Received: 30 April 2019 | Revised: 16 January 2020 | Accepted: 26 February 2020

DOI: 10.1002/rnc.4968

SPECIAL ISSUE ARTICLE

WILEY

Identification of switched autoregressive exogenous systems from large noisy datasets

Sarah Hojjatinia[1] | **Constantino M. Lagoa**[1] | **Fabrizio Dabbene**[2]

[1]School of Electrical Engineering and Computer Science, The Pennsylvania State University, University Park, Pennsylvania, USA

[2]CNR-IEIIT, Politecnico di Torino, Torino, Italy

Correspondence
Constantino Lagoa, School of Electrical Engineering and Computer Science, The Pennsylvania State University, University Park, PA.
Email: lagoa@psu.edu

Summary

The article introduces novel methodologies for the identification of coefficients of switching autoregressive moving average with exogenous input systems and switched autoregressive exogenous linear models. We consider cases where system's outputs are contaminated by possibly large values of noise for both cases of measurement noise and process noise. It is assumed that only partial information on the probability distribution of the noise is available. Given input-output data, we aim at identifying switched system coefficients and parameters of the distribution of the noise, which are compatible with the collected data. We demonstrate the efficiency of the proposed approach with several academic examples. The method is shown to be effective in the situations where a large number of measurements is available; cases in which previous approaches based on polynomial or mixed-integer optimization cannot be applied due to very large computational burden.

KEYWORDS

ARMAX, ARX, identification, noisy data, switched systems

1 | INTRODUCTION

The interest in the study of hybrid systems has been persistently growing in the last years, due to their capability of describing real-world processes in which continuous and discrete state dynamics coexist and interact. Besides classical automotive and chemical processes, emerging applications include computer vision, biological systems, and communication networks.

Moreover, hybrid systems can be used to efficiently approximate nonlinear dynamics, with broad application, ranging from civil structures to robotics and systems biology, that entail extracting information from high-volume data streams.[1-3] In the case of high-dimensional data, nonlinear order reduction or low-dimensional sparse representations techniques[4-6] are very effective in handling static data, but most do not exploit dynamical information on the data.

In the literature, several results have been obtained for the analysis and control of hybrid systems, formally characterizing important properties such as stability or reachability and proposing different control designs.[7] In parallel, researchers rapidly realized that first-principle models may be hard to derive especially with the increase of diverse application fields. This sparked interest on the problem of identifying hybrid (and, in particular, switched) models starting from experimental data; see for instance the tutorial article[8] and the survey.[9] In this article, we aim at addressing one important part of this identification problem. Namely, we aim at developing algorithms that effectively identify the switching dynamic

equations that govern the evolution of the continuous states. In other words, we aim at identifying the difference equation of each submodel in the presence of measurement and process noise.

It should be immediately pointed out that this identification problem is not a simple one, since the simultaneous presence of switching gives it a combinatorial nature. The situation becomes further complicated in the presence of noise. In this case, the problem is in general NP-hard. Several approaches have been proposed to address this difficulty.[10] The identification problem is reformulated as a mixed-integer program in Roll et al.[11] These techniques proved to be very effective in situations involving relatively small noise levels or moderate dimensions, but they do not appear to scale well, and their performance deteriorates as the noise level or problem size increase. In addition, the problem of identification of piecewise linear systems, where the "active" linear submodel depends on the value of the state, has been addressed in Ferrari-Trecate et al,[12] Juloski et al,[13] and Saxen et al.[14] The aim is to identify both the submodels and the regions where they are active. However, the problems formulated in these articles are nonconvex and only local optimality of the proposed approaches is proven.

Of particular interest are recent approaches based on convex optimization: in the work by Bako[15] relaxations based on sparsity are proposed, while in Ozay et al,[16] a moment-based approach is developed to identify the switched autoregressive exogenous system, and Hojjatinia et al[17] adapts it toward Markovian jump systems identification. These methods are surely more robust and represent the choice of reference for medium-size problems and medium values of noise and have found applications in several contexts, ranging from segmentation problems arising in computer vision to biomedical systems.

However, the methods still rely on the solution of rather large optimization problems. Even if the convex nature of these problems allows to limit the complexity growth, there are several situations for which their application becomes critical. For instance, identification problems that involve quite high-noise levels and/or large number of measurements.

