

MULTISCALE ANALYSIS AND MODELING USING WAVELETS

BHAVIK R. BAKSHI*

Department of Chemical Engineering, The Ohio State University, Columbus, OH 43210, USA

SUMMARY

Measured data from most processes are inherently multiscale in nature owing to contributions from events occurring at different locations and with different localization in time and frequency. Consequently, data analysis and modeling methods that represent the measured variables at multiple scales are better suited for extracting information from measured data than methods that represent the variables at a single scale. This paper presents an overview of multiscale data analysis and empirical modeling methods based on wavelet analysis. These methods exploit the ability of wavelets to extract events at different scales, compress deterministic features in a small number of relatively large coefficients, and approximately decorrelate a variety of stochastic processes. Multiscale data analysis methods for off-line and on-line removal of Gaussian stationary noise eliminate coefficients smaller than a threshold. These methods are extended to removing non-Gaussian errors by combining them with multiscale median filtering. Multiscale empirical modeling methods simultaneously select the most relevant features while determining the model parameters, and may provide more accurate and physically interpretable models. Copyright © 1999 John Wiley & Sons, Ltd.

KEY WORDS: wavelets; noise removal; filtering; multiscale methods; regression

1. INTRODUCTION

All data analysis and empirical modeling methods represent the measured data for each variable as a weighted sum of a set of basis functions. Most existing methods represent the data in terms of basis functions at a fixed resolution or scale in time and frequency. This single-scale representation is best for representing data containing contributions with the same localization everywhere in the time–frequency domain. In practice, it is rare for measured data to contain contributions at a single scale, since the contributing deterministic events usually occur with different localization and at different locations in time and frequency, the stochastic events are often scale- and time-dependent, and the variables may be measured at different sampling rates or may contain segments of missing data. For example, a typical signal from a chemical process may contain contributions from a variety of sources, such as sensor noise, disturbances and faults, as shown in Figure 1(a).¹ These features occupy different regions of the time–frequency space as shown in Figure 1(b), since white noise is uniformly distributed in the entire time–frequency domain, whereas a sudden change is localized in time but occupies a wide range of frequencies. Thus representation of the measurements in terms of basis functions at a single scale will not permit efficient feature extraction or noise removal from a typical

* Correspondence to: B. R. Bakshi, Department of Chemical Engineering, The Ohio State University, Columbus, OH 43210, USA. E-mail: bakshi-2@osu.edu
Contract/grant sponsor: Technical Association of the Pulp and Paper Industry; Contract/grant number: PE-357-95.
Contract/grant sponsor: American Chemical Society—Petroleum Research Fund; Contract/grant number: 30523-G9.
Contract/grant sponsor: DuPont.

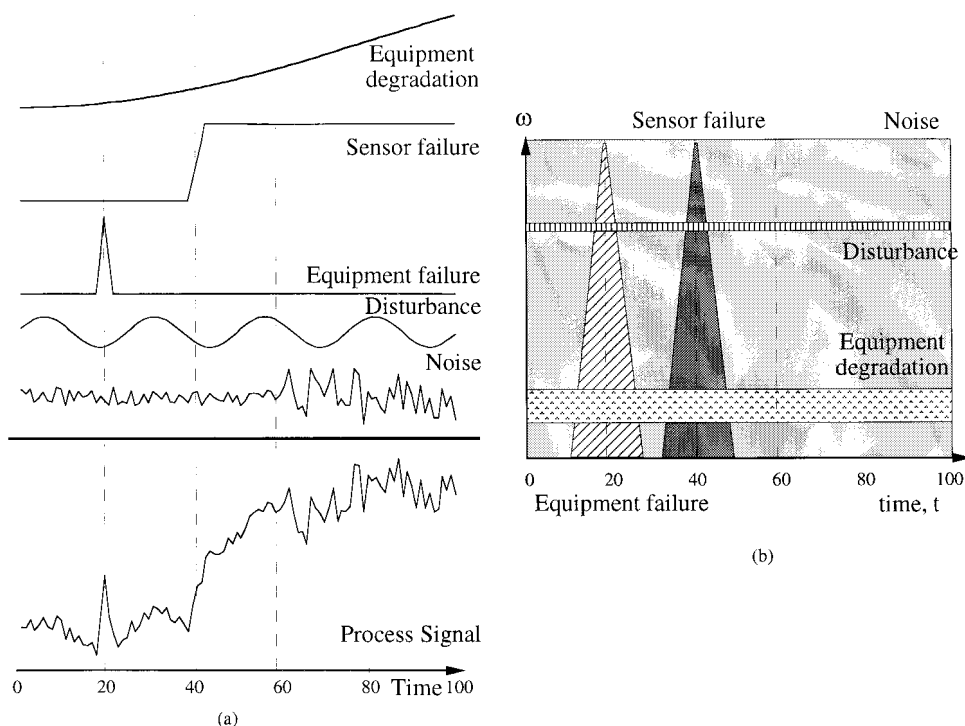


Figure 1. (a) Illustration of a typical process signal and (b) its time–frequency representation

process signal. Noise removal by eliminating the high-frequency contribution by Fourier transform will distort the localized features by excessive smoothing, since their high-frequency components will also be removed. Any single-scale method will be forced to trade off the extent of noise removal with the quality of the retained localized features. Owing to the multiscale nature of the contributing features, better extraction of various features in the measured signal and noise removal are possible by representing the signal on basis functions that are multiscale in nature, i.e. their resolution can vary in the time–frequency domain.

The ability to obtain better empirical models by selecting the most relevant variables is exploited by several methods, such as stepwise regression, inductive decision trees and multivariate adaptive regression splines. Similarly, if the measurements for each variable contain contributions that are not relevant to the predicted variable, it is also possible to improve the model quality by selecting the most relevant signal features in each variable. Such techniques for selecting the most relevant features from each variable have not received much attention, since most empirical modeling methods rely on a single-scale representation of the input and output variables, which is unable to extract features accurately from most practical signals. Most existing methods for feature extraction are also not integrated with the modeling, and rely on domain-specific knowledge, user expertise or trial and error to remove the less important signal features.

Since the development of wavelets, significant research effort has focused on developing multiscale methods for data analysis and modeling. Wavelets are a family of basis functions whose time–frequency localization or scale is not the same in the entire time–frequency domain. Thus wavelets possess multiscale character and are able to adjust their scale to the nature of the signal

features. Furthermore, wavelets can be orthonormal, and the computational complexity of the fast wavelet transform is $O(N)$ for a signal of length N , which is less than the complexity of the fast Fourier transform. Wavelets are able to capture deterministic features in a small number of relatively large coefficients, while stochastic processes contribute to all the coefficients and are approximately decorrelated. Extensions of wavelets to libraries of orthonormal basis functions that can be searched efficiently, called wavelet packets,² and time-varying wavelet packets³ have also been developed. These properties of wavelets have made them extremely popular and useful for data analysis and empirical modeling.

This paper provides an overview of wavelets and their applications to data analysis and empirical modeling, including some recent results and applications. A brief introduction to wavelets and some of their properties and extensions is presented in Section 2. Wavelet-based methods for multiscale data analysis are described in Section 3. This section focuses on techniques for removing various types of noise from a single variable, in an off-line and on-line manner. Extension of the on-line multiscale rectification approach to univariate statistical process control (SPC) is also described. Empirical modeling methods that aim to exploit the multiscale representation are surveyed in Section

