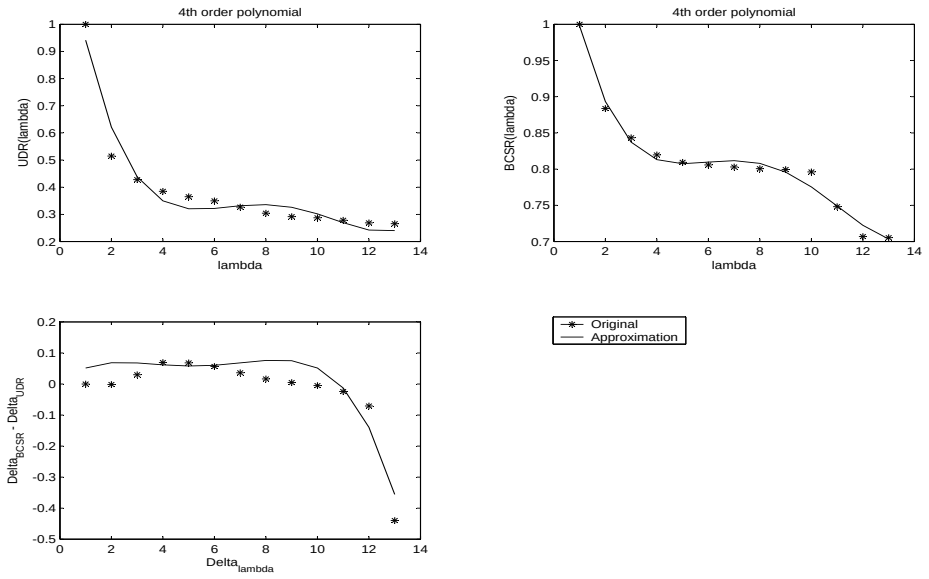


Figure 21: Fourth-order polynomial regression for $BCSR(\lambda)$ and $UDR(\lambda)$

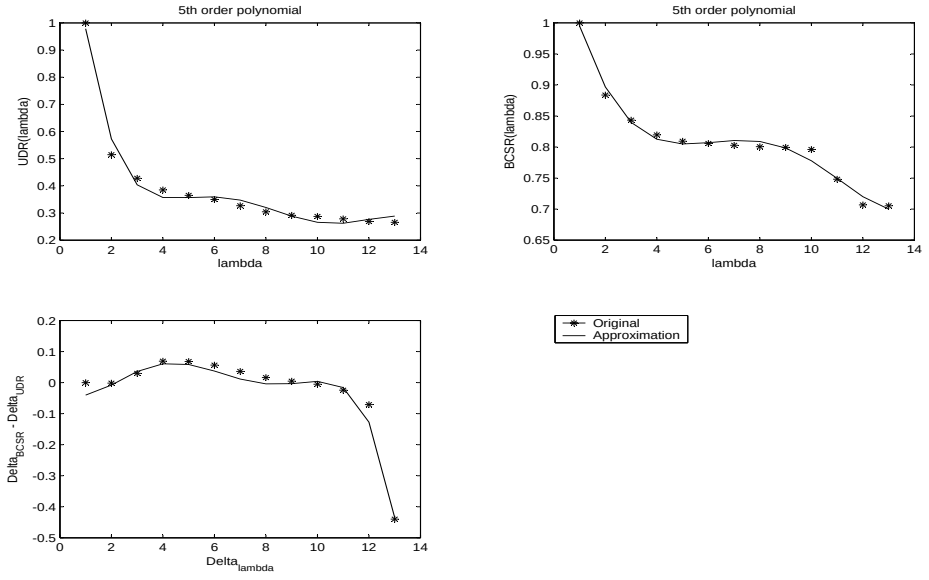


Figure 22: Fifth-order polynomial regression for $BCSR(\lambda)$ and $UDR(\lambda)$

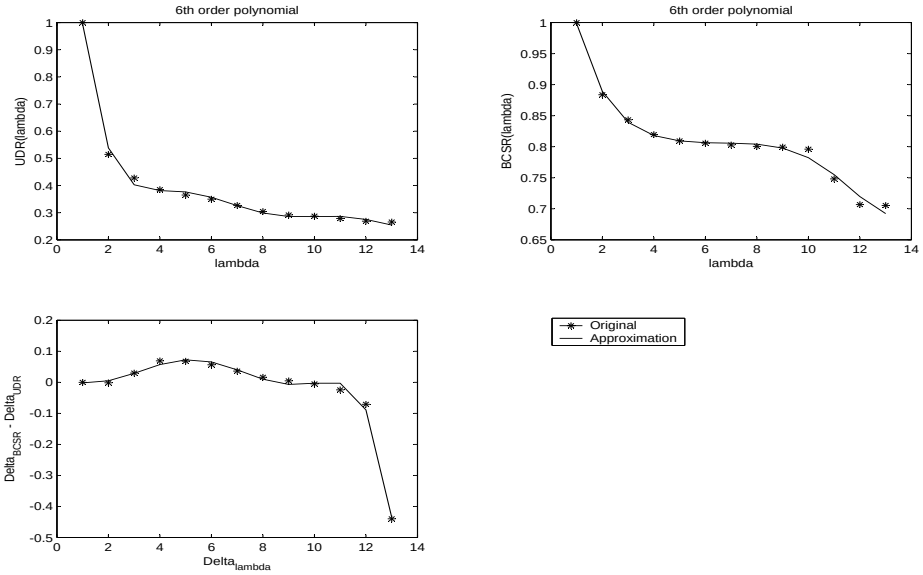


Figure 23: Sixth-order polynomial regression for $BCSR(\lambda)$ and $UDR(\lambda)$

5.4 Illustrative Case Study using Monte-Carlo Simulation

In this section, we illuminate how sensitive the each performance measure is to several variations of situation, such as the cases that the input data set varies with different levels of class-variations, different random-effect positions, and different noise-error variances, etc. The real-life data set would be very complicated, which might naturally be the combination of many cases, so that we can not fully track the impact of each characteristics of the cases above. However, it would be very meaningful to conduct the focused study for each case so that we have better understanding of the possible impact of case-by-case situations.

5.4.1 Different Levels of Class-Variation

In order to study the impact of different levels of class-variation, we generated a simple form of sine curves. On Figure 24, three different cases are considered. The difference between three cases (Case 1,2, and 3) is in the magnitude of difference in signal-class-means at every time positions. Each case presents four classes of similar pattern curves(e.g. sine curves) with different magnitude of signal-class-means. Each curve has the noise-error variance $\sigma^2 = 0.3^2$ and the variance of random-effect at 10th wavelet atom position $\tau_{10}^2 = 3^2$. Each class contains 10 curves and the length of data series is $N = 128$. Each sine curve in kth ($k = 1, 2, 3, 4$) class is generated as the form of $y_i = M_k \cdot sin(2\pi \cdot i/N), i = 1, \cdots, N$ where M_k is the magnitude of kth class. The magnitude $\{M_1, M_2, M_3, M_4\}$ for Case 1 is $\{1, 2, 3, 4\}$, $\{1, 3, 5, 7\}$ for Case 2, and $\{1, 4, 7, 10\}$ for Case 3 so that Case 3 has the most distinct classes and Case 1 has the least.

Table 6 represents the result of VGWT for each case. It shows that, as the the classes are more distinct, the performance of VGWT becomes better. That is ,all the performance measures improve. Even the worst case (Case 1), it shows only 7% of CMRE and 11.6% of UDR. This confirms the excellence of signal mean reconstruction accuracy and data reduction efficiency. When the upper limit of ORRE is decided above the minimum of ORRE, the BCSR is increased as we expected. The amount of increase of BCSR in the case of the least distinct class (case 1)is the largest among three cases. It can be explained as because the case of distinct classes already achieved quite high performance of BCSR. This

Table 6: Min-ORRE-based Statistic for Sine data with different class variation

σ^2	Min $ORRE(\lambda;\boldsymbol{\theta})$	$\hat{\lambda}$	$CMRE$	UDR	$BCSR$
Case 1	0.1789	0.5827	0.0629	0.1160	0.7858
Case 2	0.1409	0.6787	0.0254	0.1155	0.9280
Case 3	0.1268	0.7238	0.0130	0.1138	0.9658

result is presented in Table 7.

