Automatica 48 (2012) 857–865

Contents lists available at SciVerse ScienceDirect

Automatica

journal homepage: www.elsevier.com/locate/automatica

Brief paper

Structure detection and parameter estimation for NARX models in a unified EM framework☆

Tara Baldacchino [1], Sean R. Anderson, Visakan Kadirkamanathan

Department of Automatic Control and Systems Engineering, University of Sheffield, Sheffield S1 3JD, UK

ARTICLE INFO

Article history:
Received 16 June 2010
Received in revised form
4 March 2011
Accepted 29 September 2011
Available online 23 March 2012

Keywords:
Nonlinear system identification
NARX
Expectation–maximisation
Structure detection
Particle filter

ABSTRACT

In this paper, we consider structure detection and parameter estimation of the nonlinear auto-regressive with exogenous inputs (NARX) model, using the EM (expectation–maximisation) algorithm. The parameter estimation step uses particle smoothing to obtain the necessary expectations in the E-step and the parameters are then estimated in closed form in the M-step. The model structure detection is performed using an F-test, which makes use of the parameter information matrix (inverse of the covariance matrix), obtained from an augmentation of the EM algorithm. The steps for obtaining the information matrix are robust, guaranteeing a positive semi-definite information matrix to use in the structure detection step. For the case of unknown model orders, a method is proposed using the stochastic complexity (SC) information criterion for selecting between candidate models. The SC is composed of the information matrix (representing model complexity) and a likelihood estimate (representing model accuracy), which are both generated as byproducts of the augmented EM algorithm. Numerical results demonstrate that the EM approach performs well in comparison to a standard alternative based on orthogonal least squares, and also avoids the need to estimate a noise model for the case of measurement noise corrupted output signals.

1. Introduction

An important and challenging problem in nonlinear system identification is structure detection (Sjoberg et al., 1995). A typical approach to identifying a nonlinear black-box model, such as the nonlinear auto-regressive with exogenous inputs (NARX) model (Leontaritis & Billings, 1985), is to search amongst a superset of possible model terms for a parsimonious description (Haber & Unbehauen, 1990). Parameter estimation is inherently linked with this search because a model term typically requires an associated parameter estimate for testing the term's significance. Here, we present an algorithm for identification of the NARX model, where both the parameters and the structure are obtained from the expectation–maximisation (EM) algorithm (Dempster, Laird, & Rubin, 1977; Wu, 1983).

In our approach, we represent the NARX model in a state-space form, which allows for a clearer distinction of the separate process and measurement noises entering the system description (Ljung, 1999). Such a model representation naturally leads to a choice of parameter estimation framework appropriate for state-space systems, which in the linear case might include, for example, subspace methods (Van Overschee & De Moor, 1996) or EM, which also has the advantage of handling missing data (Gibson & Ninness, 2005; Isaksson, 1993; Kadirkamanathan & Anderson, 2008; Shumway & Stoffer, 1982, 2000).

The EM algorithm is a natural candidate for NARX identification, having been extended recently to handle nonlinear state and parameter estimation problems, making use of the extended and unscented Kalman filter (Roweis & Ghahramani, 2001) and the particle filter (Gopaluni, 2008; Schön, Wills, & Ninness, 2006, 2011; Zia et al., 2008). Of particular relevance here is the work by Schön et al. (2006), who used EM to estimate the parameters of a nonlinear state-space system with additive Gaussian state noise. This approach was extended to a more general state noise description by Wills, Schön, and Ninness (2008) and Schön et al. (2011). In the NARX model class considered here, the state noise is additive and Gaussian. Hence, we take a similar approach to parameter estimation to that proposed by Schön et al. (2006), with the novel distinctions that (i) we consider parameter estimation for the state-space NARX model, requiring a model-specific derivation of the EM algorithm and (ii) we consider system identification in the wider context of structure detection along with parameter estimation.

☆ The material in this paper was not presented at any conference. This paper was recommended for publication in revised form by Associate Editor Brett Ninness under the direction of Editor Torsten Söderström.

E-mail addresses: t.baldacchino@sheffield.ac.uk (T. Baldacchino), s.anderson@sheffield.ac.uk (S.R. Anderson), visakan@sheffield.ac.uk (V. Kadirkamanathan).

[1] Tel.: +44 (0)114 222 5134; fax: +44 (0)114 222 5683.

0005-1098/$ – see front matter © 2012 Elsevier Ltd. All rights reserved.
doi:10.1016/j.automatica.2012.02.021

To perform the NARX model structure detection we derive a novel term selection algorithm that is based on the F-test, which was proposed by Billings and Voon (1986) in the context of maximum likelihood NARX model identification. This type of approach is distinct from least-squares-based orthogonal search algorithms, where an error reduction ratio drives selection, either obtained from the one-step-ahead prediction (Anderson & Kadirkamanathan, 2007; Korenberg, Billings, Liu, & McIlroy, 1988) or simulated prediction (Piroddi & Spinelli, 2003). The F-test method is more closely related to methods that use estimates of parameter statistics to selectively include or exclude terms (Kukreja, Galiana, & Kearney, 2004).

The framework we propose here is based on two stages: estimation of an initial superset of model terms $\mathcal{M}_0$, followed by pruning using the F-test to a final parsimonious set $\mathcal{M}_F$. Two-stage algorithms are often used in nonlinear system identification where an initial model provides a platform for compacting the model description (Li, Peng, & Bai, 2006; Mao & Billings, 1997). The key novel feature in our approach is that the parameter information matrix (the inverse of the covariance matrix, which is used in the F-test) is obtained directly from an augmentation of the EM algorithm. Critical to this augmentation is the use of a robust method, highlighted by Duan and Fulop (2011), which guarantees a positive semi-definite information matrix, unlike other available methods used in EM (Louis, 1982; McLachlan & Krishnan, 2008). In addition, we use an information criterion, the stochastic complexity (SC) (Rissanen, 1989), to select between models of different maximum dynamic order, where the SC is also obtained from byproducts of the EM algorithm. Thus, we keep parameter estimation and structure detection within a unified EM framework. In summary, our approach exploits recent advances in using the EM algorithm for nonlinear state and parameter estimation, where we have made novel extensions for structure detection, a problem that has only received limited attention in the context of EM (Gopaluni, 2010).

The paper is structured as follows. In Section 2, we present the state-space NARX model. In Section 3 we derive the parameter estimation algorithm for the NARX state-space model based on EM. In Section 4 we propose the structure detection algorithm, using the F-value metric. We present a numerical example and benchmark comparison in Section 5 to demonstrate the method. The paper is summarised in Section 6.

2. Nonlinear system representation

A single-input single-output (SISO) NARX model (Leontaritis & Billings, 1985), with true output $z_t \in \mathbb{R}$ and measurement noise corrupted output $y_t \in \mathbb{R}$, can be represented as

$$z_{t+1} = f(\boldsymbol{x}_t, \boldsymbol{u}_t) + w_t \tag{1}$$

$$y_t = z_t + v_t, \tag{2}$$

where the state vector is composed of the true system outputs $\boldsymbol{x}_t = \left[z_t, \ldots, z_{t-n_z+1}\right]^\top$, $\boldsymbol{u}_t = \left[u_t, \ldots, u_{t-n_u+1}\right]^\top$, $f(\cdot)$ is some nonlinear function, and $u_t \in \mathbb{R}$ is the known system input. The model orders n_z and n_u are the output and input order dynamics, respectively. Process noise and measurement noise are assumed to be independent, with zero mean, white and Gaussian distributed, and are respectively represented as $w_t \sim \mathcal{N}(0, \sigma^2)$, $v_t \sim \mathcal{N}(0, \lambda^2)$.

