Biologically-Motivated System Identification: Application to Microbial Growth Modeling

Jinyao Yan[1] and J.R. Deller, Jr.[1]

Abstract— **This paper presents a new method for identification of system models that are linear in parametric structure, but arbitrarily nonlinear in signal operations. The strategy blends traditional system identification methods with three modeling strategies that are not commonly employed in signal processing: linear-time-invariant-in-parameters models, set-based parameter identification, and evolutionary selection of the model structure. This paper reports recent advances in the theoretical foundation of the methods, then focuses on the operation and performance of the approach, particularly the evolutionary model determination. The method is applied to the modeling of microbial growth by Monod Kinetics.**

I. INTRODUCTION

In parametric system identification, a critical problem is finding a suitable structure within which a good model can be found. In nonadaptive identification, guided by varying degrees of physical information, the model form is ordinarily fixed prior to estimating parameters. The term "black-box" is used to describe a model that has been selected independently of any physical system knowledge [1], [2]. Linear, time-invariant (LTI) models are frequently used because of their simplicity and the wealth of theoretical and algorithmic support that attends these structures.

Identification of nonlinear models – particularly of the black-box type in which no guidance in model selection is available – remains a challenging problem. The approach suggested in this paper employs models that are LTI in parameters (LTIiP), but which may, in general, be extremely nonlinear in the signal interactions. This approach significantly generalizes the classes of models that can be employed in signal processing applications while preserving much of the well-developed solution structure that is available for linear models. Moreover, unlike the work that has been done on nonlinear models [2], the present approach ascertains the model structure as part of the model development and estimation, a feature that could be useful in linear model estimation as well. Specifically, we describe an evolutionary algorithm-based approach to the selection of the nonlinear regressors.

As an example application of this new identification approach, we illustrate the ability of this method to identify a system that tracks the microbial growth profile of a culture population following the Monod equation [3]. The Monod equation has been used for more than 60 years to model the growth rates of microbial populations in aqueous environments as a function of sustaining nutrients. Using the Monod equation as a convenient way to generate a simulated culture trajectory, the new identification algorithm is used to estimate the parameters of a nonlinear model *ad hoc*, with no *a priori* observation or estimation models beyond the LTIiP constraint.

II. IDENTIFICATION FRAMEWORK

Consider a single-input–single-output (SISO) discrete-time system with input $x \in \mathbb{R}^{\mathbb{Z}}$ and output $\zeta \in \mathbb{R}^{\mathbb{Z}}$, each typically assumed to belong to some well-behaved space like ℓ_2. The internal processing of the system is based on a subset of a ***candidate set*** of nonlinear regressor functions, $\Xi_\psi = \{\psi_q\}$, of size $|\,\Xi_\psi\,|$. Each regressor is a mapping $\psi_q : \mathbb{R}^{r_q+s_q} \to \mathbb{R}$, operating on a set of r_q past and present system inputs, and s_q past outputs. The LTIiP ***observation model***, $\mathbb{O}_{\boldsymbol{\theta}_*}$, is given by

$$\begin{aligned}\zeta[t] &= \theta_{1*}\psi_{1*}(t,x,\zeta) + \cdots + \theta_{Q*}\psi_{Q*}(t,x,\zeta) + e_*[t] \\ &\overset{\text{def}}{=} \boldsymbol{\theta}_*^T \boldsymbol{\psi}_*(t,x,\zeta) + e_*[t], \quad t \in \mathbb{Z}, \end{aligned} \tag{1}$$

with $\boldsymbol{\theta}_* \in \mathfrak{P}$ (parameter space) $\subset \mathbb{R}^Q$, and $e_* \in \mathbb{R}^{\mathbb{Z}}$ an error sequence (properties described below) representing uncertainties in the model. The "$*$" subscript indicates a "true," but unknown, quantity associated with the observation model.[1]

