Editor's Note: This article was presented at the *Technometrics* invited session at INFORMS 2005, in San Francisco, CA, November 2005.

Wavelet-Based Data Reduction Techniques for Process Fault Detection

Myong K. JEONG

Department of Industrial and Information Engineering
The University of Tennessee
Knoxville, TN 37996
(*mjeong@utk.edu*)

Jye-Chyi LU, Xiaoming HUO, and Brani VIDAKOVIC

School of Industrial and Systems Engineering
The Georgia Institute of Technology
Atlanta, GA 30332
(*jclu@isye.gatech.edu*)

Di CHEN

Biostatics US
UCB Pharma, Inc.
Smyrna, GA 30080

This article presents new data reduction methods based on the discrete wavelet transform to handle potentially large and complicated nonstationary data curves. The methods minimize objective functions to balance the trade-off 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 datasets, the proposed methods demonstrate their competitiveness with 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 operational quality and efficiency. Signal processing and data mining techniques are more popular than ever in such fields as sensor technology and intelligent manufacturing. As datasets 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 dataset 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. Ganesan, Das, Sikdar, and Kumar (2003) have provided another motivating example from a nano-manufacturing process. This motivates the focus of this article: developing data reduction procedures for data analysis tools to be useful in handling large-sized complicated functional data. Studies in Section 4 show that the proposed procedures are more effective for data signals, with sharper changes and less noise. Applications producing this type of signal, such as the foregoing examples, can take advantage of the procedures; see Table 1 and Figure 2 for details.

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 the lobes in Figure 1, better than the Fourier transform and standard statistical procedures (e.g., splines and polynomial regressions) and provide a multiresolution approximation to the data (Mallat 1988, p. 378). 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 dataset" consisting of these features is constructed to detect specific types of known faults, then a

TECHNOMETRICS, FEBRUARY 2006, VOL. 48, NO. 1
DOI 10.1198/004017005000000553

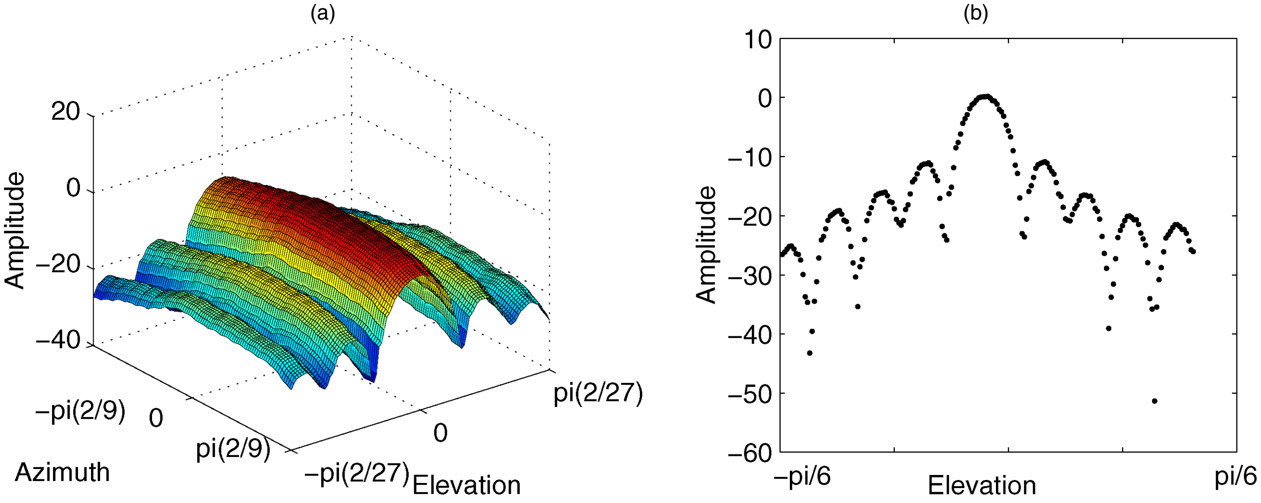

Figure 1. Data Signals From Antenna Manufacturing Processes: (a) Antenna; (b) Azimuth-Cut Data.

data reduction procedure could be derived to minimize type I 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 statistical process control (SPC) limits as the decision rule. But the wavelet coefficients selected for a given decision rule might not be suitable for other purposes of analysis. The aim of our data reduction is to produce a small set of "representative data" suitable for 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. In another method, Rying et al. (1997) applied a scale-dependent energy metric, $E_s =$ sum of squares of all wavelet coefficients (see Sec. 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, p. 289; see Fig. 12 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., Figs. 1, 4, and 7(a)] do not have large-sized random noises for demonstrating the effectiveness of data denoising procedures. Section 4 (e.g., Tables 1–4) uses simulations and real-life examples to illustrate that the ability of data denoising procedures in data reduction is limited. In contrast, the energy-metric approach is too aggressive and is 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.

Section 2 provides background information on wavelet transforms. 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 decision making analyses. Finally, Section 6 provides a few concluding remarks and future studies.

2. WAVELET TRANSFORMS

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

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

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, $\psi_{s,u}(t) = s^{1/2}\psi(st - u)$, through the dilation factor s and the translation u. Scaling function $\phi(t) \in L^2(\mathbb{R})$ is defined similarly, but $\int_{\mathbb{R}} \phi(t)\,dt \neq 0$.

Table 1. Results of Data Reduction for Testing Signals

		Threshold value		*M = no. of coefficients selected*					
Signals	*Energy*	$\hat{\lambda}_h$	$\hat{\lambda}_s$	*RRE$_h$*	*RRE$_s$*	*Visu*	*Risk*	*SURE*	*AMDL*
Nason	94.25	.3034	.6986	31	138	192	225	324	192
Heavisine	90.28	.2969	.6803	28	143	287	290	292	194
Blocks	72.36	.2658	.5099	67	379	389	407	518	391
Bumps	17.63	.1312	.3401	91	405	646	664	722	894

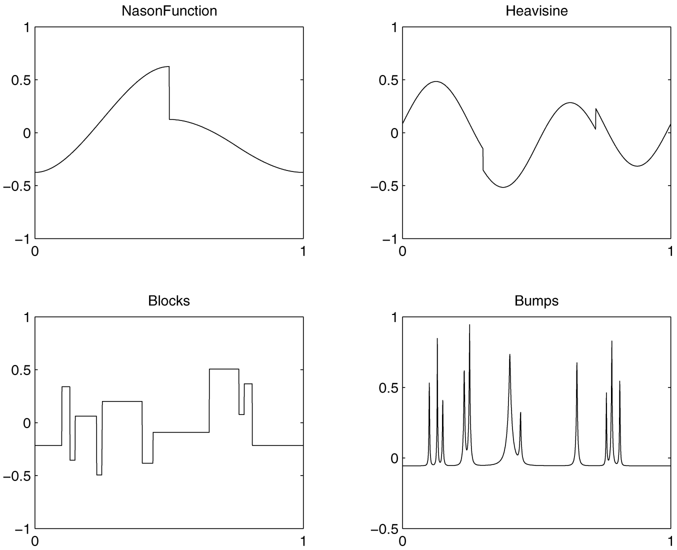

Figure 2. Four Testing Signals From the Literature.