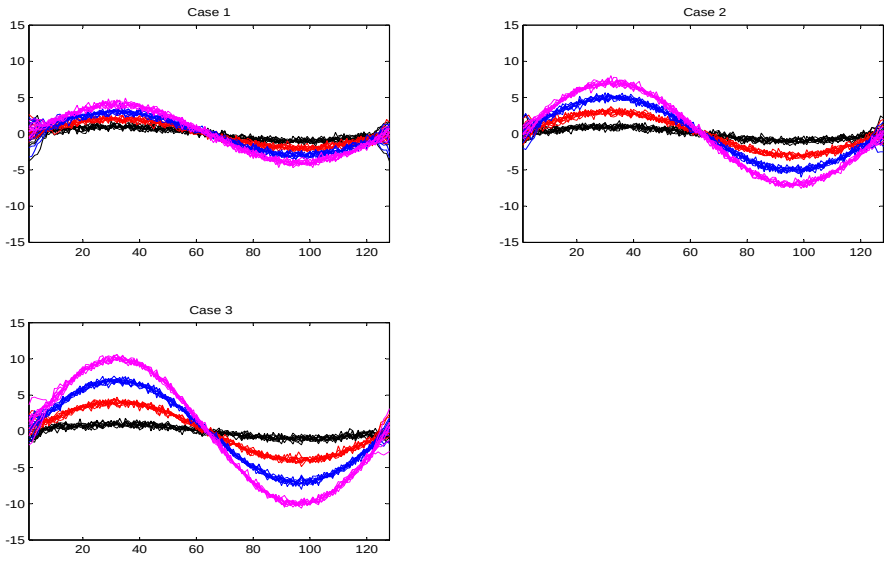


Figure 24: Different Class Separability in Similar Shape

5.4.2 Different Random-effect Positions

In order to study the impact of different random-effect position, we generated a simple form of sine curves like the previous section. In Figure 25, three different cases are considered. The difference between three cases (Case 1,2, and 3) is in the position and number of random-effect wavelet atoms. Each case presents four classes of same pattern curves(e.g. sine curves) with different random-effect wavelet atom positions. Each curve has the noise-error variance $\sigma^2 = 0.3^2$. Each class contains 10 curves and the length of data series is $N = 128$ as used in the previous section. The magnitude $\{M_1, M_2, M_3, M_4\}$ for all cases is $\{1, 3, 5, 7\}$. Case 1 has a single random-effect term at 14th wavelet atom(one of scale

Table 7: Upper-limit-ORRE-based Statistic for Sine data with different class variation

σ^2	U (ORRE Upper Limit)	$\hat{\lambda}$	$CMRE$	UDR	$BCSR$
Case 1	0.25	0.1735	0.0549	0.1953	0.8125
	0.5	0.1032	0.0362	0.4500	0.8765
Case 2	0.25	0.1928	0.02	0.2397	0.9418
	0.5	0.1054	0.0129	0.4983	0.9623
Case 3	0.25	N/A	0.0102	0.2284	0.9726
	0.5	N/A	0.0065	0.4989	0.9826

functions) with $\tau_{14}^2 = 5^2$. For Case 2, $\tau_{22}^2 = 5^2$ in the coarsest level. And Case 3 has two random-effect positions at both 14th and 22nd with $\tau_{14}^2 = 5^2$ and $\tau_{22}^2 = 5^2$. Due to the different support of each wavelet atom position, different type of curve-variation in time domain are shown in Figure 25. Especially, in Case 3, two random-effect wavelet atom positions have the support overlapped, the variation becomes more complicated in time domain.

Table 8 also represent the promising result of signal mean reconstruction error ratio and data reduction efficiency. Interestingly, this simulation results very high BCSR performance. More importantly, three different cases do not provide much difference in all performance measures. This can be explained in two ways. First, the signal curves are generated in random-effect wavelet model which has zero mean and τ_j variance (j is random effect wavelet atom position) so that our thresholding rule, which compare $\overline{d}_{.jk}$ with λ, will not much affected. Second, we set only one or two random-effect wavelet atom positions out of $N = 128$ for this simulation so that their impact to the each measure will not be serious even though they have non-zero mean. This is another example of excellence of random-effect wavelet model in the point that very complicated curve variations in time domain can be very simply modelled in our wavelet model. Table 9 represents the result of the second stage(the upper limit of ORRE). Since the first stage performed good enough in BCSR, the increase in BSCR at second stage is not very significant.

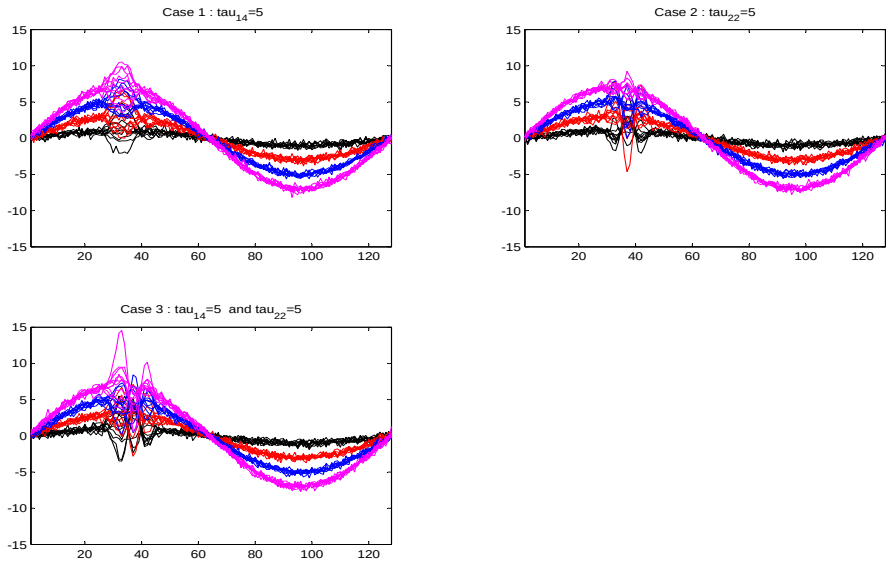


Figure 25: Different Random-effects in Similar Shape

Table 8: Min-ORRE-based Statistic for Sine data with different random-effect coefficients

σ^2	Min $ORRE(\lambda;\boldsymbol{\theta})$	$\hat{\lambda}$	$CMRE$	UDR	$BCSR$
Case 1	0.1304	0.5286	0.0209	0.1095	0.9489
Case 2	0.1409	0.6889	0.0187	0.1159	0.9498
Case 3	0.1268	0.5251	0.0224	0.1160	0.9463

5.4.3 Different Levels of Noise-error Variance

In Figure 26, three different cases of four types of RTCVD data curves are considered. The difference between three cases (Case 1,2, and 3) is in the level of noise-error variance σ^2. The variance of random-effect at 10th wavelet atom position is equally applied to all cases with $\tau_{10}^2 = 3^2$ and its impact is revealed the both end of data series. Each class contains 10 curves and the length of data series is $N = 128$. The different level of noise-error variance is 0.1^2, 0.3^2,and 0.5^2 for Case 1,2,and 3, respectively.