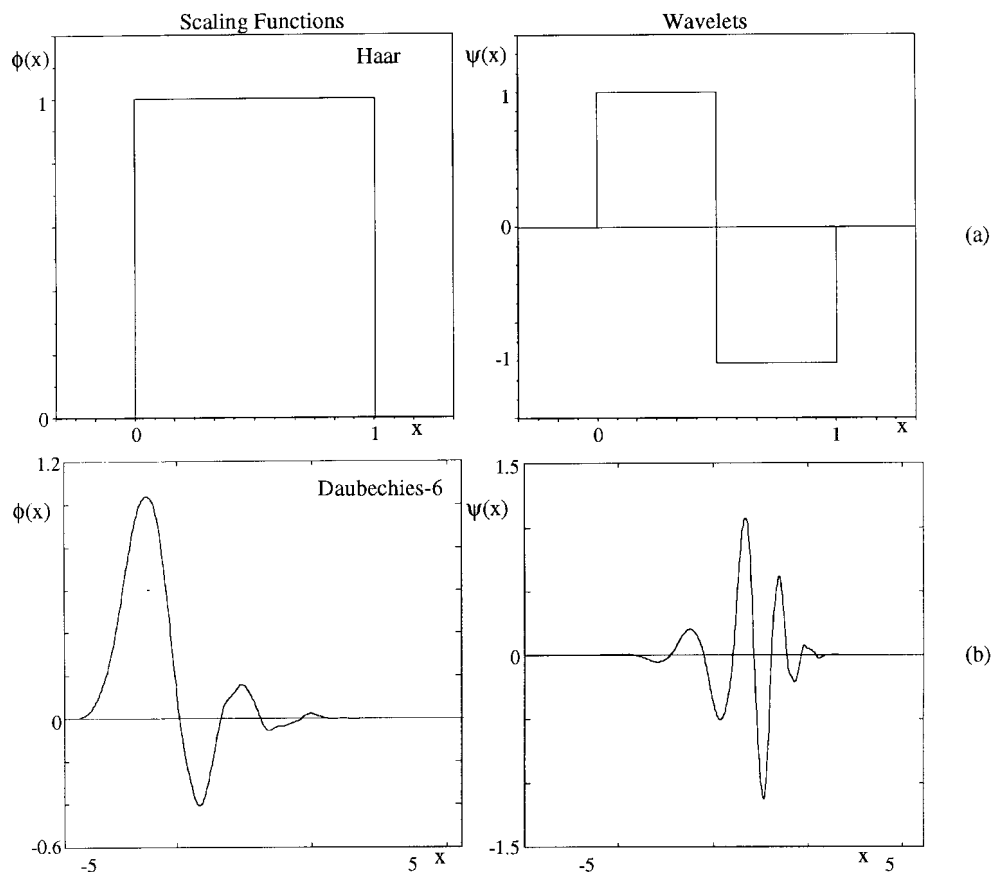


Figure 2. Typical orthonormal scaling functions and wavelets: (a) Haar; (b) Daubechies-6

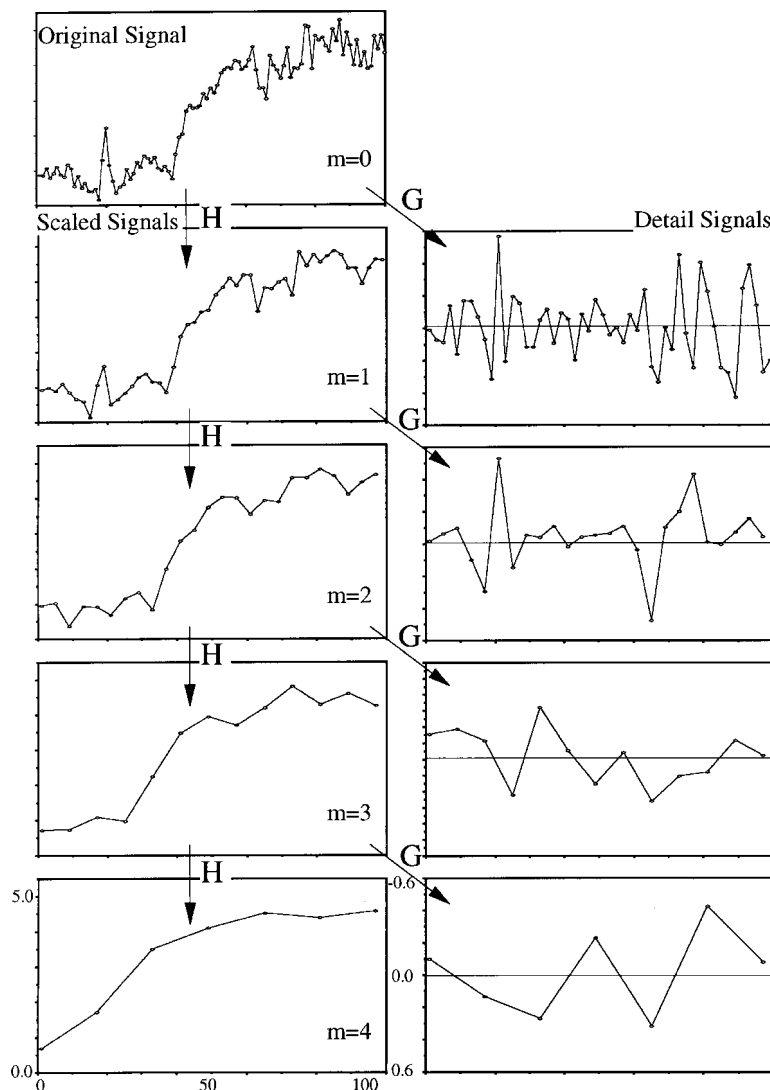


Figure 3. Wavelet decomposition using Daubechies-6 wavelet

4. This includes a multivariate method for noise removal, various approaches for selecting the signal features most relevant to the modeling, and a multiscale approach for multivariate SPC. Finally, some concluding remarks and areas of future work are discussed in Section 5.

2. WAVELETS

Wavelets are a family of basis functions that are localized in both time and frequency, and may be

represented as

$$\psi_{su}(t) = \frac{1}{\sqrt{s}} \psi\left(\frac{t-u}{s}\right) \quad (1)$$

where s and u represent the dilation and translation parameters respectively. The mother wavelet $\psi(t)$ is localized in time and frequency with an equal area above and below the t -axis, as shown in Figure 2. For most practical applications to measured data, the wavelet dilation and translation parameters are discretized dyadically as

$$s = 2^m, \quad u = 2^m k \quad (2)$$

where m and k are integers and represent the dilation and translation parameters respectively. This resulting family of wavelets is represented as

$$\psi_{mk}(t) = 2^{-m/2} \psi(2^{-m}t - k) \quad (3)$$

The translation parameter determines the location of the wavelet in the time domain, while the dilation parameter determines the location in the frequency domain as well as the scale or extent of the time–frequency localization. The wavelets represented by equation (3) may be designed to be orthonormal to each other and may have different degrees of smoothness.⁴

Any signal may be decomposed into its contributions in different regions of the time–frequency space by projecting it on the corresponding wavelet basis function, as depicted in Figure 3. The lowest-frequency content of the signal shown in the scaled signal at $m = 4$ is obtained by representing the original signal on a set of scaling functions such as those shown in Figure 2. The number of wavelet and scaling function coefficients decreases dyadically at coarser scales owing to the dyadic discretization of the dilation and translation parameters. The original signal is considered to be at $m = 0$ and the coarsest scale is at $m = L$. Fast algorithms for computing the wavelet decomposition are based on representing the projection of the signal on the corresponding basis function as a filtering operation.⁵ Convolution with a filter \mathbf{H} represents projection on the scaling function, and convolution with a filter \mathbf{G} represents projection on a wavelet. Thus the coefficients at different scales may be obtained as

$$\mathbf{a}_m = \mathbf{H}\mathbf{a}_{m-1}, \quad \mathbf{d}_m = \mathbf{G}\mathbf{a}_{m-1} \quad (4)$$

where \mathbf{d}_m is the vector of wavelet coefficients at scale m , and \mathbf{a}_m is the vector of scaling function coefficients. The original data are considered to be the scaling function coefficients at the finest scale, i.e. $\mathbf{x} = \mathbf{a}_0$. Equation (4) may also be represented in terms of the original measured data vector \mathbf{x} as

$$\mathbf{a}_m = \mathbf{H}_m \mathbf{x}, \quad \mathbf{d}_m = \mathbf{G}_m \mathbf{x} \quad (5)$$

where \mathbf{H}_m is obtained by applying the \mathbf{H} filter m times, and \mathbf{G}_m is obtained by applying the \mathbf{H} filter $m - 1$ times and the \mathbf{G} filter once. Thus the wavelet decomposition of a vector \mathbf{x} may be obtained as

