Suboptimal Identification of Nonlinear ARMA Models Using an Orthogonality Approach

Ho-En Liao and William A. Sethares

Abstract—**This paper proposes a scheme based on orthogonal projection to identify a class of nonlinear auto-regressive, moving-average (NARMA) models. The scheme decouples the nonlinear and linear identification problems, and hence there are two steps. The first step extracts nonlinearities for each delay element within the model via conditional expectations. The second step evaluates dispersion functions to weight the nonlinear functions so that the cost is minimized. This paper focuses on the second step of the proposed scheme, the characteristics of the identification scheme are studied, and simulations are provided.**

I. Introduction

IN MANY applications where data is generated by a nonlinear mechanism, linear models are unacceptable and identification schemes fail. While the theory of identification of linear dynamic systems is well established, e.g., [1]–[3], the theory of identification of nonlinear dynamic systems is not yet satisfactory though much research [4],[5] has been conducted. This is mostly due to the fact that for general nonlinear systems there are no universally applicable models. Volterra series expansions can represent very general nonlinear systems, but they are often severely overparameterized. To reduce the effort of identification, parsimonious models such as nonlinear auto-regressive, moving-average (NARMA) models [6],[7] are adopted. A Nth-order NARMA model can in general be expressed by

$$y(n) = \mathcal{F}(y(n-1),\ldots,y(n-N),w(n),\\ w(n-1),\ldots,w(n-N))$$

where $\mathcal{F}(\cdot)$ is some nonlinear function, $y(n)$ and $w(n)$ are the current output and input, and $y(n-i)$'s and $w(n-i)$'s for $i \neq 0$ are the delayed outputs and inputs.

Identification schemes for nonlinear dynamic systems usually consist two mutually dependent parts, e.g., [8]–[12], the linear identification in which correlation analysis is used, and the nonlinear identification. The nonlinear characteristics of these models are often assumed to be of polynomial form, and usually with memory. For identification of NARMA models [12], especially, the nonlinear characteristics are determined by regression methods. In this kind of scheme, a critical step is to predetermine a set of delay elements, and the crossrelations among these nonlinear delay elements. Then the regression methods based on some criterion can be applied to include or exclude a given nonlinear delay element. The identification effort will be greatly increased if such information cannot be obtained.

It is the purpose of this paper to develop a new scheme for identification of NARMA models. The goal is to identify a suitable NARMA process to match two given sequences of steady state input and output data $\mathbf{w}$ and $\mathbf{y}$ (in the following, the process $\mathbf{x}$ is used to represent either $\mathbf{w}$ or $\mathbf{y}$ unless specifically noted). The inputs and outputs, which are assumed stationary, are collected from some causal, bounded-input/bounded-output stable, time-invariant nonlinear system. The orthogonal projection based scheme allows the determination of significant delay elements and ignores the crossrelations among these elements. Hence the scheme is suboptimal. Furthermore, the nonlinrear characteristics within the model are not restricted to be of polynomial forms but can be the class of Borel functions.

Before defining cost functions for the proposed scheme, some notations used in the paper are introduced.

(1) The input sequence $\mathbf{w}$ and the output sequence $\mathbf{y}$ are both denoted by $\mathbf{x}$.
(2) The delay element $x(n-i)$ is denoted by x_i.
(3) The current output $y(n)$ is denoted by y.

Consider the following mean-squared-error criterion for the N-th order NARMA model identification

$$\mathbf{C} = E\{((y - E[y]) - \mathbf{A}(E[y \mid \mathbf{x}] - E[\mathbf{y}]))^2\}, \qquad (1\text{-}1)$$

where $E[y \mid \mathbf{x}] \equiv E[y \mid x_1] \quad E[y \mid x_2] \cdots E[y \mid x_N]]^T$ (the superscript "T" denotes the transpose) is a $N \times 1$ column vector, $E[\mathbf{y}] = [E[y] \quad E[y] \cdots E[y]]^T$ is a $N \times 1$ column vector, and $\mathbf{A}$ are $N \times 1$ row vectors which are to be determined so that $\mathbf{C}$ are minimized (as is familiar from the linear case). The cost function $\mathbf{C}$ is similar to the cost function introduced in [13] where the mean value of the output process is assumed to be zero. Note that $\mathbf{C}$ is different from the usual orthogonality-based cost function $E\{(y - \mathbf{A}_L\mathbf{x})^2\}$ defined in linear system identification, since they contain the conditional expectation of y given $\mathbf{x}$ rather than $\mathbf{x}$ itself.

The cost function as defined in (1-1) makes the identification a two-step procedure. The first is to determine the nonlinear function for each delay element. This step is not studied in this paper and some possible techniques for identification of nonlinear characteristics are given in the references.

Manuscript received April 26, 1993; revised October 11, 1994. This paper was recommended by Associate Editor Martin Hasler.

H.-E. Liao is with the Department of Electrical Engineering, Feng Chia University, Taichung, Taiwan, R.O.C.

W. A. Sethares is with the Department of Electrical and Computer Engineering, University of Wisconsin, Madison, WI 53706 USA.

IEEE Log Number 9407842.

1057–7122/95$04.00 © 1995 IEEE

There are in general two catagories for such identification, parametric methods based on least-squared-error [14] when polynomial forms are assumed, and nonparametric methods, e.g., [15]–[19].

The second step is to put weights on each nonlinear function determined in the first step. This can be achieved by solving a set of linear equations with dispersion functions serving as the coefficients as we will show later. The N-th order model thus identified has output y,

$$y = \mathbf{A}_0(E[y \mid \mathbf{x}] - E[\mathbf{y}]) + E[y] \tag{1-2}$$

for the cost defined by (1-1), where $\mathbf{A}_0$ equation represents the optimal version of $\mathbf{A}$. The identified model is pictorially shown in Fig. 1 by assuming the output process $\mathbf{y}$ has zero mean, where $a_0, a_1, \ldots, a_N$, $b_1, \ldots, b_N$ are the elements of $\mathbf{A}_0$, $f_i(w(n-i)) \equiv E[y(n) \mid w(n-i)]$ for $i = 0, 1, \ldots, N$ and $g_i(y(n-i)) \equiv E[y(n) \mid y(n-i)]$ for $i = 1, \ldots, N$. Or, explicitly, rewrite (1-2) as

$$y = \sum_{i=1}^{N} b_i g_i(y_i) + \sum_{i=0}^{N-1} a_i f_i(w_i). \tag{1-3}$$

From (1-2), it is understood that the scheme not only decouples the linear and nonlinear identification problems but also decomposes the nonlinear characteristics to each delay element. And, as mentioned above, the nonlinear functions f_i and g_i, are not restricted to polynomial but can be any Borel functions.

In Section II, dispersion functions are briefly reviewed. A. method of solving for $\mathbf{A}_0$ from the cost $\mathbf{C}$ is discussed in Section III. Section IV describes the characteristics of the two-step scheme proposed for nonlinear identification. In Section V, two simulations are conducted.

II. Dispersion Functions

Dispersion functions were first proposed by N. S. Rajbman [20]. Some of their properties and application to nonlinear system identification such as the determination of the nonlinear degree of plants can be found in [20]–[22].

