Noisy Data and Impulse Response Estimation

Soosan Beheshti, *Senior Member, IEEE*, and Munther A. Dahleh, *Fellow, IEEE*

***Abstract*—This paper investigates the impulse response estimation of linear time-invariant (LTI) systems when only noisy finite-length input-output data of the system is available. The competing parametric candidates are the least square impulse response estimates of possibly different lengths. It is known that the presence of noise prohibits using model sets with large number of parameters as the resulting parameter estimation error can be quite large. Model selection methods acknowledge this problem, hence, they provide metrics to compare estimates in different model classes. Such metrics typically involve a combination of the available least-square output error, which decreases as the number of parameters increases, and a function that penalizes the size of the model. In this paper, we approach the model class selection problem from a different perspective that is closely related to the involved denoising problem. The method primarily focuses on estimating the parameter error in a given model class of finite order using the available least-square output error. We show that such an estimate, which is provided in terms of upper and lower bounds with certain level of confidence, contains the appropriate tradeoffs between the bias and variance of the estimation error. Consequently, these measures can be used as the basis for model comparison and model selection. Furthermore, we demonstrate how this approach reduces to the celebrated AIC method for a specific confidence level. The performance of the method as the noise variance and/or the data length varies is explored, and consistency of the approach as the data length grows is analyzed.**

***Index Terms*—Least square estimate, LTI system modeling, noisy data.**

I. Introduction

A common method of LTI system modeling from data is to find the least square estimate of the impulse response. We consider the problem of overparametrization of these estimators in the presence of noisy data. This problem has been the key motivation of model selection approaches. Essentially, these approaches point to the fact that lower dimensional models can produce better parameter estimates by minimizing some risk assessment criterion. These criteria are generally proposed for the sole purpose of model set comparison and do not carry any particular information on the quality of the estimator in any individual model set. In this paper, we primarily focus on the problem of estimator quality evaluation in a given model set and employ the estimation error for this purpose. We denote the square of the estimation error in the parameter space and in the output space as the *parameter error* and the *reconstruction error*, respectively. These errors inherently include bias-variance tradeoffs and consequently can be used as risk assessments for the model selection procedure. In the absence of noise, it is logical to choose the model set with the highest order among a set of nested model sets, as the error is a decreasing function of model order. On the other hand, in the presence of the additive noise, a model set can exist with order m in which the error is smaller than that of the model set with order $m+1$. This is due only to the fact that the estimate of order m is less noisy than that of order $m+1$. Accordingly, minimizing the square error among the nested model sets results in the choice of the *least noisy* estimate among the competing estimates.

The importance of the square of the estimation error and its mean, denoted by the mean-square error (MSE), has been acknowledged in most data-based estimation approaches. The main challenge, however, is in calculating these values by using only the observed data [10]. We investigate the possibility of computing an estimate of such errors using the computable output error. For a finite data set, only probabilistic bounds of such an error can be derived for every model class. We explicitly compute such probabilistic bounds. The model selection problem is then formulated as finding the model class that minimizes the upper bound of this error for a fixed level of confidence. This approach results in a measure computed from finite observations that provides the appropriate tradeoff between bias and variance in a systematic way across all the competing model sets.

It is important to note that our method is more than an order selection approach. In cases that the true model has finite-length impulse response (FIR), we provide conditions for the consistency and can choose the true FIR as the data length grows. Nevertheless, the method also applies to the cases that the true length is known. In this case, we show that, due to the finiteness of the data, the least noisy estimate may not be the one with the same length as the true FIR. Our approach can also be used when the impulse response is infinite-length impulse response (IIR). In this case, the competing candidates are the least square estimates of the finite part of the IIR that are provided by the available data. The method chooses the least noisy one among these estimates. The provided least noisy estimate can then be utilized for modeling a rational transfer function in a zero-pole modeling process [14], [18].

The paper is arranged as follows. Section II contains the preliminaries and Section III includes the notations, motivation, and the considered problem. Section IV discusses the importance of the estimation errors. In Section V, the structure and behavior of the reconstruction error in subspaces with different orders are investigated, and the connection between the denoising approach and model selection is shown. Section VI uses the observed data to estimate the reconstruction error. Section VII

Manuscript received June 18, 2009; accepted August 20, 2009. First published September 09, 2009; current version published January 13, 2010. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Tryphon T. Georgiou.

S. Beheshti is with the Department of Electrical and Computer Engineering, Ryerson University, Toronto, ON M5B 2K3 (e-mail: soosan@ee.ryerson.ca).

M. A. Dahleh is with the Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA 02139 USA (e-mail: dahleh@mit.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TSP.2009.2032031

1053-587X/$26.00 © 2010 IEEE

Authorized licensed use limited to: MIT Libraries. Downloaded on February 25,2010 at 18:13:19 EST from IEEE Xplore. Restrictions apply.

provides the connection between the reconstruction error and the criteria in a class of exiting order selection approaches. In Sections VIII and IX more details about estimating the reconstruction error are provided, and topics such as consistency and prior information on the true order are discussed. Application of the method for two cases with FIR and IIR filters is demonstrated in Section X. The paper concludes with Section XI.

II. Preliminaries

We consider a class of stable, causal, single-input/single-output, linear time-invariant, discrete-time systems. The system is at rest and all initial conditions are zero. Input and noiseless output of the system are related as follows[1]:

$$\bar{y}[n] = \sum_{i=0}^{M-1} a_i^* u[n-i] \tag{1}$$

where a_i^*s are the real-valued coefficients of the system's impulse response.[2] The observed output is the noisy version of the output

$$y[n] = \bar{y}[n] + w[n] \tag{2}$$

where $w[n]$ is an additive white Gaussian noise (AWGN) with zero-mean and variance σ_w^2. The additive noise is independent of the input. Finite-length data, input $u^N = [u[1], \ldots, u[N]]^T$ and output $y^N = [y[1], \ldots, y[N]]^T$, are available. The goal is to provide a good representative of the true parameter

$$\theta^* = [a_0^*, a_1^*, \ldots, a_{M-1}^*]^T, \quad \theta^* \in S_M \tag{3}$$

given the observed noisy and *finite*-length data. The true parameter θ^* is a member of a compact set S_M, which is a subset of R^M. For now it is assumed that M, an upperbound for the order of the true parameter, is available and is less than the data length, $M \leq N$. These assumptions are later relaxed in Section IX.

III. Notations, Motivation, and Problem Formulation

A. Estimation in S_M

Noiseless Data $\bar{y}$ and True Parameter (Impulse Response Coefficients) θ^*: The unavailable noiseless output $\bar{y}^N = [\bar{y}(1), \bar{y}(2), \ldots, \bar{y}(N)]^T$ in (1) is related to the unknown true parameter θ^* as follows:

$$\bar{y}^N = A_{S_M} \theta^* \tag{4}$$

where A_{S_M} is the $N \times M$ Toeplitz matrix generated by the input u and is assumed to have a full rank (input is persistently exciting of order M [12], [19]).

Noisy Data y^N: Noisy data is the corrupted version of the noiseless data as given in (2).

Parameter Estimate of Order M, $\hat{\theta}_{S_M}(y^N)$: The least square estimator projects both the noiseless output and the additive Gaussian noise into the parameter's space and we have

$$\begin{aligned} \hat{\theta}_{S_M}(y^N) &= (A_{S_M}^T A_{S_M})^{-1} A_{S_M}^T y^N & (5) \\ &= \theta^* + (A_{S_M}^T A_{S_M})^{-1} A_{S_M}^T w^N. & (6) \end{aligned}$$

In the absence of the additive noise $(w^N = 0)$, this least square estimate is the true parameter itself

$$\theta^* = (A_{S_M}^T A_{S_M})^{-1} A_{S_M}^T \bar{y}^N \tag{7}$$

and therefore, if any element of θ^* is zero, or close to zero, it is captured precisely by using this projection.

However, the noisy term, $(A_{S_M}^T A_{S_M})^{-1} A_{S_M}^T w^N$, may fit too much of the noise to each component of the parameter estimate. The noise fitting issue is especially problematic and obvious if the $i\text{th}$ element of θ^* itself is zero or much smaller than the $i\text{th}$ element of this noisy term. In this case, it is better to set the estimate of the $i\text{th}$ element to zero instead of using the available noisy data. Setting the $i\text{th}$ element to zero is equivalent to searching for an estimate of the true parameter in a subset of S_M with a lower order. This observation has been the motivation of subset estimation and order selection approaches. In the following section, the related subset estimation is formulated.

B. Estimation in Subspaces of Order m, $1 \leq m \leq M$

For possible reduction of the effects of the additive noise, we consider the parameter estimates in subspaces of S_M. The subspace, in the following approach, can be any subspace of the S_M. However, for simplicity and without loss of generality our competing subspaces are nested subspaces of S_M with different orders, and the elements of subspace S_m have the following structure:

$$\theta = [a_0, a_1, \ldots, a_{m-1}, 0, \ldots, 0]^T \in S_m, \quad S_m \subset S_M. \tag{8}$$

True Parameter: The true parameter θ^* in (3) and (7) is

$$\theta^* = \begin{bmatrix} \theta_{S_m}^* \\ \Delta_{S_m} \end{bmatrix} \tag{9}$$

where Δ_{S_m} is a vector of length $N - m$, corresponding to the parameters that are not in the subspace S_m and which represents the unmodeled coefficients.

