Wavelet-based Data Reduction Techniques for Process Fault Detection

Myong K. Jeong, Jye-Chyi Lu, Xiaoming Huo, Brani Vidakovic and Di Chen

School of Industrial and Systems Engineering, Georgia Institute of Technology,
Atlanta, GA 30332-0205, U.S.A.

ABSTRACT

To handle potentially large and complicated nonstationary data curves, this article presents new data reduction methods based on the discrete wavelet transform. The methods minimize objective functions to balance the tradeoff between data reduction and modeling accuracy. Theoretic investigations provide the optimality of the methods and the large-sample distribution of a closed-form estimate of the thresholding parameter. An upper bound of errors in signal approximation (or estimation) is derived. Based on evaluation studies with popular testing curves and real-life data sets, the proposed methods demonstrate their competitiveness to the existing engineering data-compression and statistical data-denoising methods for achieving the data reduction goals. Further experimentation with a tree-based classification procedure for identifying process fault classes illustrates the potential of the data-reduction tools. Extension of the engineering scalogram to the reduced-size semiconductor fabrication data leads to a visualization tool for monitoring and understanding process problems.

KEY WORDS: Data denoising; Data mining; Quality Improvement; Scalogram; Signal processing.

1. Introduction

Recent technological advances in automatic data acquisition have created a tremendous opportunity for companies to access valuable production information for improving their operation quality and efficiency. Signal processing and data mining techniques are more popular than ever in fields such as sensor technology and intelligent manufacturing. As data sets increase in size, exploration, manipulation, and analysis become more complicated and resource consuming. Figure 1 presents an example of data taken from Nortel's wireless antenna manufacturing processes. There are more

than 30,000 data points in one antenna data set with complicated patterns. Timely synthesized information was needed for product design validation, process trouble shooting and production quality improvement. However, the local changes in the cusps and lobes of the data were difficult to handle for traditional data analysis tools. This motivates the focus of this article: *developing general-purpose data-reduction procedures for commonly used data analysis tools to be useful in handling large-size complicated functional data.* See Ganesan, Das, Sikdar and Kumar (2003) for another motivating example from a nano-manufacturing process.

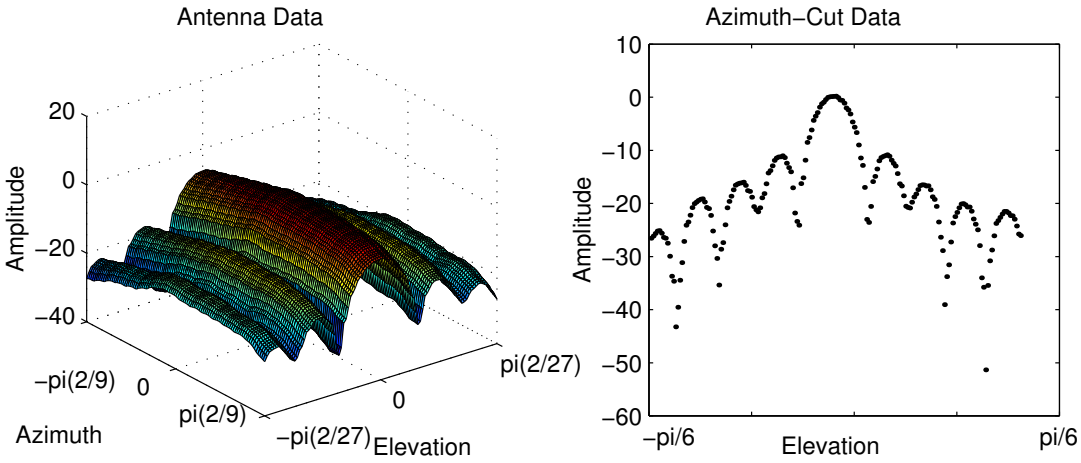

Fig. 1. Data Signals from Antenna Manufacturing Processes

Several data-reduction procedures are available in the literature. Lu (2001) summarized them into three main categories: sampling approaches, modeling and transformation techniques, and data splitting methods. Even with these methods, it is recognized that complicated functional or spatial data with nonstationary, correlated or dynamically changing patterns contributed from potential process faults are difficult to handle. Wavelet transforms model irregular data patterns such as lobes in Figure 1 better than the Fourier transform and standard statistical procedures (e.g., splines and polynomial regressions) and provide a multi-resolution approximation to the data (Mallat, 1998). Applications of wavelet-based procedures in solving manufacturing problems include: using tonnage signals to detect faults in a sheet-metal stamping process (Jin and Shi, 1999); analyzing different catalyst recycling rates to diagnose failures in a residual fluid catalytic cracking process (Wang, Chen, Yang, and McGreavy, 1999); and processing quadrupole mass spectrometry (QMS) samples of a rapid thermal chemical vapor deposition (RTCVD) process to detect significant deviations from the nominal processes (Lada, Lu and Wilson, 2002).

Using expert knowledge of a particular process, one could derive a "feature-preserving" procedure (Jin and Shi, 1999) to extract a particular data pattern represented by a few "features." Then, link these features to a specific type of process fault for monitoring production performance. More rigorously, if the "reduced-size data set" consisting of these features is constructed to detect specific types of known faults, a data-reduction procedure could be derived to minimize Type-I and/or -II errors in hypothesis testing of the occurrence of faults. For example, Jin and Shi's (2001) optimal number of wavelet coefficients used in the fault classification is based on the minimization of probabilities of misclassification errors using SPC limits as the decision rule. However, the wavelet coefficients selected for a given decision rule might not be suitable for other purposes of analysis (e.g., failure prediction, analysis of variance, and clustering analysis to improve manufacturing quality and efficiency). The aim of our data-reduction is to produce a small set of "representative data" suitable for various data and decision analyses either planned or unplanned before seeing the data.

Data-denoising procedures such as *VisuShrink* (Donoho and Johnstone, 1994) and *RiskShrink* (Donoho and Johnstone, 1995) are used as data-reduction tools in a wide range of applications (e.g., Jin and Shi, 2001; Ganesan *et al.*, 2003). See Section 3.2 for details. The following describes another method. Rying, Gyurcsik, Lu, Bilbro, Parsons, and Sorrell (1997) applied a scale-dependent energy metric, E_s = sum of squares of all wavelet coefficients (see Section 2 for a brief overview of wavelets) at atoms $\psi_{s,u}$ across all u positions at the same scale s, to the Ar^+ signals in a semiconductor fabrication experiment. The scalogram (Vidakovic 1999, page 289; see Figure 11 for an example) plots these energy metrics at different resolution scales for visualizing the data-energy distribution. These energy metrics serve as representative reduced-size data so that procedures such as linear discriminant analysis can detect and distinguish process faults in a timely manner.

The purposes of data-denoising and data-reduction are different. Data in engineering applications (e.g., Figures 1, 4 and 7(a)) do not have large-size random noises for showing the effectiveness of data-denoising procedures. Section 3 (e.g., Tables I-IV) uses simulations and real-life examples to illustrate that the ability of data-denoising procedures in data-reduction is limited. On the other hand, the energy-metric approach is too aggressive and not linked to local data characteristics. For example, any functional curve with 1,024 data points will have the same six E_s-measures. This

article develops a well motivated objective function for selecting the reduced-size data, derives the "thresholding parameter" to optimize the objective function, and evaluates the properties of the data-reduction procedures with several simulation experiments and real-life data analyses.

A background of wavelet-transforms is provided in Section 2. Section 3 presents details of the data-reduction methods. Section 4 conducts various comparisons between the proposed methods and extensions of existing methods. Section 5 gives examples of using the reduced-size data in decison-making analyses. A few concluding remarks and future studies are offered in Section 6.

2. Wavelet Transforms

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. Scaling function $\phi(t) \in L^2(\mathbb{R})$ is defined similarly, but $\int_{\mathbb{R}} \phi(t)\, dt \neq 0$.

Select 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. In practice, the following orthonormal basis of wavelet is used to represent a signal function $f(t) \in L^2(\mathbb{R})$.

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

where $\mathbb{Z}$ denotes 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, $d_{j,k} = \int_{\mathbb{R}} f(t)\psi_{j,k}(t)\, dt$ are viewed as the finer-level coefficients describing (local) details of data patterns, $J > L$ and L corresponds 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

$$y(t) = f(t) + \epsilon_t \tag{2}$$

at equally spaced discrete time points $t = t_i$'s, where ϵ_{t_i}'s are random normal $N(0, \sigma^2)$ 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 Eq. (1), $\mathbf{d} = (\mathbf{c}_L, \mathbf{d}_L, \mathbf{d}_{L+1}, \cdots, \mathbf{d}_J)$, where $\mathbf{c}_L = (c_{L,0}, \cdots, c_{L,2^L-1})$, $\mathbf{d}_L = (d_{L,0}, \cdots, d_{L,2^L-1}), \cdots, \mathbf{d}_J = (d_{J,0}, \cdots, d_{J,2^J-1})$. Using the inverse DWT, the $N \times 1$ vector $\mathbf{y}$ from the original signal curve can be "reconstructed" as $\mathbf{y} = \boldsymbol{W}'\mathbf{d}$. The process of applying the DWT to transform a data set closely resembles the process of computing the Fast Fourier Transformation (FFT).

The computational efficiency of DWT is better than the other transforms. For example, the principal component analysis (PCA) requires solving an eigenvalue system which is an expensive $O(N^3)$ operation. The FFT requires $O(N \log N)$ operations, but a fast wavelet transform only requires $O(N)$ operations. As an example, apply the data-reduction method (e.g., RRE_h) developed in Section 3.3 to a very complicated nonstationary data pattern of 1,204 data points (see Figure 8) with programs written in Matlab using a Pentium III personal computer. The total amount of time for DWT and wavelet-coefficients selection is about one second.

Finally, the process fault patterns, which are frequency or phase shifted, are invisible to time domain control limits. They can be easily detected by the wavelet transforms. Thus, wavelet transforms could be very useful in on-line process monitoring (Koh, Shi, Williams and Ni, 1999).

3. Data-Compression, -Denoising and -Reduction Methods

In order to see the difference between the proposed and exisiting methods, the following sections briefly review the background of all methods. Section 4 presents comparison details.

3.1 Signal Approximation and Data Compression Methods

In the signal processing field, the linear approximation method (see Mallat (1998, Section 9.1) for details) uses the function $\boldsymbol{f}_M = \sum_{m=0}^{M-1} < \boldsymbol{f}, \boldsymbol{g}_m > \boldsymbol{g}_m$ with a set of pre-determined vectors $\boldsymbol{g}_m$, $m = 0, 1, \ldots, M-1$, to reconstruct the original data signals, where $< \boldsymbol{f}, \boldsymbol{g}_m >$ is the inner product of the function $\boldsymbol{f}$ and the projected vector $\boldsymbol{g}_m$. In the wavelet-based approximation, $< \boldsymbol{f}, \boldsymbol{g}_m >$ is the wavelet coefficient (from the coarsest level to the finest level in the linear method).

