International Journal of Control

ISSN: 0020-7179 (Print) 1366-5820 (Online) Journal homepage: http://www.tandfonline.com/loi/tcon20

An identification algorithm for polynomial NARX models based on simulation error minimization

L. Piroddi & W. Spinelli

To cite this article: L. Piroddi & W. Spinelli (2003) An identification algorithm for polynomial NARX models based on simulation error minimization , International Journal of Control, 76:17, 1767-1781, DOI: 10.1080/00207170310001635419

To link to this article: http://dx.doi.org/10.1080/00207170310001635419

Published online: 21 May 2010.

INT. J. CONTROL, 2003, VOL. 76, NO. 17, 1767–1781

Taylor & Francis
Taylor & Francis Group

An identification algorithm for polynomial NARX models based on simulation error minimization

L. PIRODDI†* and W. SPINELLI†

Classical prediction error approaches for the identification of non-linear polynomial NARX/NARMAX models often yield unsatisfactory results for long-range prediction or simulation purposes, mainly due to incorrect or redundant model structure selection. The paper discusses some limitations of the standard approach and suggests two modifications: namely, a new index, based on the simulation error, is employed as the regressor selection criterion and a pruning mechanism is introduced in the model selection algorithm. The resulting algorithm is shown to be effective in the identification of compact and robust models, generally yielding model structures closer to the correct ones. Computational issues are also discussed. Finally, the identification algorithm is tested on a long-range prediction benchmark application.

1. Introduction

In the field of black-box non-linear modelling, the non-linear autoregressive moving average with exogenous variables (NARMAX) representation (Leontaritis and Billings 1985 a,b) has attracted considerable interest in the last years, both in theory (Korenberg *et al.* 1987, Leontaritis and Billings 1987 a, Billings *et al.* 1989, Aguirre and Billings 1995 a,b, Mao and Billings 1997, Aguirre and Jácome 1998, Palumbo and Piroddi 2000 a) and applications (Pottmann *et al.* 1993, Loh and Duh 1996, Kalkkuhl and Liccaga-Castro 1996, Maio *et al.* 1998, Palumbo and Piroddi 2000 b, Palumbo *et al.* 2001, Piroddi *et al.* 2001, Leva and Piroddi 2002). The NARMAX model yields an input–output representation of a non-linear system where the current output is obtained by means of a non-linear functional expansion of lagged inputs, outputs and noise terms. When the focus is on the deterministic input–output relationship, the simpler NARX model is often employed, with a crude simplification of the disturbance model, though the inclusion of an MA part may reduce the bias in the parameter estimation. Depending on how the functional expansion is represented and parameterized, different model structures and, consequently, identification algorithms are derived. In particular, polynomial models have been extensively used, because they are *linear-in-the-parameters* (or at least pseudo-linear, in the NARMAX case) models and can be identified with simple algorithms of the least squares family. Another reason for the popularity of polynomial models is that polynomial terms are often amenable to a direct physical interpretation: for example, in mechanical models for vibration analysis specific polynomial non-linearities are often used to describe well-known non-linear elastic or viscous behaviours. Also, some useful non-linear frequency analysis tools have been developed in the recent literature with reference to this class of non-linear models (Billings and Tsang 1989, Peyton-Jones and Billings 1989, Billings *et al.* 1990, Palumbo *et al.* 2001).

Unfortunately, the size of polynomial NARX/NARMAX models increases rapidly with the number, maximum lag and degree of elementary terms, so that it is important to select the most appropriate terms for the model. In fact, in agreement with the Occam's Razor principle (the simpler the solution, the better), a compact model size is crucial for parameter estimation confidence and model robustness: over-parameterized NARX/NARMAX models may result extremely effective in fitting the estimation data, but do not catch the real dynamics of the underlying system, and they might even exhibit unwanted non-linear behaviours (Aguirre and Billings 1995 b). Though several criteria have been proposed in the literature for the selection of suitable regression terms for polynomial NARX/NARMAX models (Korenberg *et al.* 1987, Leontaritis and Billings 1987 a, Aguirre and Billings 1995 a, Mao and Billings 1997), structure selection remains a largely unsolved problem. Empirical indices, such as the final prediction error (FPE), the Akaike information criterion (AIC) and the minimum description length (MDL) (Leontaritis and Billings 1987 a), have been applied to the problem of non-linear model selection, but they rarely yield conclusive proofs in favour of a specific model structure (Palumbo and Piroddi 2000 b). A widely used criterion is the error reduction ratio (ERR) (Billings *et al.* 1989), which evaluates the significance of model terms in terms of their ability to explain the output variance. A popular identification algorithm based on the ERR criterion is the forward-regression orthogonal estimator (FROE) (Billings *et al.* 1989), which will be used as reference in this work.

Received in final form 13 October 2003.
Communicated by A. Astolfi.
* Author for correspondence. e-mail: piroddi@elet.polimi.it
† Dipartimento di Elettronica e Informazione, Politecnico di Milano, Piazza Leonardo da Vinci 32, 20133, Milano, Italy.

International Journal of Control ISSN 0020–7179 print/ISSN 1366–5820 online © 2003 Taylor & Francis Ltd
http://www.tandf.co.uk/journals
DOI: 10.1080/00207170310001635419

The paper is devoted to the analysis of the structure selection problem for polynomial NARX models, with focus on the simulation performance of the identified models. The paper does not consider the problem of input design, although this issue is as important as the selection criterion (Leontaritis and Billings 1987 b, Chiras *et al*. 2001): it is assumed that the identification experiment cannot be designed and pre-existing data are fed to the identification algorithm. Actually, this is the case when the structure selection problem is more critical, since the data may not be informative enough.

The limitations of the FROE with respect to the structure selection problem are discussed in the paper. In particular, it is shown how the FROE may find incorrect or redundant models, especially in the presence of particular noise structures or input signals. In addition, the identified models may turn out to be extremely inaccurate in simulation and even unstable, so that their use for long-range prediction is hazardous.

As a result of the analysis, a new structure selection criterion based on the minimization of the simulation error is proposed. The proposed criterion, called simulation error reduction ratio (SRR) is compared to the ERR, and shown to yield more accurate models. A further improvement is represented by the introduction of a systematic pruning procedure for the reduction of the model dimension, to be coupled with the structure selection criterion. Some extensions to the FROE are derived as a consequence, namely the SEM (simulation error minimization), the FRP (forward regression with pruning), and the SEMP (simulation error minimization with pruning) algorithms. Some computational issues are discussed with reference to the SEMP algorithm. Finally, the performance of the SEMP algorithm on a benchmark application example is studied: a NARX model is identified for the long-range prediction of the radial crest displacement of an arch dam.

The paper is organized as follows. The NARX/NARMAX model and the FROE algorithm are reviewed in § 2. Section 3 is devoted to the analysis of the limitations and drawbacks of the FROE algorithm, with reference to the role of the ERR, oversampling, cluster analysis, etc. The SRR criterion and the pruning procedure are developed in § 4, together with the resulting extensions to the FROE. The arch dam application example is discussed in § 5. Finally, some concluding remarks are drawn in § 6.

2. Polynomial NARX modelling

2.1. *The NARX/NARMAX model*

The general form of a NARMAX model is given by the following recursive equation (Leontaritis and Billings 1985 a,b)

$$y(t) = f(\,y(t-1),\ldots,y(t-n_y),\,u(t-1),\ldots,\\ u(t-n_u),\,\xi(t-1),\ldots,\xi(t-n_\xi)) + \xi(t) \qquad (1)$$

where $f(\cdot)$ is a suitable non-linear function, $u(\cdot)$, $y(\cdot)$ and $\xi(\cdot)$ are the model input and output and a white noise term, respectively, and n_y, n_u and n_ξ are the respective maximum lags. If $f(\cdot)$ is a polynomial function of its arguments, the model can be further detailed as

$$y(t) = f^p(\,y(t-1),\ldots,y(t-n_y),u(t-1),\ldots,\\ u(t-n_u)) + v(t) \qquad (2\,a)$$

$$v(t) = f^n(\,y(t-1),\ldots,y(t-n_y),u(t-1),\ldots,\\ u(t-n_u),\xi(t-1),\ldots,\xi(t-n_\xi)) + \xi(t) \qquad (2\,b)$$

where $f^p(\cdot)$ denotes the process model, $v(\cdot)$ is a noise term and $f^n(\cdot)$ is the additive noise model. In the NARX model $v(t)$ reduces to the white noise $\xi(t)$, or, stated otherwise, $f(\cdot)$ coincides with the process model $f^p(\cdot)$

$$y(t) = f(\,y(t-1),\ldots,y(t-n_y),u(t-1),\ldots,\\ u(t-n_u)) + \xi(t) \qquad (3)$$

The deterministic part of the polynomial NARX model can be expanded in the following notation, borrowed from Peyton-Jones and Billings (1989)

$$y(t) = \sum_{m=0}^{l}\sum_{p=0}^{m}\sum_{n_1=1}^{n_y}\cdots\sum_{n_m=1}^{n_u} c_{p,m-p}(n_1,\ldots,n_m)\\ \times\prod_{i=1}^{p} y(t-n_i)\prod_{i=p+1}^{m} u(t-n_i) \qquad (4)$$

where $0\le m\le l$ is the degree of non-linearity ($m=0$ corresponds to a constant term, $m=1$ to linear terms, etc.), l is the maximum degree of non-linearity, each mth order term is the product of a pth order factor in y, a $(m-p)$th order factor in u and a coefficient $c_{p,m-p}(n_1,\ldots,n_m)$. Terms with the same type of non-linearity are characterized by the same p and m–p indices and form a cluster, denoted $\Omega_{y^p u^{m-p}}$ (Aguirre and Billings 1995 a). The sum of the coefficients of all the terms of a cluster, or cluster coefficient, is denoted $\Sigma_{y^p u^{m-p}}$ and is a measure of the real significance of the corresponding type of non-linearity in the model.