We represent the dynamics of the NARX model in (1) as the state equation,

$$\boldsymbol{x}_{t+1} = A\boldsymbol{x}_t + Bf(\boldsymbol{x}_t, \boldsymbol{u}_t) + Kw_t \tag{3}$$

where

$$A = \begin{bmatrix} 0 & 1 & \cdots & 0 \\ \vdots & & & \vdots \\ 0 & 0 & \cdots & 1 \\ 0 & 0 & \cdots & 0 \end{bmatrix}, \quad B = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix}, \quad K = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix}.$$

The measurement equation, which describes the mapping from the unobserved states to the observed output signal, is

$$y_t = C\boldsymbol{x}_t + v_t \tag{4}$$

where $C = [0, \ldots, 0, 1]$.

The nonlinear function $f(\boldsymbol{x}_t, \boldsymbol{u}_t)$ can be decomposed and represented by a linear sum of basis functions $\phi_i(\boldsymbol{x}_t, \boldsymbol{u}_t)$, which can be of varying form such as polynomial, wavelet, or radial:

$$f(\boldsymbol{x}_t, \boldsymbol{u}_t) = \sum_{i=1}^{m} \phi_i(\boldsymbol{x}_t, \boldsymbol{u}_t)\theta_i, \tag{5}$$

where m is the number of terms in the model and θ_i is the parameter that weights basis function ϕ_i. Given this representation of the NARX model, we can define the system identification task as to (i) obtain a parsimonious set of basis functions to describe the function $f(\boldsymbol{x}_t, \boldsymbol{u}_t)$ and (ii) estimate the parameters according to some criterion. Here, we use the EM algorithm to obtain maximum likelihood estimates of the model parameters and develop a novel scheme for the detection of the model structure within the EM framework.

3. Parameter estimation

The EM algorithm is an iterative process that is used to obtain a maximum likelihood (ML) estimate, via the log-likelihood function, for a parameter set $\boldsymbol{\theta} = (\theta_1 \cdots \theta_M)^\top$ when the data set is incomplete (Dempster et al., 1977). Given a set of observations $\mathcal{Y} = \{y_1 \cdots y_N\}$ and hidden data $\mathcal{X} = \{\boldsymbol{x}_1 \cdots \boldsymbol{x}_N\}$, where N is the length of the data, the observed data log-likelihood can be expressed as

$$L(\boldsymbol{\theta}) = \ln p(\mathcal{Y}|\boldsymbol{\theta}) = \ln \int_{\mathcal{X}} p(\mathcal{X}, \mathcal{Y}|\boldsymbol{\theta}) d\mathcal{X}, \tag{6}$$

where $p(\mathcal{X}, \mathcal{Y}|\boldsymbol{\theta})$ is the complete data likelihood. After some (well-known) equation manipulation, see for instance Gibson and Ninness (2005), we obtain the expression for the so-called Q-function,

$$\begin{aligned} Q(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_k) &= \int_{\mathcal{X}} p(\mathcal{X}|\mathcal{Y}, \boldsymbol{\theta}_k) \ln p(\mathcal{X}, \mathcal{Y}|\boldsymbol{\theta}) d\mathcal{X} \\ &= \mathbb{E}[\ln p(\mathcal{X}, \mathcal{Y}|\boldsymbol{\theta})|\mathcal{Y}, \hat{\boldsymbol{\theta}}_k], \end{aligned} \tag{7}$$

where $\mathbb{E}[\cdot]$ is the expectation with respect to $p(\mathcal{X}|\mathcal{Y}, \boldsymbol{\theta}_k)$ at the kth iteration of the EM algorithm. The EM algorithm alternates between finding $Q(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_k)$ in the E-step, and then maximising it in the M-step with respect to the parameters $\boldsymbol{\theta}$

$$\begin{aligned} &\textit{E-step}: Q(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_k) = \mathbb{E}[\ln p(\mathcal{X}, \mathcal{Y}|\boldsymbol{\theta})|\mathcal{Y}, \hat{\boldsymbol{\theta}}_k] \\ &\textit{M-step}: \hat{\boldsymbol{\theta}}_{k+1} = \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_k). \end{aligned} \tag{8}$$

The EM process is repeated until convergence is reached. Convergence properties are discussed in Dempster et al. (1977) and Wu (1983).

3.1. The M-step

In the M-step we are interested in maximising the function $Q(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_k)$ defined in (7) with respect to the parameter vector $\boldsymbol{\theta}$. We express the complete-data log-likelihood as the joint distribution of observed data $\mathcal{Y}$ and hidden data $\mathcal{X}$ as

$$\ln p(\mathcal{X}, \mathcal{Y}|\boldsymbol{\theta}) = \sum_{t=1}^{N} \ln p(y_t|\boldsymbol{x}_t) + \sum_{t=0}^{N-1} \ln p(\boldsymbol{x}_{t+1}|\boldsymbol{x}_t, \boldsymbol{\theta}), \tag{9}$$

where

$$p(y_t|\boldsymbol{x}_t) \sim \mathcal{N}(C\boldsymbol{x}_t, \lambda^2) \tag{10}$$

and $p(\boldsymbol{x}_{t+1}|\boldsymbol{x}_t, \boldsymbol{\theta})$ can be conveniently represented as

$$p(\boldsymbol{x}_{t+1}|\boldsymbol{x}_t, \boldsymbol{\theta}) = p(z_{t+1}|\boldsymbol{x}_t, \boldsymbol{\theta})p(\bar{\boldsymbol{x}}_{t+1}|\boldsymbol{x}_t), \tag{11}$$

where $\boldsymbol{x}_{t+1} = [\bar{\boldsymbol{x}}_{t+1} z_{t+1}]^\top$. From (3) and (4),

$$\begin{aligned} p(z_{t+1}|\boldsymbol{x}_t, \boldsymbol{\theta}) &\sim \mathcal{N}(\boldsymbol{\phi}_t\boldsymbol{\theta}, \sigma^2) \\ p(\bar{\boldsymbol{x}}_{t+1}|\boldsymbol{x}_t) &= \delta([I0]\boldsymbol{x}_t), \end{aligned} \tag{12}$$

where $\delta(\cdot)$ is the Dirac delta function and I is the identity matrix having dimensions $(n_z - 1) \times (n_z - 1)$ and $\boldsymbol{\phi}_t = [\phi_{t,1} \cdots \phi_{t,M}]$. Substituting (11) into (9) gives

$$\begin{aligned} \ln p(\mathcal{X}, \mathcal{Y}|\boldsymbol{\theta}) = &\sum_{t=1}^{N} \ln p(y_t|\boldsymbol{x}_t) + \sum_{t=0}^{N-1} \ln p(\bar{\boldsymbol{x}}_{t+1}|\boldsymbol{x}_t) \\ &+ \sum_{t=0}^{N-1} \ln p(z_{t+1}|\boldsymbol{x}_t, \boldsymbol{\theta}). \end{aligned} \tag{13}$$

Substitution of (13) into (7) leads to the following three expressions (corresponding to the three parts of (13)):

$$\begin{aligned} J_1 &= \sum_{t=1}^{N} \int \ln p(y_t|\boldsymbol{x}_t) p(\boldsymbol{x}_t|\mathcal{Y}) d\boldsymbol{x}_t \\ J_2 &= \sum_{t=0}^{N-1} \int \ln p(\bar{\boldsymbol{x}}_{t+1}|\boldsymbol{x}_t) p(\boldsymbol{x}_t|\mathcal{Y}) d\boldsymbol{x}_t \\ J_3 &= \sum_{t=0}^{N-1} \int\int \ln p(z_{t+1}|\boldsymbol{x}_t, \boldsymbol{\theta}) p(z_{t+1}, z_t|\mathcal{Y}) dz_t dz_{t+1}, \end{aligned} \tag{14}$$

where $Q(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_k) = J_1 + J_2 + J_3$.

Now, since we are interested in maximising the Q-function with respect to $\boldsymbol{\theta}$, the distributions $p(y_t|\boldsymbol{x}_t)$ and $p(\bar{\boldsymbol{x}}_{t+1}|\boldsymbol{x}_t)$ do not contribute to the maximisation step since they are conditionally independent of $\boldsymbol{\theta}$; hence

$$\begin{aligned} \frac{\partial Q(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_k)}{\partial \boldsymbol{\theta}} &= \frac{\partial J_3}{\partial \boldsymbol{\theta}} \\ &= \mathbb{E}\left[\sum_{t=0}^{N-1} \frac{\partial \ln p(z_{t+1}|\boldsymbol{x}_t, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right], \end{aligned} \tag{15}$$

where $\mathbb{E}[\cdot]$ denotes the expectation over the joint distribution $p(z_{t+1}, z_t|\mathcal{Y}, \boldsymbol{\theta}_k)$.