Wx , where W is an orthonormal matrix given as

$$\mathbf{W} = \begin{bmatrix} h_{L,1} & h_{L,2} & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & h_{L,N} \\ g_{L,1} & g_{L,2} & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & g_{L,N} \\ g_{L-1,1} & \cdot & \cdot & \cdot & g_{L-1,N/2} & \mathbf{0} & \cdot & \cdot & \cdot & \mathbf{0} \\ \mathbf{0} & \cdot & \cdot & \cdot & \mathbf{0} & g_{L-1,N/2+1} & \cdot & \cdot & \cdot & g_{L-1,N} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ g_{1,1} & g_{1,2} & \mathbf{0} & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \mathbf{0} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \mathbf{0} & \mathbf{0} & \cdot & \cdot & \cdot & \cdot & \cdot & \mathbf{0} & g_{1,N-1} & g_{1,N} \end{bmatrix} = \begin{bmatrix} \mathbf{H}_L \\ \mathbf{G}_L \\ \mathbf{G}_{L-1} \\ \cdot \\ \cdot \\ \mathbf{G}_m \\ \cdot \\ \cdot \\ \mathbf{G}_I \end{bmatrix} \quad (6)$$

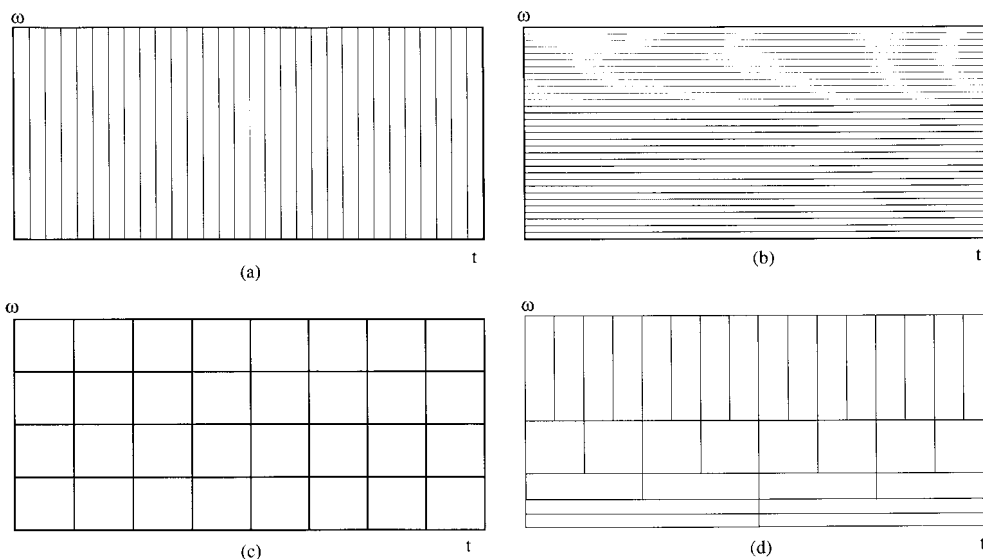


Figure 4. Tiling of time–frequency space by (a) delta functions, (b) Fourier transform, (c) linear filters and (d) wavelet transform

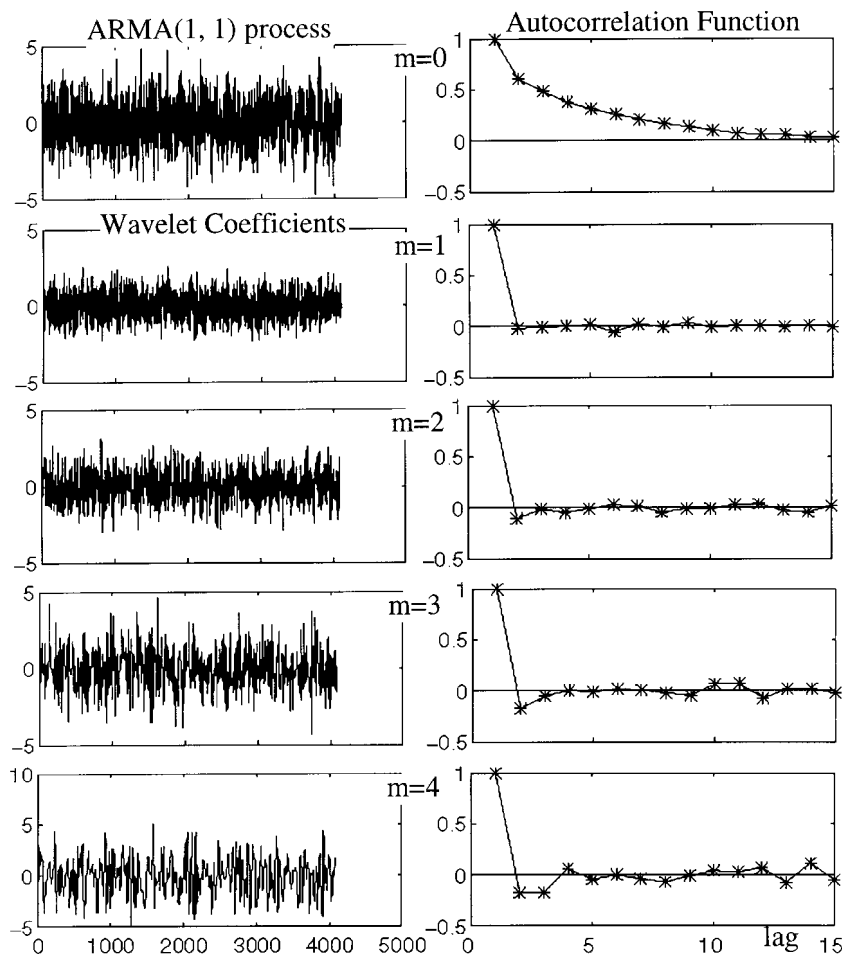


Figure 5. Approximate decorrelation of wavelet coefficients for ARMA(1,1) process, $x(t) = 0.8x(t-1) + e(t) - 0.3e(t-1)$

Here $h_{L,k}$ is the filter coefficient in the filter \mathbf{H}_L to obtain the scaled signal at the coarsest scale, while $g_{m,k}$ is the filter coefficient in the filter \mathbf{G}_m to obtain the wavelet coefficient at scale m . The original data may be reconstructed exactly from their wavelet coefficients at all scales, d_m for $m = 1, 2, \dots, L$, and scaling function coefficients at the coarsest scale, a_L .

The advantages of representing measured data on a set of wavelets may be illustrated by considering how various data analysis methods tile the time–frequency space. A set of measured data in the time domain is equivalent to representing it on a set of Dirac delta functions, resulting in a time–frequency representation as shown in Figure 4(a). All the delta functions have the same resolution and are completely localized in time, but are global in frequency. Representing the signal on a set of trigonometric basis functions by Fourier analysis results in decomposition of the time–frequency space as shown in Figure 4(b). This representation is also single-scale in nature, and the basis functions are completely localized in frequency and global in time. Linear filtering methods such as mean filtering and windowed Fourier analysis tile the time–frequency space as shown in Figure 4(c). The resulting basis functions are localized in both time and frequency but are single-scale

in nature, since the extent of time–frequency localization is constant in the entire domain. Tiling of the time–frequency space by dyadically discretized wavelet basis functions is as shown in Figure 4(d). The wavelet basis functions are localized in both time and frequency and are multiscale in nature, since their resolution changes at different locations in the time–frequency domain.

