Journal of Control, Automation and Electrical Systems (2019) 30:453–464
https://doi.org/10.1007/s40313-019-00476-9

NARX Model Identification Using Correntropy Criterion in the Presence of Non-Gaussian Noise

Ícaro B. Q. Araújo[1,2] · João P. F. Guimarães[2] · Aluísio I. R. Fontes[3] · Leandro L. S. Linhares[3] · Allan M. Martins[4] · Fábio M. U. Araújo[4]

Received: 25 September 2018 / Revised: 7 May 2019 / Accepted: 18 May 2019 / Published online: 29 May 2019

Abstract
In past years, the system identification area has emphasized the identification of nonlinear dynamic systems. In this field, polynomial nonlinear autoregressive with exogenous (NARX) models are widely used due to flexibility and prominent representation capabilities. However, the traditional identification algorithms used for model selection and parameter estimation with NARX models have some limitation in the presence of non-Gaussian noise, since they are based on second-order statistics that tightly depend on the assumption of Gaussianity. In order to solve this dependence, a novel identification method called simulation correntropy maximization with pruning (SCMP) based on information theoretic learning is introduced by this paper. Results obtained in non-Gaussian noise environment in three experiments (numerical, benchmark data set and measured data from a real plant) are presented to validate the performance of the proposed approach when compared to other similar algorithms previously reported in the literature, e.g., forward regression with orthogonal least squares and simulation error minimization with pruning. The proposed SCMP method has shown increased accuracy and robustness for three different experiments.

Keywords Nonlinear system identification · Polynomial NARX models · Model structure selection · Non-Gaussian noise · Maximum correntropy criterion

1 Introduction

The system identification is a process of constructing a mathematical model capable of representing its main characteristics through observations.

In black-box system identification, the polynomial nonlinear autoregressive moving average with exogenous input (NARMAX) (Leontaritis and Billings 1985a, b) representation has a great performance in their ability to represent nonlinear input–output relations (Yan and Deller 2016; Zhao et al. 2018) as a functional expansion of lagged input, output and noise data. In cases where the deterministic input–output relationship is the focus, a nonlinear autoregressive with exogenous input (NARX) model can be employed, using a simplification of the disturbance model (Zhao and Chen 2012). This class of representation can be used in control problems when the main goal is to find a simple, but functional, description of the system.

In practical applications, the experimental data set used in the identification procedure is often corrupted with outliers (Liu and Chen 2013; Linhares et al. 2015). Usually, the NARX parameters are estimated using the least squares (LS), which are non-optimal in the presence of noise with

✉ Ícaro B. Q. Araújo
icaro@ic.ufal.br

João P. F. Guimarães
joao.guimaraes@ifrn.edu.br

Aluísio I. R. Fontes
aluisio.rego@ifrn.edu.br

Leandro L. S. Linhares
leandro.luttiane@ifrn.edu.br

Allan M. Martins
allan@dee.ufrn.br

Fábio M. U. Araújo
meneghet@dca.ufrn.br

1 Computer Institute, Federal University of Alagoas, Maceió, AL 57072-900, Brazil

2 Department of Computer Engineering and Automation, Federal University of Rio Grande do Norte, Natal, RN 59078-970, Brazil

3 Federal Institute of Rio Grande do Norte, Pau dos Ferros, RN 59900-000, Brazil

4 Federal University of Rio Grande do Norte, Natal, RN 59078-970, Brazil

non-Gaussian distributions (Santamaria et al. 2006; Liu et al. 2007). Although there are many outlier detection methods, many approaches are not able to eliminate all of them, and the resulting data obtained after the application of such methods may still be contaminated (Liu and Chen 2013; Linhares et al. 2015). Also, regarding the NARX model structure selection (MSS), the number of candidate regressors increases rapidly along with the model order and maximum delays of the input and output signals (Cheng et al. 2009). Aguirre and Billings (1995) present examples of dynamic systems to illustrate that the models that best fit the estimation data are not necessarily models that capture the underlying dynamics appropriately. The authors also show that, despite the predictive capacity of the estimation data of superparametrized models, if a NARX model is unnecessarily complex, the model can induce spurious dynamics.

Recently, a similarity measure called correntropy was introduced (Liu et al. 2006, 2007; Santamaria et al. 2006). The correntropy idea has extended the concept of mean square error (MSE) adaptation to include descriptors of entropy and divergence, so useful in information theory. It preserves the nonparametric nature of MSE but extracts more information from the data structure and yields, therefore, solutions that are more accurate than MSE for non-Gaussian processes. Correntropy can be applied as a cost function for system identification with the advantage that it is a local criterion of similarity. Correntropy has been used in several applications including system identification problems (Liu and Chen 2013; Linhares et al. 2015; Guimaraes et al. 2016; Peng et al. 2017; Kulikova 2017; Fontes et al. 2015, 2017), with good performance in non-Gaussian noise environments.

This paper proposes an algorithm capable of estimating parameters and selecting the NARX model structure in the presence of non-Gaussian distribution noise. The proposed algorithm is called simulation correntropy maximization with pruning (SCMP), which uses correntropy as a similarity measure in order to select the structure of the mathematical NARX model and the maximum correntropy criteria (MCC) to estimate the parameters. To ensure a better performance of the MCC gradient solution, it was employed a variable kernel width (VKW-MCC) (Huang et al. 2017a) method to iteratively determine the value of the kernel width. The results obtained in non-Gaussian environment are presented to validate the advantages of the proposed approach.

The performance of SCMP has shown increased accuracy and robustness in different dynamic systems, when compared to a traditional algorithm such as the forward regression with orthogonal least squares (FROLS) and simulation error minimization with pruning (SEMP) algorithm.

The paper is organized as follows. Section 2 provides the basic framework and notation for nonlinear system identification of NARX models and briefly reviews the main approaches in the literature. Section 3 discusses the correntropy criterion and its importance in the presence of non-Gaussian noise. The proposed method is illustrated in Sect. 4 and then tested in three experiments in Sect. 5. Finally, some concluding remarks are drawn in Sect. 6.

2 NARX Models

NARX models (Leontaritis and Billings 1985a) are discrete-time representations that evidence the output value $y(k)$ as a function of previous values for the output and input signals according to Eq. (1)

$$y(k) = F^l[y(k-1), \ldots, y(k-n_y), \\ u(k-d), \ldots, u(k-d-n_u)] \quad (1)$$

where F^l is a nonlinear function with nonlinearity degree l, $y(k)$ is the output signal at an instant k, u is the input signal, d is the delay time and n_y and n_u are the maximum lags for the output and the input, respectively, and $n = n_y + n_u$.

The polynomial approximation of nonlinearity degree l for model (1) has the following structure (Chen and Billings 1989)

$$y(k) = \theta_0 + \sum_{i_1=1}^{n} \theta_{i_1} x_{i_1}(k) + \sum_{i_1=1}^{n} \sum_{i_2=i_1}^{n} \theta_{i_1 i_2} x_{i_1}(k) x_{i_2}(k) \\ + \sum_{i_1=1}^{n} \cdots \sum_{i_l=i_{l-1}}^{n} \theta_{i_1 \cdots i_l} x_{i_1}(k) \cdots x_{i_l}(k) + e(k) \quad (2)$$

where

$$x_1(k) = y(k-1), x_2(k) = y(k-2), \ldots, \\ x_{n_y+1} = u(k-d), \ldots, x_{n_y+n_u} = u(k-d-n_u)$$

with $n = n_y + n_u$.

