Variable Structure NARX Models: Application to Dissolved-Oxygen Bioprocess

Ioannis G. Sargantanis and M. N. Karim
Dept. of Chemical and Bioresource Engineering, Colorado State University Fort Collins, CO 80523

Selecting the best structure of nonlinear autoregressive models with exogeneous inputs (NARX) is important for good parameter estimation and control. If the process is also time varying, the best structure selection procedure should be done in an on-line fashion. Simultaneous model selection and parameter estimation can be achieved by employing the Givens forward selection with exponential windowing (GFEX) algorithm. An on-line implementation yields a variable-structure NARX-model identification, which can be used for control of a dissolved-oxygen (DO) process. The results of an adaptive DO control scheme, which uses a fixed model structure and was applied to a batch Bacillus subtilis fermentation, were used in this study. Time-dependent patterns of selected terms were observed that roughly corresponded to different fermentation stages. Fitting and prediction capabilities for the adaptively selected NARX model were compared with fixed ARX and NARX models. The effect of the forgetting factor and the model order was investigated. Performance is satisfactory for short-step-ahead prediction, and avoiding overparametrization can stabilize the long-step-ahead prediction.

Introduction

Modeling of biological systems presents a challenge. The dissolved-oxygen (DO) process,'' namely the process of oxygen transfer and consumption in biological systems is difficult to model, due to the time-varying, nonlinear nature of the cell metabolism and its dependence on DO level, and the presence of disturbances and variations, such as antifoam addition, and alteration of the mass-transfer coefficient and oxygen solubility during the course of a fermentation. Modeling the cell metabolic activities, and in particular the oxygen uptake rate, requires the knowledge of intricate metabolic pathways, regulatory mechanisms, and specific reaction rates, which are difficult to obtain. Also, on-line measurements are necessary for metabolic activity indicators such as cell density and carbon dioxide evolution rate, which are not always available. Deterministic input–output models can be used instead, because they require no knowledge of biological system. Their parameters can be updated on-line to allow tracking of the time-varying system dynamics.

The aim of this article is to investigate the performance of adaptively updated single input–output models used for DO control purposes. The performance is evaluated from the fitting accuracy of the model and the accuracy of the predicted future outputs. It turns out that a trade-off should be achieved between fitting and prediction capability for successful implementation of the model for control purposes. The model structure was fixed at first. A changing model structure can account for the changing metabolic state of the culture. A procedure was developed for on-line structure selection and parameter estimation for nonlinear ARX models. Different formulations of the model structure in terms of the degree of the nonlinearity of the complete model were evaluated.

The input for all models was the flow rate of oxygen in the fermenter in a constant flow-rate gas mixture with nitrogen. This manipulated variable was selected because it allows study of the DO level effect on fermentations under constant shear environment that is achieved by keeping the agitation rate and the total aeration rate constant. This selection had a significant disadvantage, namely, the introduction of time delay for the model input. The variation of the time delay due to the changing cell metabolism was estimated by allowing a high order of the input terms, such that the model structure selected can account for the varying time delay. This approach can be modified to account for explicit time-delay determina-

Correspondence concerning this article should be addressed to M. N. Karim.

tion. This way, loss of input dependence on noisy conditions and selection of purely autoregressive models can be avoided. Research to incorporate time-delay estimation with the structure selection is ongoing.

Recursive On-Line Structure Selection of NARX Models

A solution to tracking dynamics of nonlinear systems, which change with time, is to use time-varying nonlinear models and on-line parameter estimation. Linear in parameters, nonlinear autoregressive with exogeneous inputs (NARX) models (Leontaritis and Billings, 1985) can be used to describe a large class of nonlinear systems. For a discrete time SISO system, the NARX model can be written as

$$y(t) = F[\, y(t-1),\, \ldots,\, y(t-na),\, u(t-1-nk),\, \ldots,\, y(t-nb-nk)\,] + e(t), \quad (1)$$

where t is the tth time step; $y(t)$, $u(t)$, and $e(t)$ denote the output, input, and noise, respectively; na and nb are the corresponding orders of the output and input; and nk is the minimum time delay of the input. The function F[] includes cross-product and higher-order polynomial terms. The degree of the power terms in $y(t)$ and $u(t)$ is referred to as the degree of nonlinearity, l. Various possibilities of parametrizing F[] exist (Chen et al., 1989). If the function is chosen to include all the possible monomials of lagged $u(t)$ and $y(t)$, for example, $u(t-1)y(t-2)$ up to order l, a polynomial model is obtained. Introducing a bias term according to Pröll and Karim (1994), the model can be built around the n linear terms, which correspond to $l=1$, where $n = na + nb + 1$. These terms can be represented by the following vector:

$$x(t) = [\, x_1(t),\, \ldots,\, x_n(t)\,] = [1,\, y(t-1),\, \ldots,\, y(t-na),\, u(t-1-nk),\, \ldots,\, u(t-nb-nk)\,]. \quad (2)$$

Each additional term of the polynomial model is built from these n basic elements without the bias term by multiplying them together up to polynomial degree of l. The polynomial model can then be written in the following compact form in terms of the vector $x(t)$ as follows:

$$y(t) = \sum_{j1=1}^{n} \sum_{j2=jl}^{n} \cdots \sum_{jl=jl-1}^{n} \left([\, x_{jl} x_{j2} \cdots x_{jl}\,] \theta_j \right) + e(t), \quad (3)$$

where θ_j are the model parameters, $j = 1, \ldots, m_l$, where m_l is the number of nonlinear polynomial terms of the model corresponding to an l degree on nonlinearity. According to Eq. 3, the number of possible regressors in the model m_l is given by

$$m_l = \sum_{j1=1}^{n} \sum_{j2=j1}^{n} \cdots \sum_{jl=jl-1}^{n} (1). \quad (4)$$

Thereafter, the subscript l will be dropped and the number of possible terms in the model will be denoted as m. Equation 3 can be written in the following concise form:

$$y(t) = \sum_{i=1}^{m} \phi_i(t-1)\theta_i(t-1) + e(t) = \phi^T(t-1)\theta(t) + e(t). \quad (5)$$

Combining all the data at time steps 1, 2, ..., t produces the following model:

$$y(t) = \Phi(t)\theta(t) + e(t), \quad (6)$$

where $\Phi(t)$ is a $t \times m$ regression matrix that is given by $\Phi(t) = [\phi(0), \ldots, \phi(t-1)]^T$, $\theta(t)$ is a $m \times 1$ model parameter vector, $y(t) = [y(1), \ldots, y(t)]$, and $e(t) = [e(1), \ldots, e(t)]$ (both $t \times 1$ vectors).

Like most nonlinear models, the structure of the NARX model doesn't satisfy the shift-invariant property or successive expansion of candidate variables. Lattice algorithms developed for linear systems can perform on-line adjustment for the order of linear ARX or ARMAX models using order-recursion methods based on the shifting property of the regressors (Luo et al., 1994). The recursive modified Gram-Schmidt orthogonalization (Ling et al., 1986) does not require the shifting property as a prerequisite for performing order recursion, but the expansion of the candidate variables must be in a specified form for easy adjustment of the model order. Since polynomial NARX models do not have such a simple structure of model expansion, the order-recursive procedure of RMGS cannot be readily applied to adjust the structure of the model. The off-line QR decomposition algorithms with forward selection described by Chen et al., (1989) can provide a good starting point for developing on-line structure detection of models. These algorithms have the following problems:

1. Need for sequential orthogonal transformation.
2. A matrix position swapping between columns (that occurs when an important term is selected) destroys the orthogonality of the remaining vectors, hence the need for reorthogonalization.
3. The regression matrix may not be of full rank, hence problems with singularities can arise.

A new recursive QR decomposition algorithm for time-varying NARX models called Givens rotation with forward selection and exponential windowing (GFEX) developed by (Luo et al., 1994) is used to select recursively the model on-line and estimate the unknown parameters. This basic algorithm for model selection based on the Akaike information criterion (AIC) (Akaike, 1970) and an exponential window model error-tracking are presented next.

QR Decomposition

If $Q(t)$ is an orthonormal $t \times t$ matrix, that is $Q^T(t)Q(t) = I$, which obeys the following relationship:

$$Q^T(t)\Phi(t) = \begin{bmatrix} R(t) \\ O(t) \end{bmatrix}, \quad (7)$$

where $O(t)$ is a $(t-m) \times m$ zero matrix and $R(t)$ is an $m \times m$ upper triangular matrix, this decomposition is called an orthogonal triangular decomposition or a QR decomposition.

