Proceedings of the
44th IEEE Conference on Decision and Control, and
the European Control Conference 2005
Seville, Spain, December 12-15, 2005

Dealing with Collinearity in FIR models using Multiscale Estimation

Mohamed N. Nounou, *Member, IEEE*

Abstract— **In this paper, multiscale representation of data is utilized to reduce the collinearity problem often encountered in Finite Impulse Response (FIR) modeling. The idea is to decompose the input-output data at multiple scales, use the scaled signal approximations of the data to construct a FIR model at each scale, and then select among all scales the optimum estimated FIR model. The rationale behind this approach is that the number of significant cross correlation function (CCF) coefficients estimated using the scaled signal approximations of the input-output data decreases at coarser scales. This means that more parsimonious FIR models, with less collinearity and improved estimation accuracy, can be constructed at coarser scales. Of course, the estimation accuracy will deteriorate at very coarse scales. Therefore, it is very important to select the most appropriate scale for modeling purposes, which can be done by selecting the scale which results in the maximum prediction signal to noise ratio. The developed multiscale FIR modeling approach is shown to outperform existing methods, such as ordinary least squares (OLS) regression and ridge regression (RR).**

I. Introduction

One of the most commonly used empirical models in control applications is the FIR model because of its ability to describe complex dynamical systems in simple model structures. However, a disadvantage of FIR models is the fact that they require a large number of parameters, which increases the collinearity (or redundancy in the model variables) which in turn increases the variance of estimated model parameters and degrades their accuracy.

Many estimation techniques have been developed to solve this collinearity problem of FIR models, such the reduced-rank models and Ridge Regression (RR) model. The reduced-rank models include Principal Component Regression (PCR) and Partial Least Squares (PLS) regression [1-3], which use Singular Value Decomposition (SVD) to decrease the dimension of the input variables in order to create a better conditioned model and damp the variations of the FIR coefficients. RR, on the other hand, reduces the variations of FIR coefficients by imposing a penalty on the norm of their estimated values [3-5]. This penalty effectively shrinks the FIR coefficients towards zero by introducing a bias that makes the input covariance matrix full-rank.

Another challenge of constructing empirical models in general is the presence of measurement noise in the data due to malfunctioning sensors or random perturbations in the process. The presence of such noise decreases the signal-to-noise ratio (SNR) of the data, which can have a drastic effect on the accuracy of estimated models [6]. Therefore, measurement noise needs to be removed or filtered in order to improve the model accuracy. Unfortunately, errors as well as important features in the data usually have a multiscale character, i.e., span wide ranges in both time and frequency. For example, a sudden change in the data spans a wide range in the frequency domain and a narrow range in the time domain, and in contrary, a slow change spans a wide rang in the time domain and a narrow range in the frequency domain. However, most filtering techniques classify noise as high frequency features, and filter the data by retaining the features with frequency lower than a defined frequency cutoff. Since multiscale data violate this basic assumption of conventional filters. Noise removal from such data becomes a difficult task. Thus, constructing models using multiscale data requires the representation of the data at multiple scales to account for their multiscale nature.

Modeling at multiple scales has been previously shown to improve the accuracy of estimated models [6-10]. The author in [7] developed a multiscale Principle Component Analysis (MSPCA) approach that combines the ability of PCA to decorrelate measured variables with that of multiscale representation to decorrelate autocorrelated measurements. The developed MSPCA approach, which possesses improved noise-removal ability, is then used in process monitoring. The authors in [6] showed that multiscale representation acts as a noise filter which reduces the effect of noise in the data on the estimated model accuracy. Also, the authors in [8,9] used multiscale representation to shrink the variations in estimated FIR model coefficients. The author in [8] represented the OLS estimated FIR coefficients at multiple scales using wavelets and then used a recursive approach to select the set of wavelet coefficients (and shrink the unnecessary ones) to minimize a cross validation mean squares error. The authors in [9], on the other hand, used multiscale representation to shrink measurement noise in the input variables before being used in modeling.