A full polynomial expansion up to a given order l will typically have an excessive number of terms, so that direct NARX estimation based on the full polynomial model is not viable. On the other hand, experience shows that satisfactory models can be obtained which employ only a small subset of all the possible regressors of the full model. The model structure selection problem is therefore crucial in polynomial NARX identification.

2.2. *The forward-regression orthogonal estimator for polynomial NARX/NARMAX identification*

Many algorithms have been proposed in the recent literature that combine model structure selection and parameter estimation in a single optimization procedure (Leontaritis and Billings 1987 a, Aguirre and Billings 1995 a, Mao and Billings 1997). It is usually required to specify only the class of models over which the optimization search must be effected, i.e. the set of candidate regressors for inclusion in the model: by combining the elements of this set the identification procedure selects the most significant terms to include in the model, on the basis of a quantitative evaluation of the quality of the estimated model.

One of the most efficient algorithms of this category is the forward-regression orthogonal estimator (FROE) (Billings *et al.* 1989). In this approach, the model structure is iteratively incremented until a specified prediction accuracy is obtained. More specifically, parameters are estimated by means of orthogonal least squares, while the structure selection is based on the error reduction ratio (ERR) criterion. The orthogonalization technique is employed to decouple the estimation of additional parameters from that of the parameters already included in the model. In this way, at each step the significance of each candidate regressor can be separately evaluated, by computing the ERR criterion

$$[\mathrm{ERR}]_i = \frac{\hat{g}_i^2 \sum_{t=1}^{N} w_i^2(t)}{\sum_{t=1}^{N} y^2(t)} \tag{5}$$

where w_i is the ith auxiliary orthogonal regressor and $\hat{g}_i$ the corresponding estimated parameter. The regressor with the highest value of the ERR (i.e. which maximizes the increment in the output explained variance) is added to the model. At the end of the procedure the estimated parameters for the original, non-orthogonal regressors can be recovered from w_i, $\hat{g}_i$, $i = 1, \ldots, n$.

For NARMAX models the structure selection procedure is carried out by first identifying the process model, and then including additional noise terms mainly to reduce the bias in the parameters.

2.3. *Performance of the FROE algorithm*

Let us examine in more detail the functioning and performance of the FROE algorithm. Consider for this purpose the following three systems

$$\mathcal{S}_1:\ w(t) = u(t-1) + 0.5u(t-2) + 0.25u(t-1)u(t-2) - 0.3u(t-1)^3$$
$$y(t) = w(t) + \frac{1}{1-0.8q^{-1}}\xi(t), \quad \xi(\cdot) \sim \mathrm{WGN}(0, 0.02)$$

$$\mathcal{S}_2:\ w(t) = 0.5w(t-1) + 0.8u(t-2) + u(t-1)^2 - 0.05w(t-2)^2 + 0.5$$
$$y(t) = w(t) + \frac{1}{1-0.5q^{-1}}\xi(t), \quad \xi(\cdot) \sim \mathrm{WGN}(0, 0.05)$$

$$\mathcal{S}_3:\ y(t) = 0.5y(t-1) + 0.8u(t-2) + u(t-1)^2 - 0.05y(t-2)^2 + 0.5 + \xi(t), \quad \xi(\cdot) \sim \mathrm{WGN}(0, 0.05)$$

The first system has an NFIR type of input–output relation and an autoregressive process as output disturbance. System $\mathcal{S}_2$ is similar to $\mathcal{S}_1$, except for the fact that its input–output relation has an autoregressive part. Finally, $\mathcal{S}_3$ is a NARX with the same input–output relation as $\mathcal{S}_2$. Note that all the three systems can be reformulated as NARX models.

A first identification experiment is performed generating 500 data realizations from the three systems with $u(\cdot) \sim \mathrm{WGN}(0,1)$. The FROE algorithm is then applied to the data, operating on the family of NARX models of order 2 and degree 3 for system $\mathcal{S}_1$ and on the family of NARX models of order 2 and degree 2 for systems $\mathcal{S}_2$ and $\mathcal{S}_3$.

The initial regressors selected by the FROE algorithm usually amount to a considerable part of the total output variance. It is therefore important to understand if and in what conditions these are chosen correctly by the algorithm. The first two regressors selected in each case are reported in table 1. It appears that the regressors initially selected by the FROE algorithm are actually present in the system equation.

But if the identification test is repeated with a less rich input signal, e.g. a zero mean, low frequency AR(2) process (with two real poles in the range [0.75 0.9]) with unit variance, things change dramatically (table 1). Regardless of the original system structure, the autoregressive terms $y(t-1)$ and $y(t-2)$ are selected, the first one with extremely high ERR values. In each case, the FROE algorithm has basically found the 'trivial' prediction model $\hat{y}(t|t-1) \approx y(t-1)$ (the slow input dynamics imply $y(t-1) \cong y(t-2)$ and the corresponding cluster coefficient Σ_y is close to 1), which may be considered a good initial approximation for a predictor, but is a useless model for simulation (the simulated output tends asymptotically to 0).

If the last identification experiment, with a non-white input, is continued until convergence, the results reported in table 2 are obtained. Here, the selected regressors which appear also in the original system are represented in bold, whereas regressors in italic are candidates for *a posteriori* elimination from the model, since statistically indistinguishable from 0 (the associated standard deviation is comparatively high). Cluster analysis (Aguirre and Billings 1995 a) has also

		White noise input signal			Autoregressive input signal		
System	Iter.	Initial regressors	Estimated parameters	ERR	Initial regressors	Estimated parameters	ERR
$\mathcal{S}_1$	1	u(t − 2)	0.494 E−01	0.39	y(t − 1)	1.337 E+00	0.93
	2	u(t − 1)	3.040 E−01	0.16	y(t − 2)	−3.845 E−01	0.01
$\mathcal{S}_2$	1	$u(t-1)^2$	1.035 E+00	0.67	y(t − 1)	1.385 E+00	0.95
	2	y(t − 1)	4.507 E−01	0.17	y(t − 2)	−4.211 E−01	0.01
$\mathcal{S}_3$	1	$u(t-1)^2$	1.020 E+00	0.65	y(t − 1)	1.500 E+00	0.92
	2	y(t − 1)	4.981 E−01	0.20	y(t − 2)	−5.114 E−01	0.02

Table 1. Initial regressors selected by the FROE for systems $\mathcal{S}_1$, $\mathcal{S}_2$ and $\mathcal{S}_3$ with different input signals.

System	Iter.	Regressors	Estimated parameters	3·STD%	ERR
$\mathcal{S}_1$	1	y(t − 1)	7.325 E−01	16.3%	9.320 E−01
	2	$y(t-2)$	−1.092 E−01	90.6%	1.006 E−02
	3	**u(t − 1)**	1.213 E+00	14.2%	8.352 E−03
	4	$u(t-1)^2$	1.773 E−01	38.5%	7.628 E−03
	5	$\mathbf{u(t-1)^3}$	−3.786 E−01	19.1%	1.474 E−02
	6	u(t − 1) · $u(t-2)^2$	2.783 E−01	30.4%	1.364 E−03
	7	**u(t − 2)**	−6.572 E−01	36.6%	2.545 E−03
	8	$u(t-2)^2$	−8.679 E−02	54.2%	5.781 E−04
$\mathcal{S}_2$	1	**y(t − 1)**	1.012 E+00	10.5%	9.499 E−01
	2	y(t − 2)	−1.982 E−01	47.0%	8.863 E−03
	3	$\mathbf{u(t-1)^2}$	9.994 E−01	12.8%	5.137 E−03
	4	$u(t-2)^2$	−6.149 E−01	25.2%	1.393 E−02
	5	**u(t − 2)**	2.738 E−01	29.1%	1.254 E−03
	6	$\mathbf{y(t-2)^2}$	−1.561 E−02	57.2%	5.269 E−04
	7	**1**	1.926 E−01	56.9%	1.927 E−03
$\mathcal{S}_3$	1	**y(t − 1)**	5.068 E−01	13.8%	9.170 E−01
	2	$y(t-2)$	6.654 E−03	338.4%	2.134 E−02
	3	$\mathbf{u(t-1)^2}$	1.004 E+00	6.1%	1.176 E−02
	4	$u(t-2)^2$	−2.753 E−03	353.4%	2.489 E−02
	5	$u(t-1)$	−5.296 E−02	202.1%	9.568 E−03
	6	$\mathbf{y(t-2)^2}$	−5.079 E−02	14.3%	8.708 E−03
	7	**1**	4.851 E−01	13.9%	3.063 E−03
	8	**u(t − 2)**	8.631 E−01	15.4%	4.644 E−03

Table 2. Models selected by the FROE for systems $\mathcal{S}_1$, $\mathcal{S}_2$ and $\mathcal{S}_3$ ($u(\cdot) \sim$ AR).

been performed but has not given sufficient grounds for eliminating any of the regressor clusters present in the model. On the whole the identified models (re-estimated after the elimination of spurious parameters) provide a reasonable simulation performance: the FROE algorithm has compensated for the initial regressors choice, selecting appropriate regressors, either correct or of the correct cluster. However, only in the third case, where the system generating the data has a NARX structure, was the algorithm able to find the exact model. In general, over-parameterized models are estimated and a model reduction procedure must be executed *a posteriori*.

