Best Causal Mathematical Models for a Nonlinear System

Anatoli Torokhti, Phil Howlett, and Charles Pearce

***Abstract*—We provide new causal mathematical models of a nonlinear system $\mathcal{S}$ which are specifications of a nonlinear operator $\mathcal{P}_p$ of degree $p = 1, 2, \ldots$. The operator $\mathcal{P}_p$ is determined from a special orthogonalization procedure and minimization of the mean squared difference between outputs of $\mathcal{S}$ and $\mathcal{P}_p$. As a result, these models have smallest possible associated errors in the class of such operators $\mathcal{P}_p$. The causality condition is implemented through the use of specific matrices called lower trapezoidal. The associated computational work is reduced by the use of the orthogonalization procedure. We provide a strict justification of the proposed approach including theorems on an explicit representaton of the models' parameters, and theorems on the associated error representation. The possible extensions of the proposed approach and its potential applications are outlined.**

***Index Terms*—Causality, input–output map, nonlinear systems.**

I. Introduction

A. Previous Studies

CAUSALITY is an integral feature of physically realizable systems. Approaches to understanding, explanation, and formalization of physically realizable systems, and their various constructive models can be found in a number of previous works beginning with a classical paper by Russel [1]. In particular, Jones [2], Suppes [3], Petrović [4] and Verhaegen [5] related causality to both determinstic and stochastic dynamic systems. De Santis [6], Porter [7], and Bertuzzi *et al.* [8] studied causal polynomial approximation for input–output maps in Hilbert spaces. In a series of papers by Sandberg (see, for example, [9]–[12] and the bibliographies therein), new effective models of causal systems have been proposed and justified. In [13], [14], some new concepts of causality have been considered and applied to the analysis of nonlinear systems.

We note that known causal models have been developed mainly for the case of nonlinear system approximation with *any pre-assigned* accuracy.

An alternative direction of research in causal system theory has been proposed by Bode and Shannon in [15]. The approach [15] concerns optimization of causal *linear* systems. Ruzhansky and Fomin [16] extended the result [15] to the case of minimization of a cost functional with the arbitrary nonnegative weight matrix.

B. Contribution

In this paper, we propose causal mathematical models of *nonlinear* systems with smallest associated error. In general, our approach is based on the development of ideas from [15]–[19].

The differences from the known techniques are as follows. The presented models will follow from solutions to the *best approximation* problem (16), (17), given in Section IV, while the models in [13], [14] follow from the solution of problems for the input-output map approximation with *any pre-assigned* accuracy. The statements of the problem (16) and (17) and those in [17], [18] are different. As a result, the solutions are different. Unlike the known methods [15], [16], [19], we propose the *nonlinear* causal approximator of an arbitrary degree p [see (3) and (5)]. An increase of p implies the improvement at the accuracy in comparison with the models [15], [16], [19] (see Theorem 2, Corollary 1 and Remark 2 in Section V). Thus, the proposed model is equipped with a degree of freedom which is the degree p of the approximator. In contrast to the approach in [20], the presented method relates to so-called direct methods while the model in [20] is iterative. Besides, the model in [20] is not causal.

Thus, the novelty of the proposed approach consists of the new model of the system (Section III) based on the extension of the results in [13]–[19], and a new technique for the establishing its associated properties (Section V).

In particular, the proposed model implies an orthogonalization procedure presented in Section V-A1.

The general model $\mathcal{P}_p$ of the system, given in Section III, is determined by the sequences of operators. We present the constructive specification of $\mathcal{P}_p$ based on the special forms of the operators which compose $\mathcal{P}_p$. Section IV contains the rigorous statement of the problem. In Section V, we provide the determination of the parameters which define the optimal model $\mathcal{P}_p^0$. In particular, we propose and justify the orthogonalization procedure aimed at reducing the computational work associated with the optimal choice of $\mathcal{P}_p$. The representation of the error associated with $\mathcal{P}_p^0$ is also given in that section. The possible extensions of the proposed approach and its potential applications are disccused in Section VI.

II. (δ, ε)-Causality

Let (Ω, Σ, μ) be a probability space, where $\Omega = \{\omega\}$ is the set of outcomes, Σ a σ-field of measurable subsets of Ω and $\mu : \Sigma \mapsto [0,1]$ an associated probability measure on Σ with $\mu(\Omega) = 1$.

Let $\boldsymbol{\xi}, \boldsymbol{\zeta}, \boldsymbol{\vartheta} : T \times \Omega \rightarrow \mathbb{R}$ where $T = \{t_k, k = 1, \ldots, m | t_1 \leq \ldots \leq t_m\} \subset \mathbb{R}$ a collection of time instants. We write $\mathbf{x}_k =$

Manuscript received December 8, 2001; revised November 2, 2004. This paper was recommended by Associate Editor G. Setti.

A. Torokhti and P. Howlett are with the Centre for Industrial and Applicable Mathematics, The University of South Australia, Mawson Lakes, SA 5095, Australia (e-mail: anatoli.torokhti@unisa.edu.au; phil.howlett@unisa.edu.au).

C. Pearce is with the Applied Mathematics Department, The University of Adelaide, Adelaide, SA 5001, Australia (e-mail: cpearce@maths.adelaide.edu.au).

Digital Object Identifier 10.1109/TCSI.2005.846673

$\boldsymbol{\xi}(t_k,\cdot)$, $\mathbf{y}_k = \boldsymbol{\zeta}(t_k,\cdot)$ and $\mathbf{u}_k = \boldsymbol{\vartheta}(t_k,\cdot)$ so that $\mathbf{x}_k : \Omega \to \mathbb{R}$, $\mathbf{y}_k : \Omega \to \mathbb{R}$ and $\mathbf{u}_k : \Omega \to \mathbb{R}$.

Let $\mathbf{x} = (\mathbf{x}_1,\ldots,\mathbf{x}_m)^T$, $\mathbf{y} = (\mathbf{y}_1,\ldots,\mathbf{y}_m)^T$ and $\mathbf{u} = (\mathbf{u}_1,\ldots,\mathbf{u}_m)^T$, and let $\mathbf{x},\mathbf{y},\mathbf{u} \in L^2(\Omega,\mathbb{R}^m)$. Realizations of random vectors $\mathbf{x}$, $\mathbf{y}$ and $\mathbf{u}$ are denoted by $x = \mathbf{x}(\omega) \in \mathbb{R}^m$, $y = \mathbf{y}(\omega) \in \mathbb{R}^m$ and $u = \mathbf{u}(\omega) \in \mathbb{R}^m$, respectively.

As in [19], we interpret $\mathbf{x}$ as a given "idealized" input signal without any distortion, and $\mathbf{y}$ as an actual (observed) input signal. Vector $\mathbf{u}$ is treated as an output of the system. In particular, $\mathbf{y}$ can be interpreted as $\mathbf{x}$ contaminated with noise. No specific relationships between signal and noise are assumed to be known.

Each operator $S : \mathbb{R}^m \to \mathbb{R}^m$ defines an associated operator $\mathcal{S} : L^2(\Omega,\mathbb{R}^m) \to L^2(\Omega,\mathbb{R}^m)$ *via* the equation

$$[\mathcal{S}(\mathbf{x})](\omega) = S[\mathbf{x}(\omega)] \tag{1}$$

for each $\omega \in \Omega$. The operator S is interpreted here as the input–output map of the system.

Hereinafter, operators acting in spaces of random vectors are denoted by the calligraphic character letters. The terms "operator," "model," "input–output map," and "system" will be identified.