The objectives of this paper are to present some of the advantages of constructing empirical process models at multiple scales, and to present a new multiscale method that

M. N. Nounou is with the department of Chemical and Petroleum Engineering at the United Arab Emirates University, Al-Ain, UAE. P.O. Box 17555 (phone: +9713-713-3549; e-mail: mnounou@uaeu.ac.ae).

0-7803-9568-9/05/$20.00 ©2005 IEEE

improves the estimation accuracy of FIR models. One advantage is the ability of multiscale representation of data to separate important features from noise, which can improve the accuracy of estimated models. Another advantage is that the number of significant cross correlation function (CCF) coefficients relating the scaled signals of the input-output data decreases by half at coarser scales. The implication of this advantage is that more parsimonious FIR models can be constructed at coarser scales. These advantages are then used to develop a multiscale FIR (MSFIR) modeling algorithm, which helps improve the accuracy of estimated FIR models by reducing the effect of collinearity on their estimation.

The rest of the paper is organized as follows. In Section II, the FIR model representation and some of its estimation techniques are described. Then, in Section III, an introduction to wavelet-based multiscale representation of data is presented, followed by a description of some of the advantages of this representation in empirical process modeling. Then, in Section IV, the formulation and algorithm of MSFIR modeling are presented, followed by an illustrative example to show and compare the performance of MSFIR modeling to those of existing methods in Section V. Finally, in Section VI, the paper is concluded with few remarks.

II. Introduction to FIR Modeling

Consider the process input data, $u_k \in \{u_1, u_2, ..., u_n\}$, which are assumed to be noise-free, and measurements of the process output data, $y_k \in \{y_1, y_2, ..., y_n\}$, which are assumed to be contaminated with additive zero mean Gaussian noise, i.e.,

$$y_k = \tilde{y}_k + e_k \ , \qquad (1)$$

where, $\tilde{y}_k$ and e_k are the noise-free output and the additive noise at time step k. If it is assumed that the linear FIR model has the following form,

$$y_k = \sum_{i=1}^{m} \tilde{h}_i \, u_{k-i} + e_k \, , \qquad (2)$$

it is desired to estimate the noise-free process impulse response or FIR model coefficients, $\tilde{\mathbf{h}} = \{\tilde{h}_1, \tilde{h}_2, ..., \tilde{h}_m\}$. The FIR model shown in equation (2) can be written in matrix notation as follows,

$$\mathbf{Y} = \mathbf{U}\tilde{\mathbf{h}} + \mathbf{e} \qquad (3)$$

where,

$$\mathbf{Y} = \begin{bmatrix} y_n \\ y_{n-1} \\ . \\ y_2 \\ y_1 \end{bmatrix}, \mathbf{U} = \begin{bmatrix} \mathbf{u}_n^T \\ \mathbf{u}_{n-1}^T \\ . \\ \mathbf{u}_2^T \\ \mathbf{u}_1^T \end{bmatrix}, \mathbf{e} = \begin{bmatrix} e_n \\ e_{n-1} \\ . \\ e_2 \\ e_1 \end{bmatrix}, \ \mathbf{u}_k = \begin{bmatrix} u_{k\text{-}1} \\ u_{k\text{-}2} \\ . \\ u_{k\text{-}m} \end{bmatrix}.$$

Many techniques have been developed to solve this modeling estimation problem, which include OLS (Ljung, 1987), in which the output prediction error is minimized, and have the following closed form solution,

$$\{\hat{\mathbf{h}}\}_{OLS} = \left(\mathbf{U}^T \mathbf{U}\right)^{-1} \mathbf{U}^T \mathbf{Y}. \qquad (4)$$

However, inverting the input covariance matrix becomes problematic in the case of collinearity, which increases the variance of the estimated model parameters and increases the uncertainty about their estimation. RR, on the other hand, improves the estimation accuracy of model parameters by imposing a penalty of their magnitude, and has the following closed form solution [3-5],

$$\{\hat{\mathbf{h}}\}_{RR} = \left(\mathbf{U}^T \mathbf{U} + \lambda \mathbf{I}\right)^{-1} \mathbf{U}^T \mathbf{Y}. \qquad (5)$$

RR is very popular, but a Bayesian interpretation of RR [11] shows that RR shrinks the estimated FIR coefficients toward zero in order to damp their large variations, which is not very rigorous because it is known that the mean of the FIR coefficients is not zero.