By premultiplying Eq. 6 with $Q^T(t)$, one obtains

$$Q^T(t)\,y(t) = Q^T(t)\Phi(t)\theta(t) + Q^T(t)e(t) = \begin{bmatrix} R(t) \\ O(t) \end{bmatrix}\theta(t) + Q^T(t)e(t). \quad (8)$$

If $\boldsymbol{v}(t) = Q^T(t)y(t)$, then the following expression can be written:

$$\boldsymbol{v}(t) = \begin{bmatrix} \boldsymbol{v}_m(t) \\ \boldsymbol{v}_{t-m}(t) \end{bmatrix} = \begin{bmatrix} R(t) \\ O(t) \end{bmatrix}\theta(t) + Q^T(t)e(t), \quad (9)$$

where $\boldsymbol{v}_m(t)$ contains the first m components of $\boldsymbol{v}(t)$ and $\boldsymbol{v}_{t-m}(t)$ contains the remainder. The QR decomposition of the initial data set can be achieved using various methods, such as Gram-Schmidt orthogonalization, Householder transformations, and Givens rotations (Golub and Van Loan, 1989). The slower Givens rotation method is used because of the ease of developing and expanding the algorithm. Background on Givens rotation is given in the next subsection. Building the $t \times (m-1)$ matrix $[\Phi(t),\ y(t)]$, for the first m columns, $j = 1, \ldots, m$, apply Givens rotations successively with $i = j$, $k = j+1, \ldots, m$ and $m = j$. As explained later, this introduces zeros all through column j, below the diagonal, while retaining orthogonality. This sequence of Givens rotation for each column has to be repeated for the first m columns. Due to the introduction of zeros, the row elements up to the column $j-1$ are equal to zero; therefore, the Givens rotation can be applied for the row elements $n = j, \ldots, m+1$ only. The effect of this method is the following transformation:

$$[\Phi(t)\ y(t)] \Rightarrow \begin{bmatrix} R(t) & \boldsymbol{v}_m(t) \\ O(t) & \boldsymbol{v}_{t-m}(t) \end{bmatrix}. \quad (10)$$

Once the regressor matrix $\boldsymbol{\Phi}(t)$ has been decomposed and the matrices $R(t)$ and $\boldsymbol{v}_m(t)$ are available, the estimation of $\theta(t)$ can be done easily by solving the triangular system $R(t)\theta(t) = \boldsymbol{v}_m(t)$ using a backsubstitution algorithm. Hence, off-line parameter estimation can be obtained this way. Due to the orthogonality of $Q(t)$, the norm of the residual error can be calculated as follows:

$$\| e(t) \| = \| y(t) - \Phi(t)\theta(t) \| = \left\| \boldsymbol{v}(t) - \begin{bmatrix} R(t) \\ O(t) \end{bmatrix}\theta(t) \right\| = \left\| \begin{bmatrix} \boldsymbol{v}_m(t) - R(t) \\ \boldsymbol{v}_{t-m}(t) \end{bmatrix} \right\| = \| \boldsymbol{v}_{t-m}(t) \| = \left(\sum_{i=m+1}^{t} v_i^2(t) \right)^{1/2}. \quad (11)$$

Taking into account the fact that $\| \boldsymbol{v}(t) \| = \| y(t) \|$, one can conclude that

$$\| e(t) \|^2 = \| \boldsymbol{v}_{-mt}(t) \|^2 = \| y(t) \|^2 - \| \boldsymbol{v}_m(t) \|^2 = y^T(t)y(t) - \boldsymbol{v}_m^T(t)\boldsymbol{v}_m(t). \quad (12)$$

This result will be used later in the model-selection algorithm.

Recursive QR Decomposition

For recursive on-line parameter estimation, suppose that there are $t-1$ observations of the system $y(t-1) = \Phi(t-1)\ \theta(t-1)$ that contain m regression variables. Also suppose that a new observation is added to improve the previous estimates. If the QR decomposition matrices $R(t-1)$ and $\boldsymbol{v}_m(t-1)$ are available, the new matrices should be updated using the new information only. This can be represented as the following transformation:

$$\begin{bmatrix} R(t-1) & \boldsymbol{v}_m(t-1) \\ \phi^T(t-1) & y(t) \end{bmatrix} \Rightarrow \begin{bmatrix} R(t) & \boldsymbol{v}_m(t) \\ O(t) & \nu_t(t) \end{bmatrix}. \quad (13)$$

This transformation can be done by employing an orthogonal transformation matrix Q_t^T, which can be formed as the product of m Givens rotations as

$$Q_t^T = G_m^T(t)G_{m-1}^T(t)\cdots G_1^T(t), \quad (14)$$

where $G_i(t)$, $i = 1, \ldots, m$, are Given rotation matrices ($(m+1)\times(m+1)$). These matrices are used when a specific entry in a matrix needs to be zeroed (Golub and Van Loan, 1989). They have the following form:

$$G(i,k) = \begin{bmatrix} 1 & \cdots & 0 & \cdots & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & & \vdots & & \vdots \\ 0 & \cdots & c & \cdots & s & \cdots & 0 \\ \vdots & & \vdots & \ddots & \vdots & & \vdots \\ 0 & \cdots & -s & \cdots & c & \cdots & 0 \\ \vdots & & \vdots & & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & \cdots & 0 & \cdots & 1 \end{bmatrix}. \quad (15)$$

The first column involved is i and the second is k. Premultiplication of a matrix C $[(m+1)\times(m+1)]$ by $G(i,k)$ affects only rows i and k of the matrix. To introduce a zero at the position (k,l) of the matrix, one needs to select c and s as

$$c = \frac{c_{i,l}}{\sqrt{c_{i,l}^2 + c_{k,l}^2}}, \qquad s = \frac{-c_{k,l}}{\sqrt{c_{i,l}^2 + c_{k,l}^2}}. \quad (16)$$

If $g = G(i,k)^T\ C$, then

$$g_{i,h} = \begin{cases} cc_{i,h} - sc_{k,h} & j = i, \quad \forall h \\ sc_{i,h} + cc_{k,h} & j = k, \quad \forall h \\ c_{j,h} & j \neq i,k, \quad \forall h. \end{cases} \quad (17)$$

Applying Givens rotations m times, for all the elements of $\phi(t-1)$ in the last row of the matrix in the RHS of Eq. 13, that is, having sequentially $l = 1, \ldots, m$, $k = m+1$, and $i = 1, \ldots, m$ [the diagonal elements of $R(t)$] one can achieve the desired transformation by eliminating all these nonzero elements of the $m+1 =$ th row and introduce zeros. Hence, the transformation is achieved by calculating c and s by Eq. 16 and applying successively m Given rotations (Eq. 17) between rows i and $m+1$, based on column i, for $i = 1, \ldots, m$. Due to the upper triangular structure of $R(t)$, the row elements up to the column $i-1$ are equal to zero, therefore, the

Givens rotation can be applied for the row elements $j = i, \ldots, m+1$ only. To track variations in the parameters of the system, it is essential to put more weight on recent input–output and residuals data. Employing a forgetting factor λ, $0 < \lambda < 1$, the residual sum of squares (RSS), which is minimized for parameter estimation, changes from $\| e(t) \|^2$ to

$$RSS(t) = \| \bar{e}(t) \|^2 = \bar{e}^T(t)\bar{e}(t) = \sum_{i=1}^{t} \lambda^{i-1} e^2(i) = \lambda RSS(t-1) + e^2(t), \quad (18)$$

where $\bar{e}(t) = \Lambda^{1/2}(t)e(t)$, $\Lambda(t) = \text{diag}[\lambda^{t-1}, \ldots, \lambda, 1]$, is a weighted error vector. When the forgetting factor is employed, the $R(t-1)$ and $\boldsymbol{\nu}_m(t-1)$ in Eq. 13 should be weighted and multiplied by $\lambda^{1/2}$ yielding the following starting $(m+1)\times(m+1)$ matrix $C(t)$ for the transformation:

$$C(t) = \begin{bmatrix} \lambda^{1/2}\bar{R}(t-1) & \lambda^{1/2}\bar{\boldsymbol{\nu}}_m(t-1) \\ \phi^T(t-1) & y(t) \end{bmatrix} \Rightarrow \begin{bmatrix} \bar{R}(t) & \bar{\boldsymbol{\nu}}_m(t) \\ O(t) & \nu_t(t) \end{bmatrix}, \quad (19)$$

where $\bar{R}(t-1) = \boldsymbol{\Lambda}^{1/2}(t-1)R(t-1)$ and $\bar{\boldsymbol{\nu}}_m(t-1) = \boldsymbol{\Lambda}^{1/2}(t-1)\boldsymbol{\nu}_m(t-1)$. This formation can be used with the variable forgetting factor.