Noiseless Data: In each subspace, the noiseless data in (4) can be represented by

$$\bar{y}^N = [\, A_{S_m} \quad B_{S_m} \,] \begin{bmatrix} \theta_{S_m}^* \\ \Delta_{S_m} \end{bmatrix} \tag{10}$$

where A_{S_m} is the matrix with the first m columns of the Toeplitz matrix A_{S_M}, and B_{S_m} includes the rest of $N - m$ columns of the Toeplitz matrix.

Parameter Estimate of Order m, $\hat{\theta}_{S_m}(y^N)$: The least square estimate of θ^* in S_m is

$$\hat{\theta}_{S_m}(y^N) = \begin{bmatrix} (A_{S_m}^T A_{S_m})^{-1} A_{S_m}^T y^N \\ 0_{(N-m)\times 1} \end{bmatrix} \tag{11}$$

[1] For simplicity and without loss of generality we assume that both input and output are real-valued.

[2] The use of superscript "*" for the variables is adopted from the system identification literature such as [12].

Authorized licensed use limited to: MIT Libraries. Downloaded on February 25,2010 at 18:13:19 EST from IEEE Xplore. Restrictions apply.

$$= \begin{bmatrix} \theta^*_{S_m} + (A^T_{S_m} A_{S_m})^{-1} A^T_{S_m} (w^N + B_{S_m} \Delta_{S_m}) \\ 0_{(N-m)\times 1} \end{bmatrix}. \tag{12}$$

Data Estimate $\hat{y}^N_{S_m}$: Using the parameter estimate in subspace S_m results in an estimate of the observed data[3] [4], [20]

$$\begin{aligned} \hat{y}^N_{S_m} &= A_{S_M} \hat{\theta}_{S_m} && (16) \\ &= [A_{S_m}\ B_{S_m}] \hat{\theta}_{S_m}. && (17) \end{aligned}$$

Important Notation: In this paper, $E_\theta(c(Y^N))$ denotes the expected value of $c(Y^N)$ with pdf $f(Y^N;\theta)$ and $c(y^N)$ is value of the function $c(\cdot)$ at sample point y^N.

Also, to simplify notations throughout the paper, the parameter estimate is also denoted by $\hat{\theta}_{S_m}$, i.e., (y^N) in $\hat{\theta}_{S_m}(y^N)$ is eliminated.

IV. Important Factors: Errors in Subspace Selection

In subset estimation, only the first m elements of θ^* are estimated and the other $N-m$ elements are set to zero. The additive noise is now projected into a subspace of order m. However, the tradeoff is in setting the rest of the parameters to zero and possibly having a larger bias in the error. Is it better to consider the least square estimate of θ^* in a subspace of S_M or in S_M itself? To answer this question, we concentrate on the estimation error caused by the noisy data and by the choice of a subspace. In each subspace, the error caused by the estimation process is

$$\begin{aligned} e_{S_m} &= \theta^* - \hat{\theta}_{S_m} && (18) \\ &= \begin{bmatrix} -(A^T_{S_m} A_{S_m})^{-1} A^T_{S_m} (w^N + B_{S_m} \Delta_{S_m}) \\ \Delta_{S_m} \end{bmatrix} && (19) \end{aligned}$$

which includes both the noise effects (the part with w^N term) and the effects of the unmodeled coefficients Δ_{S_m}. To compare the behavior of this error in the subspaces, we can use the l^2 norm of this error and define the parameter error as

$$\textbf{Parameter error}: \quad j_{S_m} = \|\theta^* - \hat{\theta}_{S_m}\|^2_2. \tag{20}$$

Considering this distance measure for evaluation, the optimum parameter estimate is the one that minimizes this error. While the parameter error is an interesting criterion for comparison of the subspaces, it is not possible to directly calculate this error by using only the available data. On the other hand, an available error in each subspace is

$$\textbf{Data error}: \quad x_{S_m} = \frac{1}{N} \|y^N - \hat{y}^N_{S_m}\|^2_2 \tag{21}$$

which is the distance between the noisy data and its estimate in the subspaces. This output error is a decreasing function of m and is nonzero in all subspaces except in S_M. Minimizing this error always leads to choosing the space S_M itself. Therefore, the data error cannot evaluate and compare the parameter errors of different subspaces. Data error is a well-known component of the comparison criteria in existing order selection approaches, to which a penalty term (an increasing function of m) is usually added.

Another important related error is the reconstruction error, which represents the effect of the estimation error in the output space [3], [10]

$$\textbf{Reconstruction error}: \quad z_{S_m} = \frac{1}{N} \|\bar{y}^N - \hat{y}^N_{S_m}\|^2_2. \tag{22}$$

This error is the result of transforming the parameter error into the output space

$$z_{S_m} = \frac{1}{N} \|A_{S_M} (\theta^* - \hat{\theta}_{S_m})\|^2_2. \tag{23}$$

A novel approach in [3] provides an estimate of this error by using the available "data error" in a data denoising problem. The linear relation between the true parameter and the available data in the data denoising setting is analogous to the linear relation between the true parameter and the available output in the LTI system modeling. However, while here the linear transformation is through the Toeplitz matrix generated by the input, in the denoising approach the transformation matrix is orthogonal. In this paper, we focus on estimating the reconstruction error for a full rank Toeplitz matrix A_{S_M}, a transformation matrix that is not necessarily an orthogonal one. Due to the relationship between the parameter error and reconstruction error, in (23), we have

$$\frac{1}{\sigma_{\max}\left(\frac{A^T_{S_M} A_{S_M}}{N}\right)} z_{S_m} \le j_{S_m} \le \frac{1}{\sigma_{\min}\left(\frac{A^T_{S_M} A_{S_M}}{N}\right)} z_{S_m} \tag{24}$$

where $\sigma_{\min}$ and $\sigma_{\max}$ are the available minimum and maximum singular values of $A^T_{S_M} A_{S_M}/N$. Therefore, the estimate of the reconstruction error provides bounds on the parameter error. In the following sections the structure of the reconstruction error is studied. We provide probabilistic bounds on this error that capture the inherent tradeoff between the noise effects and the unmodeled coefficients effects on this error.

V. Structure of the Reconstruction Error

Using (10) and (16), the reconstruction error in (22) is in the form of

$$z_{S_m} = \frac{1}{N} \|G_{S_m} B_{S_m} \Delta_{S_m} + C_{S_m} w^N\|^2_2 \tag{25}$$

where

$$\begin{aligned} G_{S_m} &= I - A_{S_m} (A^T_{S_m} A_{S_m})^{-1} A^T_{S_m} && (26) \\ C_{S_m} &= A_{S_m} (A^T_{S_m} A_{S_m})^{-1} A^T_{S_m} && (27) \end{aligned}$$

are both projection matrices. While G_{S_m} is a projection matrix of rank $N-m$, C_{S_m} is a projection matrix of order m.

[3] The maximum likelihood (ML) estimate of θ^* in S_m is

$$\hat{\theta}_{S_m}(y^N) = \arg\max_{\theta \in S_m} f(y^N;\theta,u^N), \quad 1 \le m \le M \tag{13}$$

where $f(Y^N;\theta,u^N)$ is the probability distribution function (pdf) of the output given that the input is u and parameter θ has generated the data. Due to the structure of the data, the pdf is

$$\theta \in S_m : f(y^N;\theta,u^N) = \frac{1}{\left(\sqrt{2\pi\sigma^2_w}\right)^N} e^{-\|y^N - A_{S_m}\theta\|^2_2 / 2\sigma^2_w}. \tag{14}$$

Therefore, the ML estimate in (13) is also the least square estimate

$$\hat{\theta}_{S_m}(y^N) = \arg\min_{\theta \in S_m} \|y^N - A_{S_m}\theta\|^2, \quad 1 \le m \le M. \tag{15}$$

Authorized licensed use limited to: MIT Libraries. Downloaded on February 25,2010 at 18:13:19 EST from IEEE Xplore. Restrictions apply.

Lemma 1: The reconstruction error z_{S_m} is a sample of random variable Z_{S_m} and for this random variable we have

$$\frac{N}{\sigma_w^2}(Z_{S_m}) \sim \chi_m^2 \tag{28}$$

where χ_m^2 is a Chi-square random variable of order m. The expected value and the variance of Z_{S_m} are

$$\mathrm{E}(Z_{S_m}) = \frac{m}{N}\sigma_w^2 + \frac{1}{N}\|G_{S_m}B_{S_m}\Delta_{S_m}\|_2^2 \tag{29}$$

$$\mathrm{var}(Z_{S_m}) = \frac{2m}{N^2}(\sigma_w^2)^2. \tag{30}$$

Proof: In Appendix A. □

A. Bias-Variance Tradeoff

The MSE that is the expected value of Z_{S_m} has two terms. The first term $m\sigma_w^2/N$ is the noise dependent part and is a monotonically increasing function of m. This is a direct result of the projection with C_{S_m} which itself has rank m. The second term $1/N\|G_{S_m}B_{S_m}\Delta_{S_m}\|_2^2$ is a function of the unmodeled coefficients Δ_{S_m}. The norm of the unmodeled coefficients is a decreasing function of m and since the rank of the projection matrix G_{S_m} is also a decreasing function of m, the term $1/N\|G_{S_m}B_{S_m}\Delta_{S_m}\|_2^2$ tends to be a decreasing function of order m. This is traditionally called the bias-variance tradeoff [7], [9].