To obtain an update for $\boldsymbol{\theta}_k$, we first note that (15) is a quadratic function in $\boldsymbol{\theta}$; hence, setting (15) to zero to find the maximum leads directly to a closed-form expression for the parameter update, so at the kth iteration of the EM algorithm

$$\hat{\boldsymbol{\theta}}_{k+1} = \left(\mathbb{E}\left[\sum_{t=0}^{N-1} \boldsymbol{\phi}_t^\top \boldsymbol{\phi}_t\right]\right)^{-1} \mathbb{E}\left[\sum_{t=0}^{N-1} \boldsymbol{\phi}_t^\top z_{t+1}\right], \tag{16}$$

where details of how to calculate the expectations are given in the following section.

3.2. The E-step

In the E-step, particle filtering followed by particle smoothing is performed in order to represent the required posterior pdf $p(\boldsymbol{x}_t|\mathcal{Y}, \boldsymbol{\theta})$. In this paper, only the main equations for particle filtering and smoothing are provided. Interested readers can refer to Andrieu, Doucet, Singh, and Tadic (2004) and Arulampalam, Maskell, Gordon, and Clapp (2002) for more information on particle filters, and (Doucet, Godsill, & Andrieu, 2000) for smoothing, and for a tutorial on both filtering and smoothing see Doucet and Johansen (2008).

Filtering is performed by forming a set of N_s random samples of $\boldsymbol{x}_t$ with associated weights, represented as $\left\{\boldsymbol{x}_t^{(i)}, \boldsymbol{w}_t^{(i)}\right\}_{i=1:N_s}$. $\boldsymbol{x}_t^{(i)}$ are the particles of the states, at the kth iteration of the EM algorithm, that are used to compose the desired distribution, and $\boldsymbol{w}_t^{(i)}$ are the respective filtering weights. The filter used here is a sequential importance resampling (SIR) filter, where the predicted states are calculated using (ignoring the subscript k)

$$\boldsymbol{x}_{t+1|t}^{(i)} \sim p\left(\boldsymbol{x}_{t+1}|\boldsymbol{x}_t^{(i)}\right) \tag{17}$$

and the corresponding weights are given by

$$\boldsymbol{w}_t^{(i)} = p\left(y_t|\boldsymbol{x}_{t|t-1}^{(i)}, \boldsymbol{\theta}\right), \tag{18}$$

which then need to be normalised. Systematic resampling is independently applied N_s times from the distribution $\left\{\boldsymbol{x}_{t|t-1}^{(i)}\right\}_{i=1:N_s}$, at every time step, and the resulting state particles are

$$\Pr\left\{\boldsymbol{x}_t^{(i)} = \boldsymbol{x}_{t|t-1}^{(j)}\right\} = \boldsymbol{w}_t^{(j)} \quad \forall i, \tag{19}$$

where j is the index obtained from the resampling procedure, and all the weights are reset to be $\frac{1}{N_s}$ (Arulampalam et al., 2002).

After the filtering step, a forward–backward smoother is applied. In particle smoothing, the state particles remain unchanged and only the weights change, which are given by Doucet et al. (2000)

$$\boldsymbol{w}_{t|N}^{(i)} = \boldsymbol{w}_t^{(i)} \left[\sum_{j=1}^{Ns} \boldsymbol{w}_{t+1|N}^{(j)} \frac{p(\boldsymbol{x}_{t+1}^{(j)}|\boldsymbol{x}_t^{(i)}, \boldsymbol{\theta})}{\sum_{k=1}^{Ns} \boldsymbol{w}_t^{(k)} p(\boldsymbol{x}_{t+1}^{(j)}|\boldsymbol{x}_t^{(k)}, \boldsymbol{\theta})}\right]. \tag{20}$$

To calculate the expectations needed in (15), the joint probability density function $p(\boldsymbol{x}_t, \boldsymbol{x}_{t+1}|\mathcal{Y}, \boldsymbol{\theta}_k)$ needs to be obtained, which has weights given by (see Gopaluni, Schön, and Wills (2009) and Schön et al. (2011))

$$\boldsymbol{w}_{t,t+1}^{(ij)} = \boldsymbol{w}_{t|t}^{(i)} \boldsymbol{w}_{t+1|N}^{(j)} \frac{p(\boldsymbol{x}_{t+1}^{(j)}|\boldsymbol{x}_t^{(i)}, \boldsymbol{\theta})}{\sum_{l=1}^{N_s} \boldsymbol{w}_{t|t}^{(l)} p(\boldsymbol{x}_{t+1}^{(j)}|\boldsymbol{x}_t^{(l)}, \boldsymbol{\theta})}. \tag{21}$$

Thus the expectations required in (15) are calculated using

$$\frac{\partial Q(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_k)}{\partial \boldsymbol{\theta}} \approx \sum_{t=0}^{N-1} \sum_{i=1}^{N_s} \sum_{j=1}^{N_s} \boldsymbol{w}_{t,t+1}^{(ij)} \frac{\partial \ln p(z_{t+1}^{(j)}|\boldsymbol{x}_t^{(i)}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \tag{22}$$

at the kth iteration of the EM algorithm.

3.3. The EM parameter estimation algorithm

To start off the EM algorithm, parameter initialisation at $k = 0$ needs to be performed to obtain $\hat{\boldsymbol{\theta}}_0$. We obtain this initial parameter estimate using (16) without the expectations, which is equivalent to the least-squares estimate.

In order to terminate the iterative steps of the EM algorithm, convergence can be monitored using a suitable metric such as the change in parameters $\Delta\theta_k$, where

$$\Delta\theta_k = \frac{1}{d_k} \left\|\hat{\boldsymbol{\theta}}_k - \hat{\boldsymbol{\theta}}_{k-1}\right\|_2^2, \tag{23}$$

where $d_k = \hat{\boldsymbol{\theta}}_k^\top \hat{\boldsymbol{\theta}}_k$. A stopping criterion can be defined based on a threshold, although in practical offline modelling scenarios it is often more constructive for the investigator to visually monitor the change in parameters, which is typically problem dependent.

Algorithm 1 gives a summary of the parameter estimation algorithm.

Algorithm 1: **Parameter estimation**

(1) Initialise $\hat{\boldsymbol{\theta}}_0$.
(2) **for** k=0:stopping criterion
  (a) Perform E-step: First compute the filtered weights and particles using (18) and (17), compute the smoothed weights via (20), then calculate the joint weights from (21), based on $\hat{\boldsymbol{\theta}}_k$.
  (b) Perform M-step: Calculate $\hat{\boldsymbol{\theta}}_{k+1}$ from (16).
end for

4. Structure detection

4.1. NARX model term selection by the F-test

Billings and Voon (1986) have proposed a structure detection metric for identifying the NARX model in a maximum likelihood framework that is based on the F-test. The set of NARX model terms $\{\phi_i\}_{i=1}^{m}$ can be tested for their significance using the statistical t-test, which is given by

$$T(\phi_i)_{N-M} = \frac{\hat{\theta}_i}{\sqrt{P_{ii}}} \quad \text{for } i = 1 \cdots M, \tag{24}$$

with $N - M$ degrees of freedom; P_{ii} is the diagonal element of the ith term of the parameter covariance matrix. However, the t^2-distribution with $N - M$ degrees of freedom is equivalent to the F-distribution with $(1, N - M)$ degrees of freedom. Hence, the F-ratio test can be used to determine the significance of the individual coefficient using[2]

$$F(\phi_i) = \frac{\hat{\theta}_i^2}{P_{ii}} \quad \text{for } i = 1 \ldots M. \tag{25}$$

Structure selection is performed by eliminating terms from an initial model superset $\mathcal{M}_0$, where elimination is performed by removing terms that have an F-value less than some threshold, i.e., $F(\phi_i) < \alpha$, where α denotes the threshold and a low F-value indicates that a term does not significantly improve the fit to the data. The remaining terms give rise to the final reduced model $\mathcal{M}_F$.