In addition to representing a signal at multiple scales, wavelets possess several other properties that make them attractive for data analysis and modeling. Wavelets are able to separate deterministic and stochastic components of a signal by capturing deterministic changes in a relatively small number of large coefficients, while stochastic processes are distributed among all the coefficients. As illustrated in Figure 3, the deterministic peak at 20 is captured by the large wavelet coefficients at the two finest scales, while the white noise is present at all scales. The change in the noise variance is also captured clearly at the two finest scales. Furthermore, wavelets are the approximate eigenfunctions of a variety of mathematical operators, allowing them to approximately decorrelate most stochastic processes. Thus the wavelet decomposition of an autocorrelated stochastic process results in almost uncorrelated coefficients whose energy at different scales corresponds to the power spectrum. This property is illustrated in Figure 5 for an ARMA(1,1) stochastic process. These properties of wavelets have been exploited widely in applications such as filtering,⁶ feature extraction⁷ and statistical process control.⁸

Application of wavelets requires the user to address several practical issues. Application of wavelet basis functions smoother than the piecewise constant Haar wavelet requires information about the future to calculate the current coefficient. This non-causal nature can lead to errors at the signal ends if special boundary-corrected wavelets,⁹ which are causal while being orthonormal to all the other wavelets, are not used. The dyadic downsampling in the orthonormal wavelet decomposition does not permit on-line application of the wavelet decomposition, since signals with an odd number of samples cannot be decomposed. Since many applications require on-line data analysis and modeling, an approach for on-line wavelet decomposition is discussed in Section 3. A practical issue in using wavelets for pattern recognition is that the wavelet decomposition lacks translational invariance. Consequently, the wavelet coefficients of a signal and its non-dyadic translation are not translations of each other, and may be very different. Several approaches have been suggested to make the wavelet decomposition translationally invariant by sacrificing the orthonormality of the basis functions, as discussed in more detail in Section 3. Orthonormal wavelet basis functions are of a fixed shape and tile the time–frequency space in a predetermined and rigid manner, making them inefficient for representing several types of features in a measured signal. For example, a high-frequency oscillation would require several wavelet basis functions at a fine scale owing to their narrow localization in time, while a single trigonometric basis function would be adequate for representing this signal accurately. This rigidity of wavelets may be overcome by developing a library of wavelet packet basis functions by decomposing each detail and scaled signal from the wavelet decomposition by repeated application of the corresponding filters until further decomposition is impossible. This decomposition results in a complete binary tree of basis functions that cover a wide variety of time–frequency localizations and shapes. Any signal may be represented on a complete subset of these basis functions. Efficient algorithms for searching this wavelet packet library by branch and bound in $O(M\log N)$ time to select the best orthonormal bases for representing a given signal have been developed.² This permits efficient selection of the orthonormal basis functions that represent the signal in the most efficient manner. Additional details about wavelets and their extensions may also be obtained from Walczak,¹⁰ a tutorial paper by Alsberg *et al.*¹¹ and many books.¹²

3. MULTISCALE DATA ANALYSIS

Data analysis methods aim to reduce the contribution of errors from measured data and are essential

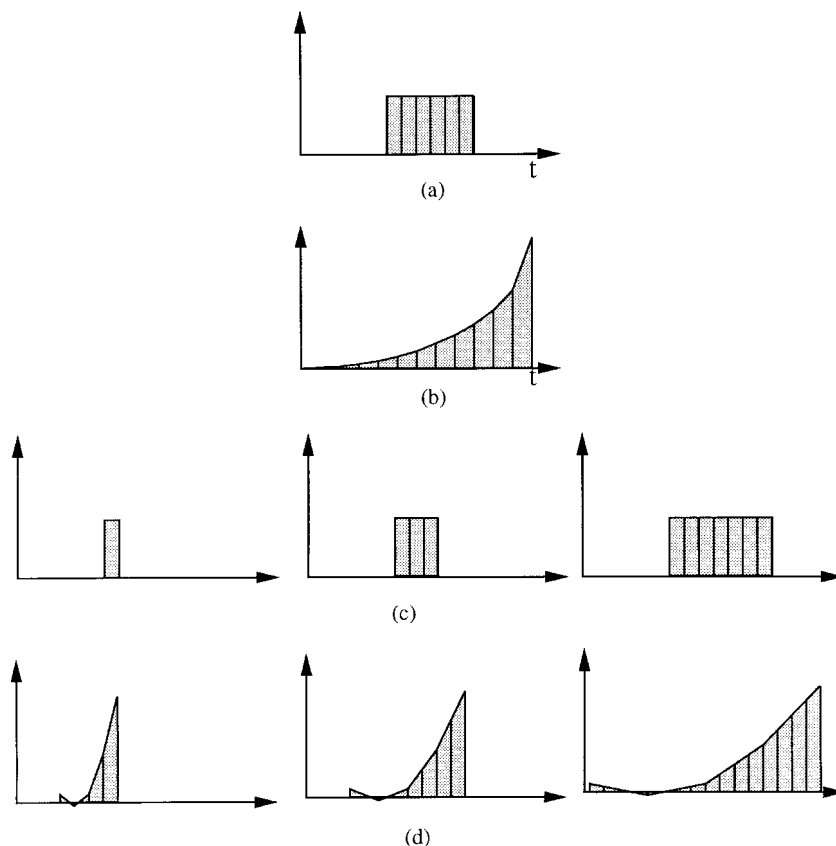


Figure 6. Nature of linear and multiscale filters: (a) mean filter; (b) exponentially weighted moving average filter; (c) Haar filter at different scales; (d) boundary-corrected Daubechies-4 filter at multiple scales

for a variety of engineering and chemometric tasks. These methods are also referred to by the terms filtering and rectification. The wide variety of existing filtering methods may be broadly classified into linear and non-linear methods. The properties of these methods and their relationship with wavelet-based multiscale filtering methods are discussed in this section. An approach for on-line multiscale decomposition is developed and used for on-line filtering and multiscale statistical process control.

3.1. Linear filtering methods

Linear filtering methods reduce errors by taking a linear combination of the measured data as

$$\hat{x}_i = \sum_{i=1}^N w_i x_i \quad (7)$$

where \hat{x}_i are the filtered data, x_i are the measured data and w_i are the weights. If the length of the window, N , is finite, these filters are called finite impulse response (FIR) filters. Filters that combine all the measurements are called infinite impulse response (IIR) filters.

Mean filter

Among the most popular FIR filtering methods is mean filtering, which attempts to eliminate random errors by computing the average of the measured samples in a moving window. The rectified signal \hat{x}_i is computed as

$$\hat{x}_i = \frac{1}{N}(x_{i-N+1} + x_{i-N+2} + \dots + x_i) \quad (8)$$

The nature of the mean filter depicted in Figure 6(a) indicates that mean filtering gives equal importance to past data within its moving window. As more samples are collected, the mean filtered value is computed in overlapping moving windows. Rectification by mean filtering works best for a constant underlying signal contaminated by uncorrelated errors. Otherwise, the rectified signal tends to track the errors or distorts the retained features. Even for uncorrelated errors, mean filtering may retain trends in the rectified data.

Exponentially weighted moving average filter

A popular IIR filter is the exponentially weighted moving average (EWMA) filter, also known as exponential smoothing, geometric moving average or a first-order pole filter. This is a recursive lowpass filtering technique with the rectified value for the i th sample, \hat{x}_i , calculated as

$$\hat{x}_i = \alpha x_i + (1 - \alpha)\hat{x}_{i-1} \quad (9)$$

where x_i is the measured value at the i th sample and α is the filter parameter, with $0 < \alpha \leq 1$. Equation (9) may also be represented as a weighted moving average of past and current observations as

$$\hat{x}_i = (1 - \alpha)^i \hat{x}_0 + \alpha(1 - \alpha)^{i-1} x_1 + \alpha(1 - \alpha)^{i-2} x_2 + \dots + \alpha x_i \quad (10)$$

where \hat{x}_0 is the starting value. The nature of the exponential filter is shown in Figure 6(b), which indicates that, unlike mean filtering, older samples are not completely eliminated from the rectification, but their contribution decreases exponentially as more samples are obtained.

Characteristics of linear filtering methods

Linear filtering methods are extremely popular owing to their ease of use and computational efficiency.¹³ They have also found wide use in statistical process control, since after proper tuning, they permit easier identification of a small shift in the mean. All linear filters represent the data at a single scale, since their filter parameters do not adapt to the nature of the signal. Thus linear filters tile the time–frequency space in a manner analogous to window Fourier transforms, as depicted in Figure 4(c), making them best for analyzing data containing features at a single scale. As discussed in Section 1, if the signal contains features at multiple scales, the linear filter will be forced to trade off the extent of error removal with the quality of the retained local features. Thus, as more errors are removed, the retained localized features in the filtered signal become smoother and more distorted. If the random errors are autocorrelated, their low-frequency contributions cause the rectified signal to track the correlation in the errors. The disadvantages of single-scale rectification methods are even more significant if the signal errors are non-stationary or time-varying. Linear filters are also not robust to the presence of non-Gaussian errors such as outliers in the measured data.

3.2. Non-linear filtering methods

Many of the disadvantages of linear filters may be overcome by transforming the measurements in a non-linear manner. These non-linear filters are multiscale in nature, since they permit the scale of the basis functions to adapt to the nature of the measurements. Some of the popular non-linear filters include median filtering, FIR median hybrid (FMH) filters and wavelet thresholding.