Linear and non-linear residual tests have also been performed on the analytical examples. In cases $\mathcal{S}_1$ and $\mathcal{S}_2$ these correlation tests show a non-white prediction error, and even some residual correlation between the prediction error and the various signals involved, which may be interpreted as a partial inadequacy of the identified model. A possible interpretation of the functioning of the FROE is as follows. Consider, e.g. system $\mathcal{S}_1$, which can be reformulated as a NARX model, by multiplying both hands of the defining equation by $(1-0.8q^{-1})$

$$y(t) = 0.8y(t-1) + u(t-1) - 0.3u(t-2) - 0.4u(t-3) + 0.25u(t-1)u(t-2) - 0.2u(t-2)u(t-3) - 0.3u(t-1)^3 + 0.24u(t-2)^3 + \xi(t)$$

It is then to be expected that the FROE algorithm will attempt to select *this* input/output structure (the optimal predictor is trivially obtained by the previous expression by eliminating $\xi(t)$). Indeed, the FROE selects terms from the correct clusters with respect to this reformulation (see table 2). However, in so doing the FROE requires additional regressors which are not present in the regressor set (and in the original system). What's more, the NARX version of the system equation is clearly redundant (eight parameters instead of five), and the eight parameters are linearly dependent, but this constraint cannot be taken into account by the FROE and bias results. For these reasons, a non-white prediction error results, which suggests that the model family is inappropriate. But, in fact, it would be wrong to extend the model family, since it already contains all the correct regressors.

3. Some limitations of the standard NARX/NARMAX identification approach

The examples discussed in the previous section raise several questions on the standard NARX/NARMAX identification approach based on the FROE algorithm.

3.1. *The model structure selection process and the ERR criterion*

Two well-known facts concerning the structure selection process as realized by the FROE algorithm are listed below:

(1) The order in which parameters are progressively included in the model influences the model selection process, as explained in Korenberg *et al.* (1987) and Billings *et al.* (1989).

(2) It is not the overall model's prediction error variance which is minimized, but rather its *reduction* achieved by the addition of a new parameter (Billings *et al.* 1989).

In summary, there is no guarantee of finding the optimal model within the family, i.e. the model with the minimum mean square prediction error (MSPE) and with the minimum number of regressors, and the algorithm may include redundant regressors in the model. This is caused mainly by the use of the ERR criterion for model selection, which can partially fail to give a consistent measure of the importance of regressors. In fact, given a selected model structure, the ERR of each regressor can vary enormously depending on the *order* in which the regressors are considered, so that the actual significance of a term is difficult to assess.

This has a number of practical consequences. For example, after the first few regressors are selected, which usually amount for most of the explained variance, the ERR values associated to additional regressors rapidly decrease, though their real importance in the model may be much greater. Therefore, the choice of additional regressors gets uncertain and sensitive to noise in the data as the identification procedure goes on. Another consequence is that, if an essential term is missing from the set of possible model regressors, it will be typically compensated by several spurious ones. Spurious terms may also be included because of noisy data (Aguirre 1994).

3.2. *Prediction vs. simulation in non-linear modelling*

The optimal 1-step ahead predictor of the general NARMAX model (1) can be easily obtained by forcing the prediction error $\varepsilon(t) = y(t) - \hat{y}(t|t-1)$ to equal the white noise term $\xi(t)$

$$\begin{aligned}\hat{y}(t \mid t-1) = f(\, & y(t-1), \ldots, y(t-n_y),\ u(t-1), \ldots, \\ & u(t-n_u), y(t-1) - \hat{y}(t-1|t-2), \ldots, \\ & y(t-n_\xi) - \hat{y}(t-n_\xi|t-n_\xi-1))\end{aligned}$$

Equivalently, the optimal predictor for the NARX model (4) is given by

$$\hat{y}(t|t-1) = f(\, y(t-1), \ldots, y(t-n_y), u(t-1), ..., u(t-n_u))$$

Note that the latter predictor is always stable, whereas the stability of the optimal predictor for the NARMAX model depends on function $f(\cdot)$.

The optimal simulation $\hat{y}_S(t)$ can be computed on the basis of the optimal predictor, by setting $y(t) = \hat{y}_S(t)$ t) and $\hat{y}(t|t-1) = \hat{y}_S(t)$ (Sjöberg *et al.* 1995). In the NARMAX case, this leads to

$$\begin{aligned}\hat{y}_S(t) = f(\hat{y}_S(t-1), \ldots, \hat{y}_S(t-n_y), u(t-1), \ldots, \\ u(t-n_u)), 0, \ldots, 0)\end{aligned}$$

which equals

$$\hat{y}_S(t) = f^p(\hat{y}_S(t-1), \ldots, \hat{y}_S(t-n_y), u(t-1), \ldots, u(t-n_u))$$

since the noise model is zero if $\xi(t) = 0$. The latter equation describes the optimal simulation model for the NARX model too. Thus, a NARX and a NARMAX with the same process model produce the same simulation output. In both cases, the stability of the simulation model depends on the original process model. Note that the simulation error $\varepsilon_S(t) = y(t) - \hat{y}_S(t)$ is not white, in general.

If the system generating the data belongs to the considered model family, there exists the 'optimal' predictor, i.e. the predictor with the minimum MSPE, which includes only the regressors included in the model. If the simulator model is computed on the basis of this optimal predictor, it is also optimal, i.e. it minimizes the mean square simulation error (MSSE). However, only suboptimal predictors are obtained in practice with a prediction error method like the FROE. This happens for several reasons:

(1) If the available data are not adequately exciting for the system to be identified (e.g. low frequency data, oversampled data), many model structures will be rated almost equivalently in terms of prediction performance, so that the selection of the 'correct' model is a numerically ill-posed problem.

(2) The model expression (4) is only an ideal abstraction of the real system.

(3) The size of the set of candidate regressors must be limited in size for practical reasons. Therefore, the model family in which the model structure is selected is typically underparameterized, so that for every exact regressor missing, many compensation terms are added to the model.

Unfortunately, the simulation model computed on the basis of a non-optimal predictor may perform in a significantly different way from the optimal simulator, even if the predictor model is nearly optimal. In the prediction framework, structural model errors are typically compensated by the presence of autoregressive terms which employ the values of the measured output

to keep the predicted output close to the real one, whereas in simulation there is no compensation for a wrong model structure.

It is worth remarking that if simulation performance is evaluated instead of prediction performance, the set of model structures with comparable accuracy tends to be much smaller. A partial explanation of this phenomenon is given by the fact that, since the simulation error accumulates in time, it is expected that models with incorrect structure will have a much larger error than the 'correct' model. This suggests that the use of the simulation error for model selection is both more precise (the set of high performance model structures is smaller) and more robust with respect to the excitation characteristics of the identification data.

3.3. *Structure selection and sampling time*

In system identification, the effects of oversampling are usually associated with numerical problems during parameter estimation, because of the conditioning of the regression matrix. In practice, however, data oversampling will also pose problems for model structure selection (Billings and Aguirre 1995). In particular, autoregressive terms are typically included in the first stages of the identification process (Billings and Aguirre 1995), largely independently of the real model structure. For instance, the term $y(t-1)$ was always selected as first in the example discussed in § 2, in the low frequency input case.

Such terms yield a big increment to the explained variance with respect to the empty model, especially with oversampled or noisy data, though their real importance in the overall model may be much smaller. In fact, it is easy to show that the value of ERR associated to $y(t-1)$ tends to 1 as $T_S \to 0$, since $y(t) \approx y(t-1)$ (Billings and Aguirre 1995). Therefore, since the sum of the ERR values of the selected regressors cannot be greater than 1, all the other terms will have very small ERR values. Consequently, it becomes difficult to select the structure in such a situation.

Note that, if $y(t-1)$ were selected as first regressor, the corresponding parameter, estimated with LS, would be computed as

$$\hat{c}_{1,0}(1) = \left(\sum_{t=1}^{N} y^2(t-1) \right)^{-1} \sum_{t=1}^{N} y(t) y(t-1)$$

and $\hat{c}_{1,0}(1) \approx 1$ for $T_S \to 0$. The corresponding simulation model would then be $\hat{y}_S(t) \approx \hat{y}_S(t-1)$, which would result in a catastrophic performance. In view of this, it is clear that a criterion based on the simulation performance rather than the prediction performance would quite certainly not select an autoregressive term at the first iteration.