Three catagories of dispersion functions [23] are defined in the following. For three random processes $x(s), s \in T_x, y(t), t \in T_y$, and $z(u), u \in T_z$, the cross-dispersion function, denoted by $\theta_{y|x}(t,s)$, is

$$\theta_{y|x}(t,s) \equiv \operatorname{cov}\{E[y(t) \mid x(s)], E[y(t) \mid x(s)]\}. \tag{2-1}$$

The auto-dispersion function $\theta_{y|y}(t,s,)$ is the special case where $y(t) = x(t)$ and $t \in \mathbf{T}_y$,

$$\theta_{y|y}(t,s) \equiv \operatorname{cov}\{E[y(t) \mid y(s)], E[y(t) \mid y(s)]\}, \tag{2-2}$$

and the generated-dispersion function, denoted by $\theta_{y|x|z}(t,\ s,\ u)$, is

$$\theta_{y|x|z}(t,s,u) \equiv \operatorname{cov}\{E[y(t) \mid x(s)], E[y(t) \mid z(u)]\}. \tag{2-3}$$

In the above definitions, $\operatorname{cov}\{\cdot,\cdot\}$ is the covariance operator of two random processes and $E[\cdot|\cdot]$ is the conditional expectation.

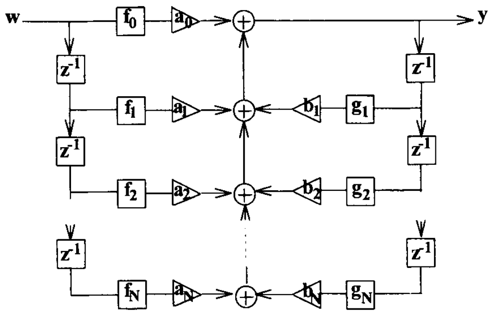

Fig. 1. The N-th order nonlinear ARMA (NARMA) model for system identification.

The cross-dispersion function $\theta_{y|x}(t,s)$ represents, for the given values of the arguments t and s, the variance of the conditional expectation of $y(t)$ relative to $x(s)$, and it characterizes the overall variance of $y(t)$ on the σ-field generated by the random process $x(s)$. Similarly, the auto-dispersion function $\theta_{y|y}(t,s)$ represents, for the given values of the argments t and s, the conditional expectation of $y(t)$ relative to $y(s)$, and it characterizes the overall variance of $y(t)$ on the σ-field generated by the random process $y(s)$. The generated-dispersion $\theta_{y|x|z}(t,s,u)$ represents, for the given values of the arguments t, s and u, the covariance between the random processes $E[y(t) \mid x(s)]$ and $E[y(t) \mid z(u)]$. In the case when all random processes are jointly stationary, the dispersion functions in (2-1) and (2-2) become functions of $(t-s)$, i.e.,

$$\theta_{y|x}(t,s) = \theta_{y|x}(\tau)$$

and

$$\theta_{y|y}(t,s) = \theta_{y|y}(\tau),$$

where $\tau = t - s$.

Dispersion functions are the nonlinear analog of correlation functions. $\theta_{y|x}(t,s)$ is an analog of the cross-covariance function $C_{yx}(t,s)(=E\{(y(t) - E[y(t)])(x(s) - E[x(s)])\})$ which characterizes the strength of the relation between the two random processes $y(t)$ and $x(s)$. Similarly $\theta_{y|y}(t,s)$ is an analog of the auto-covariance function $C_{yy}(t,s)(=E\{(y(t) - E[y(t)])(y(s) - E[y(s)])\})$ which characterizes the internal structure of the random process $y(t)$. However, $C_{yx}(t,s)$ and $C_{yy}(t,s)$ measure the relationship among the random processes $x(s)$, $y(s)$ and $y(t)$ only when they are linearly related, while $\theta_{y|x}(t,s)$ and $\theta_{y|y}(t,s)$ provide a measure of the relationship among these random processes even when they are nonlinear functions of each other.

Conditional expectation can be interpreted as an "orthogonal projection" assuming that $E[y^2(t)] < \infty$ [24], i.e., $E\{(y(t) - E[y(t) \mid x(s)])E[y(t) \mid x(s)]\} = 0$ and $E\{(y(t) - f(x(s))^2\} \geq \{(y(t) - E[y(t) \mid x(s)])^2\}$, where $y(t)$, $x(s)$ are random processes, and $f(\cdot)$ is some nonlinear (Borel) function. Therefore, dispersion functions measure the "power" of $y(t)$ after the "orthogonal projection" on some σ-fields or the "significance" $y(t)$ relative to some random variable. Some properties of dispersion functions are listed in Appendix I.

III. Solving for the Optimal Linear Weights

In this section, the optimal weights are determined by minimizing the cost function $\mathbf{C}$ in (1-1). The first subsection studies the general case for solving the optimal weights which are dependent on evaluation of dispersion functions. The second subsection gives a special case, where NMA models with uncorrelated input are assumed.

III.1. The Optimal Weights: General Case

Consider a given system with input sequence $\mathbf{w}$ and current output y. Consider the cost function $\mathbf{C}$ in (1-1)

$$\mathbf{C} = E\{((y - E[y]) - \mathbf{A}(E[y \mid \mathbf{x}] - E[\mathbf{y}]))^2\}.$$

The objective of the identification is to determine the vector $\mathbf{A}$ that minimizes the cost function $\mathbf{C}$. Expanding (1-1) for $\mathbf{C}$ gives

$$\begin{aligned}\mathbf{C} &= E\{((y - E[y]) - \mathbf{A}(E[y \mid \mathbf{x}] - E[\mathbf{y}]))\\ &\quad \times ((y - E[y]) - \mathbf{A}(E[y \mid \mathbf{x}] - E[\mathbf{y}]))^T\}\\ &= \sigma_y^2 - 2\mathbf{A}E\{(y - E[y])(E[y \mid \mathbf{x}] - E[\mathbf{y}])\}\\ &\quad + \mathbf{A}E\{(E[y \mid \mathbf{x}] - E[\mathbf{y}])(E[y \mid \mathbf{x}] - E[\mathbf{y}])^T\}\mathbf{A}^T,\end{aligned}$$

where σ_y^2 is the variance of the output process. Since $\mathbf{C} \geq 0$ is a quadratic form with respect to the row vector $\mathbf{A}$, the minimum value of the cost function can be determined by setting $d\mathbf{C}/d\mathbf{A} = \mathbf{0}$ ($\mathbf{0}$ is the null column vector), which results in

$$\begin{aligned}&E\{(E[y \mid \mathbf{x}] - E[\mathbf{y}])(E[y \mid \mathbf{x}] - E[\mathbf{y}])^T\}\mathbf{A}^T\\ &\quad = E\{(y - E[y])(E[y \mid \mathbf{x}] - E[\mathbf{y}])\}.\end{aligned}$$

The above equation can be denoted in a matrix form as

$$\mathbf{H}\mathbf{A}^T = \mathbf{R}, \tag{3-1}$$

where

$$\begin{aligned}\mathbf{H} &= E\{(E[y \mid \mathbf{x}] - E[\mathbf{y}])(E[y \mid \mathbf{x}] - E[\mathbf{y}])^T\}\\ &= E\left\{\begin{bmatrix} E[y \mid x_1] - E[y] \\ \vdots \\ (E[y \mid x_N] - E[y] \end{bmatrix}\right.\\ &\qquad \left. \times [(E[y \mid x_1] - E[y]) \cdots (E[y \mid x_N] - E[y])]\right\}.\end{aligned}$$