III. Multiscale Data Representation

A proper way of analyzing multiscale data requires their representation at multiple scales. A signal can be represented at multiple resolutions by decomposing it on a family of wavelets and scaling functions. For example, the signals in Figures 1-(b, d, and f) are at increasingly coarser scales compared to the original signal in Fig. 1(a).

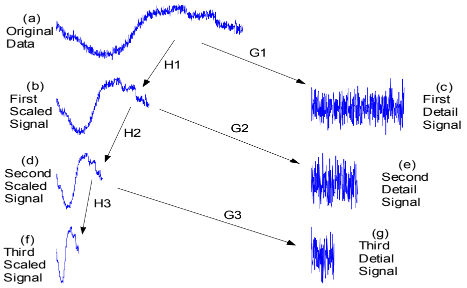

Fig. 1. A schematic diagram of data representation at multiple scales.

These scaled signals are determined by projecting the original signal on a set of orthonormal scaling functions. On the other hand, the signals in Figures 1-(c, e, and g), which are called the detail signals, capture the details between the scaled signal at particular scale and the scaled signal at the finer scale. These detail signals are determined by projecting the signal on a set of basis functions called wavelets. Therefore, the original signal can be represented as the sum of all detail signals and the last scaled signal as,

$$x(t) = \sum_{k=1}^{n2^{-L}} x_{LK} \, \varphi_{LK}(t) + \sum_{j=1}^{L} \sum_{k=1}^{n2^{-j}} dx_{jk} \psi_{jk}(t) \qquad (6)$$

where, j, k, L, and n are the dilation parameter, translation parameter, maximum number of scales or decomposition depth, and the length of the original signal, respectively. Fast wavelet transform algorithms of O(n) complexity for a discrete signal of dyadic length have been developed [12].

Just as an example to introduce some terminology, if a discrete signal, $\mathbf{Y}_o$, of length "n" in the time domain (i.e., $j = 0$) is defined as,

$$\mathbf{Y}_o = [y_o(1) \quad y_o(2) \quad . \quad y_o(k) \quad . \quad y_o(n)]^T, \qquad (7)$$

then, the scaled signal approximation of $\mathbf{Y}_o$ at scale (j), which will be denoted by $\mathbf{Y}_j$ will be written as,

$$\mathbf{Y}_j = [y_j(1) \quad y_j(2) \quad . \quad y_j(k) \quad . \quad y_j(n/2^j)]^T \quad . \qquad (8)$$

This decomposition algorithm is batch, i.e., requires the availability of the entire data set beforehand. An on-line wavelet decomposition algorithm has also been developed and used in data filtering [13].

IV. Multiscale Modeling

A. Advantages of Modeling at Multiple Scales

One advantage of modeling at multiple scales is that smaller model structures (fewer number of FIR coefficients) are needed at coarser scales. This is because the cross correlation function (CCF) relating the scaled signal approximations of the input and output data shrinks (defined at less number of time lags) at coarser scales. To illustrate this phenomenon of CCF shrinkage at multiple scales, the CCF is compared at different scales (using the Haar filters) as shown in Figure 2 for a simulated data representing the following Moving Average MA(4) model,

$$y_i = 0.9u_i + 0.8u_{i-1} + 0.7u_{i-2} + 0.5u_{i-3} + 0.3u_{i-4} \quad (9)$$

where the input is a pseudo-random binary sequence (PRBS) changing between -1 and 1. Figure 2 clearly shows that the number of important CCF coefficients decreases by half at every subsequent coarser scale. This observation, which is attributed to down-sampling, will be helpful to improve the parsimony and thus reduce the collinearity of FIR models at coarser scales.