With respect to the term 'oversampling' the following remark should be kept in mind. The sampling time should be essentially decided on the basis of the excitation/information characteristics of the data. However, it is not infrequent that the identification data are correctly sampled with respect to their information level, but result 'over-sampled' with respect to the prediction problem (see, e.g. the example in § 2 with low frequency input and the arch dam case in § 5), which turns out to be numerically ill-conditioned, thereby hampering the model structure selection and leading to redundant models. In such cases, data decimation would surely lead to a good predictor model, but would delete significant information in the data. On the other hand, if simulation performance is evaluated in the identification process, the model structure selection problem can be solved efficiently.

3.4. *Model reduction techniques*

The possible presence of redundant terms in the final model obtained with the FROE algorithm calls for model reduction. This is usually performed *a posteriori*, based on parameter standard deviation (STD) and cluster analysis. Unfortunately, such a procedure turns out to be extremely subjective and difficult to systematize. For instance, in the example discussed in § 2 the information given by the STD is extremely clear in the third case but fuzzier in the other two. There also cases in which the elimination of regressors whose parameters have a high value of STD actually worsens the model simulation performance significantly (see the application example in § 5). This happens, for example, when identification is conducted on an under-parameterized model family: many spurious regressors are inserted in the model to compensate for the absence of some of the correct terms, which are not in the regressor set, and often their estimated parameters are statistically non-significant, in spite of their being essential for good model quality. As for cluster analysis, its effectiveness is still debated in the literature. Initially, it was suggested that terms belonging to a cluster with a cluster coefficient much smaller than the parameters of the corresponding regressors should be eliminated from the model (Aguirre and Billings 1995 a). More recent research has pointed out that there are cases in which the presence of nearly compensating terms of the same cluster is in fact necessary, and that these terms should be kept in the model (Aguirre and Jácome 1998).

4. Extensions to the FROE algorithm

Based on the previous results, it should be clear that the problem of the structure selection in simulation or long-term prediction should be assessed considering the simulation performance index explicitly. Also, model

reduction must be performed on a more objective basis and with an automatic procedure.

4.1. *The SRR criterion*

In order to select the correct model structure (at least when the system belongs to the model family) a criterion for regressor selection with the following properties should be employed:

(1) At every iteration the term with the highest value of the criterion should be one of the regressors actually present in the system.

(2) At every iteration regressors which are not necessary to explain the dynamics of the actual system should have a comparably low value of the criterion with respect to correct regressors.

(3) Terms corresponding to correct regressors should have comparable values of the criterion.

A criterion which matches these requirements more closely than the ERR is the *simulation error reduction ratio* (*SRR*) criterion, defined as the decrease in MSSE obtained by the inclusion of a regressor in the model normalized with the output variance:

$$[\mathrm{SRR}]_j = \frac{\mathrm{MSSE}(\mathcal{M}_i) - \mathrm{MSSE}(\mathcal{M}_{i+1})}{\frac{1}{N}\sum_{t=1}^{N} y^2(t)}$$

where $\mathcal{M}_i$ is the model obtained at the ith iteration and $\mathcal{M}_{i+1}$ is the candidate model at the subsequent iteration, with the inclusion of the jth regressor. The SRR criterion is defined in close analogy to the ERR, which can be rewritten as

$$[\mathrm{ERR}]_j = \frac{\mathrm{MSPE}(\mathcal{M}_i) - \mathrm{MSPE}(\mathcal{M}_{i+1})}{\frac{1}{N}\sum_{t=1}^{N} y^2(t)},$$

and is used in a similar way in the identification algorithm, i.e. adding to the model at each iteration the regressor with the highest value of the criterion. Unfortunately, there does not exist a convenient computation formula for the SRR, such as (5).

To verify the efficiency of the SRR criterion as opposed to the ERR, consider the example

$$\begin{aligned} \mathcal{S}: \quad w(t) &= 0.75w(t-2) + 0.25u(t-1) \\ &\quad - 0.2w(t-2)u(t-1) \\ y(t) &= w(t) + \xi(t), \quad \xi(\cdot) \sim \mathrm{WGN}(0, 0.02) \end{aligned}$$

where the NARX model family with order 2 and nonlinearity degree 3 is employed in identification and the input signal is a low frequency AR(2) process (poles in 0.9 and 0.95), with 0 mean and variance 0.25. The following figure 1 shows the values of the two criteria for each regressor in the first four iterations of the FROE.

First of all, note that the ERR values are always positive (the prediction error always decreases with the inclusion of a new regressor), whereas SRR values can have positive or negative sign: some regressors are explicitly marked as unsafe. This is reasonable, since regressors not related to system dynamics can introduce significant variations in the model behaviour, thereby worsening the simulation performance.

A second difference is represented by the different scale of the two criteria, since in general the values of the SRR associated to regressors which are actually present in the system are greater by one or two orders than the corresponding ERR, thus reducing the possibility of confusion between regressors.

Moreover, SRR values of terms of the same cluster are in general comparable, especially if the input signal mostly excites the low frequency dynamics. Thus, one cluster at a time is selected on the basis of the SRR criterion: in our case, the regressors of cluster Ω_u are chosen at the first iteration, then those of Ω_y, Ω_{yu}, etc. Due to this feature of the SRR criterion, an error in the structure selection is generally less critical, since at least the cluster is the right one.

With the exception of the first iterations, the SRR values for regressors which are not actually present in the system are very low or negative, whereas the ERR values are comparably high (see, e.g. the high degree terms in the considered example). This is a somewhat general feature of the SRR criterion, in the sense that regressors with consistently negative values of SRR during iterations typically belong to spurious clusters. One exception to this phenomenon is represented by the terms of cluster Ω_y at the first iteration: here the negative SRR values imply that 1-term models should not contain these terms, or, stated otherwise, that they are of no use *by themselves alone* in characterizing the system dynamics.

Note that these two characteristics of the SRR may be used to reduce the computational time, as discussed in § 4.4.

Finally, the SRR criterion gives indirect indication of the fact that the optimal model is selected: at the fourth iteration most of the remaining regressors have negative or extremely low SRR, so that no significant increase can be obtained in the model simulation accuracy.

Table 3 reports the regressors actually selected according to the SRR criterion in the first three iterations. Note that while regressors $u(t-1)$ and $y(t-2)$ have comparable importance for both criteria, the third regressor displays a much smaller ERR value compared to the corresponding SRR value: its elimination worsens the prediction performance by 1%, whereas a 10% accuracy decrease in simulation would be experienced.