Thus

$$\begin{aligned}[\mathbf{H}]_{i,j} &= E\{(E[y \mid x_i] - E[y])(E[y \mid x_j] - E[y])\}\\ &\quad \text{for } i, j = 1, \ldots, N\end{aligned}$$

is a $N \times N$ square matrix. Similarly

$$\mathbf{R} = E\{y(E[y \mid \mathbf{x}] - E[\mathbf{y}])\},$$

and

$$\begin{aligned}[\mathbf{R}]_i &= E\{y(E[y \mid x_i] - E[y])\}\\ &= E\{(E[y \mid x_i] - E[y])^2\}[24]\end{aligned}$$

is a $N \times 1$ column vector. Note that the elements of $\mathbf{H}$ and $\mathbf{R}$ are the "dispersion functions" defined in (2-1), (2-2), and (2-3). Hence

$$[\mathbf{H}]_{i,j} = \theta_{y|x_i|x_j}$$

and

$$[\mathbf{R}]_i = \theta_{y|x_i|x_i} = \theta_{y|x_i}.$$

Note also that the elements of $\mathbf{R}$ are just the diagonal elements of $\mathbf{H}$. Analogous to the covariance matrix in linear case, the matrix $\mathbf{H}$ has following properties:

1. $[\mathbf{H}]_{i,j} = [\mathbf{H}]_{j,i}$ and thus $\mathbf{H}$ is symmetric;
2. Since $\mathbf{H}$ is symmetric, $\mathbf{H}$ has real eigenvalues and a set of orthonormal eigenvectors.
3. $\mathbf{H}$ is positive semi-definite.

Proof (of the third property):
Let $s = \mathbf{t}(E[y \mid \mathbf{x}] - E[\mathbf{y}])$, where $\mathbf{t} \neq \mathbf{0}$ is an arbitrary $1 \times N$ row vector. Then

$$s^2 = \mathbf{t}(E[y \mid \mathbf{x}] - E[\mathbf{y}])(E[y \mid \mathbf{x}] - E[\mathbf{y}])^T\mathbf{t}^T,$$

and

$$E[s^2] = \mathbf{t}E\{(E[y \mid \mathbf{x}]-E[\mathbf{y}])(E[y \mid \mathbf{x}]-E[\mathbf{y}])^T\}\mathbf{t}^T = \mathbf{t}\mathbf{H}\mathbf{t}^T.$$

Since $E[s^2] \geq 0$, $\mathbf{H}$ is positive semi-definite. Q.E.D.

Assume $\mathbf{H}$ is nonsingular so that the optimum of $\mathbf{A}$ (denoted by $\mathbf{A}_0$ from now on) can be solved as $\mathbf{A}_0^T = \mathbf{H}^{-1}\mathbf{R}$, the cost function $\mathbf{C}$ is minimized. The minimum cost $\mathbf{C}$ (denoted by $\mathbf{C}_0$) is

$$\mathbf{C}_0 = \sigma_y^2 - \mathbf{A}_0\mathbf{R}. \tag{3-2}$$

III.2. The Optimal Weights: A Special Case

Consider the following NMA model of order N where the output is contaminated by some noise source $\mathbf{v}$,

$$y(n) = f_1(w(n-1)) + \cdots + f_N(w(n-N)) + v(n),$$

where f_i's are assumed to be Borel functions. For notational simplicity, rewrite this as

$$y = f_1 + \cdots + f_N + v_0, \tag{3-3}$$

and assume that the inputs $w(n-i)$'s and $v(n)$ are mutually uncorrelated. Then, $\theta_{y|w_i|w_j} = 0$ for $i \neq j$ since

$$\begin{aligned}&E\{E[y \mid w_i]E[y \mid w_j]\}\\ &\quad = E\{(f_i + E[f_1 + \cdots + f_N + v_0 - f_i])\\ &\qquad \times (f_j + E[f_1 + \cdots + f_N + v_0 - f_j])\}\\ &\quad = E\{(f_i + E[f_1 + \cdots + f_N + v_0 - f_i])\}\\ &\qquad \times E\{f_j + E[f_1 + \cdots + f_N + v_0 - f_j])\}\\ &\quad = E[f_1 + \cdots + f_N + v_0]E[f_1 + \cdots + f_N + v_0]\\ &\quad = E^2[y].\end{aligned}$$

Therefore,

$$\theta_{y|w_i|w_j} = E\{E[y \mid w_i]E[y \mid w_j]\} = 0.$$

Thus (3-1) becomes

$$\begin{bmatrix}\theta_{y|w_1} & & \mathbf{0}\\ & \ddots & \\ \mathbf{0} & & \theta_{y|w_N}\end{bmatrix}\begin{bmatrix}a_1\\ \vdots\\ a_N\end{bmatrix} = \begin{bmatrix}\theta_{y|w_1}\\ \vdots \\ \theta_{y|w_N}\end{bmatrix}. \qquad (3\text{-}4)$$

Note that $\mathbf{H}$ is diagonal and it is easy to see that $a_1 = a_2 = \cdots = a_N = 1$ as desired. That is, according to (1-2), the current output of this case is

$$\begin{aligned}\mathbf{A}_0(E[y \mid \mathbf{w}] - E[\mathbf{y}]) + E[y] &= \sum_{i=1}^{N} E[y \mid w_i] - (N-1)E[y]\\ &= \sum_{i=1}^{N} f_i. \qquad (3\text{-}5)\end{aligned}$$

Futhermore, $\theta_{y|w_i} = E[f_i^2] - E^2[f_i]$. Since

$$\begin{aligned}&E\{E^2[y \mid w_i]\}\\ &\quad = E\{yE[y \mid w_i]\}\\ &\quad = E\{y(f_i + E[f_1 + \cdots + f_N + v_0 - f_i])\}\\ &\quad = E[yf_i] + E[y]E[(f_1 + \cdots + f_N + v_0) - f_i]\\ &\quad = E[(f_1 + \cdots + f_N + v_0 - f_i + f_i)f_i] + E[y]E[y - f_i]\\ &\quad = E[f_i]E[y - f_i] + E[f_i^2] + E^2[y] - E[y]E[f_i]\\ &\quad = E[f_i^2] - E^2[f_i] + E^2[y].\end{aligned}$$

Therefore,

$$\theta_{y|w_i} = E\{E^2[y \mid w_i]\} - E^2[y] = E[f_i^2] - E^2[f_i]. \qquad (3\text{-}6)$$

The variance of the output process, σ_y^2, can be computed straightforwardly using the fact that the $w(n-i)$'s and v are mutually uncorrelated. Hence, using (3-6), the result is

$$\sigma_y^2 = \theta_{y|w_1} + \cdots + \theta_{y|w_N} + \sigma_{v_0}^2, \qquad (3\text{-}7)$$

where $\sigma_{v_0}^2$ is the variance of the noise process v_0. Therefore,

$$\mathbf{C}_0 = \sigma_y^2 - \mathbf{A}_0\mathbf{R} = \sigma_{v_0}^2. \qquad (3\text{-}8)$$

III.3. Remarks

Remark 3-1: Subtraction of the mean value of the output process $\mathbf{y}$ can not be omitted in $\mathbf{C}$ defined in (1-1). Suppose another cost function $\hat{\mathbf{C}}$ defined without the subtraction, i.e., $\hat{\mathbf{C}} = E\{(y - \hat{A}E[y \mid \mathbf{x}])^2\}$, then it can be shown that $\hat{\mathbf{C}} \geq \mathbf{C}$. This can be explained in the following. Consider the special case discussed in III.2. From (3-3), $E[y \mid w_i] = f_i + E[\sum_{j\neq i} f_j]$, we can see that the result deviates from the desired function f_i by a dc-bias $E[\sum_{j\neq i} f_j]$. These dc-biases are accumulated for each nonlinear function and are removed by the subtraction of $(N-1)E[y]$ as shown in (3-5). For the general case, $E[y \mid x_i] = f_i + \sum_{j\neq i} E[f_j \mid x_i]$, the situation becomes more complicated. Nonlinear functions f_i's, for $i = 1, \ldots, N$, can not be recovered and are represented by some other forms because of the correlated data. The terms $\sum_{j\neq i} E[f_j \mid x_i]$ accumulate as in the special case where dc-biases accumulate and the effect can be reduced by the subtraction of $E[y]$ and the linear wieghting.