B. Asymptotic Behavior of Z_{S_m}

Using (31) and (32), the second-order statistics of Z_{S_m} are asymptotically

$$\lim_{N\longrightarrow\infty}\mathrm{E}(Z_{S_m}) = \lim_{N\longrightarrow\infty}\frac{1}{N}\|G_{S_m}B_{S_m}\Delta_{S_m}\|_2^2 \tag{31}$$

$$\lim_{N\longrightarrow\infty}\mathrm{var}(Z_{S_m}) = 0. \tag{32}$$

If the true parameter has order M^*, then Δ_{S_m} is zero in subspaces with orders higher than M^*. This will cause the $\mathrm{E}(Z_{S_m})$ to be zero in the limit for all $m \geq M^*$ and to be nonzero for all orders smaller than M^*. Fig. 1 shows a typical behavior of $E(Z_{S_m})$ as a function of m and as the data length grows.

It is important to mention that while the asymptotic behavior of a method is usually explored as the data length increases, it is also critical to show that the method is robust with respect to the noise variance as well. For a fixed data length, to check what happens as the noise variance changes, we fix the input power and increase the signal-to-noise ratio (SNR). In this case the unmodeled coefficients' effect is a fixed number and as SNR grows, only the noise variance goes to zero. Therefore, the behavior of $E(Z_{S_m})$ in the limit is

$$\lim_{SNR\longrightarrow\infty}\mathrm{E}(Z_{S_m}) = \frac{1}{N}\|G_{S_m}B_{S_m}\Delta_{S_m}\|_2^2 \tag{33}$$

$$\lim_{SNR\longrightarrow\infty}\mathrm{var}(Z_{S_m}) = 0. \tag{34}$$

Fig. 2 shows a typical behavior of $E(Z_{S_m})$ as a function of m and as the noise variance changes.

C. Confidence Bounds on z_{S_m}

Can knowing the pdf of Z_{S_m} help us bound the desired z_{S_m} that is a sample of this random variable? This information can definitely provide the following confidence region: Using the pdf of the random variable Z_{S_m}, for a given confidence probability p_1, there exists a D_{S_m} for which the reconstruction error z_{S_m} is bounded as follows:

$$\Pr\{|z_{S_m} - E(Z_{S_m})| \leq D_{S_m}\} = p_1. \tag{35}$$

The value of D_{S_m} is a function of p_1 and $2m/N^2(\sigma_w^2)^2$, the variance of Z_{S_m}, and can be calculated using the Chi-square CDF table. Therefore, with probability p_1 the reconstruction error is bounded with

$$\underline{z_{S_m}(p_1)} \leq z_{S_m} \leq \overline{z_{S_m}(p_1)} \tag{36}$$

where

$$\overline{z_{S_m}(p_1)} = E(Z_{S_m}) + D_{S_m} \tag{37}$$

$$= \frac{1}{N}\|G_{S_m}B_{S_m}\Delta_{S_m}\|_2^2 + \frac{m}{N}\sigma_w^2 + D_{S_m} \tag{38}$$

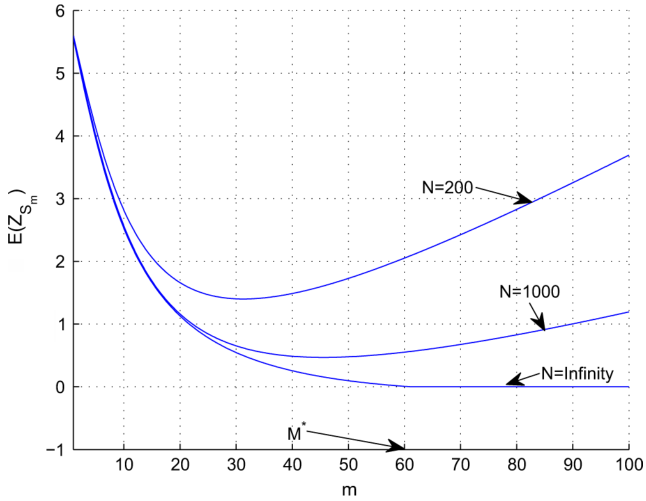

Fig. 1. Typical behavior of the expected value of the reconstruction error as the data length grows when the true parameter's order is $M^* = 60$ and the maximum subspace order is $M = 100$ (fixed SNR, fixed signal power).

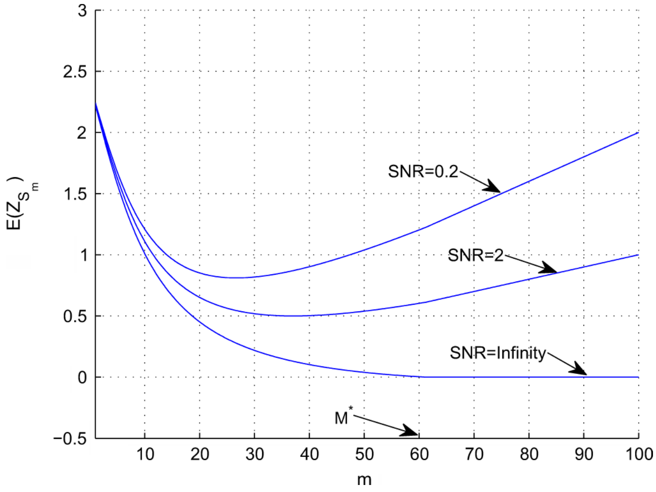

Fig. 2. Typical behavior of the expected value of the reconstruction error as the SNR (the noise variance) changes when the true parameter's order is $M^* = 60$ and the maximum subspace order is $M = 100$ (fixed data length, fixed signal power).

Authorized licensed use limited to: MIT Libraries. Downloaded on February 25,2010 at 18:13:19 EST from IEEE Xplore. Restrictions apply.

$$\underline{z_{S_m}(p_1)} = E(Z_{S_m}) - D_{S_m}. \tag{39}$$

D. Model Comparison

The bounds on z_{S_m} not only provide a probabilistic confidence region for z_{S_m}, but can also be used for comparison of the subspaces. In this case, the confidence region represents an event with probability p_1 where the upperbound $\overline{z_{S_m}(p_1)}$ is the worst case value of this event. Note that here both mean and variance of Z_{S_m} are functions of the order of the competing model set. Comparing the upperbounds of z_{S_m}, in the competing model sets, with the *same* confidence probability p_1 involves not only the mean but also the variance of the random variable in our decision making. For a fixed p_1, D_{S_m} is only a function of the variance, $2m/N^2(\sigma_w^2)^2$, and is an increasing function of m. Therefore, the upperbound has one component (the unmodeled coefficients effect) that tends to be a decreasing function of m and two terms $m/N\sigma_w^2 + D_{S_m}$ that are increasing functions of m. This is also a manifestation of another form of a bias-variance tradeoff. In this case the bias term is $E(Z_{S_m})$ and the variance term is D_{S_m} which carries the effect of the variance of Z_{S_m}.

Interestingly, the behavior of the reconstruction error and its mean shows that while capturing the true model order, as the data length grows, is important, in the presence of a finite-length data, the subset with order of the true parameter may not minimize the reconstruction error. In this case, the least noisy estimate usually have an order smaller than that of the true order.

VI. Probabilistic Bounds on Reconstruction and Parameter Errors

In the previous section we provided probabilistic bounds on the reconstruction error by using the mean and variance of Z_{S_m}. The terms of these values are functions of m, N, σ_w, and the confidence probability, except one term which is a function of the unmodeled coefficients, $1/N\|G_{S_m}B_{S_m}\Delta_{S_m}\|_2^2$.

The structure of the available data error x_{S_m} is such that it can be used to provide probabilistic bounds on $1/N\|G_{S_m}B_{S_m}\Delta_{S_m}\|_2^2$. In Appendix B it is shown that the data error is a sample of a Chi-square random variable X_{S_m}:

$$\frac{N}{\sigma_w^2} X_{S_m} \sim \chi_{N-m}^2 \tag{40}$$

where χ_{N-m}^2 is a Chi-square random variable of order $N-m$. The random variable X_{S_m} has the following expected value and variance:

$$E(X_{S_m}) = \left(1 - \frac{m}{N}\right)\sigma_w^2 + \frac{1}{N}\|G_{S_m}B_{S_m}\Delta_{S_m}\|_2^2 \tag{41}$$

$$\begin{aligned}\operatorname{var}(X_{S_m}) = &\frac{2}{N}\left(1 - \frac{m}{N}\right)(\sigma_w^2)^2 \\ &+ \frac{4\sigma_w^2}{N^2}\|G_{S_m}B_{S_m}\Delta_{S_m}\|_2^2\end{aligned} \tag{42}$$

where G_{S_m} and C_{S_m} are the projection matrices defined in (26) and (27) (shown in Appendix B).

Given the noisy data x_{S_m}, one sample of the random variable X_{S_m} is available. The variance of this random variable is of order $1/N\mathrm{th}$ of its expected value. If the data length is long enough, the variance of this random variable is close to zero. In this case, one method of estimating $\|G_{S_m}B_{S_m}\Delta_{S_m}\|_2^2$ is to assume that the available sample x_{S_m} is a good estimate of its expected value in (41). Therefore, by assuming that $E(X_{S_m}) \approx x_{s_m}$, we have

$$\frac{1}{N}\|G_{S_m}B_{S_m}\Delta_{S_m}\|_2^2 \approx x_{S_m} - \left(1 - \frac{m}{N}\right)\sigma_w^2. \tag{43}$$

However, for a finite-length data, the validity of this estimation depends on the exact behavior of the variance of X_{S_m} which is completely ignored in this estimation. In each S_m, as shown in (42), X_{S_m} has a variance which is a function of m, N and $1/N\|G_{S_m}B_{S_m}\Delta_{S_m}\|_2^2$. Therefore, the confidence in the estimate (43) of different subspaces is different. To be able to compare estimates of the unmodeled coefficients effects in subspaces of different orders, it is important that all the estimates are equally valid. The following lemma provides probabilistic bounds on the unavailable unmodeled coefficients effects. The probabilistic bounds include the effects of both the mean and the variance of X_{S_m} and are valid with the *same* p_2 probability.