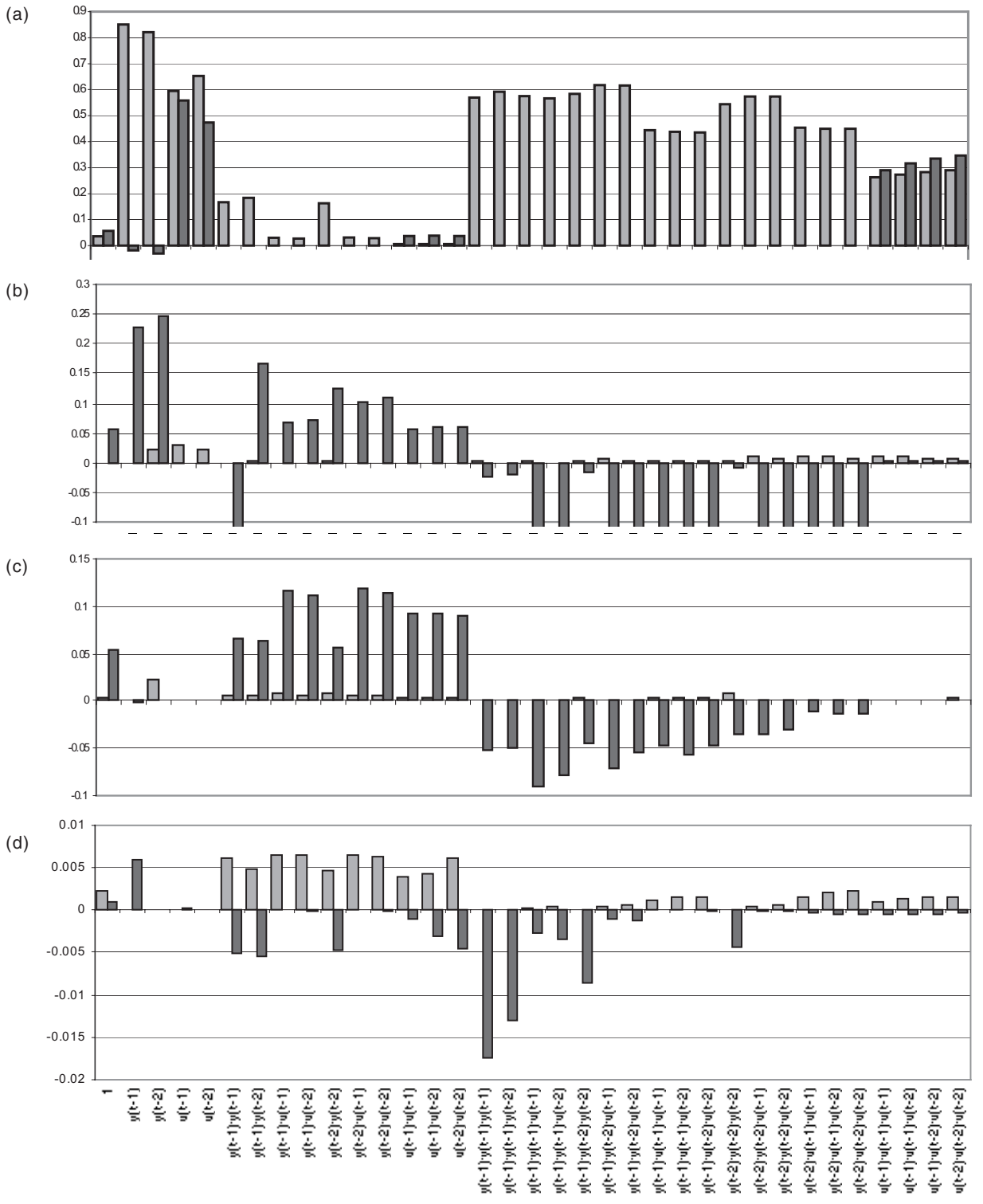

Figure 1. Value of the regressor selection criteria (ERR in light grey, SRR in dark grey): (a) 1st iteration, (b) 2nd iteration, (c) 3rd iteration, (d) 4th iteration.

If the regressors are chosen according to the ERR criterion (table 4), the first regressor has a negative SRR (the autoregressive term alone is useless in simulation) compensated by the second regressor: summing the corresponding SRR values one obtains a simulation accuracy (0.783 83) nearly equal to that achieved by the first two regressors selected according to the SRR (0.805 7). In fact, at the second iteration the models

Iter.	Selected regressor (SRR)	SRR	ERR
1	**u(t − 1)**	5.577 E−01	5.574 E−01
2	**y(t − 2)**	2.480 E−01	3.417 E−01
3	**y(t − 2)·u(t − 1)**	1.188 E−01	1.425 E−02

Table 3. Selected regressors according to the SRR criterion and corresponding values of both criteria.

Iter.	Selected regressor (ERR)	SRR	ERR
1	y(t − 1)	−1.867 E−02	8.741 E−01
2	**u(t − 1)**	8.025 E−01	1.988 E−01
3	**y(t − 2)**	2.047 E−02	2.224 E−02
	[y(t − 2)·u(t − 1)]	[1.128 E−01]	[6.582 E−03]

Table 4. Selected regressors according to the ERR criterion and corresponding values of both criteria.

identified with the two criteria are equivalent at least in terms of clusters. However, the ERR criterion wrongly picks out $y(t-2)$ as third regressor, instead of the much more significant non-linear term $y(t-2)u(t-1)$.

4.2. *The pruning procedure*

The actual importance of the included regressors may change as the structure selection process evolves, both with the ERR and the SRR criterion: regressors initially evaluated as essential, might later be compensated by other more significant regressors. Therefore, to improve the regressor selection process, those regressors which do not contribute significantly to the quality of the model must be eliminated. The proposed deletion procedure is called 'pruning', a term borrowed from the neural networks' literature to indicate a systematic method for model dimension reduction.

In principle, the least significant regressor in a model (i.e. a possible candidate for elimination) is the one whose elimination from the model minimally increases the given performance index. Based on this observation, in the proposed approach, regressor selection and elimination are coupled at each iteration step of the identification algorithm as follows. The iteration starts with the inclusion of a new regressor in the current model. Then, the following iterative pruning procedure is applied:

(1) For each selected regressor, the sub-model obtained after its elimination from the model is considered, its parameters re-estimated and the corresponding performance index computed.

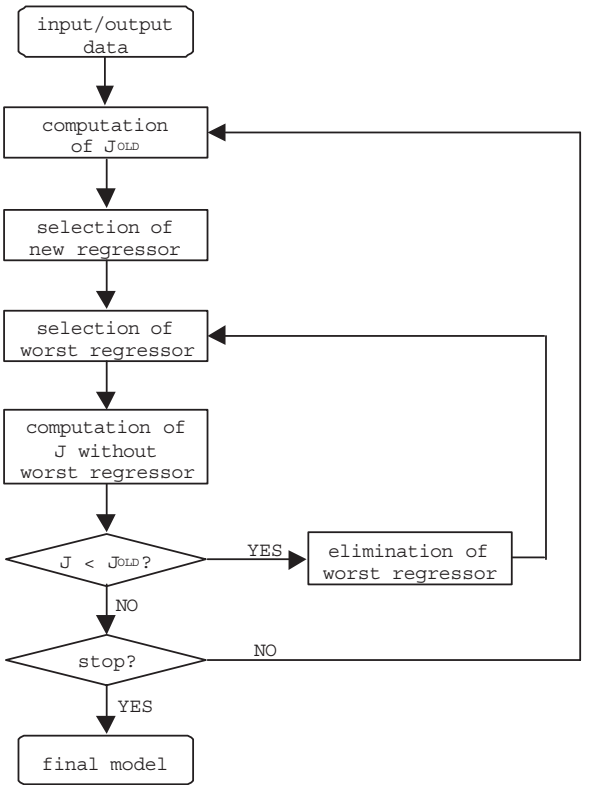

Figure 2. Flow diagram of the identification algorithm.

(2) The least significant regressor is identified as the one corresponding to the sub-model with the best value of the performance index.

(3) If the same sub-model is also better than the model obtained at the previous iteration, the regressor is actually eliminated.

(4) If a regressor elimination has taken place, a new check for eliminable regressors is done (step 1), otherwise the iteration ends.

A complete iteration of the identification algorithm either adds a new regressor to the current model or substitutes one or more of its terms with it, provided that this exchange of terms improves the model accuracy. The rationale behind the algorithm is that, given two models with the same performance index, the smaller one is to be preferred (Occam's Razor principle). In addition, the value of the performance index is always decreasing between iterations (if regressor selection and deletion use the same performance index J). A flow diagram of the identification algorithm is provided in figure 2.

The pruning mechanism is particularly effective when a correct regressor is inserted: typically, a number of spurious compensating regressors is eliminated. Notice that the pruning mechanism may occasionally

eliminate a correct regressor, but this will be typically reintroduced in a later stage.

4.3. *Identification algorithms*

Three new identification algorithms are obtained by combining the two proposed extensions.

By substituting the ERR criterion with the SRR criterion in the FROE algorithm, the *simulation error minimization* (*SEM*) identification algorithm is obtained, which operates the model structure selection directly on the basis of the minimization of the simulation error. In order to limit the computational burden, the parameter estimation of each model structure is still performed with least squares, thus minimizing the model 1-step ahead *prediction* error variance. During the identification experiments carried out, it appeared that parameter estimation based on simulation error minimization did not significantly increase the simulation performance of the identified models, while significantly enhancing the computational load.

The pruning algorithm may be applied in combination with both criteria: if the selection is carried out through the use of ERR and J is the MSPE the *forward-regression with pruning* (*FRP*) is obtained; correspondingly if the selection is carried out through the use of SRR and J is the MSSE the *simulation error minimization with pruning* (*SEMP*) is obtained.

To evaluate the effectiveness of the proposed algorithms consider again the three systems $\mathcal{S}_1$, $\mathcal{S}_2$ and $\mathcal{S}_3$ subject to the zero mean, low frequency AR(2) process with unit variance. In each case, the identification procedure has been stopped when the performance index (MSPE or MSSE) does not significantly decrease anymore. The results are those expected on the basis of the issues pointed out in § 3.

Consider the SEM algorithm first. The different approach to identification influences the model construction right from the beginning: as shown in table 5, even with low frequency input data, the initial regressors are relative to the input variable rather than output terms, and capture the fundamental input–output dynamics of the respective systems. The initial guess models are all of the NFIR type, whereas the FROE algorithm always finds an autoregressive model. The final model for system $\mathcal{S}_1$ has all the terms of the currect clusters, while that for system $\mathcal{S}_2$ contains two redundant terms. Finally, the exact model is obtained for system $\mathcal{S}_3$.