Remark 3-2: In the above discussion, the scalar case is adopted, i.e., $E[y|x_i]$ is the orthogonal projection of y on the σ-field generated by x_i. This can be extended to the vector case so that the projection is on the σ-field generated by several random processes. This means that $E[y|x_i]$ can be replaced by, for example, $E[y|x_1, x_2, \ldots, x_m]$. In this variation, the mean squared error will be reduced (this is expected as properties of orthogonal projection and, thus, properties of dispersion functions listed in Appendix I). This is especially obvious when delay elements have strong cross-relationship, i.e., they are cross terms to each other. One possible disadvantage of this extension is that the identification effort is increased since multi-dimensional estimates of nonlinear characteristics are needed.

IV. Characteristics of the Algorithm

The proposed identification strategy is a two-step scheme. First of all, the nonlinear functions for individual delay elements are determined using techniques such as nonlinear regression, polynomial fitting and splines. Then dispersion functions are evaluated so that weights are determined for each nonlinear function using (3-1). Some important characteristics of the algorithm are discussed below.

IV.1. Consistency of the Orthogonal Projection

The cost function $\mathbf{C}$ defined in (1-2) is analogous to the usual mean squared error defined in linear identification and the minimized result is closly related to the orthogonal principle although a slight modification must be made, i.e., the additional mean value of the output process must be subtracted out. To see this, rewrite $\mathbf{H}(\mathbf{A}_0^T) = \mathbf{R}$ as

$$\begin{aligned}&E\{(E[y \mid \mathbf{x}] - E[\mathbf{y}])(E[y \mid \mathbf{x}] - E[y])^T\}\mathbf{A}_0^T\\ &\quad = E\{y(E[y \mid \mathbf{x}] - E[\mathbf{y}])\},\end{aligned}$$

which results in

$$\begin{aligned}&E\{(\mathbf{A}_0(E[y \mid \mathbf{x}] - E[\mathbf{y}]))((y - E[y]) - \mathbf{A}_0(E[y \mid \mathbf{x}] - E[\mathbf{y}])\}\\ &\quad + E\{E[y](E[y \mid \mathbf{x}] - E[\mathbf{y}])\} = 0.\end{aligned}$$

Note that the second term in the above equation equals zero for $E\{E[y \mid x_i]\} = E[y]$. Therefore,

$$E\{\mathbf{A}_0(E[y \mid \mathbf{x}] - E[\mathbf{y}])((y - E[y]) - \mathbf{A}_0(E[y \mid \mathbf{x}] - E[\mathbf{y}]))\} = 0. \qquad (4\text{-}1)$$

In (4-1), the term $(y - E[y]) - \mathbf{A}_0(E[y \mid \mathbf{x}] - E[\mathbf{y}])$ is the error after $y - E[y]$ is projected on the space spanned by $(E[y \mid x_i] - E[y])$. Hence, as in the linear case, the minimization of $\mathbf{C}$ is an orthogonal projection on the space spanned by $\{(E[y \mid x_i]\}_{i=1}^{N}$. At the same time, it can be shown that the error is also orthogonal to the individual estimation in the first step of the algorithm, i.e. $[(y - E[y]) - \mathbf{A}_0(E[y \mid$

$\mathbf{x}] - E[\mathbf{y}])]$ is orthogonal to $(E[y \mid x_i] - E[y])$. This is shown below.

$$
\begin{aligned}
&E\{((y - E[y]) - \mathbf{A}_0(E[y \mid \mathbf{x}] - E[\mathbf{y}]))(E[y \mid x_i] - E[y])\}\\
&= E\{(y - \mathbf{A}_0 E[y \mid \mathbf{x}])E[y \mid x_i]\}\\
&\qquad - E\{(y - \mathbf{A}_0 E[y \mid \mathbf{x}])E[y])\}\\
&\quad - E\{(E[y] - \mathbf{A}_0 E[\mathbf{y}])E[y \mid x_i]\}\\
&\qquad + E\{(E[y] - \mathbf{A}_0 E[y])E[y]\}
\end{aligned}
$$

(note the difference between the second and the fourth terms is zero)

$$
\begin{aligned}
&= E\{(y - \mathbf{A}_0 E[y \mid \mathbf{x}])E[y \mid x_i]\}\\
&\quad - E\{(E[y] - \mathbf{A}_0 E[y])E[y \mid x_i]\}
\end{aligned}
$$

(note that $E\{yE[y \mid x_i]\} = E\{E^2[y \mid x_i]\}$)

$$
\begin{aligned}
&= (E\{E^2[y \mid x_i]\} - E^2[y])\\
&\quad - (\mathbf{A}_0(E\{E[y \mid x_i]E[y \mid \mathbf{x}] - E^2[y])\}
\end{aligned}
$$

(then by the definition of dispersion function)

$$
= \theta_{y|x_i} - \mathbf{A}_0 \begin{bmatrix} \theta_{y|x_i|x_1} \\ \vdots \\ \theta_{y|x_i|x_N} \end{bmatrix}
$$

(the second term is equal to $\theta_{y|x_i}$ from (3-1), i.e. the i-th row of $\mathbf{H}$ times $\mathbf{A}_0^T$ equals to $\theta_{y|x_i}$)

$$= 0.$$

Hence, as desired,

$$E\{((y - E[y]) - \mathbf{A}_0(E[y \mid \mathbf{x}] - E[\mathbf{y}]))(E[y \mid x_i] - E[y])\} = 0. \tag{4-2}$$

From (4-1) and (4-2), the overall minimum error in the two-step algorithm is orthogonal to both of the two estimation results consistently. This implies that although the algorithm consists of two seperate steps (nonlinear and linear identifications), the result coincides with the orthogonal principle.