4.2. Augmenting the EM algorithm to obtain the parameter covariance

To implement the F-test in (25), we require the estimate of the parameter covariance P. Under some regularity conditions, such as that the true parameter $\boldsymbol{\theta}^*$ must be interior to the parameter space, and the log-likelihood must be twice continuously differentiable in $\boldsymbol{\theta}$ (see Newey and McFadden (1994) for the full set of conditions), which are fulfilled here,

$$\sqrt{N}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) \to \mathcal{N}(0, P), \tag{26}$$

where

$$P = I(\boldsymbol{\theta}^*|\mathcal{Y})^{-1} \tag{27}$$

and $I(\boldsymbol{\theta}^*|\mathcal{Y})^{-1} = -H(\mathcal{Y}|\boldsymbol{\theta}^*)^{-1}$, where $I(\boldsymbol{\theta}^*|\mathcal{Y})$ is the observed data information matrix and $H(\mathcal{Y}|\boldsymbol{\theta}^*)$ is the observed data Hessian matrix, given by

$$H\left(\mathcal{Y}|\boldsymbol{\theta}^*\right) = \frac{\partial^2 \ln p\left(\mathcal{Y}|\boldsymbol{\theta}^*\right)}{\partial \boldsymbol{\theta}^* \partial \boldsymbol{\theta}^{*\top}}. \tag{28}$$

Two methods are described here for obtaining an estimate of the observed information matrix $I(\hat{\boldsymbol{\theta}}|\mathcal{Y})$ from an augmentation of the EM algorithm.

[2] This can also be referred to as the univariate Wald test, which is the asymptotic approximation to the well known t and F tests (Engle, 1984).

4.2.1. Naive estimation of the information matrix

The information matrix estimate described in this section was developed by Louis (1982), and it is obtained in terms of the first and second derivatives of the complete-data log likelihood function introduced within the EM framework. From McLachlan and Krishnan (2008), $I(\boldsymbol{\theta}|\mathcal{Y})$ can be expressed in terms of

$$I(\boldsymbol{\theta}|\mathcal{Y}) = \mathcal{I}_c(\boldsymbol{\theta}|\mathcal{Y}) - \mathcal{I}_m(\boldsymbol{\theta}|\mathcal{Y}), \tag{29}$$

where $\mathcal{I}_c(\cdot) = -\mathbb{E}[H_c(\cdot)]$ is the expected complete-data Hessian matrix given by

$$H_c(\mathcal{X}, \mathcal{Y}|\boldsymbol{\theta}) = \frac{\partial^2 \ln p(\mathcal{X}, \mathcal{Y}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} \tag{30}$$

and $\mathcal{I}_m(\cdot)$ is the missing information matrix given by

$$\mathcal{I}_m(\boldsymbol{\theta}|\mathcal{Y}) = \left[\mathbb{E}\left[S_c(\mathcal{X}, \mathcal{Y}|\boldsymbol{\theta}) S_c^\top(\mathcal{X}, \mathcal{Y}|\boldsymbol{\theta})\right]|\mathcal{Y}\right] - \mathbb{E}\left[S_c(\mathcal{X}, \mathcal{Y}|\boldsymbol{\theta})|\mathcal{Y}\right] \mathbb{E}\left[S_c(\mathcal{X}, \mathcal{Y}|\boldsymbol{\theta})|\mathcal{Y}\right]^\top, \tag{31}$$

where $S_c(\cdot)$ is the gradient vector of the complete-data log-likelihood function, referred to as the score statistic, expressed as

$$S_c(\mathcal{X}, \mathcal{Y}|\boldsymbol{\theta}) = \frac{\partial \ln p(\mathcal{X}, \mathcal{Y}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}. \tag{32}$$

Thus, using the expressions developed for the complete-data log-likelihood in (9)–(12), we obtain

$$S_c(\mathcal{X}, \mathcal{Y}|\hat{\boldsymbol{\theta}}) = \sigma^{-2}\left(\boldsymbol{\Phi}^\top \boldsymbol{z} - \boldsymbol{\Phi}^\top \boldsymbol{\Phi}\hat{\boldsymbol{\theta}}\right) \tag{33}$$

$$H_c(\mathcal{X}, \mathcal{Y}|\hat{\boldsymbol{\theta}}) = -\sigma^{-2}\left(\boldsymbol{\Phi}^\top \boldsymbol{\Phi}\right), \tag{34}$$

where $\boldsymbol{\Phi} = [\boldsymbol{\phi}_0^\top \cdots \boldsymbol{\phi}_{N-1}^\top]^\top$ and $\boldsymbol{z} = [z_1, \ldots, z_N]^\top$. Now, substituting the expression for $S_c(\cdot)$ defined in (33) into (31) and the resultant expression of the missing information $\mathcal{I}_m(\cdot)$ into (29) along with $\mathcal{I}_c(\cdot)$, which is defined using (34), we obtain the naive observed information estimate $I^{(N)}(\hat{\boldsymbol{\theta}}|\mathcal{Y})$

$$I^{(N)}(\hat{\boldsymbol{\theta}}|\mathcal{Y}) = \sigma^{-2}\mathbb{E}[\boldsymbol{\Phi}^\top \boldsymbol{\Phi}] - \sigma^{-4}\mathbb{E}[(\boldsymbol{\Phi}^\top \boldsymbol{z} - \boldsymbol{\Phi}^\top \boldsymbol{\Phi}\hat{\boldsymbol{\theta}}) \times (\boldsymbol{\Phi}^\top \boldsymbol{z} - \boldsymbol{\Phi}^\top \boldsymbol{\Phi}\hat{\boldsymbol{\theta}})^\top], \tag{35}$$

noting that the second part of (31) is zero at the ML estimate.

Remark 1. In our experience, the estimate $I^{(N)}(\hat{\boldsymbol{\theta}}|\mathcal{Y})$ often produces negative values of the parameter covariances. This problem has been highlighted by Duan and Fulop (2011), who attribute the negative covariances to numerical issues in the direct computational calculation of terms on the right-hand side (RHS) of (35). This poor numerical property is the reason that we label the estimate $I^{(N)}(\hat{\boldsymbol{\theta}}|\mathcal{Y})$ 'naive'. We present an alternative robust method for estimating the information matrix in the following section.

4.2.2. Robust estimation of the information matrix

The information matrix estimation method described by Louis (1982) does not ensure a positive semi-definite parameter covariance matrix. However, an alternative robust approach to estimating $I(\boldsymbol{\theta}|\mathcal{Y})$ has been developed by Duan and Fulop (2011) in the context of the EM algorithm. The estimate is based on the Newey–West method (Newey & West, 1987). We outline that robust approach in this section.

