Improved Structure Detection For Polynomial NARX Models Using a Multiobjective Error Reduction Ratio

**Samir Angelo Milani Martins · Erivelton Geraldo Nepomuceno ·
Márcio Falcão Santos Barroso**

Abstract This paper addresses the problem of structure detection for polynomial NARX models. It develops MERR, a multiobjective extension of a methodology well-known as the error reduction ratio (ERR). It is shown that it is possible to choose terms which take into account dynamics of prediction error and other types of affine information, such as fixed points or static curve. Two examples are included to illustrate the proposed methodology. A numerical example shows that the technique is able to reconstruct the structure of a system, known *a priori*. The identification of a pilot DC–DC buck converter shows that the proposed approach is capable to find models valid over a wide range of operation points. In this latter example, MERR is compared with ERR in two forms: (i) affine information is applied only in the structure selection for MERR and (ii) affine information is applied for structure selection for MERR and for parameter estimation for both MERR and ERR. In both comparisons, MERR presented nondominated solutions of Pareto set.

Keywords Multiobjective system identification ·
NARX models · Structure detection

S. A. M. Martins (✉) · E. G. Nepomuceno · M. F. S. Barroso
Department of Electrical Engineering, Federal University of São João del-Rei, Frei Orlando Square, 170-Center, São João del-Rei, MG 36307-352, Brazil
e-mail: martins@ufsj.edu.br

E. G. Nepomuceno
e-mail: nepomuceno@ufsj.edu.br

M. F. S. Barroso
e-mail: barroso@ufsj.edu.br

1 Introduction

Despite major advances in system identification, model structure detection is still a great challenge (Baldacchino et al. 2012). An approach used to identify black-box models, such as polynomial nonlinear auto-regressive with exogenous inputs (NARX) (Leontaritis and Billings 1985), consist in choosing a predefined number of model terms in a larger set of candidate terms (Korenberg et al. 1988; Mendes and Billings 2001). Once the model increases its maximum nonlinearity degree, the search space is enlarged, increasing the complexity of structure detection. Considering a nonlinear polynomial NARX model with 10 as its maximum degree and input/output maximum lag, it would yield to a set of more than 30 million candidate models in the search space. Among them, each regressor may represent a specific system behaviour or not, being classified either as genuine or spurious (Aguirre and Billings 1994, 1995).

Traditionally, structure detection uses only real dynamic data, acquired from a test station (Ljung 1987). Piroddi and Spinelly (2003) have developed an error-based simulation technique for structure detection, using dynamic data whereas Korenberg et al. (1988) have used the prediction error. In Baldacchino et al. (2012), a unified framework is developed where the whole model (parameter and structure) is obtained simultaneously, based only in dynamic data. In this sense, with a view to data acquisition, the process has to be excited by a persistently exciting input, which is not always possible.

Besides, there are many systems where the dynamic data are hampered by noise or very hard to acquire (Aguirre et al. 2000). Furthermore, all information about the process to be modeled may not be included in these data. Hence, it is important to develop new techniques that include dynamic

information and other types of information during the structure detection procedure, which is still lacking in science.

Johansen (1996) focused his research on the inclusion of auxiliary information in the parameter estimation. Information beyond dynamical data was used to compose the model, giving rise to the so-called multiobjective system identification (Nepomuceno et al. 2007; Barroso et al. 2007). Such types of information are generally used for parameter estimation, a step already consolidated (Zhao et al. 2011; Wei and Billings 2009; Previdi and Lovera 2004). Structure detection techniques, just like the ones cited above, are mostly mono-objective (Cantelmo and Piroddi 2010; Wei and Billings 2008), considering only dynamic data to select genuine regressors. The incorporation of further information in the structure detection stage is the main aim of this work.

This paper presents the *multiobjective error reduction ratio* (MERR), a multiobjective procedure for structure detection of polynomial NARX models. The methodology presented is an extension of the technique widely used, the *error reduction ratio* (ERR) (Korenberg et al. 1988) which considers only dynamic data to select models structures. When more information is available rather than only dynamic data, MERR allows an alternative way to get models structures with compromise among the objectives, i.e., regarding several types of information about the system (input/output dynamic data, static curve, and fixed points).

This paper is organized as follows. The first section has introduced the subject and has presented the paper in an overall point of view. Section 2 presents some background concepts used on this work. Section 3 shows the methodology proposed to develop MERR and used to obtain the results, which are shown in Sect. 4. Two examples are included to illustrate the proposed methodology: (i) a numerical example based on a discrete-time nonlinear SISO polynomial NARX system and (ii) a pilot DC–DC buck converter with affine static and nonlinear dynamic. Section 5 presents conclusions and future research perspectives.

2 Nonlinear System Representation

2.1 The ERR

Consider the polynomial NARX (Leontaritis and Billings 1985) model described by Eq. (1)

$$y(k) = F^{\ell}\left[y(k-1), \ldots, y(k-n_y), u(k-1), \ldots, u(k-n_u)\right], \quad (1)$$

where n_y and n_u are the maximum lags considered for the process and input terms, respectively. Moreover, $y(k)$ is a time series of the output while $u(k)$ is a time series of the input. $F^{\ell}[\cdot]$ is some nonlinear function of $y(k)$ and $u(k)$.