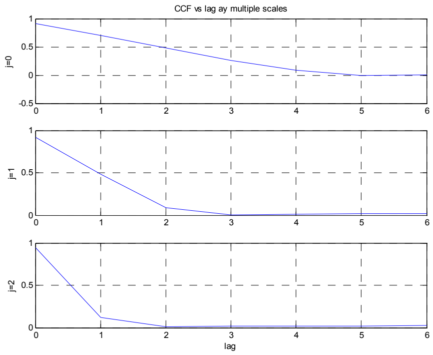

Fig. 2. The behavior of the CCF of the scaled input-output signals at multiple scales using Haar.

Another advantage of multiscale representation of data is its ability to reduce the noise content in measured data through the application of low-pass and high pass filters derived from the scaling and wavelet basis functions.

The noise reduction can be verified by comparing the SNR of the scaled signals at different scales. Theoretically, the SNR at any scales can be computed as follows,

$$SNR(j) = \frac{\mathrm{var}(\tilde{y}_j)}{\mathrm{var}(\tilde{y}_j - y_j)}, \qquad (10)$$

where $\tilde{y}_j$, is the noise free scaled signal representation of the data at scale (j). It can be easily shown through empirical simulation that the SNR of the scaled signals peaks at some intermediate scale, which can be explained as follows. At very fine scales, high frequency noise gets filtered out, which decreases the noise content and increases the SNR. However, at very coarse scales, important features start getting removed, which decreases the signal content and decreases the SNR. Therefore, there is an intermediate scale at which the SNR peaks. This observation is very useful in selecting the optimum modeling scale.

B. Multiscale FIR Modeling

In this section, the observation of CCF shrinkage is exploited to construct more parsimonious FIR models. The idea is that since the number of significant CCF coefficients decreases by half at every subsequent coarser scale, smaller and smaller FIR models with less collinearity can be constructed at coarser scales. However, there will be an intermediate scale at which the quality of FIR model is best because at very coarse scales, the FIR model size becomes too small. Therefore, it is important to select the optimum scale for model estimation, which can be done by picking the scale at which the SNR of the model prediction is a maximum.

Assume that the cross correlation function of the time domain input and output data has "p" significant coefficients. Then, the time-domain FIR model should also have "p" coefficients i.e.,

$$y_{o,k} = \sum_{i=1}^{p} h_{o,i}\, u_{o,k-i} \ , \qquad (11)$$

where, the subscripts "k", "o", and "i" denote the kth data sample (where, $k \in \{1, 2, ..., n\}$), scale zero (time domain scale), and the ith FIR model parameter (where, $i \in \{1, 2, ..., p\}$). Since the number of significant CCF coefficients decreases by half at every coarser scale, a FIR model of length (p/2) parameters is needed at the first scale, i.e.,

$$y_{1,k} = \sum_{i=1}^{p/2} h_{1,i}\, u_{1,k-i} \ . \qquad (12)$$

Similarly, the FIR model relating the scaled signal data at the second scale will have half the number of FIR model parameters at the first scale (i.e., p/4), and therefore, the second scale model cab be written as,

$$y_{2,k} = \sum_{i=1}^{p/4} h_{2,i}\, u_{2,k-i} \ . \tag{13}$$

Generalizing the above models to any scale (j), the FIR model at scale (j) will need only ($p/2^j$), and thus the jth scale FIR model can be written as,

$$y_{j,k} = \sum_{i=1}^{p/2^j} h_{j,i}\, u_{j,k-i} \ . \tag{14}$$

Note the estimated multiscale FIR model at any scale can not be directly used in the time domain because it relates the scaled signal approximations of the input-output data, and not the time domain data. To get an equivalent time-domain FIR model to that estimated at any scale (j), the following can be done. First, compute the scaled signal of an impulse function at scale j. Then, apply the decomposed impulse function as an input to the estimated MSFIR model at scale (j). The resulting output is the scaled signal approximation of the process impulse response at scale (j). Finally, reconstruct the model output obtained earlier to the time domain to get the time-domain equivalent of the FIR model parameters estimated at scale (j).