The effect of the pruning procedure can be evaluated by inspecting table 6, where the models identified with the FRP are reported. As for system $\mathcal{S}_1$, the FRP identifies a model very similar to that obtained with the FROE. This time, the spurious regressor $y(t-2)$ is eliminated automatically. Unfortunately, the FRP is not able to remove the spurious autoregressive part, since the predictive approach itself is to blame for it. On the other hand, the correct structure is found for the other systems. In particular, in the case of system $\mathcal{S}_2$ the pruning mechanism has successfully eliminated all those regressors that were absent from the system's input–output relation and that could not be removed on

System	Iter.	Regressors	Estimated parameters	3 · STD%	SRR
$\mathcal{S}_1$	1	u(t − 2)	4.854 E−01	11.5%	4.561 E−01
	2	**u(t − 1) ·** **u(t − 2)**	1.810 E−01	7.5%	3.019 E−01
	3	u(t − 1)² · u(t − 2)	1.171 E−01	15.3%	1.320 E−01
	4	**u(t − 1)**	−9.915 E−01	9.8%	5.449 E−02
$\mathcal{S}_2$	1	**u(t − 1)²**	1.169 E+00	19.3%	6.666 E−01
	2	**1**	2.579 E−01	35.1%	1.857 E−01
	3	**u(t − 2)**	7.274 E−01	32.7%	1.165 E−01
	4	*y(t − 2) · u(t − 1)*	3.818 E−02	347.0%	6.917 E−03
	5	*y(t − 2)*	1.004 E−01	149.4%	2.121 E−03
	6	**y(t − 2)²**	−8.915 E−02	38.8%	2.486 E−03
	7	**y(t − 1)**	7.071 E−01	17.2%	1.611 E−03
$\mathcal{S}_3$	1	**u(t − 1)²**	9.999 E−01	4.6%	7.797 E−01
	2	**u(t − 2)**	8.114 E−01	5.3%	1.232 E−01
	3	**1**	4.891 E−01	10.9%	6.422 E−02
	4	**y(t − 2)²**	−5.028 E−02	8.3%	5.753 E−03
	5	**y(t − 1)**	5.106 E−01	8.4%	8.336 E−03

Table 5. Selected models for systems $\mathcal{S}_1$, $\mathcal{S}_2$ and $\mathcal{S}_3$ ($u(\cdot) \sim$ AR) with the SEM algorithm.

System	Iter.	Regressors	Estimated parameters	3 · STD%
$\mathcal{S}_1$	1	y(t − 1)	5.961 E−01	15.5%
	2	**u(t − 1)**	1.296 E+00	12.2%
	3	u(t − 1)²	1.001 E−01	26.2%
	4	**u(t − 1)³**	−3.935 E−01	17.7%
	5	u(t − 1)·u(t − 2)²	2.815 E−01	28.7%
	6	**u(t − 2)**	−6.898 E−01	32.2%
$\mathcal{S}_2$	1	**y(t − 1)**	4.962 E−01	9.1%
	2	**u(t − 1)²**	1.051 E+00	13.4%
	3	**y(t − 2)²**	−5.299 E−02	18.2%
	4	**1**	4.838 E−01	24.9%
	5	**u(t − 2)**	7.805 E−01	14.1%
$\mathcal{S}_3$	1	**y(t − 1)**	5.106 E−01	8.4%
	2	**u(t − 1)²**	9.999 E−01	4.6%
	3	**y(t − 2)²**	−5.028 E−02	8.3%
	4	**u(t − 2)**	8.114 E−01	5.3%
	5	**1**	4.891 E−01	10.9%

Table 6. Selected models for systems $\mathcal{S}_1$, $\mathcal{S}_2$ and $\mathcal{S}_3$ ($u(\cdot) \sim$ AR) with the FRP algorithm.

System	Iter.	Regressors	Estimated parameters	3 · STD%
$\mathcal{S}_1$	1	**u(t − 1)·u(t − 2)**	2.626 E−01	8.0%
	2	**u(t − 1)**	1.050 E+00	9.6%
	3	**u(t − 2)**	4.442 E−01	20.8%
	4	**u(t − 1)^3**	−3.015 E−01	5.7%
$\mathcal{S}_2$	1	**1**	4.838 E−01	24.9%
	2	**u(t − 2)**	7.805 E−01	14.1%
	3	**u(t − 1)^2**	1.051 E+00	13.4%
	4	**y(t − 2)^2**	−5.299 E−02	18.2%
	5	**y(t − 1)**	4.962 E−01	9.1%
$\mathcal{S}_3$	1	**1**	9.999 E−01	4.6%
	2	**u(t − 2)**	8.114 E−01	5.3%
	3	**u(t − 1)^2**	4.891 E−01	10.9%
	4	**y(t − 2)^2**	−5.028 E−02	8.3%
	5	**y(t − 1)**	5.106 E−01	8.4%

Table 7. Selected models for system $\mathcal{S}_1$, $\mathcal{S}_2$ and $\mathcal{S}_3$ ($u(\cdot) \sim$ AR) with the SEMP algorithm.

System	Iter.	Selected regressors	Eliminated regressors	SRR
$\mathcal{S}_1$	1	**u(t − 2)**		4.561 E–01
	2	**u(t − 1)·u(t − 2)**		3.019 E–01
	3	u(t − 1)^2·u(t − 2)		1.320 E–01
	4	**u(t − 1)**	**u(t − 2)**	1.002 E–02
	5	**u(t − 2)**		4.447 E–02
	6	**u(t − 1)^3**	u(t − 1)^2 · u(t − 2)	1.037 E–02
	7	y(t − 2)·u(t − 2)^2		2.254 E–04
	8	y(t − 2)^2·u(t − 2)	y(t − 2) · u(t − 2)^2	5.806 E–05
	9	y(t − 2)·u(t − 2)^2		9.612 E–04
	10	1		2.952 E–04

Table 8. SEMP iterations on the data of system $\mathcal{S}_1$.

the basis of statistical significance. Notice that the SEM algorithm performed worse in this case and better on $\mathcal{S}_1$: this shows that the SRR criterion and the pruning procedure have different effects, and that it is worthwhile to combine them in the SEMP.

Finally, the SEMP algorithm was able to identify the correct process model in each case (see table 7). More in detail, consider the first 10 iterations on the data of system $\mathcal{S}_1$ reported in table 8. The corresponding evolutions of the MSSE and the model size are shown in figure 3: a good compromise between model accuracy and size is achieved at iteration 6, when the exact model is identified. Besides, the value of the SRR criterion significantly decreases from the 6th iteration on. Analogous results are obtained for the other two systems, where the exact model is retrieved in few iterations.

Other tests with under-parameterized model families have shown that the SEMP algorithm is generally more successful in identifying compact and robust models even when the exact model structure is not replicable.

4.4. *Computational issues related to simulation error minimization*

The obvious drawback of a simulation error based algorithm such as the SEM or the SEMP is the computational load implied by the simulation cycle calculations, which must be performed for each candidate regressor in the selection phase and for each model regressor in the pruning phase. For example, the SEMP identification in the case of system $\mathcal{S}_1$ (500 data, 35 regressors, 10 iterations) requires approximately 5 min on a Pentium III 600 MHz. 90% of this computational time is devoted to simulation cycle calculations (374 model evaluations, each requiring 0.7 s on average). The required computational time can be significantly reduced by means of the following modifications:

(1) *Selective stopping of the simulation cycle computation*. It is not always necessary to complete the cycle computation on the full set of data for all candidate models: if the partial MSSE(M) after M simulation steps ($M < N$) is such that $M/N \cdot$ MSSE(M) is greater than the overall MSSE of the best model already found, the candidate model is surely worse and the computation can be stopped. In practice, this operation can enhance the computational performance by a 10–30% factor (see § 5). The actual improvement depends on the system characteristics and on the amplitude of the candidate regressor set.

(2) *Joint SRR-cluster analysis*. The value of the SRR criterion is comparable for regressors belonging to the same cluster, especially with 'slow' inputs and 'short' regressor lags (see, e.g. figure 1). Thus the computation load may be dramatically decreased by examining the SRR only for a single regressor in each cluster. Then the SRR analysis is repeated only for the elements belonging to the cluster with the highest value of SRR.

(3) *Reduction of the candidate regressor set*. The regressor selection phase can be speeded up by reducing the candidate regressor set. This can be done, for example, on the basis of past SRR values: regressors which repeatedly present negative SRR values can be (temporarily or permanently) eliminated in successive iterations.

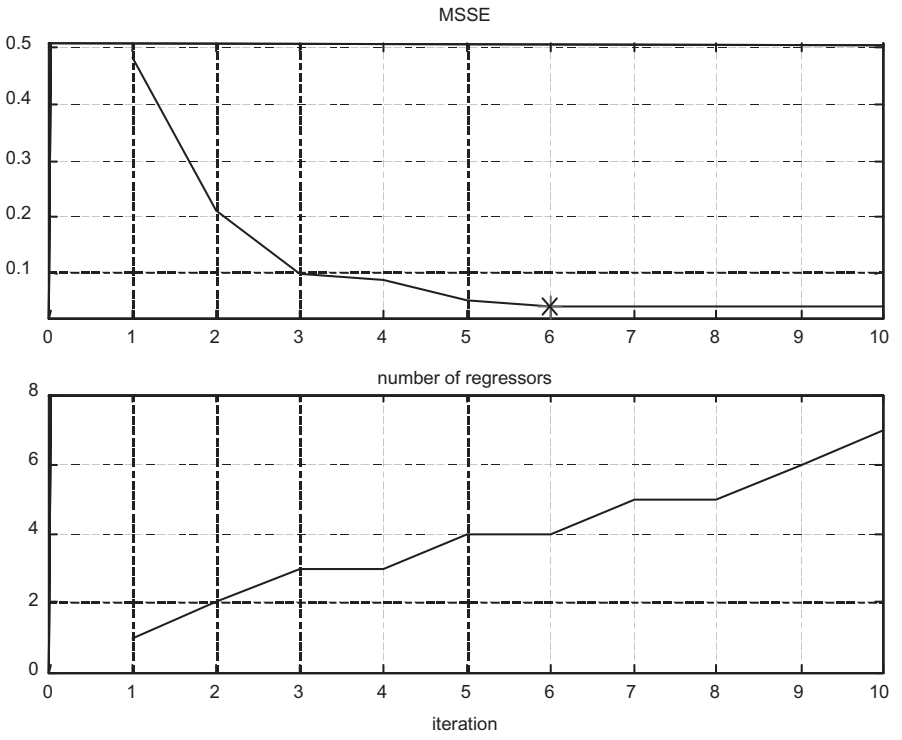

Figure 3. Stop criteria for the SEMP algorithm (system $\mathcal{S}_1$).