In this paper $F^{\ell}[\cdot]$ is taken to be a nonlinear polynomial of degree $\ell \in \mathcal{Z}^+$. In order to estimate the parameters of such a polynomial, Eq. (1) can be expressed as follows:

$$y(k) = \psi^{\mathrm{T}}(k-1)\hat{\theta} + \xi(k) = \sum_{i=1}^{n_\theta} \hat{\theta}_i \psi_i(k-1) + \xi(k), \quad (2)$$

where $\psi(k-1)$ is the vector of regressors (independent variables) that contains linear and nonlinear combinations of output and input terms up to and including time $(k-1)$, T operator is a vector transposition, θ_i is ith parameter of ith regressor ψ_i, and n_θ is the number of terms of the model. The parameters corresponding to each term in such matrices are the elements of the vector $\hat{\theta}$. Finally, $\xi(k)$ is the residual or prediction error at time k which is defined as the difference between the measured data $y(k)$ and the one-step-ahead prediction $\psi^{\mathrm{T}}(k-1)\hat{\theta}$. An auxiliary model may be expressed as

$$y(k) = \sum_{i=1}^{n_\theta} \hat{g}_i \mathbf{\Omega}_i(k-1) + \xi(k), \quad (3)$$

where $\hat{g}_i$ is the parameter associated to the regressor $\mathbf{\Omega}_i$, which composes the associated orthogonal model, and $y(k)$ is the output dynamic time series. The orthogonal model can be obtained from the original NARX model by means of Householder transformations.

ERR (Korenberg et al. 1988) can be obtained as an inner product (Chen and Billings 1989). This criterion provides indication of which terms to include in the model by ordering all the candidate terms according to a hierarchy that depends on the relative importance of each term. The average value of inner product of the output data is:

$$\langle \mathbf{y}, \mathbf{y} \rangle = \sum_{i=1}^{n_\theta} \hat{g}_i^2 \langle \mathbf{\Omega}_i, \mathbf{\Omega}_i \rangle + \langle \boldsymbol{\xi}, \boldsymbol{\xi} \rangle, \quad (4)$$

where $\langle \cdot, \cdot \rangle$ is the inner product symbol, $\mathbf{\Omega}_i$ is the i−th orthogonal regressor, and $\hat{g}_i$ its respective parameter. $\mathbf{y}$ is the dynamic output and $\boldsymbol{\xi}$ is the residue.

One of the many advantages of such algorithms is that the ERR (Korenberg et al. 1988) can be easily obtained as a by-product (Chen and Billings 1989). This technique is based on one-step-ahead dynamic prediction error, linking each search space regressor to a corresponding index. This index quantifies the contribution of the regressor to minimize $\langle \boldsymbol{\xi}, \boldsymbol{\xi} \rangle$, which is defined as:

$$\langle \boldsymbol{\xi}, \boldsymbol{\xi} \rangle = \langle \mathbf{y}, \mathbf{y} \rangle - \sum_{i=1}^{n_\theta} \hat{g}_i^2 \langle \mathbf{\Omega}_i, \mathbf{\Omega}_i \rangle. \quad (5)$$

If any regressor was included in the model ($n_\theta = 0$), $\langle \boldsymbol{\xi}, \boldsymbol{\xi} \rangle$ would be exactly the quadratic sum of the output and it would decrease ($g_i^2 \langle \mathbf{\Omega}_i^T \mathbf{\Omega}_i \rangle$) for each regressor $\mathbf{\Omega}_i$ included in the model. This reduction can be normalized, considering the

mean quadratic error, composing the ERR of the ith regressor:

$$\text{ERR}_i = \frac{\hat{g}_i^2 \langle \mathbf{\Omega}_i, \mathbf{\Omega}_i \rangle}{\langle \mathbf{y}, \mathbf{y} \rangle}. \tag{6}$$

The parameters may be calculated by means of

$$\hat{g}_i = \frac{\langle \mathbf{\Omega}_i, \mathbf{y} \rangle}{\langle \mathbf{\Omega}_i, \mathbf{\Omega}_i \rangle}, \quad i = 1, \ldots, n_\theta. \tag{7}$$

After the terms have been ordered by the ERR, an information criterion can be used to help choose a good cut-off point. This work uses the Akaike Information Criterion (AIC) defined as follows (Akaike 1974):

$$\text{AIC}(n_\theta) = N \ln[\sigma_e^2(n_\theta)] + 2n_\theta, \tag{8}$$

where N is the length of the data vector, $\sigma_e^2(n_\theta)$ is the variance of modeling error, and n_θ is the number of model parameters.

2.2 Multiobjective System Identification

Let the definition of Affine Information (Nepomuceno et al. 2007) be:

Definition 1 *(Affine Information)* Consider the parameter vector $\hat{\boldsymbol{\theta}} \in \mathfrak{R}_\theta^n$, a vector $\mathbf{v} \in \mathfrak{R}^N$ and a matrix $G \in \mathfrak{R}^{N \times n_\theta}$. Both $\mathbf{v}$ and G are assumed to be accessible. Moreover, suppose $G\hat{\boldsymbol{\theta}}$ constitutes an *estimate* of $\mathbf{v}$, such that $\mathbf{v} = G\hat{\boldsymbol{\theta}} + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \in \mathfrak{R}^N$ is an error vector. Then $[\mathbf{v}, G]$ is said to be an **affine information** pair of the system. □