Lemma 2: Using the observed data, with validation probability p_2, we have

$$\begin{aligned}L_{S_m}(x^N, y^N, \sigma_w, p_2) \le &\frac{1}{N}\|G_{S_m}B_{S_m}\Delta_{S_m}\|_2^2 \\ &\le U_{S_m}(x^N, y^N, \sigma_w, p_2)\end{aligned} \tag{44}$$

where the lower and upper bounds are functions of only the observed data (x^N, y^N) and the validation probability.

Proof: In Appendix C, calculation of the bounds is provided. □

Theorem 1: With validation probability p_2 and confidence probability p_1 the reconstruction error is bounded as follows:

$$\underline{z_{S_m}(p_1, x^N, y^N, p_2)} \le z_{S_m} \le \overline{z_{S_m}(p_1, x^N, y^N, p_2)} \tag{45}$$

where

$$\begin{aligned}\overline{z_{S_m}(p_1, x^N, y^N, p_2)} = &U_{S_m}(x^N, y^N, p_2) + \frac{m}{N}\sigma_w^2 \\ &+ D_{S_m}(p_1, \sigma_w, N)\end{aligned} \tag{46}$$

$$\begin{aligned}\underline{z_{S_m}(p_1, x^N, y^N, p_2)} = &\min\{0, L_{S_m}(x^N, y^N, p_2) \\ &+ \frac{m}{N}\sigma_w^2 - D_{S_m}(p_1, \sigma_w, N)\}.\end{aligned} \tag{47}$$

The bounds L_{S_m} and U_{S_m} on $1/N\|G_{S_m}B_{S_m}\Delta_{S_m}\|_2^2$ are provided in Lemma 2 and D_{S_m} in (35) is provided in Section V-C. The bounds on the reconstruction error are only functions of the Chi-square CDF table, σ_w^2, m, N, probabilities p_1 and p_2, and the observed data y^N.

Proof: Probabilistic bounds on the reconstruction error with confidence probability p_1 and by using the mean and variance of Z_{S_m} are provided in (36). Lemma 1 provides the values of the mean and variance of Z_{S_m}. While the variance is only a function of the noise variance, subspace order, and the data length, the expected value is a function of the unavailable unmodeled coefficients. By using the observed data, Lemma 2 provides probabilistic bounds on the unmodeled coefficients term of the expected value with validation probability p_2. □

Authorized licensed use limited to: MIT Libraries. Downloaded on February 25,2010 at 18:13:19 EST from IEEE Xplore. Restrictions apply.

A. Bounds on the Parameter Error

Using the relation between the reconstruction and parameter errors in (24) and the bounds in (47), (46) the probabilistic bounds on the parameter error are

$$\frac{1}{\sigma_{\max}\left(\frac{A_{S_M}^T A_{S_M}}{N}\right)}\underline{z_{S_m}(p_1, x^N, y^N, p_2)} \le \frac{j_{S_m}}{N}$$
$$\le \frac{1}{\sigma_{\min}\left(\frac{A_{S_M}^T A_{S_M}}{N}\right)}\overline{z_{S_m}(p_1, x^N, y^N, p_2)}. \quad (48)$$

VII. Connection Between Order Selection Approaches and the Reconstruction Error

Order selection approaches are concerned with a similar question that which subspace and its estimate can best represent the true parameter. A wide class of these methods use the available data error in (21), which is a decreasing function of m, as a part of their criterion in form of [1], [8], [17], [21]

$$k_{S_m} = x_{S_m} + f(m, N, \sigma_w). \quad (49)$$

The extra penalty term in these approaches has been provided by calculating or estimating a particular criterion. The penalty term is chosen such that it is an increasing function of m. We can connect this form of a criterion with an estimate of z_{S_m} and show that a class of these approaches behave exactly the same as a special case of our method. For example, if we set p_2 to zero in Lemma 2, it is as if we estimate the expected value of X_{S_m} with the one available sample x_{S_m}, $E(X_{S_m}) \approx x_{S_m}$. Using this assumption the estimate of the unmodeled coefficients effects is (43). Note that as it was described in Section VI, since in this estimation the variance of X_{S_m} is ignored, the confidence in this estimate in subspaces of different orders is different. Nevertheless, using this estimate of $1/N\|G_{S_m}B_{S_m}\Delta_{S_m}\|_2^2$ in (38) we have

$$\overline{z_{S_m}(p_1, x^N, y^N, 0)} = x_{S_m} + \frac{2m}{N}\sigma_w^2 + D_{S_m} - \sigma_w^2. \quad (50)$$

The last term is a constant and can be ignored in the comparison of the subspaces. However, the second and third terms both serve as penalty terms and are increasing functions of m. Particular choices of p_1 as a function of m and N leads to particular values of D_{S_m} and can cover a large class of criteria in the form of (49). For example, setting p_1 also to zero causes D_{S_m} to be zero, and the upper and lower bounds of z_{S_m} in (36) merge together

$$\underline{z_{S_m}(0, x^N, y^N, 0)} = \overline{z_{S_m}(0, x^N, y^N, 0)} \quad (51)$$
$$= x_{S_m} + \frac{2m}{N}\sigma_w^2 - \sigma_w^2 \quad (52)$$

and therefore, an estimate of the reconstruction error is

$$z_{S_m} \approx x_{S_m} + \frac{2m}{N}\sigma_w^2 - \sigma_w^2 \quad (53)$$

which is the criterion used in Akaike information criterion (AIC). In the next section we show how connecting these criteria with the z_{S_m} estimate can evaluate the behavior of these methods, including their consistency or inconsistency, from a new perspective.

VIII. Gaussian Distribution Estimation

If m is large enough, we can estimate the Chi-square distribution of Z_{S_m} with a Gaussian distribution. In this case, there is no need to check the Chi-square CDF table and with $p_1 = Q(\beta)$, the probabilistic event in (35) can be written in the form

$$Pr\{|Z_{S_m} - E(Z_{S_m})| \le \beta\sqrt{\mathrm{var} Z_{S_m}}\} = Q(\beta) \quad (54)$$

where $Q(\beta) = \int_{-\beta}^{\beta} 1/\sqrt{2\pi}e^{-x^2/2}dx$. Estimating this random variable with a Gaussian implies that D_{S_m} in (47) and (46) is simply

$$D_{S_m}(Q(\beta), \sigma_w, N) = \beta\sqrt{\mathrm{var} Z_{S_m}} = \beta\frac{\sqrt{2m}}{N}\sigma_w^2. \quad (55)$$

Note that in statistical approaches, for m as small as 10, a Chi-square distribution is well estimated with a Gaussian distribution.

On the other hand, if $N - m$ is large enough, we can estimate the Chi-square distribution of X_{S_m} with a Gaussian distribution. In this case, there is no need to use the Chi-square CDF table to calculate the bounds of the unmodeled coefficients effect in Lemma 2. The values of these bounds with validation probability $Q(\alpha)$ are provided in Appendix D.

As a result, for a range of values of m that both m and $N - m$ are large enough we have (56), shown at the bottom of the page, where the values of $L_{S_m}(x^N, y^N, Q(\alpha))$ and $U_{S_m}(x^N, y^N, Q(\alpha))$ are provided in Appendix D.

$$\overline{z_{S_m}(Q(\beta), x^N, y^N, Q(\alpha))} = U_{S_m}(x^N, y^N, Q(\alpha)) + \frac{m}{N}\sigma_w^2 + \beta\frac{\sqrt{2m}\sigma_w^2}{N}$$
$$\underline{z_{S_m}(Q(\beta), x^N, y^N, Q(\alpha))} = \max\left\{0, L_{S_m}(x^N, y^N, Q(\alpha)) + \frac{m}{N}\sigma_w^2 - \beta\frac{\sqrt{2m}\sigma_w^2}{N}\right\} \quad (56)$$

Authorized licensed use limited to: MIT Libraries. Downloaded on February 25,2010 at 18:13:19 EST from IEEE Xplore. Restrictions apply.

A. *Proper Choices of Validation Probability p_2 and Confidence Probability p_1*

The higher the probabilities p_1 and p_2, the greater is the confidence on the provided bounds. Therefore, it is important to choose the two probabilities close to one. On the other hand, the gap between the upper and lower bounds becomes larger as the probabilities approach one. To observe the behavior of the bounds as a function of the two probabilities, we can study the bounds with Gaussian estimation in (57) and (56). In this case, parameters α and β can be chosen large enough such that $p_2 = Q(\alpha)$ and $p_1 = Q(\beta)$ are close to one

$$\lim_{N \longrightarrow \infty} \alpha_N = \infty, \quad \lim_{N \longrightarrow \infty} \beta_N = \infty. \tag{57}$$

Moreover, the gap between the upper and lower bounds becomes smaller as $\alpha_N/\sqrt{N}$ and $\beta_N/\sqrt{N}$ become smaller. To guarantee tight bounds on z_{S_m} that converge to each other asymptotically, the following condition is also necessary

$$\lim_{N \longrightarrow \infty} \frac{\alpha_N}{\sqrt{N}} = 0, \quad \lim_{N \longrightarrow \infty} \frac{\beta_N}{N} = 0. \tag{58}$$

Under this condition, the upper and lower bounds approach each other as N grows and we have

$$\lim_{N \longrightarrow \infty} \underline{z_{S_m}(Q(\beta_N), x^N, y^N, Q(\alpha_N))} = \lim_{N \longrightarrow \infty} \overline{z_{S_m}(Q(\beta_N), x^N, y^N, Q(\alpha_N))}. \tag{59}$$

B. *Consistency*

Consider the case that the true parameter has order M^* $(M^* \leq M)$. In this case[4]:

Lemma 3: If the input is such that $\sigma_{\max}(A_{S_M}^T A_{S_M}/N)$ and $\sigma_{\min}(A_{S_M}^T A_{S_M}/N)$ are finite values as N grows, then we have

$$\lim_{N \longrightarrow \infty} \hat{\theta}_{S_m}(y^N) = \theta^* \quad M \geq m \geq M^*. \tag{60}$$

The convergence is in mean-square sense.