On-Line Structure Detection

For on-line structure detection three problems must be resolved:

- Determine elements or functions of elements that represent the contribution of a regression variable to the output.
- Reformulate the orthogonal vectors based on the selected optimal regressors at every selection step.
- Determine the proper criterion for stopping selecting terms.

Starting from Eq. 12 one can refer to a reduced set of regressors (parsimonious set) with the subscript s. Dividing by y^Ty, one gets the normalized RSS (NRSS) associated with this model:

$$NRSS = 1 - \frac{\boldsymbol{\nu}_{m_s}^T \boldsymbol{\nu}_{m_s}}{y^Ty} = 1 - \sum_{i=1}^{m_s} \text{ERR}_i, \quad (20)$$

where $\text{ERR}_i = \upsilon_i^2 / y^Ty$. This is defined as the error reduction ratio (ERR) of the regressor term i (Billings and Chen, 1989; Korenberg et al., 1989). Luo et al. (1994) have used the value of υ_i^2 to select significant regressor terms from all the possible regression variables by using a forward-search procedure. According to this scheme the jth optimal regressor is selected by computing υ_p^2 for all $p = j, \ldots, m$ and picking the maximum υ_p^2 term. Looking carefully, one finds that the value of each υ_i^2 represents the contribution of term i to the error reduction of any model composed of the terms $(1, \ldots, k)$, $k = i, \ldots, m$. Hence, the value of υ_i^2 calculated by QR orthogonalization depends on the arrangement of the possible regression terms done initially. In other words, it depends on the values of the previous υ_j^2 terms, $j = 1, \ldots, i-1$. Therefore, only υ_1^2 represents the true significance of the first term. To get through this difficulty, suppose that there is the jth selection step, then, as $j-1$ terms have been selected, the possible $p = j, \ldots, m$ terms have to be tested as the jth term of the model. The error associated with each possible term can be calculated using Eq. 20 after each possible term has been lined up in triangular formation next to the previous terms. This arrangement involves column swapping and retriangularization of the augmented $[R, \boldsymbol{\nu}_m]$ matrix, so that orthogonality conditions are kept. To demonstrate the retriangularization consider $j = 1$ and $p = 3$. After swapping columns j and p of $R(t)$, due to the triangular structure of $R(t)$, the new augmented matrix will be

$$\begin{bmatrix} r_{13} & r_{12} & r_{11} & r_{14} & \cdots & r_{1m} & \nu_1 \\ r_{23} & r_{22} & 0 & r_{24} & \cdots & r_{2m} & \nu_2 \\ r_{33} & 0 & 0 & r_{34} & \cdots & r_{3m} & \nu_3 \\ 0 & 0 & 0 & r_{44} & \cdots & r_{4m} & \nu_4 \\ \vdots & \vdots & \vdots & \ddots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \vdots & 0 & r_{mm} & \nu_m \end{bmatrix}. \quad (21)$$

To retriangularize the $R(t)$ matrix, one should introduce zeros in the appropriate positions below the diagonal, in this case, in the lower triangle between rows 2 and 3 and columns 1 and 2. Givens rotations can be used for this effect. In the general case, Givens rotations should be applied successively for the columns $h = 1, \ldots, p-1$, on the rows $n = h+1, \ldots, p$ for each column and with the following indices for the rotations, $i = h$, $k = n$, and $l = h$. To calculate the candidate terms' NRSS, swapping between column j and p is performed followed by retriangularization by Givens rotations. Using Eq. 20 with $m_s = j$, the NRSS of possible term p, $p = j, \ldots, m$ is recorded. The jth term of the model is then easily selected as the one that minimizes NRSS_p. The selected term then should be positioned in the jth column of R. This again involves swapping of the columns j and p of the original R (before the manipulations to calculate the NRSS_p) and retriangularization of R. When a term is selected the index j is advanced by one and the whole procedure repeated. The final problem to be solved is the criterion for ending the selection procedure. Liu et al. (1987) suggested that the selection of candidate terms continue until

$$\text{NRSS}_{m_s}(t) = 1 - \sum_{i=1}^{m_s} \text{ERR}_i(t) \le \xi_s, \quad (22)$$

where ξ_s is a preset tolerance. Generally, the selected regressor terms and their number at every computational step are different. Korenberg et al. (1989) suggested a value of 0.0005 to 0.005 for ξ_s. In the presence of noise, it is difficult to select a proper value of the tolerance ξ_s, that will work well at every time step. Instead, the employment of a more complex criterion that takes into account both the performance (error reduction) and the complexity (number of terms) of the model is desired. AIC and the Bayesian information criterion (BIC) (Akaike, 1970) are such criteria. Both penalize the number of terms used, so that overparametrization (selection of unnecessarily many terms) is avoided. The ef-

fects of overparametrization on model dynamics are discussed by Aguirre and Billings (1995). The AIC and BIC are given by the following expressions:

$$\mathrm{AIC} = n\log\left(\sigma_e^2\right) + \phi m_s$$

$$\mathrm{BIC} = n\log\left(\sigma_e^2\right) + m_s\log(n-1), \qquad (23)$$

where ϕ is a parameter (suggested value is 4) and σ_e^2 is the variance of the prediction error (*a posteriori*) calculated over n data points. Under the reasonable assumption that the mean of the error is close to zero, the variance can be approximated by the squared norm of the error divided by the number of points considered. When a forgetting factor is employed, the number of data considered depends on the weight put in each of the data. At time step, t, the number of data points $n(t)$ that should be considered in the calculation of the error variance can be given by

$$n(t) = \sum_{i=1}^{t} \lambda^{i-1} = n(t-1)\lambda + 1 \qquad \text{with} \quad n(0) = 0. \quad (24)$$

Then the RSS(t) given by Eq. 18 and $n(t)$ given by Eq. 24 can be used to yield the following expressions for the information criteria:

$$\mathrm{AIC}(t) = n(t)\log\left(\frac{\mathrm{RSS}(t)}{n(t)}\right) + \phi m,$$

$$\mathrm{BIC}(t) = n(t)\log\left(\frac{\mathrm{RSS}(t)}{n(t)}\right) + m_s\log[n(t)-1]. \quad (25)$$

To select the best model, the information criteria must be minimized with respect to the number of terms used. The reason for using two different criteria is that the AIC has been criticized because of its consistent overestimation of the true parameter vector. To use these criteria in on-line model selection, their values are recorded with each term selected. This value decreases until a minimum and then starts increasing. When one of the two criteria attains a value greater than the previously attained one, then the minimum m_m is detected. The number of selected terms m_s is equal to $m_m + 1$.

Finally, after the selection of the model is complete for each time step, the parameters should be calculated. This is done by getting the upper m_s submatrix of $R(t)$, denoted $R_s(t)$, and the first m_s elements of $\boldsymbol{\nu}_m(t)$, denoted as $\boldsymbol{\nu}_{m_s}(t)$, and solving the system $R_s(t)\theta_s(t) = \boldsymbol{\nu}_{m_s}(t)$ by backsubstitution for the parameters $\theta_s(t)$ of the parsimonious model. Care has to be taken for the proper permutation of the selected regressor terms and the reorthogonalization of $R(t)$ before calculating the parameters.

With minor modifications, the GFEX algorithm can be used to recursively estimate the parameters of a fixed model either linear or nonlinear. The modification is achieved by bypassing the error-calculating inner loop and the AIC-minimizing outer loop. At each time step, an index of the column positions corresponding to the desired model can be supplied, and after column swapping and retriangularization the parameters can be estimated by backsubstitution. Because of the column swapping the index should be provided sorted in ascending order. Several alternatives exist for the model construction, based on whether the model structure and/or time-delay changes with time. Fixed ARX or NARX model parameter estimation employs the QR decomposition technique instead of the recursive least-squares (RLS) algorithm. QR decomposition performs well in comparison to RLS (Sargantanis, 1996).

Some other modifications can be investigated, such as the bias inclusion, the output filtering, and the variable forgetting factor. For the latter option, a reliable formula has to be developed that takes into account a measure of the changing dynamics of the system, discounting the changing structure, and therefore the changing number of terms included.