Equation (2) is a general model structure. The matrix formulation represented in Eq. (3) is used for the estimation problem.

$$\boldsymbol{y} = \boldsymbol{\Psi}\hat{\boldsymbol{\theta}} + \boldsymbol{e} \quad (3)$$

where $\boldsymbol{\Psi} = \left[\boldsymbol{\psi}_1 \boldsymbol{\psi}_2 \cdots \boldsymbol{\psi}_n\right]$ is the matrix of regressors (independent variables) with column size N (number of observations), $\boldsymbol{\psi}_i$, with $i = 1, \ldots, n$ being the regressors columns, which corresponds to the different terms in the polynome, $\boldsymbol{\theta}$ is the vector with the respective parameters, and $\boldsymbol{e}$ is the system noise and modeling error.

The maximum number of candidate regressors (n_θ) grows with the increase in the degree of nonlinearity (l) and the maximum output and input delays (n_y and n_u) (and n_e, for NARMAX models).

Typical solutions for structure selection problems are the forward regression orthogonal least squares algorithm (FROLS) which uses an importance index, the error reduction ratio (ERR) (Billings et al. 1988), and also the simulation error minimization with pruning (SEMP) method associated with the simulation error reduction ratio (SRR) (Piroddi and Spinelli 2003) and its variants.

In the FROLS method, a new regressor is included in the model for each algorithm iteration according to ERR, which evaluates the improvement that can be gained by adding the regressor to the current model. The method also exploits orthogonal least squares (OLS) to decouple the estimation of the various regressors. Several variants of this method have been introduced in the literature (Falsone et al. 2015).

Instead of the prediction error minimization (PEM) paradigm, the SEMP algorithm uses simulation error minimization (SEM) paradigm, and it provides more accurate and robust identification. However, this method has a much larger computational cost, compared to FROLS method.

Despite their distinct performances regarding accuracy and computational cost, both methods do not provide good results in the presence of non-Gaussian noise since they are based on second-order moments, e.g., MSE (Principe 2010).

3 Correntropy

Correntropy is a generalized similarity measure between two arbitrary scalar random variables X and Y and defined by (Santamaria et al. 2006) as:

$$V_\sigma(X,Y) = E(k_\sigma[X,Y]) \tag{4}$$

where $E[.]$ denotes the expectation operator and $k_\sigma(.,.)$ corresponds to any positive-definite symmetric kernel. This work employs a Gaussian kernel $G_\sigma(x,y)$ defined as:

$$G_\sigma(x,y) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(x-y)^2}{2\sigma^2}\right) \tag{5}$$

where σ is kernel width parameter (or kernel width), which is a free parameter. Then, one could estimate the correntropy V_σ between two random variables X, Y as:

$$\hat{V}_\sigma(X,Y) = \frac{1}{\sqrt{2\pi}\sigma}\frac{1}{N}\sum_{i=1}^{N}\exp\left(-\frac{(x(i)-y(i))^2}{2\sigma^2}\right) \tag{6}$$

Some interesting properties were presented in Santamaria et al. (2006) and Liu et al. (2007) when the Gaussian kernel is used. It makes correntropy symmetric, positive, bounded, and able to extract high-order statistical information from data. To illustrate this last and important property, one can use the Taylor series expansion of the Gaussian function in (4) to obtain the following representation:

$$\begin{aligned} V_\sigma(X,Y) &= E[G_\sigma(x,y)] \\ &= \frac{1}{\sqrt{2\pi}\sigma}\sum_{k=0}^{\infty}\frac{(-1)^k}{2^k\sigma^{2k}k!}E\left[(X-Y)^{2k}\right] \end{aligned} \tag{7}$$

As shown in Eq. (7), the Gaussian kernel makes correntropy a weighted sum of all the even moments from the random variable $(X-Y)$. The kernel width σ appears as a weighting parameter that controls which moments are used, being an effective mechanism to reject outliers in non-Gaussian noise environments. Increasing the kernel width value makes correntropy tend to correlation (Principe 2010).

3.1 Maximum Correntropy Criteria

In order to take advantage of robustness provided by correntropy in non-Gaussian environments, it is possible to define the cost function J as the correntropy between the real output signal $\boldsymbol{y}$ and the estimated output signal $\hat{\boldsymbol{y}}$.

$$J = \hat{V}_\sigma(y,\hat{y}) = \frac{1}{\sqrt{2\pi}\sigma N}\sum_{i=1}^{N}\exp\left(-\frac{\left(y(i)-\hat{y}(i)\right)^2}{2\sigma^2}\right) \tag{8}$$

One could maximize $V_\sigma(y,\hat{y})$, and consequently, the similarity between the desired signal and the system output, which will minimize the error $e = y - \hat{y}$. This approach is called maximum correntropy criteria.

Since $\hat{\boldsymbol{y}} = \boldsymbol{\psi}\hat{\boldsymbol{\theta}}$, one could make

$$\boldsymbol{\theta}(k+1) = \boldsymbol{\theta}(k) + \eta\nabla J \tag{9}$$

where

$$\nabla J = \frac{\partial J}{\partial \boldsymbol{\theta}} = V_\sigma(Y,\hat{Y}) = E[G_\sigma(\boldsymbol{Y}, \boldsymbol{\Psi}\hat{\boldsymbol{\theta}})] \tag{10}$$

Then, a simple iterative gradient solution can be used to determine the update rule for $\boldsymbol{\theta}$

$$\begin{aligned} \boldsymbol{\theta}(k+1) = \boldsymbol{\theta}(k) + \frac{\eta}{N\sqrt{2\pi}\sigma^3} \times \\ \sum_{i=k-N+1}^{k}\left[\exp\left(-\frac{\boldsymbol{e}(i)^2}{2\sigma^2}\right)\boldsymbol{e}(i)\boldsymbol{\Psi}^T(i)\right] \end{aligned} \tag{11}$$

where $\boldsymbol{\Psi}(i)$ is the ith line of regressor matrix $\boldsymbol{\Psi}$ which corresponds to the ith observation.

It is also possible to approximate the sum to the current value $(N = 1)$ inspired by the stochastic gradient as in Singh

and Príncipe (2010).

$$\boldsymbol{\theta}(k+1) = \boldsymbol{\theta}(k) + \frac{\eta}{N\sqrt{2\pi}\sigma^3} \exp\left(-\frac{\boldsymbol{e}(k)^2}{2\sigma^2}\right) \boldsymbol{e}(k)\boldsymbol{\Psi}^T(k) \tag{12}$$

The practical consequence of the use of correntropy to treat non-Gaussian cases (in general, since its expansion contains all the even statistical moments, etc) is not to capture nonlinearities in the generating model. Rather due to the exponential decaying of the kernel, outliers in the noise are naturally discarded by the cost function. Bimodal and alpha-stable noise types are very harmful to mean square cost functions.

The kernel width σ appears as a free parameter. According to (7), it can be stated that σ affects directly the convergence rate, robustness, and steady-state performance of the adaptive filtering (Principe 2010).

Due to the importance of the kernel width, this free parameter must be properly chosen to ensure good performance. The definition of a fixed and optimal value for the kernel width is not a trivial task, once it changes according to data and application nature (Huang et al. 2017b; Santamaria et al. 2006). In order to overcome this drawback, an adaptive kernel width algorithm is used to properly determine the value for the kernel width iteratively.

The kernel width acts as a zoom lens controlling the observation window which the similarity between two random variables is assessed. The kernel width plays a key role in the MCC performance, since this parameter affects the stability of weight tracks, convergence speed, and presence of local optima (Singh 2010). A good option to deal with the selection of the kernel width is to use an adaptive method to make its adjustment.