Given observations of x and ζ sufficient to compute outputs on time interval $t \in \mathfrak{T}$, we pose an ***estimation model***,

$$\zeta^p[t] = \theta_1\psi_1(t,x,\zeta) + \cdots + \theta_Q\psi_Q(t,x,\zeta) \overset{\text{def}}{=} \boldsymbol{\theta}^T\boldsymbol{\psi}(t,x,\zeta)$$

in which each ψ_q is drawn from the set Ξ_ψ (see Footnote 1) and $\boldsymbol{\theta} \in \mathfrak{P}$. The superscript on ζ^p is meant to connote "prediction", as this estimation model corresponds to the classical prediction-error method of Ljung [2] and others. The residual sequence associated with the observation model at discrete time t is a function of the parameters, as well as the regressor functions chosen,

$$\varepsilon(t,\boldsymbol{\theta},\boldsymbol{\psi}) = \zeta[t] - \zeta^p(t) = \zeta[t] - \boldsymbol{\theta}^T\boldsymbol{\psi}(t,x,\zeta)\,. \tag{2}$$

The objective is to determine the appropriate regressor functions concomitantly with the estimation of parameters

*This work was supported in part by the U.S. National Science Foundation under Cooperative Agreement DBI-0939454. Any opinions, conclusions or recommendations expressed are those of the authors and do not necessarily reflect the views of the NSF.

[1] Department of Electrical Engineering, and the NSF BEACON Research Center, Michigan State University, East Lansing, MI 48824, USA {yanjinya,deller}@egr.msu.edu

[1] Because the index q in ψ_q has been defined as an enumeration of the elements of the candidate set Ξ_ψ, the functions in (1) should be indexed as $\psi_{q_i*}, i = 1, \ldots, Q$, but we use the slightly abusive notation ψ_{q*} for simplicity. It is to be understood that ψ_{q*} is the q^{th} element *selected from* Ξ_ψ, rather than the q^{th} element of Ξ_ψ.

978-1-4244-7929-0/14/$26.00 ©2014 IEEE

TABLE I

ADAPTATION OF SYSTEM MODELING TO A GENETIC ALGORITHM

Cell Biology	System Model
Chromosome	LTIiP model
Gene	Regressor function
Regulator of gene	Parameter

for a particular modeling application. The regressor set will be chosen according to a genetic algorithm based on an evolutionary view of the selection process.

The method used to identify the parameters $\boldsymbol{\theta}$ plays a critical role in the evolutionary model selection process. In the interests of focusing on the model determination and performance, we will only sketch the parameter estimation procedures which are described in [4], with foundations in archival papers [5], [6], [7], [8], [9], [10], [11].

Set-membership* (*SM*) *identification refers to a class of algorithms that use *a priori* knowledge about a model to constrain the parameter solutions to certain sets. In the present paper we employ the the ***quasi-optimal bounding ellipsoid*** (QOBE) algorithm described in [12], [13]. QOBE can be regarded as a blending of the classical recursive least squares (RLS) approach with knowledge of bounds on model errors. However, starting with the obviation of statistical modeling of the errors [e_* in (1)] in QOBE, the estimation strategy is profoundly different from that in RLS. QOBE uses a sequence of pointwise error bounds, $|\, e_*[t] \,| < \gamma_t$, $t \in \mathfrak{T}$, to constrain the feasible parameter values to an hyperellipsoidal set, $\mathcal{H}_t$, of solutions in $\mathbb{R}^Q$ at each t. The solution set is only updated at time t if observed data contain innovation – defined as the ability to shrink the hypervolume of the set. For a set of observations $\{x[t], \zeta[t]\}_{t \in \mathfrak{T}}$, and a model with a fixed set of regressor functions, $\boldsymbol{\psi}$, the result of parameter estimation via QOBE is an hyperellipsoidal set of feasible parameter vectors that are consistent with the measurements and the known error bounds. The center of the ellipsoid can be used as a point estimate if desired and it has the interpretation of a RLS estimate with weights selected by the set shrinkage optimization. Several properties of these sets can be used to infer ***evolutionary fitness*** of a particular regressor function set. In this paper we use the volume of the final hyperellipsoid, and the squared error associated with the central estimate.