Results

In this section, the modeling of the DO process using data obtained in batch fermentation runs with *B. subtilis* cells is presented. The basis of all modeling results is the data of a fermentation run employing Medium 3 (Park, 1993) at 500 rpm and 37°C. The sampling time for identification was 24 s, the control algorithm was adaptive pole placement (Sargantanis, 1996), and the time delay was adjusted manually. A variable 15% to 3% DO profiles was employed. The particular experiment was chosen because of the different ranges of dynamic regimes (DO setpoints and cell metabolic states) and control qualities. The plots of the output and input signals is given in Figure 1.

Linear model results

Since the GFEX formulation is suitable for linear models too, the first task is to establish connection between the QR-based and the RLS-based parameter estimation techniques. A fixed model with orders $na = 2$, $nb = 2$, and $nk = 2$ was

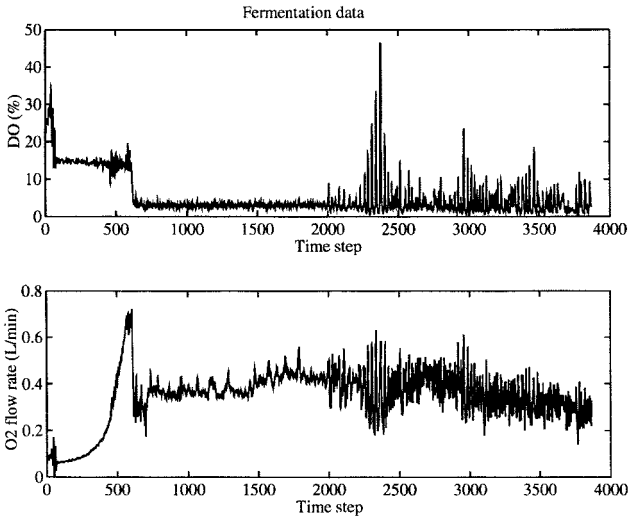

Figure 1. Experimental DO and O_2 flow-rate profiles for a *B. subtilis* fermentation.

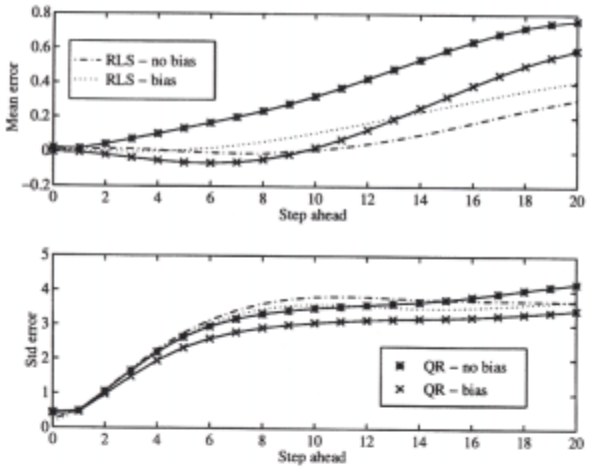

Figure 2. Comparison of RLS- and QR-based parameter estimation of a fixed-order (*na* = 2, *nb* = 2, *nk* = 2) ARX model.

tested. The inclusion of bias (con = 1) was tried afterward. The results showing the comparison of the statistics of the fitting and prediction errors of these options with those of the standard (regularized constant trace) (Lee et al., 1991) RLS implementation are plotted in Figure 2. In this plot, zero step ahead denotes fitting, while positive x-axis values denotes step-ahead prediction. For the GFEX algorithm $\lambda = 0.99$, $l = 1$, $na = 5$, $nb = 8$, $nk = 0$, and the normalization factor is 100. The fitting error is worse (mean and variance) for the QR-estimated models than with RLS. The mean error of the prediction improves with the incorporation of the bias term, a result that is opposite from what was observed for the RLS estimated model. The standard deviation (σ) of the fitting error is higher with QR decomposition. Short-term prediction is better with QR decomposition, and the improvement is more pronounced when the bias term was included. The σ of the prediction error is reduced for all prediction ranges when the bias is included, as was observed for the RLS algorithm. For long-term prediction, comparing RLS with QR, the performance of the QR decomposition is better when the bias is included, while it deteriorates when bias is not included. The bias term thus should always be considered for fixed ARX models estimated using QR decomposition. In the analyses performed for nonlinear models, the bias term is always included as a candidate term.

Results with $l = 1$ (linear models) and $na = 5$, $nb = 8$, $nk = 0$, and con = 1 were obtained. Figure 3 illustrates the time sequence of the selected terms, where a dot denotes that a particular term has been selected. Also the number of possible terms is plotted with time. Figure 4 shows the dependence of the output on input terms as selected by the GFEX algorithm. The value 1 denotes selection of an input regressor, while 0 denotes no selection. The model loses all depen-

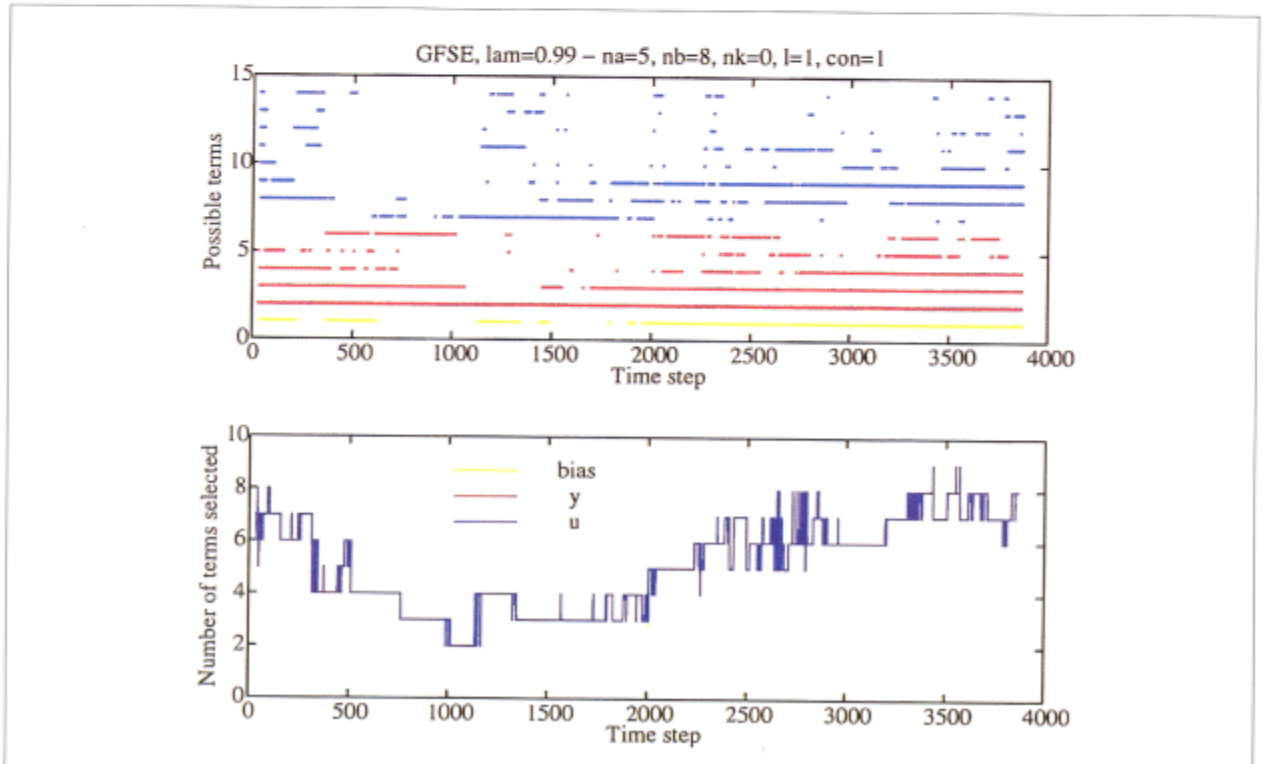

Figure 3. Profile of the selected terms with GFEX–QR decomposition, *na* = 5, *nb* = 8, con = 1, *nk* = 0, *l* = 1, $\lambda = 0.99$.

The legend placed on the bottom subplot corresponds to the selected terms' cluster memberships for the top subplot.