Median and FIR median hybrid filters

Median filtering takes the median of the measurements in a moving window of odd length as the central filtered value. This approach is robust to outliers and is better at retaining sudden changes in the measured signal while decreasing the contribution of random errors. The ability of median filtering to retain changes in the signal may be improved by combining it with FIR filters.¹⁴ The resulting FMH filter computes the filtered value as

$$\hat{x}_i = \text{median}(\text{FIR}(x_{i-1}, x_{i-2}, \dots, x_{i-k}), x_i, \text{FIR}(x_{i+1}, x_{i+2}, \dots, x_{i+k})) \quad (11)$$

The filter length for the median and FMH filter needs to be chosen based on the maximum length of the outliers. If the filter length is shorter than the duration of the longest patch of outliers, it will not be removed. In contrast, if the filter length is too long, it may distort some of the important features. Repeated application of these filters to the filtered signal results in a 'root' signal that passes through the filter unchanged. Both median and FMH filtering are non-causal in nature, since they require measurements from the future for computing the filtered value for the current measurement. This requirement restricts them to off-line application to a batch of data. Furthermore, these filters are best for filtering random errors from a piecewise constant underlying signal.¹⁴

Wavelet thresholding

A deterministic signal contaminated by stationary stochastic errors may be filtered by a wavelet-thresholding approach consisting of the following three steps.⁶

- *Decompose* measured data on a selected family of basis functions.
- *Eliminate* coefficients smaller than a selected threshold.
- *Reconstruct* the rectified signal.

This approach exploits the ability of wavelets to capture deterministic features in a relatively small number of large wavelet coefficients, while distributing the stochastic component among all the coefficients according to its power spectrum. The threshold for data rectification may be determined from knowledge about the variance and stochastic nature of the errors, and the known properties of stochastic processes in the wavelet domain. The variance of the errors at each scale represents the energy of the stochastic process in the corresponding frequency band and will change according to the power spectrum of the errors. Thus, if the errors are known to be white or uncorrelated, then the constant power spectrum at all frequencies indicates that the threshold should be identical at each scale and proportional to the standard deviation of the errors in the original signal. For autocorrelated errors the property of wavelets to approximately decorrelate them permits wavelet thresholding to filter data contaminated by stationary autocorrelated processes. For autocorrelated errors the threshold at each scale changes according to the variation of the power spectrum of the errors in the corresponding frequency band.¹⁵

In practice, the variance or the stochastic nature of the errors may not be known. Several methods have been suggested for estimating the threshold for rectification from the measured data.^{6,16} The *VisuShrink* method⁶ determines the threshold at each scale as

$$t_m = \sigma_m \sqrt{2 \log N} \quad (12)$$

where σ_m is the standard deviation of the errors at scale m , and N is the length of the signal. The statistical properties of this wavelet-thresholding approach have been studied by Donoho *et al.*⁶ The factor $\sqrt{2 \log N}$ is included in equation (12) to improve the visual appearance of the rectified signal, since for uncorrelated Gaussian errors n_i , the probability of not eliminating a coefficient representing the error decreases as the number of samples increases:

$$\text{pr}\left\{\max(|n_i| > \sqrt{\log N})\right\} \rightarrow 0, \quad N \rightarrow \infty \quad (13)$$

The error between the actual error-free signal and the rectified signal is guaranteed to be within $O(\log N)$ of the error between the error-free signal and the signal rectified with *a priori* knowledge about the smoothness of the underlying signal. Thus, the rectified signal by wavelet thresholding is nearly optimal for a wide variety of theoretical objectives, such as various error norms and smoothness, which is much broader than the range of objectives satisfied by existing methods.

The standard deviation of the errors at each scale, σ_m , may be determined from knowledge of the nature of the errors, if available. In most practical situations, information about the stochastic character and variance of the errors is not available and σ_m has to be estimated from the measured data. Since most of the coefficients in the multiscale decomposition of a measured signal correspond to the errors, and those corresponding to the underlying signal are larger in magnitude, the variance of the errors at each scale may be estimated by the robust median absolute deviation as

$$\sigma_m = \frac{1}{0.6745} \text{median}(|d_m|) \quad (14)$$

where d_m denotes the wavelet coefficients at the selected scale. The accuracy of equation (14) for determining the variance of the noise in a noisy signal decreases at coarser scales, since the number of available data points decreases, and the contribution from the underlying signal may increase.

Another practical approach for estimating the threshold for minimizing a selected error norm is based on cross-validation.¹⁶ This approach divides the measured data into sets of training and testing data and selects the threshold that minimizes the error for testing data. Since the fast wavelet decomposition algorithm requires a signal of dyadic length, twofold cross-validation is performed by dividing the signal into odd-numbered points x_o and even-numbered points x_e . First the odd-numbered points are used as training data and are rectified by using different values of the threshold. The values of the even-numbered testing data are estimated as the mean of the adjacent odd-numbered rectified points:

$$\hat{x}_{e,i} = 0.5(\hat{x}_{o,i} + \hat{x}_{o,i+1}) \quad (15)$$

The mean square error of approximation for the testing data is then computed as

$$e_e = \sum_{i=1}^{N/2} (x_{e,i} - \hat{x}_{e,i})^2 \quad (16)$$

The same process is repeated with the even-numbered points as the training data. Finally, the threshold with the minimum prediction error is selected as the optimum. This cross-validation method is claimed to give better rectification but is computationally more expensive, since different values of

the threshold need to be evaluated for selecting the best value. A golden-section search routine is suggested for this search.¹⁶

The number of scales to which the signal is decomposed also determines the quality of the rectification. Usually, the relative contribution of the signal to the errors increases at lower frequencies or coarser scales. If the measured signal is decomposed to too fine a scale (not deep enough), then the errors may still be present in the last scaled signal, which will contaminate the rectified signal. In contrast, if the signal is decomposed to too coarse a scale (too deep), then the detail signal at the coarsest scales may not contain significant contribution from the errors, and estimation of the threshold by equation (12) may result in elimination of coefficients that represent the underlying signal, resulting in excessive smoothing of the rectified signal. The optimum scale for rectification of a given signal may be determined by observation or as the scale that minimizes the error on testing data by cross-validation.

Rectification by wavelet thresholding has a tendency to create spurious features near large changes in the measured signal. These spurious features are due to a localized Gibbs phenomenon caused by a mismatch between the location of the change in the signal and that in the basis function. Thus, for Haar wavelets, if a step change is at a dyadic location, it will be represented exactly by the multiscale rectification, since the location of the step will match the step change in the basis function. In contrast, if the step is at a non-dyadic location, the rectified signal will exhibit significant spurious features, since several Haar wavelets will be needed to capture it exactly. Consequently, translation of the measured signal affects the location and size of the spurious features.

The variation in the wavelet decomposition with signal translation is widely recognized⁵ and several approaches have been suggested for making the wavelet representation translationally invariant. The approach of Mallat and Zhong¹⁷ decomposes the signal on uniformly discretized wavelets and represents the multiscale decomposition in a translationally invariant manner in terms of its wavelet transform extrema. This approach sacrifices the orthonormality of the basis functions, making it inconvenient for the characterization and removal of stochastic errors. The filtered signal can be reconstructed from the wavelet extrema by a computationally intensive iterative method. The methods developed by Liang and Parks¹⁸ and Coifman and Donoho¹⁹ retain the orthonormality of the basis functions in each translation by decomposing all translations of the measured signal while considering the signal to be a cyclical list. The approach of Liang and Parks selects the signal translation that minimizes an additive cost criterion such as entropy. Since a signal may contain multiple features at arbitrary locations, it is usually impossible to find a single translation that captures all the features accurately without any spurious features. Consequently, Coifman and Donoho do not attempt to find the best signal translation, but generate the rectified signal as the average of the rectification at all translations, since the spurious features at different translations tend to approximately cancel each other. Averaging the rectification at all translations also improves the smoothness of the rectified signal, such that rectification with even the piecewise constant Haar wavelets results in an infinitely smooth rectified signal. Despite the benefits of the translationally invariant multiscale filtering, it suffers from the following disadvantages.

- The method requires a batch of data of dyadic length and cannot be used for on-line rectification.
- If the values of the samples at the beginning and end are very different, an artificial discontinuity is created, resulting in errors near the signal ends.