The observed-data score statistic $S(\boldsymbol{\theta}|\mathcal{Y})$ can be expressed as

$$S(\mathcal{Y}|\boldsymbol{\theta}) = \frac{\partial \ln p(\mathcal{Y}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}. \tag{36}$$

Louis (1982) shows that (36) can be written in terms of the complete-data smoothed scores as follows:

$$S(\mathcal{Y}|\boldsymbol{\theta}) = \mathbb{E}\left[\left.\frac{\partial \ln p(\mathcal{X}, \mathcal{Y}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right| \mathcal{Y}, \boldsymbol{\theta}\right] = \sum_{t=1}^{N} a_t(\boldsymbol{\theta}), \tag{37}$$

Algorithm 2: **Structure detection for the case of fixed dynamic order n_z and n_u**

(1) Define the superset of initial NARX model terms. $\mathcal{M}_0 = \{\phi_i\}_{i=1}^{m}$,
(2) Estimate parameters for the initial model $\mathcal{M}_0$ using Algorithm 1.
(3) Model structure selection:
  (a) Robustly estimate the parameter information matrix for $\mathcal{M}_0$ using (40).
  (b) Estimate the parameter covariance matrix for $\mathcal{M}_0$ using (27).
  (c) Obtain the F-value, $F(\phi_i)$, corresponding to each model term using (25).
  (d) Define a threshold for the F-test, α, and obtain the set of final model terms, $\mathcal{M}_F = \{\phi_i : \phi_i \in \mathcal{M}_0, F(\phi_i) > \alpha\}$.
(4) (a) Estimate parameters for $\mathcal{M}_F$ using Algorithm 1.
  (b) Estimate the information matrix using (40) and the covariance matrix using (27).
  (c) Obtain the log-likelihood $\ln p(\mathcal{Y}|\boldsymbol{\theta})$ from (44).

where $a_t(\boldsymbol{\theta}) = \mathbb{E}\left[\frac{\partial \ln p(x_t, y_t|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} | \mathcal{Y}, \boldsymbol{\theta}\right]$, so, in our case,

$$a_t(\hat{\boldsymbol{\theta}}) = \sigma^{-2}\mathbb{E}\left[\boldsymbol{\phi}_t^\top z_{t+1} - \boldsymbol{\phi}_t^\top \boldsymbol{\phi}_t \hat{\boldsymbol{\theta}}\right] = \sigma^{-2}\left(\mathbb{E}\left[\boldsymbol{\phi}_t^\top z_{t+1}\right] - \mathbb{E}\left[\boldsymbol{\phi}_t^\top \boldsymbol{\phi}_t\right]\hat{\boldsymbol{\theta}}\right), \tag{38}$$

which is a by-product of the EM algorithm obtained from (16). Duan and Fulop (2011) show that

$$\mathrm{Var}(S(\mathcal{Y}|\boldsymbol{\theta})) = -\mathbb{E}\left[H(\mathcal{Y}|\boldsymbol{\theta})\right] = I(\boldsymbol{\theta}|\mathcal{Y}), \tag{39}$$

and $I(\boldsymbol{\theta}|\mathcal{Y})$ can be approximated using the Newey–West method (Newey & West, 1987). So, the robust estimate $I^{(R)}(\hat{\boldsymbol{\theta}}|\mathcal{Y})$ is

$$I^{(R)}(\hat{\boldsymbol{\theta}}|\mathcal{Y}) \approx \Omega_0 + \sum_{j=1}^{\tau} \omega(j)(\Omega_j + \Omega_j^\top), \tag{40}$$

where τ is needed to account for the dependence between lagged terms of a_t. The term Ω_j is a sample auto-covariance given by

$$\Omega_j = \sum_{t=1}^{N-j} a_t(\hat{\boldsymbol{\theta}}) a_{t+j}^\top(\hat{\boldsymbol{\theta}}), \tag{41}$$

and

$$\omega(j) = 1 - \frac{j}{\tau + 1} \tag{42}$$

are the modified Bartlett weights used to smooth the sample auto-covariance function. $I^{(R)}$ is guaranteed positive semi-definite since it follows from the positive semi-definiteness of the sample auto-covariance function (Newey & West, 1987). The number of lags is chosen based on letting $\tau \to \infty$ as $N \to \infty$ but keeping $\frac{\tau}{N} \to 0$. This guarantees that the Newey–West estimate of $I(\hat{\boldsymbol{\theta}}|\mathcal{Y})$ is consistent.

Remark 2. The Newey–West method ensures a positive semi-definite estimate of $I(\hat{\boldsymbol{\theta}}|\mathcal{Y})$. In a practical scenario, the estimate is likely to generate a positive definite covariance matrix, as noted by Duan and Fulop (2011), because in a typical numerical problem the eigenvalues of the information matrix will be non-zero.

4.3. Structure detection for the case of unknown model orders

Typically, the dynamic orders of the model n_z and n_u are unknown at the outset of identification. A possible solution to this problem is to search over a large range of dynamic orders in a one-pass attempt at identification. Wei, Billings, and Liu (2004) have highlighted the difficulty of searching over a large set of model terms, suggesting instead a preselection of model orders. To overcome the dynamic order selection problem here, we suggest a method inspired by the typical approach for obtaining models in linear system identification using an information criterion (IC) (Ljung, 1999). We iteratively increase the maximum dynamic order and at each stage perform structure detection using Algorithm 2 in order to select a parsimonious description, which is added to a set of preferred models $\mathfrak{M}$. The optimal model is selected from a comparison using an IC.

There are many information criteria that might be used, such as Akaike's information criterion (AIC) (Akaike, 1974), or Schwarz's (also known as the Bayesian information criterion—BIC) (Schwarz, 1978), which form a trade-off between maximising the model's accuracy and minimising the model's complexity. A natural choice here is one that makes use of quantities generated in the augmented EM procedure described above, specifically the stochastic complexity (SC), which was developed by Rissanen (1989). The SC uses the information matrix to penalise complexity, and it is related to the BIC, which it converges to in the limit (Hansen & Yu, 2001), and is also known in other literature as the *maximum a posteriori* criterion (Djuric, 2002). So, here we define the optimal model as the one that minimises the SC,

$$\mathcal{M}^* = \underset{\mathcal{M}\in\mathfrak{M}}{\arg\min}\left(-\ln p(\mathcal{Y}|\boldsymbol{\theta}) + \frac{1}{2}\ln|I(\boldsymbol{\theta}|\mathcal{Y})|\right), \tag{43}$$

where the estimate of the information matrix is obtained as a byproduct of Algorithm 2, and the log-likelihood is obtained by approximating (6) using the particle filter, where (6) can also be expressed as (ignoring $\boldsymbol{\theta}$ for clarity) (Kadirkamanathan, Li, Jaward, & Fabri, 2002)

$$\begin{aligned}\ln p(\mathcal{Y}) &= \ln \prod_{t=1}^{N-1} p(y_{t+1}|y_{1:t}) \\ &= \ln \prod_{t=1}^{N-1} \int p(y_{t+1}|\boldsymbol{x}_{t+1}) p(\boldsymbol{x}_{t+1}|y_{1:t}) d\boldsymbol{x}_{t+1} \\ &= \sum_{t=1}^{N-1} \ln\left[\frac{1}{N_s}\sum_{i=1}^{N_s} p(y_{t+1}|\boldsymbol{x}_{t+1|t}^{(i)})\right],\end{aligned} \tag{44}$$

where $\boldsymbol{x}_{t+1|t}^{(i)}$ are the predicted state particles and $p(y_{t+1}|\boldsymbol{x}_{t+1|t}^{(i)})$ is obtained from (18).

In practice, the use of certain types of basis function to describe the NARX model will typically require the selection of additional context-specific structural parameters. For example, in the case of polynomials this would mean selecting polynomial order and for radial basis functions (RBFs) selection of the widths (centres are dependent on dynamic order and are therefore catered for by Algorithm 3). In such cases it would be straightforward to augment Algorithm 3 to handle additional parameters, retaining the spirit of the method—model selection using SC. For instance, an additional loop could be placed around dynamic order selection to handle polynomial order, and RBF width selection could be incorporated into the EM parameter estimation stage (Gopaluni, 2010). In order to retain generality we have omitted those particulars from explicit consideration here.

The procedure for identifying the structure using SC is described in Algorithm 3.

Algorithm 3: **Structure detection for the case of unknown dynamic order**

(1) Define minimum and maximum dynamic orders over which to search, n_{min} and n_{max}.
(2) Initialise the set of identified models, $\mathfrak{M} = \oslash$.
(3) For $j = n_{min}, \ldots, n_{max}$,
  (a) define $n_z = n_u = j$,
  (b) estimate model $\mathcal{M}_F^{(j)}$ using Algorithm 2,
  (c) add $\mathcal{M}_F^{(j)}$ to the set of preferred models, $\mathfrak{M} = \mathfrak{M} \cup \mathcal{M}_F^{(j)}$.
(4) From (43) obtain the optimal model $\mathcal{M}^*$ that minimises the SC.

