APPLIED STOCHASTIC MODELS IN BUSINESS AND INDUSTRY
Appl. Stochastic Models Bus. Ind., 2003; **19**:231–244 (DOI: 10.1002/asmb.495)

Thresholded scalogram and its applications in process fault detection

Myong K. Jeong*,†, Di Chen and Jye-Chyi Lu

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

SUMMARY

Scalograms provide measures of signal energy at various frequency bands and are commonly used in decision making in many fields including signal and image processing, astronomy and metrology. This article extends the scalogram's ability for handling noisy and possibly massive data. The proposed thresholded scalogram is built on the fast wavelet transform, which can capture non-stationary changes in data patterns effectively and efficiently. The asymptotic distribution of the thresholded scalogram is derived. This leads to large sample confidence intervals that are useful in detecting process faults statistically, based on scalogram signatures. Application of the scalogram-based data mining procedure (mainly, classification and regression trees) demonstrates the potential of the proposed methods for analysing complicated signals for making engineering decisions. Copyright © 2003 John Wiley & Sons, Ltd.

KEY WORDS: asymptotic normality; classification; data mining; discrete wavelet transform; pattern recognition; scalogram; signal processing

1. INTRODUCTION

Many data signals collected from manufacturing processes are non-stationary and correlated. Many researchers have recommended wavelet-based methods to analyse this type of data (see e.g. Reference [1]). Wavelet transforms of a signal have multi-resolution representation allowing decision-makers to use the information contained in each resolution for signal classification. For example, process fault patterns, which are frequency or phase shifted and invisible to time domain monitoring or control procedures, could be easily detected by wavelet transforms. In particular, Koh *et al.* [2] indicates that because of its computational efficiency the discrete wavelet transform (DWT) is very useful in on-line (real-time) process monitoring.

One deficiency in the procedures developed from the wavelet coefficients provided from the DWT is the lack of any shift-invariance. To elaborate, consider two signals slightly shifted in time. Energy values (e.g. the sum of squared wavelet coefficients) at various frequency scales (or

*Correspondence to: Myong K. Jeong, School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA 30332-0205, U.S.A.
† E-mail: mkjeong@isye.gatech.edu

 Accepted 1 July 2003

resolution levels) show no difference between the two signals, i.e. the energy is shift-invariant. However, when these signals are transformed and decomposed via the DWT, there is clearly an appreciable difference between the two representations based on individual wavelet coefficients. Therefore, direct assessment of the wavelet coefficients often leads to inaccurate decisions. Thus, a scale-wise energy representation such as a scalogram provides a more robust signal feature for fault detection against time-shift than the DWT coefficients directly.

Scalograms, defined in Section 2, are commonly used in many fields such as signal and image processing [3] and astronomy and metrology [4]. In DWT applications, scalograms measure signal energy contained at various frequency bands with different sizes of scale in wavelet transforms. Intuitively, it is possible that the scalogram will be useful in monitoring process changes with data collected in time sequences. See Section 4 for some successful examples.

In estimating a signal's functional pattern with noisy data, Donoho and Johnstone [5] proposed a data denoising procedure based on the idea of thresholding out secondary wavelet coefficients representing data noises. In many applications in data mining, the large size of non-stationary data makes computations inefficient (see Reference [6] for an example). Extending the usefulness of the popular scalogram to noisy and possibly massive data, this article develops a thresholded scalogram and studies its properties and applicability to engineering decision making.

The rest of this paper proceeds as follows. Section 2 reviews the basics of wavelets and establishes the statistical model of the thresholded scalogram. Section 3 discusses the asymptotic properties of the estimates of the thresholded scalogram. Section 4 describes the applications of the developed procedure in detecting and classifying process faults. Some concluding remarks are in Section 5.

2. WAVELET TRANSFORMS AND THRESHOLDED SCALOGRAMS

2.1. Wavelet transforms

Wavelets are localized basis functions that are translated and dilated versions of some fixed mother wavelet. Some signals often show a non-stationary and transient nature and carry small yet informative components embedded in larger repetitive signals. Wavelets provide flexible time-frequency resolution and comprise an efficient alternative in quantifying such transient signals.

Suppose the functions

$$\{\phi_{L,k}(t), \psi_{j,k}(t)\}_{(j \geqslant L, k \in \mathbb{Z})}$$

are an orthonormal basis in $L^2(\mathbb{R})$. Any function $f(t)$ in $L^2(\mathbb{R})$ can be represented as follows:

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

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

used:

$$\tilde{f}(t) = \sum_{k=0}^{2^L-1} c_{L,k}\phi_{L,k}(t) + \sum_{j=L}^{J}\sum_{k=0}^{2^j-1} d_{j,k}\psi_{j,k}(t) \tag{1}$$

where $J > L$ and L corresponds to the lowest decomposition level.