Select the scaling and wavelet functions as $\{\phi_{L,k}(t) = 2^{L/2}\phi(2^L t - k); k \in \mathbb{Z}\}$ and $\{\psi_{j,k}(t) = 2^{j/2}\psi(2^j t - k); j \geq L, k \in \mathbb{Z}\}$. In practice, the following orthonormal basis of wavelet is used to represent a signal function $f(t) \in L^2(\mathbb{R})$:

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

where $\mathbb{Z}$ denotes the set of all integers $\{0, \pm1, \pm2, \ldots\}$, the coefficients $c_{L,k} = \int_{\mathbb{R}} f(t)\phi_{L,k}(t)\,dt$ are considered 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 $\mathbf{y} = (y(t_1), \ldots, y(t_N))'$ taken from $f(t)$ or obtained as a realization of

$$y(t) = f(t) + \epsilon_t \qquad (2)$$

at equally spaced discrete time points $t = t_i$'s, where the ϵ_{t_i}'s are random normal $\mathrm{N}(0, \sigma^2)$ noises. The discrete wavelet transform (DWT) of $\mathbf{y}$ is defined as $\mathbf{d} = \mathbf{W}\mathbf{y}$, where $\mathbf{W}$ is the orthonormal $N \times N$ DWT matrix. From (1), $\mathbf{d} = (\mathbf{c}_L, \mathbf{d}_L, \mathbf{d}_{L+1}, \ldots, \mathbf{d}_J)$, where $\mathbf{c}_L = (c_{L,0}, \ldots, c_{L,2^L-1})$, $\mathbf{d}_L = (d_{L,0}, \ldots, d_{L,2^L-1}), \ldots,$ and $\mathbf{d}_J = (d_{J,0}, \ldots, d_{J,2^J-1})$. Using the inverse DWT, the $N \times 1$ vector $\mathbf{y}$ from the original signal curve can be "reconstructed" as $\mathbf{y} = \mathbf{W}'\mathbf{d}$. The process of applying the DWT to transform a dataset closely resembles the process of computing the fast Fourier transformation (FFT).

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

Finally, the process fault patterns, which are frequency- or phase-shifted, are invisible to time domain control limits and 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

To demonstrate the difference between the proposed and existing 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 signal processing, the linear approximation method (see Mallat 1988, sec. 9.1, for details) uses the function $\mathbf{f}_M = \sum_{m=0}^{M-1} \langle \mathbf{f}, \mathbf{g}_m \rangle \mathbf{g}_m$ with a set of pre-determined vectors $\mathbf{g}_m$, $m = 0, 1, \ldots, M-1$, to reconstruct the original data signals, where $\langle \mathbf{f}, \mathbf{g}_m \rangle$ is the inner product of the function $\mathbf{f}$ and the projected vector $\mathbf{g}_m$. In the wavelet-based approximation, $\langle \mathbf{f}, \mathbf{g}_m \rangle$ is the wavelet coefficient (from the coarsest level to the finest level in the linear method).

The nonlinear approximation method (Mallat 1988, sec. 9.2) selects the M projection vectors [e.g., M-largest wavelet coefficients (in absolute values)] adaptively 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 predetermined 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 foregoing 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.

The literature includes limited studies on determining the number of vectors (M) used in the model $\mathbf{f}_M$ adaptively based on signal characteristics. The approximate minimum description length (AMDL) method proposed by Saito (1994) selects M to minimize the following objective function:

$$\mathrm{AMDL}(M) = 1.5M \log_2 N + .5N \log_2 \left[\sum_{i=1}^{N} (y_i - \widehat{y}_{i,M})^2 \right],$$

where $\widehat{y}_{i,M}$ is the approximation model similar to (1) constructed from the M largest-magnitude wavelet coefficients and the data y_i consist of $y(t)$ evaluated at $t = t_i$ from the model (2). As addressed by 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," such as entropy and mutual information (see Ihara 1993 and 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 $\mathbf{d} = \mathbf{W}\mathbf{y}$ to the data $\mathbf{y}$

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. Because $\mathbf{W}$ is an orthonormal transform, $\eta_{j,k}$'s are still iid $\mathrm{N}(0, \sigma^2)$ (Vidakovic 1999, p. 169).

Donoho and Johnstone (1995) developed several wavelet-based "shrinkage" techniques to find a smooth estimate ($\hat{\mathbf{f}}$) of $\mathbf{f}$ from the "noisy" data, $\mathbf{y}$. In particular, their hard-thresholding policy found 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 nonzero coefficients selected to estimate the underlying function $\mathbf{f}$ (using $\hat{\mathbf{f}} = \mathbf{W}^{-1}\hat{\boldsymbol{\theta}}$). The optimal estimate $\hat{\theta}_i$ was found to be equal to d_i if $|d_i| > \tau$ and to 0 otherwise. 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 are used in Section 4 for comparison studies.

Shrinkage methods require an estimate of the standard deviation σ for calculating the threshold value; for example, *VisuShrink*'s threshold is $(2 \ln N)^{1/2}\sigma$. Different estimates of σ will lead to distinct thresholds and different numbers of wavelet coefficients. This article uses a robust estimate, $\hat{\sigma} = \mathrm{median}(|d_{J,k}| : 1 \le k \le N/2)/.6745$, suggested by Donoho and Johnstone (1994), where J is the finest resolution level. The next section proposes two new data reduction methods that do not require estimation of σ.

3.3 Data Reduction Methods: *RRE$_h$* and *RRE$_s$*

Usually, data denoising, AMDL, and nonlinear signal approximation methods retain the largest number of coefficients M_λ based on some derivations of the threshold λ (see Cherkassky and Shao 2001 and Portilla and Simoncelli 2000 for other schemes in data denoising research). Our methods also follow this principle by assuming that larger wavelet coefficients will better characterize signal patterns in terms of their energy and thus will retain more information.

Definition 1. The energy of a finite sequence $\mathbf{f} = (f_1, \dots, f_N)$ is defined by $\xi = \|\mathbf{f}\|^2$. 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 $\mathbf{f} \in L^2(\mathbb{R})$, an upper bound of the approximation error for $\mathbf{f}_M$, is $\|\mathbf{f} - \mathbf{f}_M\|^2 \le [(N - M)/M]\xi$, and an upper bound of the estimation error for $\hat{\mathbf{f}}_M$ is $\mathrm{E}\|\mathbf{y} - \hat{\mathbf{f}}_M\|^2 \le [(N - M)/M]\mathrm{E}(\hat{\xi})$.

Data reduction and denoising methods are distinct for different purposes. As seen in (4), data denoising procedures aim to find the estimate $\hat{\boldsymbol{\theta}}$ (and $\hat{\mathbf{f}}$) for reducing "modeling error" of $\boldsymbol{\theta}$ (and $\mathbf{f}$). Thus the data denoising methods are 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 sufficiently representative in capturing key data characteristics for subsequent planned or unplanned decision analyses. Theorem 2 shows that our data reduction methods depend explicitly on the "data energy" representing data characteristics, whereas *VisuShrink* depends on the variance (σ^2) representing data noises.

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

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

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

Using "normalizing constants" to make the two balancing terms compatible is critical; Table 2 gives the results of empirical studies to illustrate its impact. The weighting parameter ω

Table 2. Impacts of Normalization for Data Reduction

	With normalization			*Without normalization*		
Signals	*Relative error*	*M/N*	*RRE$_h$*	*Relative error*	*M/N*	*RRE$^*_h$*
Bumps ($SNR^* = \infty$)	2.18E–02	.090	.112	2.81E–19	.770	.770
Bumps ($SNR^* = 15$)	2.94E–02	.066	.096	6.18E–04	.456	.456
Bumps ($SNR^* = 7$)	3.97E–02	.066	.106	2.98E–03	.432	.435
Bumps ($SNR^* = 3$)	9.45E–02	.066	.161	1.60E–02	.395	.411
RTCVD	1.77E–02	.130	.147	8.89E–07	.578	.578
Antenna	4.25E–02	.180	.222	3.27E–05	.644	.644

is user-selected or provided by such methods as generalized cross-validation (GCV) (Craven and Wahba 1979). However, further studies are needed to develop the GCV-like selection of ω in our problem and understanding its properties. For simplicity, here we use $\omega = 1$, which places equal weight on both components in follow-up studies (see Remark 4 in Sec. 4 for the guideline of determining the weight parameter ω). In what follows we use engineering and statistical experience to motivate the objective function (5). Our discussion focuses 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 that of Mallat (1988, pp. 378–391), the "relative error,"

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

is commonly used in comparing signal approximation quality. This is similar to the first term in (5). This article uses a thresholding parameter λ to decide which wavelet-domain data to keep and which to discard in decision making analyses using the terms $\hat{d}_{h,i}(\lambda) = I(|d_i| > \lambda)d_i$, $i = 1, \dots, N$. 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, in which the parameter τ was not originally set as the threshold for data reduction purposes in the construction of the objective function (4). Recall that in the discussion after (4) that the denoising procedures are aimed at estimating the 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.