An enlightening example, which serves as a practical motivation for our developments, arises in healthcare applications: the availability of *activity tracking devices* allows to gather a large amount of information of the physical activity of an individual. Physical activity is a dynamic behavior, which in principle can be modeled as a dynamical system.[18] Moreover, its characteristics may significantly change depending on the time of the day, position, and so on. This motivated the approach of modeling it as a switching system.[19]

In this article, we focus on cases involving a very large number of sample points, possibly affected by large levels of noise. In this situation, polynomial/moments-based approaches become ineffective, and different methodologies need to be devised. The approach we propose builds upon the same premises as Ozay et al[16] and Hojjatinia et al:[20] the starting point is the algebraic procedure due to Ma and Vidal,[21] where it has been shown that for noiseless processes it is possible to identify the different subsystems in a switching system by recurring to a generalized principal component analysis (GPCA). In particular, we infer the parameters of each subsystem from the null space of a matrix $V_n(r)$ constructed from the input-output data r via a nonlinear embedding (the Veronese map).

The approach was extended to the case where process noise is present in Ozay et al,[16] showing how the entries of this matrix depend polynomially on the unknown noise terms. Then, the problem was formulated in an unknown-but-bounded setting, looking for an admissible noise sequence rendering the matrix $V_n(r)$ rank deficient. This problem was then relaxed using polynomial optimization methods.

In this work, we follow the same line of reasoning, but then take a somewhat different route. First, we consider random noise, and we assume that *some* information on the noise is available. Then, instead of relaxing the problem, we exploit the availability of a large number of measurements and its "averaged behavior." This allows us to devise an algorithm characterized by an extremely low complexity in terms of required operations. The ensuing optimization problem involves only the computation of the singular vector associated with the minimum singular value of a matrix that can be efficiently computed and whose size does not depend on the number of measurements.

1.1 | Article organization

In Section 2, previous results on switched system identification when no noise is present are reviewed. Section 3 concentrates on the problem of switched system identification in the presence of measurement noise. The results are extended to the case of process noise in Section 4. Procedures for simultaneous estimation of systems parameters and noise parameters are described in Section 5. Several examples that illustrate the performance of the proposed approach are provided in Section 6. Finally, some concluding remarks are provided in Section 7.

1.2 | Notation

Given a scalar random variable $x \in \mathbb{R}$, we denote by m_d its dth moment $\mathrm{E}[x^d]$, where $\mathrm{E}[\cdot]$ refers to expectation. The moments of x may be computed according to the following integral

$$m_d = \mathrm{E}[x^d] = \int_{-\infty}^{\infty} x^d \, f(x) \, dx, \tag{1}$$

where $f(x)$ is the probability density function of x. In addition, the variance of x is indicated by $\mathrm{Var}(x)$.

When some of the parameters θ of the distribution are not known, we use the notation $f(x|\theta)$ to denote the dependence of the probability density function on these unknown parameters. Throughout this article, we assume that $f(x|\theta)$ is a continuous function of θ. Obviously, this implies that the moments of the random variable are known continuous functions of θ.

For example, if x has a normal distribution with zero mean and we assume that the variance $\theta = \sigma^2$ is not known then we have

$$f(x|\theta) = \frac{1}{\sqrt{2\pi\theta}} e^{-x^2/2\theta}.$$

The moments of x as a function of θ are given by

$$m_d = E[x^d] = \begin{cases} 0 & \text{if } d \text{ is odd} \\ \theta^{d/2} \, (d-1)!! & \text{if } d \text{ is even} \end{cases} \tag{2}$$

where !! denotes double factorial ($n!!$ is the product of all numbers from n to 1 that have the same parity as n).