C. *Multiscale FIR (MSFIR) Modeling Algorithm*

Based on the above formulation, the following MSFIR modeling algorithm is proposed:

1. Compute the cross correlation function for the available input-output data set, and determine its settling length, (p).
2. Using the time domain data, estimate an FIR model of length "p", and compute the SNR of its prediction as follows,

   $$SNR = \mathrm{var}(\hat{y}) / \mathrm{var}(y - \hat{y})$$

3. Compute the scaled signals for the input and output data at multiple scales, and at each scale (j) construct a FIR model of length $p/2^j$ using OLS regression and compute the predicted output SNR as above.
4. Choose the multiscale FIR model with highest SNR as the optimum MSFIR model.
5. Compute the time-domain equivalent of the estimated MSFIR as described earlier at the end of subsection B.

V. Illustrative Example

In this section, the performance of the MSFIR approach described in Section III is illustrated and compared to those of some of the existing methods, such as OLS and RR. In this example, the various techniques are compared by computing the mean squared errors of the estimated FIR model parameters with respect to their true noise-free values, i.e.,

$$MSE = \frac{1}{m}\sum_{i=1}^{m}\left(\hat{h}_i - \tilde{h}_i\right)^2 \tag{15}$$

where $\hat{h}$ and $\tilde{h}$ are the estimated and noise-free FIR model parameter vectors, respectively, and the estimated process gain, i.e.,

$$G = \sum_{i=1}^{m} \hat{h}_i \quad . \tag{16}$$

The process used in this simulation has the following second order plus dead time (SOPDT) model [14]:

$$\frac{Y(s)}{U(s)} = \frac{-5e^{-4s}}{(5s+1)(3s+1)}, \tag{17}$$

which when discretized using a sampling interval of 0.2 min, has an impulse response with a settling time of around 250 sampling intervals as shown in Figure 3. The discretized process model is used to generate data by applying a 2000-sample PRBS input signal to the process model to give noise-free output, which is then contaminated with additive zero mean Gaussian noise. Different levels of noise contents (standard deviations of 0.1, 0.5, and 1) have been used to test the robustness of the MSFIR algorithms. Fig. 4 shows a portion of the input and output data in which the standard deviation of the output noise is 0.5.

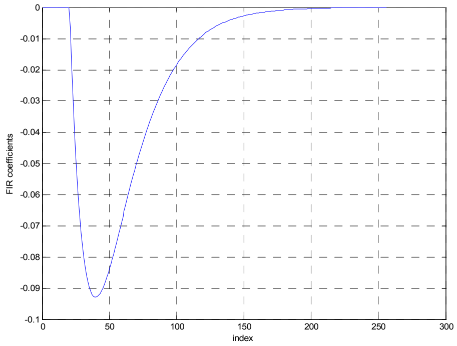

Fig. 3. The impulse response of the SOPDT process used in the illustrative example.

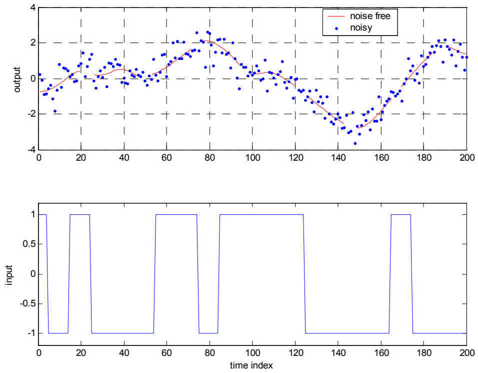

Fig. 4. A 200-sample data set of the input and output data used in the simulation of example 1 using noise standard deviation of 0.5.

To statistically compare the performances of MSFIR with those of OLS and RR, a Monte Carlo simulation of 100 realizations is performed for each noise level, and the results are summarized in Tables I, II, and III. Table I, which lists the MSE of estimated model parameters for the various methods, shows that for each noise level, there is a scale at which the estimated FIR model parameters have a smaller MSE than both OLS and RR. Also, the magnitude of improvement over OLS and RR is very clear for all noise contents.