IV.2. Nonincreasing Error Performance

In the minimization problem for linear identification, the mean-squared-error will be reduced as the order of the model is increased. This is not true in general for nonlinear system identification. The following shows that the scheme proposed has nonincreasing error as the order of the model is increased and gives a formula for the performance difference between two consecutive orders. Let the minimum mean-squared-errors of N-th order and $(N-1)$-th order estimations be $\mathbf{C}_0^{(N)}(=\sigma_y^2 - \mathbf{A}_0^{(N)}\mathbf{R}_N)$ and $\mathbf{C}_0^{(N-1)}(=\sigma_y^2 - \mathbf{A}_0^{(N-1)}\mathbf{R}_{N-1})$ respectively. Then

$$
\begin{aligned}
\mathbf{C}_0^{(N-1)} - \mathbf{C}_0^{(N)} &= \mathbf{A}_0^{(N)}\mathbf{R}_N - \mathbf{A}_0^{(N-1)}\mathbf{R}_{N-1}\\
&= \mathbf{R}_N^T\mathbf{H}_N^{-1}\mathbf{R}_N - \mathbf{R}_{N-1}^T\mathbf{H}_{N-1}^{-1}\mathbf{R}_{N-1}.
\end{aligned} \tag{4-3}
$$

The second term in the right-hand-side of the above equation can be rewritten as

$$\mathbf{R}_{N-1}^T\mathbf{H}_{N-1}^{-1}\mathbf{R}_{N-1} = \mathbf{R}_N^T\tilde{\mathbf{H}}_N^{-1}\mathbf{R}_N,$$

where

$$
\tilde{\mathbf{H}}_N^{-1} = \begin{bmatrix} & \mathbf{H}_{N-1}^{-1} & & \vdots & \mathbf{0} \\ \cdots & \cdots & \cdots & \vdots & \cdots \\ & \mathbf{0}^T & & \vdots & 0 \end{bmatrix}.
$$

Similarly, $\mathbf{H}_N$ and $\mathbf{R}_N^T$ can be partitioned as

$$
\begin{aligned}
\mathbf{H}_N &= \begin{bmatrix} & \mathbf{H}_{N-1} & & \vdots & \theta \\ \cdots & \cdots & \cdots & \vdots & \cdots \\ & \theta^T & & \vdots & \theta_N \end{bmatrix};\\
\mathbf{R}_N^T &= [\mathbf{R}_{N-1}^T \mid \theta_N]^T;
\end{aligned}
$$

where $bf\theta = [\theta_{y|x_1|x_N} \cdots \theta_{y|x_{N-1}|x_N}]^T$, $\mathbf{R}_{N-1}^T = [\theta_{y|x_1} \cdots \theta_{y|x_{N-1}}]$ and $\theta_N = \theta_{y|x_N}$. Then (4-3) becomes

$$\mathbf{C}_0^{(N-1)} - \mathbf{C}_0^{(N)} = \mathbf{R}_N^T\{\mathbf{H}_N^{-1} - \tilde{\mathbf{H}}_N^{-1}\}\mathbf{R}_N.$$

Furthermore, the above equation can be simplified to (using the formula shown in Appendix II),

$$\mathbf{C}_0^{(N-1)} - \mathbf{C}_0^{(N)} = \frac{(\mathbf{A}_0^{N-1}\theta - \theta_N)^2}{\theta_N - \theta^T H_{N-1}^{-1}\theta}, \tag{4-4}$$

and in Appendix III, $\theta_N - bf\theta^T H_{N-1}^{-1} bf\theta$ is shown to be greater than zero. Hence, as the order of the model is increased, the mean-squarred-error is nonincreasing. Equation (4-4) also gives a measure of whether an extra delay element should be added to the current model without the need to solve the linear equations of higher order.

IV.3. Interaction of the Delay Elements

As it is mentioned in Section II, the larger the dispersion functions are, the more significant the corresponding delay elements are. For a second order model, suppose the identified model is

$$y - E[y] \approx a_1(E[y \mid x_1] - E[y]) + a_2(E[y \mid x_2] - E[y]).$$

Then the contributed "power" difference between the two delay elements x_1 and x_2 to this model is

$$
\begin{aligned}
&E\{a_1^2(E[y \mid x_1] - E[y])^2\} - E\{a_2^2(E[y \mid x_2] - E[y])^2\}\\
&= a_1^2\theta_{y|x_1} - a_2^2\theta_{y|x_2}.
\end{aligned}
$$

It can be shown (by solving a_1 and a_2 for the second order model explicitly) that whenever $\theta_{y|x_1} > \theta_{y|x_2}$, $a_1^2\theta_{y|x_1} - a_2^2\theta_{y|x_2} > 0$. This result coincides with the previous argument. But this is not always true for higher order cases, i.e., $\tilde{a}_1^2\theta_{y|x_1} - \tilde{a}_2^2\theta_{y|x_2}$ is sign indefinite even if $\theta_{y|x_1} > \theta_{y|x_2}$, where $\tilde{a}_1$ and $\tilde{a}_2$ are the solutions for the higher order model. The dispersion functions $\theta_{y|x_1|x_i}$ for $i = 2, \ldots, N$ and $\theta_{y|x_2|x_j}$ for $j = 1, 3, \ldots, N$ of a N-th order model will influence the "power" distribution. $\theta_{y|x_i|x_j}$ is the measure of the correlation between the two random processes $E[y \mid x_i]$ and $E[y \mid x_j]$. If $\theta_{y|x_i|x_j}$ is very large, $\theta_{y|x_i|x_j} \approx \theta_{y|x_i} < \theta_{y|x_j}$ for example, then some amount of "power" contributed by x_i will be consumed by x_j after the orthogonal projection on the span of $\{E[y \mid x_1], \ldots, E[y \mid x_N]\}$. Therefore, this allocation of

$\theta_{y|x_i|x_j}$ for $i \neq j$ causes the sign of $\tilde{a}_i^2\theta_{y|x_i} - \tilde{a}_j^2\theta_{y|x_j}$ to be indefinite. The rule of thumb is that when $\theta_{y|x_k}$ for some delay element $x(n-k)$ is null (or very small compared with the others since $\theta_{y|x_j}$ may be contaminated by the noise) then $x(n-k)$ should not be included into the model. Such small $\theta_{y|x_j}$ in the matrix **H** makes the solution of (3-1) very ill-conditioned where the condition number of **H** is defined as the ratio of the maximum eigenvalue, $\lambda_{\max}$ to the minimum eigenvalue $\lambda_{\min}$ of **H**. Since, by Raleigh's quotient,

$$\frac{\lambda_{\max}}{\lambda_{\min}} \geq \frac{\theta_{\max}}{\theta_{\min}},$$

where $\theta_{\max} \equiv \max\{\theta_{y|x_i}\}_{i \in s}$ and $\theta_{\min} \equiv \min\{\theta_{y|x_i}\}_{i \in s}$ is some positive integer set.

V. Simulations

Given two sequences of data **w** and **y** (input and output in steady state) of an unknown system, the goal is then to build a class of NARMA models as depicted in Fig. 1 so that the output of the model can mimic the output of the given system. The proposed scheme in this paper consists of two steps, evaluating the nonlinear function for each delay element and then solving a set of linear equations composed of dispersion functions. Two simulations are given in Sections V.2 and V.3, which identify a NMA model and a NAR model respectively. The procedures used in the simulations are explained in the following subsection. The procedure, though naive, gives convincing results. In the simulations, the output processes are assumed to be ergodic so that the mean values can be evaluated by averaging the time sequences.

V.1. Simulation Procedures

A. *Evaluation of the Nonlinear Functions, $E[y \mid x_i]$*:

1) "Quantize" each delay element by dividing it into M small intervals of equal width δ. Let this sequence of intervals be denoted by $\{I_m\}_{m=1}^{M}$.
2) Evaluate the average value of the output process within each interval. This gives the "estimation" of $E[y \mid x_i \in I_m]$ for $m = 1, \ldots, M$ and denote these discrete nonlinear functions as $f_i(x(n-i))$ (or $g_i(x(n-i))$. Hence $f_i(x(n-i))$ has M values, one corresponding to each interval.
3) Use least mean square polynomial (of degree fifteen) fit to estimate the "continuous version" of $f_i(x(n-i))$ (or $g_i(x(n-i))$) and denote it as $p_i(x(n-i))$ (or $q_i(x(n-i))$).