Taking (2) over a set of data yields

$$\mathbf{y} = \Psi\hat{\boldsymbol{\theta}} + \boldsymbol{\xi}. \tag{9}$$

According to the definition above, $[\mathbf{y}, \Psi]$ is an *affine information* pair, where $\mathbf{y} \in \mathfrak{R}^N$, $\Psi \in \mathfrak{R}^{N \times n_\theta}$, and $\boldsymbol{\epsilon} = \boldsymbol{\xi}$. The vector $\hat{\boldsymbol{\theta}}$ is usually estimated by minimizing convex functionals of the form

$$J_{\text{LS}}(\hat{\boldsymbol{\theta}}) = \parallel \boldsymbol{\xi} \parallel_2^2 = (\mathbf{y} - \Psi\hat{\boldsymbol{\theta}})^{\text{T}} (\mathbf{y} - \Psi\hat{\boldsymbol{\theta}}), \tag{10}$$

where the last functional is minimized by the least-squares estimator. In a multiobjective approach, the problem is to minimize

$$\mathbf{J}(\hat{\boldsymbol{\theta}}) = \left[J_1(\hat{\boldsymbol{\theta}}) \ldots J_m(\hat{\boldsymbol{\theta}}) \right]^{\text{T}}, \tag{11}$$

where $\mathbf{J}(\cdot) : \mathfrak{R}^n \mapsto \mathfrak{R}^m$. The outcome is a set of solutions—called the *Pareto-set*—that describes the trade-off among these objectives, namely the minimization of each cost function. In this paper, the cost functions $J_1(\hat{\boldsymbol{\theta}}), \ldots, J_m(\hat{\boldsymbol{\theta}})$ take into account auxiliary information about the system.

Generally, there is no unique solution (model) that simultaneously minimizes all the different cost functions $J_j(\cdot)$. Rather, several solutions (models) are found with the property that the improvement of any objective necessarily implies

loss in some other objective. These are the *efficient solutions* or the *Pareto set solutions*. Any parameter vector which is an efficient solution will be referred to as a *Pareto model*. Thus, Pareto models are "the best" in the sense that there is no ordering among them, and that there is always some Pareto model that is better than any non-efficient solution, when compared in all optimization objectives. In the case of all functionals J_j being convex[1], the Pareto set can be found by defining (Chankong and Haimes 1983)

$$W = \{ \mathbf{w} \mid \mathbf{w} \in \mathfrak{R}^m, \ w_j \geq 0 \text{ and } \sum_{j=1}^m w_j = 1 \} \tag{12}$$

and solving the convex optimization problem

$$\hat{\boldsymbol{\theta}}^* = \arg \min_{\hat{\theta}} \langle \mathbf{w}, \mathbf{J}(\hat{\boldsymbol{\theta}}) \rangle. \tag{13}$$

For each vector $\mathbf{w}$, which defines a particular combination of weights to the various cost functions involved, a solution $\hat{\boldsymbol{\theta}}^*$ belonging to the Pareto set $\hat{\boldsymbol{\Theta}}^*$ is found. The entire Pareto set is associated to the set of all realizations of $\mathbf{w} \in W$.

Example 1 Following the Definition 1 (Nepomuceno et al. 2007), it is possible to express the following pair of affine information. (a) *Dynamic data (input/output)*: $[\mathbf{y}, \Psi]$, where $\mathbf{y}$ is the output data and Ψ is the regressor matrix. (b) *Fixed points*: $[\sigma, S]$, where σ is the normalized set of cluster coefficients and $S \in \mathfrak{R}^{\ell+1 \times n_\theta}$ is a constant matrix, that maps parameters to the cluster coefficients, that is $\hat{\boldsymbol{\sigma}} = S\hat{\boldsymbol{\theta}}$. (c) *Static function*: $[\bar{\mathbf{y}}, QR]$, where $\bar{\mathbf{y}}$ is the steady state of output, R is a constant matrix of ones and zeros that maps the parameter vector to the cluster coefficients, and $Q = \begin{bmatrix} \mathbf{q}_1 & \ldots & \mathbf{q}_{n_{\text{sf}}} \end{bmatrix}$, in which n_{sf} different steady-state points of input and output were considered. □

3 The Multiobjective Error Reduction Ratio (MERR)

Using the definition of affine information, (2) can be rewritten as

$$\mathbf{v_j} = G_j \hat{\boldsymbol{\theta}} + \boldsymbol{\xi}_j \tag{14}$$

where $[v_j, G_j]$ is the jth pair of affine information and Eq. (4) can be changed to

$$\langle \mathbf{v_j}, \mathbf{v_j} \rangle = \sum_{i=1}^{n_\theta} \hat{g}_i^2 \langle \mathbf{\Omega}_{j,i}, \mathbf{\Omega}_{j,i} \rangle + \langle \boldsymbol{\xi}_j, \boldsymbol{\xi}_j \rangle. \tag{15}$$

where $\mathbf{\Omega}_j$ is the jth orthogonal vector of the respective affine information and

$$\langle \boldsymbol{\xi}_j, \boldsymbol{\xi}_j \rangle = \langle \mathbf{v_j}, \mathbf{v_j} \rangle - \sum_{i=1}^{n_\theta} \hat{g}_i^2 \langle \mathbf{\Omega}_{j,i}, \mathbf{\Omega}_{j,i} \rangle. \tag{16}$$

---

[1] Notice that this is not true if any functional J_j is not convex.