These disadvantages may be overcome by the approach described in the next subsection.

The wavelet-thresholding approach works best for the reduction of stationary Gaussian errors, but may be extended for filtering of non-Gaussian errors by combining it with multiscale median filtering.²⁰ The resulting robust wavelet-thresholding approach applies a median filter of a selected size, such as three, to the original data. This filtered signal does not contain outliers of length one and

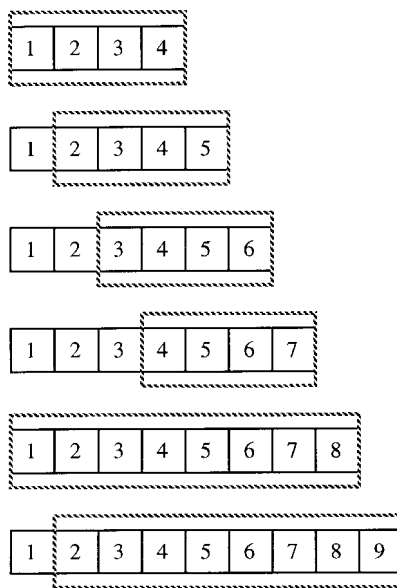


Figure 7. Moving window for on-line multiscale filtering

is decomposed by the wavelet **H** and **G** filters to get the scaled and detail signal at a coarser scale. This scaled signal is again filtered by the median filter to eliminate outliers of longer duration. This process is continued until the scale at which the longest outlier is expected to be present. Gaussian errors are reduced by thresholding the detail signals as described in previous paragraphs. Some outliers may 'leak' in this approach, since the wavelet coefficients will be contaminated by outliers of longer duration than the scale of the wavelet coefficients. This leakage may be reduced by filtering the filtered signal again. Other extensions of the wavelet-thresholding approach have also been developed for removing other types of errors.²¹ Most of the applications of wavelets in chemometrics are based on the off-line multiscale filtering approach described above.^{10,22,23}

On-line multiscale filtering

The multiscale filtering approach described above cannot be used on-line, since the dyadic discretization of orthonormal wavelets results in a complete multiscale decomposition only when the number of samples in the measured signal is dyadic. Thus signals of non-dyadic length cannot be decomposed until a dyadic number of samples are obtained. This time delay increases as the number of scales increases.

Owing to this off-line nature of existing multiscale filtering methods, they cannot be used for process operation tasks that require on-line filtering, such as feedback control, statistical process control and continuous process monitoring, instead of the current single-scale filtering methods such as exponential smoothing and mean filtering. These disadvantages of multiscale filtering are overcome in this subsection by extending it to on-line multiscale filtering. The resulting method can be applied to any problem where single-scale filtering methods are used currently, and performs better than existing methods for rectification without process models.²⁴

The methodology for on-line multiscale rectification decomposes the measured data on the selected family of wavelets or wavelet packets in the largest possible window of dyadic length. As new

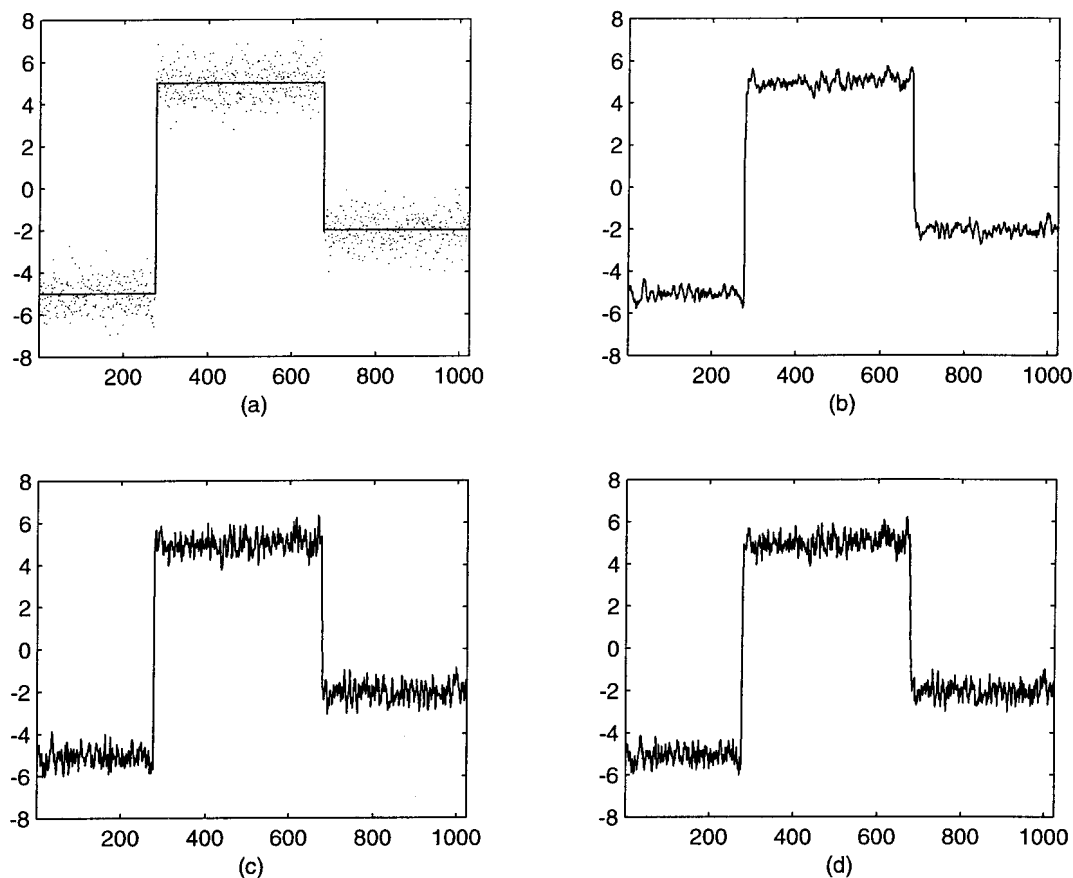


Figure 8. Performance of on-line multiscale filter and linear filters: (a) unfiltered noisy data (points) and noise-free data (full line), $MSE = 0.5$; (b) on-line multiscale filtering with Haar wavelets and depth of 4, $MSE = 0.1216$; (c) mean filtering with window of length 3, $MSE = 0.2635$; (d) exponentially weighted moving average with $\alpha = 0.42$, $MSE = 0.2194$

samples are collected, the window is translated so that the most recent sample is at a dyadic location for at least one translated window. As more samples are collected, the window size is increased to the largest possible dyadic length, as depicted in Figure 7, and reflecting the assumption that the nature of the random errors does not change over time. This translationally invariant multiscale decomposition method requires only $O(M \log N)$ computations, since only a few new coefficients need to be computed for every new sample. The threshold for eliminating coefficients corresponding to errors may be computed by the methods for wavelet thresholding, such as the median absolute deviation. If it is not necessary to obtain on-line rectification, the results of rectification in each window may be averaged to yield a signal similar to that obtained by translationally invariant denoising, but without the end effects, since the on-line approach does not assume the signal to be a cyclical list.