Let $\mathcal{P} : L^2(\Omega,\mathbb{R}^m \times \mathbb{R}^m) \to L^2(\Omega,\mathbb{R}^m)$ be a model for $\mathcal{S}$ with the associated operator $P : \mathbb{R}^m \times \mathbb{R}^m \to \mathbb{R}^m$ such that $u = P(y,u) = \begin{bmatrix} f_1(\tau_1(y,u)) \\ \cdots \\ f_m(\tau_m(y,u)) \end{bmatrix}$, where $\tau_k : \mathbb{R}^m \times \mathbb{R}^m \to \mathbb{R}^k \times \mathbb{R}^k$ and $f_k : \mathbb{R}^k \times \mathbb{R}^k \to \mathbb{R}$. If

$$\tau_k(y,u) = (y_1,\ldots,y_k,u_1,\ldots,u_k)^T \tag{2}$$

for all $k = 1,\ldots,m$ then the operator P is called causal.

This definition can equivalently be written as follows.

Definition 1: Let τ_k be defiend by (2). The operator P is called causal if for any $\hat{y}$, $\hat{u}$, $\check{y}$ and $\check{u}$

$$\tau_k(\hat{y},\hat{u}) = \tau_k(\check{y},\check{u}) \quad \Rightarrow \quad f_k(\hat{y},\hat{u}) = f_k(\check{y},\check{u}).$$

In many real problems, information is often obtained with some error, caused by the influence of external factors, data and instrument inexactness, etc. In this sense, the definition above is rather idealistic. A more realistic definition of causality for the operator P is as follows.

Definition 2: Let τ_k be defiend by (2). The operator P is called (δ,ε)-causal if for any $\delta \geq 0$ there exists $\varepsilon \geq 0$ such that for arbitrary $\hat{y}$, $\hat{u}$, $\check{y}$ and $\check{u}$

$$\|\tau_k(\hat{y},\hat{u}) - \tau_k(\check{y},\check{u})\|^2 \leq \delta \quad \Rightarrow \quad |f_k(\hat{y},\hat{u}) - f_k(\check{y},\check{u})| \leq \varepsilon$$

where $k = 1,\ldots,m$.

It is clear that the (0, 0)-causal operator is causal in the sense of Definition 1.

The set of (δ,ε)-causal systems is denoted by $\mathbb{C}_{\delta,\varepsilon}$.

III. Model of System

1) Preliminary Formulation of the Problem: We wish to find a mathematical model of the system which possesses the following properties. First, the model should be defined constructively, i.e., in an algorithmical form. Second, the model should be causal. Third, the model should approximate the system with the best possible accuracy. Fourth, the model should have some degrees of freedom to adjust to associated conditions such as a computational cost and a desirable accuracy of representation.

To pose the problem in the rigorous form, we need some preparatory work which we present in this section. The rigorous statement of the problem is given in Section IV.

2) General Model: We begin with a general representation of the system model.

An idea is to represent the model as a sum of composition of the operators $\mathcal{A}_q$, $\mathcal{B}_q$ and $\mathcal{C}_q$ with $q = 1,\ldots,p$. The operators $\mathcal{A}_q$, $\mathcal{B}_q$ and $\mathcal{C}_q$ are introduced in (5) to satisfy the conditions which are given in Section IV. In particular, the operators $\mathcal{A}_1,\ldots,\mathcal{A}_p$ are to minimize the related mean squared error. The operators $\mathcal{B},\ldots,\mathcal{B}_p$ are introduced to reduce the associated computational work by implementing the orthogonalization procedure. The operators $\mathcal{C}_1,\ldots,\mathcal{C}_p$ are to specify a transformation of vectors to a form suitable for computation. In particular, $\mathcal{C}_1,\ldots,\mathcal{C}_p$, given by (7) , reduce the model to a Volterra-like polynomial form (9), (10). An alternative choice for $\mathcal{C}_1,\ldots,\mathcal{C}_p$ is considered in Section VI. It will be shown in Section III-A3 that the proposed model is reduced to the form (12), where $\mathcal{T}_i$ and $\mathcal{G}_i$ are derived from operators $\mathcal{A}_q$ and $\mathcal{B}_q$. The operators $\mathcal{T}_i$ and $\mathcal{G}_i$ will be determined in Section V. We note that the model (12) requires associated derivations and cannot be introduced straight away.

We denote by $\mathbb{W}$ some set of vectors and write $\mathcal{L}_{\mathbb{R}^m} = L^2(\Omega,\mathbb{R}^m)$. Let

$$\mathcal{V} : \mathcal{L}_{\mathbb{R}^m} \to \mathcal{L}_{\mathbb{R}^m} \quad \mathcal{C}_q : \mathcal{L}_{\mathbb{R}^m} \times \mathcal{L}_{\mathbb{R}^m} \to \mathcal{L}_{\mathbb{W}}$$
$$\mathcal{B}_q : \mathcal{L}_{\mathbb{W}} \to \mathcal{L}_{\mathbb{R}^m} \quad \text{and} \quad \mathcal{A}_q : \mathcal{L}_{\mathbb{R}^m} \to \mathcal{L}_{\mathbb{R}^m}$$

for $q = 1,\ldots,p$ with $p \in \mathbb{N}$. Let

$$\mathbf{u} = \mathcal{P}_p(\mathbf{y},\mathbf{u}) \tag{3}$$

where $[\mathcal{P}_p(\mathbf{y},\mathbf{u})](\omega) = P_p(y,u)$ with

$$\mathcal{P}_p(\mathbf{y},\mathbf{u}) = a + \sum_{q=1}^{p} \mathcal{A}_q\mathcal{B}_q\mathcal{C}_q(\mathbf{y},\mathbf{g}) \tag{4}$$

$$P_p(y,u) = a + \sum_{q=1}^{p} A_qB_qC_q(y,g) \tag{5}$$

and where $\mathbf{g} = \mathcal{V}(\mathbf{u})$ and operators $\mathcal{V}$, $\mathcal{A}_q$, $\mathcal{B}_q$, and $\mathcal{C}_q$ are determined by the equations which are similar to (1), with $u = \mathbf{u}(\omega)$, $g = \mathbf{g}(\omega)$, $V : \mathbb{R}^m \to \mathbb{R}^m$, $C_q : \mathbb{R}^m \times \mathbb{R}^m \to \mathbb{W}$, $B_q : \mathbb{W} \to \mathbb{R}^m$ and $A_q : \mathbb{R}^m \to \mathbb{R}^m$.

We consider $\mathcal{P}_p$ as the model of the system $\mathcal{S}$.

The model associated with the operator$\mathcal{P}_p$ is represented in Fig. 1. We note that the representation (5) is motivated by the known structure of the p-degree approximator studied, for example, in [7], [8], [13], [14]. Indeed, if we choose $\mathbb{W} = (\mathbb{R}^m)^q$, denote $\mathbf{v}^q = \mathcal{C}_q(\mathbf{y},\mathbf{g})$ with $\mathbf{v}^q \in L^2(\Omega,(\mathbb{R}^m)^q)$, put $\mathcal{B}_q$ a q-linear operator and $\mathcal{A}_q$ the identity, then

$$\mathcal{P}_p(\mathbf{y},\mathbf{u}) = a + \sum_{q=1}^{p} \mathcal{B}_q(\mathbf{v}^q). \tag{6}$$

Such a model has been exploited in a number of works, in particular, in [7], [8], [13], [14]. At the same time, despite the natural interpretation of $\mathcal{P}_p$ in the form (6), the model (5) is not suitable

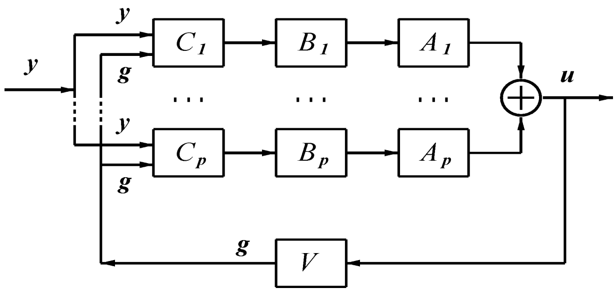

Fig. 1. Representation of the operator $\mathcal{P}_p$.

for computation. Next, we show that, on the basis of Lemma 1, the model (5) is reduced to the computationally adjusted form (12).