Several adaptive kernel width methods can be found in the literature. They are based on the fact that the statistics of the error changes continuously during the model parameters estimation. Therefore, the main goal is to improve the identification method by adapting the kernel width to best suit the error signal at each iteration (Singh and Príncipe 2011). The adaptive kernel width MCC (AMCC) algorithm was proposed in Wang et al. (2015a) aiming to improve the convergence speed, mainly when the initial model parameter vector is far from being optimal The switch kernel width method of correntropy (SMCC) updates the kernel width based on the instantaneous error between the estimated and the desired signal in order to adjust such parameter for each iteration (Wang et al. 2015b). Recently, the technique developed in Huang et al. (2017b) called variable kernel width-maximum correntropy criterion (VKW-MCC) has been suggested as a solution capable of searching for the best kernel width at each iteration, thus implying reduced error. This strategy is able to provide fast convergence rate and stable steady-state performance. The choice of this method is justified by its superior performance when compared to others adaptive kernel methods found in the literature (Huang et al. 2017a).

3.1.1 Variable Kernel Width MCC (VKW-MCC)

The VKW-MCC algorithm calculates the kernel width at each iteration by maximizing $\exp(-e^2/2\sigma^2)$ with respect to the kernel width σ (Huang et al. 2017b). For this purpose, the authors employ a modified cost function to reduce the interference of the kernel width. Instead of making $J(k) = E[G_{\sigma(e)}]$, the following statement is considered:

$$J(k) = E\left[\sigma^2 G_\sigma(\boldsymbol{e})\right]. \tag{13}$$

Using the gradient ascent approach, the modified MCC algorithm is given as:

$$\boldsymbol{\theta}(k+1) = \boldsymbol{\theta}(k) + \mu \exp\left(-\frac{\boldsymbol{e}(k)}{2\sigma^2}\right) \boldsymbol{e}(k)\Psi^T(k) \tag{14}$$

with $\mu = \frac{\eta}{N\sqrt{2\pi}\sigma^3}$. The choice of the μ value is performed empirically because the cost function is nonlinear and not convex. However, since the gradient expression is always less than the gradient of a convex function, the choice of μ is not difficult to do empirically. It may even be estimated, in critical cases, as a proportion of the greater self-value of the $\boldsymbol{e}^2$.

Assuming that the noise is not impulsive, the work developed in Huang et al. (2017b) has also shown that the optimal kernel width in the kth iteration is given by:

$$\sigma_k = k_\sigma \left|\boldsymbol{e}_k\right|, \tag{15}$$

where k_σ is a positive constant. In order to ensure a robust response to impulsive noise (Huang et al. 2017c), the VKW-MCC method computes $E[|\boldsymbol{e}(k)|]$ instead of $|\boldsymbol{e}(k)|$ in (15), i.e.,

$$\overline{\boldsymbol{e}}(k) = \tau\overline{\boldsymbol{e}}(k-1) + (1-\tau)min\left(A_{e,k}\right). \tag{16}$$

where τ is a smoothing factor that can assume any value between 0 and 1 and $A_{e,k}$ is a set of values $|\boldsymbol{e}(k)|$ in the form:

$$A_{e,k} = [\,|\boldsymbol{e}(k)| \;\; |\boldsymbol{e}(k-1)| \;\; \cdots \;\; |\boldsymbol{e}(k-N_w+1)|\,]. \tag{17}$$

being N_w the length of the estimation window. Then, Eq. (15) can be rewritten as:

$$\sigma(k) = k_\sigma \overline{\boldsymbol{e}}(k). \tag{18}$$

The authors in Huang et al. (2017b) also mention that the VKW-MCC algorithm lacks of robustness when σ_k is too large. To prevent this from happening, the kernel width must be within an interval $[0, \sigma_0]$.

4 Simulation Correntropy Maximization with Pruning (SCMP)

Similarly to FROLS and SEMP methods for MSS problem, the SCMP adds a new regressor at each iteration and it selects the one that best fits the data.

Initially, it is considered a matrix that has all the candidate regressors $\boldsymbol{\Psi}_M$ with dimension $N \times n_\theta$, where N is the number of observations and n_θ is the number of regressors. The matrix $\boldsymbol{P} = [\boldsymbol{p}_1 \quad \dots \quad \boldsymbol{p}_j]$, with $j \leq n_\theta$, represents the set of candidate regressors that are in the current model of iteration j, and is initialized as $\boldsymbol{P} = \{\}$. The matrix $\boldsymbol{Q} = [\boldsymbol{q}_1 \quad \dots \quad \boldsymbol{q}_{n_\theta - j}]$ represents the set of candidate regressors that are outside the current model during iteration j. $\boldsymbol{Q}$ is initialized as $\boldsymbol{Q} = \boldsymbol{\Psi}_M$.

At each iteration, a term j of matrix $\boldsymbol{Q}$ is sequentially added to matrix $\boldsymbol{P}$. For each term added to the model, the algorithm finds the parameters $\boldsymbol{\theta} \in \mathbb{R}^{\boldsymbol{P}}$ using an estimator. At the end of the $j \leq n_\theta$ iterations, the matrix $\boldsymbol{P}$ is a subset of $\boldsymbol{\Psi}_M$ with the best set of regressors.

The FROLS and SEMP methods use LS-based estimators. The SCMP uses the iterative gradient solution for maximizing the correntropy, as described in (12). In this step, the equation of the estimator is represented by

$$\boldsymbol{\theta}(n+1) = \boldsymbol{\theta}(n) + \mu \exp\left(-\frac{\boldsymbol{e}(n)}{2\sigma_e^2}\right) \boldsymbol{e}(n) \boldsymbol{P}(n)_{test}^T \tag{19}$$

where $\hat{\boldsymbol{y}}$ is the estimated output and $n = 1, \dots, N$ and $\boldsymbol{P}(n)_{test}$ is the nth line of the current matrix regressor.

The model with structure characterized by the regressor matrix $\boldsymbol{P}_j$ and $\boldsymbol{\theta}_j$ can be evaluated in terms of its prediction or simulation performance. The SCMP algorithm uses the correntropy cost function (8) in the prediction and simulation performance (similar to the SEMP method) of the current model to evaluate each candidate regressor. In this case, the goal is to maximize the similarity between the data set and the model. For this to happen, the SCMP uses the simulation similarity maximization rate (SSMR), which has robustness to non-Gaussian noise, and it is described by the following expression:

$$[\text{SSMR}]_i = \frac{\hat{V}_\sigma\left(\boldsymbol{y}, \hat{\boldsymbol{y}}_{M_{i+1}}\right) - \hat{V}_\sigma\left(\boldsymbol{y}, \hat{\boldsymbol{y}}_{M_i}\right)}{\frac{1}{N}\sum_{k=1}^{N} \boldsymbol{y}^2(k)} \tag{20}$$

where M_i is the current model and M_{i+1} is the model with the regressor under test and $\hat{V}_\sigma$ is the MCC described in (8). This expression indicates the portion of the output variance explained by the addition of a new term to the model (similarity to ERR and SRR).

After the regressor addition, a recursive test for redundant terms is performed, and terms are eliminated as long as the combined addition and pruning function improves the model accuracy. A complete iteration of the SCMP is thus guaranteed to maximize the SSMR. The pseudo-algorithm is described in (1).