(4) *Reduction of the number of the model evaluations.* In the regressor selection phase all the candidate regressors are tested for possible inclusion in the model and the best one is chosen. To speed up the algorithm it is possible to establish a threshold value for the minimal MSSE improvement in order to accept a regressor: instead of examining *all* the candidate regressors and choosing the best one, the evaluation stops when the first regressor satisfying a given threshold, e.g. a 5% improvement, is found. Clearly, the value of this threshold becomes an important tuning knob in the algorithm: the smaller it is, the shorter the iteration. However, shorter iterations mean more identification steps and longer pruning operations.

With reference to the SEMP identification in the case of system $\mathcal{S}_1$, the selective stopping of the simulation cycle computation has reduced the overall computation time by a 33% factor (200 s). The joint SRR-cluster analysis is also very effective: it reduces the total model evaluations to 193 (52% reduction).

5. A case study

The SEMP algorithm has been successfully employed in a number of experimental applications with real data (Piroddi *et al.* 2001, Leva and Piroddi 2002). In these cases, the assumption that the exact model structure is of the NARX family is obviously unrealistic. Nevertheless, quite robust and accurate models can be derived.

In the following, the identification of a NARX model for a long-range prediction application will be described. The data analysis problem, originally formulated for the *6th Benchmark Workshop on Numerical Analysis of Dams* (Salzburg, Austria, 17–19 October, 2001), concerns the numerical modelling of the Schlegeis Arch Dam behaviour in terms of radial crest displacement dynamics, as a function of the water level, six concrete temperature measurements and the air temperature, all available on a daily basis in the period 1992–2000, except the radial crest displacement, which was available only up to 1998. The objective was to predict the radial crest displacements in the target period 1999–2000. Some preliminary results are available in (Palumbo *et al.* 2001).

The set of available inputs is redundant. In particular, as can be easily established through linear analysis, some of the temperatures are only slightly correlated with the radial crest displacement, and the remaining ones are strongly inter-correlated. In view of this, only the water level, one concrete temperature and the air temperature will be considered in the following. These inputs will be denoted $u_i(\cdot)$,

$i = 1, \ldots, 3$, whereas the radial crest displacement is the model output $y(\cdot)$.

Since the level of accuracy obtainable with linear models is not satisfactory (the best linear ARX model yields MSSE = 18.92), the identification of non-linear NARX models is tried, focusing on the following set of elementary regressors: $u_1(t)$, $u_1(t-1)$, $u_1(t-2)$, $u_2(t)$, $u_2(t-1)$, $u_3(t)$, $u_3(t-1)$, $y(t-1)$, $y(t-2)$. Non-linear input terms up to order 3 are included in the regressor set. Mixed terms in the different inputs are discarded. This is motivated by a specification of the benchmark application, which requires the model to be separable into two additive components, related to the hydrostatic effect (u_1) and thermal effect (u_2, u_3), respectively. The regressor set amounts to 40 possible regressors. The identification process with the SEMP algorithm is summarized in table 9.

After six iterations of the SEMP algorithm the simulation performance of the NARX model is already better than the best linear model found. The final model is obtained after 15 iterations: further iterations improve the model only slightly. The pruning procedure has eliminated four regressors during the identification process, so that the resulting model has 11 parameters.

The estimated model (table 10) has a MSSE = 1.76 on the identification data (1992–1998), i.e. one order of magnitude more accurate than the best linear model obtained. Figure 4 displays the performance on the validation data (1999–2000), measured by a MSSE = 1.29, which is 10% better than the more accurate model proposed at the workshop where the problem was originally presented. It is also worth remarking that the best model presented at the workshop required as many as five concrete temperatures, as opposed to the model proposed here, which obtains better performance with just one concrete temperature.

By careful inspection of table 10 one can spot the following possible sources of spurious model terms:

- regressors with high uncertainty

$$u_2(t-1) \quad \rightarrow \quad \text{STD} = 99.7\%$$
$$u_2(t)^2 \quad \rightarrow \quad \text{STD} = 104.3\%$$

- terms of the same cluster which nearly compensate

$$u_1(t), u_1(t-2) \in \Omega_{u_1} \rightarrow \Sigma_{u_1} = 0.0281 \text{ (5\% of the larger cluster term coefficient)}$$
$$u_1(t)u_1(t-2), u_1(t-1)u_1(t-2) \in \Omega_{u_1^2} \rightarrow \Sigma_{u_1^2} = 0.0389 \text{ (7\% of the larger cluster term coefficient)}$$
$$u_1(t)^3, u_1(t-1)^2 u_1(t-2) \in \Omega_{u_1^3} \rightarrow \Sigma_{u_1^3} = 0.000\,36 \text{ (1\% of the larger cluster term coefficient)}$$

These terms, which classical standard deviation and cluster analysis would attempt to eliminate, are in fact necessary terms in the model, in that their elimination greatly affects the model accuracy in simulation. For example, the elimination of $u_2(t-1)$ or $u_2(t)^2$ worsens the MSSE on the identification data by a 35.87% or 24.7% factor, respectively. As for cluster analysis, if both regressors of the $\Omega_{u_1^2}$ cluster were eliminated the model performance would be MSSE = 12.11. Analogous results hold for the other compensating clusters.

It is worth remarking that, in this example, the effect of adopting the MSSE as performance index for structure selection is much more relevant than the usage of the pruning algorithm: the SEM achieves a 12-term

Iter.	Selected regressors	Eliminated regressors	MSSE	SRR
1	$u_1(t-2)$		58.442	6.259 E−01
2	$y(t-2)$		32.242	1.677 E−01
3	$u_3(t-1)$		28.615	2.322 E−02
4	$u_1(t)$	$u_1(t-2)$	25.151	2.218 E−02
5	$u_1(t-1)u_1(t-2)$		19.791	3.432 E−02
6	1		11.138	5.540 E−02
7	$u_1(t-2)$		8.128	1.927 E−02
8	$u_1(t)u_1(t-2)$	$u_1(t-1)u_1(t-2)$	7.780	2.224 E−03
9	$u_1(t-1)u_1(t-2)$		4.113	2.348 E−02
10	$u_2(t)^2$		3.273	5.377 E−03
11	$u_2(t-1)$		2.907	2.347 E−03
12	$y(t-1)$	$y(t-2)$	2.530	2.410 E−03
13	$u_1(t)^3$		2.357	1.113 E−03
14	$u_1(t-1)^2u_1(t-2)$	$u_2(t)^2$	2.200	1.004 E−03
15	$u_2(t)^2$		1.764	2.791 E−03

Table 9. Iterations of the SEMP algorithm on the arch dam data.

Iter.	Regressors	Parameters	STD %
1	$u_3(t-1)$	−1.863 E−02	13.2%
2	$u_1(t)$	5.846 E−01	6.1%
3	1	−5.526 E−03	52.3%
4	$u_1(t-2)$	−5.565 E−01	6.6%
5	$u_1(t)u_1(t-2)$	5.954 E−01	11.4%
6	$u_1(t-1)u_1(t-2)$	−5.868 E−01	11.6%
7	$u_2(t-1)$	2.674 E−03	99.7%
8	$y(t-1)$	9.791 E−01	0.4%
9	$u_1(t)^3$	3.471 E−02	33.0%
10	$u_1(t-1)^2u_1(t-2)$	−3.435 E−02	33.6%
11	$u_2(t)^2$	−1.839 E−03	104.3%

Table 10. Model identified with the SEMP algorithm.