The nonlinear approximation method (Mallat (1998, Section 9.2)) selects the M projection vectors adaptively (e.g., M-largest wavelet coefficients (in absolute values)) using the data signal information to improve the approximation error. In both linear and nonlinear approximation methods, M is fixed by the decision-maker, or by the pre-determined error bound (e.g., $\epsilon(M) = \sum_{i=1}^{N}[f(t_i) - f_M(t_i)]^2/N$). The wavelet coefficients selected from the above approximation methods are usually treated as "compressed data" for reconstructing the original data signals. In this article, they are treated as "reduced-size" data in decision-making analyses.

There were limited studies in the literature for deciding the number of vectors (M) used in the model $\boldsymbol{f}_M$ adaptively based on signal characteristics. The following presents $AMDL$ (Approximate Minimum Description Length) method proposed by Saito (1994). The $AMDL$ selects M to minimize the following objective function:

$$\text{AMDL}(M) = 1.5M \log_2 N + 0.5N \log_2 \left[\sum_{i=1}^{N} (y_i - \widehat{y}_{i,M})^2 \right],$$

where $\widehat{y}_{i,M}$ is the approximation model similar to Eq. (1) constructed from the M largest-magnitude wavelet coefficients and the data y_i is $y(t)$ evaluated at $t = t_i$ from the model (2). As addressed in Antoniadis, Gijbels and Grégoire (1997), the AMDL(M) function is similar to the Akaike information quantity commonly used in statistical model selection procedures including linear regression models. There are several similar model selection methods in the signal processing literature based on objective functions related to quantities defined in "information theory" (e.g., entropy or mutual information (see Ihara (1993); Liu and Ling (1999) for examples)).

3.2 Data Denoising: Wavelet Shrinkage Methods

Data-denoising methods are developed based on statistical models. Specifically, applying the DWT $\mathbf{d} = \boldsymbol{W}\boldsymbol{y}$ to the data $\boldsymbol{y}$ generated from the model (2), we obtain

$$\mathbf{d} = \boldsymbol{\theta} + \boldsymbol{\eta}, \tag{3}$$

where $\mathbf{d}, \boldsymbol{\theta}$ and $\boldsymbol{\eta}$ represent the collections of all coefficients, parameters and errors, transformed from the data $y(t_i)$, the true function $f(t_i)$ and the error $\epsilon(t_i)$ in the time-domain, 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).

Donoho and Johnstone (1995) developed several wavelet-based "shrinkage" techniques to find a smooth estimate ($\hat{\boldsymbol{f}}$) of $\boldsymbol{f}$ from the "noisy" data, $\boldsymbol{y}$. In particular, their hard-thresholding policy finds the estimate of θ_i to minimize the objective function

$$\sum_{i=1}^{N}(d_i - \theta_i)^2 + \tau^2 \sum_{i=1}^{N} |\theta_i|_0, \tag{4}$$

where $\sum_{i=1}^{N} |\theta_i|_0$ is the number of non-zero coefficients selected to estimate the underlying function $\boldsymbol{f}$ (using $\hat{\boldsymbol{f}} = \boldsymbol{W}^{-1}\hat{\boldsymbol{\theta}}$). The optimal estimate $\hat{\theta}_i$ is found to be equal to d_i if $|d_i| > \tau$; otherwise, $\hat{\theta}_i = 0$. Although the parameter τ was not set as the threshold originally, it becomes the threshold in the estimate of θ_i through the minimization process.

Because smaller coefficients are usually contributed from data noises, thresholding out these coefficients has an effect of "removing data noises." Thus the shrinkage methods are called data-denoising methods. The *VisuShrink* (Donoho and Johnstone, 1994), *RiskShrink* (Donoho and Johnstone, 1995) and *SURE* (Donoho and Johnstone, 1995) are three popular thresholding methods commonly used in practice. They represent different ways to find the optimal choice of the threshold τ based on another set of criteria. For example, *RiskShrink* minimizes a theoretical upper bound of the asymptotic risk to find τ. See Donoho and Johnstone (1994, 1995) for details. These data-denoising methods will be used in Section 4 for comparison studies.

Shrinkage methods require an estimate of the standard deviation σ for calculating the threshold value (e.g., $VisuShrink$'s threshold is $(2\ln N)^{1/2}\sigma$). Different estimates of σ will lead to distinct thresholds and different number of wavelet coefficients. This article uses a robust estimate, $\hat{\sigma} = median(|d_{J,k}| : 1 \le k \le N/2)/0.6745$, suggested by Donoho and Johnstone (1994), where J is the finest resolution level. The next section proposes two new data-reduction methods which do not require the estimation of σ.

3.3 Data-Reduction Methods - RRE_h and RRE_s

All data-denoising, $AMDL$ and nonlinear signal approximation methods retain the largest M_λ number of coefficients based on some derivations of the threshold λ. Our methods will also follow this principle by assuming that large wavelet coefficients will better characterize signal patterns in terms of their energy and thus retain more information.

Definition 1. The energy of a finite sequence $\boldsymbol{f} = (f_1, \cdots, f_N)$ is defined by $\xi = ||\boldsymbol{f}||^2$. Correspondingly, the empirical estimate of the energy of a data signal is $\hat{\xi} = ||\mathbf{y}||^2 = ||\mathbf{d}||^2$.

The following theorem gives an upper bound of the approximation (or estimation) error using the largest M wavelet coefficients. These errors represent the "reconstruction error" in our data-reduction methods.

Theorem 1. *For $\boldsymbol{f} \in L^2(\mathbb{R})$, an upper bound of the approximation error for $\boldsymbol{f}_M$, is $\|\boldsymbol{f} - \boldsymbol{f}_M\|^2 \leq [(N-M)/M]\ \xi$, and an upper bound of the estimation error for $\hat{\boldsymbol{f}}_M$ is $E\left\|\boldsymbol{y} - \hat{\boldsymbol{f}}_M\right\|^2 \leq [(N-M)/M]\ E(\hat{\xi})$.*

Data-reduction and -denoising methods are distinct for different purposes. As seen in Eq. (4), data-denoising procedures aim to find the estimate $\hat{\boldsymbol{\theta}}$ (and $\hat{\boldsymbol{f}}$) for reducing "modeling error" of $\boldsymbol{\theta}$ (and $\boldsymbol{f}$). The data-denoising methods are therefore more aggressive in reducing the modeling errors. Conversely, data-reduction methods select the "reduced-size" data with a more aggressive data-reduction ratio. However, the selected reduced-size data should be representative enough in capturing key data characteristics for subsequent planned or unplanned decision analyses. Theorem 2 below shows that our data-reduction methods depend on the "data energy" representing data characteristics instead of the variance (σ^2) representing data noises in the data-deniosing procedures.

The following data-reduction criterion is developed for balancing two ratios: (1) the relative data-energy in the approximation model and (2) the relative number of coefficients used (i.e., the data-reduction ratio).

$$RRE_h(\lambda) = \frac{\mathrm{E}||\boldsymbol{d} - \hat{\boldsymbol{d}}_h(\lambda)||^2}{\mathrm{E}||\boldsymbol{d}||^2} + \omega \frac{\mathrm{E}||\hat{\boldsymbol{d}}_h(\lambda)||_0}{N}, \tag{5}$$

where $||\hat{\boldsymbol{d}}_h(\lambda)||_0 = \sum_{i=1}^{N} |\hat{d}_{h,i}(\lambda)|_0$ is the number of coefficients selected, and $|\hat{d}_{h,i}(\lambda)|_0 = 1$, if $\hat{d}_{h,i}(\lambda) \neq 0$; $|\hat{d}_{h,i}(\lambda)|_0 = 0$, otherwise. Theorem 2 finds the optimal λ to minimize Eq. (5).

The use of "normalizing constants" to make the two balancing terms compatible is critical. See Table II of empirical studies for understanding its impact. The weighting parameter ω is user-selected or provided by methods such as generalized cross-validation (GCV) method (Weyrich and Warhola, 1998). However, results from Weyrich and Warhola (1998) illustrate the need for further studies for developing the GCV-like selection of ω in our problem and understanding its properties.

For simplicity, this article will use $\omega = 1$, which places equal weights in both components in follow-up studies. The following uses engineering and statistical experience to motivate the objective function (5). Our discussion will focus on the *hard-thresholding-based method* RRE_h. A similarly motivated method RRE_s based on the soft-thresholding policy is presented in the Appendix.

In engineering applications such as Mallat (1998, pages 378-391), the "relative error,"

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

is commonly used in comparing signal approximation quality. This is similar to the first term in Eq. (5). This article utilizes a thresholding parameter λ to decide which wavelet-domain data to keep and which to discard in decision-making analyses using the terms $\hat{d}_{h,i}(\lambda) = I(|d_i| > \lambda)d_i$, $i = 1, \ldots, N$. Ideally, only a small portion of the data is kept to meet the data-reduction goal. This is quite different from the data-denoising procedure where the parameter τ was not set as the threshold originally for the data-reduction purpose in the construction of the objective function (4). Recall that in the discussion under Eq. (4) that the denoising procedures aimed to estimate θ_i's. Their threshold τ for the estimate $\hat{\theta}_i$ is decided from *another set of criteria* such as minimizing a theoretical upper bound of the asymptotic risk.

Eq. (5)'s second component serves as a penalty term for limiting the size of data used in follow-up decision analyses. 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, like the data-denoising method of finding estimate $\hat{\theta}$, ridge regression finds the optimal estimate of regression coefficients to minimize 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 the one in Eq. (5). Note that this objective function is not normalized as was done in Eq. (5). More importantly, ridge regression does not use a threshold to select which data to keep in follow-up decision analyses.

The following presents a few analytical properties of the proposed data-reduction method. The closed-form solution of the optimization of Eq. (5) becomes handy in practical implementations. See the Appendix for the proof of the theorem.

Theorem 2. *Consider the model stated in (3). Then, we have*

(i) the objective function $RRE_h(\lambda)$ is minimized uniquely at $\lambda = \lambda_{N,h}$ where

$$\lambda_{N,h} = (\frac{1}{N}\mathrm{E}||\boldsymbol{d}||^2)^{1/2}; \tag{6}$$

The moment estimate of $\lambda_{N,h}$,

$$\hat{\lambda}_{N,h} = \left(\frac{1}{N}\sum_{i=1}^{N} d_i^2\right)^{1/2} = (\frac{\hat{\xi}}{N})^{\frac{1}{2}}, \tag{7}$$

(ii) $(\hat{\lambda}_{N,h} - \lambda_{N,h}) \xrightarrow{w.p.1} 0$;

(iii) $\sqrt{N}(\hat{\lambda}_{N,h} - \lambda_{N,h})/\sigma^*_{N,h} \xrightarrow{d} N(0,1)$, *where*

$$(\sigma^*_{N,h})^2 = \frac{1}{4N}\left(\frac{4\sigma^2\sum_{i=1}^{N}\theta_i^2 + 2N\sigma^4}{\sum_{i=1}^{N}\theta_i^2 + \sigma^2}\right).$$

Consider a few well-known testing signal curves with 1,024 data points in each curve (see Figure 2 for their "normalized" forms (in the same scale and zero mean)) taken from the literature (e.g., Donoho and Johnstone, 1995). Table I shows the relationship between the energy value of signals and the number (M) of wavelet-domain data selected. Note that our methods normalize the signal to have zero mean and apply the thresholding rules to all resolution levels of the wavelet coefficients while the denoising techniques do not threshold the coefficients in the coarser level ($c_{L,k}$'s; L in Eq. (1) is pre-selected, e.g., $L = 4$ for $N = 1024$; Donoho and Johnstone, 1995).