That $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$ are sought concurrently represents a significant departure from the conventional QOBE development. Here this deviation is handled by automatic bound estimation (ABE) [14], and the underbounding safeguard (UBS) [15]. We refer the algorithm as QOBE-ABE-UBS, or “QAU”.

III. EVOLUTIONARY MODEL SELECTION

A. *“Cell Biology” of the LTIiP Model*

The genetic algorithm aspects of this work represent a novel adaptation of standard methods in the fields of evolutionary algorithms (e.g., [16]). The present problem has an unusual, but very coherent, mapping into the genetic algorithm framework. Specifically, a chromosome in the genetic algorithm encodes a LTIiP model as an string of bits from a binary alphabet $\{0, 1\}$. Let $\Xi_\psi = \{\psi_q\}$, contain the regressor functions available to create models. The decoding of a chromosome into its corresponding model is thus an array indexing operation where the chromosome is broken into genes with equal lengths, and each gene comprises one member of the regressor function set, say ψ_q, in the model. By analogy, a regressor function, as a “building block” of the model which is indicated by a particular gene, plays the role of a phenotype in the model organism. The parameters represent regulators of the genes, the desired model being the linear mix of genes that give the model the highest potential for survival (see Table I). The parameter of each chromosome is determined by the QAU algorithm.

To demonstrate, a LTIiP model is represented by a binary string (chromosome) as follows

$$\underbrace{\overbrace{110000}^{\text{gene}}\overbrace{000010}^{\text{gene}}}_{\text{chromosome}} \leftrightarrow \underbrace{\zeta[t] = \theta_1 \overbrace{x[t]}^{\psi_{i_1}} + \theta_2 \overbrace{x[t-1]\zeta[t-1]}^{\psi_{i_2}}}_{\text{LTIiP model}} \tag{3}$$

where $i_1, i_2 \in \{1, 2, \ldots, |\Xi_\psi|\}$. In this example, the chromosome is a binary string of length 12. Broken into two sub-strings of equal length, each six-bit string represents a regressor function (phenotype) through a prescribed array indexing operation. The parameters of the LTIiP model are determined by fitting the model to the observations $\{x[t], \zeta[t]\}_{t \in \mathfrak{T}}$.

A viable model is one with parameter values that allow it to effectively produce the observed ζ from the observed x. Like evolutionary biology, survival depends on the inherent suitability of an individual's genetic makeup to meet the challenges of the environment (reflected in x and ζ), and also in the realization of that genetic potential through an effective parameter set.

B. *Set Measures*

The set-theoretic aspects of the identification constrain the sets of parameters to those that are feasible in light of the observations and the error constraints. In turn, they determine the range and statistical viability of phenotypes, and ultimately the plight of the chromosomes. Hence, starting with a population $\mathcal{P}$, a set of candidate models $\{\zeta^{p,j}[t] = (\boldsymbol{\theta}^j)^T \boldsymbol{\psi}^j\{t, x, \zeta\}\}$ of size $|\mathcal{P}|$, the performance of each model (corresponding to a chromosome) is evaluated via an objective function derived from the set measure properties of the QAU algorithm.

For simplicity, in the experiments below the error energy associated with the central estimate is used as the objective function

$$g^j = 1 / \sum_{t \in \mathfrak{T}} (\zeta[t] - \zeta^{p,j}[t])^2 \tag{4}$$

where $\zeta^{p,j} \in \mathcal{P}$, $j \in \{1, 2, \ldots, |\mathcal{P}|\}$ is the estimated output using the ellipsoid center. Thus g^j is the reciprocal of the error energy associated with candidate model j. Sigma scaling is used to map the object values g to the fitness values

f [16]

$$f^j = \begin{cases} 1 + \frac{g^j - \overline{g}}{2\sigma}, & \sigma \neq 0 \\ 1, & \sigma = 0 \end{cases} \tag{5}$$

where $\overline{g}$ and σ are the mean and the standard deviation of the object values of the population, respectively. A model that can produce a satisfactory estimate of the output will have a high fitness value, f.