Thus the ERR for ith regressor and jth affine information is expressed as

$$\text{ERR}_{i,j} = \frac{\hat{g}_i^2 \langle \boldsymbol{\Omega}_{j,i}, \boldsymbol{\Omega}_{j,i} \rangle}{\langle \mathbf{v}_j, \mathbf{v}_j \rangle}. \tag{17}$$

Then the m affine information pairs can be simultaneously taken into account to yield the multiobjective ERR, named as MERR for the ith regressor. Using the weighted sum of m affine information pairs yield:

$$\text{MERR}_i = \hat{g}_i^2 \sum_{j=1}^{m} w_j \frac{\langle \boldsymbol{\Omega}_{j,i}, \boldsymbol{\Omega}_{j,i} \rangle}{\langle \mathbf{v}_j, \mathbf{v}_j \rangle}. \tag{18}$$

Similarly, the parameters may be expressed as

$$\hat{g}_i = \sum_{j=1}^{m} w_j \frac{\langle \boldsymbol{\Omega}_{j,i}, \mathbf{v}_j \rangle}{\langle \boldsymbol{\Omega}_{j,i}, \boldsymbol{\Omega}_{j,i} \rangle}, \quad i = 1, \ldots, n_\theta. \tag{19}$$

Considering just the incorporation of dynamic data in the Eq. (18), MERR becomes exactly the traditional ERR. In this case, the problem would have only one weight ($w_1 = 1$), associated to the dynamic residue variance.

When there is not enough information embedded in dynamic data for obtaining a representative model, affine information may be incorporated in the structure detection procedure by means of the presented technique. In addition, these kinds of information can be available either by means of real data or theoretically. Information, such as static curve, dynamic behavior, fixed points, and static gain, can be inserted using MERR technique, once such type of information can be written using the prediction error formulation, in function of the model parameters.

The inclusion of each regressor is done sequentially by means of a forward algorithm. First of all, the regressor with maximum MERR composes the model with one regressor. Afterwards, MERR is calculated for the remaining non-included regressors, without the regressor previously selected. The regressor with maximum MERR is included in the model. This procedure is repeated until the desired model size, i.e., the desired number of regressors is obtained.

Example 2 Taking (18) for two affine information: dynamic data and fixed point $[\mathbf{v}_1, G_1] = [\mathbf{y}, \Psi]$ and $[\mathbf{v}_2, G_2] = [\sigma, S]$, it yields

$$\text{MERR}_i = \hat{g}_i^2 \sum_{j=1}^{2} w_j \frac{\langle \boldsymbol{\Omega}_{j,i}, \boldsymbol{\Omega}_{j,i} \rangle}{\langle \mathbf{v}_j, \mathbf{v}_j \rangle}$$
$$= \hat{g}_i^2 \left(w_1 \frac{\langle \boldsymbol{\Omega}_{1,i}, \boldsymbol{\Omega}_{1,i} \rangle}{\langle \mathbf{y}, \mathbf{y} \rangle} + w_2 \frac{\langle \boldsymbol{\Omega}_{2,i}, \boldsymbol{\Omega}_{2,i} \rangle}{\langle \sigma, \sigma \rangle} \right), \tag{20}$$

where $\boldsymbol{\Omega}_{1,i}$ and $\boldsymbol{\Omega}_{2,i}$ are the orthogonal regressors for Ψ and S, respectively. □

Example 3 Taking (18) for three affine information: dynamic data, static function, and fixed point $[\mathbf{v}_1, G_1] = [\mathbf{y}, \Psi]$, $[\mathbf{v}_2, G_2] = [\bar{\mathbf{y}}, QR]$ and $[\mathbf{v}_3, G_3] = [\sigma, S]$, it yields

$$\text{MERR}_i = \hat{g}_i^2 \sum_{j=1}^{3} w_j \frac{\langle \boldsymbol{\Omega}_{j,i}, \boldsymbol{\Omega}_{j,i} \rangle}{\langle \mathbf{v}_j, \mathbf{v}_j \rangle}$$
$$= \hat{g}_i^2 \left(w_1 \frac{\langle \boldsymbol{\Omega}_{1,i}, \boldsymbol{\Omega}_{1,i} \rangle}{\langle \mathbf{y}, \mathbf{y} \rangle} + w_2 \frac{\langle \boldsymbol{\Omega}_{2,i}, \boldsymbol{\Omega}_{2,i} \rangle}{\langle \bar{\mathbf{y}}, \bar{\mathbf{y}} \rangle} + w_3 \frac{\langle \boldsymbol{\Omega}_{3,i}, \boldsymbol{\Omega}_{3,i} \rangle}{\langle \sigma, \sigma \rangle} \right), \tag{21}$$

where $\boldsymbol{\Omega}_{1,i}$, $\boldsymbol{\Omega}_{2,i}$, and $\boldsymbol{\Omega}_{3,i}$ are the orthogonal regressors for Ψ, QR, and S, respectively. □

3.1 MERR Analysis

Since MERR is a multiobjective problem solved by the weighted sum approach, the ERR can be directly derived, using only dynamic data (in this case, $w_1 = 1$). On the other hand, varying the weights associated to each objective (affine residues variance of each information), models belonging to the Pareto set can be obtained. Clearly, each solution of the Pareto optimal curve may have a different structure. Non-optimal solutions may be generated and these solutions do not compose the Pareto optimal set, once they do not represent the system as well as other solutions.

The use of auxiliary information by means of the presented approach helps one find a model structure which is more compatible with represent specific types of system informations. In some cases, using only dynamic data is not enough to achieve a reliable model. Furthermore, some systems may have several limitations on dynamic data acquisition or even a low signal to noise ratio. These kinds of problems can be minimized using the MERR approach.