At time t, assume that signal $y(t)$ can be decomposed into signal plus noise: $y(t) = f(t) + \varepsilon_t$, where $f(t)$ is the true signal space and ε_t is random noise. Consider a sequence of data $y = (y(t_1), \ldots, y(t_N))^\top$ obtained as a realization of $y(t)$ at equally spaced discrete time points. Assume that $y(t_j)$ are independently $N(f(t_j), \sigma^2)$ distributed, $j = 1, \ldots, N$. In this article, $\top$ is used to denote the vector transpose. The DWT of $\mathbf{y}$ is defined as $\mathbf{d} = \mathbf{W}\mathbf{y}$, where $\mathbf{W}$ is the orthonormal $N \times N$ DWT-matrix. Let $\mathbf{f} = (f(t_1), \ldots, f(t_N))^\top$ and $\boldsymbol{\varepsilon} = (\varepsilon(t_1), \ldots, \varepsilon(t_N))^\top$. Then, we have that

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

where $\boldsymbol{\theta} = \mathbf{W}\mathbf{f}$, and $\boldsymbol{\eta} = \mathbf{W}\boldsymbol{\varepsilon}$ is $\mathbf{N}_N(0, \mathbf{I_N}\sigma^2)$ distributed.

In Equation (1), let

$$\mathbf{d} = (\mathbf{c}_L, \mathbf{d}_L, \mathbf{d}_{L+1}, \ldots, \mathbf{d}_J)^\top \tag{3}$$

where $\mathbf{c}_L = (c_{L,0}, \ldots, c_{L,2^L-1})^\top$, $\mathbf{d}_L = (d_{L,0}, \ldots, d_{L,2^L-1})^\top, \ldots, \mathbf{d}_J = (d_{J,0}, \ldots, d_{J,2^J-1})^\top$ are wavelet coefficients at various scales or subbands. The total number of wavelet coefficients is equal to the number of signal measurements, i.e. $N = 2^{J+1}$. The $c_{L,k}$'s capture the low frequency oscillations, while $d_{j,k}$'s capture the high frequency oscillations. The coefficients $d_{J,k}$'s represent the finest scale and the $c_{L,k}$'s represent the coarsest scale [7]. To simplify the notation, we use $\mathbf{d} = (d_1, d_2, \ldots, d_N)^\top$ instead of using c_{Lk}, d_{jk} for the components of $\mathbf{d}$. Using the inverse DWT, the $N \times 1$ vector $\mathbf{y}$ of the original signal curve can be reconstructed as $\mathbf{y} = \mathbf{W}^\top\mathbf{d}$.

2.2. Thresholded scalograms

Scalograms represent the scale-wise distribution of energies. Scalograms at scale j are defined as [8, p. 289]

$$S_{dj} = \sum_{k=0}^{2^j-1} d_{jk}^2,\ j = L, L+1, \ldots, J, \quad \text{and} \quad S_{cL} = \sum_{k=0}^{2^L-1} c_{L,k}^2$$

where S_{cL} is the energy at the coarsest level. Thus, the total energy of the signal, $\|\mathbf{y}\|^2$, can be decomposed among the resolution levels as follows:

$$\|\mathbf{y}\|^2 = \|\mathbf{W}\mathbf{y}\|^2 = \|\mathbf{d}\|^2 = \sum_{j=L}^{J} S_{dj} + S_{cL}$$

For analysing potentially massive data and for removing secondary noises, we propose the following thresholded scalogram:

$$S_j^*(\lambda) = \sum_{k=0}^{m_j-1} \mathbf{I}(|d_{jk}| > \lambda) d_{jk}^2 \tag{4}$$

where λ is the threshold value that can be decided by various methods (see e.g. References [5, 9, 10, 11]) and $m_j = 2^j$ is the number of wavelet coefficients at the jth resolution level. This screening of smaller wavelet coefficients makes the detection of process fault more robust in a noisy environment. Next, we will present the optimal threshold value based on the new criterion that requires balancing data denoising and data reduction goals.

2.3. Thresholding parameter

For a given λ, let $\hat{\mathbf{d}}(\lambda) = (\hat{d}_1(\lambda), \ldots, \hat{d}_N(\lambda))^\top$, where $\hat{d}_i(\lambda) = I(|d_i| > \lambda)d_i$, $i = 1, \ldots, N$, be the thresholded wavelet coefficients. Only coefficients larger than λ are kept in subsequent analyses for engineering decision making. Thus, in examples illustrated in Section 4, the computation in decision-making analysis (e.g. the decision tree in Section 4.2) using the thresholded scalograms is much more efficient than the use of all original data in the time domain.

In many engineering applications, the relative error,

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

is commonly used in comparing signal approximation quality. See Reference [12, pp. 378–391] for an example of using the relative error in evaluating signal approximation methods. In the equation given below, this type of relative error is used to quantify the modeling accuracy in our data reduction procedure.

Jeong *et al.* [11] used several real-life data sets and testing curves commonly cited in the literature (e.g. Reference [9]) to formulate the following objective function for data denoising and data reduction:

$$R_0(\lambda) = \frac{E(||\mathbf{d} - \hat{\mathbf{d}}(\lambda)||^2)}{E(||\mathbf{d}||^2)} + E\left(\frac{||\hat{\mathbf{d}}(\lambda)||_0}{N}\right)$$

where $||\hat{\mathbf{d}}(\lambda)||_0 = \sum_{i=1}^{N} |\hat{d}_i(\lambda)|_0$, and $|\hat{d}_i(\lambda)|_0 = 1$, if $\hat{d}_i(\lambda) \neq 0$; $|\hat{d}_i(\lambda)|_0 = 0$, otherwise. Note that $||\hat{\mathbf{d}}(\lambda)||_0$ is nothing but the number of non-zero $\hat{d}_i(\lambda)$'s. For simplicity, $R_0(\lambda)$ uses an equal weight between the relative error in its first component and the data reduction ratio in its second component. Jeong *et al.* [11] derived the following theorem for understanding the properties of the optimal λ.