Algorithm 1 SCMP

Initialization: σ, μ, ρ

$\boldsymbol{P} = [\,]$
$\boldsymbol{Q} = \boldsymbol{\Psi}_M$
for $i = 1 : n_\theta$ **do**
  for $j = 1$: size of $\boldsymbol{Q}$ **do**
    $\boldsymbol{P}_{test} = [\boldsymbol{P}\ \boldsymbol{q}_j]$
    Calculation of $\boldsymbol{\theta}_j$ using 19
    Calculation of J_j using 20
  end for
  l = position of argmax J
  if $J_l > J_{l_{old}}$ & $|J_{l_{old}} - J_l| > \rho$ **then**
    $\boldsymbol{P} = [P\ q_l]$
    $\boldsymbol{q}_l = [\,]$
  else
    END
  end if
  for $k = 1 : $ size of $\boldsymbol{P}$ **do**
    $\boldsymbol{R} = \boldsymbol{P}$ without p_k
    Calculation of $\boldsymbol{\theta}_k$ using 19
    Calculation of Jp_k using 20
  end for
  m = position of argmax Jp
  if $Jp_m > J_{l_{old}}$ **then**
    $\boldsymbol{P} = [\boldsymbol{P}$ without $\boldsymbol{p}_m]$
    GOTO 16
  else
    GOTO 4
  end if
end for

In Algorithm 1, it is possible to observe that the identification process is divided into two steps: the estimation process and the choosing of the best candidate model that fits the data. To highlight, Fig. 1 shows the proposed method components diagram.

The SCMP algorithm, as described in Algorithm 1, uses two kernel size values. The first one (σ_e) is used in the parameter estimation step, which uses (19). The second one (σ_s) is used to verify the quality of the model being evaluated during the current iteration, according to (8). It is interesting to note that the values of σ_e and σ_s are not necessarily equal. Thus, to reduce the amount of free parameters, SCMP uses the adaptive kernel methodology during the parameter estimation step.

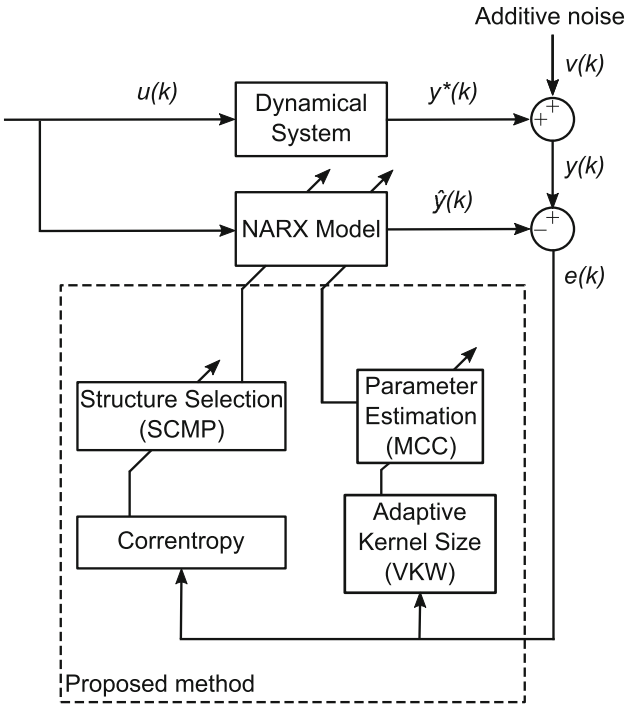

Fig. 1 Proposed identification architecture used in this paper. The correntropy function is used to select the structure while the maximum correntropy criteria is used to estimate the parameters of the NARX model. The output of the dynamic system is polluted with impulsive noise

5 Non-Gaussian Noise

To evaluate the performance of the proposed algorithm in a non-Gaussian noise environment, this paper uses three different strategies to create impulsive noise. The first approach is achieved from the summation

$$\Upsilon(\varrho, \mu_1, \sigma_1, \mu_2, \sigma_2) = (1-\varrho)\mathcal{N}(\mu_1, \sigma_1) + \varrho\mathcal{N}(\mu_2, \sigma_2) \tag{21}$$

where $\varrho\mathcal{N}(\mu, \sigma)$ is a Gaussian distribution with mean μ and variance σ. The ϱ represents the percentage of the samples concentrated in that mode, i.e., $\Upsilon(\varrho = 0.1, \mu_1 = 0, \sigma_1 = 1, \mu_2 = 2, \sigma_2 = 1)$, 90 % from the data is concentrated in the mode with 0 mean and variance 1 while 10 % is in the mode with mean 2 and variance 1.

The second approach is by using a stable distribution (Shao and Nikias 1993) to model the noise. The paper (Weron and Weron 1995) highlights that Lévy α-stable random variables can be achieved by their characteristic function

$$\log\phi(t) = \begin{cases} -\sigma^{\alpha}|t|\{1 - j\beta\text{sign}(t)\tan\frac{\pi\alpha}{2}\} + j\nu t, & \alpha \neq 1 \\ -\sigma|t|\{1 + j\beta\text{sign}(t)\frac{2}{\pi}\log|t|\} + j\mu t, & \alpha = 1 \end{cases} \tag{22}$$

where α is an index of stability in (22) and may vary as $(0 < \alpha \leq 2)$. The smaller the value of alpha, the longer the distribution tail will be, controlling how impulsive it is. On the other hand, using $\alpha = 2$ makes it equivalent to a Gaussian distribution. The $\beta \in [-1, 1]$ is a skewness parameter, and $\nu > 0$ is a scale parameter. $\mu \in \mathbb{R}$ is a location parameter while j is the imaginary unit. Lastly, sign is the sign function, which is defined as

$$sign(u) = \begin{cases} -1 & u < 1 \\ 0 & u = 0 \\ 1 & u > 1 \end{cases} \tag{23}$$

In this paper, all the simulations use $\beta = 0$ and $\mu = 0$. The generalized signal-to-noise error (GSNR) (Nikias and Shao 1995) is used to calculate the γ parameter, which is given by

$$\text{GSNR} = 10\log_{10}\left(\frac{P_S}{\gamma}\right) \tag{24}$$

where P_S is the power of the clean signal and $\gamma = \nu^{\alpha}$, measuring the dispersion from the noise. Then, in the second approach, the impulsiveness of the noise is controlled by the α parameter while the general strength is selected by the GSNR value.

The third and last approach simulates an intermittent connection resulted from a faulty wiring, which could bring the value from a sensor to 0 from time to time. Given L samples from a signal, this method would make a percentage of this signal to 0. All positions could be selected with equal probability.

6 Experimental Results

In this section, three experiments are discussed to demonstrate the performance of the SCMP algorithm. The first is an experiment with a numerical system described in Billings (2013). The second is an experiment using data from a benchmark (Wigren and Schoukens 2013). The third and last experiment consists in using a real plant (Quanser 2011b) to evaluate the proposed method

In order to analyze the quality of the obtained models from each method, the root-mean-square error (RMSE) is computed for a validation set as follows (Kulikova 2017):

$$\text{RMSE} = \sqrt{\frac{1}{MN}\sum_{j=1}^{M}\sum_{i=1}^{N}\left(y(i)^j - \hat{y}(i)^j\right)^2} \tag{25}$$

The experiment was repeated $M = 400$ times (where M is the number of Monte Carlo trials) with a number of observa-

tions (or discrete points) $N = 500$. The error was calculated using the actual output y and the estimated output $\hat{y}$.

6.1 Experiment 1

To illustrate the performance of the proposed SCMP algorithm, consider the following numerical system described by Eq. (26), which was proposed by Billings (2013)

$$y(k) = 0.605y(k-1) - 0.163y(k-2)^2 + 0.588u(k-1) - 0.240u(k-2) + \varepsilon \quad (26)$$

where ε is a additive noise (bimodal or α-stable) used for this experiment.

In this experiment, the matrix of candidate regressors has the following characteristics: output delay $n_y = 3$, input delay $n_u = 3$; and nonlinearity degree $l = 3$.