Since the Pareto optimal set was already obtained, the decision problem comes up. This is not a trivial problem and there are also some techniques available for that. Nepomuceno et al. (2007) present a decision technique based on the minimization of the norm of a vector composed by the normalized objectives, while the reference Barroso et al. (2007) uses minimal correlation criterion. In the present work, all the models obtained will be analyzed and validated either using the known structure (numerical example) or by means of the normalized root mean squared error (RMSE) index, defined as:

$$\text{RMSE} = \frac{\sqrt{\sum_{k=1}^{N} [\mathbf{v}(k) - G\hat{\boldsymbol{\theta}}(k)]^2}}{\sqrt{\sum_{k=1}^{N} [\mathbf{v}(k) - \bar{\mathbf{v}}]^2}}, \tag{22}$$

where $G\hat{\boldsymbol{\theta}}$ is the estimative of $\mathbf{v}(k)$, whose average is $\bar{\mathbf{v}}$. This index measures the error coherently with the data used to validate the model and should be calculated by the use of a specific validation dataset. Models are considered representative when this index is lesser than one, meaning that the error

is, on average, lesser than the error given by the mean of the time series. Finally, the multiobjective feature of the MERR has to be highlighted, since the structure detection techniques are mostly mono-objective. These techniques are not able to quantify the contribution of each regressor, explaining different system characteristics.

All simulations were performed on Matlab®, by means of an Asus® laptop, using 4 Gb of RAM, Intel® core $i3$ at 2.27 GHz, running Windows 7 ® Home Premium. 2 was used as the maximum order for input and output regressors, as well as the maximum nonlinearity degree of each model.

4 Application Examples

4.1 Numerical Example

Consider the following polynomial NARX system, presented originally by the reference Bonin et al. (2010):

$$y(k) = 0.5y(k-1) + 0.8u(k-2) + u(k-1)^2$$
$$- 0.05y(k-2)^2 + 0.5 + e(k), \quad (23)$$

The input u is given by a linear combination of an autoregressive regressor and a white Gaussian noise with mean zero and variance one, composing a low-pass filter:

$$u(k) = 0.5u(k-1) + \aleph(k), \quad (24)$$

where $\aleph(\cdot)$ is the white Gaussian noise.

This kind of excitation may confuse the identification algorithm in the estimation process, once the difference between regressors like $y(t)$ and $y(t-1)$ is considerably reduced Bonin et al. (2010). The noise added is also white and Gaussian with mean zero and variance 0.05.

The static data used for identification were generated by applying term clustering theory on the Eq. (23). The model is composed by constant, linear in u and y, and quadratic in u and y clusters.

MERR was applied to identify this system, considering $[\mathbf{y}, \Psi]$ and $[\bar{\mathbf{y}}, QR]$ as affine information pairs (dynamic and static data). Weights $w_1 = [0.1; 0.35; 0.7; 0.9; 1]$ and $w_2 = [0.9; 0.65; 0.3; 0.1; 0]$ were used for dynamic and static information pairs, respectively. These weights were chosen aiming a good diversity of the Pareto set. Observe that, for $w_1 = 1$, MERR is equivalent to ERR. From now, the models will be indexed as MERR^z, where z is the zth column of the weights vectors.

Besides, according to the AIC, 7 regressors would have to be inserted in the model to identify this system. Although this is not true (the model structure is already known and it has five regressors), MERR will be calculated for a model with seven regressors.

In order to avoid bias in the parameter estimation procedure, parameters were estimated through the use of extended least squares, composing a NARMAX structure. However, the presented technique can be applied either for NARX or NARMAX structures or even to other linear in parameters representation.

The Table 1 shows the results obtained when MERR was applied to the numerical system proposed, and the parameters associated to each regressor. Considering five regressors obtained in the first five iterations ($i = 1, \cdots, 5$ on the Table 1), MERR was able to reconstruct the original system using either dynamic or dynamic and static data (MERR^3, MERR^4 and MERR^5).

To check out the proposed technique in different noise realizations, the model MERR^3 was simulated, considering 100 noise realizations. Table 2 presents the results obtained. In all simulations 7 was considered the model length and the mean and standard deviation of MERR obtained for each noise realization were calculated. Further, in all simulation (as can be noticed by the third column) the genuine regressors were chosen by the MERR. This means that, for this case, MERR is useful not just to a specific noise realization.

The use of a low-pass filter as an input hindered the identification process. The AIC has suggested 7 regressors as an estimative of the model's length. However, the MERR is robust enough to quantify genuine regressors with a higher ratio. Furthermore, the presented approach quantifies regressors like $u(k-1)$ and $y(k-2)$ as good candidates. Considering dynamic data, these regressors are not important. However, their clusters (clusters linear in u and y) are important to represent the static behavior and, once the MERR considers static information, these regressors could be a genuine candidate. It is important to notice that as weight of static information increases, the percentage of MERR for $y(k-1)$ increases. This effect is due to the fact that static information is less susceptible to noise.

4.2 The Pilot DC–DC Buck Converter

In order to compare the proposed technique MERR with the classic ERR, two studies were undertaken. First, MERR is compared with classic ERR, both with parameters estimated via mono-objective extended least squares. After that, the MERR is compared with ERR, when the parameters were estimated using auxiliary information for both of them. This second comparison aims at verifying if there is a gain of quality only from the structure selection.