Table 10 clearly shows the impact of noise-error variance. It shows that, as noise-error variance σ^2 increases, CMRE and BCSR are getting worse. Table 11 also shows that the increase of BCSR is not very significant compared to other case studies. The term N/A

Table 9: Upper-limit-ORRE-based Statistic for Sine data with different random-effect coefficients

σ^2	U (ORRE Upper Limit)	$\hat{\lambda}$	$CMRE$	UDR	$BCSR$
Case 1	0.20	0.2062	0.0175	0.1828	0.9556
	0.25	0.1682	0.0159	0.2341	0.9587
Case 2	0.20	0.2082	0.0163	0.1837	0.9558
	0.25	0.1735	0.0148	0.2352	0.9599
Case 3	0.20	0.2113	0.0178	0.1822	0.9548
	0.25	0.1704	0.0163	0.2337	0.9582

is inserted because there is no solution since the upper limit U is below the minimum of ORRE (e.g. U should be greater or equal to 0.3157 for Case 3).

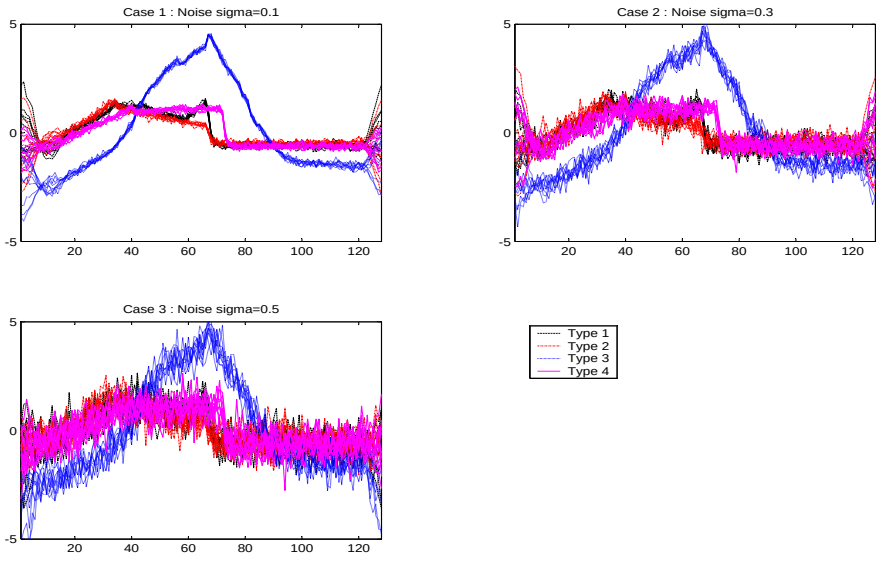


Figure 26: Different Noise-error Variance in RTCVD Data

5.5 Real-life Example: Tonnage data

Sheet-metal stamping has been known as a very complicated and sensitive manufacturing process. In recent years, stamping tonnage sensors have been used widely to measure the stamping force for each stamped part for the purpose of stamping process monitoring and fault diagnosis. Stamping process performance can be illuminated from rich information

Table 10: Min-ORRE-based Statistic for RTCVD data

σ^2	Min $ORRE(\lambda;\boldsymbol{\theta})$	$\hat{\lambda}$	$CMRE$	UDR	$BCSR$
0.1^2	0.1454	0.20	0.0121	0.1333	0.9865
0.3^2	0.2368	0.18	0.1069	0.1299	0.8868
0.5^2	0.3157	0.39	0.1829	0.1328	0.8033

Table 11: Upper-limit-ORRE-based Statistic for RTCVD data

σ^2	U (ORRE Upper Limit)	$\hat{\lambda}$	$CMRE$	UDR	$BCSR$
0.1^2	0.25	0.0383	0.0071	0.2429	0.9923
	0.3	0.0328	0.0065	0.2935	0.9930
	0.35	0.0288	0.0052	0.3448	0.9948
0.3^2	0.25	0.1359	0.0845	0.1655	0.9080
	0.3	0.1024	0.0684	0.2316	0.9241
	0.35	0.0862	0.0588	0.2912	0.9352
0.5^2	0.25	N/A	N/A	N/A	N/A
	0.3	N/A	N/A	N/A	N/A
	0.35	0.2298	0.1539	0.1961	0.8349

and characteristics contained in the tonnage signal. Figure 27 shows the total tonnage or stamping force which is the sum of the outputs of all tonnage sensors mounted on the press. In Figure 27 (a), the horizontal axis is the crank angle which is transformed to appropriate data index with size $N = 256$. In order to monitor successfully stamping process and diagnose any fault type in the past experience, various tonnage signal-analysis techniques have been studied. (Koh, Shi, and Black 1996; Jin and Shi 1999). Especially, due to the massive amount of data from high stamping productivity and the limited storage capacity for historical data, the efficient tonnage data compression techniques, retaining high reconstruction accuracy and high discriminative ability, becomes hot and critical issue recently.

Figure 27 (a) represents three classes of tonnage signals, which has 25 signals in each class. The significant difference among classes mainly resides in the middle lobes at data index from 90 to 115. Figure 27 (b) focuses the middle lobes for visually better distinction. Table 12 shows that data reduction results with the optimal $\lambda_0 = 4021.7$. The result is very impressive with only 0.8% of signal class-mean reconstruction error, only 1.17% of original

Table 12: Min-ORRE-based Statistic for Tonnage data

Min $ORRE(\lambda;\boldsymbol{\theta})$	$\hat{\lambda}$	$CMRE$	UDR	$BCSR$
0.0197	4021.7	0.008	0.0117	0.9290

data usage and 92.9% preservation of original between-class-variation in the reduced-size data. This impressive output is due to the very smooth and similar signal patterns among classes, even in each class. Table 13 shows that, according to our second stage procedure, only less than 1% increase in UDR can achieve almost 7% increase in BCSR (when U=0.025). It is quite successful data reduction process keeping high accuracy of signal representation and preserving almost all distinct features of original data in the reduced size data.

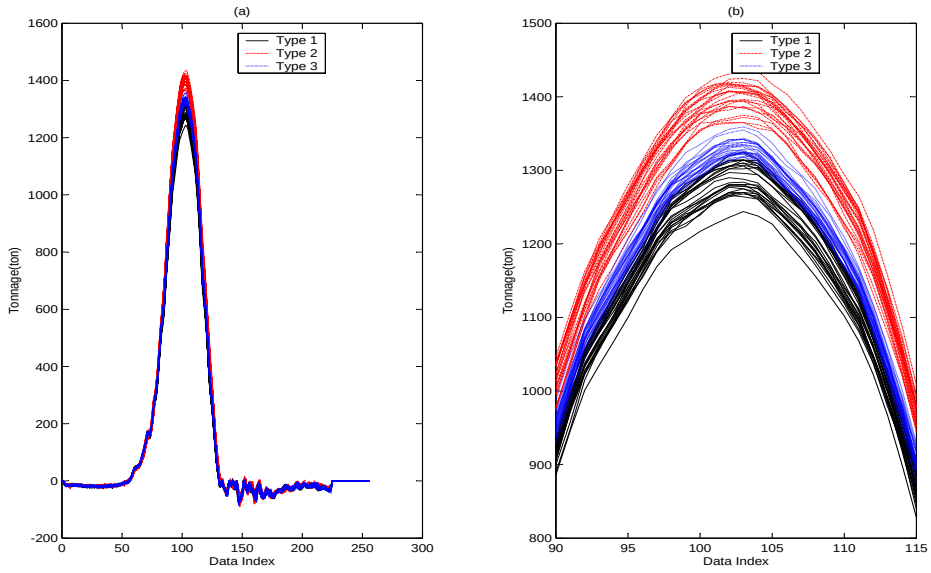


Figure 27: Three Different Types of Tonnage Signal Class

Table 13: Upper-limit-ORRE(U)-based Statistic for Tonnage data

U	$\hat{\lambda}$	$CMRE$	UDR	$BCSR$
0.025	1304.8	0.0035	0.0215	0.9965
			($\triangle_{UDR} = 0.0098$)	($\triangle_{BCSR} = 0.0675$)
0.05	537.1527	0.0014	0.0486	0.9985
			($\triangle_{UDR} = 0.0369$)	($\triangle_{BCSR} = 0.0695$)

CHAPTER VI

CONCLUSION AND FUTURE RESEARCH

6.1 Summary of Results

6.1.1 Wavelet-based Random-effect Model

This section proposed the *Wavelet-based Random-effect Model* to characterize the between-curve variation among multiple complicated functional data with sharp changes. The demand for a new model was addressed with a real life example (Antenna data). The difference from the traditional random-effect model is that our model using wavelet has locally focused impact in time domain. A simple way to find the random-effect wavelet atom positions using a simple QQ-plot is suggested. Several cases of different number of random-effect positions are compared and a real life data set (Antenna data) was successfully modelled using random-effect wavelet model.