The input used to estimate the parameters is a sinusoidal signal described by Eq. (27). The input used to validate the data through a free simulation is described by Eq. (28).

$$u(t) = 0.5\sin(0.7t) + 0.25\sin(1.4t) + 0.5\sin(0.35t) \quad (27)$$

$$u_{\text{val}}(t) = 0.25\sin(0.7t) + 0.1\sin(1.4t) + 0.5\sin(0.35t) + 0.2\sin(t) + 0.4\sin(0.5t) \quad (28)$$

This sinusoid has a frequency approximately equal to the cutoff frequency of the system added to a sinusoid with the frequency equal to the half cutoff frequency plus one sinusoid with a frequency equal to twice the frequency of court.

The stopping criterion in all algorithms occurs when the improvement drops below a threshold value ρ. In FROLS algorithm, $\rho = 0.5$. In SEMP and SCMP, $\rho = 5e^{-4}$. Besides, in the results for this experiment, SCMP employs the step size $\mu = 0.1$ from the gradient ascendant method and $\sigma_s = 0.1$ as free parameters.

To simulate the presence of additive noises with non-Gaussian distributions, two configurations described in Sect. 5 were used. The first one is described by Eq. (21) with parameters $\Upsilon(\varrho = 0.1, \mu_1 = 0, \sigma_1 = 1, \mu_2 = 2, \sigma_2 = 1)$. Figure 2 shows a typical validation using this kind of noise, while Fig. 3 shows the results using an α-stable distribution with $\alpha = 1$ and GSNR = 5 dB to model the noise. In all cases, it is possible to state that SCMP presents better fitting than other methods.

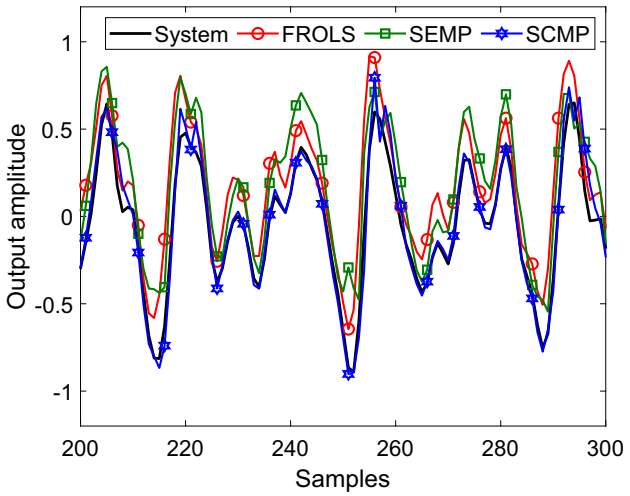

Fig. 2 Experiment 1: Typical simulation from the numerical system of Eq. 26 with bimodal noise following the parameters $\Upsilon(\varrho = 0.1, \mu_1 = 0, \sigma_1 = 0.1, \mu_2 = 2, \sigma_2 = 0.1)$

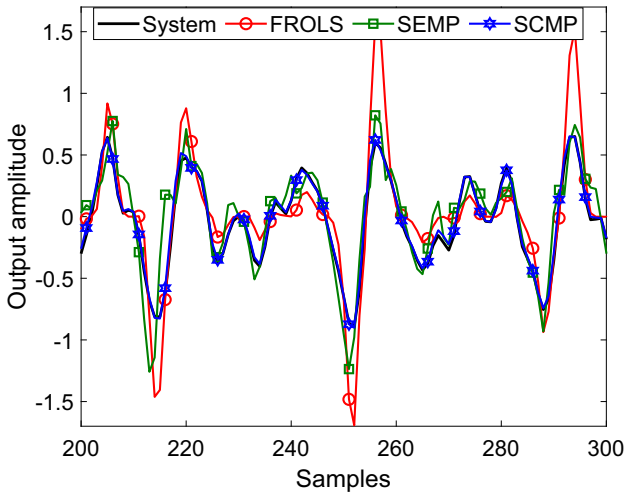

Fig. 3 Experiment 1: Typical simulation from the numerical system of Eq. 26 with α-stable noise with GSNR = 5 dB and $\alpha = 1$

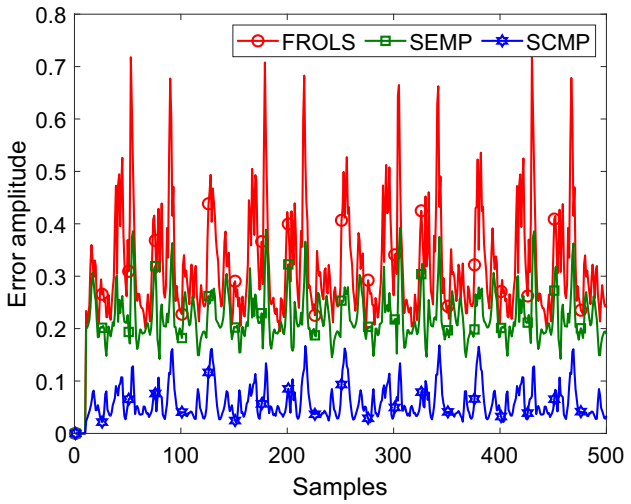

Fig. 4 Experiment 1: Mean from 400 trials of the error between the system output and algorithms for in the presence of bimodal noise with parameters $\Upsilon(\varrho = 0.1, \mu_1 = 0, \sigma_1 = 1, \mu_2 = 2, \sigma_2 = 1)$ in numerical experiment of Eq. 26

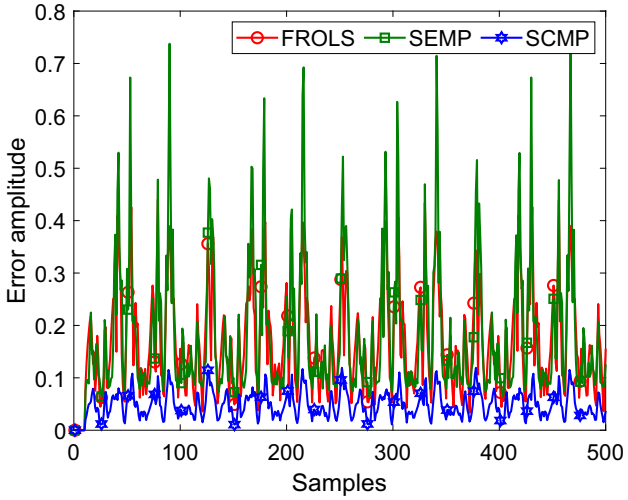

Fig. 5 Experiment 1: Mean from 400 trials of the error between the system output and algorithms in the presence of α-stable noise with GSNR = 5 dB and $\alpha = 1$ in numerical experiment of Eq. 26

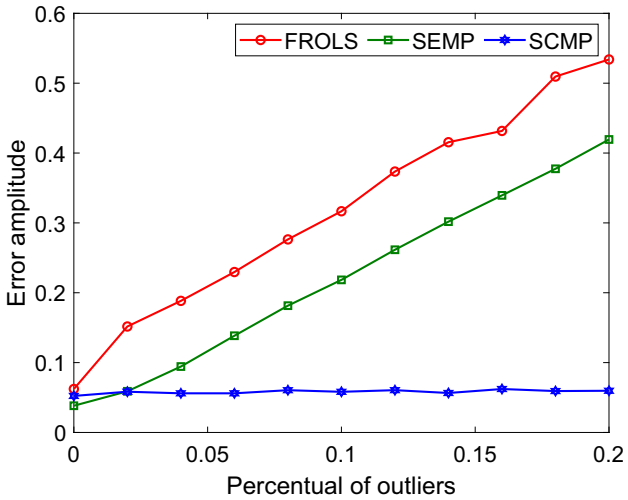

Fig. 6 Experiment 1: Mean from 400 trials of the mean error in the presence of a growing outlier percentage in numerical experiment of Eq. 26

Using this same noise configuration, an average from 400 experiments were made to produce Figs. 4 and 5 . As expected, using correntropy makes the SCMP more accurate in the presence of non-Gaussian noise when compared to FROLS and SEMP.