Proof: In Appendix E. □

On the other hand, when the conditions in (57) and (58) are satisfied, $E(Z_{S_m})$ is squeezed between the upperbound and lowerbound of z_{S_m} with probabilities that go to one and at the same time the bounds approach each other. So the bounds converge to $E(Z_{S_m})$ in the limit. Asymptotic behavior of $E(Z_{S_m})$ is discussed in Section V-B. As was shown in that section, since there is a nonzero unmodeled coefficients effect for subspaces with order $m < M^*$, $E(Z_{S_m})$ cannot be zero for these subspaces in the limit. However, as is shown in (31), $E(Z_{S_m})$ is zero in the limit for all subspaces that include the true parameter, $m \geq M^*$. As a result, the smallest value of m for which the asymptotic values of the bounds converge to zero is the true parameter M^*. Therefore, not only the estimate in the subspace of order M^* approaches the true parameter, but also we choose the correct order by comparing the provided probabilistic upperbound of z_{S_m} which proves the consistency of the method.

C. *A Note on the Noise Variance*

All the order selection approaches are functions of the noise variance. There are different approaches to estimate the noise variance for the cases where it is unknown. If it is possible to access the output for when the system is at rest, the most popular variance estimation approach is the median approach. In this method the standard deviation estimate is $\hat{\sigma}_w = \mathrm{MAD}/0.6745$ where MAD is the median of absolute value of the gathered noise. If collecting noise only data is not possible, the proposed method in [3], to our knowledge, is the only method proposed for such a scenario and can be used to simultaneously choose the optimum order and estimate the noise variance. Starting with a range of possible values for the noise variance, the method suggests estimating the reconstruction error for this range of noise variances in the competing subspaces. The noise variance and the subspace for which the probabilistic upperbound is at its minimum will be chosen.

IX. Prior Information on the Order of the True Parameter

In order to calculate the least square subset estimate $\hat{\theta}_{S_m}$, it is only sufficient for m to be less than or equal to the data length N. Otherwise, if we choose a subset with more parameters than the data length, the least square estimate cannot be calculated. Therefore, regardless of the order of the true parameter, the highest possible order for comparison of the least square estimates is $M = N$. This limitation is only due to the finiteness of the data and is independent from the true order of the impulse response.[5]

Here we discuss the behavior of the method when no upperbound on the true order is available (i.e., the upperbound is unknown and can be infinity). Therefore, the data length may be smaller than the unknown true order. In this case M is no longer the upperbound of the true impulse response length. It is only a chosen highest order for the competing subspaces such that $M \leq N$. The competing nested subspaces of S_M include parameter estimates of different orders. The goal is to compare these subspaces and choose the least noisy estimate. In this setting, θ^* is not in S_M and we have

$$\theta^* = \begin{bmatrix} \theta_{S_M}^* \\ \Delta \end{bmatrix} \tag{61}$$

where $\theta_{S_M}^* = \theta^*[0 : M-1]$ and $\Delta = \theta^*[M : \infty)$ captures the coefficients of θ^* outside of S_M. The parameter error can be rewritten as

$$j_{S_m} = \|\theta_{S_M} - \hat{\theta}_{S_m}(y^N)\|_2^2 + \|\Delta\|_2^2. \tag{62}$$

The method proposed in this paper focuses on the first term of this error and is able to provide bounds on this term. On the other hand, the second term $\|\Delta\|_2^2$ is a constant term for all the competing subspaces[6] and can be ignored in comparing the subspaces. Therefore, our method is able to compare the parameter error of the competing subspace regardless of the true order, and provide the least noisy estimate among the competing estimates.

[4] In a given model set, consistency is to guarantee that the coefficient estimates converge to the true coefficients in that model set. This form of consistency for impulse response estimation, through state space representation, for occasions when the first m taps$(m < N)$ is modeled is discussed in [16]

[5] Note that our parameters in the least-square estimation are the impulse response coefficients and not coefficients of a transfer function.

[6] There are different deterministic approaches that deal with this type of unmodeled dynamic. They either consider the noise to belong to a deterministic set or use a known bound on the unmodeled coefficients [5], [6], [15], [22]–[24].

Authorized licensed use limited to: MIT Libraries. Downloaded on February 25,2010 at 18:13:19 EST from IEEE Xplore. Restrictions apply.

How does this approach behave as the data length grows? If the true parameter has an unknown finite order M^*, then there exists a data length for which $N \geq M^*$. We can choose M, the highest order of the competing subspaces, as an increasing function of N $(M(N))$. In this case, as N grows, there exists a data length such that $M(N) \geq M^*$. As a result, we are back to the setting that M is an upperbound of the true order and the consistency argument holds. On the other hand, if the true parameter has an infinite order, as the data length grows, the length of the least noisy estimate keeps growing. In this case, the provided least noisy estimate can be utilized for modeling a rational transfer function in a zero-pole modeling process [14], [18].

X. Modeling and Order Selection for LTI Systems in Application

The following summarizes the steps of the proposed model selection process.

- Competing subsets are $S_1, S_2, \ldots, S_M$. Each S_m represents impulse responses with m taps in the form of (8). In each S_m, the estimate of the impulse response, $\hat{\theta}_{S_m}(y^N)$, in (11) is calculated.
- Validation and confidence probabilities p_1 and p_2 are chosen. In each S_m, the probabilistic bounds on the reconstruction error and the parameter error are calculated [(46) and (47) or (56) and (57)].
- By comparing the probabilistic worst case, the optimum subspace is chosen

$$S_{m^*} = \arg\min_{S_m} \overline{z_{S_m}(Q(\beta), x^N, y^N, Q(\alpha))}. \tag{63}$$

The optimum order is m^* and the least noisy impulse response estimate is $\hat{\theta}_{S_{m^*}}(y^N)$.

The choice of validation and confidence probabilities is discussed in Section VIII-A. Here, we provide the results for the following three choices of these probabilities $p_1 = Q(\alpha)$ and $p_2 = Q(\beta)$:

Case 1) Set $\alpha = \beta = 4$ which is such that the probabilities are large, $Q(\alpha) = Q(\beta) = 0.999934$.

Case 2) Set the probabilities to zero, equivalently $\alpha = \beta = 0$. In this case the upper and lower bounds in (48) are merged as discussed in Section VII and the estimate of the reconstruction error in (53) is the same as AIC [1].

Case 3) Set $\alpha = 0$ and $\beta = \sqrt{m}\log(N)$. In this case the second probability is a function of the subspace order.[7]

A. FIR Models

Consider a class of models in (1) with the following structure:

$$y[n] = \sum_{i=0}^{M} a_i^* u[n-i] + w[n] \tag{64}$$

where the FIR filter has length 30 and

$$a_i^* = .3(.5)^{i-1} + 3(i-1)(.8)^{i-1}, \quad 0 \leq i \leq 30. \tag{65}$$

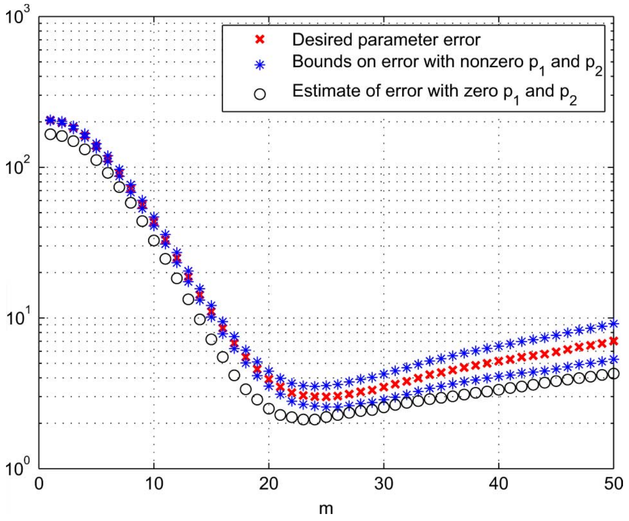

Fig. 3. Parameter error as a function of m, for data of length $N = 200$ when SNR $= 10$ dB; bounds for $p_1 = Q(4)$ and $p_2 = Q(4)$; and bounds for $p_1 = p_2 = 0$ (where the upperbound and lowerbound collapse into one).

In this simulation, the input is an independent identically distributed (i.i.d.) Bernoulli sequence of ±1, the length of data is $N = 200$, and the SNR is 10 dB. Fig. 3 shows the desired parameter error as a function of subspace order m. As the figure shows, estimating more than 23 taps of the impulse response increases the error as it starts fitting more of the noise effects. The figure shows the bounds provided by the observed data for case 1 (where $\alpha = \beta = 4$) and case 2 (where $\alpha = \beta = 0$). As the figure shows, the nonzero probabilities provide tight bounds in this case which are valid with the high confidence and validation probabilities of $Q(4) \approx 1$ (condition in (57) is satisfied). The tightness of the bounds could have been predicted because $Q(4)/\sqrt{200}$ is small enough and the conditions in (58) are almost satisfied.