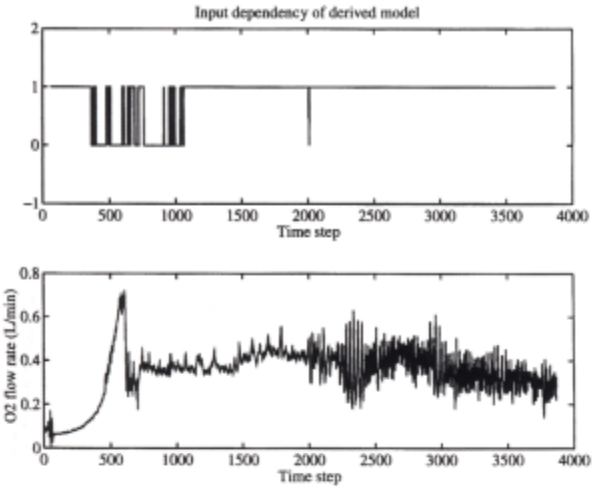

Figure 4. Dependence of the output on input regressors of a linear variable structure model selected by the GFEX procedure.

dence on input terms for an extended period of the exponential growth phase of the microorganisms. At these time steps the model depends only on output terms. This is a disadvantage of the original GFEX algorithm when it is used for control purposes. It is observed that the number of selected terms follows the same pattern, decreasing as the cells go the exponential phase and then increasing gradually to around 7 or 8 terms when the cells begin to die. It might be suggested that the number of terms selected can be an indicator of the metabolic state of the cells, or of the excitation conditions for the model building. Small number of terms implies lack of sufficient excitation. When the control performance is good, then the excitation level is low, and problems can arise that lead to loss of input dependency of the model and therefore deterioration of the control.

Nonlinear model results

The model can be expanded to include nonlinear terms. Models with polynomial degree $l = 2$ or $l = 3$ will be studied. The number of possible terms increases geometrically with the polynomial degree. For $na = 3$, $nb = 5$, and $l = 2$ the number of possible terms are 45 with the constant bias term. The Givens rotation speed is proportional to the square of the number of possible terms considered (Golub and Van Loan, 1989). When the original algorithm, which performs column swapping and retriangularization for the evaluation of all possible regressors, is used, the order dependence is increased to the cube of n (number of possible terms). Therefore, the speed is prohibitively slow for a large model regressor structure. The speed of the algorithm can be in-

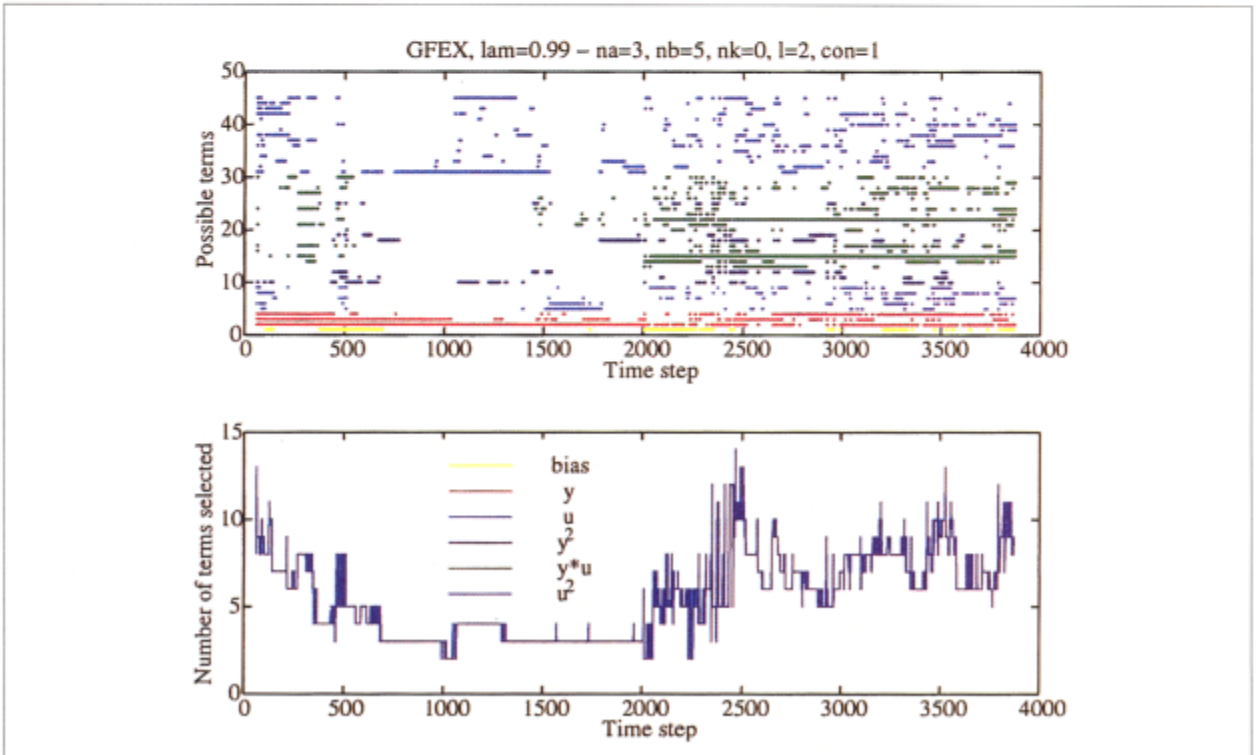

Figure 5. Profile of the selected terms with GFEX–QR decomposition, $na = 3$, $nb = 5$, con = 1, $nk = 0$, $l = 2$, $\lambda = 0.99$.

The legend placed on the bottom subplot corresponds to the selected terms' cluster memberships for the top subplot.

creased by an order of magnitude if one observes that for the term-selection step the entire matrix may not be calculated on swapping and retriangularization, but only the diagonal pivot element and the last column element that provides for the error-reduction calculation. This is implied by the properties of Givens rotations. By swapping columns i and j in an already triangular matrix, the elements of only column i and rows $i+1$ to j need to be zeroed, because these are the only ones that affect the pivot element (i, i) and the the last column element $(i, m+1)$. After these rotations, the subdiagonally introduced zeros for columns $i+1$ to $j-1$ need to be zeroed only if subsequent model terms need to be evaluated for their error-reduction contribution, which is not required. Nevertheless, the matrix should be preserved through the testing of possible regressor terms, thus necessitating the use of a copy of the augmented $[\boldsymbol{R},\boldsymbol{r}_m]$ matrix. Therefore, a considerable speed increase requires only a moderate increase in memory.

Simulations were performed for the second- and third-order polynomial models. The model orders (na, nb) were reduced for the increased degree of nonlinearity. The model selections and the number of terms are depicted graphically in Figures 5, 6, and 7 for the following conditions: (Figure 5: $na=3$, $nb=5$, $l=2$; Figure 6: $na=2$, $nb=3$, $l=3$; Figure 7: $na=3$, $nb=4$, $l=3$). Note that for the last simulation, 120 possible terms could be selected. The graphical depiction includes different gray-scale denotation of the different term clusters (6 for second-order and 9 for third-order models). These clusters include, that is, for the second-order models, the following: bias, y, u, y^2, yu, and u^2. The number of terms of each cluster selected with time provide an easier description of the nonlinear dynamics of the process. Summing the term cluster's parameters can then be plotted with respect to time. Such a plot for the simulation of Figure 5 is provided in Figure 8. Zero values account for the nonselection of the particular cluster.