4.2.1 MERR Versus Classic ERR with Mono-objective Parameter Estimative (ELS)

A pilot DC–DC buck converter (Aguirre et al. 2000) was modeled in which its real dynamic data and theoretical static

Table 1 MERR—numerical example

System	MERR¹ (p) [θ̂]	MERR² (p) [θ̂]	MERR³ (p) [θ̂]	MERR⁴ (p) [θ̂]	MERR⁵ (p) [θ̂]
$0.5y(k-1)$	97.98 (1)	80.30 (1)	55.22 (1)	46.78 (1)	25.72 (3)
	[0.4790]	[0.4962]	[0.4932]	[0.4932]	[0.5098]
$0.8u(k-2)$	0.265 (2)	0.346 (7)	3.053 (4)	4.510 (4)	5.234 (4)
	[0.8936]	[0.7904]	[0.8007]	[0.8007]	[0.8436]
$u(k-1)^2$	0.085 (7)	4.749 (2)	17.63 (2)	20.76 (2)	44.85 (1)
	[0.2121]	[0.9962]	[0.9960]	[0.9960]	[1.0185]
$-0.05y(k-2)^2$	–	–	18.91 (3)	24.22 (3)	20.65 (2)
	0	0	[−0.0491]	[−0.0491]	[−0.0490]
0.5	0.053 (6)	2.918 (4)	2.050 (5)	1.386 (5)	1.128 (5)
	[1.3504]	[0.5446]	[0.5355]	[0.5355]	[0.5199]
$\underline{y(k-2)}$	0.124 (4)	–	0.155 (7)	0.028 (7)	–
	[0.0427]	0	[0.0089]	[0.0089]	0
$\underline{u(k-1)}$	–	1.591 (6)	0.812 (6)	0.066 (6)	–
	0	[0.0099]	[0.0118]	[0.0118]	0
$\underline{u(k-2)u(k-1)}$	–	–	–	–	0.022 (6)
	0	0	0	0	[−0.0573]
$\underline{u(k-2)y(k-1)}$	–	–	–	–	0.017 (7)
	0	0	0	0	[−0.0084]
$\underline{u(k-2)^2}$	0.300 (3)	4.945 (3)	–	–	–
	[−0.0444]	[−0.0489]	0	0	0
$\underline{u(k-2)y(k-2)}$	0.165 (5)	0.879 (5)	–	–	–
	[−0.0395]	[0.0037]	0	0	0

The letter p stands for the iteration in which the regressor was selected by the technique. The variable $\hat{\theta}$ is the estimated parameter associated to each regressor. Underlined regressors are not genuine

Table 2 Model MERR³ obtained for 100 noise realizations

	MERR (mean and standard deviation)	Times included as a genuine regressor by the MERR³
$y(k-1)$	0.300 ± 0.200	100
$u(k-2)$	0.028 ± 0.005	100
$u(k-1)^2$	0.552 ± 0.200	100
$y(k-2)^2$	0.058 ± 0.020	100
C	0.042 ± 0.100	100

The third column means that the genuine regressors were chosen for all simulations

data ($[\mathbf{y}, \Psi]$ and $[\bar{\mathbf{y}}, QR]$ as affine information pairs) were used.

In order to acquire dynamic data, the system was excited by a pseudo-random binary signal. This input excites the local non-linearities of the system, which is expected to be in the data. The static information pair was obtained theoretically and the input/output relation can be written as:

$$\bar{y} = \frac{4}{3}v_d - \frac{v_d}{3}\bar{u} \qquad (25)$$

where $v_d = 24$ V, $\bar{u}$ and $\bar{y}$ are the input and output operating in steady state. 84 static and dynamic points were used for identification.

The models were obtained using $w_1 = [0.1; 0.3; 0.7; 0.9; 1]$ and $w_2 = [0.9; 0.7; 0.3; 0.1; 0]$ for dynamic and static data, respectively. In order to simplify, the models will be called MERR^z, where z is the zth column of the vector w_r, $r = 1, 2$. The structures obtained for $z = 1, 2$ were the same. However, it should be emphasized that the regressors for these models do not necessarily have the same MERR, since the technique uses different weights on each objective. Thereby, a set of four (4) different models was obtained. The AIC was used to establish how many regressors should be included in the model, resulting in 9 regressors. Table 3 exhibits the regressors obtained for the models and its respective parameters and the value of the MERR, considering parameters estimated via ELS. The (p) index is the iteration in which the regressor was chosen by the MERR. Figs. 1 and 2 show the system and models static and dynamic behavior, respectively. Clearly, the objectives considered are conflicting. However, models with static improvement could be obtained, with some or even without any penalty considering its dynamic (MERR⁴ model). There are, therefore, important information in static data, which are relevant for obtaining global models, considering different characteristics of system, such as its static and dynamic. This ratifies the requirement of including more than one affine information pairs in the structure detection procedure. To conclude, it has to be

Table 3 MERR—DC–DC Buck converter

Models' regressors	MERR[1] (p) [$\hat{\theta}$]	MERR[2] (p) [$\hat{\theta}$]	MERR[3] (p) [$\hat{\theta}$]	MERR[4] (p) [$\hat{\theta}$]	MERR[5] (p) [$\hat{\theta}$]
$u(k-1)$	0.000036 (3) [−2.1340]	0.000856 (3) [−2.1340]	0.009267 (7) [−0.7677]	0.017904 (4) [−5.6780]	0.000596 (9) [3.2458]
$u(k-2)$	0.000001 (5) [−2.8396]	0.000010 (5) [−2.8396]	– [0]	0.000103 (5) [9.3834]	– [0]
$y(k-1)$	99.85087 (1) [1.6168]	99.82074 (1) [1.6168]	99.89977 (1) [−0.4849]	99.91307 (1) [−0.0396]	99.91860 (1) [−0.6867]
$y(k-2)$	0.148949 (2) [−1.1554]	0.174741 (2) [−1.1554]	0.064284 (2) [0.1546]	0.042932 (2) [−1.0472]	0.040465 (2) [−0.3253]
$u(k-1)^2$	– [0]	– [0]	0.001337 (3) [−3.0944]	0.000291 (9) [−0.8788]	0.014101 (4) [−2.8685]
$u(k-2)^2$	– [0]	– [0]	0.005358 (4) [0.3150]	0.000057 (6) [−2.0131]	– [0]
$u(k-2)u(k-1)$	– [0]	– [0]	0.001061 (5) [1.2441]	– [0]	– [0]
$y(k-1)^2$	0.000003 (8) [−0.0863]	0.000076 (8) [−0.0863]	– [0]	– [0]	0.005679 (3) [−0.0081]
$y(k-2)^2$	0.000001 (6) [−0.0744]	0.000008 (6) [−0.0744]	– [0]	0.000315 (8) [0.0152]	0.000065 (5) [−0.0404]
$y(k-2)y(k-1)$	0.000001 (7) [0.1519]	0.000037 (7) [0.1519]	– [0]	– [0]	0.001073 (6) [0.0593]
$u(k-1)y(k-1)$	– [0]	– [0]	0.000115 (8) [0.7952]	0.001063 (7) [0.5945]	0.000605 (8) [0.6153]
$u(k-2)y(k-2)$	0.000002 (9) [0.2078]	0.000045 (9) [0.2078]	0.000323 (9) [−0.3438]	– [0]	– [0]
C	0.000063 (4) [13.9972]	0.001604 (4) [13.9972]	0.000807 (6) [14.0006]	0.002313 (3) [14.0005]	0.000218 (7) [14.0007]