5. Numerical example

To investigate the consistency and accuracy of the proposed structure detection and parameter estimation algorithms, we ran a Monte Carlo (MC) simulation in which 100 input–output realisations of a test system were generated.

To generate test data we simulated a discrete-time nonlinear SISO polynomial NARX system, with a polynomial order of $l = 3$ and second-order dynamics for both the input and output, given by

$$\begin{aligned} z_t &= -0.5z_{t-2} + 0.7z_{t-1}u_{t-1} + 0.6u_{t-2}^2 \\ &\quad - 0.7z_{t-2}u_{t-2}^2 + w_t \\ y_t &= z_t + v_t, \end{aligned} \tag{45}$$

where w_t and v_t were zero-mean Gaussian random variables with variances $\sigma^2 = \lambda^2 = 0.01$. This gave an signal-to-noise ratio (SNR) of about 10 dB. The input was set to a zero-mean uniform random sequence $(-1, 1)$.

5.1. Case 1: comparison of EM-PF (EM-particle filter) with FRO-ERR

In this section, we demonstrate that the EM-PF algorithm described in Algorithm 2 produces accurate and consistent estimates of structure and parameters. We benchmark the EM-PF algorithm against a standard and often used method: the forward regression orthogonal (FRO) estimator with error reduction ratio (ERR), described in Korenberg et al. (1988). The model orders were assumed known in this case, and so the initial model $\mathcal{M}_0$ was defined as all possible polynomial terms produced by the definitions of $n_z = 2$, $n_u = 2$, and $l = 3$, which gave a superset composed of 34 terms. For Algorithm 2, the number of particles used in the particle filter and smoother was set to 100.

Algorithm 2 always selected the true model terms, $\theta_1 = z_{t-2}$, $\theta_2 = u_{t-2}^2$, $\theta_3 = z_{t-1}u_{t-1}$, and $\theta_4 = z_{t-2}u_{t-2}^2$, as having a significant contribution to fitting the data when the threshold $\alpha = 6$, for all the 100 datasets, as seen in Fig. 1(a). Similarly, the FRO estimator also always selected the true model terms when the threshold $\rho = 0.01$ is used, as shown in Fig. 1(b). Fig. 3 shows the average (over the 100 MC simulations) convergence over 50 EM iterations for the initial models and 50 EM iterations for the final models. This gave a sufficient number of iterations for the parameters to converge to a steady value.

The parameter estimates obtained from Algorithm 1 were accurate and distributed evenly about the true value: Fig. 2(a) shows a histogram spread of the parameter estimates obtained in the 100 simulations using Algorithm 1. We compared the MC non-parametric empirical estimation of the parameter distribution with the parametric distribution of a single typical estimate, where the covariance was obtained using the augmentation of the EM algorithm defined in (40). The empirical and algorithmic covariances appear to be very similar, emphasising the utility of the EM augmentation, which is of central importance to the structure detection in Algorithm 2. We performed the same analysis for the parameter estimates obtained from the FRO method, which are shown in Fig. 2(b). The parameter means obtained by the FRO over the MC runs appear to be more biased than in the EM-PF—see Table 1. The bias could possibly be reduced if a NARMAX (MA: moving average) model was used (Korenberg et al., 1988). However, that approach would necessarily increase the complexity of identification and produce a model with extra terms.

Table 1
Comparison of true and estimated parameters from final models averaged across all 100 Monte Carlo trials with associated standard deviations. In parentheses are shown the standard deviations obtained from estimating the covariance matrix using (27) in the case of EM, and the information matrix generated as part of the least-squares procedure in the case of FRO.

$\boldsymbol{\theta}^*$	EM-PF estimates	FRO estimates
−0.5	$-0.505 \pm 0.046(\pm 0.071)$	$-0.467 \pm 0.048(\pm 0.032)$
0.7	$0.687 \pm 0.042(\pm 0.089)$	$0.646 \pm 0.045(\pm 0.036)$
0.6	$0.601 \pm 0.024(\pm 0.035)$	$0.582 \pm 0.057(\pm 0.021)$
−0.7	$-0.664 \pm 0.086(\pm 0.140)$	$-0.595 \pm 0.094(\pm 0.070)$

We compared parameter covariances obtained from the MC runs with estimates derived from the parameter estimation methods (EM and orthogonal least squares). For EM-PF, we used (27) and (40), and for FRO the inverse of the information matrix that is generated as part of the least-squares estimate. The covariances obtained from both algorithms are all within the same order of magnitude, demonstrating a consistency across empirical results and theoretical estimates, as seen in Table 1. The EM augmentation for estimating parameter covariance tends to slightly overestimate the variance compared to the MC result, whilst the least-squares approach tends to slightly underestimate the variance.

5.2. Case 2: unknown order of dynamics

In this section, we deal with the situation of unknown dynamics, and demonstrate the use of Algorithm 3 for selecting between models identified assuming either second-order or third-order dynamics, i.e., $\mathcal{M}_F^{(2)}$ or $\mathcal{M}_F^{(3)}$. We applied Algorithm 3 to a set of 25 data sets generated from simulation of the system defined in (45). As above, the threshold for term selection of the second-order dynamic model was set to $\alpha = 6$, whilst for the third-order dynamic model it was set to $\alpha = 1$ (based on inspection of the F-value change with increasing number of model terms). The true model terms were included in the superset of each initial model, $\mathcal{M}_0^{(2)}$ and $\mathcal{M}_0^{(3)}$, where the number of terms in each set was 34 and 86, respectively.

The correct four model terms were always selected for model $\mathcal{M}_F^{(2)}$; this was not the case for $\mathcal{M}_F^{(3)}$, where the correct terms were included in the model in 76% of trials, and many of these models also included additional terms—the exact model was identified in 40% of trials (results shown in Fig. 4(a)). In the remaining trials (24%) for $\mathcal{M}_F^{(3)}$, the identified model omitted a term. The reason why the selection process often included additional terms (or missed a correct term) for $\mathcal{M}_F^{(3)}$ was possibly due to the fact that a large set of initial terms tends to reduce the quality of the parameter estimates, disrupting the structure selection process, as noted by Kukreja et al. (2004).

In a realistic scenario, the correct model would be unknown, which motivates the use of the information criterion based on stochastic complexity (SC) used in Algorithm 3. The optimal model $\mathcal{M}^*$ selected using Algorithm 3 was the true model in all but one case—the SC values are shown in Fig. 4(b). In the single erroneous case, a model with one additional term was preferred. A clear result observed from the comparison of SC values was that when the

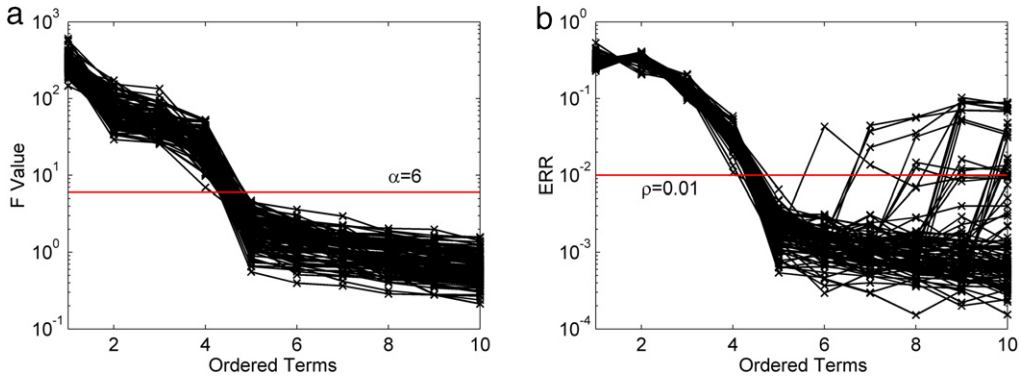

Fig. 1. (a) F-values of ordered model terms with threshold $\alpha = 6$, from 100 Monte Carlo trials. (b) ERR values of model terms from 100 Monte Carlo trials, with threshold $\rho = 0.01$.