It can be seen that similar patterns of terms selected exist for both second- and third-order nonlinear models. The number of terms selected doesn't increase proportionally to the increase in the number of possible regressors, as can be seen by comparing Figure 6 and Figure 7. The pattern of the number of terms selected also follows the same behavior as with the linear model (Figure 3). It is noteworthy that in the late exponential phase the number of terms selected is 3 to 4 regardless of the degree of nonlinearity and the number of possible regressors. Figure 9 illustrates the lack of input excitation for a long period in the late exponential growth phase of the bacteria, and also the predominance of the mixed cluster yu in the late phases of the fermentation. It is also noted that the coefficients of the y^2 cluster tend to attain high numerical values (note that the y values are normalized), such that there might be a numerical instability of the long-term output predictions obtained using variable structure models. This behavior is consistent with the effects of overparametrization, where many possible terms are selected, whose parameter estimates tend to be high. To test against overparametrization,

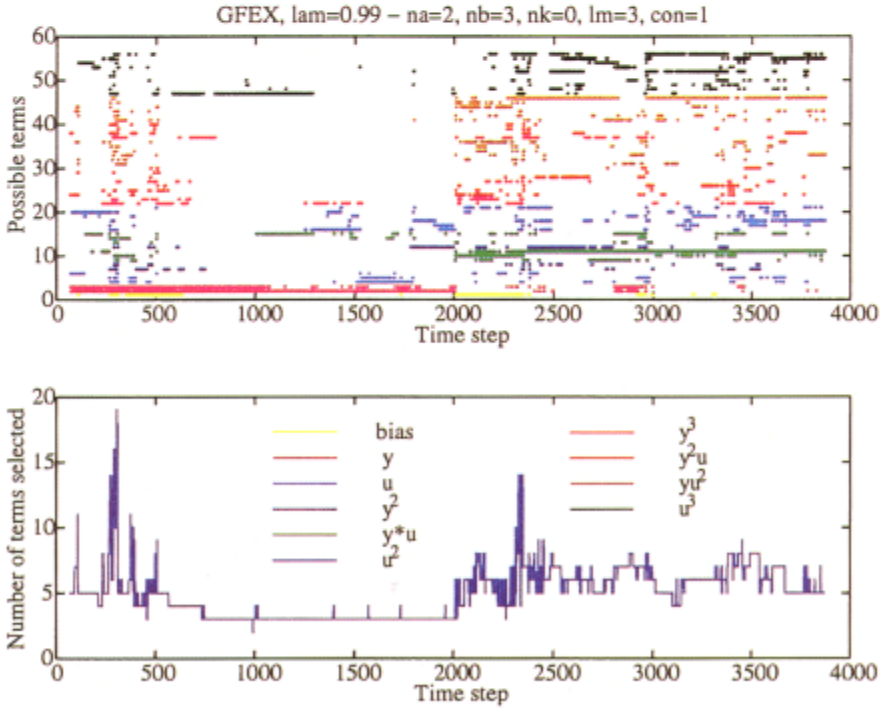

Figure 6. Profile of the selected terms with GFEX–QR decomposition, $na=3$, $nb=4$, con$=1$, $nk=0$, $l=3$, $\lambda=0.99$.

The legend placed on the bottom subplot corresponds to the selected terms' cluster memberships for the top subplot.

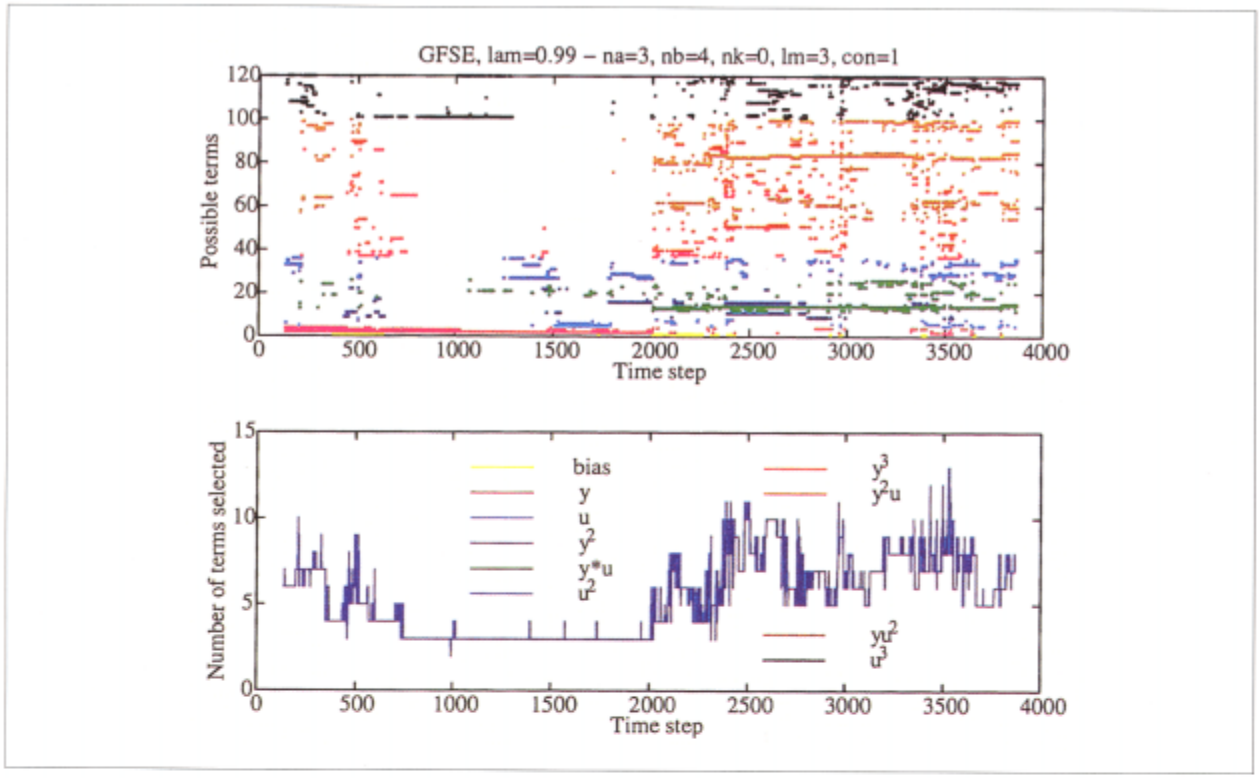

Figure 7. Profile of the selected terms with GFEX–QR decomposition, $na=3$, $nb=4$, con $=1$, $nk=0$, $l=3$, $\lambda=0.99$.

The legend placed on the bottom subplot corresponds to the selected terms' cluster membership for the top subplot.

when excessive noise or nonpersistent excitation conditions occur, the condition number of the $R_s(t)$ matrix after model selection is plotted with time in Figure 10 for various models. It can be seen that the degree of nonlinearity increases the condition number. Nevertheless, the condition number never attains disproportionally high values to account for the blowup of long-term output prediction, which occurs with nonlinear models at certain times. Thus, some heuristics are needed to discourage these prediction failures, and modifications to remediate these failures have to be introduced. This can be an area of future research.

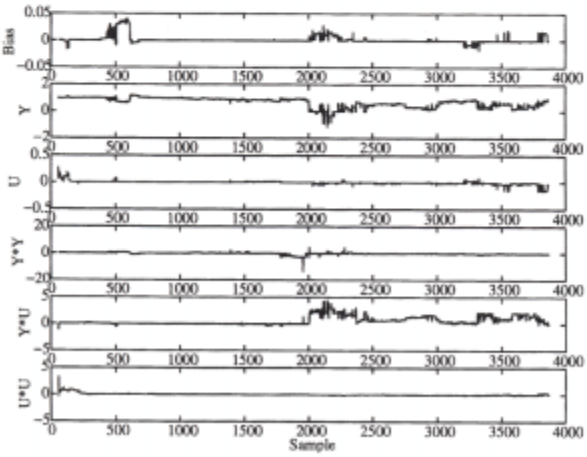

Figure 8. Plot of the term clusters coefficients using GFEX–QR decomposition, $na=3$, $nb=5$, con $=1$, $nk=0$, $l=2$, $\lambda=0.99$.

The effect of the exponential forgetting factor on structure selection was also studied. The selected structures with $\lambda=0.98$ and with $\lambda=0.998$ are shown in Figures 11 and 8, respectively, for second-order models. These figures are compared with Figure 5, which corresponds to $\lambda=0.99$. Both pattern of selected terms and number of these terms are influenced by λ. It is concluded that the patterns become sharper as the forgetting factor increases. On the other hand, the number of terms selected increases, especially in the early exponential phase and by the end of the fermentation run. Thus, effects of overparametrization should be stronger with increasing λ.