Each chromosome is assigned a fitness value derived from the objective function which is then used in the selection to bias the new population towards more fit individuals. Highly fit individuals have a high probability of being selected for reproduction. The process continues through subsequent generations. The average fitness of the population increases as more fit individuals appear and interbreed, and the less fit individuals die out. The evolutionary selection algorithm is terminated when a certain number of generations is reached or the fitness values in the population reaches a prescribed maximum. Evolutionary operations, mutation, reproduction and replacement, are used. Algorithm 1 summarizes the process, and the details regarding evolutionary operators can be found in [4].

Algorithm 1: Evolutionary Model Selection

```
Data: Observation subsequences $x, \zeta$
Result: Best fitting model
Initialization :
  ① population $\mathcal{P}$ of size $|\mathcal{P}|$
  ② maximum generation $N_{\max}$
  ③ crossover and mutation rate
  ④ QAU algorithm initialization
/* Regressor selection starts */
for $t \leftarrow 1$ to $N_{\max}$ do
    for $j \leftarrow 1$ to $|\mathcal{P}|$ do
        QAU ;
        calculate objective values $g^j$ of each individual;
        calculate fitness values $f^j$ of each individual;
    end
    selection;
    crossover;
    mutation;
    replacement;
end
```

IV. Experiments and Discussion

A. System Description

The evolutionary identification algorithm is applied to the identification of a simulated microbial growth process. The growth rate is modeled by Monod Kinetics [3], [17], [18] as follows (τ denotes continuous time)

$$\frac{dM}{d\tau} = \frac{aM(\tau)S}{S+b} - M(\tau)x(\tau) \tag{6}$$

$$\frac{dS}{d\tau} = \frac{caM(\tau)S(\tau)}{S(\tau)+b} + (S_{\text{in}} - S(\tau))x(\tau) \tag{7}$$

where $M(\tau)$ represents the microbial concentration in the process at time τ; $S(\tau)$ is the substrate concentration in the process; $x(\tau)$ is the dilution rate and also system input; a is the maximum growth rate; b is the saturation parameter; c is the yield factor; and S_{in} is the inlet substrate concentration. $M(\tau)$ and $S(\tau)$ are state variables; a, b, and c are system parameters; and S_{in} is a constant. We assume that S is observed at discrete time instants

$$\zeta[t] = S(t) + e_*[t], \quad t = 1, 2, 3, \ldots \tag{8}$$

where e_* is the measurement noise sequence, which is unknown, but pointwise-bounded. The system parameters are set to: $a = 0.55$, $b = 0.15$, $c = 2$, and $S_{\text{in}} = 0.8$. The input x is assumed known. A set of 2000 sampled input-output pairs $\{x[t], \zeta[t]\}_{t \in \mathfrak{T}}$ is generated from the model above and plotted in Fig. 1.