3) Specification of $\mathcal{P}_p$: The next step in our preparatory work is a specialization of operators $\mathcal{A}_q$, $\mathcal{B}_q$, and $\mathcal{C}_q$, and the related constructive representation of the operator $\mathcal{P}_p$.

We note that different specifications of $\mathcal{A}_q$, $\mathcal{B}_q$ and $\mathcal{C}_q$ define different forms of $\mathcal{P}_p$ in (5). Here, we consider the case when $V = I$ with I the identity operator,[1] $\mathbb{W} = (\mathbb{R}^m \times \mathbb{R}^m)^q$, $B_q : (\mathbb{R}^m \times \mathbb{R}^m)^q \to \mathbb{R}^m$ is a $2q$-linear operator and C_q is given by

$$C_q(y,u) = (y,u)^q \tag{7}$$

where

$$(y,u)^q = ((y,u),\ldots,(y,u)) \in (\mathbb{R}^m \times \mathbb{R}^m)^q.$$

For such V and C_q, the model (5) is reduced to

$$\mathcal{P}_p(\mathbf{y},\mathbf{u}) = a + \sum_{q=1}^{p} \mathcal{A}_q \mathcal{B}_q(\mathbf{y},\mathbf{u})^q. \tag{8}$$

In Lemma 1, we show that the choice of $\mathcal{C}_q$ in the form (7) implies a Volterra-like polynomial representaion for $\mathcal{P}_p$. In the Section VI-A5, we consider a different form for $\mathcal{C}_q$ which implies a Fourier-like polynomial representaion for $\mathcal{P}_p$. Some other possible forms for $\mathcal{A}_q$ and $\mathcal{B}_q$ are also considered in Section VI-A5.

We denote $v_q = y_{j_1}\ldots y_{j_q} u_{k_1}\ldots u_{k_{q-1}} u$ and $z_q = y_{j_1}\ldots y_{j_{q-1}} u_{k_1}\ldots u_{k_q} y$.

Lemma 1: Let $A_q \in \mathbb{R}^{m\times m}$ and let B_q be a q-linear operator. There exist matrices $D_{j_1}, D_{j_1,\ldots,j_q,k_1,\ldots,k_{q-1}} \in \mathbb{R}^{m\times m}$ and $F_{k_1}, F_{j_1,\ldots,j_{q-1},k_1,\ldots,k_q} \in \mathbb{R}^{m\times m}$ such that

$$\begin{aligned} P_p(y,u) &= a + A_1 \sum_{j_1=1}^{m} D_{j_1} y_{j_1} u + \sum_{q=2}^{p} A_q \sum_{j_1=1}^{m} \ldots \\ &\quad \sum_{j_q=1}^{m} \sum_{k_1=1}^{m} \ldots \sum_{k_{q-1}=1}^{m} D_{j_1,\ldots,j_q,k_1,\ldots,k_{q-1}} v_q \qquad (9) \\ &= a + A_1 \sum_{k_1=1}^{m} F_{k_1} u_{k_1} y + \sum_{q=2}^{p} A_q \sum_{j_1=1}^{m} \ldots \\ &\quad \sum_{j_{q-1}=1}^{m} \sum_{k_1=1}^{m} \ldots \sum_{k_q=1}^{m} F_{j_1,\ldots,j_{q-1},k_1,\ldots,k_q} z_q. \qquad (10) \end{aligned}$$

[1]The case when V is the zero operator will follow directly from our next derivations.

Proof: The proof is given in Appendix. □

The representations (9) and (10) are equivalent. Without loss of generality, we shall now consider $P_p(y,u)$ given by (9) only.

We denote $W_{j_1,\ldots,j_q,k_1,\ldots,k_{q-1}} = A_q D_{j_1,\ldots,j_q,k_1,\ldots,k_{q-1}}$ and $W_{j_1,\ldots,j_q,k_1,\ldots,k_{q-1}} = \{w_{kl}\}$ with $k,l = 1,\ldots,m$ and $q = 2,\ldots,p$.

To define a structure of the (δ,ε)-causal model P_p, we need the following definition.

Definition 3: The matrix $W_{j_1,\ldots,j_q,k_1,\ldots,k_{q-1}}$ is called i-lower trapezoidal if $w_{kl} = 0$ for $l = 1,\ldots,m$ and $k = 1,\ldots,i-1$ where $i = \max\{j_1,\ldots,j_{q-1},k_1,\ldots,k_{q-1}\}$, and also $w_{kl} = 0$ for $k = i, i+1,\ldots,m$ and $k < l$.

In particular, the i-lower trapezoidal matrix $W_{j_1} = A_q D_{j_1} = \{w_{kl}\} \in \mathbb{R}^{m\times m}$ is defined as above, and in this case $i = j_1$.

For example, a 3 × 3 matrix $\{w_{kl}\}$ is 2-lower trapezoidal if $w_{1l} = 0$ for $l = 1, 2, 3$, and $w_{23} = 0$, i.e., if it has the form

$$\begin{bmatrix} 0 & 0 & 0 \\ \bullet & \bullet & 0 \\ \bullet & \bullet & \bullet \end{bmatrix}$$

where "•" means the entry which is not necessarily zero.

The following statement establishes the structure of the (δ,ε)-causal model P_p.

Proposition 1: If matrices W_{j_1} and $W_{j_1,\ldots,j_q,k_1,\ldots,k_{q-1}}$ are i-lower trapezoidal then P_p is (δ,ε)-causal.

Proof: The proof follows directly from the above definitions. □

Next stage in this section is to reduce the model P_p to a more compact form. We do this as follows.

If we consider W_{j_1} and $W_{j_1,\ldots,j_q,k_1,\ldots,k_{q-1}}$ for each combination of $j_1,\ldots,j_q$ and $k_1,\ldots,k_{q-1}$, then for each $q = 1,\ldots,p$ we obtain m^{2q-1} matrices. For $q = 1,\ldots p$, there are

$$N_{yu} = m + \sum_{i=2}^{p} m^{2i-1} = \sum_{i=1}^{p} m^{2i-1} \tag{11}$$

matrices altogether and if we denote them by $T_1G_1,\ldots,T_{N_{yu}}G_{N_{yu}}$, where $T_i \in \mathbb{R}^{m\times m}$, $G_i \in \mathbb{R}^{m\times m}$, with the corresponding operands denoted by $w_1,\ldots,w_{N_{yu}} \in \mathbb{R}^m$ then we can write

$$\mathcal{P}_p(\mathbf{y},\mathbf{u}) = a + \sum_{i=1}^{N_{yu}} \mathcal{T}_i \mathcal{G}_i(\mathbf{w}_i) \text{ and } P_p(y,v) = a + \sum_{i=1}^{N_{yu}} T_i G_i w_i \tag{12}$$

where $\mathcal{G}_i, \mathcal{T}_i : \mathcal{L}_{\mathbb{R}^m} \to \mathcal{L}_{\mathbb{R}^m}$ and $[\mathcal{G}_i(\mathbf{w}_i)](\omega) = G_i w_i$ with $w_i = \mathbf{w}_i(\omega)$.

Now we are in a position to pose the problem rigorously.

IV. Rigorous Statement of Problem

Without loss of generality, we assume that all random vectors have zero mean.

Let

$$\begin{aligned} \mathbf{h}_i &= \mathcal{G}_i(\mathbf{w}_i) \\ \mathbf{h} &= (\mathbf{h}_1,\ldots,\mathbf{h}_{N_{yu}}) \\ h &= \mathbf{h}(\omega) \\ \mathbf{w} &= (\mathbf{w}_1,\ldots,\mathbf{w}_{N_{yu}}) \\ w &= \mathbf{w}(\omega). \end{aligned} \tag{13}$$

We write

$$J = J\left(a, \mathcal{T}_1, \ldots, \mathcal{T}_{n_{yu}}\right) = E\left[\|\mathcal{S}(\mathbf{x}) - \mathcal{P}_p(\mathbf{y},\mathbf{u})\|^2\right] = E\left[\left\|\mathcal{S}(\mathbf{x}) - a - \sum_{i=1}^{N_{yu}} \mathcal{T}_i(\mathbf{h}_i)\right\|^2\right] \quad (14)$$

where E is the expectation operator, $\|\cdot\|$ is the Euclidean norm and each T_i is i-lower trapezoidal. The latter condition is essential for the next derivations.