Based on the observation from Table I, in general, if the signal has a larger value of energy, its threshold value will be higher (see the threshold values for RRE_h (and RRE_s) for examples), and it is more likely to have a smaller M. There are some exceptions. For example, if most of the signal energy is kept in a few larger wavelet coefficients, the signal has a set of very "unbalanced" wavelet coefficients. When there are a larger number of smaller coefficients, the number of thresholded coefficients is smaller. This leads to a smaller M. For example, the threshold values $\hat{\lambda}_h$ in Nason and Heavisine signals are very close, but the energy for the Heavisine is slightly more unbalanced. This leads to a slightly smaller M in RRE_h for the Heavisine signal. See Vidakovic (2000) for a technique to compare signals with different unbalancing characteristics.

TABLE I

Results of Data Reduction for Testing Signals

Signals	Energy	Threshold value		M = #Coefficients Selected					
		$\hat{\lambda}_h$	$\hat{\lambda}_s$	RRE_h	RRE_s	$Visu$	$Risk$	$SURE$	$AMDL$
Nason	94.25	0.3034	0.6986	31	138	192	225	324	192
Heavisine	90.28	0.2969	0.6803	28	143	287	290	292	194
Blocks	72.36	0.2658	0.5099	67	379	389	407	518	391
Bumps	17.63	0.1312	0.3401	91	405	646	664	722	894

Table II presents the impact of not using the normalizing constants in Eq. (5), denoted as ${RRE_h}^*$, where $SNR^* = std(\boldsymbol{f})/\sigma$ represents the noise level of data, $std(\boldsymbol{f})$ is the standard deviation of the discretized signal points, and σ is the standard deviation of noise. Smaller SNR^* means that the data is noiser. Note that RRE_h in Table II is the sum of the first two columns, relative error and M/N, representing the metric defined in (5). Without the normalization the ${RRE_h}^*$ procedure has very poor data-reduction ratio for all cases studied, and its performance is similar to the use of data-denoising methods for the data-reduction purpose. That is, it over-emphasizes on reducing the modeling error by sacrifying their data-reduction ability. The relative errors of ${RRE_h}^*$ are very small with plots similar to Figures 3 to 6 produced by data-denoising methods (see Tables III and IV for details).

4. Comparisons of Data Reduction Methods

Although methods described in Sections 3.1 and 3.2 were not developed for data-reduction purposes, practitioners did use them for selecting "reduced-size" data to perform various decision analyses. This section will compare all six methods presented in Section 3 in terms of their modeling error and data-reduction ability. The data patterns for comparisons include two real-life data curves (Figures 4 and 5) and four well-known testing signals from the wavelet literature (Figure 2). The four "noise-free" testing signals characterize different types of important features arising in imaging, seismography, manufacturing and other engineering fields. The symmlet-8 wavelet family is used in wavelet transforms for all cases.

TABLE II

IMPACTS OF NORMALIZATION FOR DATA REDUCTION

Signals	With Normalization			Without Normalization		
	Relative error	M/N	RRE_h	Relative error	M/N	RRE_h
Bumps ($SNR^* = \infty$)	2.18E-02	0.090	0.112	2.81E-19	0.770	0.770
Bumps ($SNR^* = 15$)	2.94E-02	0.066	0.096	6.18E-04	0.456	0.456
Bumps ($SNR^* = 7$)	3.97E-02	0.066	0.106	2.98E-03	0.432	0.435
Bumps ($SNR^* = 3$)	9.45E-02	0.066	0.161	1.60E-02	0.395	0.411
RTCVD	1.77E-02	0.130	0.147	8.89E-07	0.578	0.578
Antenna	4.25E-02	0.180	0.222	3.27E-05	0.644	0.644

Tables III-V present the comparison results with the following summary measures: (1) Reduction ratio (%) : $RR = (1 - M/N) \times 100$; (2) $RelErr = \|\boldsymbol{f} - \hat{\boldsymbol{f}}_M\|/\|\boldsymbol{f}\|$ for the case without random errors and $RelErr = \|\boldsymbol{y} - \hat{\boldsymbol{f}}_M\|/\|\boldsymbol{y}\|$ for the case with random errors; and (3) $AMDL$-measure.

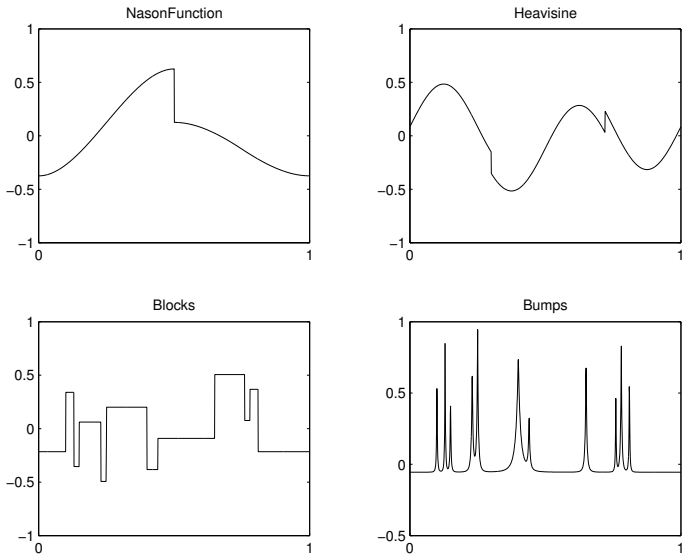

Fig. 2. Four Testing Signals from the Literature

Figure 3 shows the results for the bumps-signal. $VisuShrink$, $RiskShrink$, $SURE$ and AMDL(M) procedures achieve very small modeling errors (see Table III for the very small $RelErr$ in the 10^{-16} level). RRE_s did as well as the others when relative errors are compared. RRE_h missed some

details in the smoother signal between peaks. However, all the shapes and locations of the 11 peaks were identified and modeled well by the more aggressive RRE_h method, which has a 90% data-reduction ratio as opposed to the 60% in RRE_s and below 40% in all other methods. Note that the values for $AMDL$-measure are quite different from data-reduction and -denoising measures. Although the $RelErr$ in $SURE$ is the second best, its $AMDL$-measure is much worse than $VisuShrink$, $RiskShrink$ and even RRE_s methods. It is interesting to note that though $SURE$ and AMDL(M) methods have similar $RelErr$ and data-reduction ratios, their $AMDL$-measures are very different. Thus, AMDL(M) and our RRE_h and RRE_s methods work very differently for these curves.

Similar conclusions were observed for several other testing signals (not shown here). Examples from Section 5 show that RRE_h and RRE_s methods did give accurate decision results even with a more aggressive data-reduction emphasis. The following examples test if the proposed methods work well in the two real-life data sets where errors were involved. See Remark #1 for the studies of noisy bumps signals.

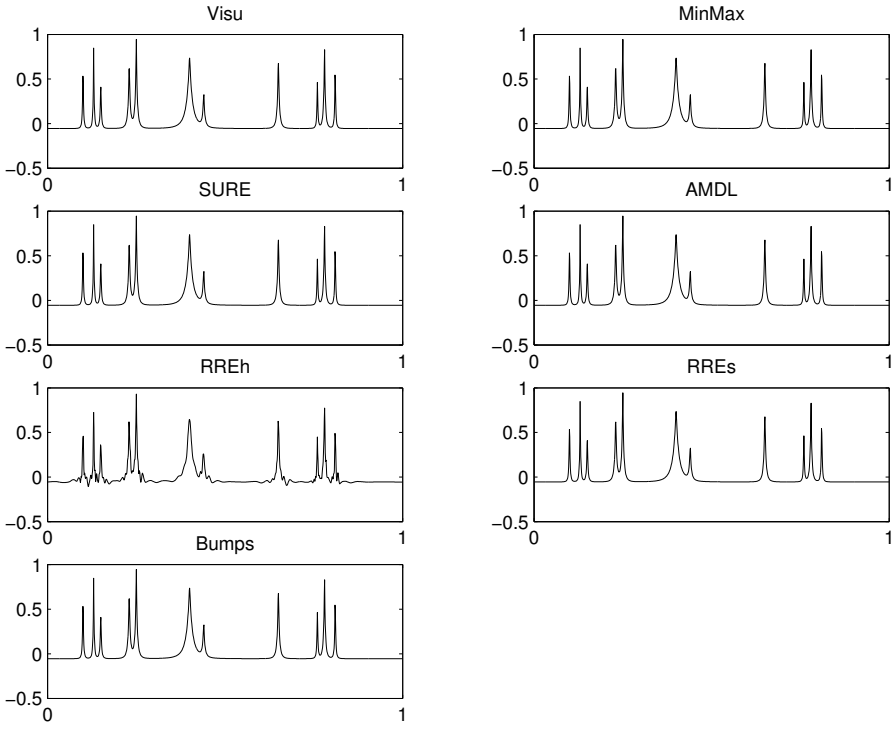

Fig. 3. Reconstruction of the "Noise-free" Bumps Signal

TABLE III

Results for the Noise-free Bumps Signal

Method	M	$RelErr$	RR	$AMDL$
$VisuShrink$	646	$1.50E-16$	36%	16390.6
$RiskShrink$	664	$1.23E-18$	35%	13108.3
$SURE$	722	$2.22E-21$	29%	26321.8
$AMDL$	894	$3.91E-25$	13%	5506.6
RRE_h	91	$2.18E-02$	91%	32151.2
RRE_s	405	$1.51E-09$	60%	24682.6

Example 4.1 (RTCVD Data). The RTCVD process deposits thin films on the wafer by a temperature driven surface chemical reaction. As feature size decreases, functional operation of semiconductors (e.g., transistors) becomes increasingly unreliable due to variations of deposition processes. Therefore, controlling the process variability is critical. Quadrupole mass spectrometry is commonly used in semiconductor manufacturing processes for monitoring thin-film deposition quality. The data shown in Figure 4 is one of the several nominal RTCVD process runs in a research project (Rying, 2001) for developing a new measurement technique for online process monitoring. Although there are only 128 data points in the curve, and the data change-pattern is not very complicated, this case study serves as a basis for developing process monitoring and fault detection/classification tools applicable in various engineering applications. See Section 5.2 for more details. More importantly, wavelet transforms are proven to be useful in locating change-points (e.g., the two peaks) for developing an integrated metric essential for the new measurement technique. See Rying (2001) for details.

Results in Figure 4 and Table IV show that the RRE_h could be too aggressive in data reduction (87% ratio) due to its non-smoothing fit in the straight rising component (data between 20 to 30 points). However, it did roughly pick up the two peaks and other change-points. The AMDL(M) did a much better job in balancing the data-reduction ratio and the modeling error in this case. The errors of the three data-denoising methods are smaller, but the reduction ratios are lower. It is difficult to distinguish these small amount of modeling errors in the plots by visual inspection.