Equation (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, Tibshirani, and Friedman 2001, p. 59) and neural network (Hastie et al. 2001, p. 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:

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

where ω is a weighting parameter like that in (5). Note that this objective function is not normalized, as was done in (5). More important, 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 (5) becomes handy in practical implementations. The proof of the theorem is given in the Appendix.

Theorem 2. Consider the model stated in (3). Then we have the following:

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

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

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

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

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

(c) $\sqrt{N}(\hat{\lambda}_{N,h} - \lambda_{N,h})/\sigma^*_{N,h} \xrightarrow{\text{d}} \mathrm{N}(0, 1)$, 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 (in the same scale and with mean 0) with 1,024 data points in each curve (Fig. 2) taken from the literature (e.g., Donoho and Johnstone 1995). Table 1 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 mean 0 and apply the thresholding rules to all resolution levels of the wavelet coefficients, whereas the denoising techniques do not threshold the coefficients in the coarser level ($c_{L,k}$'s; L in (1) is preselected, e.g., $L = 4$ for $N = 1{,}024$) (Donoho and Johnstone 1995).

Based on the observations from Table 1, in general, if the signal has a larger value of energy, then its threshold value will be higher (see, e.g., the threshold values for RRE_h and RRE_s) and will be more likely to have a smaller M. There are some exceptions to this, however. For example, if most of the signal energy is kept in a few larger wavelet coefficients, then the signal has a set of very "unbalanced" wavelet coefficients. When there is 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. Vidakovic (2000) provided a technique to compare signals with different unbalancing characteristics.

Table 2 presents the impact of not using the normalizing constants in (5), denoted by RRE_h^*, where $SNR^* = \mathrm{std}(\mathbf{f})/\sigma$ represents the noise level of data, $\mathrm{std}(\mathbf{f})$ is the standard deviation of the discretized signal points, and σ is the standard deviation of noise. Smaller SNR^* means that the data are noiser. Note that RRE_h in Table 2 is the sum of the first two columns, relative error and M/N, representing the metric defined in (5). Without normalization, the RRE_h^* procedure has a very poor data reduction ratio for all cases studied, and its performance is similar to that of data denoising methods for data reduction purposes. That is, it overemphasizes reducing modeling error by sacrificing their data reduction ability. The relative errors of RRE_h^* are very small, with plots similar to Figures 3–6 produced by data denoising methods (see Tables 3 and 4 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 used them for selecting "reduced-size" data to perform various decision analyses. This section compares the six methods presented

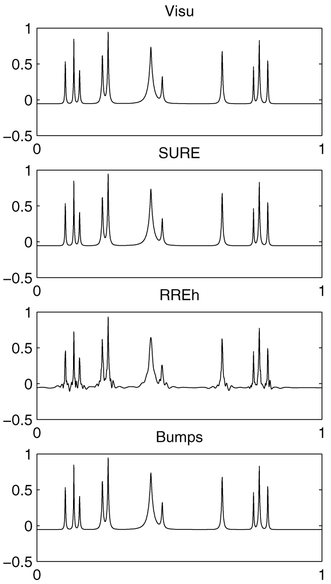

Figure 3. Reconstruction of the “Noise-Free” Bumps Signal.

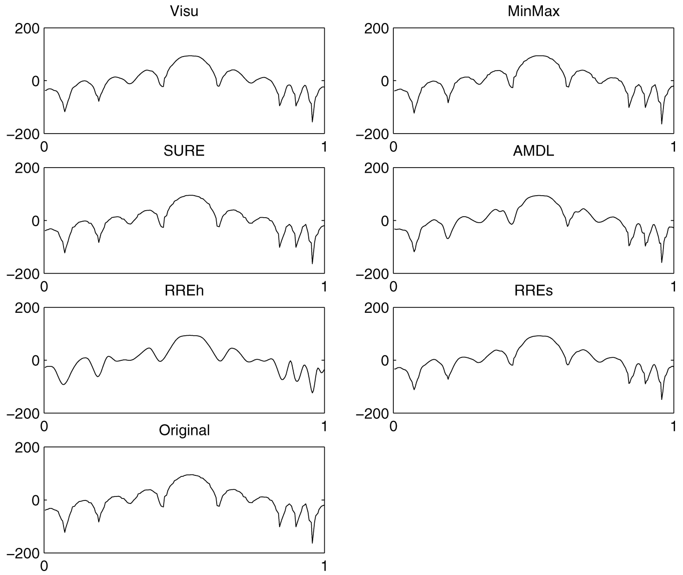

Figure 5. Reconstruction of the Antenna Data.

in Section 3 in terms of their modeling error and data reduction ability. The data patterns for comparisons include two real-life data curves (Figs. 4 and 5) and four well-known testing signals from the wavelet literature (Fig. 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.

Tables 3–5 present comparison results with the following summary measures: (1) reduction ratio (%): $RR = (1 - M/N) \times 100$; (2) $RelErr = \|\mathbf{f} - \hat{\mathbf{f}}_M\|/\|\mathbf{f}\|$ for the case without random errors and $RelErr = \|\mathbf{y} - \hat{\mathbf{f}}_M\|/\|\mathbf{y}\|$ for the case with random errors; and (3) AMDL measure.

Figure 3 shows the results for the bumps signal. The *VisuShrink*, *RiskShrink*, *SURE*, and AMDL(M) procedures achieve very small modeling errors (see Table 3 for the very

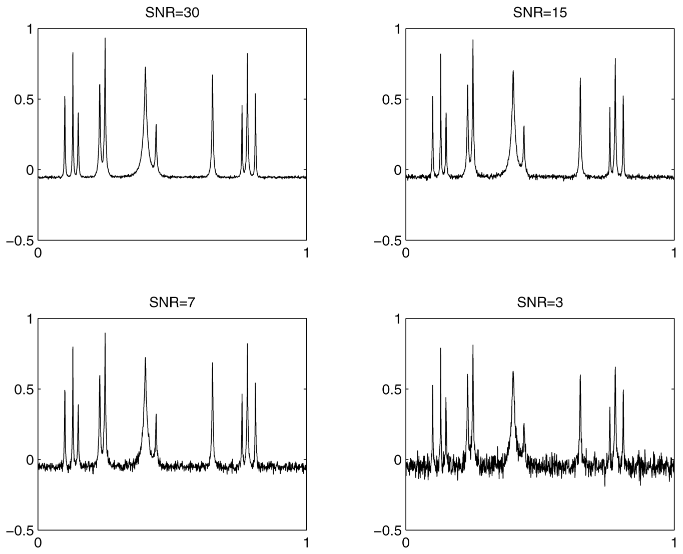

Figure 6. Noisy Bumps Signal at Various Noise Levels.