This SCMP outlier rejection capability is highlighted in Fig. 6. The average error from 400 trials is presented together with the percentage of outliers, ϱ, which varies from 0 to 0.2:

As the number of outliers increases, the performance of both the FROLS and the SEMP algorithms deteriorates since they are based in second-order statistics. The proposed SCMP algorithm was able to keep the error amplitude stable during all the ϱ tested range.

Table 1 summarizes the comparison between SCMP, FROLS, and SEMP. It is possible to notice that SCMP has superior performance in experiment 1. The SCMP maintains the structure selection capability, even in the presence of non-Gaussian noises, as can be observed by the reduced number of parameters and by the lower RMSE values for all cases where the kernel is well adjusted. Also in Table 1, it is possible to observe the standard deviation of the MSE from the validation data and the average number of model parameters obtained by each method.

Regarding the presence of noise with non-Gaussian distributions, the addition of the correntropy as a cost function, both for parameter estimation and for the selection of the regressors, adds two free parameters, the kernel width of the estimator and the cost function, which needs to be adjusted. Simulations were performed in order to demonstrate the problem of a poor fit of such parameter with variations of the kernel values, and the results are given in Table 1. In case with σ_e being too small ($\sigma_e = 0.001$), the algorithm tends to diverge from a solution. On the other hand, using a large kernel width for parameter estimation ($\sigma_e = 5$), as point out by Eq. (7), makes correntropy tend to correlation, which makes the algorithm have similar performance than the second-order methods. By proper tuning the kernel width ($\sigma_e = 0.1$), the SCMP was able to achieve better results. Since this is not always a simple task, this paper implements the VKW adaptive kernel width strategy, which was described by Sect. 3.1.1.

Table 1 Experiment 1: RMSE of validation, standard deviation (Std) of the MSE of the validation data and number of terms (n_θ) used in the numerical experiment of Eq. 26

Method	Noise distribution								
	Gaussian			Bimodal			α-stable		
	RMSE	Std	n_θ	RMSE	Std	n_θ	RMSE	Std	n_θ
FROLS	0.0850	0.0022	2	0.4106	0.1219	84	0.2440	0.0323	52
SEMP	0.0545	0.0016	4	0.2484	0.0206	7	0.4770	0.7454	9
SCMP ($\sigma_e = 0.001$)	0.3602	2.4e−04	10	0.3602	2.4e−04	11	0.3602	2e−04	7
SCMP ($\sigma_e = 0.1$)	0.1090	0.0035	25	0.1129	0.0042	19	0.0685	0.0039	18
SCMP ($\sigma_e = 5$)	0.1314	0.0014	1	0.1526	0.0126	1	0.1352	0.0039	1
SCMP (VKW)	0.0798	0.0154	8	0.0908	0.0147	9	0.1023	0.0292	9

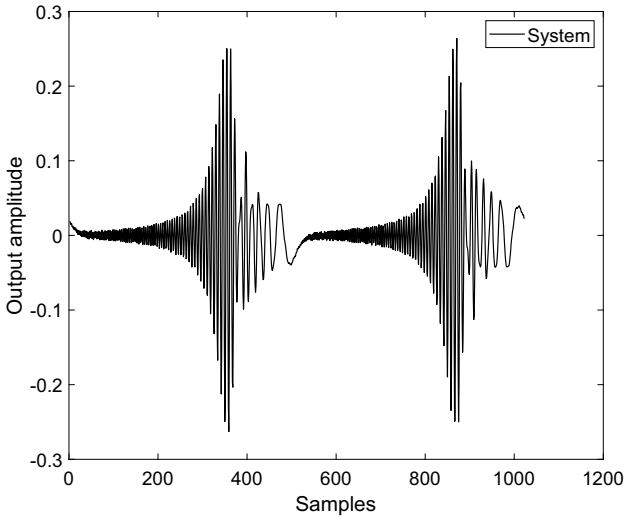

Fig. 7 Experiment 2: Typical simulation from the Silver Box benchmark

Table 2 Experiment 2: RMSE of validation, standard deviation (Std) of the MSE of the validation data and number of terms (n_θ) used in the Benchmark Silver Box experiment with Gaussian noise

Method	Noise distribution		
	Gaussian		
	RMSE	Std	n_θ
FROLS	1.7377e−4	0.0113	84
SEMP	9.4502e−7	6.6261e−04	33
SCMP(VKW)	9.6388e−4	0.0271	22

The SCMP using the VKW strategy has achieved similar performance to the best kernel width selection overcoming this free parameter issue.

6.2 Experiment 2

The second case study and benchmark considered in this paper concern an electronic implementation of a nonlinear system, denoted as Silver Box (Wigren and Schoukens 2013). This system simulates a second-order mass–spring–damper mechanical system, with a nonlinear spring constant with the purpose of relating the displacement $y(t)$ to the force $u(t)$.

The matrix of candidate regressors Ψ_M has an output delay $n_y = 3$, input delay $n_u = 3$, and nonlinearity degree $l = 3$. In all algorithms, the threshold stopping criterion is $\rho = 10^{-8}$. The SCMP parameters for this experiment are $\mu = 1$ and $\sigma_s = 1$.

Figure 7 shows the output of the system and the outputs of the models obtained by the methods covered in this paper. The MSE obtained by each method is described in Table 2.

Table 2 shows that SCMP performance is similar to FROLS and SEMP in this benchmark case.

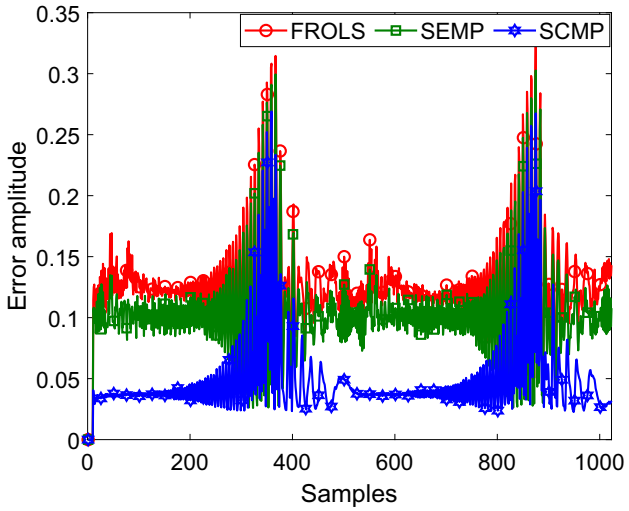

Fig. 8 Experiment 2: Mean from 100 trials of the error between the system output and algorithms in the presence of bimodal noise with parameters $\Upsilon(\varrho = 0.1, \mu_1 = 0, \sigma_1 = 0.05, \mu_2 = 1, \sigma_2 = 0.05)$ for Silver Box benchmark experiment

Table 3 Experiment 2: RMSE of validation, standard deviation (Std) of the MSE of the validation data and number of terms (n_θ) used in the Silver Box benchmark with bimodal noise

Method	Noise distribution		
	Bimodal		
	RMSE	Std	n_θ
FROLS	0.1474	0.0199	84
SEMP	0.1172	8.88e−04	45
SCMP(VKW)	0.0683	0.0023	17

Figure 8 shows the average error signal from 100 experiments with a simulated bimodal noise with parameters $\Upsilon(\varrho = 0.1, \mu_1 = 0, \sigma_1 = 0.05, \mu_2 = 1, \sigma_2 = 0.05)$. As expected, correntropy makes the SCMP more accurate in the presence of non-Gaussian noise when compared to FROLS and SEMP. Table 3 shows the results with this non-Gaussian noise presence.