It is assumed that a structure of the input-output map $\mathcal{S}$ of the system is unknown, and that an information on the system is given by certain covariance matrices formed from $\mathcal{S}(\mathbf{x})$, $\mathbf{y}$ and $\mathbf{u}$. Such an assumption is traditionally used in the problems dealing with transformation of stochastic signals [15], [16], [18]–[25]. The methods for the estimation of covariance matrices can be found, for example, in [26]–[33]. This estimation technique is an area of special study and is not a subject of this paper.

We write $\mathbb{O}_{m\times m}$ for the $m \times m$ zero matrix and denote $E_{h_i h_j} = E[\mathbf{h}_i \mathbf{h}_j^T]$.

The problem is to find $a^0, \mathcal{G}_1, \ldots, \mathcal{G}_{N_{yu}}, \mathcal{T}_1^0, \ldots, \mathcal{T}_{n_{yu}}^0$ such that

$$E_{h_i h_j} = \mathbb{O}_{m\times m}, \quad \text{for} \quad i \neq j, \quad i,j = 1,\ldots,N_{yu} \quad (15)$$

$$J\left(a^0, \mathcal{T}_1^0, \ldots, \mathcal{T}_{N_{yu}}^0\right) = \min_{a,\mathcal{T}_1,\ldots,\mathcal{T}_{N_{yu}}} J\left(a, \mathcal{T}_1, \ldots, \mathcal{T}_{N_{yu}}\right) \quad (16)$$

subject to

$$\mathcal{P}_p^0 \in \mathbb{C}_{\delta,\varepsilon} \quad (17)$$

where

$$\mathcal{P}_p^0(\mathbf{y},\mathbf{u}) = a^0 + \sum_{i=1}^{N_{yu}} \mathcal{T}_i^0(\mathbf{h}_i). \quad (18)$$

The condition (15) will allow us to simplify and reduce the computational work needed to determine $\mathcal{T}_1^0, \ldots, \mathcal{T}_{n_{yu}}^0$ (see Sections IV and V).

The operator $\mathcal{P}_p^0$ is called the (δ,ε)-causal optimal model of the system $\mathcal{S}$.

V. Determination of (δ,ε)-Causal Optimal Model $\mathcal{P}_p^0$

1) Generic Scheme: Let $E_{h_k h_k}^{\dagger}$ be the Moore–Penrose pseudo-inverse of $E_{h_k h_k}$.

Lemma 2: The operators $\mathcal{G}_1, \ldots, \mathcal{G}_{N_{yu}}$, which satisfy (15), are given by

$$\mathcal{G}_1 = I \quad \text{and} \quad \mathcal{G}_i(\mathbf{w}_i) = \mathbf{w}_i - \sum_{k=1}^{i-1} \mathcal{H}_{ik}\mathcal{G}_k(\mathbf{w}_k) \quad (19)$$

where $i = 2, \ldots, N_{yu}$ and $\mathcal{H}_{ik} : \mathcal{L}_{\mathbb{R}^m} \to \mathcal{L}_{\mathbb{R}^m}$ is defined by

$$[\mathcal{H}_{ik}(\mathbf{h}_k)](\omega) = H_{ik} h_k \quad (20)$$

and

$$H_{ik} = E_{w_i h_k} E_{h_k h_k}^{\dagger} + K_{ik}\left(I - E_{h_k h_k} E_{h_k h_k}^{\dagger}\right) \quad (21)$$

with $H_{jk} \in \mathbb{R}^{m\times m}$, $h_k = \mathbf{h}_k(\omega)$ and $K_{ik} \in \mathbb{R}^{m\times m}$ arbitrary.

Proof: Let $\mathbf{h}_1 = \mathbf{w}_1$ and $\mathbf{h}_i = \mathbf{w}_i - \sum_{k=1}^{i-1} \mathcal{H}_{ik}(\mathbf{h}_k)$ for $i = 2, \ldots, N_{yu}$. The required condition (15) implies

$$E\left[\left(\mathbf{w}_i - \sum_{k=1}^{i-1} \mathcal{H}_{ik}(\mathbf{h}_k)\right)\mathbf{h}_k^T\right] = \mathbb{O}_{m\times m} \text{ and } H_{ik}E_{h_k h_k} = E_{w_i h_k}.$$

The solution to the latter equation is given by (21). Thus, the desirable $\mathcal{G}_1, \ldots, \mathcal{G}_{N_{yu}}$ are given by (19). □

A possible and natural choice for K_{ik} is $K_{ik} = \mathbb{O}_{m\times m}$.

Example 1: Let $p = 1$. By (34) and (12), $P_1(y,u) = a + \sum_{i=1}^m A_1 D_{j_1} y_{j_1} u = a + \sum_{i=1}^m T_i G_i w_i$ with $w_i = y_i u$ and $i = 1, \ldots, m$, and therefore h_i (with $i = 2, \ldots, m$) is determined by $E_{w_i w_k}$ and $E_{w_k w_k}$ for $k = 1, \ldots, i-1$. The matrices $E_{w_i w_k}$ and $E_{w_k w_k}$ are assumed to be estimated by the techniques given, in particular, in [26]–[33]. Of course, for such estimates, it is not necessary to suppose that u itself is known. To estimate $E_{w_i w_k}$ and $E_{w_k w_k}$ by [26]–[33], one can set, for example, $u = s = S(x)$ or $u = y$. Then $E_{w_i w_k} = E[\mathbf{y}_i \mathbf{s}\mathbf{s}^T \mathbf{y}_k]$ or $E_{w_i w_k} = E[\mathbf{y}_i \mathbf{y}\mathbf{y}^T \mathbf{y}_k]$, where $\mathbf{s} = \mathcal{S}(\mathbf{x})$. As a result, the assumption made in Section III can be reduced to the assumption on the knowledge of covariance matrices formed from $\mathcal{S}(\mathbf{x})$ and $\mathbf{y}$.

We note that $u = s$ (or $u = y$) can be used to specify the associated covariance matrices only. In general, u is determined by $u = P_p(y,u)$ and can be obtained numerically (see [34, pp. 66–67], and Example 2).

Theorem 1: Let $\mathcal{G}_1, \ldots, \mathcal{G}_{N_{yu}}$ be defined by Lemma 2. Let $E_{h_i h_i}$ be positive definite and let

$$E_{h_i h_i} = L_i L_i^T$$

the Cholesky factorization for $E_{h_i h_i}$ with L_i lower triangular. Let $\mathbf{s} = \mathcal{S}(\mathbf{x})$ and

$$E_{s h_i} L_i^{-T} = M_{1i} + M_{2i} + M_{3i}, \quad (22)$$

where M_{1i} is strictly upper triangular (i.e., with the zero entries on the main diagonal), M_{2i} is i-lower trapezoidal and M_{3i} supplements M_{2i} to form a lower triangular matrix. Then the (δ,ε)-causal optimal model $\mathcal{P}_p^0$ is given by

$$a^0 = \mathbb{O}_{m\times 1} \quad \text{and} \quad T_i^0 = M_{2i} L_i^{-1} \quad (23)$$

with $i = 1, \ldots, N_{yu}$.