B. *Solving the Row Matrix* $\mathbf{A}_0$:

1) Pointwise multiply $f_i(x(n-i))$ and $f_j(x(n-j))$ for each interval I_m. There are M values after the pointwise multiplication.
2) Average these M values of the above multiplication and, then, subtract the average value of the output process. This results in $\theta_y|x_i|x_j$.
3) Construct the matrix $\tilde{\mathbf{H}}$ and solve (3-4) for $\mathbf{A}_0$.

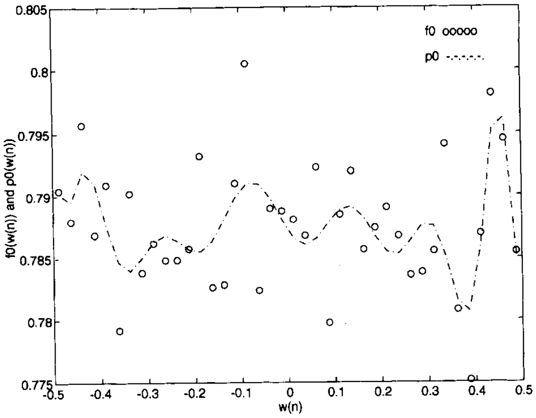

Fig. 2. The estimated nonlinear function for the delay element $w(n)$.

For each simulation, 5000 pairs of input and output data are given (generated from the "unknown" system). The first 500 data points are discarded to let the system converge to its steady state and the rest of data is used for the identification. The final results are obtained by averaging 20 independent experiments. The model output, denoted by $\tilde{\mathbf{y}}$, will, according to (1-2), be

$$\tilde{y}(n) = \sum_{i \in S} a_i p_i(w(n-i)) + E[y]\left(1 - \sum_{i \in S} a_i\right),$$

for NMA processes and

$$\tilde{y}(n) = \sum_{i \in S} a_i q_i(y(n-i)) + E[y]\left(1 - \sum_{i \in S} a_i\right),$$

for NAR processes. Where a_i represents the elements of $\mathbf{A}_0$. Mean-squared-errors are then computed, by feeding new sets of data to the "unknown systems" and the identified models with the same initial conditions, to investigate the performance of the identified models.

V.2.. Identification of NMA Process

Let y be the output process generated by the following system,

$$y(n) = \exp(-|w(n-1)|) + 0.25\sin(2\pi w(n-2)) + v(n),$$

where the input sequence **w** is uncorrelated and uniformly distributed, $(-0.5,\ 0.5)$, and **v** is added white noise (normally distributed with zero mean) uncorrelated to **w**. The output process **y** of the above system has variance equal to 0.0436. Assume that the third order NMA model of Fig. 1 is used to identify the given system. Following the procedures described in Section V.1, Figs. 2–4 depict the evaluated nonlinear functions for the delay elements $w(n)$, $w(n-1)$ and $w(n-2)$, and we have

$$\mathbf{H} = \begin{bmatrix} 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0129 & 0.0001 \\ 0.0 & 0.0001 & 0.0318 \end{bmatrix}$$

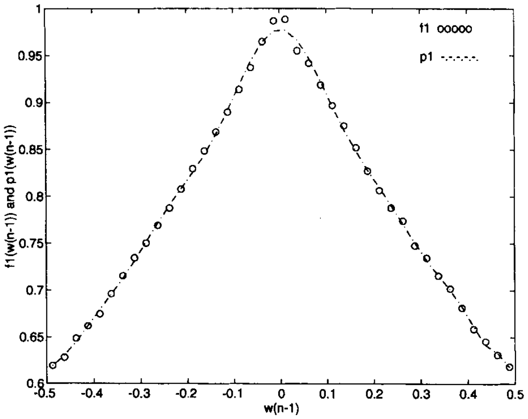

Fig. 3. The estimated nonlinear function for the delay element $w(n-1)$.

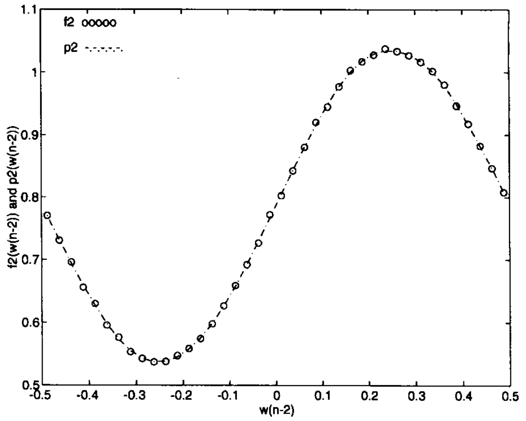

Fig. 4. The estimated nonlinear function for the delay element $w(n-2)$.

(note that "0.0's" in $\mathbf{H}$ are not exactly zeros but some very small values, each less than 10^{-4}),

$$\mathbf{A}_0 = [2.4273 \ 0.9954 \ 1.0001],$$

and the mean-squared-error for this identification (mse_{NMA3}) is

$$\text{mse}_{\text{NMA3}} = 2.7493 \times 10^{-4}.$$

The sum of the diagonal elements of $\mathbf{H}$ approximates the variance of the output processes, that is σ_y^2, which coincides with (3-7). The off-diagonal elements of $\mathbf{H}$ are all very small, which is also expected from (3-4) of Section III.2. Since the first diagonal element of $\mathbf{H}$ is very small, we can eliminate the first column and the first row of $\mathbf{H}$. Actually it can be seen from Fig. 2 that the nonlinear function of $w(n)$ is noisy, which implies that $w(n)$ is insignificant in representing the unknown system. $\mathbf{A}_0$ for this case is

$$\mathbf{A}_0 = [0.9943 \ 0.9977],$$

and

$$\text{mse}_{\text{NMA2}} = 1.3428 \times 10^{-4}.$$

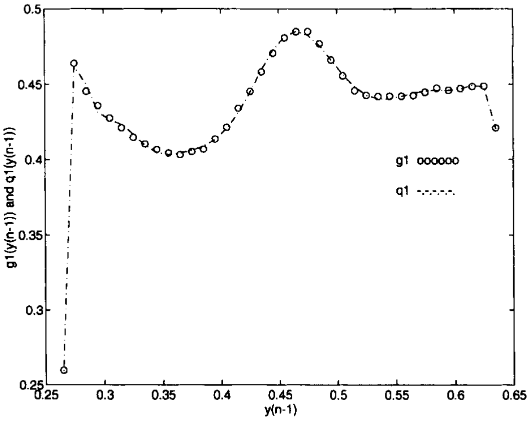

Fig. 5. The estimated nonlinear function for the delay element $y(n-1)$.

The mean square error mse_{NMA2} approximates the variance of the added noise ($\sigma_v^2 = 10^{-4}$), hence this result is convincing, i.e., see (3-8). The importance of the subtraction of the mean value of the output process in (1-2) is observed from Fig. 3 and Fig. 4. In Fig. 4, especially, a sine wave is clearly displayed, which is a shifted version of the sine wave in the given system, that is the term $0.25\sin(2\pi w(n-2))$. This shift action is caused by the term, $\exp(-|w(n-1)|)$. The subtraction of the mean value in (1-2) supresses the dc-bias in the evaluation as discussed in Remark 3-1.