6.1.2 Vertical Energy Threshold(VET) without Class Information

For the sake of efficient, reliable and effective data reduction, this section proposed *Vertical Energy Threshold* method. This vertical-energy based thresholding method is easy to understand and implement, where a closed-form expression of the estimate of the thresholding parameter is provided. This parameter depends on the positions of random effects and their variations. Analytical properties such as the strong consistency and the large-sample normal distribution of the parameter estimate are derived. Based on the simulation studies and real-life examples from the antenna data collected from Nortel's manufacturing process, the proposed method is more effective in capturing the key patterns in the multiple data curves with very limited number of coefficients than other single-curve based data denoising methods. The reduced-size data of wavelet coefficients are shown to be very useful in separating the characteristics among several classes of data curves for the clustering analysis.

6.1.3 Vertical Group-wise Threshold(VGWT) with Class Information

Based on class membership information of each signal obtained, this thesis proposed the *Vertical Group-Wise Threshold* method to increase the discriminative capability of the reduced-size data so that the reduced data set retains salient differences between classes as much as possible. The selection problem of class-wise thresholding scheme (intersection, union, and voting) was also briefly addressed. A new thresholding function using intersection and a class-separability measure were proposed for finding the optimal threshold. A two-stage procedure based on these tools successfully increased the class separability with reasonably small loss of data reduction efficiency. Also, investigations on how several different situations can impact the performance of reconstruction accuracy, data reduction ratio, and class separability in the reduced-size data, were carried out using Monte-carlo simulations. A real-life example (Tonnage data) showed our proposed method is promising.

6.2 Future Research

Future work is needed to explore a more rigorous framework to find the random-effect wavelet atom positions in wavelet-based random-effect model, and to use a quantitative measure to decide the most suitable number of random-effect positions. The vertical energy threshold(VET) in a soft thresholding version can also be studied and compared to our hard thresholding one. Other than wavelets, several multiscale methods such as beamlets and wedgelets can be explored in our proposed objective functions. The most suitable degree of high order polynomial regression to characterize the component of the objective functions in vertical group-wise thresholding method and its performance in class separability approximation can be researched in a more mathematical framework. The extension of our proposed methods to 2-D data (image and spatial data) will be very demanding and important in the area of signal processing. Also, its statistical analysis will have a large contribution to the statistical data mining field.

APPENDIX A

SOME ANCILLARY STUFF

A.1 Meta-functions : $P_{jk(\lambda)}$, $Q_{jk}(\lambda)$ *and* $R_{jk}(\lambda)$

Let's assume the distribution of $\bar{d}_{.jk}$ as $N(\mu_{jk}, \sigma^2_{jk})$, and define the two common functions $\phi(y)$ and $\Phi(y)$ as

$$\begin{aligned} \phi(y) &= \frac{1}{\sqrt{2\pi}} \exp(-0.5y^2) \\ \Phi(y) &= \int_{-\infty}^{y} \phi(z)dz \end{aligned}$$

Then, let the meta-functions $P_{jk(\lambda)}$, $Q_{jk}(\lambda)$ and $R_{jk}(\lambda)$ defined as

$$\begin{aligned} P_{jk}(\lambda) &= E[\bar{d}^2_{.jk} \cdot I(|\bar{d}_{.jk}| > \lambda)] \\ Q_{jk}(\lambda) &= E[\bar{d}_{.jk} \cdot I(|\bar{d}_{.jk}| > \lambda)] \\ R_{jk}(\lambda) &= E[I(|\bar{d}_{.jk}| > \lambda)] \end{aligned}$$

These meta-functions will be used in computation of $CMRE(\lambda)$, $UDR(\lambda)$, and $BCSR(\lambda)$. They can be derived to computable formulas using two common functions $\phi(y)$ and $\Phi(y)$, as below.

$$\begin{aligned} P_{jk}(\lambda) &= \left\{1 - \Phi\left(\frac{\lambda_{jk} - \mu_{jk}}{\sigma_{jk}}\right) + \Phi\left(\frac{-\lambda_{jk} - \mu_{jk}}{\sigma_{jk}}\right)\right\}(\sigma^2_{jk} + \mu^2_{jk}) \\ &\quad -2\sigma_{jk}\left\{\phi\left(\frac{-\lambda_{jk} - \mu_{jk}}{\sigma_{jk}}\right) - \phi\left(\frac{\lambda_{jk} - \mu_{jk}}{\sigma_{jk}}\right)\right\} \\ &\quad -\sigma^2_{jk}\left\{\left(\frac{-\lambda_{jk} - \mu_{jk}}{\sigma_{jk}}\right) \cdot \phi\left(\frac{-\lambda_{jk} - \mu_{jk}}{\sigma_{jk}}\right) - \left(\frac{\lambda_{jk} - \mu_{jk}}{\sigma_{jk}}\right) \cdot \phi\left(\frac{\lambda_{jk} - \mu_{jk}}{\sigma_{jk}}\right)\right\} \end{aligned}$$

$$\begin{aligned} Q_{jk}(\lambda) &= \left\{1 - \Phi\left(\frac{\lambda_{jk} - \mu_{jk}}{\sigma_{jk}}\right) + \Phi\left(\frac{-\lambda_{jk} - \mu_{jk}}{\sigma_{jk}}\right)\right\} \cdot \mu_{jk} \\ &\quad -\sigma_{jk}\left\{\phi\left(\frac{-\lambda_{jk} - \mu_{jk}}{\sigma_{jk}}\right) - \phi\left(\frac{\lambda_{jk} - \mu_{jk}}{\sigma_{jk}}\right)\right\} \end{aligned}$$

$$R_{jk}(\lambda) = \left\{1 - \Phi\left(\frac{\lambda_{jk} - \mu_{jk}}{\sigma_{jk}}\right) + \Phi\left(\frac{-\lambda_{jk} - \mu_{jk}}{\sigma_{jk}}\right)\right\}$$

The proof of derivations above is following.