The letter p stands for the iteration in which the regressor was selected by the technique. The variable $\hat{\theta}$ is the estimated parameter associated to each regressor
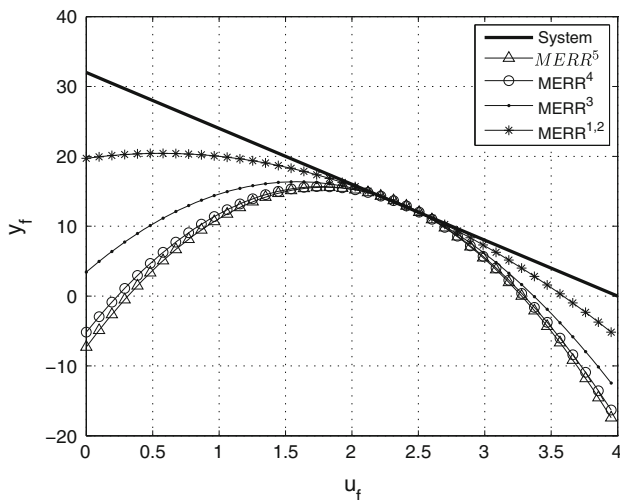

Fig. 1 Static curve. The *continuous line* represents the system, *triangle* the model obtained $w_1 = 1$, *open circle* is the static curve of the model obtained when $w_1 = 0.9$, *dot* shows the static curve of the model obtained by $w1 = 0.7$ and *asterisks* is the model chosen when $w1 = 0.35$. Parameter estimated using ELS
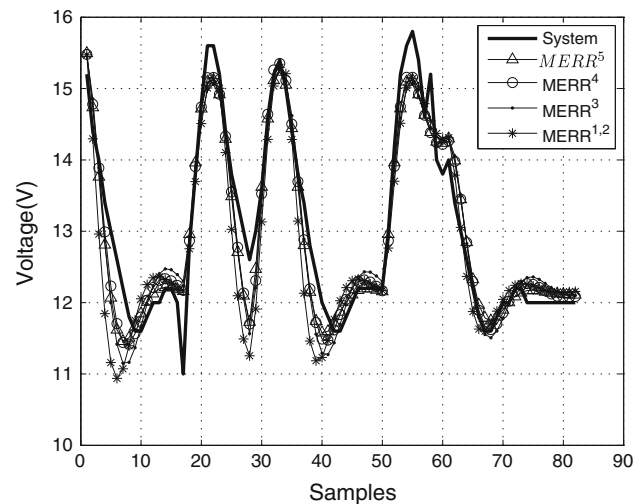

Fig. 2 Dynamic time series. The *continuous line* represents the system, *triangle* is the free prediction of the model obtained $w_1 = 1$, *open circle* exhibits the model obtained when $w_1 = 0.9$, *dot* shows the behavior of the model obtained by $w1 = 0.7$ and *asterisks* shows the model chosen when $w_1 = 0.35$. Parameter estimated using ELS

Table 4 Models validation

	MERR[1,2]	MERR[3]	MERR[4]	MERR[5]
RMSE$_1$	0.4802 (0.5105)	0.3545 (0.4849)	0.2935 (0.4853)	0.3004 (0.4853)
RMSE$_2$	0.4930 (0.0053)	1.1493 (0.0040)	1.4959 (0.0045)	1.5824 (0.0050)

The sub-index 1 stands for dynamic whereas the sub-index 2 stands for static

highlighted that MERR presents an efficient solutions curve, which allows an amount of *optimal* structures to represent the system, considering different operation points. This is considerably important when using the model in a specific identification purpose and also in other applications, such as in model-based control.

4.2.2 MERR Versus ERR with Auxiliary Information in the Parameter Estimation

Once the structure was obtained, the parameters are now estimated using both static and dynamic data, comparing the models performance. The decision-maker used for parameter estimation was the minimal correlation criterion (Barroso et al. 2007) and the results are presented in Table 4, in parenthesis. Outside of the parenthesis, the RMSE static and dynamic obtained when the parameters was estimated via ELS are presented. Although the static behavior is improved when parameters were estimated through the use of a multiobjective approach, the dynamic error also increases. This means that, in an optimization point of view, these techniques obtain models that map different regions of the Pareto set, ratifying the performance of the technique proposed in this paper. In both comparisons, the MERR presented nondominated solutions of Pareto set. The model MERR[5] is the model obtained via ERR, when $w_1 = 1$ and $w_2 = 0$.