The effect of the degree of nonlinearity and of the exponential forgetting factor on the time delay selected by the model structure is shown in Figure 12. Time delay is the not selected explicitly, but is the lower time delay of the input terms that are selected (either linear or nonlinear). When there is no input dependence (see Figure 4), then the time delay is set to zero. From Figure 12 one can see that the selected time delay is 0 or 1, which is a low value compared with the usual time delay associated with the particular system. Most of the zero readings are associated with loss of

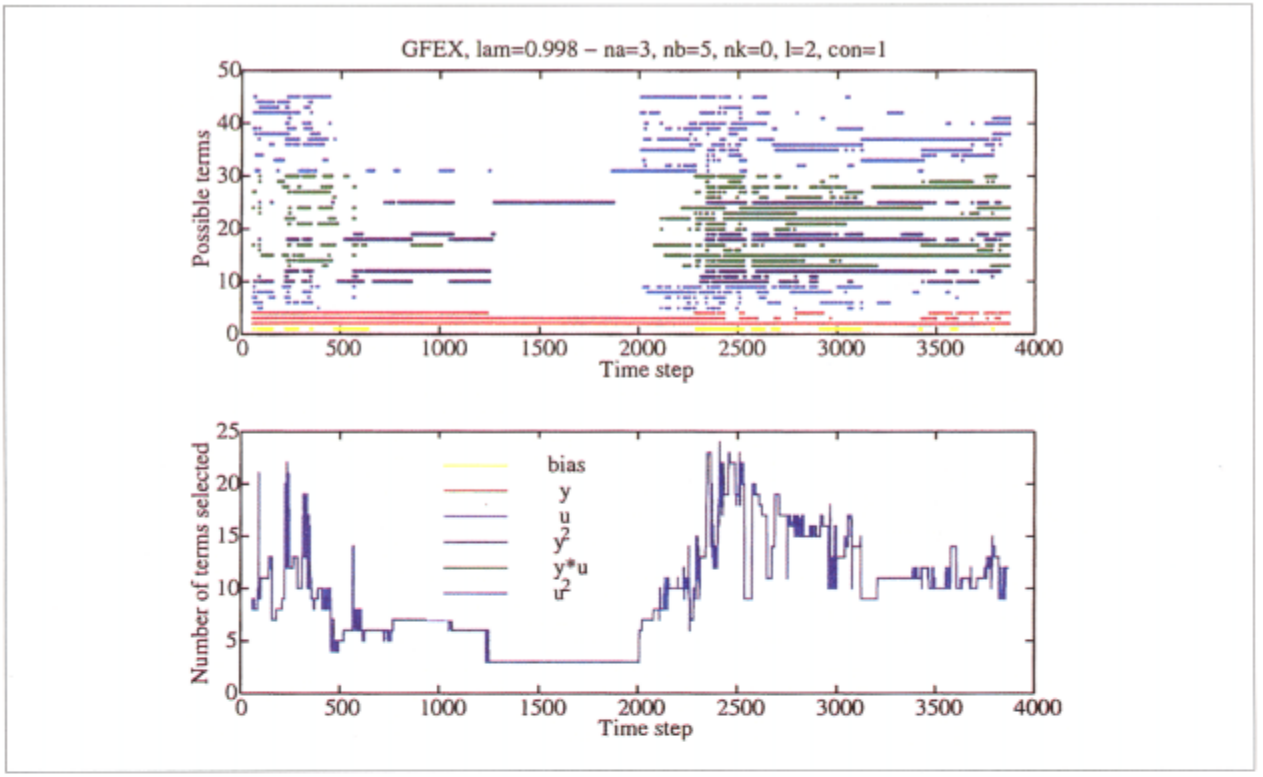

Figure 9. Profile of the selected terms with GFEX–QR decomposition, $na = 3$, $nb = 5$, con = 1, $nk = 0$, $l = 2$, $\lambda = 0.998$.

The legend placed on the bottom subplot corresponds to the selected terms' cluster memberships for the top subplot.

excitation on the input. With increasing degree of nonlinearity, the time delay attains the value of zero less frequently. The biggest effect of the time delay comes from the employed exponential forgetting factor. At $\lambda = 0.998$, that is, very small exponential weighting of the data, zero time-delay spreads over many sample intervals, while with $\lambda = 0.99$, the area of zero time-delay, and, thus of input excitation loss, diminishes.

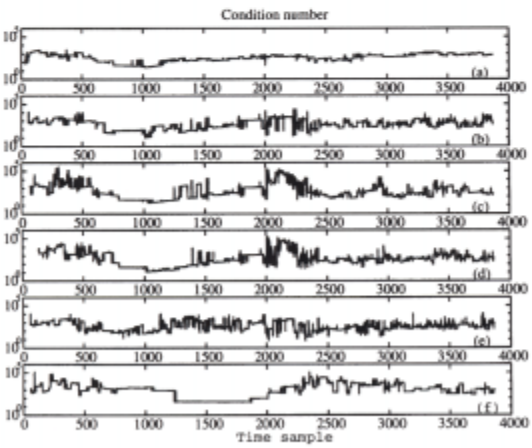

Figure 10. Condition number—con = 1, $nk = 0$: (a) $na = 5$, $nb = 8$, $l = 1$, $\lambda = 0.99$; (b) $na = 3$, $nb = 5$, $l = 2$, $\lambda = 0.99$; (c) $na = 2$, $nb = 3$, $l = 3$, $\lambda = 0.99$; (d) $na = 3$, $nb = 4$, $l = 3$, $\lambda = 0.99$; (e) $na = 3$, $nb = 5$, $l = 2$, $\lambda = 0.98$; (f) $na = 3$, $nb = 5$, $l = 2$, $\lambda = 0.998$.

Finally, the prediction capabilities of the various model structures are presented in Figure 13. Comparing Figure 13 with Figure 2, one notes that Figure 13 doesn't go far in prediction horizon. This occurs because the prediction blows to very big values, which becomes more obvious with the degree of nonlinearity involved. For the mean error, at short-step-ahead prediction (less than 6), the variable structure model are better than fixed ARX models. Afterward, the prediction error diverges fast. For the error standard deviation, the prediction is comparable or better than the prediction of a QR-estimated fixed ARX model for short step-ahead prediction. As with the mean, the standard deviation of the error deviates rapidly to large values. Linear models surprisingly outperform the nonlinear models at most steps ahead. This can be attributed to avoidance of overparametrization, as well as for allowance of higher input and output model orders than the nonlinear models. Indeed, the third-order model performs better when more terms are allowed. For the forgetting factor, the best results were obtained at $\lambda = 0.99$. Overall, there should be an optimum forgetting factor, or an optimal time sequence of λ for each process at hand, which allows good prediction capabilities, correct model structure selection, and which avoids input excitation loss and overparametrization. For employing NARX variable-structure

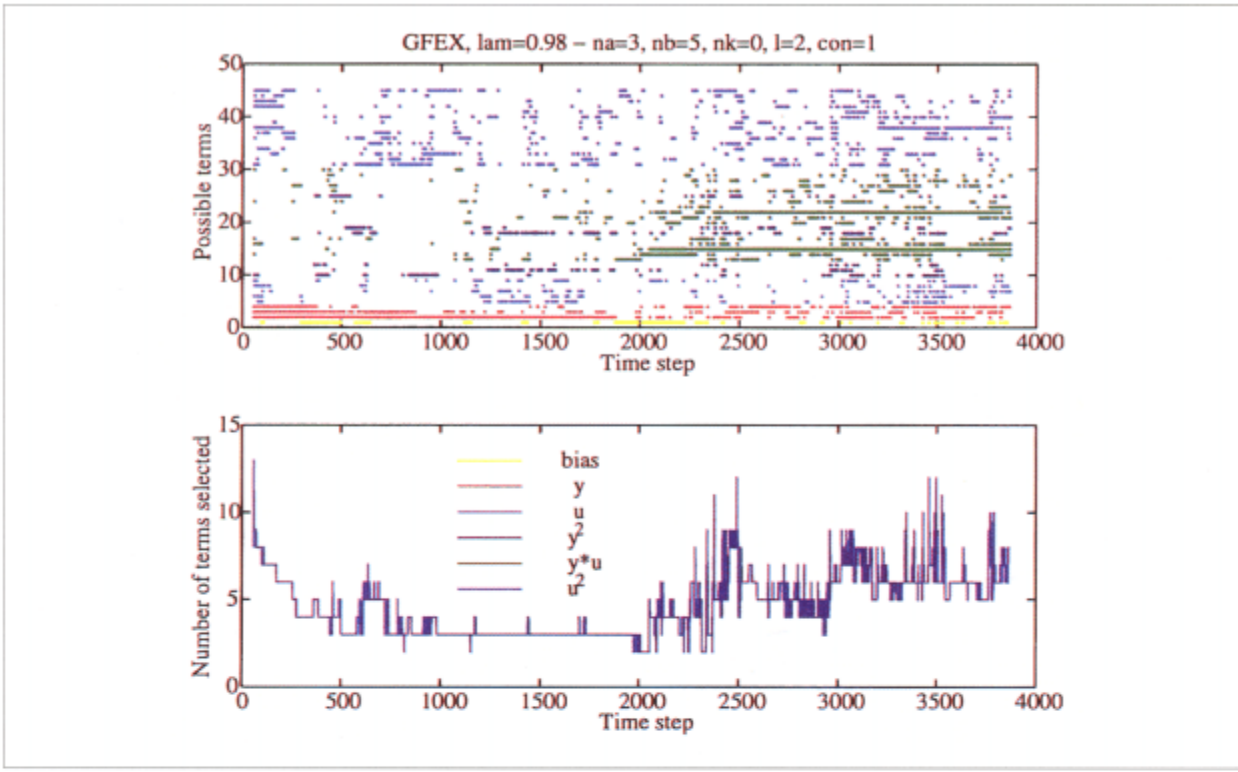

Figure 11. Profile of the selected terms with GFEX–QR decomposition, $na = 3$, $nb = 5$, con = 1, $nk = 0$, $l = 2$, $\lambda = 0.98$.