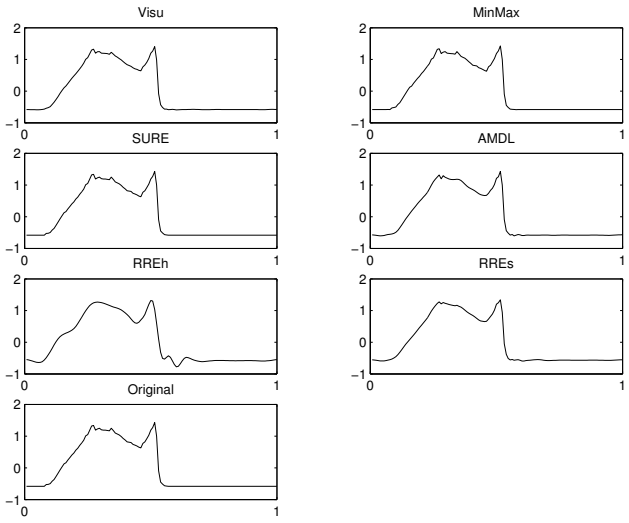

Fig. 4. Reconstruction of the RTCVD Signal

Example 4.2 (Antenna Data). The increasing popularity of wireless communication has produced an increasing demand for high quality antenna equipment. Eighteen sets of antenna data like Figure 1 were collected at Nortel for developing a procedure to monitor antenna manufacturing quality. Figure 5 shows the reconstructed antenna curves based on various data-reduction methods. Excluding the RRE_h method, all methods model the complicated peak and valley patterns very well. The RRE_h provides a reasonable fitting other than the valleys between the second and the third peaks from the main lobe in the middle. Surprisingly, the AMDL(M) has an excellent data-reduction ratio (81%) as good as the RRE_h. See Table IV for details.

Remarks and Discussions:

1. We also test the robustness of the above data-reduction methods against random noises. In a series of experiments, various amount of random normal noises were added to the testing signals. Figure 6 shows the noisy bumps with different values of SNR^*'s. Table V summarizes model fitting and data-reduction results from all methods in the cases of $SNR^* = 3$, $SNR^* = 7$, and $SNR^* = 15$. Smaller SNR^* means a noisier signal. For the signals with larger SNR^* (less noisy), the noise level (σ) is lower and the threshold value should be lower (e.g., the threshold value of $VisuShrink$ is $(2\ln N)^{\frac{1}{2}}\sigma$). This leads to a larger number of selected coefficients. For this reason the denoising methods are less effective in data-reduction and use a larger number of wavelet coefficients in the model. See the drops of data-reduction ratio for $SURE$ in Table V from

TABLE IV

Results for the RTCVD and Antenna Data

Method	RTCVD		Antenna	
	RR	$RelErr$	RR	$RelErr$
$VisuShrink$	50%	$9.92E-05$	59%	$1.70E-03$
$RiskShrink$	46%	$2.37E-06$	45%	$1.07E-04$
$SURE$	36%	$8.69E-08$	27%	$1.46E-05$
$AMDL$	75%	$5.35E-04$	81%	$7.47E-03$
RRE_h	87%	$1.77E-02$	82%	$4.25E-02$
RRE_s	68%	$2.27E-03$	67%	$5.55E-03$

TABLE V

Results for the Noisy Bumps Signal

Method	$SNR^* = \infty$		$SNR^* = 15$		$SNR^* = 7$		$SNR^* = 3$	
	RR	$RelErr$	RR	$RelErr$	RR	$RelErr$	RR	$RelErr$
$Visu$	36%	1.50E-16	85%	1.12E-02	88%	4.18E-02	91%	1.54E-01
$Risk$	35%	1.23E-18	78%	2.52E-03	83%	1.21E-02	86%	6.24E-02
$SURE$	29%	2.22E-21	54%	8.00E-04	70%	8.42E-03	78%	4.91E-02
$AMDL$	13%	3.91E-25	87%	6.37E-03	90%	2.39E-02	95%	1.36E-01
RRE_h	91%	2.18E-02	93%	3.00E-02	93%	4.00E-02	93%	9.45E-02
RRE_s	60%	1.51E-09	85%	1.17E-02	88%	3.94E-02	76%	7.63E-02

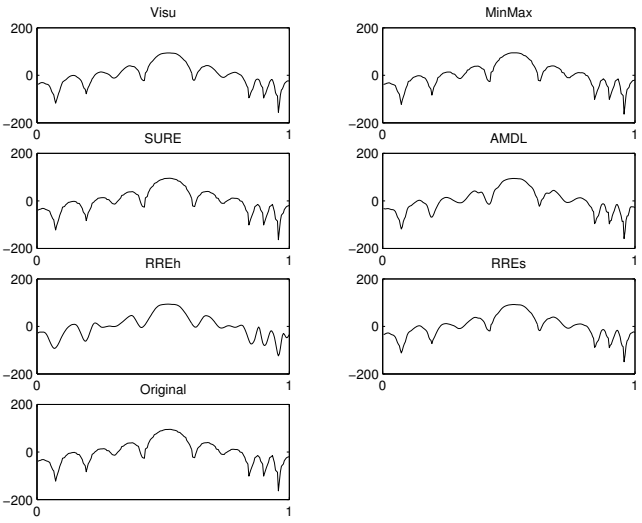

Fig. 5. Reconstruction of the Antenna Data

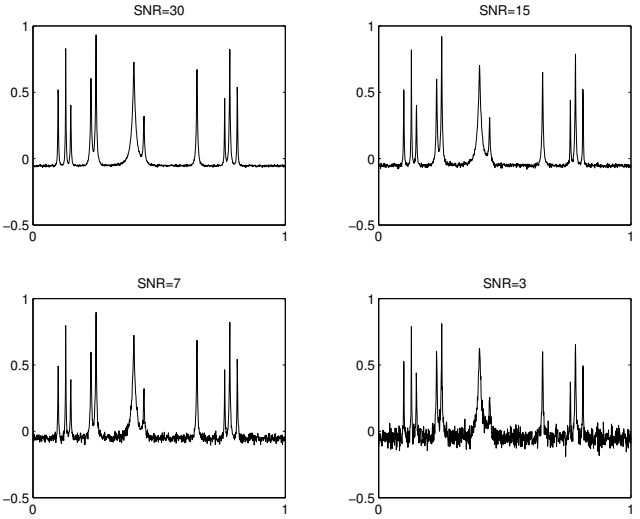

Fig. 6. Noisy Bumps Signal at Various Noise Levels

$SNR^* = 3$ to $SNR^* = \infty$ cases for a specific example. With noisy data, the difference in modeling errors from these six methods is smaller than the difference in the case without added noises where SNR^* is equal to ∞. The reduction ratio stays the same for the RRE_h but improved considerably for all other methods. However, they pay a price to have much larger modeling errors (see Table V) as compared with the results given in Table III. Surprisingly, the modeling errors from $VisuShrink$ and AMDL(M) methods in the case for $SNR^* = 3$ (the most noisy case studied) are larger than the errors in the proposed RRE_h and RRE_s methods.

2. In engineering applications such as Lada *et al.* (2002), replicated signal curves exhibit patterns as shown in Figure 7(a) from the RTCVD experiment. This type of process variation could be easily experienced from the example of circle signals from x-ray image of products. With a certain amount of process variations, the resulting circles could have different radii and distinct centers, but they are all similar circles. This type of process variation is quite different to the data noise generated from model (2), where normal random noise is added to a deterministic functional curve. See Figure 7(b) for an example. Thus, in the decision-tree evaluation experiment (presented in Section 5.1), the replicates of data curves will be generated from "engineering variations." In addition, statistical normal random noises are added. Figures 7(c) and (d) show one example of the original and the replicated curve from the data generation procedure.

3. In deciding which wavelet family is most suitable for representing a data signal, the more "disbalancing" type (more separation in the larger and smaller wavelet coefficients) of wavelet family that is used, the more efficient the data-reduction will be. Because "symmlet-8" showed excellent disbalancing properties on most of the curves studied in our evaluation studies and application examples in Sections 4 and 5, we used it as the "default" choice of the wavelet family in our data-reduction exercises.

In summary, RRE_h, AMDL(M) and RRE_s are more suitable for data-reduction purposes. However, RRE_h could be too aggressive in some cases where certain details are ignored; AMDL(M) is not suitable for signal curves "without noise," (e.g., results given in Table III). $VisuShrink$, $RiskShrink$ and $SURE$ are not very effective in data reduction but their modeling quality is excellent. When larger amounts of normal random noises are added to the deterministic signal curves, the difference between these six methods in their modeling quality and data-reduction ratio becomes smaller. This could be due to the fact that all methods performed worse in modeling the data with more noise. The next section further examines the effectiveness of the data-reduction methods with various decision rules.

5. Illustrations of Decisions Based on Reduced-size Data

This section presents two examples to illustrate the use of selected reduced-size data in decision analyses. Note that there are several difficulties in these illustrations. As addressed in Remark #2

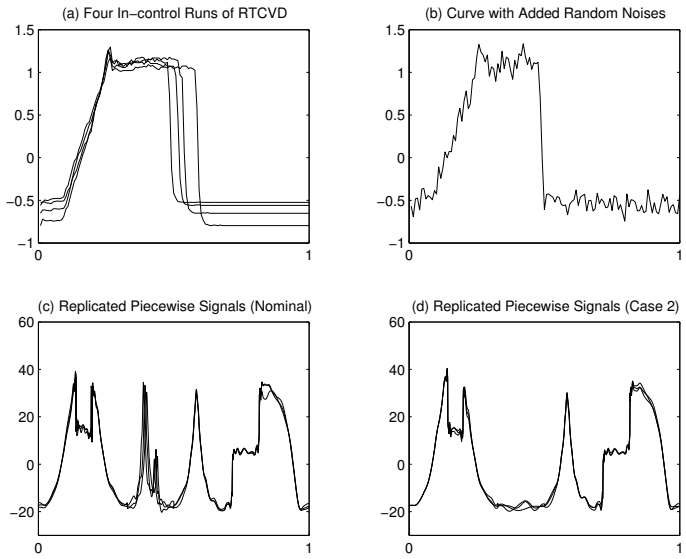

Fig. 7. Different Types of Signal Replications

of Section 4, engineering variations used for generating replicated data curves are quite different from statistical random noises. Learning from the experience in the repeated measurements of biomedical studies, Jung and Lu (2004) proposed a wavelet-based random-effect model. Because the research in this area is relatively recent, this section uses the "shifting method" described in the second paragraph of Section 5.1 for generating replicated data curves. Another major difficulty is the selecting of the reduced-size data (wavelet-coefficients) in the case of multiple curves for classification/clustering analysis or for functional analysis of variance. Note that if a data-reduction method is applied to the multiple curves one curve at a time, the selected wavelet coefficients will be different for distinct curves. Then these curves cannot be studied or compared together due to different wavelet-bases of reduced-size data sets. See Jung and Lu (2004) for a vertical thresholding procedure to tackle this problem. Due to these difficulties, it is premature to compare data-reduction methods in terms of errors in decision rules. Thus this section only illustrates the potential use of selected reduced-size data.