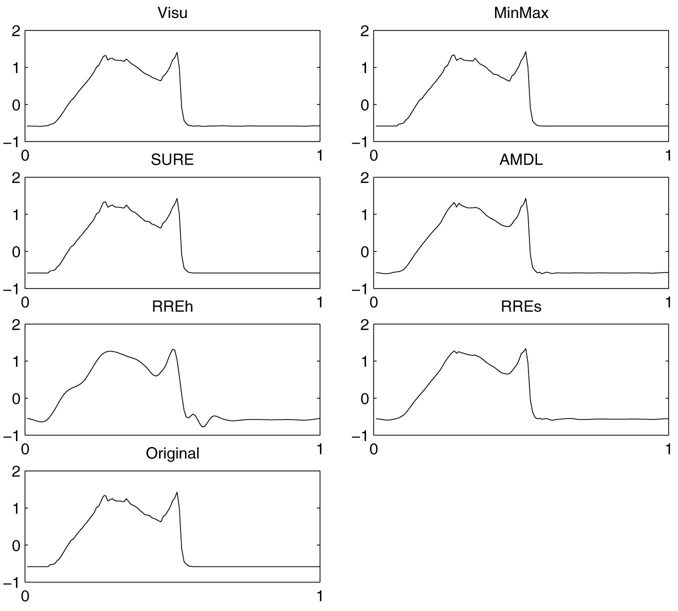

Figure 4. Reconstruction of the RTCVD Signal.

Table 3. Results for the Noise-Free Bumps Signal

Method	*M*	*RelErr*	*RR*	*AMDL*
VisuShrink	646	1.50E–16	36%	16,390.6
RiskShrink	664	1.23E–18	35%	13,108.3
SURE	722	2.22E–21	29%	26,321.8
AMDL	894	3.91E–25	13%	5,506.6
RRE_h	91	2.18E–02	91%	32,151.2
RRE_s	405	1.51E–09	60%	24,682.6

Table 4. Results for the RTCVD and Antenna Data

	RTCVD		*Antenna*	
Method	*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 5. Results for the Noisy Bumps Signal

	$SNR^* = \infty$		$SNR^* = 15$		$SNR^* = 7$		$SNR^* = 3$	
Method	*RR*	*RelErr*	*RR*	*RelErr*	*RR*	*RelErr*	*RR*	*RelErr*
Visu	36%	1.50E–16	85%	1.06E–02	88%	3.96E–02	91%	1.60E–01
Risk	35%	1.23E–18	78%	2.45E–03	82%	1.12E–02	85%	6.12E–02
SURE	29%	2.22E–21	57%	8.95E–04	65%	5.39E–03	73%	3.85E–02
AMDL	13%	3.91E–25	88%	6.92E–03	91%	2.74E–02	95%	1.50E–01
RRE_h	91%	2.18E–02	93%	2.91E–02	93%	3.93E–02	93%	9.70E–02
RRE_s	60%	1.51E–09	85%	1.20E–02	86%	2.85E–02	75%	7.59E–02

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 of the shapes and locations of the 11 peaks were identified and well modeled by the more aggressive RRE_h method, which has a 91% data reduction ratio as opposed to the 60% of RRE_s and <40% of 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 that of the *VisuShrink*, *RiskShrink*, and even RRE_s methods. It is interesting to note that though the *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 results were observed for several other testing signals (not shown here). Examples from Section 5 show that the RRE_h and RRE_s methods did give accurate decision results even with a more aggressive data reduction emphasis. The following examples test whether the proposed methods work well in the two real-life datasets in which errors were involved. Remark 1 discusses the studies of noisy bumps signals.

Example 1 (RTCVD data). The RTCVD process deposits thin films on the wafer through a temperature-driven surface chemical reaction. As feature size decreases, the functional operation of semiconductors (e.g., transistors) becomes increasingly unreliable because of variations in deposition processes. Therefore, controlling process variability is critical. QMS is commonly used in semiconductor manufacturing processes to monitor thin-film deposition quality. The data shown in Figure 4 are for one of the several nominal RTCVD process runs in a research project (Rying 2001) aimed at developing a new measurement technique for on-line 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 important, wavelet transforms are 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 4 show that the RRE_h could be too aggressive in data reduction (87% ratio) because of its nonsmoothing fit in the straight rising component (data between 20 and 30 points); however, it did roughly pick up the two peaks and other changepoints. 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 through visual inspection.

Example 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 that in 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 third peaks from the main lobe in the middle. Surprisingly, the AMDL(M) has an excellent data reduction ratio (81%), as good as that of the RRE_h; see Table 4 for details.

Remark 1. We also tested the robustness of the foregoing data reduction methods against random noise. 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^*. Table 5 summarizes model fitting and data reduction results from all methods in the cases where $SNR^* = 3$, $SNR^* = 7$, and $SNR^* = 15$.

The reported results are the means of performance measures from 100 simulation runs. The maximum coefficients of variation are 6% for AMDL and 2% for other procedures with respect to *RR*, and 29% for AMDL and 10% for other procedures with respect to *RelErr*. Smaller SNR^* means a noisier signal. For the signals with larger SNR^* (i.e., less noisy), the noise level (σ) is lower, and the threshold value should be lower [e.g., the threshold value of *VisuShrink* is $(2\ln N)^{1/2}\sigma$]. 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. A specific example is the drop in data reduction ratio for *SURE* in Table 5 from the $SNR^* = 3$ to $SNR^* = \infty$ cases. 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 data reduction ratio stays the same for the RRE_h, but improves considerably for all other methods. However, they pay a price for the much larger modeling errors (see Table 5) compared with the results given in Table 3. Surprisingly, the modeling errors from the *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.

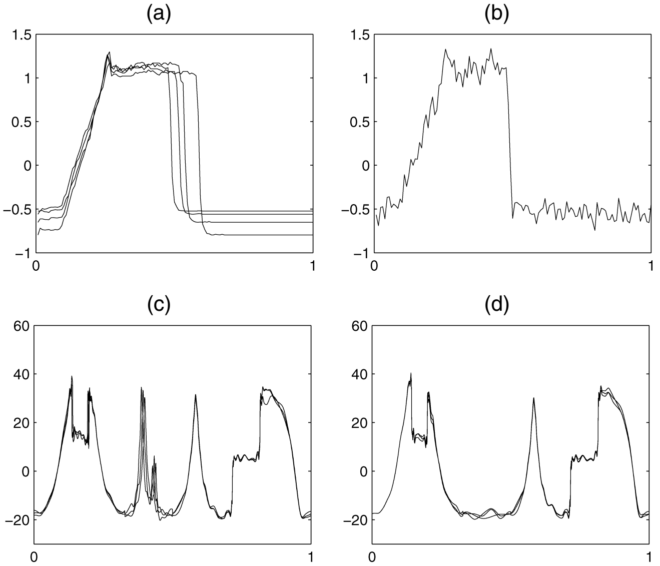

Figure 7. Different Types of Signal Replications: (a) Four In-Control Runs of RTCVD; (b) Curve With Added Random Noises; (c) Replicated Piecewise Signals (nominal); (d) Replicated Piecewise Signals (case 2).