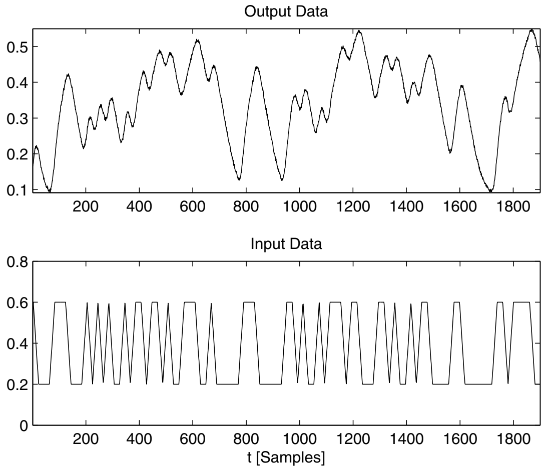

Fig. 1. Simulated Monod Kinetics: Sampled input and output data

B. Evolutionary System Identification

The proposed evolutionary model selection algorithm is applied using the input, output pairs generated above to identify the system. In the initialization of the algorithm, the population size is 20 per generation, the number of iterations N_{max} is 100. We use one-point crossover. The crossover rate is 1.0 and mutation rate is 0.003 per site. The length of the chromosome is 24 genes, with six-bit coding for each gene, yielding 2^6 possible regressor functions ψ_q. Hence, there are 679120 different estimation models. The set Ξ_ψ of regressor functions contains various linear, and nonlinear expressions of combinations of short-delay samples of x and ζ. Specifically, Ξ_ψ contains ψ_q functions which operate on input samples x to delay 3 and output samples ζ to delay 4. The nonlinear expression are polynomial combinations of x and ζ. For instance, $x[t-1]$, $x[t-2]\zeta[t-3]$.

The result of fitting the observations $\{x[t], \zeta[t]\}_{t \in \mathfrak{T}}$ with the model selected by the evolutionary system identification algorithm is shown in Fig. 2. The model selected can be observed to exhibit excellent tracking ability.

For comparison, the same input-output observations are fitted using an autoregressive with exogenous input (ARX)

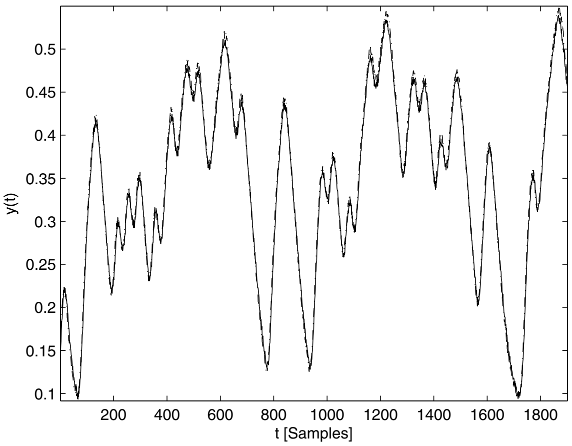

Fig. 2. System identification result using proposed evolutionary model selection algorithm: True data (dash-dot curve) and estimated data (continuous curve).

model of the form

$$\hat{\zeta}[t] = a_1\zeta[t-1]+\cdots+a_s\zeta[t-s]+b_0x[t]+\cdots+b_rx[t-r] \quad (9)$$

The model also corresponds to a special case of evolutionary algorithm for model selection when $\{\psi_q\}$ are just delayed samples of x and ζ. The parameters are estimated using least square error optimization. The results of fitting the observations $\{x[t], \zeta[t]\}_{t\in\mathfrak{T}}$ with $s = 4$, $r = 3$ are shown in Fig. 3. The system output is significantly better reconstructed in Fig. 2. Thus, using evolutionary nonlinear model selection is superior to fitting the observations with the linear model (ARX). Note that the nonlinear character of the model was discovered automatically via evolution. No prior knowledge of the Monod kinetics was available to the estimator.

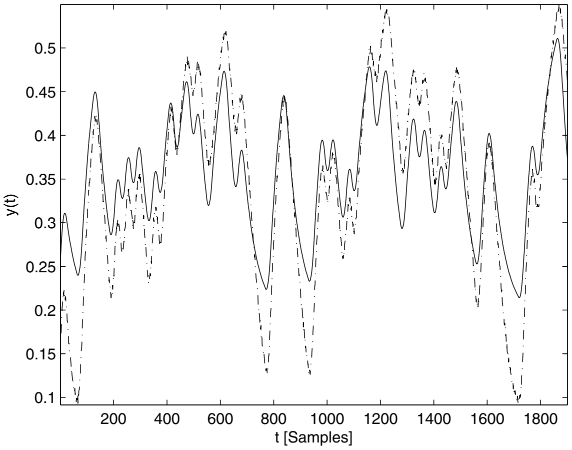

Fig. 3. System identification result using ARX model: True data (dash-dot curve) and estimated data (continuous curve).