Theorem 1

Consider the model stated in (2). Then,

(i) the objective function $R_0(\lambda)$ is minimized uniquely at $\lambda = \lambda_R$ where

$$\lambda_R = \left(\frac{1}{N}\sum_{i=1}^{N} \theta_i^2 + \sigma^2\right)^{1/2} \tag{5}$$

The moment estimate of λ_R,

$$\hat{\lambda}_R = \left(\frac{1}{N}\sum_{i=1}^{N} d_i^2\right)^{1/2} \tag{6}$$

has the following properties:

(ii) $(\hat{\lambda}_R - \lambda_R) \overset{\text{w.p.1}}{\to} 0$;

(iii) $\sqrt{N}(\hat{\lambda}_R - \lambda_R)/\sigma_N^* \overset{d}{\to} N(0,1)$, where

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

Remark

The quantity λ in Equation (4) could be replaced by any estimate such as $\hat{\lambda}_R$ from the above method or $\hat{\lambda}_D$ from Reference [9]. When the estimate converges to its parameter λ_R or λ_D with probability one, the following lemma and theorems will carry through. Thus, for the remainder of this article, we will use λ_N and $\hat{\lambda}_N$ to represent these parameters and estimates.

3. ASYMPTOTIC PROPERTIES OF THRESHOLDED SCALOGRAMS

Replacing λ by $\hat{\lambda}_N$ in (4), we obtain the following applicable thresholded scalogram:

$$\hat{S}_{Nj}^*(\hat{\lambda}_N) = \sum_{k=0}^{m_j-1} I(|d_{jk}| > \hat{\lambda}_N) d_{jk}^2$$

For the convenience of notation, let

$$S_{Nj}^* = \sum_{k=0}^{m_j-1} I(|d_{jk}| > \lambda_N) d_{jk}^2 \tag{7}$$

which represents the thresholded scalogram with respect to a thresholding parameter λ_N.

Note that d_{jk}'s are independent, but are not identically distributed because of different means. First, we will use the following lemma to show that the difference between S_{Nj}^* and $\hat{S}_{Nj}^*$ converges to zero with probability one. Then, we will focus on the derivation of the asymptotic distribution of S_{Nj}^*.

Lemma 1

Assume that $\{d_i : i = 1, 2, \ldots\}$ is a series of independent random variables with different means and the same variance. For a fixed index i, let $\hat{Y}_{Ni} = I(|d_i| \geqslant \hat{\lambda}_N)$ and $Y_{Ni} = I(|d_i| \geqslant \lambda_N)$, where $\hat{\lambda}_N$ and λ_N are defined in (5) and (6), respectively. Then,

$$\hat{Y}_{Ni} - Y_{Ni} \overset{\text{w.p.1}}{\to} 0, \quad \text{as } N \to \infty$$

Proof

For any $\varepsilon > 0$,

$$\Pr\left\{\lim_{N\to\infty} |\hat{Y}_{Ni} - Y_{Ni}| > \varepsilon\right\}$$

$$= \Pr\left\{\lim_{N\to\infty} (\hat{Y}_{Ni} = 1, Y_{Ni} = 0)\right\} + \Pr\left\{\lim_{N\to\infty} (\hat{Y}_{Ni} = 0, Y_{Ni} = 1)\right\}$$

$$= \Pr\left\{\lim_{N\to\infty} (|d_i| \geqslant \hat{\lambda}_N, |d_i| < \lambda_N)\right\} + \Pr\left\{\lim_{N\to\infty} (|d_i| < \hat{\lambda}_N, |d_i| \geqslant \lambda_N)\right\}$$

$$= \Pr\left\{\lim_{N\to\infty} (\hat{\lambda}_N \leqslant |d_i| < \lambda_N)\right\} + \Pr\left\{\lim_{N\to\infty} (\lambda_N \leqslant |d_i| < \hat{\lambda}_N)\right\}$$

$$\leqslant \Pr\left\{\lim_{N\to\infty} (\hat{\lambda}_N < \lambda_N)\right\} + \Pr\left\{\lim_{N\to\infty} (\lambda_N < \hat{\lambda}_N)\right\}$$

$$= \Pr\left\{\lim_{N\to\infty} (\hat{\lambda}_N - \lambda_N \neq 0)\right\}$$

$$= 0$$

The last equation follows from $(\hat{\lambda}_N - \lambda_N) \overset{\text{w.p.1}}{\to} 0$. □

Based on this lemma, the corresponding result for the thresholded scalogram can be obtained immediately. Theorem 2 states this result without proof.