The mean and variance of the optimum m^* for each of the three choices of p_1 and p_2 are provided in Figs. 4 and 5. The figures show the result for the data length of $\mathrm{N} = 60$ and $\mathrm{N} = 200$, and as a function of data SNR for ranges from 0 to 40 dB. The results are provided by averaging 100 trials. The two figures confirm some interesting facts. As the figures show, the higher the SNR, the more likely is the choice of the correct order. Both higher SNR and larger data length guarantee convergence to the true impulse response length. As these figures show, when the data length is 60, the true impulse response length is chosen after SNR of 33, and when the data length is 200, it starts choosing the correct order at a lower SNR of 25. This confirms that the asymptotic convergence to the true parameter is a function of both the SNR and the data length. Fig. 4 shows that, for example, with SNR $= 10$ dB and data length of $N = 60$, on average case 1 chooses $m^* = 17$ with variance of order 6. However, with the same SNR, if the data length is increased to $N = 200$, as shown in Fig. 5, the method on average chooses $m^* = 23$ with a smaller variance of 4. For a given SNR, as the data length increases, the variance of m^* becomes smaller and its mean becomes larger. Therefore, in case 1 since the confidence and validation probabilities are close to one, with longer data of length 200, we achieve the correct choice of the impulse

[7] With this choice of probabilities the upper bound $\overline{z_{S_m}(Q(\beta), x^N, y^N, 0)}$ criterion is identical to Bayesian information criterion (BIC) [17] and two-stage MDL [2] .

Authorized licensed use limited to: MIT Libraries. Downloaded on February 25,2010 at 18:13:19 EST from IEEE Xplore. Restrictions apply.

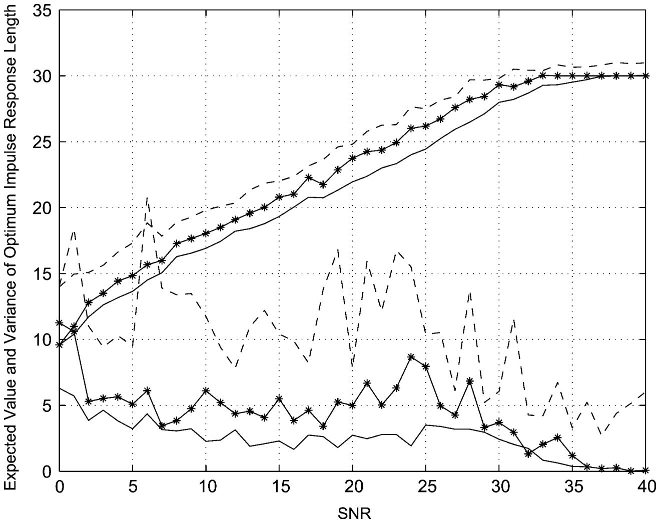

Fig. 4. FIR Filter: Expected values and variance of optimum impulse response length m^* with data length $N = 60$ for SNR between 0 and 40 dB. Solid line with star "*" is the results of case 1; dashed line "- -" is the results of case 2; solid line is the results of case 3. Three top lines are the expected values, and lower lines are the variances.

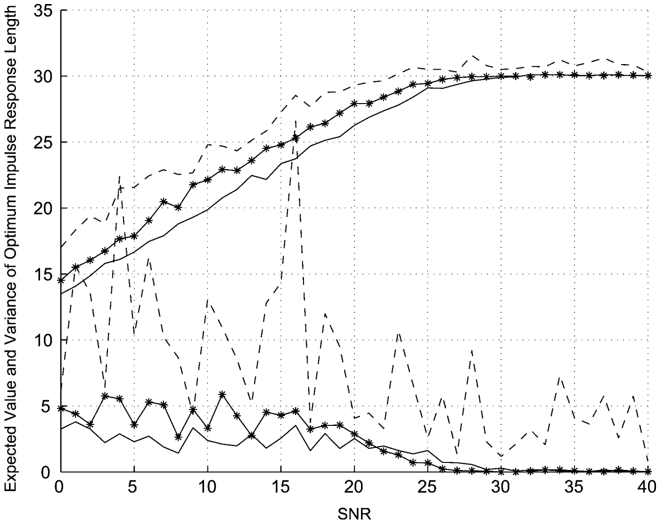

Fig. 5. Fig. 4 repeated for increased data length from $N = 60$ to $N = 200$.

response length of 30 at a lower SNR of 25 dB (when $N = 60$, this SNR is higher and at 34 dB).

On the other hand, the proper choice of probabilities in case 1 results in a better choice of order selection, and the method reaches the true impulse response length, 30, faster than the other two cases as the SNR grows. Case 2 has a tendency to over model as the SNR grows, and it has the highest uncertainty and variance in the choice of the optimum subspace compared to the other cases. This confirms the inconsistency of AIC as the variance of this order selection method is significant even for $m > 30$. Case 3 performs much the same as case 1. However, compared to case 1, it has a tendency of under modeling for lower SNRs.

The proposed method provides the least noisy estimate among the competing estimates. Nevertheless, in applications that there is a penalty for the length of the chosen parameter, we can also choose orders less than the optimum m^* depending on our error tolerance [11]. For example in the case presented in Fig. 3, the optimum order that minimizes the criterion in case 1 is $m^* = 24$ with error bounded between 2.5 and 4. On the other hand, if in this case we can tolerate parameter errors up to 10, then, as the figure shows, we can reduce the order to as low as $m = 15$. This capability of our quality evaluation method is an important strength of the proposed approach that is missing in the existing model selection approaches.

It is important to mention that the figures illustrate that even if it was known *a priori* that the length of the true FIR is 30, the least noisy estimate may still have length of less than 30. For example, as Fig. 4 shows, this estimate has a length less than 30 for the data length of 60 and the SNR is less than 33.

B. IIR Models

Consider a class of models in (1) with the following structure:

$$y[n] = \sum_{i=0}^{\infty} a_i^* u[n-i] + w[n]. \tag{66}$$

The true order of this system is not finite and the true order is larger than the data length. This case was discussed in Section IX. In this case, the proposed method can provide the least noisy estimate among the possible least square impulse response estimates with length one to the largest possible length of N.

In the following simulation, the impulse response is

$$a_i^* = .3(.5)^{i-1} + 3(i-1)(.8)^{i-1}, \quad 0 \le i. \tag{67}$$

This is a stable system with two poles.

The mean and variance of the optimum m^* for each of the three choices of p_1 and p_2 are provided in Fig. 6. As the figure shows, case 2 still has the highest variance due to setting both probabilities to zero. Since the true order is not finite, with confidence probabilities close to one, the method chooses larger and larger orders, as the data length grows.

XI. Conclusion

We studied the parameter error resulting in least-square impulse response estimation from noisy data. It was shown that the additive noise affects both mean and variance of the error and the tradeoff between the noise fitting and the unmodeled coefficients points to an optimum model set for the estimate of the true unknown impulse response. We derived probabilistically-validated bounds on this error that incorporate the effects of both mean and variance of the estimates. Such bounds were achieved by using only the observed finite data and are expressed as a function of a probabilistic confidence. The bounds were proposed as the basis for model quality assessment of an estimate as well as for model selection among competing model sets through the minimization of the upper bound with a fixed level of confidence. The optimum estimate is the least noisy one among the competing estimates. Moreover, if an acceptable error range is satisfied with an estimate in a subset of the optimum model set, as it was illustrated through simulation results, the estimate with a smaller length than the optimum one can be chosen and the procedure will provide a more parsimonious model estimate. We also demonstrated that the criteria

Authorized licensed use limited to: MIT Libraries. Downloaded on February 25,2010 at 18:13:19 EST from IEEE Xplore. Restrictions apply.

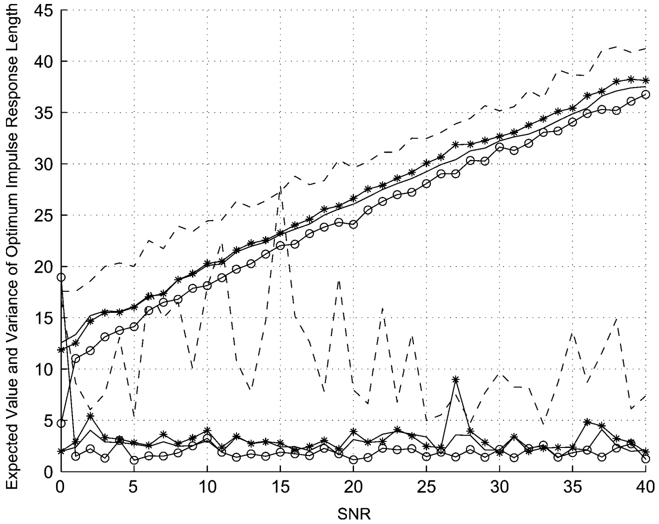

Fig. 6. IIR Filter: Expected values and variance of optimum order m^* with data length $N = 200$ for SNR between 0 and 40 dB. Solid line with star "*" is the results of case 1; dashed line "- -" is the results of case 2; solid line with "o" is the results of case 3; solid line is the results for $\alpha = \beta = 8$. Four top lines are the expected values and lower lines are the variances.

used in several model selection approaches coincide with the provided bounds for particular choices of the confidence level. For example, AIC is a special case of the upperbound where the confidence and validation probabilities are set to zero. This can easily explain the inconsistency of AIC. It was illustrated that the proper choice of confidence and validation probabilities guarantees the consistency of the method as the data length grows. Furthermore, we showed the consistency of the method as a function of the SNR.