Let's assume the distribution of y as $N(\mu, \sigma^2)$. Then simple notations of $P_{jk(\lambda)}$, $Q_{jk}(\lambda)$ and $R_{jk}(\lambda)$ without subscripts j and k can be derived as

$$\begin{aligned}
R(\lambda) &= E[I(|y| > \lambda)] \\
&= \int_{-\infty}^{\infty} \frac{1}{\sigma}\phi(\frac{y-\mu}{\sigma})dy - \int_{-\lambda}^{\lambda} \frac{1}{\sigma}\phi(\frac{y-\mu}{\sigma})dy \\
&= 1 - \left[\int_{\frac{-\lambda-\mu}{\sigma}}^{\frac{\lambda-\mu}{\sigma}} \phi(z)dz\right] \\
&= 1 - \left\{\Phi(\frac{\lambda-\mu}{\sigma}) - \Phi(\frac{-\lambda-\mu}{\sigma})\right\} \\
&= 1 - \Phi(\frac{\lambda-\mu}{\sigma}) + \Phi(\frac{-\lambda-\mu}{\sigma})
\end{aligned}$$

$$\begin{aligned}
Q(\lambda) &= E[yI(|y| > \lambda)] \\
&= \int_{-\infty}^{\infty} y\frac{1}{\sigma}\phi(\frac{y-\mu}{\sigma})dy - \int_{-\lambda}^{\lambda} y\frac{1}{\sigma}\phi(\frac{y-\mu}{\sigma})dy \\
&= \mu - \left[\int_{\frac{-\lambda-\mu}{\sigma}}^{\frac{\lambda-\mu}{\sigma}} (\mu + \sigma z)\frac{1}{\sigma}\phi(z)\sigma dz\right] \\
&= \mu - \left[\mu\int_{\frac{-\lambda-\mu}{\sigma}}^{\frac{\lambda-\mu}{\sigma}} \phi(z)dz + \sigma\int_{\frac{-\lambda-\mu}{\sigma}}^{\frac{\lambda-\mu}{\sigma}} z\phi(z)dz\right] \\
&= \mu - \left[\mu\int_{\frac{-\lambda-\mu}{\sigma}}^{\frac{\lambda-\mu}{\sigma}} \phi(z)dz + \sigma\left\{\int_{-\infty}^{\frac{\lambda-\mu}{\sigma}} z\phi(z)dz - \int_{-\infty}^{\frac{-\lambda-\mu}{\sigma}} z\phi(z)dz\right\}\right] \\
&= \mu - \left[\mu(\Phi(\frac{\lambda-\mu}{\sigma}) - \Phi(\frac{-\lambda-\mu}{\sigma})) + \sigma(-\phi(\frac{\lambda-\mu}{\sigma}) + \phi(\frac{-\lambda-\mu}{\sigma})\right] \\
&= \left\{1 - \Phi(\frac{\lambda-\mu}{\sigma}) + \Phi(\frac{-\lambda-\mu}{\sigma})\right\}\mu - \sigma\left\{-\phi(\frac{\lambda-\mu}{\sigma}) + \phi(\frac{-\lambda-\mu}{\sigma})\right\}
\end{aligned}$$

$$
\begin{aligned}
P(\lambda) &= E[y^2 I(|y| > \lambda)] \\
&= \int_{-\infty}^{\infty} y^2 \frac{1}{\sigma}\phi(\frac{y-\mu}{\sigma})dy - \int_{-\lambda}^{\lambda} y^2 \frac{1}{\sigma}\phi(\frac{y-\mu}{\sigma})dy \\
&= \mu^2 + \sigma^2 - \left[\int_{\frac{-\lambda-\mu}{\sigma}}^{\frac{\lambda-\mu}{\sigma}} (\mu^2 + 2\sigma z + \sigma^2 z^2)\frac{1}{\sigma}\phi(z)\sigma dz\right] \\
&= \mu^2 + \sigma^2 - \left[\mu^2 \int_{\frac{-\lambda-\mu}{\sigma}}^{\frac{\lambda-\mu}{\sigma}} \phi(z)dz + 2\sigma \int_{\frac{-\lambda-\mu}{\sigma}}^{\frac{\lambda-\mu}{\sigma}} z\phi(z)dz + \sigma^2 \int_{\frac{-\lambda-\mu}{\sigma}}^{\frac{\lambda-\mu}{\sigma}} z^2\phi(z)dz\right] \\
&= \mu^2 + \sigma^2 - \left[\mu^2 \int_{\frac{-\lambda-\mu}{\sigma}}^{\frac{\lambda-\mu}{\sigma}} \phi(z)dz + 2\sigma \left\{\int_{-\infty}^{\frac{\lambda-\mu}{\sigma}} z\phi(z)dz - \int_{-\infty}^{\frac{-\lambda-\mu}{\sigma}} z\phi(z)dz\right\} + \sigma^2 \int_{\frac{-\lambda-\mu}{\sigma}}^{\frac{\lambda-\mu}{\sigma}} z^2\phi(z)dz\right] \\
&= \mu^2 + \sigma^2 \\
&\quad -\mu^2\left\{\Phi(\frac{\lambda-\mu}{\sigma}) - \Phi(\frac{-\lambda-\mu}{\sigma})\right\} - 2\sigma\left\{-\phi(\frac{\lambda-\mu}{\sigma}) + \phi(\frac{-\lambda-\mu}{\sigma})\right\} \\
&\quad -\sigma^2\left\{-(\frac{\lambda-\mu}{\sigma})\phi(\frac{\lambda-\mu}{\sigma}) + (\frac{-\lambda-\mu}{\sigma})\phi(\frac{-\lambda-\mu}{\sigma}) + \Phi(\frac{\lambda-\mu}{\sigma}) - \Phi(\frac{-\lambda-\mu}{\sigma})\right\} \\
&= \left\{1 - \Phi\left(\frac{\lambda-\mu}{\sigma}\right) + \Phi\left(\frac{-\lambda-\mu}{\sigma}\right)\right\}(\sigma^2 + \mu^2) \\
&\quad -2\sigma\left\{\phi\left(\frac{-\lambda-\mu}{\sigma}\right) - \phi\left(\frac{\lambda-\mu}{\sigma}\right)\right\} \\
&\quad -\sigma^2\left\{\left(\frac{-\lambda-\mu}{\sigma}\right)\cdot\phi\left(\frac{-\lambda-\mu}{\sigma}\right) - \left(\frac{\lambda-\mu}{\sigma}\right)\cdot\phi\left(\frac{\lambda-\mu}{\sigma}\right)\right\}
\end{aligned}
$$

where

$$
\begin{aligned}
\int_{\frac{-\lambda-\mu}{\sigma}}^{\frac{\lambda-\mu}{\sigma}} \phi(z)dz &= \Phi(\frac{\lambda-\mu}{\sigma}) - \Phi(\frac{-\lambda-\mu}{\sigma}), \\
\int_{-\infty}^{\frac{\lambda-\mu}{\sigma}} z\phi(z)dz &= -\phi(\frac{\lambda-\mu}{\sigma}), \\
\int_{-\infty}^{\frac{-\lambda-\mu}{\sigma}} z\phi(z)dz &= -\phi(\frac{-\lambda-\mu}{\sigma}), \\
\int_{\frac{-\lambda-\mu}{\sigma}}^{\frac{\lambda-\mu}{\sigma}} z^2\phi(z)dz &= \frac{1}{\sqrt{2\pi}}\left\{-z\cdot e^{-0.5z^2}|_{\frac{-\lambda-\mu}{\sigma}}^{\frac{\lambda-\mu}{\sigma}} + \int_{\frac{-\lambda-\mu}{\sigma}}^{\frac{\lambda-\mu}{\sigma}} \phi(z)dz\right\} \\
&= -(\frac{\lambda-\mu}{\sigma})\phi(\frac{\lambda-\mu}{\sigma}) + (\frac{-\lambda-\mu}{\sigma})\phi(\frac{-\lambda-\mu}{\sigma}) + \Phi(\frac{\lambda-\mu}{\sigma}) - \Phi(\frac{-\lambda-\mu}{\sigma})
\end{aligned}
$$

.