Theorem 2

Under the same conditions in Theorem 1, for fixed j, $j = L, L+1, \ldots, J$, we have

$$(\hat{S}^*_{Nj} - S^*_{Nj}) \overset{\text{w.p.1}}{\to} 0, \quad \text{as } N \to \infty$$

Next, we derive the asymptotic distribution of S^*_{Nj}. Recall that d_{jk}'s in S^*_{Nj} are independently distributed as normal with mean θ_{jk} and common variance σ^2. Then, we have

$$E(S^*_{Nj}) = \sum_{k=0}^{m_j-1} E(I(|d_{jk}| \geqslant \lambda_N) d^2_{jk}) = \sum_{k=0}^{m_j-1} \int_{|t|\geqslant\lambda_N} t^2 \frac{1}{\sigma}\phi\left(\frac{t-\theta_{jk}}{\sigma}\right) \mathrm{d}t$$

$$= \sum_{k=0}^{m_j-1} (\theta^2_{jk} + \sigma^2) - \sum_{k=0}^{m_j-1} \int_{|t|<\lambda_N} t^2 \frac{1}{\sigma}\phi\left(\frac{t-\theta_{jk}}{\sigma}\right) \mathrm{d}t$$

and

$$\begin{aligned}
\mathrm{Var}(S_{Nj}^*) &= \sum_{k=0}^{m_j-1} \mathrm{Var}(I(|d_{jk}| \geqslant \lambda_N) d_{jk}^2) \\
&= \sum_{k=0}^{m_j-1} (E(I(|d_{jk}| \geqslant \lambda_N) d_{jk}^4) - E^2(I(|d_{jk}| \geqslant \lambda_N) d_{jk}^2)) \\
&= \sum_{k=0}^{m_j-1} \left[E(d_{jk}^4) - \int_{|t|<\lambda_N} t^4 \frac{1}{\sigma} \phi\left(\frac{t-\theta_{jk}}{\sigma}\right) \mathrm{d}t \right] \\
&\quad - \sum_{k=0}^{m_j-1} \left[\theta_{jk}^2 + \sigma^2 - \int_{|t|\leqslant\lambda_N} t^2 \frac{1}{\sigma} \phi\left(\frac{t-\theta_{jk}}{\sigma}\right) \mathrm{d}t \right]^2 \\
&= \sum_{k=0}^{m_j-1} \left[(3\sigma^4 + 6\sigma^2\theta_{jk}^2 + \theta_{jk}^4) - \int_{|t|<\lambda_N} t^4 \frac{1}{\sigma} \phi\left(\frac{t-\theta_{jk}}{\sigma}\right) \mathrm{d}t \right] \\
&\quad - \sum_{k=0}^{m_j-1} \left[\theta_{jk}^2 + \sigma^2 - \int_{|t|\leqslant\lambda_N} t^2 \frac{1}{\sigma} \phi\left(\frac{t-\theta_{jk}}{\sigma}\right) \mathrm{d}t \right]^2
\end{aligned}$$

These support the proof of the following theorem.

Theorem 3

Let $\eta_j = E(S_{Nj}^*) \geqslant 0$ and assume that $\mathrm{Var}(S_{Nj}^*)/m_j \to \sigma_j^2$, as $m_j \to \infty$. Then, under the same conditions in Theorem 1, we have

$$\frac{\eta_j(\ln S_{Nj}^* - \ln \eta_j)}{\sqrt{m_j}\,\sigma_j} \xrightarrow{d} N(0,1), \quad \text{as } m_j \to \infty$$

Proof

Let $X_{jk} = d_{jk}^2 I(|d_{jk}| > \lambda_N)$. These X_{jk}'s are independent random variables with the finite mean $E(X_{jk}) = \mu_{jk}$ and the finite variance $\mathrm{Var}(X_{jk}) = \sigma_{jk}^2$. Then, $\eta_j = E(S_{Nj}^*) = \sum_{k=0}^{m_j-1} \mu_{jk}$ and $\mathrm{Var}(S_{Nj}^*) = \sum_{k=0}^{m_j-1} \sigma_{jk}^2$. To show the asymptotic normality of $(S_{Nj}^* - \eta_j)/(\sqrt{m_j}\sigma_j)$ it is sufficient to verify the following Lindeberg condition [13, p. 30], for $\varepsilon > 0$,

$$\frac{1}{m_j} \sum_{k=0}^{m_j-1} \int_{|t^2 I(|t|>\lambda_N) - \mu_{jk}| > \varepsilon\sqrt{m_j}} [t^2 I(|t| > \lambda_N) - \mu_{jk}]^2 \phi\left(\frac{t-\theta_{jk}}{\sigma}\right) \mathrm{d}t \to 0, \quad m_j \to \infty$$