Detecting and classifying process fault types are important in engineering applications. When manufacturing processes become complicated, human operators have difficulty identifying the sources of process problems. Effective use of process data (e.g., control signals and various stages of process performance measurements) in a timely manner could drastically reduce process defects,

production costs or more serious process problems. Section 5.1 shows the possibility of making decisions on process fault types with the classification and regression tree (CART) method. Section 5.2 presents an interesting idea of using wavelet's multi-resolution property to construct a visualization plot for understanding process fault problems.

5.1 Fault Classification Using the CART Method

CART is very popular in data mining applications (e.g., customer relationship management). It is a tree with nodes at various levels organized in a series of hierarchical binary-decisions. Each decision is based on the "cut-off value" of a chosen variable. See Breiman, Friedman, Olshen and Stone (1984) for details of tree-building and pruning procedures.

To evaluate the error rate in applying CART to the reduced-size data for classifying process fault types, various replicated data curves were generated from a very difficult signal pattern (see Figure 8) taken from Mallat (1998; page 378). In our experiment, the entire curve is shifted to the left (or right) in 5 (or 10, 15, 20, 25, 30) time-units (out of a total of $N = 1,024$ units) for generating a new curve with added random $N(0, \sigma^2)$ noises using a small value of σ $(= 0.1)$.

Figure 8 presents seven fault classes of curves. Some of them are considerably more difficult than the others for decision trees to correctly identify fault classes. For example, the only difference between fault class 4 and the original curve is a smaller amount of vertical drop of the first rectangle-shaped dip around 147 to 204 time units. Class 1 could also be considered a difficult case where the first dip is filled smoothly. Three hundreds of replicated curves were generated for each of the eight cases. Thus there are 2,400 data curves totally in this study.

For dealing with multiple classes of replicated data curves, our study uses the union positions of all selected coefficients (obtained from application of the RRE_s method to individual data curves) to create the reduced-size data. Because the RRE_s method has better modeling accuracy than the RRE_h method it is our choice here. Although its data-reduction ratio is not as good as the RRE_h method in general, it does achieve a 91.89% reduction ratio in this example. That is, only 83 out of 1,024 wavelet coefficients are used in CART applications. In the decision analysis, CART is supposed to identify all these fault types based on the reduced-size data.

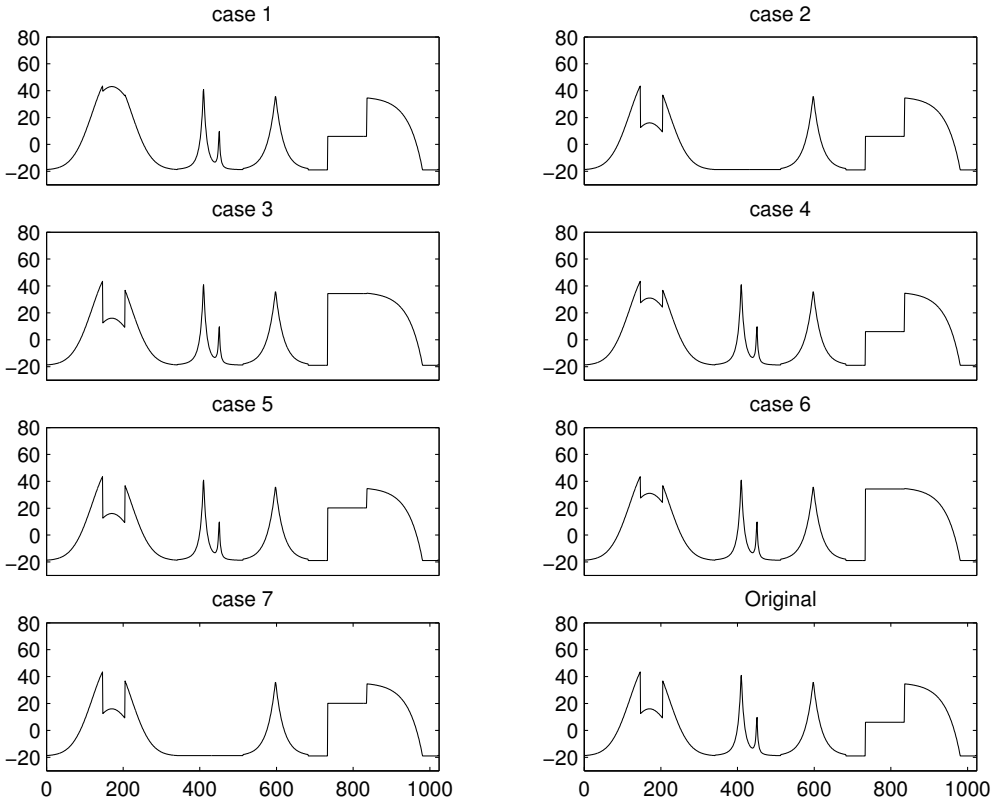

Fig. 8. Mallat's Piecewise Signals

There is no good guideline available on how to divide the 2,400 samples into training and testing data sets. Fukunaga (1990) provided arguments in favor of using more samples for testing than for training the classifier to challenge the classification rules. Therefore, our experiment used 1/3 of the data randomly selected from each case for training and 2/3 data for testing. Figure 9 shows the CART tree constructed using the reduced-size training data. This tree has eight terminal nodes for locating data curves in different classes, nominal or case 1 to 7.

The decision nodes picked by CART for decisions have certain interesting interpretations. The first split is $c_{5,6} \leq -28.967$ where $c_{5,6}$ is the $6th$ position coefficient in the coarsest resolution level. This coefficient covers the support $[161, 192]$ in the time domain, which is somewhere close to the first rectangle-dip. Note that fault class 1 does not have the dip and fault class 4 has a less shallow dip. The coefficient selected for the split at node 2 is $c_{5,17}$. The coefficient $c_{5,17}$ covers the support $[513, 544]$, which is slightly to the right of the middle of the curve. This coefficient presents a possibility of missing the the second and third peaks critical to the fault detection and

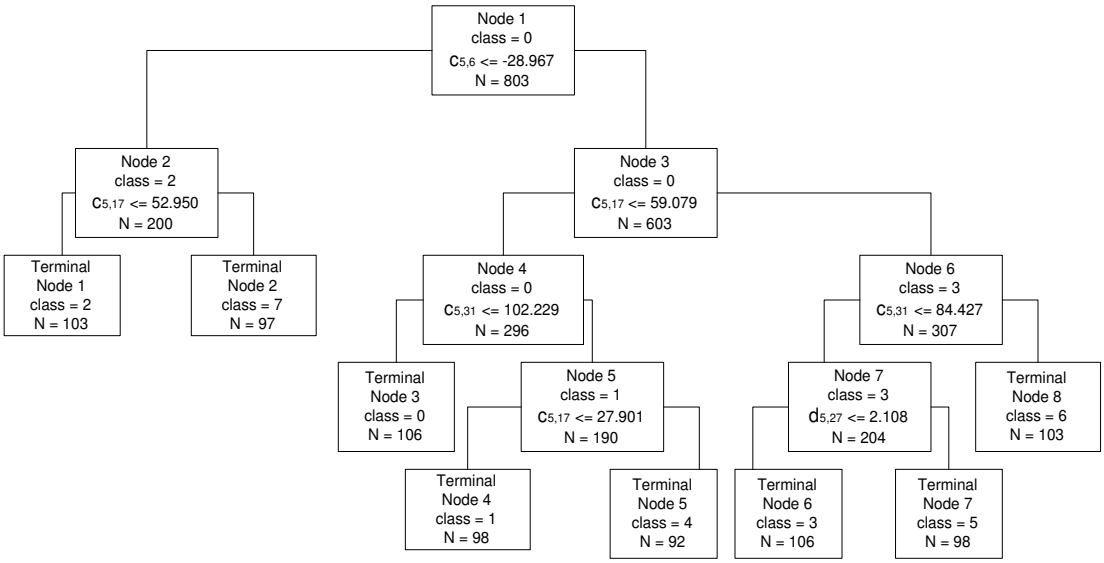

Fig. 9. CART Tree in the Wavelet Domain

classification. Similar interpretation could be obtained for other coefficients selected by CART. In practice, the majority of patterns could be identified by the coefficients at the coarser resolution level while only a few patterns will require information from coefficients at finer levels for decisions (e.g., $d_{5,27}$ of node 7). The use of combinations of coarser- and finer-level coefficients at different hierarchies of CART provides a multi-resolution oriented decision-making opportunity not available in the time domain based on the original data.

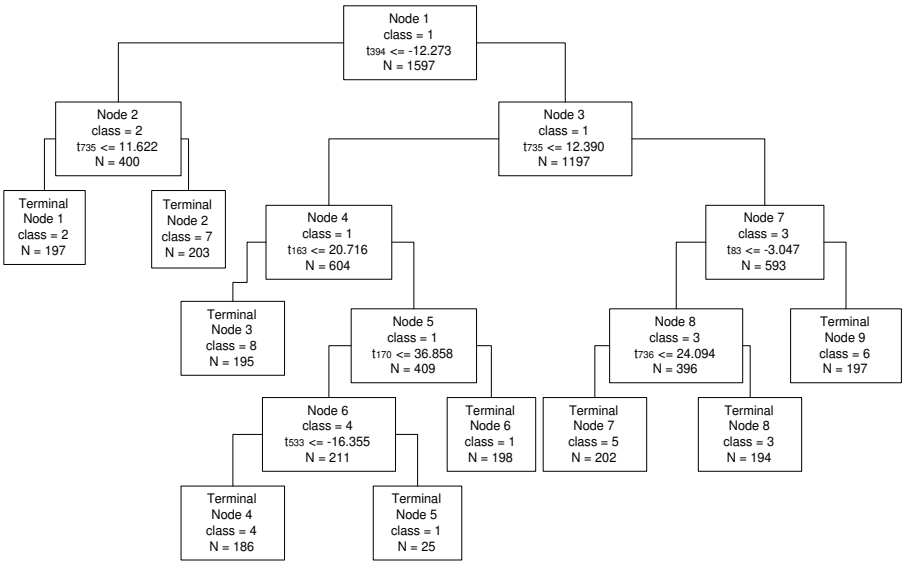

Fig. 10. CART Tree in the Time Domain

TABLE VI

MISCLASSIFICATION ERROR(%)

	Training data		Testing data	
Class	wavelet	time	wavelet	time
original	0.00	0.00	2.06	3.09
1	5.10	4.08	8.42	8.91
2	0.00	0.00	0.51	0.00
3	0.00	0.00	0.00	0.00
4	5.43	3.26	6.25	12.02
5	0.00	0.00	0.00	0.00
6	0.00	0.00	0.00	0.51
7	0.00	0.00	0.49	0.00
total error	1.25	0.87	2.25	3.13

As an illustration for the time saving in using the reduced-size data for decision analyses, Figure 10 shows the CART tree constructed using $N = 1,024$ points in the time domain. The larger size data in the time domain inceased the time needed to construct the decision tree by a factor of ten compared to working with the reduced-size data (55 versus 5 seconds; It took only one second to obtain the reduced-size data set by applying the DWT and the RRE_s method). The interpretation of Figure 10 is somewhat different from the one for Figure 9. In node 1, the first split is $t_{394} \leq -12.283$ where t_{394} is the value of the signal at time 394. In node 2, if $t_{735} \leq 11.622$, then the signal is classified into class 2; otherwise, the signal is classified into class 7. Thus this tree compares the height of the signal at a particular time point rather than the "energy" preserved in the wavelet-coefficients in certain support area as illustrated in Figure 9.