Proof: Let us denote $\tilde{T} = [T_1 \ldots T_{N_{yu}}]$ and $h = [h_1^T \ldots h_{N_{yu}}^T]^T$. By the assumption, $E_{h_i h_i}$ is positive definite therefore $E_{h_i h_i}^{\dagger} = E_{h_i h_i}^{-1}$. Moreover, $\mathcal{G}_1, \ldots, \mathcal{G}_{N_{yu}}$ defined by Lemma 2 are such that matrix E_{hh} is block diagonal, i.e.,

$$E_{hh} = \begin{bmatrix} E_{h_1 h_1} & \mathbb{O} & \mathbb{O} & \ldots & \mathbb{O} & \mathbb{O} \\ \mathbb{O} & E_{h_2 h_2} & \mathbb{O} & \ldots & \mathbb{O} & \mathbb{O} \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ \mathbb{O} & \ldots & \ldots & \ldots & \mathbb{O} & E_{h_{N_{yu}} h_{N_{yu}}} \end{bmatrix}.$$

Then

$$J=\mathrm{tr}\{E_{ss}+aa^T\}+\mathrm{tr}\Big\{\Big(\tilde{T}-E_{sh}E_{hh}^{-1}\Big)E_{hh}\times\Big(\tilde{T}^T-E_{hh}^{-1}E_{hs}\Big)\Big\}$$
$$-\mathrm{tr}\{E_{sh}E_{hh}^{-1}E_{hs}\}$$
$$=\mathrm{tr}\left\{E_{ss}+aa^T-E_{sh}E_{hh}^{-1}E_{hs}\right\}$$
$$+\sum_{i=1}^{N_{yu}}\mathrm{tr}\left\{(T_iL_i-M_{1i}-M_{2i}-M_{3i})\right.$$
$$\left.\times\left(L_i^TT_i^T-M_{1i}^T-M_{2i}^T-M_{3i}^T\right)\right\}.$$

Here

$$E_{sh}E_{hh}^{-1}E_{hs}=\sum_{i=1}^{N_{yu}}E_{sh_i}E_{h_ih_i}^{-1}E_{h_is}$$
$$=\sum_{i=1}^{N_{yu}}E_{sh_i}L_i^{-T}\left(E_{sh_i}L_i^{-T}\right)^T$$
$$=\sum_{i=1}^{N_{yu}}(M_{1i}+M_{2i}+M_{3i})$$
$$\times(M_{1i}+M_{2i}+M_{3i})^T$$
$$=\sum_{i=1}^{N_{yu}}\left(M_{13i}M_{13i}^T+M_{13i}M_{2i}^T\right.$$
$$\left.+M_{2i}M_{13i}^T+M_{2i}M_{2i}^T\right)$$

where $M_{13i}=M_{1i}+M_{3i}$.

Next

$$\mathrm{tr}(T_iL_i-M_{1i}-M_{2i}-M_{3i})\left(L_i^TT_i^T-M_{1i}^T-M_{2i}^T-M_{3i}^T\right)$$
$$=\mathrm{tr}\left[(T_iL_i-M_{2i})-M_{13i}\right]\left[\left(L_i^TT_i^T-M_{2i}^T-M_{13i}^T\right]\right.$$
$$=\|T_iL_i-M_{2i}\|^2-\mathrm{tr}\left(T_iL_iM_{13i}^T+M_{13i}L_i^TT_i^T\right)$$
$$+\mathrm{tr}\left(M_{2i}M_{13i}^T+M_{13i}M_{2i}^T+M_{13i}M_{13i}^T\right)$$

and

$$\mathrm{tr}\left\{\left(T_iL_iM_{13i}^T+M_{13i}L_i^TT_i^T\right\}=0.\right.$$

The latter is true because T_i is i-lower trapezoidal. Therefore

$$J=\mathrm{tr}\{E_{ss}+aa^T\}-\sum_{i=1}^{N_{yu}}\mathrm{tr}\left\{M_{2i}M_{2i}^T\right\}+\sum_{i=1}^{N_{yu}}\|T_iL_i-M_{2i}\|^2.$$

Thus, $a=a^0$ and $T_i=T_i^0$ minimize J. Each matrix T_i^0 is i-lower trapezoidal therefore $\mathcal{P}_p^0\in\mathbb{C}_{\delta,\varepsilon}$. □

Example 2: Let $p=1$. Then

$$P_1(y,u)=a+\sum_{i=1}^{m}A_1D_{j_1}y_{j_1}u=a+\sum_{i=1}^{m}T_iG_iw_i$$

with $w_i=y_iu$, $i=1,\ldots,m$ and therefore by Theorem 1, $P_1^0(y,u)=\sum_{j_1=1}^{m}T_i^0G_iw_i$, i.e., $u=\sum_{j_1=1}^{m}T_i^0G_iy_iu$. The solution to the latter equation in terms of $T_i^0G_iy_i$ follows, for example, from the fixed point theorem [34, pp. 66–67]. In other words, the feedback model can be reduced to the nonfeedback one. □

Remark 1: The solution (23) to the problem (16) implies that the matrices E_{sh_i} and $E_{h_ih_i}$ are known. The methods of their estimation is based on the approach described in Example 1. □

Theorem 2: The error $\Delta_{p,y,u}$ associated with the (δ,ε)-causal model $\mathcal{P}_p^0$ is given by

$$\Delta_{p,y,u}=E\left[\left\|\mathcal{S}(\mathbf{x})-\mathcal{P}_p^0(\mathbf{y},\mathbf{u})\right\|^2\right]=\left\|E_{ss}^{\frac{1}{2}}\right\|^2-\sum_{i=1}^{N_{yu}}\|M_{2i}\|^2. \quad (24)$$

Proof: We have $\Delta_p=J(a^0,\mathcal{T}_1^0,\ldots,\mathcal{T}_{N_{yu}}^0)$ and then (24) follows from the above proof when $\mathcal{G}_1,\ldots,\mathcal{G}_{n_{yu}}$ are determined by (19)–(21), and a^0 and T_i^0 are substituted instead of a and T_i in the expression for J given by (14). □

2) Particular Case: $V=\mathbb{O}_{m\times m}$ in (3)–(5). Let us consider the model $\mathcal{P}_p$ without a feedback, i.e., when $V=\mathbb{O}_{m\times m}$ in (3)–(5). Then similarly to Lemma 1, it is proved that there exist matrices $F_{j_1\ldots j_{q-1}}$ such that

$$P_p(y)=a+A_1B_1y+\sum_{q=2}^{p}A_q\sum_{j_1=1}^{m}\cdots\sum_{j_{q-1}=1}^{m}F_{j_1,\ldots,j_{q-1}}f_{j_1,\ldots,j_{q-1}} \quad (25)$$

where we write $P_p(y)$ instead of $P_p(y,u)$ and where $f_{j_1,\ldots,j_{q-1}}=y_{j_1}\ldots y_{j_{q-1}}y$.

By analogy with (12), the latter is reduced to the representation

$$\mathcal{P}_p(\mathbf{y})=a+\sum_{i=1}^{N_y}\mathcal{T}_i\mathcal{G}_i(\mathbf{w}_i)\quad\Rightarrow\quad P_p(y)=a+\sum_{i=1}^{N_y}T_iG_iw_i. \quad (26)$$

Remark 2: Here, we use the same notation $\mathcal{G}_i$, $\mathcal{T}_i$, G_i, T_i, $\mathbf{w}_i$ and w_i as in (12) but for *different operators, matrices and vectors*, which are now constructed from $A_q, B_1, F_{j_1,\ldots,j_{q-1}}$ and $f_{j_1,\ldots,j_{q-1}}$, respectively. We also define matrix M_{2i} similar to that in (22) but with h_i determined from (13) and Lemma 2 for the case when $V=\mathbb{O}_{m\times m}$. Another difference from (12) is the number N_y of terms in (26) which is essentially smaller than N_{yu} in (11) and (12), namely

$$N_y=\sum_{i=1}^{p}m^{i-1}. \quad (27)$$

These conditions lead to the accuracy associated with the model, $\mathcal{P}_p^0$ as follows.