It follows that

$$\int_{|t^2 I(|t|>\lambda_N)-\mu_{jk}|>\varepsilon\sqrt{m_j}} [t^2 I(|t|>\lambda_N)-\mu_{jk}]^2 \phi\left(\frac{t-\theta_{jk}}{\sigma}\right) \mathrm{d}t$$

$$= O\left(\int_{t^2>\varepsilon\sqrt{m_j}} t^4 \phi\left(\frac{t-\theta_{jk}}{\sigma}\right) \mathrm{d}t\right) = O\left(\int_{t>\varepsilon^{1/2} m_j^{1/4}} t^4 \phi\left(\frac{t-\theta_{jk}}{\sigma}\right) \mathrm{d}t\right)$$

$$= O\left(\varepsilon^2 m_j \phi\left(\frac{\varepsilon^{1/2} m_j^{1/4}-\theta_{jk}}{\sigma}\right)\right) = O\left(\varepsilon^2 m_j \exp\left\{-\frac{\varepsilon\sqrt{m_j}}{2\sigma^2}\right\}\right)$$

Therefore, for $\varepsilon > 0$, as $m_j \to \infty$,

$$\int_{|t^2 I(|t|>\lambda_N)-\mu_{jk}|>\varepsilon\sqrt{m_j}} [t^2 I(|t|>\lambda_N)-\mu_{jk}]^2 \phi\left(\frac{t-\theta_{jk}}{\sigma}\right) \mathrm{d}t = O\left(\varepsilon^2 m_j \exp\left\{-\frac{\varepsilon\sqrt{m_j}}{2\sigma^2}\right\}\right) \to 0$$

Recall the delta method: if $(T_N - \eta_N)/\tau_N \to^d N(0,1)$, then $[h(T_N) - h(\eta_N)]/[\tau_N h'(\eta_N)] \xrightarrow{d} N(0,1)$ provided h is a continuous function such that $h'(\eta_N)$ exists and $h'(\eta_N) \neq 0$. By applying the delta method with $h(\eta_N) = \ln \eta_N$ and $h'(\eta_N) = 1/\eta_N$, we obtain the stated result. □

Corollary

Let $\eta^*_{Nj} = E(\hat{S}^*_{Nj}) \geqslant 0$, and assume that $\mathrm{Var}(\hat{S}^*_{Nj})/m_j \to (\sigma^*_{Nj})^2$, as $m_j \to \infty$. Then, under the same conditions in Theorem 1, we have

$$\frac{\eta^*_{Nj}(\ln \hat{S}^*_{Nj} - \ln \eta^*_{Nj})}{\sqrt{m_j}\,\sigma^*_{Nj}} \xrightarrow{d} N(0,1), \quad \text{as} \quad N \to \infty \quad \text{and} \quad m_j \to \infty$$

4. APPLICATION OF SCALOGRAMS FOR FAULT DETECTION AND CLASSIFICATION

4.1. Fault detection using thresholded scalograms

The asymptotic distribution of the thresholded scalograms can be used to establish the approximated $100(1-\alpha)\%$ confidence interval, $\ln \hat{S}^*_{Nj} \pm z_{\alpha/2}\hat{\sigma}^*_{Nj}/(\hat{\eta}^*_{Nj})$, where z_α is the upper $100\alpha\%$ percentile of the standard normal distribution. By connecting the point-wise interval values for each resolution level as shown in Figure 1, we can construct a set of lower and upper bounds of thresholded scalograms for the nominal run. This will serve as a tool to statistically detect process faults at several resolution levels. This idea is applied to a rapid thermal chemical vapour deposition (RTCVD) process that deposits thin films on semiconductor wafers using a temperature-driven surface chemical reaction. As feature size decreases, the functional operation of devices (e.g. transistors) will become increasingly susceptible to failure because of variations in deposition processes. Therefore, detecting a process condition different from the nominal is critical.

Quadruple mass spectrometry (QMS) is commonly used in the semiconductor manufacturing processes for monitoring thin-film deposition quality. Figure 2 presents some of the data collected by the QMS in a research project [14] to develop an *in situ* measurement technique for online process monitoring. The subfigures represent one of the 21 nominal RTCVD process runs and four sets of data from different faulty processes. Although only 128 data points are 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 to many engineering applications. More important, wavelet transforms have been proven to be useful in locating those change points for the *in situ* deposition thickness measurement tool [10] for thin films.

Comparably, the scalogram values for the data in Fault 3 class are much different from the nominal one at all resolution levels. Because of the similarity of the data curves in the original time domain, Fault classes 1 and 2 have similar scalogram values at the finer resolution levels, but not at the coarsest resolution level. The sharp drop in the data curve in Fault class 1 may partly explain why the value of its coarsest level scalogram is much different from its nominal value as compared with the value obtained from Fault class 2. Fault class 2 and the nominal curves have similar finer and coarsest level scalogram values, but this is not seen at the middle level of the scalograms. Results plotted in Figure 1 show that these four classes of data curves are clearly out of bounds at almost all resolution levels except at the coarsest level for Fault 2 class.