6.3 Experiment 3

This experiment consists in a real plant, the Quanser Servor Base Unit (Quanser 2011b). The input signal is the voltage applied to the system, and the output is the position of the motor.

The matrix of candidate regressors Ψ_M has an output delay $n_y = 2$, input delay $n_u = 2$ and nonlinearity degree $l = 2$. The FROLS threshold stopping criterion is 10^{-2}, and for SEMP and SCMP it is 10^{-4}. The SCMP parameters in this experiment are $\mu = 0.2$ and $\sigma_s = 1.0$.

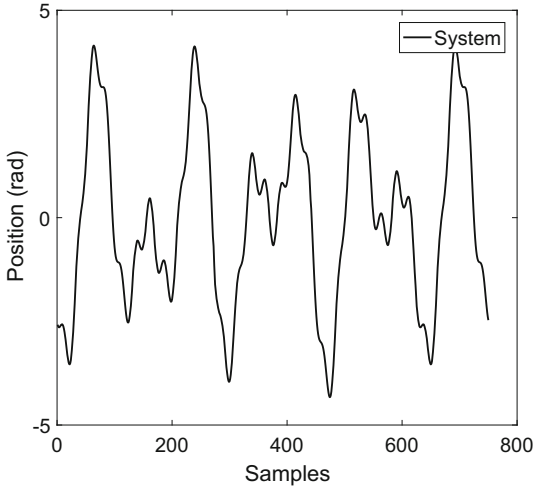

Fig. 9 Experiment 3: Typical simulation from the Quaser Servo system

Table 4 Experiment 3: RMSE of validation, standard deviation (Std) of the MSE of the validation data and number of terms (n_θ) used in Quanser Servor experiment with Gaussian noise

Method	Noise distribution		
	Gaussian		
	RMSE	Std	n_θ
FROLS	7.4112e−4	0.0039	2
SEMP	4.4723e−4	0.0039	2
SCMP(VKW)	2.6784e−2	0.0066	2

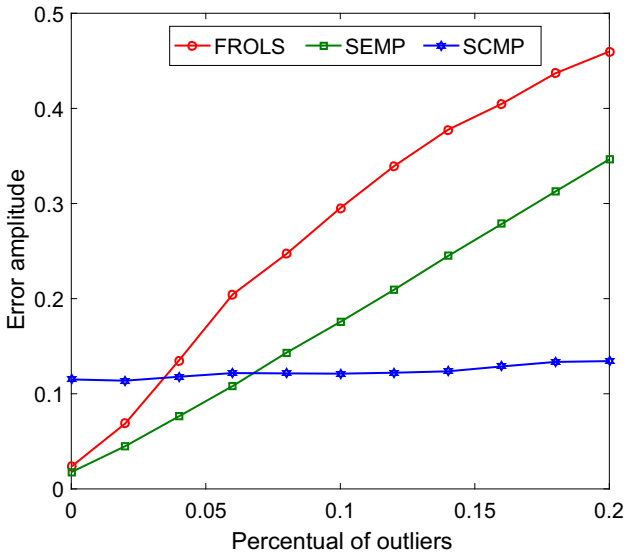

Fig. 10 Experiment 3: Mean from 400 trials of the mean error in the presence of a growing outlier percentage in real case Quanser Servo

Figure 9 illustrates an output of the measured data from the system. The MSE obtained by each method is described in Table 4.

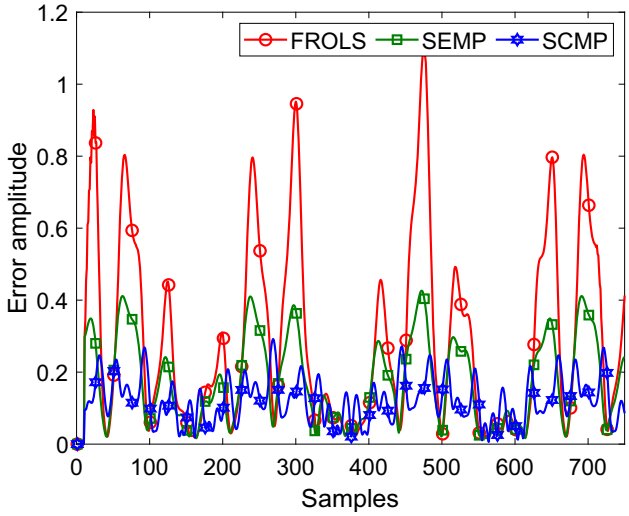

Fig. 11 Experiment 3: Mean from 100 trials of the error between the system output and algorithms in the presence of bimodal noise with 10% of outliers resulted from a faulty wiring for Quanser Servo Base Unit

Table 5 Experiment 3: RMSE of validation, standard deviation (Std) of the MSE of the validation data and number of terms (n_θ) used Quanser Servo Base Unit experiment

Method	Noise distribution		
	Bimodal		
	RMSE	Std	n_θ
FROLS	0.4870	0.2776	15
SEMP	0.2217	0.0128	5
SCMP(VKW)	0.1524	0.0112	2

In this experiment, as described in Sect. 5, the non-Gaussian noise can be interpreted as an intermittent connection resulted from a faulty wiring, which brings the sensor value to 0. With no outliers, FROLS and SEMP had a better performance than SCMP proposed method, as can be seen in Table 4. But, the increment of outliers percentage, ϱ, in the output signal makes the FROLS and SEMP methods lose its capabilities of identify the system while the SCMP maintains its identification capacity. Figure 10 shows this behavior.

Figure 11 shows the average error signal from 400 experiments of this system. Table 5 shows the results of this experiment.

In the presence of Gaussian noise, all the methods used found linear models with few parameters. This makes sense, because the system, according to Quanser (2011a), can be represented by a linear second-order model. However, in the presence of non-Gaussian noises, only SCMP presented models with few parameters, which shows its natural ability to reject outliers.

7 Conclusion

This paper has presented a novel algorithm for nonlinear system identification called simulation correntropy maximization with pruning-SCMP, which uses correntropy as a cost function to both select the structure and estimate the parameters of NARX models in non-Gaussian noise environment.

The proposed method was able to achieve better performance when compared to other methods such as FROLS and SEMP in three different system identification tasks.

One of the drawbacks of using correntropy and, consequently, the SCMP, is a free parameter called kernel width that influences the convergence rate and robustness of the proposed method. This paper addresses this issue by employing a variable kernel width strategy denominated VKW, which was able to achieve good results.

Future work includes the application of the proposed algorithm to NARX MIMO (multiple-input multiple-output) models. Moreover, the presented algorithm may be extended to other nonlinear model representations.

Acknowledgements This work was financed by CAPES—Brazilian Federal Agency for Support and Evaluation of Graduate Education within the Ministry of Education of Brazil.

References

Aguirre, L. A., & Billings, S. A. (1995). Dynamical effects of overparametrization in nonlinear models. *Physica D: Nonlinear Phenomena*, *80*(1), 26–40. https://doi.org/10.1016/0167-2789(95)90053-5.

Billings, S. A. (2013). *Nonlinear system identification: NARMAX methods in the time, frequency, and spatio-temporal domains*. London: Wiley.

Billings, S. A., Korenberg, M. J., & Chen, S. (1988). Identification of non-linear output-affine systems using an orthogonal least-squares algorithm. *International Journal of Systems Science*, *19*(8), 1559–1568. https://doi.org/10.1080/00207728808964057.

Chen, S., & Billings, S. A. (1989). Representations of non-linear systems: The narmax model. *International Journal of Control*, *49*(3), 1013–1032. https://doi.org/10.1080/00207178908559683.