The misclassification rates in the wavelet and time domains and in the training and testing samples are shown in Table VI. The CART tree in the time domain was almost perfect with respect to the training data, but it adapted too much to the features specific to the training data and lost its generalization power. Hence, it did not work well when applied to the testing data. The misclassification rate for the CART built from the reduced-size data is comparable to the

one obtained using the original time domain data in the training samples but is smaller (2.25% versus 3.13%) in the testing samples. The existence of noise in signals makes classification in the time domain difficult. Our RRE_s-based method reduces the data size and removes some noises simultaneously for a more efficient and effective signal classification.

Remark: Our procedures were compared with the principal coordinates approach based on the function data-analytic method proposed in Hall, Poskitt and Presnell (2001). Their method approximates the signal using the first M Karhunen-Loève basis functions with M decided from the cross-validation for minimizing the error in a specific decision method (e.g., the CART classification in our application here). Applied to all eight data signal classes as studied in Section 5.1, CART's total misclassification rates for their and our methods are 2.82% and 2.25%, respectively. Although our data-reduction method RRE_s is not designed for any specific decision method and their method is designed for CART classification, our misclassification error 2.25% is slightly smaller than theirs. This shows the potential of our procedure. Similar observations were obtained from normal-distribution-based quadratic discriminant analysis (QDA) advocated in Hall *et al.* (2001), which has much higher total misclassification rates (about 25% in both methods). Because their method requires more computing effort, is more difficult to interpret the selected coordinates (in the sense of the reduced-size data), and might not be appropriate when the data signal is noisy and the number of replicates is limited (smaller than L), our procedures are more useful in data reduction for various types of decisions.

5.2 Multi-resolution Fault Detection Using Thresholded Scalogram

One deficiency that wavelet-bases inherently possess is the lack of a shift-invariant property. For example, for two "replicated" data curves with a slight shift in time (i.e, perturbation to left/right (e.g., see Figure 7 (a))), when the two signals are decomposed via the DWT we can see appreciable differences between their wavelet coefficients. Direct assessment from a particular wavelet coefficient often leads to inaccurate decisions. For two signals with a slight shift in time, energy metrics E_s at each resolution-scale show no difference between the two signals. That is, the scale-based energy representation provides a more robust (against small shift in time) signal feature for fault detection.

One of the advantages in wavelet transforms is the multi-resolution decomposition of compli-

cated data signals. Information contained in each resolution could be useful in different types of fault detections. For example, the coarser-scale coefficients represent the global shape of the signal in the lower (coarser) resolution level, while the fine-scale coefficients represent the details of the signal in the higher (finer) resolution level. We therefore propose to use the following scalogram (Vidakovic 1999, page 289) for fault detection:

$$S_{d_j} = \sum_{k=0}^{m_j-1} d_{jk}^2, \quad j = L, L+1, \ldots, J,$$

where m_j is the number of wavelet coefficients in the jth resolution level. We use the notation S_{c_L} for the energy at the coarser level (i.e., $S_{c_L} = \sum_{k=0}^{2^L-1} c_{L,k}^2$). Scalogram is a commonly used tool in signal and image processing (Rioul and Vetterli, 1991), astronomy, and meterology studies (see Scargle, 1997 for an example). It measures the signal energy contained in the specific frequency band with a given scale.

For handling potentially large size data and for removing secondary noises, we propose the following "thresholded scalogram":

$$S_{d_j}^*(\hat{\lambda}) = \sum_{k=0}^{m_j-1} I(|d_{jk}| > \hat{\lambda}) d_{jk}^2,$$

where $\hat{\lambda}$ is the threshold value decided (from data) in various methods introduced in Section 3. Similarly, $S_{c_L}^*(\hat{\lambda}) = \sum_{k=0}^{2^L-1} I(|c_{Lk}| > \hat{\lambda}) c_{Lk}^2$. The screening of smaller wavelet coefficients makes the detection of process fault more robust in a noisy environment.

Figure 11 presents a thresholded scalogram plot (in a $\log_2$-scale) of the RTCVD experimental data from three fault classes. See Figure 12 for the data curves obtained from the nominal and three fault classes. Comparably, the scalogram values for the data in the Fault 3 class are much different from the nominal ones at any resolution levels. Due to similarity of data signals in the original time domain, Fault classes 1 and 2 have similar scalogram values in the finer resolution levels d_6 and d_7 but not in the coarser resolution levels c_5 and d_5. Comparing them with the nominal case, Fault class 2 and the nominal curves have similar scalogram-value in (c_5), but not in d_5 and d_6. Possibly due to the sharp drop of the data curve in Fault class 1, its c_5 value is quite different from the nominal one.

Let S_j^* represent the thresholded scalogram, $S_{d_j}^*$ and $S_{c_L}^*$. The following derives the needed (approximated) distribution theorem for constructing a set of "lower and upper bounds" of values of

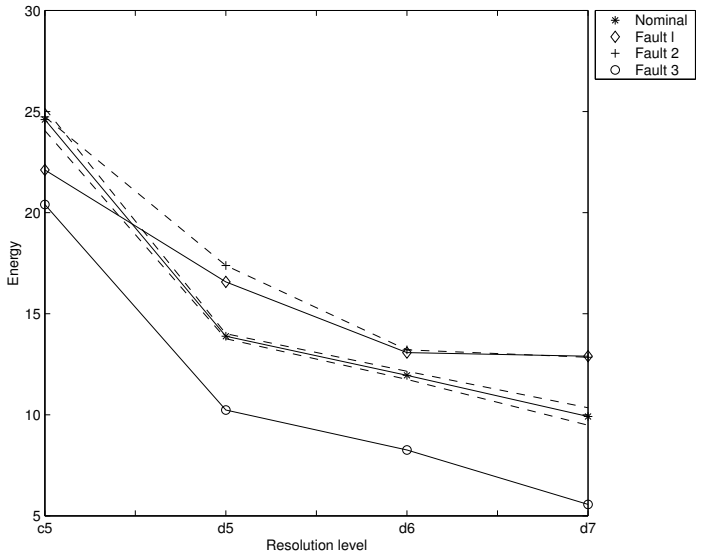

Fig. 11. Thresholded Scalograms with Pointwise Confidence Intervals

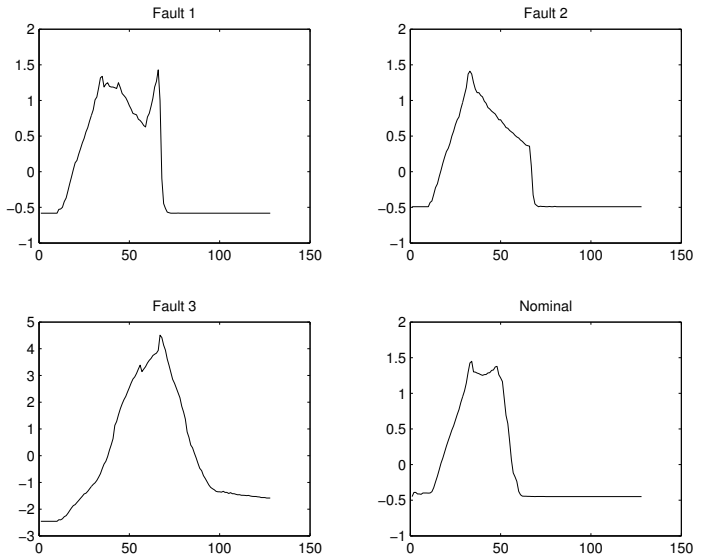

Fig. 12. RTCVD Signals in Fault Classes.

the thresholded scalograms in process monitoring. The proof is based on a probability argument to establish the asymptotic equivalence between $S_j^*(\hat{\lambda})$ and $S_j^*(\lambda)$, and the validation of the Lindeberg condition (as seen in the proof of Theorem 2) for $S_j^*(\lambda)$. See Jeong, Chen and Lu (2003) for details.

Theorem 3. *If* $\mu_j^* = E(S_j^*) \geq 0$ *and* $\sigma_{m_j}^2 = Var(S_j^*) < \infty$, *then*

$$(\log_2 S_j^* - \log_2 \mu_j^*)/\sigma_{m_j} \xrightarrow{D} N[0, 1/(\mu_j^* \ln 2)^2] \text{ as } m_j \to \infty. \tag{8}$$

Based on the approximated normal distribution, the $(1-\alpha)100\%$ confidence interval for the $\log_2$-scale thresholded scalogram is obtained as $\log_2 S_j^* \pm z_{\alpha/2}\hat{\sigma}_{m_j}/[\hat{\mu}_j^*(\ln 2)]$, where z_α is the usual upper $\alpha \times 100\%$th percentile value of the standard normal distribution. The values of this confidence interval will serve as the "monitoring bounds" for our scalogram plots. Figure 11 shows the bounds connected in a pointwise manner from the 95% confidence intervals calculated at selected resolution levels.

Because the RRE_h has a much better data-reduction ratio (see Table IV for details) in analyzing the RTCVD data, it was used in this example for the thresholding. Even with a limited data size, the monitoring bounds constructed from the approximated distribution are rather tight. Results plotted on Figure 11 show that these three fault classes of data curves are clearly out of the bounds in almost all resolution levels except the coarsest level (c_5) for the Fault 2 curve.

6. Conclusion and Future Research

This article proposes an idea of handling a special type of large and complicated functional data in data analysis and decision making. Properties of the proposed data-reduction methods are investigated by testing four popular signals in the statistics and engineering literature and two real-life examples. Results from the classification trees show that the proposed methods give similar accuracy (or better in some cases) but a more favorable computational efficiency compared to the results obtained from analyzing the original larger size data.

Future work is needed to explore the strengths and weaknesses in other decision rules (e.g., cluster analysis in data mining) and to extend the proposed idea to traditional quality improvement and SPC areas (e.g., analyze design of experiment data based on reduced-size information, analysis of variance of time-sequence or spatial data based on thresholded wavelet coefficients, and multi-

resolution SPC for spatial image data in process monitoring).

Acknowledgments

The authors thank the anonymous referees and Associate Editor for very valuable comments and suggestions. This research was supported by National Science Foundation grants EEC-0080315, DMS-0072960 and DMI-0400071.