On-line multiscale filtering using Haar wavelets is similar to mean filtering, but with an adaptive window of dyadic length that is selected according to the nature of the features in the measured signal, as shown in Figure 6(c). For wavelets other than Haar the nature of the boundary-corrected filter for the scaling function at the right boundary is similar to an exponential filter and will approximately subsume exponential smoothers of dyadic lengths, as depicted in Figure 6(d). The performance of on-

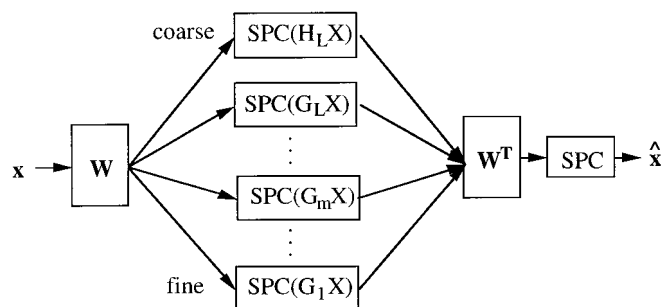


Figure 9. Methodology for multiscale statistical process control

line multiscale filtering is compared with that of exponential smoothing and mean filtering in Figure 8. The mean square error (MSE) measures the error between the noise-free signal and the filtered signal. The smaller error of on-line multiscale filtering as compared with other on-line filters has also been verified for other types of signals by Monte Carlo simulation.²⁴

The properties of wavelets and on-line multiscale filtering may also be exploited for improved univariate statistical process control. The performance of an SPC chart depends on its ability to detect any change from the normal state of operation. Existing methods for univariate SPC plot the measured data at the finest scale as a Shewhart chart, or represent the data at coarser scales by linear filtering by a mean filter, exponential smoother, or by taking the cumulative sum. These control charts are best for detecting certain types of deterministic changes in the data. For example, Shewhart charts are best for detecting large mean shifts, while EWMA and CUSUM charts are best for detecting small mean shifts. Since existing control charts differ only in the scale at which they represent the measurements, they may be unified by the on-line multiscale filtering approach described above. This results in a method that can automatically specialize to a control chart at any dyadic scale that is best for detecting the abnormal feature.⁸

The methodology for multiscale SPC consists of decomposing the measurements on the selected wavelet and applying separate detection limits to the coefficients at each scale, as depicted in Figure 9. The detection limits are determined from the wavelet decomposition of data representing normal operation. For uncorrelated measurements the detection limits will be almost equal at each scale, whereas for autocorrelated measurements the limits change with scale. If the current wavelet coefficient at any scale violates the limit, it may indicate abnormal operation in the form of a deterministic change. Whether a deterministic change is really present in the data is verified by reconstructing the signal based on coefficients that violate the detection limit. The detection limit for the reconstructed signal is computed from the limits of the scales at which abnormal operation was detected. Thus the scale of the reconstructed signal and its detection limits adapt to the nature of the measured signal and may change for each measurement. Unlike existing control charts, this multiscale approach is able to specialize to a Shewhart chart by selecting all the scales, or to a mean filter of dyadic length for Haar wavelets, or to an EWMA chart for smoother wavelets. The statistical properties of univariate MSSPC are discussed by Top and Bakshi.⁸

4. MULTISCALE EMPIRICAL MODELING

The benefits of the multiscale representation may also be exploited for extracting empirical models from measured data that are multiscale in nature. It is well known that the quality of empirical models

depends on the quality of the measured data used for the modeling. Preprocessing the variables by filtering or compression is a popular approach for improving the quality of the measured data by reducing the contribution of less relevant signal features. Since most data contain contributions at multiple scales, multiscale methods are expected to result in better empirical models than those obtained by existing methods.

The simplest approach for exploiting the benefits of the multiscale representation is to preprocess each variable independently by wavelet thresholding, followed by determining the empirical model from the filtered signal. Unfortunately, this approach does not integrate the filtering and modeling task and may require some heuristic knowledge or trial and error to extract the features most relevant to the modeling. Furthermore, since the univariate denoising approach described in Section 3 does not exploit the multivariate nature of the empirical modeling problem, it may be possible to improve the quality of the noise removal by an approach that is both multiscale and multivariate in nature.

Latent variable regression methods have an inherent ability to reduce errors by retaining only those latent variables that capture the relationship between the variables. Unfortunately, this approach can never completely eliminate the errors, since the scores and loadings retained will still be contaminated by the errors. Theoretical analysis indicates that while PCA does decrease the contribution of errors in the data matrix, it is impossible to eliminate the entire error while capturing the underlying relationship between the variables.²⁵ The imbedded error or the error retained after eliminating small principal components is related to the real error and the number of components retained. This indicates that all the error can be removed by PCA only if none of the principal components is retained. Unfortunately, not retaining any components will not capture the underlying relationship between the variables at all, defeating the very purpose of a latent variable regression method.

The noise removal ability and quality of a latent variable regression model may be enhanced by combining it with wavelet thresholding. Noise removal by eliminating the less important latent variables and wavelet thresholding of the latent variables complement each other, since the former removes noise based on the relationship between the variables, whereas the latter approach removes noise based on extracting features from the measurements. Existing multiscale empirical modeling methods^{26–29} exploit this synergy between latent variable models and wavelet thresholding as described in the rest of this section. Different approaches may be developed depending on whether both the input and output variables are represented in the wavelet domain, and on the method used for selecting the most relevant coefficients. The multiscale modeling approach is a type of errors-in-variables modeling, since both methods attempt to simultaneously remove errors and determine the model parameters.

Multiscale PCA^{26,29} combines the properties of wavelet analysis and PCA by decomposing each variable on a selected wavelet, followed by computing the PCA of the matrix of coefficients. This methodology is analogous to that for MSSPC shown in Figure 9, with the univariate Shewhart chart at each scale and for the reconstructed signal replaced by PCA. The relationship between the variables is decorrelated by PCA, while the relationship between the stochastic measurements is approximately decorrelated by the wavelet decomposition. If the noise is non-white or scale-dependent, developing a separate PCA model for the coefficients at each scale will permit better removal of the noise by using scale-dependent threshold values. A subset of principal components and wavelet coefficients is selected at each scale, by using thresholding criteria depending on the nature of the application. For example, if the objective of MSPCA is multivariate noise removal, the threshold for the latent variables at each scale may be determined by any univariate thresholding method discussed in Section 3. MSPCA may also be used for multivariate SPC,²⁶ in which case the threshold is provided by the SPC detection limit applied to the appropriate control chart. The number of principal components to be selected depends on the nature of the correlation between the variables. Since the wavelet

decomposition does not affect the relationship between the variables, the same number of components is selected at each scale. If an independent PCA model is developed at each scale, the final step in MSPCA is to combine the results at each scale. This is accomplished by reconstructing the signal based on the selected coefficients at each scale, computing the PCA for the reconstructed signal and eliminating the less relevant components. This step further improves the quality of noise removal and maintains orthonormality of the scores and loadings for the reconstructed signal. This last step also improves the ability of multivariate SPC by MSPCA to detect the return of a process to normal operation.

The approach for MSPCA may also be extended to other empirical modeling methods by decomposing all the variables and replacing the PCA at each scale in Figure 9 by the selected input–output modeling method. Alternatively, only the input variables may be decomposed as suggested by Alsberg *et al.*²⁷ The most relevant wavelet coefficients may be selected by thresholding the input and output latent variables. The threshold may be determined by the methods described in Section 3. Alternatively, the coefficients may be sorted according to their importance determined by various criteria such as mutual information or correlation between the input and output wavelet coefficients. The model may then be developed by including coefficients starting with the most important, with the stopping point determined by cross-validation. Application of this approach to practical problems indicates that it can result in models that provide more physical insight into the system being modeled and may also have more accurate parameters and better predictive ability.^{26–28}

While multiscale modeling by thresholding the latent variables can perform better than existing methods, it does not truly integrate the modeling and thresholding steps and lacks a rigorous statistical basis. Proper integration of modeling and thresholding requires solution of the following optimization problem:

$$\begin{aligned} & \underset{\hat{z}_A}{\text{minimize}} \left\{ (z - \hat{z})^T V^{-1} (z - \hat{z}) \right\} \\ & \text{subject to} \\ & A\hat{z} = 0 \\ & \hat{z} = W^T [\text{threshold}(Wz)] \end{aligned} \quad (17)$$

where \hat{z} denotes the reconstructed input and output variables after thresholding. This is a computationally expensive problem owing to the non-linearity of the thresholding constraint. The problem formulated in equation (17) is similar to errors-in-variables modeling with an additional thresholding constraint.