Cheng, Y., Wang, L., & Hu, J. (2009). A two-step method for nonlinear polynomial model identification based on evolutionary optimization. In *2009 World congress on nature biologically inspired computing (NaBIC)* (pp. 613–618). https://doi.org/10.1109/NABIC.2009.5393428.

Falsone, A., Piroddi, L., & Prandini, M. (2015). A randomized algorithm for nonlinear model structure selection. *Automatica*, *60*, 227–238. https://doi.org/10.1016/j.automatica.2015.07.023.

Fontes, A. I., Martins, Ad M, Silveira, L. F., & Principe, J. C. (2015). Performance evaluation of the correntropy coefficient in automatic modulation classification. *Expert Systems with Applications*, *42*(1), 1–8.

Fontes, A. I., Rego, J. B., Martins, Ad M, Silveira, L. F., & Príncipe, J. C. (2017). Cyclostationary correntropy: Definition and applications. *Expert Systems with Applications*, *69*, 110–117.

Guimaraes, J. P. F., Fontes, A. I. R., Rego, J. B. A., Silveira, L. F. Q., & Martins, A. M. (2016). Performance evaluation of the maximum correntropy criterion in identification systems. In *2016 IEEE conference on evolving and adaptive intelligent systems (EAIS)* (pp. 110–113). https://doi.org/10.1109/EAIS.2016.7502500.

Huang, F., Zhang, J., & Zhang, S. (2017a). Adaptive filtering under a variable kernel width maximum correntropy criterion. *IEEE Transactions on Circuits and Systems II: Express Briefs*, *64*(10), 1247–1251. https://doi.org/10.1109/TCSII.2017.2671339.

Huang, F., Zhang, J., & Zhang, S. (2017b). Adaptive filtering under a variable kernel width maximum correntropy criterion. *IEEE Transactions on Circuits and Systems II: Express Briefs*, *64*(10), 1247–1251. https://doi.org/10.1109/TCSII.2017.2671339.

Huang, F., Zhang, J., & Zhang, S. (2017c). Nlms algorithm based on a variable parameter cost function robust against impulsive interferences. *IEEE Transactions on Circuits and Systems II: Express Briefs*, *64*(5), 600–604. https://doi.org/10.1109/TCSII.2016.2594069.

Kulikova, M. (2017). Square-root algorithms for maximum correntropy estimation of linear discrete-time systems in presence of non-gaussian noise. *Systems & Control Letters*, *108*, 8–15. https://doi.org/10.1016/j.sysconle.2017.07.016.

Leontaritis, I. J., & Billings, S. A. (1985a). Input–output parametric models for non-linear systems part I: Deterministic non-linear systems. *International Journal of Control*, *41*(2), 303–328. https://doi.org/10.1080/0020718508961129.

Leontaritis, I. J., & Billings, S. A. (1985b). Input-output parametric models for non-linear systems part II: Stochastic non-linear systems. *International Journal of Control*, *41*(2), 329–344. https://doi.org/10.1080/00207185089611130.

Linhares, L. L. S., Fontes, A. I. R., Martins, A. M., Araújo, F. M. U., & Silveira, L. F. Q. (2015). Fuzzy wavelet neural network using a correntropy criterion for nonlinear system identification. *Mathematical Problems in Engineering*, *2015*, 12. https://doi.org/10.1155/2015/678965.

Liu, Y., & Chen, J. (2013). Correntropy-based kernel learning for nonlinear system identification with unknown noise: An industrial case study. In *10th IFAC international symposium on dynamics and control of process systems* (vol. 46, pp. 361–366). https://doi.org/10.3182/20131218-3-IN-2045.00025.

Liu, W., Pokharel, P. P., & Principe, J. C. (2006). Correntropy: A localized similarity measure. In *The 2006 IEEE international joint conference on neural network proceedings* (pp. 4919–4924). https://doi.org/10.1109/IJCNN.2006.247192.

Liu, W., Pokharel, P. P., & Principe, J. C. (2007). Correntropy: Properties and applications in non-Gaussian signal processing. *IEEE Transactions on Signal Processing*, *55*(11), 5286–5298. https://doi.org/10.1109/TSP.2007.896065.

Nikias, C. L., & Shao, M. (1995). *Signal processing with alpha-stable distributions and applications*. New York: Wiley.

Peng, S., Chen, B., Sun, L., Ser, W., & Lin, Z. (2017). Constrained maximum correntropy adaptive filtering. *Signal Processing*, *140*, 116–126. https://doi.org/10.1016/j.sigpro.2017.05.009.

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. https://doi.org/10.1080/00207170310001635419.

Principe, J. C. (2010). *Information theoretic learning: Renyi's entropy and kernel perspectives*. Berlin: Springer.

Quanser. (2011a). SRV02 rotary servo base unit—Instructor workbook. Quanser.

Quanser. (2011b). SRV02 rotary servo base unit—User manual. Quanser.

Santamaria, I., Pokharel, P. P., & Principe, J. C. (2006). Generalized correlation function: Definition, properties, and application

to blind equalization. *IEEE Transactions on Signal Processing*, *54*(6), 2187–2197. https://doi.org/10.1109/TSP.2006.872524.

Shao, M., & Nikias, C. L. (1993). Signal processing with fractional lower order moments: Stable processes and their applications. *Proceedings of the IEEE*, *81*(7), 986–1010. https://doi.org/10.1109/5.231338.

Singh, A. (2010). Cost functions for supervised learning based on a robust similarity metric. Master's thesis, University of Florida, Florida.

Singh, A., Príncipe, J. C. (2010). A closed form recursive solution for maximum correntropy training. In *2010 IEEE international conference on acoustics, speech and signal processing* (pp. 2070–2073). https://doi.org/10.1109/ICASSP.2010.5495055.

Singh, A., & Príncipe, J. C. (2011). Information theoretic learning with adaptive kernels. *Signal Processing*, *91*(2), 203–213.

Wang, W., Zhao, J., Qu, H., Chen, B., & Principe, J. C. (2015a). An adaptive kernel width update method of correntropy for channel estimation. In *2015 IEEE international conference on digital signal processing (DSP)* (pp. 916–920). https://doi.org/10.1109/ICDSP.2015.7252010.

Wang, W., Zhao, J., Qu, H., Chen, B., & Principe, J. C. (2015b). A switch kernel width method of correntropy for channel estimation. In *IEEE international joint conference on neural networks (IJCNN)* (pp. 1–7). https://doi.org/10.1109/IJCNN.2015.7280632.

Weron, A., & Weron, R. (1995). Computer simulation of lévy α-stable variables and processes. In P. Garbaczewski, M. Wolf, & A. Weron (Eds.), *Chaos—The interplay between stochastic and deterministic behaviour* (pp. 379–392). Heidelberg: Springer.

Wigren, T., & Schoukens, J. (2013). Three free data sets for development and benchmarking in nonlinear system identification. In *2013 European control conference (ECC)* (pp. 2933–2938). https://doi.org/10.23919/ECC.2013.6669201.

Yan, J., & Deller, J. R. (2016). Narmax model identification using a set-theoretic evolutionary approach. *Signal Processing*, *123*, 30–41. https://doi.org/10.1016/j.sigpro.2015.12.001.

Zhao, W. X., & Chen, H. F. (2012). Identification of wiener, hammerstein, and narx systems as Markov chains with improved estimates for their nonlinearities. *Systems & Control Letters*, *61*, 1175–1186. https://doi.org/10.1016/j.sysconle.2012.08.008.

Zhao, W., Chen, H. F., Bai, E. W., & Li, K. (2018). Local variable selection of nonlinear nonparametric systems by first order expansion. *Systems & Control Letters*, *111*, 1–8. https://doi.org/10.1016/j.sysconle.2017.10.001.