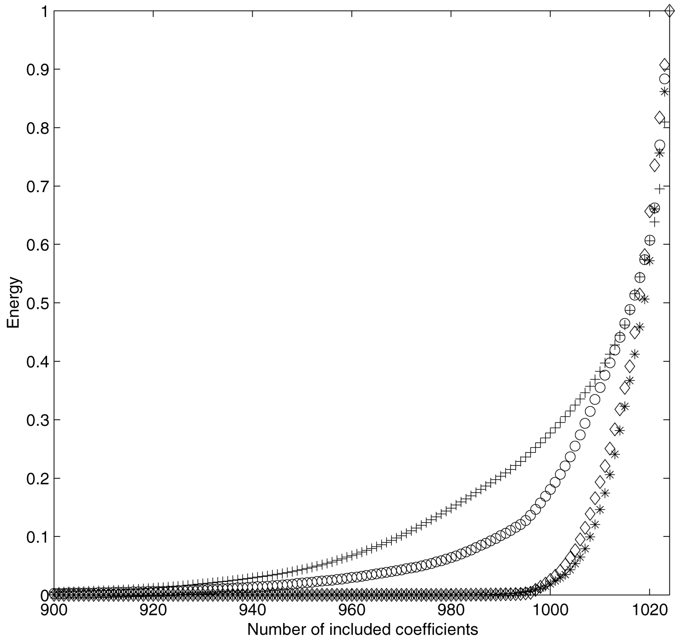

Figure 8. Cumulative Energy of Data Signals ($*$ Nason; ◇ Heavisine; + bumps; ○ blocks).

Remark 2. In engineering applications such as that of 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 images of products. With a certain amount of process variation, the resulting circles could have different radii and distinct centers, but they are all similar circles. This type of process variation is quite different from 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 Sec. 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 7(d) show one example of the original and replicated curves from the data generation procedure.

Remark 3. In deciding which wavelet family is most suitable for representing a data signal, the more "disbalancing" type (i.e., more separation in the larger and smaller wavelet coefficients) of wavelet family 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.

Remark 4. We present the following guidelines for choosing the weighting parameter ω. Figure 8 shows plots of cumulative energy of ordered wavelet coefficients from smallest to the largest (in absolute value). Note that Nason curve has the smoothest pattern, with a single sharp change (see Fig. 2). Thus only a few larger coefficients are needed to characterize the energy of the signal. Figure 8 shows that Nason curve has the largest slope for the rising pattern, because of the inclusion of the few largest coefficients in the last part of the energy cumulation. In contrast, the bumps signal has many finer-level coefficients characterizing "bumps." The larger size coefficients will be more numerous from the bumps signal than from other signals. Thus the Nason signal is more "disbalanced" than the other three signals (see Vidakovic 2000 for more details on signal disbalancy).

In general, more disbalanced signals have better data reduction ratios (see Table 1). More importanty, for this type of signal, because the signal energy is so focused in a few coefficients, if these coefficients are included, then the relative error will be small. Then even if more coefficients are added, the relative error will not change much. Thus, more disbalanced signals will be less sensitive to the choice of ω. Our experience from trying several values of ω with the four testing curves validate this observation. For a less disbalanced signal, such as bumps, if the relative error for the default choice $\omega = 1$ is not satisfactory, then smaller ω can be used to improve the fit. One might want to check the sensitivity of decision analyses for a few choices of ω's.

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., the results in Table 3). *VisuShrink*, *RiskShrink*, and *SURE* are not very effective in data reduction but have excellent modeling qualities. When larger amounts of normal random noise are added to the deterministic signal curves, the difference between these six methods in terms of modeling quality and data reduction ratio becomes smaller. This could be because 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 of Section 4, engineering variations used for generating replicated data curves are quite different from statistical random noise. Another major difficulty is the selection of the reduced-size data in the case of multiple curves. 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 because of the different wavelet bases of reduced-size datasets. Jung, Lu, and Jeong (2004) have presented a vertical thresholding procedure to tackle this problem. These difficulties make it 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.

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, and more serious process problems. Section 5.1 discusses the possibility of making decisions on process fault types using the classification and regression tree (CART) method. Section 5.2 presents the interesting idea of using the wavelet's multiresolution property to construct a visualization plot for understanding process problems.

5.1 Fault Classification Using the Classification and Regression Tree 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 "cutoff 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 Fig. 8) taken from Mallat (1988, p. 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) to generate a new curve with added random $\mathrm{N}(0, \sigma^2)$ noises using a small value of σ (=.1).

Figure 9 presents seven fault classes of curves, some of which are considerably more difficult than others for decision trees to correctly identify fault classes in. For example, the only difference between fault class 4 and the original curve is the smaller amount of vertical drop of the first rectangle-shaped dip at around 147–204 time units. Class 1 could also be considered a difficult case where the first dip is filled smoothly. We generated 300 replicated curves for each of the eight cases. Thus there are 2,400 data curves in this study.

To deal 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 that of 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 of these fault types based on the reduced-size data.

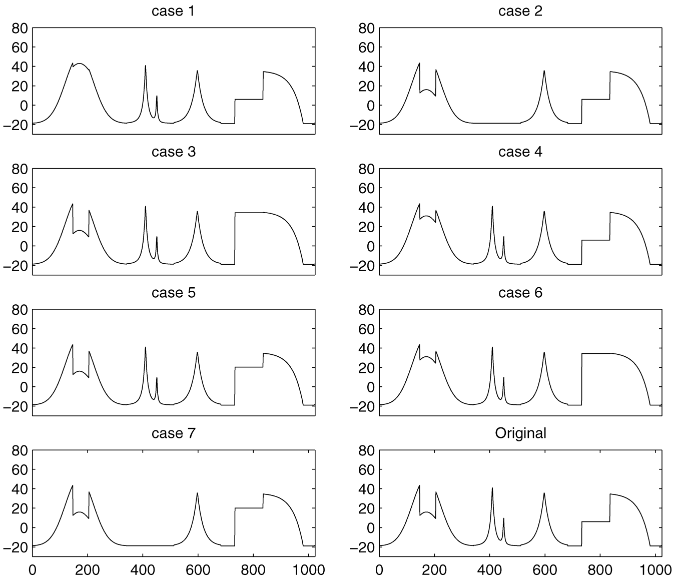

Figure 9. Mallat's Piecewise Signals.

There are no good guidelines available on how to divide the 2,400 samples into training and testing datasets. 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 10 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 cases 1–7.

The decision nodes picked by CART for decisions have certain interesting interpretations. The first split is $c_{5,6} \le -28.967$, where $c_{5,6}$ is the sixth 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 second and third peaks critical to fault detection and classification. Similar interpretation could be obtained for other coefficients selected by CART. In practice, most patterns could be identified by the coefficients at the coarser resolution level, whereas only a few patterns will require information from coefficients at finer levels for decisions (e.g., $d_{5,27}$ of node 7). Using combinations of coarser- and finer-level coefficients at different hierarchies of CART provides a multiresolution-oriented decision making opportunity not available in the time domain based on the original data.

As an illustration for the time savings achieved by using the reduced-size data for decision analyses, Figure 11 shows the CART tree constructed using $n = 1{,}024$ points in the time domain. The larger-sized data in the time domain increased the