Appendix

Extension of the RRE_h method to a soft-thresolding-based method RRE_s:

A similar idea presented for RRE_h can be extended from the soft-thresholding idea. In the wavelet-shrinkage literature it has been shown that hard-thresholding results in a larger variance of estimates, while soft-thresholding has a larger bias. Hard-thesholding is also very sensitive to small changes in the data. Soft-thresholding has various advantages such as continuity of the shrinkage rule. See Bruce and Gao (1996) for a comparison study between these two thresholding policies in data-denoising applications. See Tables III to V for their comparisons in data-reduction applications. The analytical properties of RRE_s can be derived similarly as presented in Theorem 4. Denote by $\hat{\boldsymbol{d}}_s(\lambda) = (\hat{d}_{s,1}(\lambda), \ldots, \hat{d}_{s,N}(\lambda))^\top$, where $\hat{d}_{s,i}(\lambda) = I(|d_i| > \lambda) sign(d_i)(|d_i| - \lambda)$, $i = 1, \ldots, N$. Then,

$$RRE_s(\lambda) = \frac{\mathrm{E}||\boldsymbol{d} - \hat{\boldsymbol{d}}_s(\lambda)||^2}{(\mathrm{E}||\boldsymbol{d}||^2)^{\frac{1}{2}}} + \omega \frac{\mathrm{E}||\hat{\boldsymbol{d}}_s(\lambda)||_1}{(\mathrm{E}||\boldsymbol{d}||_1)^{\frac{1}{2}}}, \tag{9}$$

where $||\hat{\boldsymbol{d}}_s(\lambda)||_1 = \sum_{i=1}^N |\hat{d}_{s,i}(\lambda)|$.

Theorem 4. *Consider the model stated in Eq. (3). Then we have*

(i) the objective function $RRE_s(\lambda)$ is minimized uniquely at $\lambda = \lambda_{N,s}$ where

$$\lambda_{N,s} = 0.5 * (\frac{\mathrm{E}||\boldsymbol{d}||^2}{\mathrm{E}||\boldsymbol{d}||_1})^{1/2}; \tag{10}$$

The empirical estimate of $\lambda_{N,s}$,

$$\hat{\lambda}_{N,s} = 0.5 * \left(\frac{\sum_{i=1}^N d_i^2}{\sum_{i=1}^N |d_i|} \right)^{1/2} = 0.5 * (\frac{\hat{\xi}}{l_1})^{\frac{1}{2}}, \tag{11}$$

where l_1 is the L_1-norm of $\boldsymbol{d}$.

(ii) $(\hat{\lambda}_{N,s} - \lambda_{N,s}) \xrightarrow{w.p.1} 0$;

Proof of Theorem 1. In this proof, we focus on the stochastic case first, and address the modification of the proof for the deterministic case in the end. Let $d^2_{(1)} \geq d^2_{(2)} \geq \cdots d^2_{(N)}$ be the ordered energies of wavelet coefficients. Because

$$E(\hat{\xi}) = E\left\|\mathbf{y}\right\|^2 = E\left\|\mathbf{d}\right\|^2 = \sum_{i=1}^{N} E(d_i^2) = \sum_{i=1}^{N} E(d_{(i)}^2) \geq \sum_{i=1}^{M} E(d_{(i)}^2) \geq ME(d_{(M)}^2),$$

the inequalities, $E(d^2_{(M)}) \leq E(\hat{\xi})/M$ holds for $M = 1, 2, \cdots, N$. Therefore,

$$E\left\|\mathbf{y} - \hat{\boldsymbol{f}}_M\right\|^2 = \sum_{i=M+1}^{N} E(d_{(i)}^2) \leq \sum_{i=M+1}^{N} E(\hat{\xi})/i \leq (N-M)E(\hat{\xi})/M.$$

For the deterministic case, replace $d_{(i)}$'s with $\theta_{(i)}$'s, $E(\hat{\xi})$ with $\xi = \|\mathbf{f}\|^2 = \|\boldsymbol{\theta}\|^2$, and delete the expectations. The error bound will be derived as stated in Theorem 1.

Proof of Theorem 2. Denote

$$H_i(\lambda) = \mathrm{E}(I(|d_i| \leq \lambda) d_i^2) = \int_{-\lambda}^{\lambda} t^2 \frac{1}{\sigma} \phi\left(\frac{t-\theta_i}{\sigma}\right) dt$$

and

$$h_i(\lambda) = \mathrm{E}(|\hat{d}_{h,i}(\lambda)|_0) = \mathrm{E}(I(|d_i| > \lambda) = 1 - \int_{-\lambda}^{\lambda} \frac{1}{\sigma} \phi\left(\frac{t-\theta_i}{\sigma}\right) dt,$$

where $\phi(x) = (2\pi)^{-1/2} \exp(-t^2/2)$, the standard normal density. It follows that

$$\begin{aligned}
\mathrm{E}||\boldsymbol{d} - \hat{\boldsymbol{d}}_h(\lambda)||^2 &= \sum_{i=1}^{N} \mathrm{E}(d_i - I(|d_i| > \lambda) d_i)^2 = \sum_{i=1}^{N} \mathrm{E}(I(|d_i| \leq \lambda) d_i^2) = \sum_{i=1}^{N} H_i(\lambda), \\
\mathrm{E}||\hat{\boldsymbol{d}}_h(\lambda)||_0 &= \sum_{i=1}^{N} \mathrm{E}(|\hat{d}_i(\lambda)|_0) = \sum_{i=1}^{N} \mathrm{E}(I(|d_i| > \lambda)) = \sum_{i=1}^{N} h_i(\lambda).
\end{aligned}$$

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

$$RRE_h(\lambda) = \sum_{i=1}^{N} H_i(\lambda)/\mathrm{E}||\boldsymbol{d}||^2 + \frac{1}{N} \sum_{i=1}^{N} h_i(\lambda).$$

Because of

$$\frac{dh_i(\lambda)}{d\lambda} = -\frac{1}{\sigma}\left[\phi\left(\frac{\lambda-\theta_i}{\sigma}\right) + \phi\left(\frac{-\lambda-\theta_i}{\sigma}\right)\right] < 0$$

and

$$\frac{dH_i(\lambda)}{d\lambda} = \frac{\lambda^2}{\sigma}\left[\phi\left(\frac{\lambda-\theta_i}{\sigma}\right) + \phi\left(\frac{-\lambda-\theta_i}{\sigma}\right)\right] = -\lambda^2 \frac{dh_i(\lambda)}{d\lambda},$$

we know that

$$\frac{dRRE_h(\lambda)}{d\lambda} = \left(-\lambda^2/\mathrm{E}(||\boldsymbol{d}||^2) + \frac{1}{N}\right)\sum_{i=1}^{N}\frac{dh_i(\lambda)}{d\lambda} = 0,$$

only if

$$\lambda = \lambda_{N,h} = \left(\frac{1}{N}\mathrm{E}||\boldsymbol{d}||^2\right)^{1/2}.$$

Since d_i's are independently $N(\theta_i, \sigma^2)$ distributed, $N\hat{\lambda}^2_{N,h}/\sigma^2 = \sum_{i=1}^{N} d_i^2/\sigma^2$ is $\chi^2(N, \delta_N)$ distributed with degree of freedom N and non-centrality parameter $\delta_N = \sum_{i=1}^{N}\theta_i^2/\sigma^2$. It follows that $\mathrm{E}(\hat{\lambda}^2_{N,h}) = \sigma^2(\delta_N/N + 1) = \lambda_N$ and $\mathrm{Var}(\hat{\lambda}^2_{N,h}) = \sigma^4(4\delta_N + 2N)/N^2 \to 0$, as $N \to \infty$. Note that $f(t)$ is continuous on $[0, T]$, and then $\max_{0\le t\le T}|f(t)| = K \le \infty$. Because DWT is orthonormal, $|\theta_i|$, $i = 1, 2, \ldots, N$, should be uniformly bounded, as $N \to \infty$. Without loss of generality, we assume that $|\theta_i| < K$, $i = 1, 2, \cdots, N$. Therefore,

$$\lim_{N\to\infty}\sum_{i=1}^{N}\frac{\theta_i^2}{i^2} < K^2 \lim_{N\to\infty}\sum_{i=1}^{N}\frac{1}{i^2} < \infty,$$

and we know that

$$\lim_{N\to\infty}\sum_{i=1}^{N}\frac{\mathrm{Var}(d_i^2)}{i^2} < (4\sigma^2K^2 + 2\sigma^4)\lim_{N\to\infty}\sum_{i=1}^{N}\frac{1}{i^2} < \infty.$$

Therefore, from the Kolmogorov Theorem (Serfling, 1980, p.27), we know that $(\hat{\lambda}_{N,h} - \lambda_{N,h}) \xrightarrow{w.p.1} 0$ (i.e. the result (ii) is true).

In order to show the asymptotic normality of $\sqrt{N}(\hat{\lambda}^2_{N,h} - \lambda^2_{N,h})/\sigma(\hat{\lambda}^2_{N,h})$, it is sufficient to verify the following Lindeberg condition (Serfling. 1980, p.30), for every $\varepsilon > 0$

$$\frac{1}{N}\sum_{i=1}^{N}\int_{|t^2-\mu_i|>\varepsilon\sqrt{N}}(t^2-\mu_i)^2\phi\left(\frac{t-\theta_i}{\sigma}\right)dt \to 0, \quad N \to \infty, \tag{12}$$

where $\mu_i = \mathrm{E}(d_i^2) = \theta_i^2 + \sigma^2$. It follows that

$$\begin{aligned}
\int_{|t^2-\mu_i|>\varepsilon\sqrt{N}}(t^2-\mu_i)^2\phi\left(\frac{t-\theta_i}{\sigma}\right)dt &= O\left(\int_{t^2>\varepsilon\sqrt{N}} t^4\phi\left(\frac{t-\theta_i}{\sigma}\right)dt\right) \\
&= O\left(\int_{t>\varepsilon^{1/2}N^{1/4}} t^4\phi\left(\frac{t-\theta_i}{\sigma}\right)dt\right) \\
&= O\left(\varepsilon^2 N\phi\left(\frac{\varepsilon^{1/2}N^{1/4}-\theta_i}{\sigma}\right)\right) \\
&= O\left((\varepsilon^2 N\exp\left\{-\frac{\varepsilon\sqrt{N}}{2\sigma^2}\right\}\right).
\end{aligned}$$

Therefore, for every $\varepsilon > 0$, as $N \to \infty$,

$$\frac{1}{N}\sum_{i=1}^{N}\int_{|t^2-\mu_i|>\varepsilon\sqrt{N}}(t^2-\mu_i)^2\phi\left(\frac{t-\theta_i}{\sigma}\right)dt = O\left((\varepsilon^2 N\exp\left\{-\frac{\varepsilon\sqrt{N}}{2\sigma^2}\right\}\right)\to 0,$$

and we know that $\sqrt{N}(\hat{\lambda}^2_{N,h}-\lambda^2_{N,h})/\sigma(\hat{\lambda}^2_{N,h})$ is asymptotically normal. Then, from the delta method, if $(T_N-\eta_N)/\tau_N \xrightarrow{d} N(0,1)$, then $[h(T_N)-h(\eta_N)]/[\tau_N h'(\eta_N)] \xrightarrow{d} N(0,1)$ provided h is continuous function such that $h'(\eta_N)$ exists and $h'(\eta_N)\neq 0$. In our situation, let $T_N=\hat{\lambda}^2_{N,h}$, $\eta_N=\lambda^2_{N,h}$, and $\tau_N=\sigma_N(\hat{\lambda}^2_N)$, $h(\eta)=\sqrt{\eta}$ and $h'(\eta)=1/2\sqrt{\eta}$, by applying the delta method, we can get the stated results of (iii).