A.2 Class Separability

First, let's consider $E[tr(SSB(\lambda))]$ term based on Between-class variability using Class mean like following.

$$E\left[\sum_{j=1}^{N}\sum_{k=1}^{K} n_k(\overline{d}_{.jk}-\overline{d}_{.j.})^2 \cdot \prod_{k=1}^{K} I(|\overline{d}_{.jk}|>\lambda)\right]$$

For ease of computation of $E[tr(SSB(\lambda))]$, we investigate the followings.

$$\begin{aligned}\sum_{k=1}^{K} n_k(\overline{d}_{.jk}-\overline{d}_{.j.})^2 &= \sum_{k=1}^{K} n_k(\overline{d}^2_{.jk}-2\overline{d}_{.jk}\overline{d}_{.j.}+\overline{d}^2_{.j.}) \\ &= \sum_{k=1}^{K} n_k\overline{d}^2_{.jk}-2\overline{d}_{.j.}\sum_{k=1}^{K} n_k\overline{d}_{.jk}+n_T\overline{d}^2_{.j.} \\ &= \sum_{k=1}^{K} n_k\overline{d}^2_{.jk}-n_T\overline{d}^2_{.j.} \quad since \quad \overline{d}_{.j.}=\frac{\sum_{k=1}^{K} n_k\overline{d}_{.jk}}{n_T}\end{aligned}$$

$$\sum_{k=1}^{K} n_k(\overline{d}_{.jk}-\overline{d}_{.j.})^2 \cdot \prod_{k=1}^{K} I(|\overline{d}_{.jk}|>\lambda)) = \sum_{k=1}^{K}\left[n_k\overline{d}^2_{.jk}\prod_{k=1}^{K} I(|\overline{d}_{.jk}|>\lambda)\right]-n_T\overline{d}_{.j.}\prod_{k=1}^{K} I(|\overline{d}_{.jk}|>\lambda)$$

Then, the $E[tr(SSB(\lambda))]$ term is rearranged like below.

$$\begin{aligned}& E\left[\sum_{j=1}^{N}\sum_{k=1}^{K} n_k(\overline{d}_{.jk}-\overline{d}_{.j.})^2 \cdot \prod_{k=1}^{K} I(|\overline{d}_{.jk}|>\lambda))\right] \\ = \;& \sum_{j=1}^{N}\sum_{k=1}^{K} n_k E\left[\overline{d}^2_{.jk}\prod_{k=1}^{K} I(|\overline{d}_{.jk}|>\lambda)\right]-n_T\sum_{j=1}^{N} E\left[\overline{d}^2_{.j.}\prod_{k=1}^{K} I(|\overline{d}_{.jk}|>\lambda)\right] \\ = \;& \sum_{j=1}^{N}\sum_{k=1}^{K} n_k E\left[\overline{d}^2_{.jk} I(|\overline{d}_{.jk}|>\lambda)\right]\prod_{r\neq k} E\left[I(|\overline{d}_{.jr}|>\lambda)\right]-n_T\sum_{j=1}^{N} E\left[\overline{d}^2_{.j.}\prod_{k=1}^{K} I(|\overline{d}_{.jk}|>\lambda)\right]\end{aligned}$$

If we look at the second term in the last equation above, it can be rearranged like below.

$$\begin{aligned}& n_T\sum_{j=1}^{N} E\left[\overline{d}^2_{.j.}\prod_{k=1}^{K} I(|\overline{d}_{.jk}|>\lambda)\right] \\ = \;& n_T\sum_{j=1}^{N} E\left[(\frac{\sum_{k=1}^{K} n_k\overline{d}_{.jk}}{n_T})^2\prod_{k=1}^{K} I(|\overline{d}_{.jk}|>\lambda)\right] \\ = \;& n_T\sum_{j=1}^{N} E\left[\frac{1}{n_T^2}(\sum_{k=1}^{K} n_k^2\overline{d}^2_{.jk}+2\sum_{a\neq b} n_a n_b\overline{d}_{.ja}\overline{d}_{.jb})\prod_{k=1}^{K} I(|\overline{d}_{.jk}|>\lambda)\right]\end{aligned}$$

$$
\begin{aligned}
&= \sum_{j=1}^{N} \frac{1}{n_T} E\left[\sum_{k=1}^{K} (n_k^2 \overline{d}_{.jk}^2 I(|\overline{d}_{.jk}| > \lambda) \prod_{r \neq k} I(|\overline{d}_{.jr}| > \lambda))\right] \\
&\quad + \sum_{j=1}^{N} \frac{2}{n_T} E\left[\sum_{a \neq b} n_a n_b \overline{d}_{.ja} I(|\overline{d}_{.ja}| > \lambda) \overline{d}_{.jb} I(|\overline{d}_{.jb}| > \lambda) \prod_{r \neq a,b} I(|\overline{d}_{.jr}| > \lambda)\right] \\
&= \frac{1}{n_T} \sum_{j=1}^{N} \sum_{k=1}^{K} n_k^2 E(\overline{d}_{.jk}^2 I(|\overline{d}_{.jk}| > \lambda)) \prod_{r \neq k} E(I(|\overline{d}_{.jr}| > \lambda)) \\
&\quad + \frac{2}{n_T} \sum_{j=1}^{N} \sum_{a \neq b} n_a n_b E(\overline{d}_{.ja} I(|\overline{d}_{.ja}| > \lambda)) E(\overline{d}_{.jb} I(|\overline{d}_{.jb}| > \lambda)) \prod_{r \neq a,b} E(I(|\overline{d}_{.jr}| > \lambda))
\end{aligned}
$$

Then, according to the definition of meta-functions, $E[tr(SSB(\lambda))]$ term in new notations is

$$
\begin{aligned}
&E\left[\sum_{j=1}^{N} \sum_{k=1}^{K} n_k (\overline{d}_{.jk} - \overline{d}_{.j.})^2 \cdot \prod_{k=1}^{K} I(|\overline{d}_{.jk}| > \lambda)\right] \\
&= \sum_{j=1}^{N} \sum_{k=1}^{K} n_k E\left[\overline{d}_{.jk}^2 I(|\overline{d}_{.jk}| > \lambda)\right] \prod_{r \neq k} E\left[I(|\overline{d}_{.jr}| > \lambda)\right] - n_T \sum_{j=1}^{N} E\left[\overline{d}_{.j.}^2 \prod_{k=1}^{K} I(|\overline{d}_{.jk}| > \lambda)\right] \\
&= \sum_{j=1}^{N} \sum_{k=1}^{K} n_k E\left[\overline{d}_{.jk}^2 I(|\overline{d}_{.jk}| > \lambda)\right] \prod_{r \neq k} E\left[I(|\overline{d}_{.jr}| > \lambda)\right] \\
&\quad - \frac{1}{n_T} \sum_{j=1}^{N} \sum_{k=1}^{K} n_k^2 E(\overline{d}_{.jk}^2 I(|\overline{d}_{.jk}| > \lambda)) \prod_{r \neq k} E(I(|\overline{d}_{.jr}| > \lambda))) \\
&\quad - \frac{2}{n_T} \sum_{j=1}^{N} \sum_{a \neq b} n_a n_b E(\overline{d}_{.ja}^2 I(|\overline{d}_{.ja}| > \lambda)) E(\overline{d}_{.jb}^2 I(|\overline{d}_{.jb}| > \lambda)) \prod_{r \neq a,b} E(I(|\overline{d}_{.jr}| > \lambda)) \\
&= \sum_{j=1}^{N} \sum_{k=1}^{K} n_k P_{jk}(\lambda) \prod_{r \neq k} R_{jk}(\lambda) \\
&\quad - \frac{1}{n_T} \sum_{j=1}^{N} \sum_{k=1}^{K} n_k^2 P_{jk} \prod_{r \neq k} R_{jk}(\lambda)) \\
&\quad - \frac{2}{n_T} \sum_{j=1}^{N} \sum_{a \neq b} n_a n_b Q_{ja}(\lambda) Q_{jb}(\lambda) \prod_{r \neq a,b} R_{jr} \lambda \\
&= \sum_{j=1}^{N} \sum_{k=1}^{K} n_k \left(1 - \frac{n_k}{n_T}\right) P_{jk}(\lambda) \prod_{r \neq k} R_{jr}(\lambda) \\
&\quad - \frac{2}{n_T} \sum_{j=1}^{N} \sum_{a \neq b} n_a n_b Q_{ja}(\lambda) Q_{jb}(\lambda) \prod_{r \neq a,b} R_{jr} \lambda
\end{aligned}
$$

.