V.3. Identification of NAR Process

Let $\mathbf{y}$ be the output of the following NAR system

$$y(n) = \exp(-|y(n-2)|) + 0.25\cos(2\pi y(n-3)) + v(n),$$

where v is added white noise (normally distributed with zero mean). The output process of this "unknown" system has variance 0.0062. Assume that the third order NAR model of Fig. 1 is used to identify the given system. Following the procedures described in Section V.1, Figs. 5–7 pictorically show the estimated nonlinear functions for the delay elements $y(n-1)$ (g_1 and q_1), $y(n-2)$ (g_2 and q_2) and $y(n-3)$ (g_3 and q_3), and we have

$$\mathbf{H} = \begin{bmatrix} 0.0023 & 0.0011 & 0.0006 \\ 0.0011 & 0.0088 & 0.0048 \\ 0.0006 & 0.0048 & 0.0058 \end{bmatrix},$$

$$\mathbf{A}_0 = [0.5749 \quad 0.7631 \quad 0.3064],$$

and the mean-squared-error for this identification (mse_{NAR}) is

$$\text{mse}_{\text{NAR}} = 5.9310 \times 10^{-4}.$$

The small mean-squared-error shows the scheme works well. Notice that in the estimation of the nonlinear functions, the fitted polynomials have to be carefully chosen so that they are bounded in the region where the output takes values. Actually, one could use other types of basis functions. The simulation results are also effected by the distribution of the output process (which is unknown), i.e. intervals where there is not enough data will have inaccurate estimates of the conditional expectation and cause larger errors.

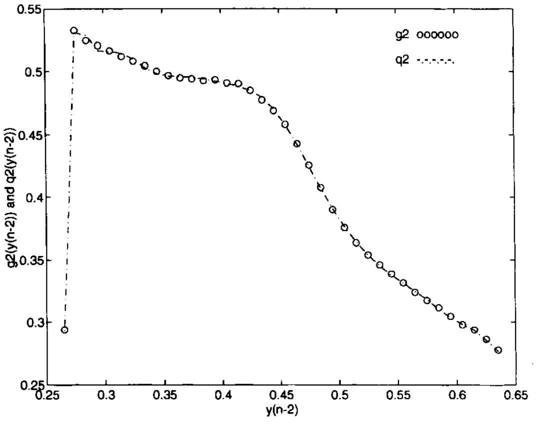

Fig. 6. The estimated nonlinear function for the delay element $y(n-2)$.

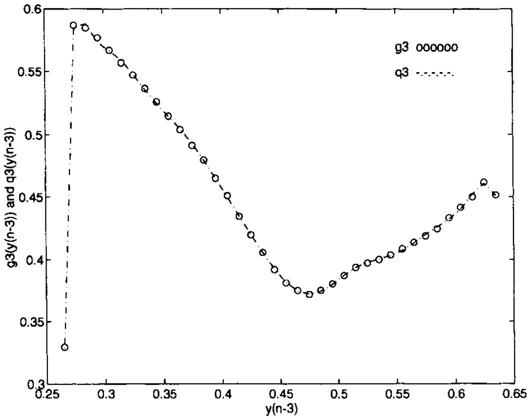

Fig. 7. The estimated nonlinear function for the delay element $y(n-3)$.

VI. Conclusion

This paper proposed a novel scheme for suboptimal nonlinear system identification using the cost **C** defined in (1-1). Though the scheme is a two-step strategy, it is unified. The first step is the extraction of nonlinear characteristics for each delay element via conditional expectation that has the meaning of "orthogonal projection" and hence minimizes the two-norm error. The second step, also based on projection, results in a linear combination of the nonlinear functions evaluated in the first step. Therefore, the proposed scheme not only decouples the linear and nonlinear identification problems but also decouples the nonlinear characteristics to each delay element. Furthermore, as mentioned in Remark 3-2, this subclass of NARMA model can be generalized to other similar NARMA models if cross-relations among delay elements are considered.

Some areas for further investigation are:

1) The estimates of the conditional expectations (for estimating nonlinear functions in the first step, and the evaluation of dispersion functions in the second step) in the identification scheme are not studied analytically. Unfortunately, there is no unbiased method for such estimates. Nonparametric identification for nonlinear characteristics was proposed by Greblicki [25]–[28], where Hammerstein systems and Wiener systems are considered. This may extend to the model shown in Fig. 1.
2) In the above simulations, least square polynomial fitting is utilized to approximate the nonlinear functions. These polynomials must be carefully chosen such that the identified model is stable. Hence appropriate polynomials and degrees, or the study of nonlinear functions which are not based on polynomials should be further considered.

Appendix I

Property 1: $0 \leq \theta_{y|x}(t,s) \leq \sigma_y^2(t)$, where $\sigma_y^2(t)$ is the variance of the random process $y(t)$.

Property 2: If $y(t)$ is independent of $x(s)$ then $\theta y|x(t,s) = 0$.

Property 3: $\theta_{y|x}(t,s) = \sigma 2_y(t)$ if $y(t)$ and $x(s)$ are related by some (possibly nonlinear) functional relation $f(\cdot)$, i.e., if there is a function $f(\cdot)$ such that $y(t) = f(x(s),t,s)$.

Property 4: $\theta_{y|x_1}(t,s)\theta_{y|x_2}(t,u) \geq \theta^2_{y|x_1|x_2}(t,s,u)$ and $\theta_{y|x_1}(t,s)\theta_{y|x_2}(t,s) \geq 2\theta^2_{y|x_1|x_2}$.

Property 5: $\theta_{y|x_1}(t,s) \leq \theta_{y|x_1,x_2}(t,s,u)$, where

$$\theta_{y|x_1,x_2}(t,s,u) = \operatorname{cov}\{E[y(t) \mid x_1(s), x_2(u)], E[y(t) \mid x_1(s), x_2(u)]\}.$$

Appendix II

Let **V** be a $N \times N$ square matrix. Consider a partitioned form of **V**

$$\mathbf{V} = \begin{bmatrix} & \mathbf{A} & & \vdots & \mathbf{b}_1 \\ \cdots & \cdots & \cdots & \vdots & \cdots \\ & \mathbf{b}_2 & & \vdots & c \end{bmatrix},$$

where A is $(N-1) \times (N-1)$ square matrix, $\mathbf{b}_1$ and $\mathbf{b}_2^T$ are $1 \times (N-1)$ column vector and c is a scalar. Then

$$\mathbf{V}^{-1} = \begin{bmatrix} & \tilde{\mathbf{A}} & & \vdots & \tilde{\mathbf{b}}_1 \\ \cdots & \cdots & \cdots & \vdots & \cdots \\ & \tilde{\mathbf{b}}_2 & & \vdots & \tilde{c} \end{bmatrix}, \qquad \text{(A-1)}$$

where

$$\tilde{\mathbf{A}} = \left(\mathbf{A} - \frac{1}{c}\mathbf{b}_1\mathbf{b}_2\right)^{-1};$$

$$\tilde{\mathbf{b}}_1 = -\frac{1}{c}\left(\mathbf{A} - \frac{1}{c}\mathbf{b}_1\mathbf{b}_2\right)^{-1}\mathbf{b};$$

$$\tilde{\mathbf{b}}_2 = -\frac{1}{c}\mathbf{b}_2\left(\mathbf{A} - \frac{1}{c}\mathbf{b}_1\mathbf{b}_2\right)^{-1};$$

and

$$\tilde{c} = \frac{1}{c}\left[1 + \frac{1}{c}\mathbf{b}_2\left(\mathbf{A} - \frac{1}{c}\mathbf{b}_1\mathbf{b}_2\right)^{-1}\mathbf{b}_1\right].$$

APPENDIX III

Fact: Let $\mathbf{V}$ be symmetric and positive definite, $\mathbf{V}^T = \mathbf{V}$ and $\mathbf{V} > 0$. Consider the partitioned form of V as shown in $(A-1)$ with $\mathbf{b}_1 = \mathbf{b}_2^T = \mathbf{b}$. Then $c - \mathbf{b}^T\mathbf{A}^{-1}\mathbf{b} > 0$.