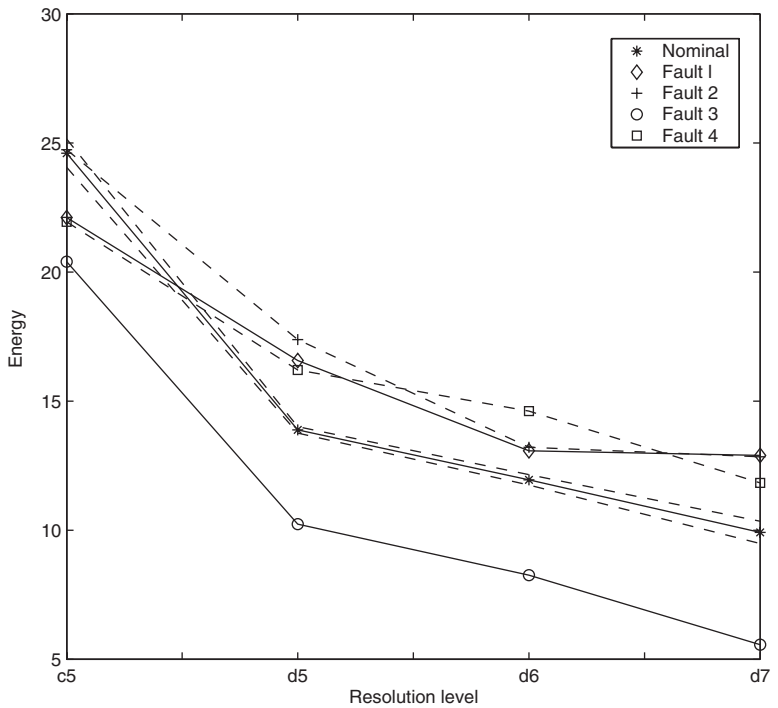

Figure 1. Point-wise confidence intervals of thresholded scalograms for the nominal run.

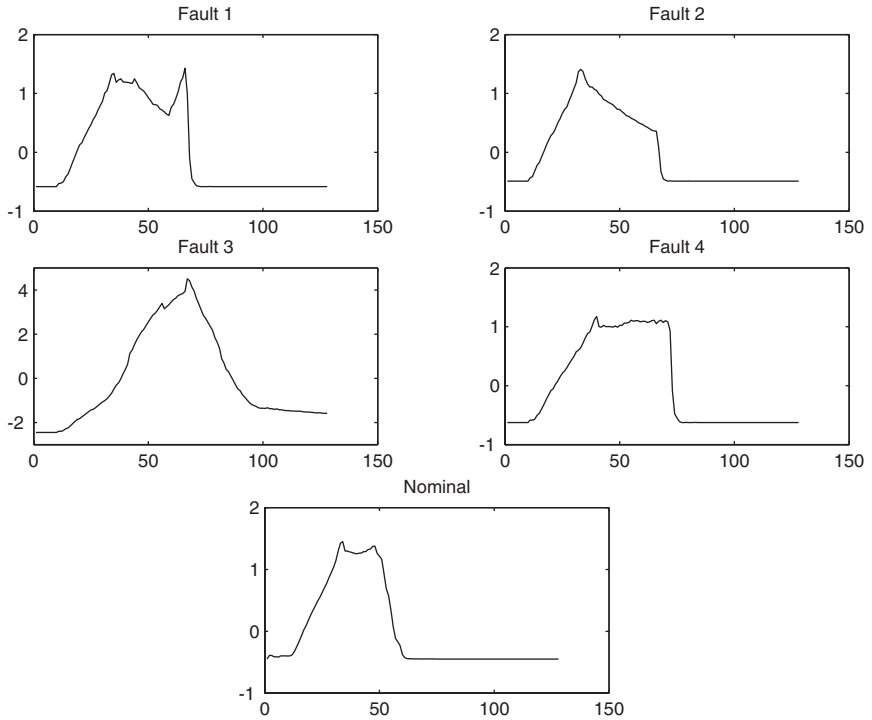

Figure 2. RTCVD signals.

4.2. Data mining using thresholded scalograms

This subsection presents another example with a testing curve [12, p. 378]. Figure 3 shows the curves from the nominal run (the original signal pattern as given in Mallat [12]) and three fault-situations artificially created for experimenting with the applicability and sensitivity of the proposed metric. Note that these testing curves have many sharp peaks and drops that are difficult for most statistical techniques to model well.

Figure 4 shows the results of a clustering analysis based on thresholded scalograms. Thresholded scalograms at the fifth level and at the sixth level of the data were used as features for clustering. This plot shows that these four signals can be well discriminated based on thresholded scalograms. This example illustrates the potential of the scalograms for signal classification.

Next, we apply a commonly used data mining tool classification and regression trees (CART) to the thresholded scalograms for analysing these signals. See Reference [15] for details of CART tree-building and pruning procedures.

For applying CART to the four classes of data curves presented in Figure 3, various curve-replicates were generated. In our experiment, all of the testing curves were shifted to the left (or right) in 5 (or 10, 15, 20, 25, 30) time-units (out of a total of $N = 1024$ units) for generating a new curve. Moreover, Gaussian random noises with $\sigma = 0.1$ are also added. Shifting the curves to the left and right artificially tested the invariant property of thresholded scalograms. For all curves in these four classes, the above data replication method was applied in order to generate