Proof of Theorem 4. Denote

$$V_i(\lambda)=\mathrm{E}(|\hat{d}_{s,i}(\lambda)|)=\mathrm{E}\left((|I(|d_i|>\lambda)sign(d_i)(|d_i|-\lambda)|\right).$$

According to the intervals of d_i, the term $I(|d_i|>\lambda)sign(d_i)(|d_i|-\lambda)$ can be defined as follows:

$$I(|d_i|>\lambda)sign(d_i)(|d_i|-\lambda)=\begin{cases} d_i+\lambda, & d_i<-\lambda \\ 0, & -\lambda<d_i<\lambda \\ d_i-\lambda, & d_i>\lambda. \end{cases}$$

Then,

$$\begin{aligned} V_i(\lambda) &= \mathrm{E}(|I(d_i>\lambda)(d_i-\lambda)|)+\mathrm{E}(|I(d_i<-\lambda)(d_i+\lambda)|) \\ &= \int_{\lambda}^{\infty}\frac{|t-\lambda|}{\sigma}\phi\left(\frac{t-\theta_i}{\sigma}\right)dt+\int_{-\infty}^{-\lambda}\frac{|t+\lambda|}{\sigma}\phi\left(\frac{t-\theta_i}{\sigma}\right)dt. \end{aligned}$$

Since,

$$\begin{aligned} \mathrm{E}(d_i-\hat{d}_{s,i}(\lambda))^2 &= \mathrm{E}\left[(d_i-I(|d_i|>\lambda)sign(d_i)(|d_i|-\lambda))^2\right] \\ &= \mathrm{E}\left[I(|d_i|\leq\lambda)d_i^2\right]+\lambda^2\mathrm{E}\left[I(|d_i|>\lambda)\right] \\ &= H_i(\lambda)+\lambda^2 h_i(\lambda), \end{aligned}$$

$RRE_s(\lambda)$ can be written as

$$RRE_s(\lambda)=\left(\sum_{i=1}^{N}H_i(\lambda)+\lambda^2\sum_{i=1}^{N}h_i(\lambda)\right)/\mathrm{E}(||\boldsymbol{d}||^2)^{\frac{1}{2}}+\sum_{i=1}^{N}V_i(\lambda)/\mathrm{E}(||\boldsymbol{d}||_1)^{\frac{1}{2}}.$$

Since

$$\begin{aligned}\frac{dV_i(\lambda)}{d\lambda} &= -\frac{\lambda}{\sigma}\phi\left(\frac{\lambda-\theta_i}{\sigma}\right) - \int_{\lambda}^{\infty}\frac{1}{\sigma}\phi\left(\frac{t-\theta_i}{\sigma}\right)dt + \frac{\lambda}{\sigma}\phi\left(\frac{\lambda-\theta_i}{\sigma}\right) \\ &\quad + \frac{\lambda}{\sigma}\phi\left(\frac{-\lambda-\theta_i}{\sigma}\right) - \int_{-\infty}^{-\lambda}\frac{1}{\sigma}\phi\left(\frac{t-\theta_i}{\sigma}\right)dt - \frac{\lambda}{\sigma}\phi\left(\frac{-\lambda-\theta_i}{\sigma}\right) \\ &= -\mathrm{E}(|d_i| > \lambda) = -h_i(\lambda),\end{aligned}$$

$$\begin{aligned}\frac{dRRE_s(\lambda)}{d\lambda} &= \left[-\lambda^2\sum_{i=1}^{N}\frac{dh_i(\lambda)}{d\lambda} + 2\lambda\sum_{i=1}^{N}h_i(\lambda) + \lambda^2\sum_{i=1}^{N}\frac{dh_i(\lambda)}{d\lambda}\right]/\mathrm{E}(||\boldsymbol{d}||^2)^{\frac{1}{2}} - \left[\sum_{i=1}^{N}h_i(\lambda)\right]/\mathrm{E}(||\boldsymbol{d}||_1)^{\frac{1}{2}} \\ &= \left(\frac{2\lambda}{\mathrm{E}(||\boldsymbol{d}||^2)^{\frac{1}{2}}} - \frac{1}{\mathrm{E}(||\boldsymbol{d}^{|}|_1)^{\frac{1}{2}}}\right)\sum_{i=1}^{N}h_i(\lambda) = 0,\end{aligned}$$

only if

$$\lambda = \lambda_{N,s} = \frac{1}{2}\left(\frac{\mathrm{E}||\boldsymbol{d}||^2}{\mathrm{E}||\boldsymbol{d}^{|}|_1}\right)^{1/2}.$$

Also, similar to the proof of (ii) of Theorem 2, we know that $(\hat{\lambda}_{N,s} - \lambda_{N,s}) \xrightarrow{w.p.1} 0$ from the Kolmogorov Theorem and Slutsky's Theorem, i.e. the result (ii) is true.

References

[1] Antoniadis, A., Gijbels, I., and Grégoire, G. (1997), "Model S election Using Wavelet Decomposition and Applications," *Biometrika*, 84(4), 751–763.

[2] Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984), *Classification and Regression Trees*, New York: Chapman and Hall.

[3] Bruce, A. G., and Gao, H.-Y. (1996), "Understanding WaveShrink: Variance and Bias Estimation," *Biometrika*, 83(4), 727–745.

[4] Donoho, D. L., and Johnstone, I. M. (1994), "Ideal Spatial Adaptation by Wavelet Shrinkage," *Biometrika*, 81(4), 425–455.

[5] Donoho, D. L., and Johnstone, I. M. (1995), "Adapting to Unknown Smoothness via Wavelet Shrinkage," *Journal of the American Statistical Association*, 90(432), 1200–1224.

[6] Donoho, D. L., and Huo, X. (2001), "Beamlets and Multiscale Image Analysis," in *Multiscale and Multiresolution Methods*, Editors T.J. Barth, T. Chan, and R. Haimes, Springer Lecture Notes in Computational Science and Engineering, 20, 149–196.

[7] Fukunaga, K. (1990), *Introduction to Statistical Pattern Recognition*, London: Morgan Kaufmann Inc.

[8] Ganesan, R., Das, T. K., Sikdar, A., and Kumar, A. (2003), "Wavelet Based Detection of Delamination Defect in CMP Using Nonstationary Acoustic Emission Signal," *IEEE Transactions on Semiconductor Manufacturing*, 16(4), to appear.

[9] Hall, P., Poskitt, D. S., and Presnell, B. (2001), "A Functional Data-Analytic Approach to Signal Discrimination," *Technometrics*, 43(1), 1–9.

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

[11] Ihara, I. (1993), *Information Theory for Continuous System,* New Jersey: World Scientific.

[12] Jeong, M. K., Chen, D., and Lu, J.-C. (2003), "Fault Detection Using Thresholded Scalogram," *Applied Stochastic Models in Business and Industry*, 19(3), 231-244.

[13] Jeong, M. K., and Lu, J.-C. (2004), "Adaptive SPC Procedures for Complicated Functional Data," *Technical Report*, The School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA.

[14] Jin, J., and Shi, J. (1999), "Feature-Preserving Data Compression of Stamping Tonnage Information Using Wavelets," *Technometrics*, 41(4), 327–339.

[15] Jin, J., and Shi, J. (2001), "Automatic Feature Extraction of Waveform Signals for In-Process Diagnostic Performance Improvement," *Journal of Intelligent Manufacturing*, 12, 257–268.

[16] Jung, U., and Lu, J.-C. (2004), "A Wavelet-based Random-effect Model for Multiple Sets of Complicated Functional Data," *Technical Report*, The School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA. See http://www.isye.gatech.edu/

[17] Koh, C. K. H., Shi, J., Williams, W. J., and Ni, J. (1999), "Multiple Fault Detection and Isolation Using the Haar Transform, Part 2: Application to the Stamping Process," *Transactions of the ASME*, 295–299.

[18] Lada, E. K., Lu, J.-C., and Wilson, J. R. (2002), "A Wavelet Based Procedure for Process Fault Detection," *IEEE Trans. on Semiconductor Manufacturing*, 15(1), 79-90.

[19] Liu, B., and Ling, S. F. (1999), "On the Selection of Informative Wavelets for Machinery Diagnosis," *Mechanical Systems and Signal Processing*, 13(1), 145-162.

[20] Lu, J.-C. (2001), "Methodology of Mining Massive Data Set for Improving Manufacturing Quality/Efficiency," a chapter (pp. 255-288) for the book entitled *Data Mining for Design and Manufacturing: Methods and Applications* edited by D. Braha as a volume in a series of "Massive Computing" that is organized by James Abello (AT&T Labs Research), Panos Pardalos (Univ. of Florida) and Mauricio Resende (ATT Labs Research, New York: Kluwer Academic Publishers.

[21] Mallat, S. G. (1998), *A Wavelet Tour of Signal Processing*, San Diego: Academic Press.

[22] Rioul, O., and Vetterli, M. (1991),"Wavelets and Signal Processing," *IEEE Signal Processing Magazine*, October, 14–38.

[23] Rying, E. A., Gyurcsik, R. S., Lu, J. C., Bilbro, G., Parsons, G., and Sorrell, F. Y. (1997), "Wavelet Analysis of Mass Spectrometry Signals for Transient Event Detection and Run-to-run Process Control," in *Proceedings of the Second International Symposium on Process Control, Diagnostics, and Modeling in Semiconductor Manufacturing*, editors: Meyyappan, M., Economou D., J., Bulter, S. W., 37-44.

[24] Rying, E. A. (2001), "A Novel Focused Local-Learning Wavelet Network with Application to In-situ Selectivity and Thickness Monitoring during Selective Silicon Epitaxy," unpublished Ph.D. thesis, Department of Electrical and Computer Engineering, North Carolina State University, Raleigh, NC.

[25] Saito, N. (1994), "Simultaneous Noise Suppression and Signal Compression Using a Library of Orthonormal Bases and the Minimum Description Length Criterion," in *Wavelets in Geophysics*, E. Foufoula-Georgiou and P. Kumar, Eds. New York: Academic Press, 299–324.

[26] Scargle, J.D. (1997), "Wavelet Methods in Astronomical Time Series Analysis," in *Application of Time Seires Analysis in Astronomy and Meteorology*, T. S. Rao, M. B. Priestly, and O. Lessi, Eds. New York: Chapman and Hall, 226–248.

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

[28] Vidakovic, B. (1999), *Statistical Modeling by Wavelets*, New York: John Wiley & Sons.

[29] Vidakovic, B. (2000), "Unbalancing Data With Wavelet Transformations," *Technical Report*, Department of Statistics, Duke University, Durham, NC.

[30] Wang, X. Z., Chen, B. H., Yang, S. H., and McGreavy, C. (1999), "Application of Wavelets and Neural Networks to Diagnostic System Development, 2, An integrated Framework and its Application," *Computers and Chemical Engineering*, 23, 945–954.

[31] Weyrich, N. and Warhola, G. T. (1998), "Wavelet Shrinkage and Generalized Cross Validation for Image Denoising," *IEEE Transactions on Image Processing*, 7(1), 82-90.