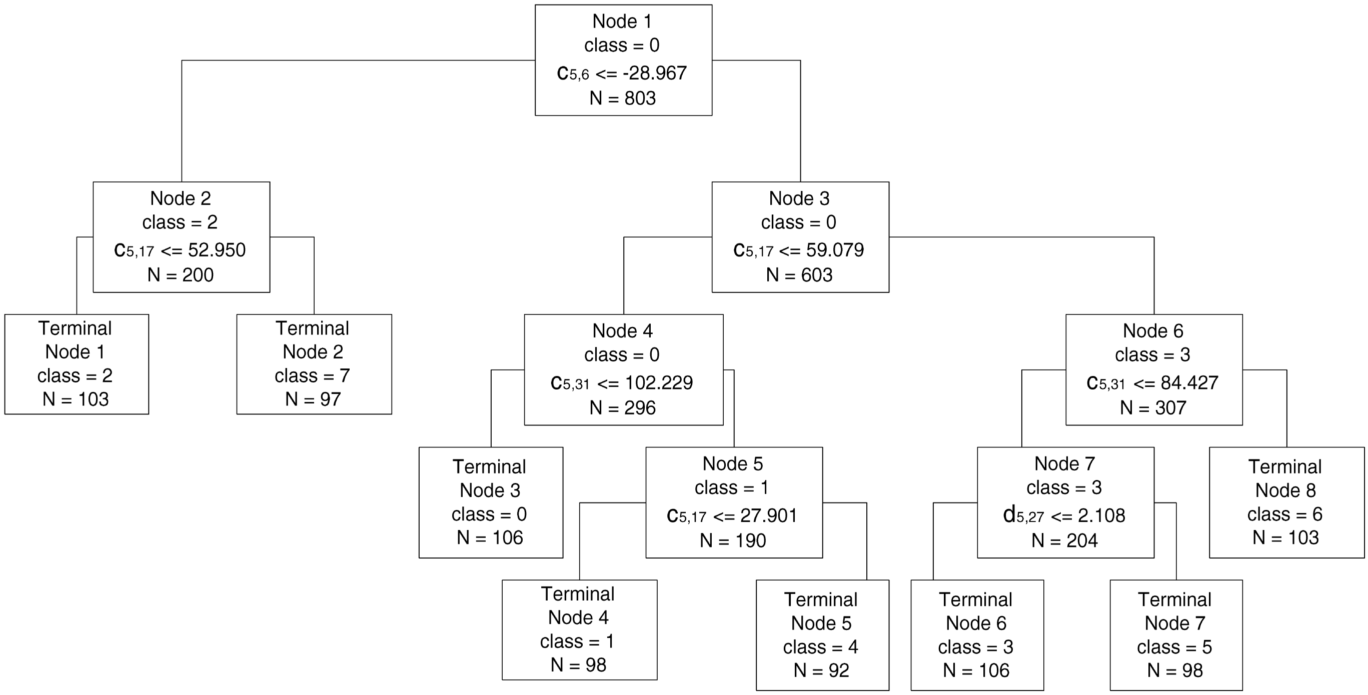

Figure 10. CART Tree in the Wavelet Domain.

time needed to construct the decision tree by a factor of 10 compared with working with the reduced-size data (55 vs. 5 seconds); it took only 1 second to obtain the reduced-size dataset by applying the DWT and the RRE_s method. The interpretation of Figure 11 is somewhat different from that of Figure 10. In node 1, the first split is $t_{394} \le -12.283$, where t_{394} is the value

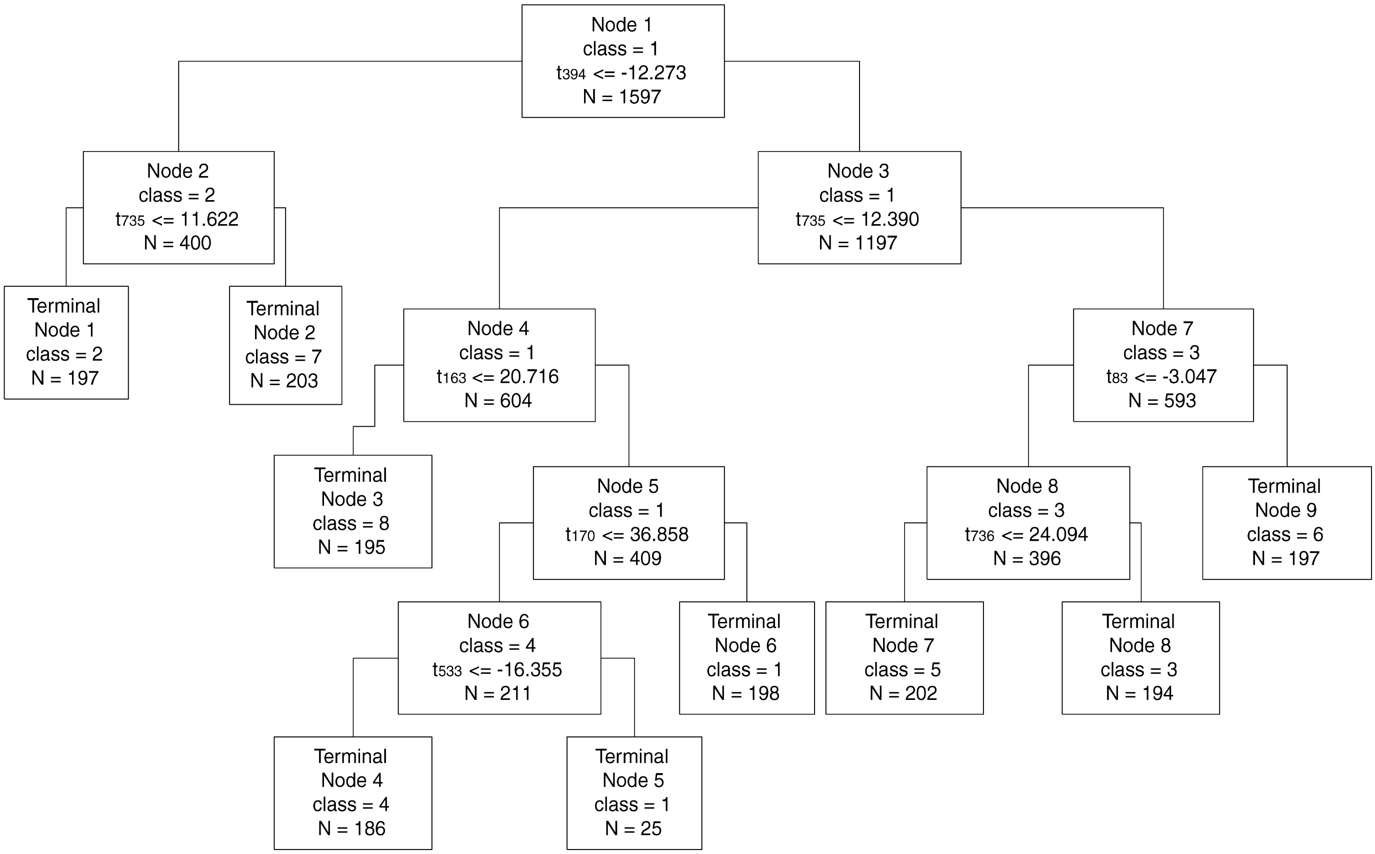

Figure 11. CART Tree in the Time Domain.

Table 6. Misclassification Error (%)

	Training data		Testing data	
Class	*Wavelet*	*Time*	*Wavelet*	*Time*
Original	0	0	2.06	3.09
1	5.10	4.08	8.42	8.91
2	0	0	.51	0
3	0	0	0	0
4	5.43	3.26	6.25	12.02
5	0	0	0	0
6	0	0	0	.51
7	0	0	.49	0
Total error	1.25	.87	2.25	3.13

of the signal at time 394. In node 2, if $t_{735} \le 11.622$, then the signal is classified as class 2; otherwise, the signal is classified as 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 10.

The misclassification rates in the wavelet and time domains and in the training and testing samples are given in Table 6. The CART tree in the time domain was almost perfect with respect to the training data, but it over-adapted 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 that obtained using the original time domain data in the training samples but is smaller (2.25% vs. 3.13%) in the testing samples. The existence of noise in signals makes classification in the time domain difficult.

Remark 5. Our procedures were compared with the principal coordinates approach based on the function data-analytic method proposed by Hall, Poskitt, and Presnell (2001). Their method approximates the signal using the first M Karhunen–Loève basis functions, with M determined from 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 specified in Section 5.1, CART's total misclassification rates for their and our methods are 2.82% and 2.25%. 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. Similar observations were obtained from normal distribution-based quadratic discriminant analysis advocated by Hall et al. (2001), which has a much higher total misclassification rate (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.