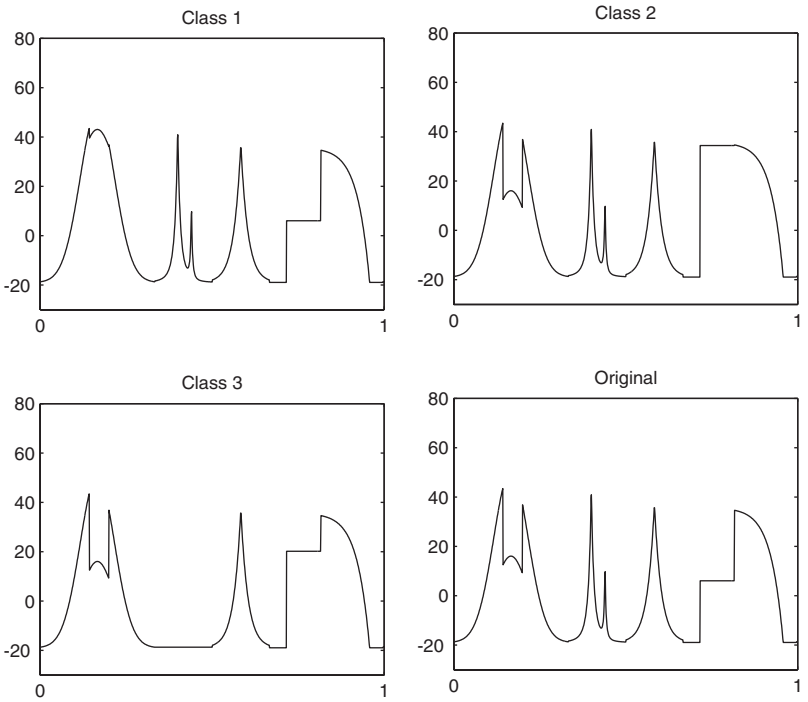

Figure 3. Four classes of piecewise signals.

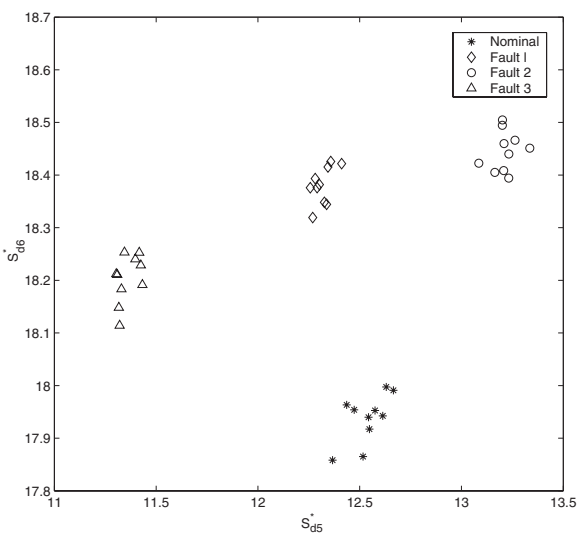

Figure 4. Clustering using thresholded scalograms.

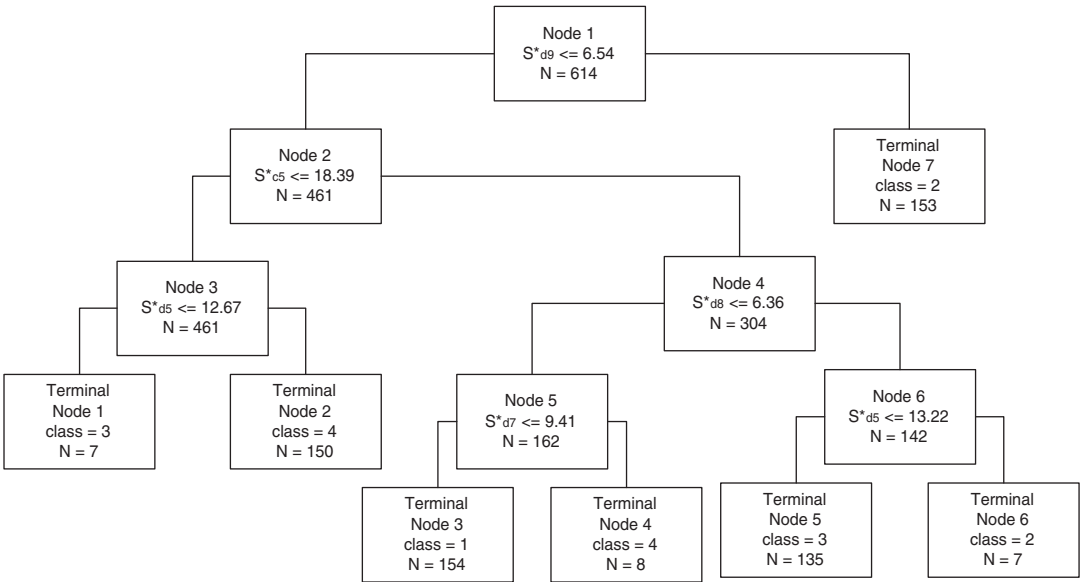

Figure 5. CART tree using thresholded scalograms.

1200 total replicated-curves (300 in each case). CART will then identify all these fault types based on the scalogram data.