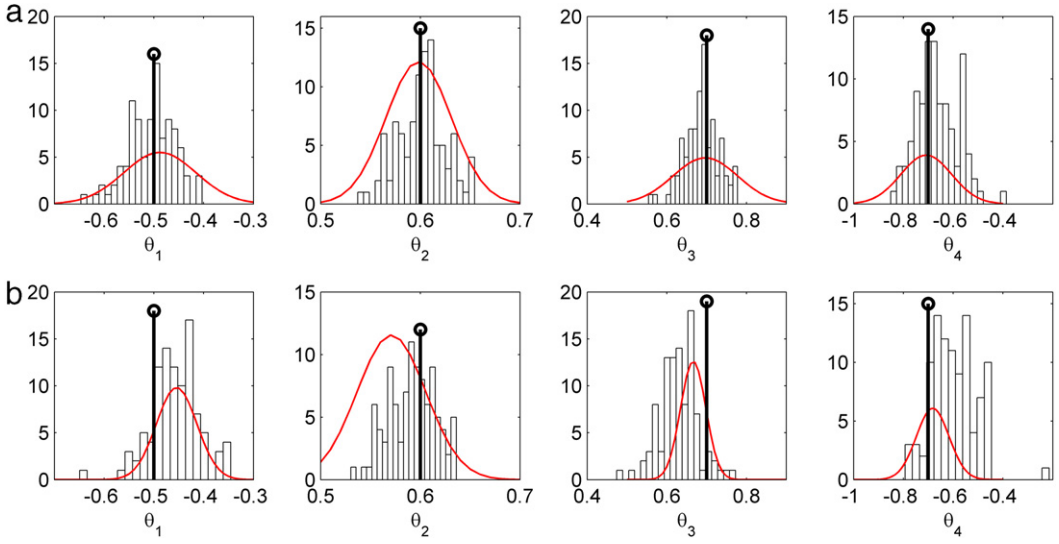

Fig. 2. (a) True parameter values (stem), Monte Carlo spread of parameters from all 100 trials (histogram) using Algorithm 1 and estimated distribution of parameters obtained from one typical Monte Carlo trial, where the parameter covariance was estimated by the augmentation of the EM algorithm defined in (40) (red line). (b) True parameter values (stem), Monte Carlo spread of parameters from all 100 trials (histogram) using FRO and estimated distribution of parameters obtained from one typical Monte Carlo trial (red line). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

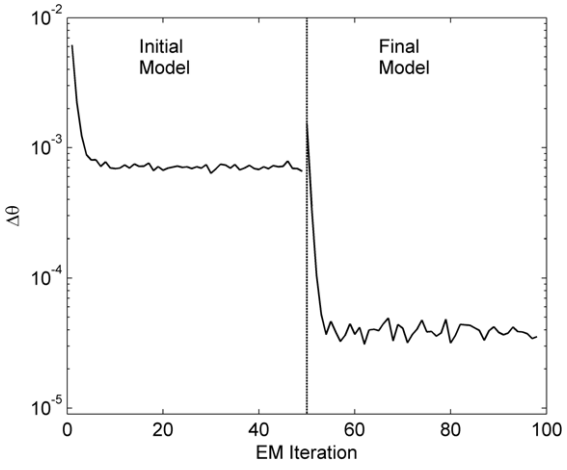

Fig. 3. Average parameter convergence over 50 EM iterations for the initial models and 50 EM iterations for the final models.

identified model was missing a correct term, the SC values were significantly larger than those of the true model, which implies that Algorithm 3 robustly selects the correct model in the case of missing terms.

6. Summary

We have developed a structure detection and parameter estimation procedure for NARX models that is contained within a unified EM framework. The method incorporates parameter estimation using a particle-smoother EM algorithm and model term selection based on the F-test. The F-test is based on an augmentation of the EM algorithm, where the information matrix is robustly estimated, which in turn then leads to an estimate of the covariance matrix from the information matrix inverse. Numerical results from a Monte Carlo example show that the algorithm consistently and successfully detects the model structure and accurately estimates the model parameters. For the realistic scenario of unknown dynamic order, we demonstrate the use of an information criterion (stochastic complexity) for selecting between candidate models, which uses the information matrix (to penalise model complexity) and a likelihood estimate (to quantify model accuracy), each generated as a byproduct of the augmented EM algorithm. Numerical results demonstrated that the EM approach performed well in comparison to the often-used FRO-ERR method, and did not require the additional identification of a noise model to reduce parameter bias from handling measurement noise corrupted output signals.

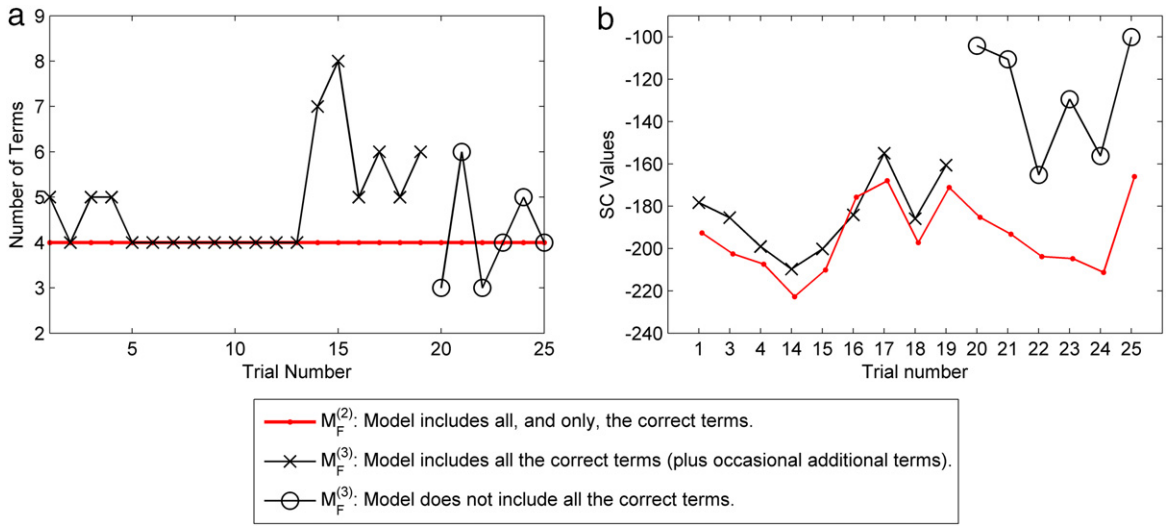

Fig. 4. (a) Comparison of the number of terms selected in the final model identified by Algorithm 2, for maximum dynamic orders of $n_y = n_u = 2$ or 3, where results are partitioned into models that include all the true terms and models with missing terms. (b) Comparison of the stochastic complexity (SC) values for the correctly identified models $M_F^{(2)}$ and the erroneous models $M_F^{(3)}$ shown in (a)—the SC values are minimal for the correct model in all cases except trial 16, demonstrating the high success rate of the selection algorithm for unknown dynamic order (Algorithm 3).

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*, 716–723.

Anderson, S. R., & Kadirkamanathan, V. (2007). Modelling and identification of non-linear deterministic systems in the delta-domain. *Automatica*, *43*, 1859–1868.

Andrieu, C., Doucet, A., Singh, S. S., & Tadic, V. B. (2004). Particle methods for change detection, system identification, and control. *Proceedings of the IEEE*, *92*, 423–438.

Arulampalam, M., Maskell, S., Gordon, N., & Clapp, T. (2002). A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking. *IEEE Transactions on Signal Processing*, *50*, 174–188.

Billings, S., & Voon, W. (1986). A prediction-error and stepwise-regression estimation algorithm for nonlinear systems. *International Journal of Control*, *44*, 803–822.

Dempster, A., Laird, N., & Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society B*, *39*, 1–38.

Djuric, P. M. (2002). Asymptotic MAP criteria for model selection. *Signal Processing, IEEE Transactions on*, *46*, 2726–2735.