5 Conclusions

This paper presented and developed the MERR, a technique used for structure detection in polynomial NARX models. In general, techniques for structure detection are mono-objective and they just take into account dynamic data. In fact, further information are important to be considered in a structure detection procedure.

A set of stable efficient models was obtained through the use of MERR, with a relevant representativity of the system. A substantial improvement on the system's static curve was also observed, given by the obtained models. Furthermore, such models showed up to be global and representative, once other information about the system can be considered for structure detection. It is worth to state that the use

of affine information only on parameter estimation increases substantially the quality of models, as one can see in Table 4. Although, solutions produced by MERR are nondominated in this case, which clarified the limit of using affine information only in parameter estimation. Structure selection with affine information has been shown to be an efficient tool to find out the best structure as shown in Sect. 4.1, even in the presence of noisy. Moreover, although the present methodology has been described only for Polynomial NARX, MERR may be extended for other representations such as Volterra series and forward neural networks (Aguirre et a. 2001).

As a future research topic, one expects to make an adaptation on the AIC index, aiming at incorporating other information about the system, besides dynamical data. Moreover, the MERR will be tested using a multiobjective problem, where more than two kinds of information about the system will be incorporated in the model. Finally, the decision problem will be approached in order to provide, following a suitable criterion, one or more models to the user.

References

Aguirre, L. A., & Billings, S. A. (1994). Model reference control of regular and chaotic dynamics in the Duffing-Ueda oscillator. *IEEE Transactions on Circuits and Systems, Part I, 41*(7), 477–480.

Aguirre, L. A., & Billings, S. A. (1995). Improved structure selection for nonlinear models based on term clustering. *International Journal of Control, 62*(3), 569–587.

Aguirre, L. A., Donoso-Garcia, P. F., & Santos, R. (2000). Use of a priori information in the identification of global nonlinear models—A case study using a buck converter. *IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications, 47*(7), 1081–1085.

Aguirre, L. A., Freitas, U. S., Letellier, C. S., & Maquet, J. (2001). Structure-selection techniques applied to continuous-time nonlinear models. *Physica D 158*(1–4), 1–18. doi:10.1016/S0167-2789(01)00313-X.

Akaike, H.: New look at statistical-model identification. *IEEE Transactions on Automatic Control 19*(6), 716–723 (1974). doi:10.1109/TAC.1974.1100705.

Baldacchino, T., Anderson, S. R., & Kadirkamanathan, V. (2012). Structure detection and parameter estimation for NARX models in a unified EM framework. *Automatica, 48*(5), 857–865.

Barroso, M. F. S., Takahashi, R., & Aguirre, L. (2007). Multi-objective parameter estimation via minimal correlation criterion. *Journal of Process Control, 17*(4), 321–332.

Bonin, M., Seghezza, V., & Piroddi, L. (2010). NARX model selection based on simulation error minimisation and LASSO. *IET Control Theory and Applications 4*(7), 1157–1168. doi:10.1049/iet-cta.2009.0217.

Cantelmo, C., & Piroddi, L. (2010). Adaptive model selection for polynomial NARX models. *IET Control Theory and Applications*, *4*(12), 2693–2706.

Chankong, V., & Haimes, Y. Y. (1983). *Multiobjective decision making: Theory and methodology*. New York: North-Holland (Elsevier).

Chen, S., & Billings, S. A. (1989). Representations of non-linear systems: The NARMAX model. *International Journal of Control*, *49*(3), 1013–1032.

Johansen, T. A. (1996). Identification of non-linear systems using empirical data and prior knowledge—An optimization approach. *Automatica*, *32*(3), 337–356.

Korenberg, M., Billings, S., Liu, Y., & Mcilroy, P. (1988). Orthogonal parameter estimation algorithm for non-linear stochastic systems. *International Journal of Control*, *48*(1), 193–210.

Leontaritis, I. J., & Billings, S. A. (1985). Input–output parametric models for non-linear systems—Part I: Deterministic non-linear systems. *International Journal of Control*, *41*(2), 303–328.

Ljung, L. (1987). *System identification: Theory for the user*. London: Prentice-Hall.

Mendes, E. M. A. M., & Billings, S. A. (2001). An alternative solution to the model structure selection problem. *IEEE Transactions on Systems Man and Cybernetics: Part A—Systems and Humans, 31*(6), 597–608. doi:10.1109/3468.983416.

Nepomuceno, E. G., Takahashi, R. H. C., & Aguirre, L. A. (2007). Multiobjective parameter estimation for non-linear systems: Affine information and least-squares formulation. *International Journal of Control*, *80*(6), 863–871.

Piroddi, L., & Spinelli, W. (2003). An identification algorithm for polynomial NARX models based on simulation error minimization. *International Journal of Control*, *76*(17), 1767–1781.

Previdi, F., & Lovera, M. (2004). Identification of non-linear parametrically varying models using separable least squares. *International Journal of Control*, *77*(16), 1382–1392.

Wei, H. L., & Billings, S. A. (2008). Model structure selection using an integrated forward orthogonal search algorithm assisted by squared correlation and mutual information. *International Journal of Modelling Identification and Control*, *3*(4), 341–356.

Wei, H. L., & Billings, S. A. (2009). Improved parameter estimates for non-linear dynamical models using a bootstrap method. *International Journal of Control*, *82*(11), 1991–2001.

Zhao, W. X., Chen, H. F., & Zhou, T. (2011). New results on recursive identification of NARX systems. *International Journal of Adaptive Control and Signal Processing*, *25*(10), 855–875.