The legend placed on the bottom subplot corresponds to the selected terms' cluster memberships for the top subplot.

models in model predictive control (MPC), the problems of loss of input excitation and the prediction blowout should be solved. Explicit time-delay determination and safeguards against ill-conditioning can be incorporated in the structure selection and parameter-estimation scheme to allow for successful control implementation.

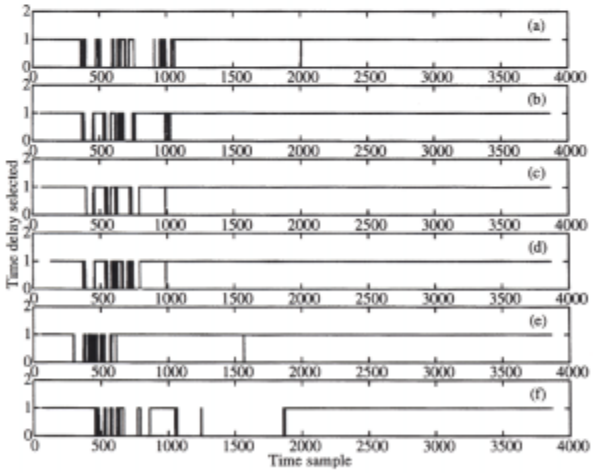

Figure 12. Time-delay—con = 1, $nk = 0$: (a) $na = 5$, $nb = 8$, $l = 1$, $\lambda = 0.99$; (b) $na = 3$, $nb = 5$, l = 2, $\lambda = 0.99$ (c) $na = 2$, $nb = 3$, $l = 3$, $\lambda = 0.99$; (d) $na = 3$, $nb = 4$, $l = 3$, $\lambda = 0.99$; (e) $na = 3$, $nb = 5$, $l = 2$, $\lambda = 0.98$; (f) $na = 3$, $nb = 5$, $l = 2$, $\lambda = 0.998$.

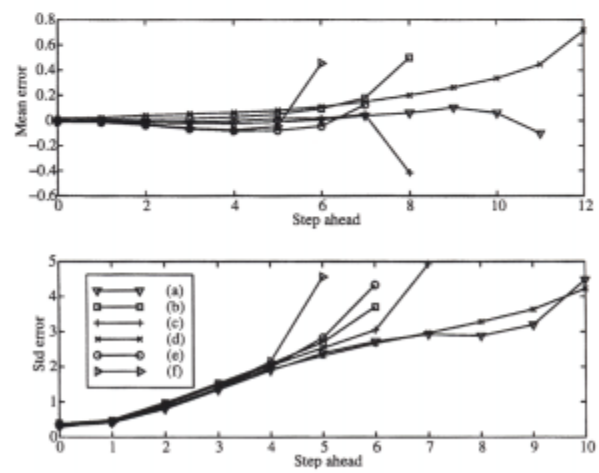

Figure 13. Fitting and prediction-error statistics—con = 1, $nk = 0$; (a) $na = 5$, $nb = 8$, $l = 1$, $\lambda = 0.99$; (b) $na = 3$, $nb = 5$, $l = 2$, $\lambda = 0.99$; (c) $na = 2$, $nb = 3$, $l = 3$, $\lambda = 0.99$; (d) $na = 3$, $nb = 4$, $l = 3$, $\lambda = 0.99$; (e) $na = 3$, $nb = 5$, $l = 2$, $\lambda = 0.98$; (f) $na = 3$, $nb = 5$, $l = 2$, $\lambda = 0.998$.

Conclusions

On-line selection of NARX model structures was performed using a QR decomposition algorithm that employs the GFEX scheme for model selection. Backsubstitution solution provides the model parameters. The procedure was shown to yield time-varying model structures in time-varying batch bacterial cultivation experimental data. Distinct structure patterns were observed in the course of the fermentation. QR decomposition performs well in time-varying parameter estimation. Fitting and short step-ahead prediction were satisfactory, but overparametrization and blowout of the long step-ahead prediction were observed. The forgetting factor turned out to be a tuning parameter to enhance performance. Further research in the fields of explicit time-delay estimation, smoothing of abrupt model structure changes, and on-line detection of poor performance is needed for successful real-time process-control application.

Acknowledgments

The authors are grateful for the financial support from the National Science Foundataion (Grant No: BES-96225926), the Colorado Institute for Research in Biotechnology (CIRB), and the Colorado State University Experimental Station.

Notation

c = Givens matrix second element
e = model error
$\boldsymbol{e}$ = model error vector
I = Identity matrix
s = Givens matrix first element
$\boldsymbol{v}$ = QR tranformed y
y = output vector
ϕ = regressor vector for parameter estimation

Symbols

$_s$ = parsimonious model selection
$^-$ = exponentially weighted

Literature Cited

Aguirre, L. A., and S. A. Billings, ''Improved Structure Selection for Nonlinear Models Based on Term Clustering,'' *Int. J. Control*, **62**, 569 (1995).

Akaike, H., ''Statistical Predictor Identification,'' *Ann. Stat. Math.*, **22**, 203 (1970).

Billings, S. A., and S. Chen, ''Extended Model Set, Global Data and Threshold Model Identification of Severely Non-Linear Systems,'' *Int. J. Control*, **50**, 1897 (1989).

Chen, S., S. A. Billings, and W. Luo, ''Orthogonal Least-Squares Methods and Their Application to Non-Linear System Identification,'' *Int. J. Control*, **50**, 1873 (1989).

Golub, G. H., and C. F. Van Loan, *Matrix computations*, 2nd ed., The Johns Hopkins University Press, Baltimore, MD (1989).

Korenberg, M. J., S. A. Billings, Y. P. Liu, and P. J. McIIoy, ''Orthogonal Parameter Estimation Algorithms for Non-Linear Stochastic Systems,'' *Int. J. Control*, **48**, 193 (1989).

Lee, S. C., Y. B. Hwang, H. N. Chang, and Y. K. Chang, ''Adaptive Control of Dissolved Oxygen Concentration in a Bioreactor,'' *Biotechnol. Bieong.*, **37**, 597 (1991).

Leontaritis, I. J., and S. A. Billings, ''Input-Output Parameter Models for Non-Linear Systems. Part 1: Deterministic Non-Linear Systems; Part 2: Stochastic Non-Linear Systems,'' *Int. J. Control*, **41**, 303 (1985).

Ling, F., D. Manolakis, and J. G. Proakis, ''A Recursive Modified Gram-Schmidt Algorithm for Least Squares Estimation,'' *IEEE Trans. Acoust., Speech, Signal Processing*, **ASSP-34**, 829 (1986).

Liu, Y. P., M. J. Korenberg, S. A. Billings, and M. B. Fadzil, ''The Nonlinear Identification of a Heat Exchanger, IEEE Conf. on Decision and Control, Los Angeles (1987).

Luo, W., S. A. Billings, and K. M. Tsang, ''On-Line Structure Detection and Parameter Estimation with Exponential Windowing for Nonlinear Systems,'' *Tech. Rep. 503*, Univ. of Sheffield, Sheffield, UK (1994).

Park, K., ''Environmental Effects on Cell Metabolism and Recombinant Protein Production by *Bacillus subtilis*,'' PhD Thesis, Colorado State Univ., Fort Collins (1993).

Pröll, T., and M. N. Karim, ''Model-Predictive pH Control Using Real-Time NARX Approach,'' *AIChE J.*, **40**, 269 (1994).

Sargantanis, I. G., ''Model Based Control with Variable Structure: Application to Dissolved Oxygen Control for β-lactamase Production,'' PhD Thesis, Colorado State University, Fort Collins (1996).

Manuscript received Jan. 30, 1998, and revision received Apr. 1, 1999.