A more efficient and statistically rigorous formulation of the multiscale modeling problem has been suggested recently by formulating multiscale modeling as a Bayesian estimation problem.³⁰ This approach maximizes the probability of the estimates given the measured data. By applying Bayes rule, the objective function may be represented as

$$\underset{\hat{z}_A}{\text{maximize}} \frac{P(z|\hat{z})P(\hat{z})}{P(z)} \quad (18)$$

If the errors are Gaussian and the probability distribution of the noise-free measurements, $P(\hat{z})$, is uniform, equation (18) reduces to the objective function in equation (17). If a better estimate of $P(\hat{z})$ is available, it is possible to improve the quality of the filtered data and empirical model. If the noise-free variables are deterministic, then their wavelet coefficients are known to consist of only a few large coefficients. Consequently, their probability distribution at scale m , $P(G_m\hat{z})$, may be

approximated as an exponential function, and equation (18) may be represented as

$$\underset{A, G_m \hat{z}}{\text{minimize}} \left\{ (G_m z - G_m \hat{z})^T V_m^{-1} (G_m z - G_m \hat{z}) + \sigma_{\hat{m}}^{-1} |G_m \hat{z}| \right\} \quad (19)$$

where V_m denotes the covariance of the noise at scale m , and $\sigma_{\hat{m}}$ is the estimated standard deviation of the noise-free wavelet coefficients at scale m . The standard deviation at each scale may be estimated by the VisuShrink criterion described in Section 3. If the noise-free variables are stochastic in nature, it may be better to assume the probability distribution at each scale to be Gaussian. In this case, equation (19) is modified to

$$\underset{A, G_m \hat{z}}{\text{minimize}} \left\{ (G_m z - G_m \hat{z})^T V_m^{-1} (G_m z - G_m \hat{z}) + (G_m \hat{z})^T V_m^{-1} (G_m \hat{z}) \right\} \quad (20)$$

subject to the constraints in equation (18), where V_m denotes the covariance of the noise-free wavelet coefficients at scale m . The second term in equations (19) and (20) regularizes the wavelet coefficients at each scale, which has an effect similar to soft thresholding.³¹ This problem formulation for multiscale modeling is statistically sound, and application to various modeling tasks indicates superior results.³⁰

5. CONCLUSIONS

Most practical measured data contain contributions at multiple scales due to events with different localization in time and frequency, scale and time-dependent stochastic processes, and variables measured at different rates or containing segments of missing data. Multiscale methods are ideally suited for the analysis and modeling of such multiscale data. The development of wavelets has resulted in several novel techniques for improved data analysis and empirical modeling. These methods exploit the multiscale representation of wavelet basis functions and their ability to compress deterministic features in a small number of large coefficients and to approximately decorrelate a wide variety of stochastic processes.

Efficient removal of stationary Gaussian noise from an underlying deterministic signal is possible by eliminating wavelet coefficients smaller than a threshold. This approach has been extended to the removal of other types of errors, including non-Gaussian errors, and for on-line multiscale filtering. The on-line approach subsumes existing linear filtering methods and can automatically adapt the scale of the filter to the nature of the signal features. Using the on-line multiscale filtering methods for statistical process control results in a general method that is better at detecting deterministic changes by automatically specializing to existing control charts depending on the nature of the data.

The benefits of the multiscale representation have been extended to empirical modeling by combining wavelet thresholding with empirical modeling. A popular approach for multiscale modeling is to threshold the latent variables. Unfortunately, this approach may not reach the true optimum of the multiscale modeling problem without computationally expensive optimization. Interpretation of multiscale modeling as Bayesian estimation is expected to result in a statistically rigorous and computationally efficient methodology.

This paper has provided an overview of multiscale data analysis and empirical modeling based only on wavelets. Each of the methods described in this paper may also be extended to using a library of basis functions, such as wavelet packets, to obtain better results.³¹⁻³⁴ Wavelets provide a mathematical framework for understanding several existing methods and for developing new techniques that are able to handle multiscale data efficiently. So far, most research has focused on

multiscale univariate filtering and feature extraction. It is expected that future work on multiscale modeling, multiscale methods for handling missing data and for solving other process operation and chemometric tasks will permit more efficient and better extraction of the information contained in measured data.

ACKNOWLEDGEMENTS

Partial financial support from the Technical Association of the Pulp and Paper Industry (PE-357-95), the American Chemical Society—Petroleum Research Fund (30523-G9) and DuPont is gratefully acknowledged.

REFERENCES

1. J. T. -Y. Cheung, 'Representation and extraction of trends from process data', ScD Thesis, Massachusetts Institute of Technology (1992).
2. R. R. Coifman and M. V. Wickerhauser, *IEEE Trans. Info. Theory*, **IT-38**, 713–718 (1992).
3. C. Herley, J. Kovacevic, K. Ramchandran and M. Vetterli, *IEEE Trans. Signal Process.* **41**, pp. 3341–3359 (1993).
4. I. Daubechies, *Commun. Pure Appl. Math.* **41**, 909–996 (1988).
5. S. G. Mallat, *IEEE Trans. Pattern Anal. Machine Intell.* **PAMI-11**, 674–693 (1989).
6. D. L. Donoho, I. M. Johnstone, G. Kerkyacharian and D. Picard, *J. R. Statist. Soc. B*, **57**, 301–369 (1995).
7. B. R. Bakshi and G. Stephanopoulos, *Comput. Chem. Engng.* **18**, 267–302 (1994).
8. S. Top and B. R. Bakshi, 'Improved statistical process control using wavelets', Proc. Foundations of Computer Aided Process Operation, Snowbird, UT, July (1998).
9. A. Cohen, I. Daubechies and V. Pierre, *Appl. Comput. Harmonic Anal.* **1**, 54–81 (1993).
10. B. Walczak, Proc. Second Int. Chemometrics Research Meet, Veldhoven, May (1998).
11. B. K. Alsberg, A. M. Woodward and D. B. Kell, *Chemometrics Intell. Lab. Syst.* **37**, 215 (1997).
12. T. Nguyen and G. Strang, *Wavelets and Filter Banks*, Wellesley-Cambridge Press, Wellesley, MA (1996).
13. M. T. Tham and A. Parr, *Chem. Engng. Prog.* **90**(5), 46 (1994).
14. P. Heinonen and Y. Neuvo, *IEEE Trans. Acoust., Speech, Signal Process.* **ASSP-35**, 832 (1987).
15. I. M. Johnstone and B. W. Silverman, *J. R. Statist. Soc. B*, **59**, 319–351 (1997).
16. G. P. Nason, *J. R. Statist. Soc. B*, **58**, 463–479 (1996).
17. S. G. Mallat and S. Zhong, *IEEE Trans. Pattern Anal. Machine Intell.* **PAMI-14**, 710–732 (1992).
18. J. Liang and T. W. Parks, *IEEE Trans. Signal Process.* **44**, 225 (1996).
19. R. R. Coifman and D. L. Donoho, in *Wavelets and Statistics*, ed. by A. Antoniadis and G. Oppenheim, Springer, New York (1995).
20. A. G. Bruce, D. L. Donoho, H.-Y. Gao and R. D. Martin, *Proc. SPIE*, **2242**, 325–336 (1994).
21. R. von Sachs and K. Schneider, *Appl. Comput. Harmonic Anal.* **3**, 268–282 (1996).
22. B. Walczak and D. L. Massart, *Chemometrics Intell. Lab. Syst.* **36**, 81 (1997).
23. C. R. Mittermayr, S. G. Nikolov, H. Hutter and M. Grasserbauer, *Chemometrics Intell. Lab. Syst.* **34**, 187 (1996).
24. M. N. Nounou and B. R. Bakshi, *AIChE J.* **45**, 1041–1058 (1999).
25. E. R. Malinowski, *Factor Analysis in Chemistry*, Wiley, New York (1991).
26. B. R. Bakshi, *AIChE J.* **44**, 1596 (1998).
27. B. K. Alsberg, A. M. Woodward, M. K. Winson, J. J. Rowland and D. B. Kell, *Anal. Chim. Acta*, **368**, 29–44 (1998).
28. J. Trygg and S. Wold, Proc. Second Int. Chemometrics Research Meet., Veldhoven, May (1998).
29. B. R. Bakshi, P. Bansal and M. N. Nounou, *Comput. Chem. Engng.* **21**, S1167–S1172 (1997).
30. M. N. Nounou and B. R. Bakshi, 'Multiscale linear modeling with application to inferential modeling and system identification', AIChE Ann. Meet., Miami Beach, FL (1998).
31. S. S. B. Chen, D. L. Donoho and M. A. Saunders, *SIAM J. Sci. Comput.* **20**, 33–61 (1999).
32. R. R. Coifman and N. Saito, *C. R. Acad. Sci. Paris, Ser. I*, **319**, 191–196 (1994).
33. M. V. Wickerhauser, *Adapted Wavelet Analysis from Theory to Software*, A. K. Peters, Wellesley, MA (1994).
34. B. Walczak and D. L. Massart, *Chemometrics Intell. Lab. Syst.* **38**, 39–50 (1997).