TABLE I
COMPARISON OF THE PREDICTION MEAN SQUARE ERRORS FOR THE VARIOUS MODELING METHODS ($\times 10^4$)

Modeling Method	$\sigma_e = 0.1$	$\sigma_e = 0.5$	$\sigma_e = 1$
RR	0.41	1.8	2.2
OLS	11	290	1200
MSFIR (j=1)	1.30	39	160
MSFIR (j=2)	0.150	1.3	20
MSFIR (j=3)	0.45	0.54	1.54
MSFIR (j=4)	1.70	1.8	1.89
MSFIR (j=5)	4.50	4.6	4.4

TABLE II
COMPARISON OF THE PROCESS GAINS ESTIMATED BY THE VARIOUS MODELING METHODS (TRUE VALUE = -5)

Modeling Method	$\sigma_e = 0.1$	$\sigma_e = 0.5$	$\sigma_e = 1$
RR	-4.834	-4.328	-4.760
OLS	-5.001	-5.050	-4.981
MSFIR (j=1)	-4.999	-5.055	-4.981
MSFIR (j=2)	-5.005	-5.070	-4.981
MSFIR (j=3)	-5.0674	-5.129	-4.958
MSFIR (j=4)	-5.055	-5.086	-5.065
MSFIR (j=5)	-4.921	-5.017	-5.385

Table II, on the other hand, which lists the process gains estimated by the various methods, shows that the gains estimated by MSFIR and OLS are very close to the true process gain, and that RR is much worse even when its prediction is comparable to that of MSFIR, as in the case where noise standard deviation is unity. Also, Table III shows that the in most cases, the correct optimum scales (which also matched the minimum parameter MSE) were selected by maximizing the SNR of the model prediction.

TABLE III
PERCENTAGES EACH SCALE SELECTED AS OPTIMUM USING THE SNR CRITERION

Scale	$\sigma_e = 0.1$	$\sigma_e = 0.5$	$\sigma_e = 1$
j=0	0	0	0
j=1	0	0	0
j=2	40	0	0
j=3	60	80	0
j=4	0	20	100
j=5	0	0	0

The improvement achieved by MSFIR can also be seen from Figure 5, which compares the estimated FIR model coefficients using RR and MSFIR at the optimum scale (scale 3) for a noise standard deviation of 0.5.

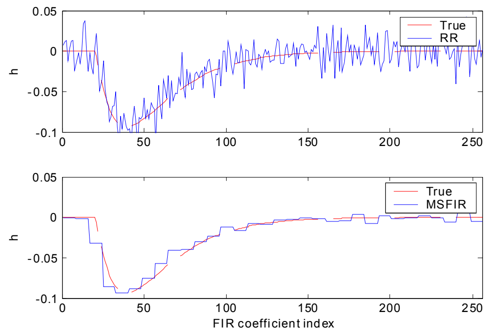

Fig. 5. Comparison of the estimated MSFIR model at optimum scale (scale 3) with that obtained using RR, for the case where the noise standard deviation is 0.5.

Also, to show the advantage of constructing MSFIR models, the time-domain equivalents of the MSFIR models at the first five scales are compared in Figure 6, which shows that the accuracy of estimated models improves at coarser scales until an optimum scale (scale three in this case), after which it deteriorates. Here, scale three was selected as the optimum scale bases on the maximum SNR criterion, and as Figure 7 shows, this scale also matches with the least parameter MSE.

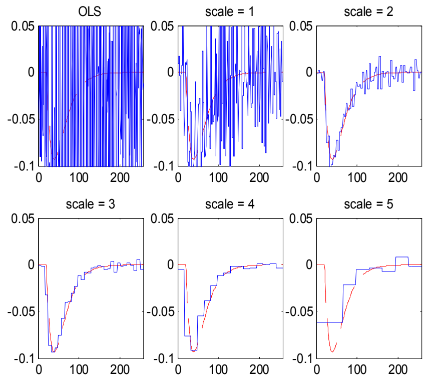

Fig. 6. Comparison the estimated FIR coefficients at different scales for the case where the noise standard deviation is 0.5.