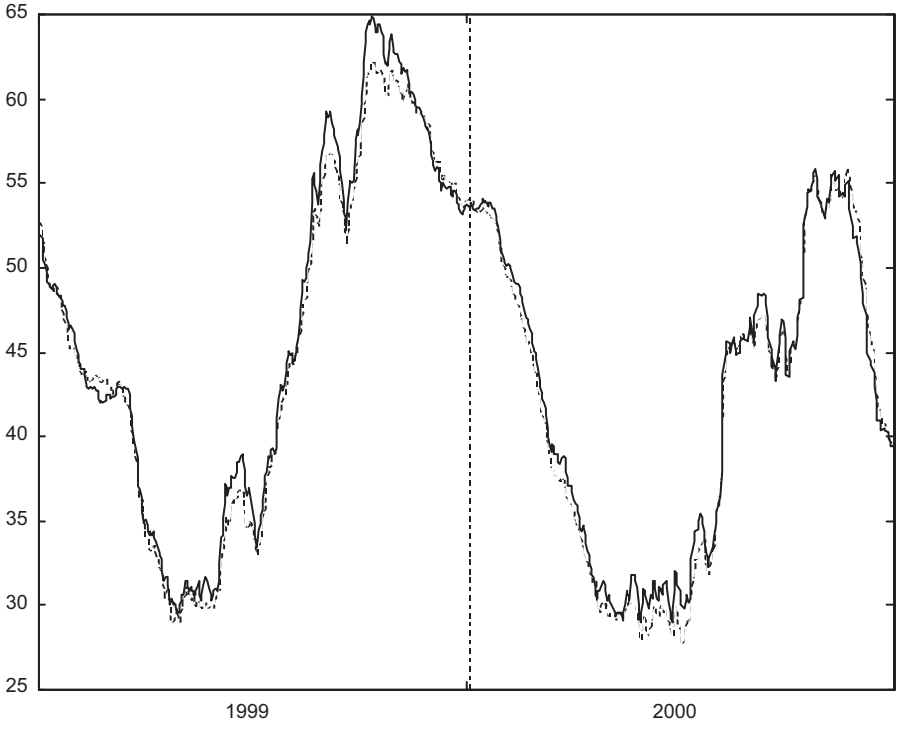

Figure 4. Model performance on the validation data (1999–2000): radial crest displacement (mm) (continuous line) and NARX model simulation output (dashed line).

model, which has a very similar structure, in terms of clusters, to that of table 10, and whose performance on the identification data is 14.34% worse. On the other hand, the FRP obtains a 15-term model, with a MSSE = 2.53 on the identification data (almost 44% worse). Finally, if the original FROE is used, the model accuracy is still low (MSSE = 2.55) at the 20th iteration and many spurious regressors are included in the model.

The model structure identified automatically by the SEMP algorithm is pretty much in agreement with traditional modelling assumptions in the field. Interestingly enough, the source of major differences in the models proposed at the workshop lies in the modelling of the thermal component, where various heuristic functions of the available temperatures are used. This is yet another proof that a systematic structure selection procedure is necessary.

Finally, as for the computational time, the standard SEMP takes about 65 min on a Pentium III 600 MHz to perform 20 iterations on the available 2558 data, operating the structure selection on a 40 term regressor set. Of this time 90% is employed for the 834 model evaluations in simulation. The simulation cycle computation can be reduced from 4 to 3 s on average by means of the selective stopping, lowering the computational time to 50 min. This can be further reduced to about 30 min with the joint SRR-cluster analysis (509 model evaluations).

6. Conclusions

A critical analysis of classical prediction error approaches to the identification of NARX/NARMAX models has pointed out some drawbacks, particularly if long-range prediction or simulation models are desired. To overcome these limitations, a new algorithm has been devised which employs a combination of two factors: a pruning mechanism to keep the model dimension small during iterations, and a simulation error based criterion for regressor selection. The proposed algorithm generally obtains compact and robust models, both in simulation examples and with real practice data. When the system generating the identification data belongs to the model family, the algorithm is particularly effective in the reconstruction of the correct model structure.

The identification algorithm requires a significant computational time, but this can be considerably reduced using several heuristic methods in the structure selection phase, some of which are suggested in the paper.

As a final remark, note that the proposed (simulation-based) approach is not viable for the identification of chaotic systems, owing to the extreme sensitivity of their behaviour to initial conditions.

Acknowledgement

Paper supported by MURST project 'Identification and Control of Industrial Systems'.

References

Aguirre, L. A., 1994, Some remarks on structure selection for nonlinear models. *International Journal of Bifurcation and Chaos*, **4**, 1707–1714.

Aguirre, L. A., and Billings, S. A., 1995 a, Improved structure selection for nonlinear models based on term clustering. *International Journal of Control*, **62**, 569–587.

Aguirre, L. A., and Billings, S. A., 1995 b, Dynamical effects of overparametrization in nonlinear models. *Physica D*, **80**, 26–40.

Aguirre, L. A., and Jácome, C. R. F., 1998, Cluster analysis of NARMAX models for signal-dependent systems. *IEE Proceedings—Control Theory Applications*, **145**, 409–414.

Billings, S. A., and Aguirre, L. A., 1995, Effects of the sampling time on the dynamics and identification of nonlinear models. *International Journal of Bifurcation and Chaos*, **5**, 1541–1556.

Billings, S. A., Chen, S., and Korenberg, M. J., 1989, Identification of MIMO non-linear systems using a forward-regression orthogonal estimator. *International Journal of Control*, **49**, 2157–2189.

Billings, S. A., and Tsang, K. M., 1989, Spectral analysis for non-linear systems—Part I: parametric non-linear spectral analysis—Part II: interpretation of non-linear frequency response functions. *Mechanical Systems and Signal Processing*, **3**, 319–359.

Billings, S. A., Tsang, K. M., and Tomlinson, G. R., 1990, Spectral analysis for non-linear systems—Part III: case study examples. *Mechanical Systems and Signal Processing*, **4**, 3–21.

Chiras, N., Evans, C., and Rees, D., 2001, Nonlinear gas turbine modeling using NARMAX structures. *IEEE Transactions on Instrumentation and Measurement*, **50**, 893–898.

Kalkkuhl, J., and Liccaga-Castro, E., 1996, Nonlinear control based on the NARX plant representation. *Proceedings IEEE International Symposium on Industrial Electronics, ISIE '96*, **1**, 115–120.

Korenberg, M., Billings, S. A., Liu, Y. P., and Mcilroy, P. J., 1987, Orthogonal parameter estimation algorithm for non-linear stochastic systems. *International Journal of Control*, **48**, 193–210.

Leontaritis, I. J., and Billings, S. A., 1985 a, Input–output parametric models for non-linear systems—Part I: deterministic non-linear systems. *International Journal of Control*, **41**, 303–328.

Leontaritis, I. J., and Billings, S. A., 1985 b, Input–output parametric models for non-linear systems—Part II: stochastic non-linear systems. *International Journal of Control*, **41**, 329–344.

Leontaritis, I. J., and Billings, S. A., 1987 a, Model selection and validation methods for non-linear systems. *International Journal of Control*, **45**, 311–341.

Leontaritis, I. J., and Billings, S. A., 1987 b, Experimental design and identifiability for non-linear systems. *International Journal of Systems Science*, **18**, 189–202.

Leva, A., and Piroddi, L., 2002, NARX-based technique for the modeling of magneto-rheological damping devices. *Smart Materials and Structures*, **11**, 79–88.

Loh, C. H., and Duh, J. Y., 1996, Analysis of nonlinear system using NARMA models. *JSCE—Structural Engineering/Earthquake Engineering*, **13**, 11–21.

Maio, I. A., Stievano, I. S., and Canavero, F. G., 1998, NARX approach to black-box modeling of circuit elements. *Proceedings IEEE International Symposium on Circuits and Systems, ISCAS 98*, **3**, 411–414.

Mao, K. Z., and Billings, S. A., 1997, Algorithms for minimal model structure detection in nonlinear dynamic system identification. *International Journal of Control*, **68**, 311–330.

Palumbo, P., and Piroddi, L., 2000 a, Harmonic analysis of nonlinear structures by means of Generalized Frequency Response Functions coupled with NARX models. *Mechanical Systems and Signal Processing*, **14**, 243–265.

Palumbo, P., and Piroddi, L., 2000 b, Seismic behaviour of buttress dams: nonlinear modelling of a damaged buttress based on ARX/NARX models. *Journal of Sound and Vibration*, **239**, 405–422.

Palumbo, P., Piroddi, L., Lancini, S., and Lozza, F., 2001, NARX modelling of radial crest displacements of the Schlegeis Arch Dam. *6th Benchmark Workshop on Numerical Analysis of Dams*, Salzburg, Austria, 17–19 October.

Peyton-Jones, J. C., and Billings, S. A., 1989, Recursive algorithm for computing the frequency response of a class of non-linear difference equation model. *International Journal of Control*, **50**, 1925–1940.

Piroddi, L., Feriani, A., and Lozza, F., 2001, Black-box modelling of an anti-seismic isolator. *5th World Congress on Joints, Bearings and Seismic Systems for Concrete Structures, JBSS '01*, Rome, Italy, 7–11 October.

Pottmann, M., Unbehauen, H., and Seborg, D. E., 1993, Application of a general multi-model approach for identification of highly nonlinear processes—A case study. *International Journal of Control*, **57**, 97–120.

Sjöberg, J., Zhang, Q., Ljung, L., Benveniste, A., Delyon, B., Glorennec, P., Hjalmarsson, H., and Juditsky, A., 1995, Nonlinear black-box modeling in system identification: a unified overview. *Automatica*, **12**, 1691–1724.