Corollary 1: The error $\Delta_{p,y}$ associated with the (δ,ε)-causal model $\mathcal{P}_p^0$ with $V=\mathbb{O}_{m\times m}$ is given by

$$\Delta_{p,y}=E\left[\left\|\mathcal{S}(\mathbf{x})-\mathcal{P}_p^0(\mathbf{y})\right\|^2\right]=\left\|E_{ss}^{\frac{1}{2}}\right\|^2-\sum_{i=1}^{N_y}\|M_{2i}\|^2. \quad (28)$$

Proof: The proof follows directly from the above. □

A comparison of (24) and (28), and (11) and (27) shows that the accuracy associated with the optimal feedback model (12) is better than that of the optimal nonfeedback model (26).

Remark 3: The known models [15], [16] follow from (26) if $p = 1$ (or $N_y = 1$) and $G_1 = I$, i.e if $P_p(y) = A_1 y$. The quadratic model [19], satisfying the causality condition (17), and its associated error follow from the above in a similar way. As we have mentioned in Section I-B, unlike the models [15], [16] and the described extension of the model [19], the proposed (δ, ε)-causal models (18) and (25) are equipped with the degree of freedom p. In particular, for $p = 2, 3, \ldots$, the errors given by (24) and (28) are less than the errors associated with the mentioned approximators [15], [16], [19]. In other words, it may occur that the accuracy of approaches [15], [16], [19] are not satisfactory but the accuracy of the proposed method can be improved due to increasing degree p. Another advantage of the models (18) and (25) is the orthogonalization procedure. An associated benefit is discussed in the next Section VI-A1.

VI. Discussion

A performance of a particular model of the system is characterized by related computational work and associated accuracy. In this Section, we discuss these two characteristics in relation to the proposed method.

1) Computational Work: The technique presented in Section IV provides the optimality and (δ, ε)-causality of the model $\mathcal{P}_p^0$. The specification of the operators $\mathcal{G}_1, \ldots, \mathcal{G}_{N_{yu}+1}$ by Lemma 2 allows us to reduce matrix E_{hh} to the block diagonal form. The latter leads to the representation of $T_1^0, \ldots, T_{N_{yu}+1}^0$ in the form of simple independent expressions given by (23). Computation of the matrix T_{j+1}^0 in (23) consists of the vector orthogonalization by Lemma 2 which requires $O(N_{yu} m^2)$ flops (since $E_{h_i h_i}^{\dagger} = E_{h_i h_i}^{-1}$), and the Cholesky decomposition which requires $O(m^3)$ flops. In total, $\mathcal{P}_p^0$ requires $O(N_{yu} m^3) = O(m^{2p+2})$ flops only. If $V = \mathbb{O}_{m\times m}$ in (3)–(5) then the nonfeedback $\mathcal{P}_p^0$ requires $O(m^{p+2})$ flops.

2) Associated Accuracy: Theorem 2 establishes that the accuracy of the model $\mathcal{P}_p^0$ increases with the increase in the number of terms N_{yu} in $\mathcal{P}_p^0$, and consequently, with the increase of the degree p of $\mathcal{P}_p^0$. Thus, one can regulate the accuracy by varying p. Next, the degree p of the model $\mathcal{P}_p^0$ can be varied depending on k, i.e., $p = p(k)$ where k follows from $u_k = f_k(\tau_k(y, u))$. An increase of $p(k)$ for a few initial values of k will improve the accuracy of the estimate.

3) Possible Extensions and Applications: The technique presented above can be extended for different types of operators $\mathcal{A}_q$, $\mathcal{B}_q$ and $\mathcal{C}_q$ in (5). In particular, let $\mathbb{W} = \mathbb{R}^m \times \mathbb{R}^m$ and let unlike (7), the operators $\mathcal{C}_1, \ldots, \mathcal{C}_p$ be chosen so that

$$\mathcal{C}_1(\mathbf{y}, \mathbf{u}) = (\mathbf{y}, \mathbf{u}), \ldots, \mathcal{C}_p(\mathbf{y}, \mathbf{u}) = \boldsymbol{cos}\left[(p-1)(\mathbf{y}, \mathbf{u})\right], \quad (29)$$

where

$$\left[\boldsymbol{cos}\left(k(\mathbf{y}, \mathbf{u})\right)\right](\omega) = \cos\left(k\left[\mathbf{y}(\omega), \mathbf{u}(\omega)\right]\right) \quad (30)$$

and

$$\cos\left(k\left[\mathbf{y}(\omega), \mathbf{u}(\omega)\right]\right) = \left[\cos(ky_1), \ldots, \cos(ky_m), \cos(ku_1), \ldots, \cos(ku_m)\right]^T \quad (31)$$

with $k = 1, \ldots, p-1$.

Second, in (5), we set $\mathcal{B}_k : L^2(\Omega, \mathbb{R}^m \times \mathbb{R}^m) \to L^2(\Omega, \mathbb{R}^m)$ linear for each $k = 1, \ldots, p$, not k-linear as in Lemma 1.

A motivation for such a choice of operators $\mathcal{C}_1, \ldots, \mathcal{C}_p$ and $\mathcal{B}_1, \ldots, \mathcal{B}_p$ follows from the observation that the model $\mathcal{P}_p^0$ with $\mathcal{C}_1, \ldots, \mathcal{C}_p$ defined by (7), requires computation of N_{yu} matrices $G_1, \ldots G_{N_{yu}}$ by Lemma 2 and N_{yu} matrices $T_1^0, \ldots T_{N_{yu}}^0$ by (23). The number N_{yu} is given by (11). The model $\mathcal{P}_p$ with $\mathcal{C}_1, \ldots, \mathcal{C}_p$ and $\mathcal{B}_1, \ldots, \mathcal{B}_p$ as above requires computation of only p matrices $B_1, \ldots, B_p$ and p matrices $A_1, \ldots, A_p$ with $p \ll N_{yu}$. The latter diminishes a related computational cost.

We note that the model $\mathcal{P}_p$ with $\mathcal{C}_1, \ldots, \mathcal{C}_p$ defined by (29)–(31), operators $\mathcal{B}_1, \ldots, \mathcal{B}_p$ determined similarly to those in Lemma 2, and operators $\mathcal{A}_1, \ldots, \mathcal{A}_p$ obtained from a solution to the minimization problem similar to (16), can be referred to as an operator generalization of the truncated Fourier series (i.e., the Fourier polynomial) in a separable Hilbert space [35]–[37]. Unlike the original Fourier polynomial [35]–[37], $\mathcal{A}_q$ and $\mathcal{B}_q$ are operators, not scalars.

Our preliminary investigations show that such a device leads first, to a reduction of associated computational work, and second, to a slight increase of the associated error, in comparison with $\mathcal{P}_p^0$ determined by (7), Lemmas 1, 2 and Theorem 1. In further work, we plan to investigate a compromise between accuracy and computational load in more detail.

An attractive specialization of the model (12) is based on a determination of operators $\mathcal{T}_1, \ldots, \mathcal{T}_{N_{yu}}$ from a solution of the interpolation problem similar to that in [38] instead of their determination from the solution of the minimization problem (14)–(16).

Another possibility for a modification of the proposed technique is based on a choice of $\mathcal{C}_1, \ldots, \mathcal{C}_p$ in the form of so-called partitioning operators [39]. In such a case, $\mathbf{y}$ and $\mathbf{u}$ are partitioned into 'shorter' subvectors. This leads to the reduction of associated computational cost.

An important extension follows from considering the problem (16) for $S = I$ subject to the restriction on the rank of operators $\mathcal{T}_1, \ldots, \mathcal{T}_{N_{yu}}$ in (14)–(16). The solution has a direct application to data compression and filtering [22], [24].

Other potential applications of the proposed technique include areas in target detection [23], a blind channel equalization problem [24], [40], combating speckle in SAR images [41] and pattern recognition [42].