Appendix A
Proof of Lemma 1

From (25), we have

$$z_{S_m} = \frac{1}{N}\|G_{S_m}B_{S_m}\Delta_{S_m} + C_{S_m}w^N\|_2^2. \tag{68}$$

Since the projection matrices G_{S_m} in (26) and C_{S_m} in (27) are orthogonal, the inner product of the two vectors $G_{S_m}B_{S_m}\Delta_{S_m}$ and $C_{S_m}w^N$ is zero and therefore

$$z_{S_m} = \frac{1}{N}\|G_{S_m}B_{S_m}\Delta_{S_m}\|_2^2 + \frac{1}{N}\|C_{S_m}w^N\|_2^2. \tag{69}$$

While the first term is a constant term which will be part of the mean of Z_{S_m}, the second term is a Chi-square random variable of order m, that is the order of the projection matrix C_{S_m}. To elaborate this fact, consider F_{S_m} an $N \times m$ unitary matrix that spans the subspace generated by A_{S_m}. We have

$$A_{S_m} = F_{S_m}H_{A_{S_m}}, \quad F_{S_m}^T F_{S_m} = I_{m\times m} \tag{70}$$

where $H_{A_{S_m}}$ is the $m \times m$ full rank matrix that generates A_{S_m} by the unitary matrix. In this case, the projection matrix is

$$C_{S_m} = A_{S_m}(A_{S_m}^T A_{S_m})^{-1}A_{S_m}^T \tag{71}$$

$$= F_{S_m}H_{A_{S_m}}(H_{A_{S_m}}^T F_{S_m}^T F_{S_m} H_{A_{S_m}})^{-1} \times H_{A_{S_m}}^T F_{S_m}^T \tag{72}$$

$$= F_{S_m}H_{A_{S_m}}H_{A_{S_m}}^{-1}(H_{A_{S_m}}^T)^{-1}H_{A_{S_m}}^T F_{S_m}^T \tag{73}$$

$$= F_{S_m}F_{S_m}^T. \tag{74}$$

Therefore,

$$\frac{1}{N}\|C_{S_m}w^N\|_2^2 = \frac{1}{N}\|F_{S_m}F_{S_m}^T w^N\|_2^2 \tag{75}$$

$$= \frac{1}{N}(w^N)^T F_{S_m}F_{S_m}^T F_{S_m}F_{S_m}^T w^N \tag{76}$$

$$= \frac{1}{N}(w^N)^T F_{S_m}F_{S_m}^T w^N \tag{77}$$

$$= \left\|\frac{1}{\sqrt{N}}F_{S_m}^T w^N\right\|_2^2. \tag{78}$$

This is a sample of a Chi-square random variable that is generated by m independent Gaussian random variable u_is with zero mean and variance σ_w/N

$$\frac{1}{\sqrt{N}}F_{S_m}^T w^N = \begin{bmatrix} u_1 \\ \vdots \\ u_m \end{bmatrix}. \tag{79}$$

This completes the proof of Lemma 1.

Appendix B
Structure of X_{S_m}

Similar to the argument for the projection matrix C_{S_m} in proof of Lemma 1, there exists R_{S_m} a full rank $N \times (N-m)$ unitary matrix that spans the column space of the projection matrix G_{S_m} and just as what is shown in (74), we have

$$G_{S_m} = R_{S_m}R_{S_m}^T, \quad R_{S_m}^T R_{S_m} = I_{(N-m)\times(N-m)}. \tag{80}$$

Therefore, from (21) we have

$$x_{S_m} = \frac{1}{N}\|G_{S_m}B_{S_m}\Delta_{S_m} + G_{S_m}w^N\|_2^2 \tag{81}$$

$$= \frac{1}{N}\|G_{S_m}(B_{S_m}\Delta_{S_m} + w^N)\|_2^2 \tag{82}$$

$$= \frac{1}{N}\|R_{S_m}R_{S_m}^T(B_{S_m}\Delta_{S_m} + w^N))\|_2^2 \tag{83}$$

$$= \frac{1}{N}\|R_{S_m}^T(B_{S_m}\Delta_{S_m} + w^N))\|_2^2. \tag{84}$$

Similar to what is shown in (78), (84) is obtained from (83) since for any unitary matrix such as R_{S_m} and any matrix A we have $\|R_{S_m}A\|_2^2 = \|A\|_2^2$. Also, we have

$$\frac{1}{\sqrt{N}}R_{S_m}^T w^N = \begin{bmatrix} v_1 \\ \vdots \\ v_{N-m} \end{bmatrix}. \tag{85}$$

Each element v_i has a zero mean with variance σ_w^2/N and v_is are independent. Therefore, we have

$$x_{S_m} = \sum_{i=1}^{N-m}(\delta_i + v_i)^2 \tag{86}$$

Authorized licensed use limited to: MIT Libraries. Downloaded on February 25,2010 at 18:13:19 EST from IEEE Xplore. Restrictions apply.

where the sum of means of the Chi-square random variable is

$$\sum_{i=1}^{N-m} \delta_i^2 = \frac{1}{N}\|G_{S_m} B_{S_m} \Delta_{S_m}\|_2^2. \tag{87}$$

Consequently, the data error is a Chi-square random variable of order $N-m$ and the expected value and variance of the data error are [13]

$$E(X_{S_m}) = (N-m)\text{var}(v_i) + \sum_{i=1}^{N-m} \delta_i^2 \tag{88}$$

$$\text{var}(X_{S_m}) = 2\text{var}(v_i)\left((N-m)\text{var}(v_i) + 2\sum_{i=1}^{N-m}\delta_i^2\right). \tag{89}$$

By using (87) and $\text{var}(v_i) = \sigma_w^2/N$, the mean and variance of this random variable are (41) and (42).

Appendix C
Proof of Lemma 2

The following method provides bounds on the desired $1/N\|G_{S_m} B_{S_m} \Delta_{S_m}\|_2^2$ through probabilistic validation. Validation of $\|G_{S_m} B_{S_m} \Delta_{S_m}\|_2^2$ is such that x_{S_m} is in the neighborhood of its mean with probability p_2. The procedure is as follows: Given a validation probability p_2, for each possible value of $\|G_{S_m} B_{S_m} \Delta_{S_m}\|_2^2$ we have

$$\Pr\{|X_{S_m} - E(X_{S_m})| \leq J_{S_m}\} = p_2 \tag{90}$$

where the bound J_{S_m} is a function of $\|G_{S_m} B_{S_m} \Delta_{S_m}\|_2^2, \sigma_w^2, m,$ and p_2. The value of J_{S_m} is calculated by using the Chi-square CDF table and therefore for any given $\|G_{S_m} B_{S_m} \Delta_{S_m}\|_2^2$ with probability p_2 samples of the random variable X_{S_m} are bounded as follows:

$$E(X_{S_m}) - J_{S_m} \leq x_{S_m} \leq E(X_{S_m}) + J_{S_m}. \tag{91}$$

Therefore, for each possible $\|G_{S_m} B_{S_m} \Delta_{S_m}\|_2^2$, samples of X_{S_m} are bounded between $E(X_{S_m}) \pm J_{S_m}$ with probability p_2, where both $E(X_{S_m})$ and J_{S_m} are functions of the unmodeled coefficients. Given the available sample x_{S_m}, we validate those $\|G_{S_m} B_{S_m} \Delta_{S_m}\|_2^2$ for which x_{S_m} is inside the bounds $E(X_{S_m}) \pm J_{S_m}$. Due to the Chi-square structure of X_{S_m}, this validation provides $U_{S_m}(x^N, y^N, p_2)$ and $L_{S_m}(x^N, y^N, p_2)$, upper and lower bounds on $1/N\|G_{S_m} B_{S_m} \Delta_{S_m}\|_2^2$ as a function of y^N and p_2 in (44).

Note that setting p_2 in (90) to zero is the same as ignoring the variance of X_{S_m}. In this case $J_{S_m} = 0$ for all the subspaces and instead of the probabilistic bounds on $\|G_{S_m} B_{S_m} \Delta_{S_m}\|_2^2$, we have the estimate in (43).

Appendix D
Bounds on $1/N\|G_{S_m} B_{S_m} \Delta_{S_m}\|_2^2$ With Gaussian Distribution Estimation

Using a Gaussian distribution estimate for the Chi-Square distribution of X_{S_m}, the J_{S_m} in (90) is $\alpha\sqrt{\text{var}(X_{S_m})}$ where p_2 is $Q(\alpha)$ and the variance of X_{S_m} is provided in (42). Therefore, the validation step is simply validating $1/N\|G_{S_m} B_{S_m} \Delta_{S_m}\|_2^2$ for which the observed x_{S_m} is bounded with

$$E(X_{S_m}) - \alpha\sqrt{\text{var}(X_{S_m})} \leq x_{S_m} \leq E(X_{S_m}) + \alpha\sqrt{\text{var}(X_{S_m})} \tag{92}$$

where both the expected value and variance of X_{S_m} are provided in (41) and (42).