Some of the curves from fault conditions are considerably more difficult for decision trees to correctly identify. For example, the only difference between class 1 and the original curve is a smaller amount of vertical drop in the first rectangle-shape dip located around 147 to 204 time units. In Figure 5, the notation S^*_{c5} represents the energy at the coarsest level and S^*_{dj} the energy at the finer resolution level j. The first split is $S^*_{d9} \leqslant 6.54$ where S^*_{d9} is the energy at the finest resolution level. If $S^*_{d9} > 6.54$, then the signal is classified into class 2; otherwise, go to node 2. Similar interpretations could be obtained for other nodes.

Table I shows the importance-rankings of variables selected from CART. The scores reflect the contribution each variable made in classifying or predicting the target variable; in each case the contribution stems from the role of each variable as both a primary splitter and as a surrogate to any of the primary splitters. The relative importance of input variables can be measured by

$$\hat{I}_j = \sum_{t=1}^{J-1} \hat{i}_t^2 I(v_t = j)$$

where the summation is over all the non-terminal nodes t of the J-terminal node tree, v_t is the splitting variable associated with node t, and $\hat{i}_t^2$ is the empirical improvement in misclassification error as a result of a split [15]. The influence from the most influential variable is arbitrarily assigned the value 100. In our example, S^*_{d5} is ranked most important. Note that the thresholded scalograms, S^*_{d5} and S^*_{d6}, at two finer resolution levels (but not at the coarsest level S^*_{c5}) are more important than the others in this example. The misclassification

Table I. Variable importance.

Variable	S^*_{d5}	S^*_{d6}	S^*_{d8}	S^*_{c5}	S^*_{d9}	S^*_{d7}
Score	100.00	91.32	70.42	69.57	65.68	55.19

Table II. Misclassification error (%).

Class	Training data	Testing data
Original	0.65	0.69
1	1.91	0.00
2	0.00	0.00
3	0.70	3.82

rates in the training and testing samples are shown in Table II. All classification errors are less than 4%, with many error-free cases. Our scalogram-based CART method performed well despite complicated testing curves, noises, and left-and-right shifting that could have made our fault classification problem difficult.

5. CONCLUSIONS

This article proposed the use of thresholded scalograms to detect and classify faulty processes. The properties of thresholded scalograms were explored via theoretical and empirical investigations. One real-life example and one simulated case study were presented to illustrate the potential of the proposed method. We believe that when data size becomes larger, our procedure will become more powerful and important.

ACKNOWLEDGEMENT

The authors sincerely acknowledge the contributions made by the reviewers and editors through their detailed reviews in improving the clarity and presentation of our paper.

REFERENCES

1. Jin J, Shi J. Feature-preserving data compression of stamping tonnage information using wavelets. *Technometrics* 1999; **41**(4):327–339.
2. Koh CKH, Shi J, Williams WJ, Ni J. Multiple fault detection and isolation using the Haar transform, Part 2: application to the stamping process. *Transactions of the ASME* 1992; 295–299.
3. Rioul O, Vetterli M. Wavelets and signal processing. *IEEE Signal Processing Magazine* October 1991; 14–38.
4. Scargle JD. Wavelet methods in astronomical time series analysis. In Rao TS, Priestly MB, Lessi O (eds). *Application of Time Series Analysis in Astronomy and Meteorology*. Chapman & Hall: New York, 1997; 226–248.
5. Donoho DL, Johnstone IM. Ideal spatial adaptation by wavelet shrinkage. *Biometrika* 1994; **81**(4):425–455.
6. Pittner S, Kamarthi V. Feature extraction from wavelet coefficients for pattern recognition tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 1999; **21**(1):83–88.
7. Morettin P. Wavelets in statistics. *Resenhas* 1997; **3**(2):211–272.
8. Vidakovic B. *Statistical Modeling by Wavelets*. Wiley: New York, 1999.

9. Donoho DL, Johnstone IM. Adapting to unknown smoothness via wavelet shrinkage. *Journal of the American Statistical Association* 1995; **90**(432):1200–1224.
10. Lada EK, Lu JC, Wilson JR. A wavelet based procedure for process fault detection. *IEEE Transactions on Semiconductor Manufacturing* 2002; **15**(1):79–90.
11. Jeong MK, Lu JC, Huo X, Vidakovic B, Chen D. Wavelet-based data reduction techniques for fault detection and classification. *Technical Report in the School of Industrial and Systems Engineering*, Georgia Institute of Technology, Atlanta, GA, 2002.
12. Mallat SG. *A Wavelet Tour of Signal Processing*. Academic Press: San Diego, 1998.
13. Serfling RJ. *Approximation Theorems of Mathematical Statistics*. Wiley: New York, 1980.
14. Rying EA, Gyurcsik RS, Lu JC, Bilbro G, Parsons G, Sorrell FY. Wavelet analysis of mass spectrometry signals for transient event detection and run-to-run process control. In Meyyappan M, Economou DJ, Bulter SW (eds), *Proceedings of the Second International Symposium on Process Control, Diagnostics, and Modeling in Semiconductor Manufacturing*. The Electrochemical Society; Pennington, New Jersey, 1997; 37–44.
15. Breiman L, Friedman JH, Olshen RA, Stone CJ. *Classification and Regression Trees*. Chapman & Hall: New York, 1984.