4) Degrees of Freedom: It follows from the above that the performance of the model (5) can be varied by choosing $\mathcal{A}_q$, $\mathcal{B}_q$ and $\mathcal{C}_q$. Hence, the proposed model is flexible and has the degrees of freedom implied by $\mathcal{A}_q$, $\mathcal{B}_q$, $\mathcal{C}_q$ and p.

5) Particular Case: Filtering of Stochastic Signals: If $\mathcal{S} = I$ then (5)–(9), (12) and (23) represent the filter with $\mathbf{x}$ and $\mathbf{y}$ the reference signal and observed data, respectively.

VII. Conclusion

The approach proposed and justified in this paper provides models of a nonlinear system which are causal and optimal in the sense of minimizing the associated mean squared errors. The models are generated by the module given by (5) and are defined by the sequences of operators $\{\mathcal{A}_q\}$, $\{\mathcal{B}_q\}$ and $\{\mathcal{C}_q\}$ with $q = 1, 2, \ldots, p$. For the particular choice of operators

$\{\mathcal{A}_q\}$, $\{\mathcal{B}_q\}$ and $\{\mathcal{C}_q\}$ given in Section III-A3, the model is represented by the operator $\mathcal{P}_p$ given in (12). The computational implementation of $\mathcal{P}_p$ implies the orthogonalization procedure (19)–(21) and the solution of the minimization problem (23). The causality condition has been incorporated into the models through so-called i-lower trapezoidal matrices. Explicit equations for the errors associated with the models have been established. It has been shown that the models have a degree of freedom which is the degree p of the operator $\mathcal{P}_p$.

Possibilities for alternative determinations of operators $\{\mathcal{A}_q\}$, $\{\mathcal{B}_q\}$ and $\{\mathcal{C}_q\}$ have been discussed (Section VI).

APPENDIX

Proof: Let $\{e_1, \ldots, e_n\}$ be the standard basis in $\mathbb{R}^n$. Then

$$\begin{aligned} B_q(y,u)^q &= B_q(y,\ldots,y,u,\ldots,u) \\ &= B_q\left(\sum_{j_1=1}^m y_{j_1}e_{j_1},\ldots,\sum_{j_q=1}^m y_{j_q}e_{j_q},\right. \\ &\qquad \left.\sum_{k_1=1}^m u_{k_1}e_{k_1},\ldots,\sum_{k_q=1}^m u_{k_q}e_{k_q}\right) \\ &= \sum_{j_1=1}^m\cdots\sum_{j_q=1}^m\sum_{k_1=1}^m\cdots \\ &\quad \sum_{k_q=1}^m B_q\left(e_{j_1},\ldots,e_{j_q},e_{k_1},\ldots,e_{k_q}\right) \\ &\quad \times y_{j_1}\cdots y_{j_q}u_{k_1}\cdots u_{k_q} \end{aligned}$$

since B_q is the $2q$-linear operator.

Let us denote

$$B_q\left(e_{j_1},\ldots,e_{j_q},e_{k_1},\ldots,e_{k_q}\right) = b_{j_1,\ldots,j_q,k_1,\ldots,k_q} \in \mathbb{R}^m.$$

Then, we can define the matrices $D_{j_1} \in \mathbb{R}^{m\times m}$ and $D_{j_1,\ldots,j_q,k_1,\ldots,k_{q-1}} \in \mathbb{R}^{m\times m}$ by the formulas

$$D_{j_1}g = \sum_{k_q=1}^m b_{j_1,k_1}g_{k_1} \quad \text{and}$$

$$D_{j_1,\ldots,j_q,k_1,\ldots,k_{q-1}}g = \sum_{k_q=1}^m b_{j_1,\ldots,j_q,k_1,\ldots,k_q}g_{k_q}$$

respectively, where $g = (g_1,\ldots g_m)^T \in \mathbb{R}^s$. Therefore

$$\begin{aligned} B_1(y,u) &= \sum_{j_1=1}^m\sum_{k_1=1}^m B_q\left(e_{j_1},e_{k_1}\right)y_{j_1}u_{k_1} \\ &= \sum_{j_1=1}^m y_{j_1}\sum_{k_1=1}^m b_{j_1,k_1}u_{k_1} \\ &= \sum_{j_1=1}^m D_{j_1}y_{j_1}u \end{aligned} \tag{32}$$

and for $q = 2,\ldots p$,

$$\begin{aligned} &B_q(y,u)^q \\ &= \sum_{j_1=1}^m\cdots\sum_{j_q=1}^m\sum_{k_1=1}^m\cdots\sum_{k_{q-1}=1}^m y_{j_1}\cdots y_{j_q} \\ &\quad \times u_{k_1}\cdots u_{k_{q-1}}\left(\sum_{k_q=1}^s b_{j_1,\ldots,j_q,k_1,\ldots,k_q}u_{k_q}\right) \\ &= \sum_{j_1=1}^m\cdots\sum_{j_q=1}^m\sum_{k_1=1}^m\cdots\sum_{k_{q-1}=1}^m D_{j_1,\ldots,j_q,k_1,\ldots,k_{q-1}}y_{j_1}\cdots y_{j_q} \\ &\quad \times u_{k_1}\cdots u_{k_{q-1}}u, \end{aligned} \tag{33}$$

and then (9) follows.

The representation of $P_p(y,u)$ in form (10) follows from the above scheme in a similar way due to the symmetry of y and u in (34). □

Example 3: For $p = 1$, the formula (9) takes the form

$$P_1(y,u) = A^{(0)} + A_1B_1(y,u) = A^{(0)} + A_1\sum_{j_1=1}^m D_{j_1}y_{j_1}u, \tag{34}$$

where $B_1 = \{b_{i,j_1,j_2}\}_{i,j_1,j_2=1}^{i,j_1,j_2=m}$ is the $m\times m\times m$ tensor and $D_1 = \{b_{i,1,j_2}\}_{i,j_2=1}^{i,j_2=m} \in \mathbb{R}^{m\times m},\ldots,D_m = \{b_{i,m,j_2}\}_{i,j_2=1}^{i,j_2=m} \in \mathbb{R}^{m\times m}$ or $D_1 = \{b_{i,j_1,1}\}_{i,j_1=1}^{i,j_1=m} \in \mathbb{R}^{m\times m},\ldots,D_m = \{b_{i,j_1,m}\}_{i,j_1=1}^{i,j_1=m} \in \mathbb{R}^{m\times m}$. □

ACKNOWLEDGMENT

The authors are grateful to anonymous reviewers for their suggestions which have led to the improvement of this paper's presentation.

REFERENCES

[1] B. Russel, "On the notion of cause," in *Proc. Aristotelian Soc.*, vol. 13, 1913, pp. 1–25.
[2] R. Jones, "Causality and determinism in physics," *Amer. J. Phys.*, vol. 64, pp. 208–215, 1996.
[3] P. Suppes, *A Probabilistic Theory of Causality*. Amsterdam, The Netherlands: North-Holland, 1970.
[4] L. Petrović, "Causality and markovian representation," *Stat. Prob. Lett.*, vol. 26, pp. 223–227, 1996.
[5] M. Verhaegen, "A subspace model identification solution to the identification of mixed causal, anti-causal LTI systems," *SIAM J. Matrix Anal.*, vol. 17, no. 2, pp. 332–347, 1996.
[6] R. M. De Santis, "Causality theory in systems analysis," *Proc. IEEE*, vol. 64, no. 1, pp. 36–44, Jan. 1976.
[7] W. Porter, "An overview of polynomic system theory," *Proc. IEEE*, vol. 64, no. 1, pp. 36–44, Jan. 1976.
[8] A. Bertuzzi, A. Gandolfi, and A. Germani, "Causal polynomial approximation for input-output maps on hilbert space," *Math. Syst. Theory*, vol. 14, pp. 339–352, 1981.
[9] I. W. Sandberg, "Conditions for the causality of nonlinear operators defined on a linear space," *Quart. Appl. Math.*, vol. 23, no. 1, pp. 87–91, 1965.
[10] ——, "Criteria for iniform approximatiion using dynamical neural networks," in *Proc. 4th IEEE Mediterranean Symp. Control Automation*, Crete, Greece, June 1996, pp. 1–5.
[11] ——, "Notes on uniform approximation of timre-varying systems on finite time intervals," *IEEE Trans. Circuits Syst. I, Fundam. Theory Appl.*, vol. 45, no. 8, pp. 863–865, Aug. 1998.