A.3 $ORRE(\lambda)$ ***for VGWT***

The nominator of the cluster mean reconstruction error term ($CMRE(\lambda)$) using cluster mean in new notations is

$$
\begin{aligned}
& E\left[\sum_{j=1}^{N}\sum_{k=1}^{K} n_k(\overline{d}_{.jk} - \overline{d}_{.jk}\cdot\prod_{k=1}^{K} I(|\overline{d}_{.jk}|>\lambda))^2\right] \\
= & \sum_{j=1}^{N}\sum_{k=1}^{K} n_k E\left[(\overline{d}_{.jk} - \overline{d}_{.jk}\cdot\prod_{k=1}^{K} I(|\overline{d}_{.jk}|>\lambda))^2\right] \\
= & \sum_{j=1}^{N}\sum_{k=1}^{K} n_k E\left[(\overline{d}_{.jk} - \overline{d}_{.jk}\cdot\prod_{k=1}^{K} I(|\overline{d}_{.jk}|>\lambda))^2 \;\mid\; |\overline{d}_{.jk}|>\lambda \;\; for \;\; all\; k\right] \\
& \quad \times P(|\overline{d}_{.jk}|>\lambda \;\; for \;\; all\; k) \\
& + \sum_{j=1}^{N}\sum_{k=1}^{K} n_k E\left[(\overline{d}_{.jk} - \overline{d}_{.jk}\cdot\prod_{k=1}^{K} I(|\overline{d}_{.jk}|>\lambda))^2 \;\; |(|\overline{d}_{.jk}|>\lambda \;\; for \;\; all\; k)^c\right] \\
& \quad \times P((|\overline{d}_{.jk}|>\lambda \;\; for \;\; all\; k)^c) \\
= & \sum_{j=1}^{N}\sum_{k=1}^{K} n_k \left[0 + E(\overline{d}^2_{.jk})\cdot P((|\overline{d}_{.jk}|>\lambda \;\; for \;\; all\; k)^c)\right] \\
= & \sum_{j=1}^{N}\sum_{k=1}^{K} n_k \cdot (\mu^2_{jk}+\sigma^2_{jk})\cdot(1-\prod_{r=1}^{K} R_{jk}(\lambda)) \\
= & \sum_{j=1}^{N}\left[(1-\prod_{r=1}^{K} R_{jk}(\lambda))\cdot\sum_{k=1}^{K} n_k\cdot(\mu^2_{jk}+\sigma^2_{jk})\right]
\end{aligned}
$$

The denominator of $CMRE(\lambda)$ can easily derived from the above derivation.

The used-data-ratio $UDR(\lambda)$ using cluster mean in new notations is

$$
\begin{aligned}
E\left[\frac{\sum_{j=1}^{N}\prod_{k=1}^{K} I(|\overline{d}_{.jk}|>\lambda)}{N}\right] &= \frac{1}{N}\sum_{j=1}^{N} E\left[\prod_{k=1}^{K} I(|\overline{d}_{.jk}|>\lambda)\right] \\
&= \frac{1}{N}\sum_{j=1}^{N}\left[\prod_{k=1}^{K} E(I(|\overline{d}_{.jk}|>\lambda))\right] \\
&= \frac{1}{N}\sum_{j=1}^{N}\left[\prod_{k=1}^{K} R_{jk}(\lambda)\right]
\end{aligned}
$$

A.4 ***Proof of monotonicity of*** $E[tr(SSB(\lambda))]$

Let's define $\lambda_L = \lambda_S + \varepsilon$, where λ_i is any certain value in $\mathbb{R}^+$ (i.e., $0 < \lambda_i < \infty$) and $\varepsilon > 0$, and $J(\lambda_i) = E[tr(SSB(\lambda_i))]$. Considering several sets of $j's$ that are associated with λ_L

and λ_S, we can define S, L, and $L^c|S$, i.e., sets of the form

$$\begin{aligned} S &= \{j; \prod_k^K I(|\bar{d}_{.jk}| > \lambda_S) = 1\} \\ L &= \{j; \prod_k^K I(|\bar{d}_{.jk}| > \lambda_L) = 1\} \\ L^c|S &= \{j; j \in S, \textit{ but not } j \in L\} \end{aligned}$$

and realize $S \supset L$ since $\lambda_S < \lambda_L$. Also we define J^S, J^L, and $J^{L^c|S}$ in the manner that

$$J^a = \sum_{j \in a} \sum_{k=1}^K n_k (\bar{d}_{.jk} - \bar{d}_{.j.})^2$$

Then

$$\begin{aligned} J(\lambda_S) &= E[\sum_{j=1}^N \sum_{k=1}^K n_k (\bar{d}_{.jk} - \bar{d}_{.j.})^2 \prod_k^K I(|\bar{d}_{.jk}| > \lambda_S)] \\ &= E[\sum_{j \in S} \sum_{k=1}^K n_k (\bar{d}_{.jk} - \bar{d}_{.j.})^2] = J^S \\ &= E[\sum_{j \in L} \sum_{k=1}^K n_k (\bar{d}_{.jk} - \bar{d}_{.j.})^2] + E[\sum_{j \in L^c|S} \sum_{k=1}^K n_k (\bar{d}_{.jk} - \bar{d}_{.j.})^2] \\ &= J^L + J^{L^c|S} \\ &= J(\lambda_L) + J^{L^c|S} \end{aligned}$$

Realizing that $J^{L^c|S}$ is always greater than or equal to 0, it is proved that $J(\lambda_S) \geq J(\lambda_L)$. That is, $J(\lambda_i)$ is monotonically decreasing function.

A.5 Invariance Property

The MLE satisfies the invariance principle; *If $\hat{\theta}$ is the MLE of θ, then for any function $\tau(\theta)$ the MLE of $\tau(\theta)$ is $\tau(\hat{\theta})$.*

Note that the invariance property also holds for vector θ. Then invariance principle states that if $Y = \tau(X)$ denotes a change in measurement scale such that X and Y have the same underlying structure, then inference about a parameter is invariant under the transformation.

A.6 Iterative Secant Method

The slop of the curve $y = f(x)$ at the point $x = x_k$ can be approximated in the case if the exact derivative $f'(x_k)$ is difficult to compute. The backward difference formula approximates the derivative by the slop of the secant line between two points $(x_k, f(x_k))$ and $(x_{k-1}, f(x_{k-1}))$:

$$f'(x_k) = \frac{f(x_k) - f(x_{k-1})}{x_k - x_{k-1}}$$

With the backward approximation, the Newton-Raphson method becomes the secant method:

$$x_{k+1} = g(x_k) = x_k \cdot \frac{f(x_k)(x_k - x_{k-1})}{f(x_k) - f(x_{k-1})}$$

Starting with two initial approximations x_0 and x_1, use the mapping above to find x_2, etc. If the sequence $\{x_k\}$ converges to a fixed point $x = x_*$ such that $x_* = g(x_*)$, then the value x_* is the root of the nonlinear equation $f(x) = 0$.

REFERENCES

[1] Antoniadis, A., Gijbels, I., and Gregoire, G., "Model selection using wavelet decomposition and applicaitons," *Biometrika*, vol. 84, no. 4, pp. 751–763, 1997.

[2] Antoniadis, A., Gregoire, G., and Nason, G. P., "Density and hazard rate estimation for right censored data using wavelets," *J. R. Statist. Soc. B*, 1999.

[3] Bakshi, B. R., "Multiscale pca with application to multivariate statistical process monitoring," *AIChe Journal*, vol. 44, no. 7, pp. 1596–1610, 1998.