Doucet, A., Godsill, S., & Andrieu, C. (2000). On sequential Monte Carlo sampling methods for Bayesian filtering. *Statistics and Computing*, *10*, 197–208.

Doucet, A., & Johansen, A. (2008). A tutorial on particle filtering and smoothing: fifteen years later. In *Technical report department of statistics*. University of British Columbia.

Duan, J.-C., & Fulop, A. (2011). A stable estimator of the information matrix under EM for dependent data. *Statistics and Computing*, *21*, 83–91.

Engle, R. F. (1984). In Z. Griliches, & M. D. Intriligator (Eds.), *Wald, likelihood ratio, and Lagrange multiplier tests in econometrics*: *2. Handbook of econometrics* (pp. 775–826). Elsevier, (chapter 13).

Gibson, S., & Ninness, B. (2005). Robust maximum-likelihood estimation of multivariable dynamic systems. *Automatica*, *41*, 1667–1682.

Gopaluni, B., Schön, T.B., & Wills, A. (2009). Particle filter approach to nonlinear system identification under missing observations with a real application. In *Proceedings of the 15th IFAC symposium on system identification (SYSID)*. Saint-Malo, France.

Gopaluni, R. (2008). A particle filter approach to identification of nonlinear processes under missing observations. *The Canadian Journal of Chemical Engineering*, *86*, 1081–1092.

Gopaluni, R. B. (2010). Nonlinear system identification under missing observations: the case of unknown model structure. *Journal of Process Control*, *20*, 314–324.

Haber, R., & Unbehauen, H. (1990). Structure identification of nonlinear dynamic systems—a survey on input/output approaches. *Automatica*, *26*, 651–667.

Hansen, M. H., & Yu, B. (2001). Model selection and the principle of minimum description length. *Journal of the American Statistical Association*, *96*, 746–774.

Isaksson, A. J. (1993). Identification of ARX-models subject to missing data. *IEEE Transactions on Automatic Control*, *38*, 813–819.

Kadirkamanathan, V., & Anderson, S. R. (2008). Maximum-likelihood estimation of delta-domain model parameters from noisy output signals. *IEEE Transactions on Signal Processing*, *56*, 3765–3770.

Kadirkamanathan, V., Li, P., Jaward, M. H., & Fabri, S. G. (2002). Particle filtering-based fault detection in non-linear stochastic systems. *International Journal of Systems Science*, *33*, 259–265.

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

Kukreja, S. L., Galiana, H. L., & Kearney, R. E. (2004). A bootstrap method for structure detection of NARMAX models. *International Journal of Control*, *77*, 132–143.

Leontaritis, I. J., & Billings, S. A. (1985). Input–output parametric models for non-linear systems part 1: deterministic nonlinear systems. *International Journal of Control*, *2*, 303–328.

Li, K., Peng, J. X., & Bai, E. W. (2006). A two-stage algorithm for identification of nonlinear dynamic systems. *Automatica*, *42*, 1189–1197.

Ljung, L. (1999). *System identification—theory for the user* ((2nd ed.)). Upper Saddle River, NJ, Prentice Hall.

Louis, T. A. (1982). Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society, Series B*, *44*, 226–233.

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

McLachlan, G. J., & Krishnan, T. (2008). *The EM algorithm and extensions*. Wiley Interscience.

Newey, W. K., & McFadden, D. (1994). In R. Engle, & D. McFadden (Eds.), *Large sample estimation and hypothesis testing*: *4. Handbook of econometrics* (pp. 2111–2245). Elsevier, (chapter 36).

Newey, W. K., & West, K. D. (1987). A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica*, *55*, 703–708.

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

Rissanen, J. (1989). *Stochastic complexity in statistical inquiry*. Singapore: World Scientific.

Roweis, S. T., & Ghahramani, Z. (2001). Learning nonlinear dynamical systems using the expectation–maximization algorithm. In S. Haykin (Ed.), *Kalman filtering and neural networks*. New York: John Wiley and Sons.

Schön, T. B., Wills, A., & Ninness, B. (2006). Maximum likelihood nonlinear system estimation. In *Proceedings of the 14th IFAC symposium on system identification* (pp. 1003–1008). Australia: Newcastle.

Schön, T. B., Wills, A., & Ninness, B. (2011). System identification of nonlinear state-space models. *Automatica*, *47*, 39–49.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*, 461–464.

Shumway, R. H., & Stoffer, D. S. (1982). An approach to time series smoothing and forecasting using the EM algorithm. *Journal of Time Series Analysis*, *3*, 253–264.

Shumway, R. H., & Stoffer, D. S. (2000). *Time series analysis and its applications*. Berlin: Springer.

Sjoberg, J., Zhang, Q., Ljung, L., Benveniste, A., Delyon, B., Glorennec, P., et al. (1995). Nonlinear black box modeling in system identification: a unified overview. *Automatica*, *31*, 1691–1724.

Van Overschee, P., & De Moor, B. (1996). *Subspace identification of linear systems: theory, implementation, applications*. Kluwer Academic Publishers.

Wei, H. L., Billings, S. A., & Liu, J. (2004). Term and variable selection for nonlinear system identification. *International Journal of Control*, *77*, 86–110.

Wills, A., Schön, T., & Ninness, B. (2008). Parameter estimation for discrete-time nonlinear systems using EM. In *Proceedings of the 17th IFAC world congress*. (pp. 4012–4017). South Korea.

Wu, C. (1983). On the convergence properties of the EM algorithm. *The Annals of Statistics*, *11*, 95–103.

Zia, A., Kirubarajan, T., Reilly, J., Yee, D., Punithakumar, K., & Shirani, S. (2008). An EM algorithm for nonlinear state estimation with model uncertainties. *IEEE Transactions on Signal Processing*, *56*, 921–936.

Tara Baldacchino received her BEng (Hons.) degree in Electrical Engineering from the Faculty of Engineering, University of Malta, Malta, in 2006. In 2008 she obtained an MSc in Control Systems and in 2011 she completed her Ph.D. degree, both in the Department of Automatic Control and Systems Engineering, University of Sheffield, UK. Her research interests include nonlinear systems modelling, specifically using MCMC techniques and particle filtering.

Sean R. Anderson received his MEng degree in Control Systems Engineering from the Department of Automatic Control and Systems Engineering (ACSE), University of Sheffield, UK, in 2001, and his Ph.D. degree from the Department of Chemical and Process Engineering in 2005 (also at the University of Sheffield). From 2005 to 2010 he was a Research Associate in the Centre for Signal Processing in Neuroimaging and Systems Neuroscience, University of Sheffield, studying motor control and sensory processing in animals and humans from a systems engineering perspective, and translating bioinspired control algorithms to neurorobotic systems. He is currently a Lecturer in the Department of ACSE at Sheffield working on systems modelling and control with applications in biological processes. His research interests include the identification of continuous-time and discrete-time nonlinear dynamic systems, the study of adaptive and optimal control in biological systems, and bioinspired control in neurorobotics.

Visakan Kadirkamanathan received his BA and Ph.D. degrees in Electrical and Information Engineering from the University of Cambridge, UK. He held Research Associate positions at the University of Surrey and the University of Cambridge before joining the Department of Automatic Control and Systems Engineering, The University of Sheffield, as a Lecturer in 1993. He is currently the Head of the Department and a Professor of Signal and Information Processing, affiliated to the Centre for Signal Processing and Complex Systems. His research interests include nonlinear signal processing, system identification, intelligent control, and fault diagnosis, with applications in systems biology, aerospace systems, and wireless communication. He has coauthored a book on intelligent control and has published more than 140 papers in refereed journals and proceedings of international conferences. Professor Kadirkamanathan is the Co-Editor of the International Journal of Systems Science and has served as an Associate Editor for the IEEE Transactions on Neural Networks and the IEEE Transactions on Systems, Man, and Cybernetics, Part B. He was also the Conference Chair of the 4th IAPR International Conference on Pattern Recognition in Bioinformatics (PRIB 2009) held in 2009 at Sheffield, UK.