[12] ——, “Time-delay polynomial networks and quality of approximation,” *IEEE Trans. Circuits Syst. I, Fundam. Theory Appl.*, vol. 47, no. 1, pp. 40–49, Jan. 2000.
[13] P. G. Howlett, A. P. Torokhti, and C. E. M. Pearce, “The modeling and numerical simulation of causal nonlinear systems,” *Nonlinear Anal.*, vol. 47, no. 8, pp. 5559–5572, 2001.
[14] ——, “A philosophy for the modeling of realistic nonlinear systems,” in *Proc. Amer. Math. Soc.*, vol. 132, 2003, pp. 353–363.
[15] H. W. Bode and C. E. Shannon, “A simplified derivation of linear least square smoothing and prediction theory,” *Proc. IRE*, vol. 38, pp. 417–425, 1950.
[16] M. V. Ruzhansky and V. N. Fomin, “Optimal filter construction for a general quadratic cost functional,” *Bull. St. Petersburg Univ., Math.*, vol. 28, pp. 50–55, 1995.
[17] S. Chen and S. A. Billings, “Representation of nonlinear systems: NARMAX model,” *Int. J. Contr.*, vol. 49, no. 3, pp. 1013–1032, 1989.
[18] S. Chen, S. A. Billings, and W. Luo, “Orthogonal least squares methods and their application to nonlinear system identification,” *Int. J. Contr.*, vol. 50, no. 5, pp. 1873–1896, 1989.
[19] A. Torokhti and P. Howlett, “On the best quadratic approximation of nonlinear systems,” *IEEE Trans. Circuits Syst. I, Fundam. Theory Appl.*, vol. 48, no. 5, pp. 595–602, May 2001.
[20] ——, “Method of recurrent best estimators of second degree for optimal filtering of random signals,” *Signal Process.*, vol. 83, pp. 1013–1024, 2003.
[21] L. L. Scharf, “The SVD and reduced rank signal processing,” *Signal Process.*, vol. 25, pp. 113–133, 1991.
[22] Y. Yamashita and H. Ogawa, “Relative Karhunen-Loéve transform,” *IEEE Trans. Signal Process.*, vol. 44, no. 2, pp. 371–378, Feb. 1996.
[23] B. Javidi and J. Wang, “Optimum filter for detecting a target in multiplicative noise and additive noise,” *J. Opt. Soc. Amer. A*, vol. 14, no. 4, pp. 836–844, 1997.
[24] Y. Hua and W. Q. Liu, “Generalized Karhunen–Loéve transform,” *IEEE Signal Processing Letters*, vol. 5, no. 6, pp. 141–143, 1998.
[25] V. J. Mathews and G. L. Sicuranza, *Polynomial Signal Processing*, New York: Wiley, 2001.
[26] L. I. Perlovsky and T. L. Marzetta, “Estimating a covariance matrix from incomplete realizations of a random vector,” *IEEE Trans. Signal Process.*, vol. 40, no. 12, pp. 2097–2100, Dec. 1992.
[27] M. Jansson and P. Stoica, “Forward-only and forward-backward sample covariances—a comparative study,” *Signal Process.*, vol. 77, no. 3, pp. 235–245, 1999.
[28] J.-P. Delmas, “On eigenvalue decomposition estimators of centro-symmetric covariance matrices,” *Signal Process.*, vol. 78, no. 1, pp. 101–116, 1999.
[29] G. Kauermann and R. J. Carroll, “A note on the efficiency of sandwich covariance matrix estimation,” *J. Amer. Stat. Assoc.*, vol. 96, pp. 1387–1396, 2001.
[30] C. J. Champion, “Empirical Bayesian estimation of normal variances and covariances,” *J. Multivar. Anal.*, vol. 87, pp. 60–79, 2003.
[31] M. K. Schneider and A. S. Willsky, “A Krylov subspace method for covariance approximation and simulation of a random process and fields,” *Multidim. Syst. Signal Process.*, vol. 14, pp. 295–318, 2003.
[32] T. Kubokawa and M. S. Srivastava, “Estimating the covariance matrix: a new approach,” *J. Multivar. Anal.*, vol. 86, no. 1, pp. 28–47, 2003.
[33] O. Ledoit and M. Wolf, “A well-conditioned estimator for large-dimensional covariance matrices,” *J. Multivar. Anal.*, vol. 88, no. 2, pp. 365–411, 2004.
[34] A. N. Kolmogorov and S. V. Fomin, *Introductory Real Analysis*. New York: Dover, 1970.
[35] J. M. Holtzman, *Nonlinear System Theory. A Functional Analysis Approach*. Englewood Cliffs, NJ: Prentice-Hall, 1970, p. 70.
[36] M. Cotlar and R. Cignoli, *An Introduction to Functional Analysis*. Amsterdam, The Netherlands: North-Holland, 1974, pp. 114–116.
[37] L. V. Kantorovich and G. P. Akilov, *Functional Analysis*. Oxford, U.K.: Pergamon, 1982.
[38] P. Howlett and A. Torokhti, “Weak interpolation and approximation of nonlinear operators on the space $\mathcal{C}([0,1])$,” *Numer. Funct. Anal. Optim.*, vol. 19, no. 9 & 10, 1998.
[39] A. Torokhti, P. Howlett, and S. Lucas, “Method of partitioned data orthogonalization for principal component determination,” in *Proc. XVI Int. Symp. Mathematical Theory of Networks Systems*, Leuven, Belgium, 2004.
[40] K. Abed-Meraim, W. Qiu, and Y. Hua, “Blind system identification,” *Proc. IEEE*, vol. 85, no. 8, pp. 1310–1322, Aug. 1997.
[41] Q. Lin and J. P. Allebach, “Combating speckle in SAR images: vector filtering and sequential classification based on a multiplicative noise model,” *IEEE Trans. Geosci. Remote Sensing*, vol. 28, no. 4, pp. 647–653, Apr. 1990.
[42] K. Fukunaga, *Introduction to Statistical Pattern Recognition*. Boston, MA: Academic, 1990.

Anatoli Torokhti is an Associate Professor at the School of Mathematics and Statistics, University of South Australia, Mawson Lakes, SA, Australia.

His research interests are in the areas of mathematical modelling of nonlinear systems, mathematical signal processing, operator approximation, applied statistics, and operator equations.

Phil Howlett received the B.Sc , BSc.Hons., and Ph.D. degrees of in mathematics , from the University of Adelaide, Adelaide, Australia, in 1964, 1965, and 1971, respectively.

He is currently a Professor in the School of Mathematics, University of South Australia, Mawson Lakes, SA, Australia and the Director of the Centre for Industrial and Applied Mathematics at the same University. He is also the Director of the Australian Mathematics-in-Industry Study Group. His research interests are applied optimal control theory, optimization, singular perturbations, and operator approximation.

Charles Pearce received the the B.Sc degrees in mathematics and physics from the University of New Zealand, in 1961, the M.Sc degree in mathematics from Victoria University, Wellington, Australia, in 1962, and the Ph.D. degree in mathematical statistics from the Australian National University, Australia, in 1966.

He is currently a Professor in Applied Mathematics at The University of Adelaide, Adelaide, Australia. His research interests are harmonic analysis, teletraffic, queueing theory, and probability theory.

Dr. Pearce is the National President of the Australian Society for Operation Research and a Member of the Editorial Boards of eleven journals in applied mathematics.