5.2 Multiresolution Fault Detection Using a Thresholded Scalogram

One deficiency inherent in wavelet bases 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; see Fig. 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 shifts in time) signal feature for fault detection.

One advantage of wavelet transforms is the multiresolution decomposition of complicated data signals. Information contained in each resolution could be useful in different types of fault detection; for example, the coarser-scale coefficients represent the global shape of the signal in the lower (coarser) resolution level, whereas the fine-scale coefficients represent the details of the signal in the higher (finer) resolution level. We therefore propose using the following scalogram (Vidakovic 1999, p. 289) for fault detection:

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

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., $S_{c_L} = \sum_{k=0}^{2^L-1} c_{L,k}^2$). The scalogram is a commonly used tool in signal and image processing (Rioul and Vetterli 1991), astronomy, and meteorology 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-sized data and 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 determined (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 12 presents a thresholded scalogram plot (in a $\log_2$-scale) of the RTCVD experimental data from three fault classes. Figure 13 shows the data curves obtained from the nominal and three fault classes. Comparably, the scalogram values for the data in the fault class 3 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 because of 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 the thresholded scalograms in process monitoring. The

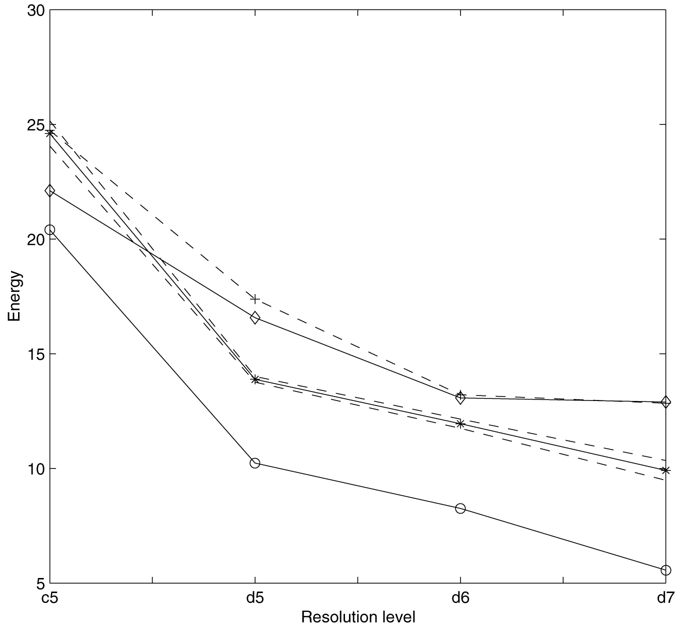

Figure 12. Thresholded Scalograms With Pointwise Confidence Intervals (∗ nominal; ◇ fault 1; + fault 2; ○ fault 3).

proof is based on a probability argument to establish the asymptotic equivalence between $S_j^*(\hat{\lambda})$ and $S_j^*(\lambda)$, and on validation of the Lindeberg condition (as seen in the proof of Thm. 2) for $S_j^*(\lambda)$ (see Jeong et al. 2003 for details).

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

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

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.

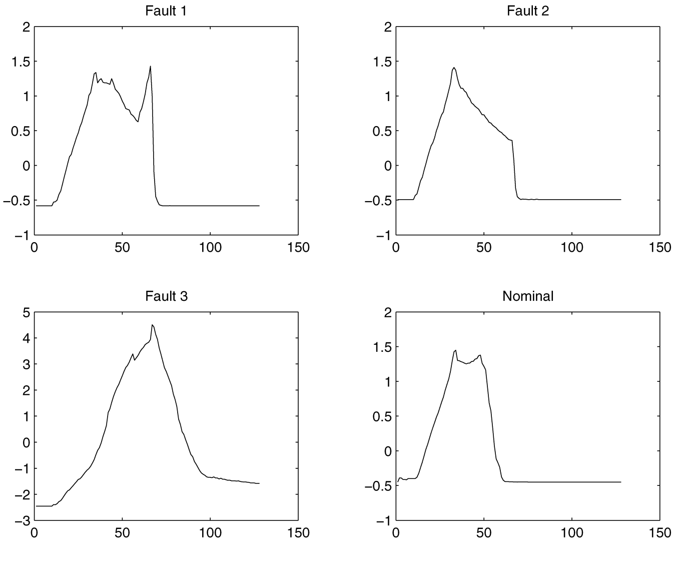

Figure 13. RTCVD Signals in Fault Classes.

Figure 12 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 4 for details) in analyzing the RTCVD data, it was used in this example for the thresholding. Even with limited data size, the monitoring bounds constructed from the approximated distribution are rather tight. Results plotted in Figure 12 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 has proposed an approach to handling a special type of large and complicated functional data in data analysis and decision making. Properties of the proposed data reduction methods have been 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 more favorable computational efficiency compared with the results obtained from analyzing the original larger-sized 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 multiresolution SPC for spatial image data in process monitoring). We will also consider extending the above to high-dimensional data, for example, imagery dataset. Newly developed multiscale methods for high-dimensional data, such as beamlets, wedgelets (Donoho and Huo 2001), and so on will be explored.

ACKNOWLEDGMENTS

The authors thank the anonymous referees and the associate editor for their very valuable comments and suggestions. This research was supported by National Science Foundation grants EEC-0080315, DMS-00-72960, and DMI-0400071.

APPENDIX: EXTENSION OF THE *RRE_h* METHOD TO A SOFT–THRESHOLDING–BASED METHOD, *RRE_s*

A similar idea presented for RRE_h can be extended to the soft-thresholding idea. In the wavelet shrinkage literature, it has been shown that hard thresholding results in a larger variance of estimates, whereas soft thresholding has a larger bias. Hard thresholding is also very sensitive to small changes in the data. Soft thresholding has various advantages, such as continuity of the shrinkage rule. Bruce and Gao (1996) provided a comparison study of these two thresholding policies in data denoising applications; see Tables 3–5 for their comparisons in data reduction applications. The analytical properties of RRE_s can be derived as presented in Theorem A.1. Denote

$\hat{\mathbf{d}}_s(\lambda) = (\hat{d}_{s,1}(\lambda), \dots, \hat{d}_{s,N}(\lambda))^\top$, where $\hat{d}_{s,i}(\lambda) = I(|d_i| > \lambda) \times \operatorname{sign}(d_i)(|d_i| - \lambda)$, $i = 1, \dots, N$. Then

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

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

Theorem A.1. Consider the model stated in (3). Then we have the following:

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

$$\lambda_{N,s} = .5 \cdot \left(\frac{\mathrm{E}\|\mathbf{d}\|^2}{\mathrm{E}\|\mathbf{d}\|_1}\right)^{1/2}; \tag{A.2}$$

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

$$\hat{\lambda}_{N,s} = .5 \cdot \left(\frac{\sum_{i=1}^N d_i^2}{\sum_{i=1}^N |d_i|}\right)^{1/2} = .5 \cdot \left(\frac{\hat{\xi}}{l_1}\right)^{1/2}, \tag{A.3}$$

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

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

Proof of Theorem 1

In this proof we first focus on the stochastic case first, then address the modification of the proof for the deterministic case. Let $d_{(1)}^2 \ge d_{(2)}^2 \ge \cdots \ge d_{(N)}^2$ be the ordered energies of wavelet coefficients. Because

$$\begin{aligned}
\mathrm{E}(\hat{\xi}) &= \mathrm{E}\|\mathbf{y}\|^2 = \mathrm{E}\|\mathbf{d}\|^2 \\
&= \sum_{i=1}^N \mathrm{E}(d_i^2) = \sum_{i=1}^N \mathrm{E}\big(d_{(i)}^2\big) \\
&\ge \sum_{i=1}^M \mathrm{E}\big(d_{(i)}^2\big) \ge M\mathrm{E}\big(d_{(M)}^2\big),
\end{aligned}$$

the inequality, $\mathrm{E}(d_{(M)}^2) \le \mathrm{E}(\hat{\xi})/M$, holds for $M = 1, 2, \dots, N$. Therefore,

$$\begin{aligned}
\mathrm{E}\|\mathbf{y} - \hat{\mathbf{f}}_M\|^2 &= \sum_{i=M+1}^N \mathrm{E}\big(d_{(i)}^2\big) \le \sum_{i=M+1}^N \frac{\mathrm{E}(\hat{\xi})}{i} \\
&\le \frac{(N-M)\mathrm{E}(\hat{\xi})}{M}.
\end{aligned}$$