Proof: Note that

$$\det(\mathbf{V}) = \det(\mathbf{A})\det(c - \mathbf{b}^T\mathbf{A}^{-1}\mathbf{b}).$$

Therefore,

$$(c - \mathbf{b}^T\mathbf{A}^{-1}\mathbf{b}) = \det(c - \mathbf{b}^T\mathbf{A}^{-1}\mathbf{b}) = \frac{\det(\mathbf{V})}{\det(\mathbf{A})}.$$

Since $\mathbf{V} > 0$, $\det(\mathbf{V}) > 0$ and $\det(\mathbf{A}) > 0$, and, thus, $(c - \mathbf{b}^T\mathbf{A}^{-1}\mathbf{b}) > 0$. Q.E.D.

REFERENCES

[1] J. P. Norton, *An Introduction to Identification*. London: Academic, 1986.

[2] L. Lijun, *System Identification: Theory for the User*. Englewood Cliffs, NJ: Prentice-Hall, 1987.

[3] T. Södersträm and P. Stoica, *System Identification*. Englewood Cliffs, NJ: Prentice Hall, 1989.

[4] R. Haber and L. Keviczky, "Identification of nonlinear dynamic systems—survey paper," preprints, *4th IFAC Symp. Identific., Syst. Parameter Estimation*, Tbilisi, U.S.S.R., 1976, pp. 62–112.

[5] R. Haber and H. Unbehauen, "Structure identification of nonlinear dynamic systems—a survey on input-output approches," *Automatica*, vol. 26, pp. 651–677, 1990.

[6] I. J. Leontaritis and S. A. Billings, "Input-output parametric models for nonlinear systems, part I: Deterministic non-linear systems," *Int. J. Control*, vol. 41, no. 2, pp. 303–328, 1985.

[7] ——, "Input-output parametric models for non-linear systems, part II: Stochastic non-linear systems," *Int. J. Control*, vol. 41, no. 2, pp. 329–344, 1985.

[8] S. A. Billings and S. Y. Fakhouri, "Identification of a class of nonlinear systems using correlation analysis," *Proc. IEE*, vol. 125, pp. 691–697, 1978.

[9] ——, "Theory of separable processes with applications to the identification of nonlinear systems," *Proc. IEE*, vol. 125, pp. 1051–1057, 1978.

[10] ——, "Identification of systems containing linear dynamic and static nonlinear elements," *Automatica*, vol. 18, pp. 15–26, 1982.

[11] M. J. Korenberg, "Identifying noisy cascades of linear and static nonlinear systems," in *Proc. 7th IFAC Symp. on Identific., Syst. Parameter Estimation*, York, U.K., 1985, pp. 421–426.

[12] M. Korenberg, S. A. Billings, Y. P. Liu, and P. J. McIlroy, "Orthogonal parameter estimation algorithm for non-linear stochastic systems," *Int. J. Control*, 1988, vol. 48, no. 1, pp. 193–210, 1988.

[13] I. S. Durgaryan and F. F. Pashchenko, "Non-parametric identification of nonlinear systems," *IFAC Identific., Syst. Parameter Estimation*, vol. 1, no. 7, pp. 433–437, 1985.

[14] S. D. Conte and C. Boor, *Elementary Numerical Analysis: An Algorithmic Approach*, 3rd ed. New York: McGraw-Hill, 1980.

[15] G. A. F. Seber and C. J. Wild, *Nonlinear Regression*. New York: Wiley, 1989.

[16] R. L. Eubank, *Spline Smoothing and Nonparametric Regression*. New York: Marcel Dekker, 1988.

[17] L. Györfi, W. Härdle, P. Sarda, and P. Vieu, *Nonparametric Curve Estimation from Time Series*. Berlin: Springer-Verlag, 1989.

[18] J. S. Bendat, *Nonlinear System Analysis and Identification*. New York: Wiley, 1990.

[19] W. Härdle, *Applied Nonparametric Regression*. Cambridge, UK: Cambridge Univ. Press, 1990.

[20] N. S. Rajbman and A. T. Terekhin, "Variance ratio methods for random processes and their application to the analysis of nonlinear control plants," *Automat., Remote Control*, vol. 35, no. 3, pp. 496–506, 1965.

[21] N. S. Rajbman and V. M. Chadeev, *Identification of Industrial Processes*. Amsterdam: North-Holland, 1980.

[22] H. E. Liao and W. A. Sethares, "On nonparametric nonlinear system identification using dispersion function," in *IEEE 36th Midwest Symp. Circuits Syst.*, Detroit, MI, Aug. 1993, pp. 978–982.

[23] P. Eykhoff, Ed., *Trends and Progress in System Identification*, IFAC Series for Graduates, Research Workers, and Practicing Engineers, vol. 1, pp. 189–195, 1981.

[24] K. Chung, *A Course in Probability Theory*, 2nd ed. New York: Academic, 1974.

[25] W. Greblicki and M. Pawlak, "Identification of discrete Hammerstein system using kernel regression estimates," *IEEE Trans. Automat. Control*, vol. AC-31, pp. 74–77, Jan. 1986.

[26] ——, "Non-parametric identification of Hammerstein systems," *IEEE Trans. Inform. Theory*, vol. 45, pp. 409–418, Mar. 1989.

[27] W. Greblicki, "Non-parametric orthogonal series identification of Hammerstein systems," *Int. J. Syst. Sci.*, vol. 20, pp. 2355–2367, 1989.

[28] ——, "Nonparametric identification of Wiener systems," *IEEE Trans. Inform. Theory*, vol. 38, no. 5, pp. 1487–1493, Sept. 1992.

Ho-En Liao received the B.A. degree in electronic engineering from Chung Yuan Christian University, Taiwan, in 1984, and the M.S. and Ph.D. degrees in electrical engineering from the University of Wisconsin, Madison, in 1989 and 1993, respectively.

He is currently an Associate Professor at the Department of Electrical Engineering at Feng Chia University, Taiwan. His research interests include system identification, DS, adaptive filters, and digital communications.

W. A. Sethares received the B.A. degree in mathematics from Brandeis University, and the M.S. and Ph.D. degrees from Cornell University.

He has been a Systems Engineer at Raytheon Company, and is currently with the Department of Electrical and Computer Engineering at the University of Wisconsin, Madison. His research interests include adaptive systems in signal processing, acoustics, and communications.