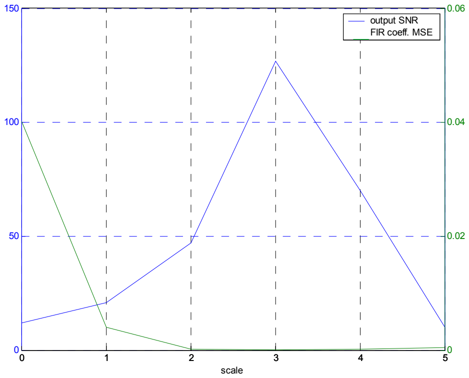

Fig. 7. Estimating the optimum scale as the maximum output SNR for the case where the noise standard deviation is 0.5.

VI. CONCLUSIONS

The objective of this paper was twofold: it discussed some of the advantages of constructing empirical process models at multiple scales, and presented a new multiscale approach for estimating FIR models. One advantage of multiscale representation is that it helps separate measurement noise from important features in the data. This advantage reduces the effect of measurement errors on the accuracy of estimated models. Another advantage of multiscale representation is the fact that the number of significant CCF coefficients relating the scaled signal approximations of the input-output data decreases by half at every subsequent coarser scale. These advantages are exploited to develop a multiscale FIR (MSFIR) modeling algorithm. The developed algorithm estimates smaller (and thus less collinear) FIR models at multiple scales using the scaled signals of the input and output data. Then, from all scales, the model which results in the maximum prediction signal to noise ratio is selected as the optimum model. The performance of the developed MSFIR modeling algorithm is shown to outperform existing FIR model estimation methods, such as OLS and RR, through a simulated example.

REFERENCES

[1] Wise, B.M. and Ricker, N.L. (1992). Identification of Finite Impulse Response Models by Principal Component Regression: Frequency-Response Properties", Process Control and Quality, 4, 77-86.

[2] Ricker, N.L. (1988). The Use of Biased Least-Squares Estimators for Parameters in Discrete-Time Pulse-Response Models. *Ind. Eng. Chem. Res.,* 27, 2, 243.

[3] Frank, I.E. and J.H. Friedman (1993) A statistical View of Some Chemometric Regression Tools. *Technometrics*, 35, 2, 109-148.

[4] Hoerl, A.E. and Kennard, R.W. (1970). Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*, 8, 27-52.

[5] McGregor, J., Kourti, T., and Kresta, J. (1991). Multivariate Identification: A Study of Several Methods. *IFAC ADCHEM Proc.*, Toulouse, France.

[6] Palavajjhala, S., Motrad, R., and Joseph, B. (1996) Process Identification Using Discrete Wavelet Transform: Design of Prefilters. *AICHE J.*, 42, 3, 777-790.

[7] Bakshi, B. R. (1998). Multiscale PCA with Application to Multivariate Statistical Process Monitoring. *AIChE Journal*, 44, 7, 1596-1610.

[8] Nikolaou, M. and Vuthandam, P. (1998). FIR Model Identification: Achieving Parsimony through Kernel Compression with Wavelets. *AICHE J.*, 44, 1, 141-150.

[9] Robertson, A.N., Park, K.C., and Alvin, K.F. (1998). Extraction of Impulse Response Data via Wavelet Transform for Structural System Identification. *Journal of Vibration and Acoustics*. 120, 252-260, 1998.

[10] Bakshi, B. R. (1999). Multiscale Analysis and Modeling Using Wavelets. *Journal of Chemometrics*, 13, 3-4, 415-434.

[11] Nounou, M. N., Bakshi, B. R., Goel, P. K., and Shen, X. (2002). Process Modeling by Bayesian Latent Variable Regression. *AIChE J.*, 48, 8, 1775-1793.

[12] Mallat, S.G. (1989). A Theory of Multiresolution Signal Decomposition: The Wavelet Representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11, 7, 764.

[13] Nounou, M.N. and Bakshi, B.R., (1999). Online Multiscale Filtering of Random and Gross Errors without Process Models. *AIChE Journal*, 45, 5, 1041-1058.

[14] Seborg, D., Edgar, T., and Mellichamp, D. (1989). *Process Dynamics and Control*, John Wiley and Sons Inc., New York.