Using (92), to find the upperbound for $1/N\|G_{S_m} B_{S_m} \Delta_{S_m}\|_2^2$ we should solve the following inequality:

$$E(X_{S_m}) - \alpha\sqrt{\text{var}(X_{S_m})} \leq x_{S_m}. \tag{93}$$

This inequality provides the upperbound as long as p_2, or equivalently α, has been chosen sufficiently large such that

$$\alpha \geq \frac{N}{\sqrt{2(N-m)}}\left(1 - \frac{m}{N} - \frac{x_{S_m}}{\sigma_w^2}\right). \tag{94}$$

In this case, with validation probability $Q(\alpha)$, the upperbound is

$$U_{S_m}(x^N, y^N, Q(\alpha)) = x_{S_m} - m_w + \frac{2\alpha^2\sigma_w^2}{N} + K_{S_m}(\alpha) \tag{95}$$

and

$$K_{S_m}(\alpha) = 2\alpha\frac{\sigma_w}{\sqrt{N}}\sqrt{\frac{\alpha^2\sigma_w^2}{N} + x_{S_m} - \frac{1}{2}m_w} \tag{96}$$

$$m_w = \left(1 - \frac{m}{N}\right)\sigma_w^2. \tag{97}$$

Solving for the lower bound $L_{S_m}(x^N, y^N, Q(\alpha))$ using the following inequality:

$$x_{S_m} \leq E(X_{S_m}) + \alpha\sqrt{\text{var}(X_{S_m})} \tag{98}$$

the lower bound is zero if

$$(m_w - \alpha\sqrt{v_{S_m}}) \leq x_{S_m} \leq (m_w + \alpha\sqrt{v_{S_m}}) \tag{99}$$

where

$$v_{S_m} = \frac{2}{N}\left(1 - \frac{m}{N}\right)\sigma_w^4. \tag{100}$$

Otherwise, the lower bound is

$$L_{S_m}(x^N, y^N, Q(\alpha)) = x_{S_m} - m_w + \frac{2\alpha^2\sigma_w^2}{N} - K_{S_m}(\alpha). \tag{101}$$

Appendix E
Proof of Lemma 3

In Section V-B it was shown that as N grows, $E(Z_{S_m})$ converges to zero for all subspaces with $m \geq M^*$. At the same time, since both j_{S_m} and z_{S_m} are positive random variables from (24), we have

$$\frac{1}{\sigma_{\max}\left(\frac{A_{S_M}^T A_{S_M}}{N}\right)}E(Z_{S_m}) \leq E(J_{S_m}) \leq \frac{1}{\sigma_{\min}\left(\frac{A_{S_M}^T A_{S_M}}{N}\right)}E(Z_{S_m}). \tag{102}$$

Authorized licensed use limited to: MIT Libraries. Downloaded on February 25,2010 at 18:13:19 EST from IEEE Xplore. Restrictions apply.

If the input is such that both $\sigma_{\max}(A_{S_M}^T A_{S_M}/N)$ and $\sigma_{\min}(A_{S_M}^T A_{S_M}/N)$ are finite values in limit, then we have $\lim_{N\longrightarrow\infty} E(J_{S_m}) = 0$ in the subspaces that $E(Z_{S_m})$ is zero in the limit. Therefore, with the expected value of the parameter error $(\|\hat{\theta}_{S_m} - \theta_0\|^2)$ converging to zero, the estimate $\hat{\theta}_{S_m}$ converges in mean square to the true parameter.

References

[1] H. Akaike, "A new look at the statistical model identification," *IEEE Trans. Autom. Control*, vol. AC-19, no. 6, pp. 716–723, Dec. 1974.

[2] Y. Barron, J. Rissanen, and B. Yu, "The minimum description length principle in coding and modeling," *IEEE Trans. Inf. Theory*, vol. 44, no. 6, pp. 2743–2760, Oct. 1998.

[3] S. Beheshti and M. A. Dahleh, "A new information theoretic approach to signal denoising and best basis selection," *IEEE Trans. Signal Process.*, vol. 53, no. 10, pp. 3613–3624, Oct. 2005.

[4] S. Beheshti and M. A. Dahleh, "On model quality evaluation of stable LTI systems," in *Proc. 39th IEEE Conf. Decision Control*, 2000, pp. 2716–2721.

[5] A. Garulli, A. Vicino, and G. Zappa, "Conditional central algorithms for worst case set-membership identification and filtering," *IEEE Trans. Autom. Control*, vol. 45, no. 1, pp. 14–23, Jan. 2000.

[6] L. Giarre and M. Milanese, "Model quality evaluation in H_2 identification," *IEEE Trans. Autom. Control*, vol. 42, no. 5, pp. 691–698, May 1997.

[7] L. Gilbert and A. R. Barron, "Information theory and mixing least-squares regressions," *IEEE Trans. Inf. Theory*, vol. 52, no. 8, pp. 3396–3410, Aug. 2006.

[8] E. Hannan, "The determination of the order of an auto-regression," *J. Roy. Statist. Soc.*, pp. 190–195, 1979.

[9] A. O. Hero, J. A. Fessler, and M. Usman, "Exploring estimator bias-variance tradeoffs using the uniform CR bound," *IEEE Trans. Signal Process.*, vol. 44, no. 8, pp. 2026–2041, Aug. 1996.

[10] H. Krim, D. Tucker, S. Mallat, and D. Donoho, "On denoising and best signal representation," *IEEE Trans. Inf. Theory*, vol. 45, no. 7, pp. 2225–2238, Nov. 1999.

[11] A. P. Liavas, P. A. Regalia, and J. Delmas, "Blind channel approximation: Effective channel order estimation," *IEEE Trans. Signal Process.*, vol. 47, no. 12, pp. 3336–3344, Dec. 1999.

[12] L. Ljung*, System Identification: Theory for the User*. Englewood Cliffs, NJ: Prentice-Hall, 2007.

[13] J. K. Patel and C. B. Read*, Handbook of the Normal Distribution*. New York: Marcel Dekker, 1996.

[14] P. A. Regalia, "An unbiased equation error identifier and reduced-order approximations," *IEEE Trans. Signal Process.*, vol. 42, no. 6, pp. 1397–1412, Jun. 1994.

[15] W. Reinelt, A. Garulli, and L. Ljung, "Comparing different approaches to modeling error modeling in robust identification," *Automatica*, vol. 38, pp. 787–803, 2002.

[16] E. Reynders, R. Pintelon, and G. De Roeck, "Consistent impulse-response estimation and system realization from noisy data," *IEEE Trans. Signal Process.*, vol. 56, no. 7, pp. 2696–2705, Jul. 2008.

[17] G. Schwarz, "Estimating the dimension of a model," *Ann. Statist.*, vol. 6, pp. 461–464, 1978.

[18] H. C. So and Y. T. Chan, "Analysis of an LMS algorithm for unbiased impulse response estimation," *IEEE Trans. Signal Process.*, vol. 51, no. 7, pp. 2008–2013, Jul. 2003.

[19] T. Soderstrom and P. Stoica*, System Identification*. Englewood Cliffs, NJ: Prentice-Hall, 1989.

[20] M. Wax and T. Kailath, "Detection of signals by information theoretic criteria," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-33, no. 2, pp. 387–392, Apr. 1985.

[21] Y. Yang and A. R. Barron, "An asymptotic property of model selection criteria," *IEEE Trans. Inf. Theory*, vol. 44, no. 1, pp. 95–116, Jan. 1998.

[22] G. Zames, "On the metric complexity of causal linear systems ϵ -entropy and ϵ -dimension for continuous time," *IEEE Trans. Autom. Control*, vol. 24, no. 2, pp. 222–230, Apr. 1979.

[23] B. Ninness and G. C. Goodwin, "Estimation of model quality," *Automatica*, vol. 31, pp. 1771–1797, 1995.

[24] L. Y. Wang and G. G. Yin, "Persistent identification of systems with unmodeled dynamics and exogenous disturbances," *IEEE Trans. Autom. Control*, vol. 45, no. 7, pp. 1246–1256, Jul. 2000.

Soosan Beheshti (M'03–SM'06) received the B.S. degree from Isfahan University of Technology, Isfahan, Iran, and the M.S. and Ph.D. degrees from the Massachusetts Institute of Technology (MIT), Cambridge, in 1996 and 2002 respectively, all in electrical engineering.

From September 2002 to June 2005, she was a Postdoctoral Associate and a Lecturer at MIT. Since July 2005, she has been with the Department of Electrical and Computer Engineering, Ryerson University, Toronto, ON, Canada, where she is currently an Assistant Professor. Her research interests include signal and information processing, and system dynamics and modeling.

Munther A. Dahleh (S'84–M'87–SM'97–F'01) was born in 1962. He received the B.S. degree from Texas A & M university, College Station, in 1983 and the Ph.D. degree from Rice University, Houston, TX, in 1987, all in electrical engineering.

Since then, he has been with the Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology (MIT), Cambridge, where he is now a Full Professor. He is also currently the Associate Director of the Laboratory for Information and Decision Systems. He was a visiting Professor at the Department of Electrical Engineering, California Institute of Technology, Pasadena, in spring 1993. He has held consulting positions with several companies in the U.S. and abroad. He is interested in problems at the interface of robust control, filtering, information theory, and computation, which include control problems with communication constraints and distributed mobile agents with local decision capabilities.

Dr. Dahleh has been the recipient of the Ralph Budd award in 1987 for the best thesis at Rice University, the George Axelby outstanding paper award (for a paper coauthored with J. B. Pearson in 1987), an NSF presidential young investigator award (1991), the Finmeccanica career development chair (1992) and the Donald P. Eckman award from the American Control Council in 1993, the Graduate Students Council teaching award in 1995, the George Axelby outstanding paper award (for a paper coauthored with Bamieh and Paganini in 2004), and the Hugo Schuck Award for Theory (for a paper he coauthored with Martins). He was a plenary speaker at the 1994 American Control Conference, at the Mediterranean Conference on Control and Automation in 2003, and at the MTNS in 2006. He was an Associate Editor for the IEEE TRANSACTIONS ON AUTOMATIC CONTROL and for *Systems and Control Letters*.

Authorized licensed use limited to: MIT Libraries. Downloaded on February 25,2010 at 18:13:19 EST from IEEE Xplore. Restrictions apply.