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

Proof of Theorem 2

Denote

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

and

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

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

$$\begin{aligned}
\mathrm{E}\|\mathbf{d} - \hat{\mathbf{d}}_h(\lambda)\|^2 &= \sum_{i=1}^N \mathrm{E}\big(d_i - I(|d_i| > \lambda)d_i\big)^2 \\
&= \sum_{i=1}^N \mathrm{E}\big(I(|d_i| \le \lambda)d_i^2\big) = \sum_{i=1}^N H_i(\lambda)
\end{aligned}$$

and

$$\mathrm{E}\|\hat{\mathbf{d}}_h(\lambda)\|_0 = \sum_{i=1}^N \mathrm{E}\big(|\hat{d}_i(\lambda)|_0\big) = \sum_{i=1}^N \mathrm{E}\big(I(|d_i| > \lambda)\big) = \sum_{i=1}^N h_i(\lambda).$$

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

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

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

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

only if

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

Because the d_i's are independently $\mathrm{N}(\theta_i, \sigma^2)$ distributed, $N\hat{\lambda}_{N,h}^2/\sigma^2 = \sum_{i=1}^N d_i^2/\sigma^2$ is $\chi^2(N, \delta_N)$ distributed with degree of freedom N and noncentrality parameter $\delta_N = \sum_{i=1}^N \theta_i^2/\sigma^2$. It follows that $\mathrm{E}(\hat{\lambda}_{N,h}^2) = \sigma^2(\delta_N/N + 1) = \lambda_N$ and $\operatorname{var}(\hat{\lambda}_{N,h}^2) = \sigma^4(4\delta_N + 2N)/N^2 \to 0$, 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 the DWT is orthonormal, $|\theta_i|$, $i = 1, 2, \dots, N$, should be uniformly bounded as $N \to \infty$. Without loss of generality, we assume that $|\theta_i| < K$, $i = 1, 2, \dots, N$. 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

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

Thus, from the Kolmogorov theorem (Serfling 1980, p. 27), we know that $(\hat{\lambda}_{N,h} - \lambda_{N,h}) \xrightarrow{\text{w.p.1}} 0$; that is, the result (b) is true.

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 Lindeberg condition (Serfling 1980, p. 30) that, for every $\varepsilon > 0$,

$$\frac{1}{N}\sum_{i=1}^{N}\int_{|t^2-\mu_i|>\varepsilon\sqrt{N}}(t^2-\mu_i)^2\phi\left(\frac{t-\theta_i}{\sigma}\right)dt \to 0, \qquad N\to\infty, \quad \text{(A.4)}$$

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 $N\to\infty$,

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

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

Proof of Theorem A.1

Denote

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

According to the intervals of d_i, the term $I(|d_i| > \lambda) \times \operatorname{sign}(d_i)(|d_i| - \lambda)$ can be defined as

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

Then

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

Because

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

$RRE_s(\lambda)$ can be written as

$$\begin{aligned}RRE_s(\lambda) = &\left(\sum_{i=1}^{N}H_i(\lambda) + \lambda^2\sum_{i=1}^{N}h_i(\lambda)\right)\Big/\mathrm{E}(\|\mathbf{d}\|^2)^{1/2}\\ &+ \left(\sum_{i=1}^{N}V_i(\lambda)\right)\Big/\mathrm{E}(\|\mathbf{d}\|_1)^{1/2}.\end{aligned}$$

Because

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

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

only if

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

In addition, similar to the proof of result (b) of Theorem 2, we know that $(\hat{\lambda}_{N,s} - \lambda_{N,s}) \xrightarrow{\text{w.p.1}} 0$ from the Kolmogorov theorem and Slutsky's theorem; that is, result (b) is true.

[Received November 2002. Revised May 2005.]

REFERENCES

Antoniadis, A., Gijbels, I., and Grégoire, G. (1997), "Model Selection Using Wavelet Decomposition and Applications," *Biometrika*, 84, 751–763.

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

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

Cherkassky, V., and Shao, X. (2001), "Signal Estimation and Denoising Using VC-Theory," *Neural Networks*, 14, 37–52.

Craven, P., and Wahba, G. (1979), "Smoothing Noisy Data With Spline Functions," *Numerische Mathematik*, 31, 377–403.

Donoho, D. L., and Huo, X. (2001), "Beamlets and Multiscale Image Analysis," in *Multiscale and Multiresolution Methods*, eds. T. J. Barth, T. Chan, and R. Haimes, New York: Springer-Verlag, pp. 149–196.

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

——— (1995), "Adapting to Unknown Smoothness via Wavelet Shrinkage," *Journal of the American Statistical Association*, 90, 1200–1224.

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

Ganesan, R., Das, T. K., Sikdar, A., and Kumar, A. (2003), "Wavelet-Based Detection of Delamination Defect in CMP Using a Nonstationary Acoustic Emission Signal," *IEEE Transactions on Semiconductor Manufacturing*, 16, 677–685.

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

Hastie, T., Tibshirani, R., and Friedman, J. (2001), *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, New York: Springer-Verlag.

Ihara, I. (1993), *Information Theory for Continuous Systems*, NJ: World Scientific.

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

Jeong, M. K., and Lu, J.-C. (2004), "Adaptive SPC Procedures for Complicated Functional Data," technical report, School of Industrial and Systems Engineering, Georgia Institute of Technology.

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

——— (2001), "Automatic Feature Extraction of Waveform Signals for In-Process Diagnostic Performance Improvement," *Journal of Intelligent Manufacturing*, 12, 257–268.

Jung, U., Lu, J.-C., and Jeong, M. K. (2004), "A Wavelet-Based Random-Effect Model for Multiple Sets of Complicated Functional Data," technical report, School of Industrial and Systems Engineering, Georgia Institute of Technology, available at *http://www.isye.gatech.edu/*.

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*, 121, 295–299.

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

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

Lu, J.-C. (2001), "Methodology of Mining Massive Data Set for Improving Manufacturing Quality/Efficiency," in *Data Mining for Design and Manufacturing: Methods and Applications*, ed. D. Braha, New York: Kluwer Academic, pp. 255–288.

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

Portilla, J., and Simoncelli, E. P. (2000), "Image Denoising via Adjustment of Wavelet Coefficient Maginitude Correlation," in *Center for Neural Science, and Proceedings of the 7th International Conference on Image Processing*, Vancouver, Canada, pp. 277–280.

Rioul, O., and Vetterli, M. (1991),"Wavelets and Signal Processing," *IEEE Transactions in Signal Processing*, 8, 14–38.

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 doctoral thesis, North Carolina State University, Dept. of Electrical and Computer Engineering.

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*, Pennington, New Jersey, pp. 37–44.

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*, eds. E. Foufoula-Georgiou and P. Kumar, New York: Academic Press, pp. 299–324.

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

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

Vidakovic, B. (1999), *Statistical Modeling by Wavelets*, New York: Wiley.

——— (2000), "Unbalancing Data With Wavelet Transformations," technical report, Duke University, Dept. of Statistics.

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