V. Conclusion

A biologically-inspired framework for nonlinear system identification based on set-theoretic estimation has been presented. Whereas conventional model identification focuses on the estimation of parameters, the framework presented here simultaneously addresses model selection and parameter estimation. The approach synergistically integrates three modeling and identification strategies that are not commonly employed in signal processing: (i) LTIiP models, (ii) set-based parameter identification, and, (iii) evolutionary strategies for optimization over fitness measures derived from the set solutions.

As an application, a highly-nonlinear system commonly used to model microbial growth rates was identified without knowledge of the true system dynamics. The evolutionary algorithm produces excellent tracking of the microbial population growth relative to a conventional linear time-series model.

References

[1] J. Sjöberg, Q. H. Zhang, L. Ljung, et al., "Nonlinear black-box modeling in system identification: A unified overview," *Automatica*, vol. 31, pp. 1691–1724, 1995.

[2] L. Ljung, *System Identification: Theory for the User*, Prentice-Hall PTR, 2^{nd} edition, 1999.

[3] J. Monod, "The growth of bacterial cultures," *Annual Reviews in Microbiology*, vol. 3, pp. 371–394, 1949.

[4] J. Yan, J. R. Deller Jr., D. B. Fleet, et al., "Evolutionary identification of nonlinear parametric models with a set-theoretic fitness criterion," in *Proc. 2013 IEEE China Summit and International Conf. Signal and Information Processing*, 2013, vol. 1, pp. 44–48.

[5] P. L. Combettes, "The foundations of set theoretic estimation," *Proc. IEEE*, vol. 81, pp. 182–208, 1993.

[6] E. Walter, J. Norton, H. Piet-Lahanier, and M. Milanese, Eds., *Bounding Approaches to System Identification*, Perseus Publishing, 1996.

[7] J. R. Deller Jr., M. Nayeri, and S. F. Odeh, "Least-square identification with error bounds for real-time signal processing and control," *Proc. IEEE*, vol. 81, pp. 815–849, 1993.

[8] J. R. Deller Jr. and Y. F. Huang, "Set-membership identification and filtering for signal processing applications," *Circuits, Systems, and Signal Processing*, vol. 21, pp. 69–82, 2002.

[9] J. R. Deller Jr, "Set membership identification in digital signal processing," *ASSP Magazine, IEEE*, vol. 6, no. 4, pp. 4–20, 1989.

[10] E. Fogel, "System identification via membership set constraints with energy constrained noise," *IEEE Trans. Automatic Control*, vol. 24, pp. 752–758, 1979.

[11] E. Fogel and Y. F. Huang, "On the value of information in system identification: Bounded noise case," *Automatica*, vol. 18, pp. 229–238, 1982.

[12] J. R. Deller Jr., S. Gollamudi, S. Nagaraj, et al., "Convergence analysis of the quasi-OBE algorithm and related performance issues," *International J. Adaptive Control and Signal Processing*, vol. 21, pp. 499–527, 2007.

[13] J. R. Deller Jr., M. Nayeri, and M. S. Liu, "Unifying the landmark developments in optimal bounding ellipsoid identification," *International J. Adaptive Control and Signal Processing*, vol. 8, pp. 43–60, 1994.

[14] T. M. Lin, M. Nayeri, and J. R. Deller Jr., "A consistently convergent OBE algorithm with automatic estimation of error bounds," *International J. Adaptive Control and Signal Processing*, vol. 12, pp. 305–324, 1998.

[15] D. Joachim and J. R. Deller Jr., "Adaptive optimal bounded-ellipsoid identification with an error under-bounding safeguard: Applications in state estimation and speech processing," in *Proc. IEEE International Conf. Acoustics, Speech, and Signal Processing, 2000*, 2000, vol. 1, pp. 372–375.

[16] M. Melanie, *An Introduction to Genetic Algorithms*, MIT Press, 1998.

[17] Q. Zhang, "Using wavelet network in nonparametric estimation," *IEEE Trans. Neural Networks*, vol. 8, pp. 227–236, 1997.

[18] G. D'ans, D. Gottlieb, and P. Kokotovic, "Optimal control of bacterial growth," *Automatica*, vol. 8, pp. 729–736, 1972.