[4] Berry, M. J. A. and Lidoff, G., *Data Mining Techniques for Marketing, Sales, and Customer Support*. New York: Wiley, 1997.

[5] Box, G. E. P. and Meyer, R. D., "An analysis for unreplicated fractional factorial," *Technometrics*, vol. 28, 1986.

[6] Braha, D. e. a., *Data Mining for Design and Manufacturing*. Kluwer, 2001.

[7] Chen, Z., *Data Mining and Uncertain Reasoning*. New York: Wiley, 2001.

[8] Cios, K. J., Pedrycz, W., and Swiniarski, R., *Data Mining Methods for Knowledge Discovery*. Boston: Kluwer, 1998.

[9] Donoho, D. L. and Johnstone, I. M., "Ideal spatial adaptation by wavelet shrinkage," *Biometrika*, vol. 81, no. 4, pp. 425–455, 1994.

[10] Donoho, D. L. and Johnstone, I. M., "Adapting to unknown smoothness in wavelet shrinkage," *Journal of the American Statistical Association*, vol. 90, 1995.

[11] Duda, R. O., Hart, P. E., and Stork, D. G., *Pattern Classification*. Reading, Massachusetts: Wiley-Interscience, 2001.

[12] Fan, J., "Test of significance based on wavelet thresholding and neyman's truncation," *Journal of American Statistical Association*, vol. 91, 1996.

[13] Faraway, J. J., "Regression analysis for a functional response," *Technometrics*, vol. 39, pp. 254–261, 1997.

[14] Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P., and Ramasamy, U., *Advances in Knowledge Discovery and Data Mining*. Cambridge, MA: AAAI/MIT Press, 1996.

[15] Ganesan, R., Das, T. K., Sikder, A. K., and Kumar, A., "Wavelet-based identification of delamination defect in cmp (cu-low k) using nonstationary acoustic emission signal," *IEEE Trans. on Semiconductor Manufacturing*, vol. 16, pp. 677–685, Nov. 2003.

[16] Gardner, M. M., Lu, J. C., Gyurcsik, R. S., Wortman, J. J., Hornung, B. B., Heinisch, H. H., Rying, E. A., Rao, S., Davis, J. C., and Mozumder, P. K., "Equipment fault detection using spatial signatures," *IEEE Transaction on Components, Packaging, and Manufacturing Technology- Part C*, vol. 20, 1997.

[17] Hall, P., Poskitt, D. S., and Presnell, B., "A functional data-analytic approach to signal discrimination," *Technometrics*, vol. 43, no. 1, pp. 1–9, 2001.

[18] Han, J. and Kamer, M., *Data Mining: Concepts and Techniques*. San Francisco: Morgan Kaufmann, 2000.

[19] Hastie, T., Tibshirani, R., and Friedman, J., *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2001.

[20] Jeong, M. K. and Lu, J. C., "Wavelet-based spc procedure for complicated functional data," tech. rep., The School of Industrial and Systems Engineering, Georgia Tech, 2004.

[21] Jeong, M. K., Lu, J. C., Huo, X., Vidakovic, B., and Chen, D., "Wavelet-based data reduction techniques for fault detection and classification," tech. rep., The School of Industrial and Systems Engineering, Georgia Tech, 2003.

[22] Jin, J. and Shi, J., "Feature-preserving data compression of stamping tonnage information using wavelets," *Technometrics*, vol. 41, no. 4, pp. 327–339, 1999.

[23] Jin, J. and Shi, J., "Automatic feature extraction of waveform signals for inprocess diagnostic performance improvement," *Journal of Intelligent Manufacturing*, vol. 12, 2001.

[24] Kennedy, R. L., van Roy Lee, Y., Reed, C. C., and Lippman, R. P., *Solving Data Mining Problems through Pattern Recognition*. Upper Saddle River, NJ: Prentice-Hall, 1997.

[25] Koh, C. K. H., Shi, J., Williams, W., and Ni, J., "Multiple fault detection and isolation using the haar transform," *ASME Transactions, Journal of Manufacturing Science and Engineering*, vol. 121, no. 2, 1999.

[26] Lada, E. K., Lu, J. C., and Willson, J. R., "A wavelet- based procedure for process fault detection," *IEEE Transactions on Semiconductor Manufacturing*, vol. 15, no. 1, pp. 79–90, 2002.

[27] Learned, R. E. and Willsky, A. S., "A wavelet packet approach to transient signal classification," *Appl. Comput. Harm. Anal*, vol. 2, 1995.

[28] Lenth, R. V., "Quick and easy analysis of unreplicated factorials," *Technometrics*, vol. 31, no. 4, 1989.

[29] Mallat, S. G., *A Wavelet Tour of Signal Processing*. San Diago: Academic Press, 1998.

[30] Moulin, P., "Wavelet thresholding techniques for power spectrum estimation," *IEEE Trans. Signal Processing*, vol. 42, 1994.

[31] Pittner, S. and Kamarthi, S. V., "Feature extraction from wavelet coefficients for pattern recognition tasks," *IEEE Transactions on Pattern Analysis And Machine Intelligence*, vol. 21, no. 1, pp. 83–88, 1999.

[32] Ramsay, J. O. and Silverman, B. W., *Funtional Data Analysis*. Springer, 1991.

[33] Rao, R. M. and Bopardikar, A. S., *Wavelet Transforms: Introduction to Theory and Applications*. Massachusetts: Addison-Wesley, 1998.

[34] Rying, E. A., Ozturk, M. C., Bilbro, G. L., and Lu, J. C., "In situ selectivity and thickness monitoring during selective silicon epitaxy using quadrupole mass spectrometry," *Rapid Thermal and Other Short-Time Processing Technologies I in the Proceedings of the 197th Meeting of the Electrochemical Society, Session I1*, 2003.

[35] Saito, N., *Simultaneous Noise Suppression and Signal Compression Using a Library of Orthonormal Bases and the Minimum Description Length Criterion : Wavelets in Geophysics*. New York: Academic Press, 1994.

[36] Serfling, R. J., *Approximation Theorems of Mathematical Statistics*. New York: John Wiley, 1980.

[37] Skillicorn, D., "Strategies for parallel data mining," *IEEE Concurrency*, 1999.

[38] Vidakovic, B., "Nonlinear wavelet shrinkage with bayes rules and bayes factors," *J. Am, Statist. Ass*, vol. 93, 1998.

[39] Vidakovic, B., *Statistical Modeling by Wavelets*. John Wiley & Sons, 1999.

[40] Westphal, C. and Blaxton, T., *Data Mining Solutions: Methods and Tools for Solving Real-World Problems*. New York: Wiley, 1998.

[41] Weyrich, N. and Warhola, G. T., "Wavelet shrinkage and generalized cross validation for image denoising," *IEEE Transactions on Image Processing*, vol. 7, no. 1, pp. 82–90, 1998.

[42] Woodall, W. H., Spitzner, D. J., Montgomery, D. C., and Gupta, S., "Using control charts to monitor process and product profiles," *submitted to Journal of Quality Technology*, 2003.

[43] Zhang, D., Lin, X., Raz, J., and Sowers, M., "Semiparametric stochastic mixed models for longgitudinal data," *Journal of the American Statistical Association*, vol. 93, 1998.

VITA

Uk Jung is a Ph.D. candidate at the Department of Industrial and Systems Engineering, Georgia Institute of Technology. He was born January 19, 1972, in Pusan, Korean. He received his Bachelor of Science degree in Industrial Engineering from Sungkyunkwan University, Seoul, Korea in 1999, and his Master of Science degree in Operations Research from Georgia Institute of Technology in 2000. His research interests include Industrial Statistics, Data Mining, and Operations Research. Also, his interest involves business and management consulting. Besides being a researcher, he has served as the President of the Korean Student Association in Georgia Tech.