2 | NOISELESS SWITCHED SYSTEM IDENTIFICATION: A REVIEW

As a motivation for the approach presented in this article, we review and slightly reformulate earlier results on an algebraic approach to the switched system identification. We refer the reader to Vidal et al[22] for details on this formulation. Consider a switched autoregressive exogenous (SARX) system of the form

$$x_k = \sum_{j=1}^{n_a} a_{j\delta_k} \, x_{k-j} + \sum_{j=1}^{n_b} b_{j\delta_k} \, u_{k-j}, \tag{3}$$

where $x_k \in \mathbb{R}$ and $u_k \in \mathbb{R}$ are the output and input at time k, respectively. The variable $\delta_k \in \{1, \dots, n\}$ denotes the subsystem active at time k, where n is the total number of subsystems. Furthermore, $a_{j\delta_k}$ and $b_{j\delta_k}$ denote unknown coefficients corresponding to mode δ_k. Assume that the values of u_k, $k = -n_b + 1, \dots, N-1$ and x_k, $k = -n_a + 1, \dots, N$ are available.

As a first step toward an identification algorithm, we start by noting that Equation (3) can be written in compact form as

$$\mathbf{t}_{\delta_k}^{\top} \, \mathbf{r}_k = 0, \tag{4}$$

where we introduced the (known) regressor vector at time k

$$\mathbf{r}_k = \left[x_k, \; x_{k-1}, \dots, x_{k-n_a}, \; u_{k-1}, \dots, u_{k-n_b} \right]^{\top}$$

and the vector of (unknown) coefficients at time k

$$\mathbf{t}_{\delta_k} = \left[-1, \; a_{1\delta_k}, \dots, a_{n_a\delta_k}, \; b_{1\delta_k}, \dots, b_{n_b\delta_k} \right]^{\top}.$$

Hence, independently of which of the n submodels is active at time k, we have that the following equality should hold

$$p_n(\mathbf{r}_k) = \prod_{i=1}^{n} \mathbf{t}_i^\top \mathbf{r}_k = \nu_n(\mathbf{r}_k)^\top \mathbf{c}_n = 0, \tag{5}$$

where the vector of parameters corresponding to the ith submodel is denoted by $\mathbf{t}_i \in \mathbb{R}^{n_a+n_b+1}$, $\nu_n(\cdot)$ is Veronese map of degree n,[23] and $\mathbf{c}_n$ is a vector whose entries are polynomial functions of unknown parameters $\mathbf{t}_i$ (see Reference 24 for explicit definition).

The Veronese map, also known as polynomial embedding in machine learning, contains all monomials of order n in lexicographical order. That is, given a vector $x \in \mathbb{R}^s$ and $n > 0$, we have

$$\nu_n(x) = \begin{bmatrix} \vdots \\ x_1^{\alpha_1} x_2^{\alpha_2} \dots x_s^{\alpha_s} \\ \vdots \end{bmatrix}, \quad \sum_{i=1}^{s} \alpha_i = n, \alpha_i \geq 0,$$

and $\nu_n(x) \in \mathbb{R}^\ell$, with $\ell = \binom{n+s}{n}$. Equation (5) holds for all k, and these equalities can be expressed in matrix form as follows

$$\mathbf{V}_n(\mathbf{r})\ \mathbf{c}_n = \left[\nu_n(\mathbf{r}_1)^\top, \dots, \nu_n(\mathbf{r}_N)^\top\right]^\top \mathbf{c}_n = 0 \tag{6}$$

where $\mathbf{r}$, without subscript, denotes the set of all regressor vectors. Clearly, we are able to identify $\mathbf{c}_n$ (and hence, under general conditions, the system's parameters[24]) if and only if $\mathbf{V}_n(\mathbf{r})$ is rank deficient. In that case, the vector $\mathbf{c}_n$ can be found by computing the null space of $\mathbf{V}_n(\mathbf{r})$. To better clarify this procedure and fix the notation, we illustrate it in the following simple example.

Example 1. Consider a system of order 1 ($n_a = n_b = 1$) which switches between two different subsystems ($n = 2$) , that is,

$$\begin{aligned} &\text{subsystem 1}: \ x_k = a_1\, x_{k-1} + b_1\, u_{k-1} \\ &\text{subsystem 2}: \ x_k = a_2\, x_{k-1} + b_2\, u_{k-1}. \end{aligned} \tag{7}$$

We can rewrite the system as in Equation (4). The regressor vector $\mathbf{r}_k$ at time k

$$\mathbf{r}_k = \left[x_k\ x_{k-1}\ u_{k-1}\right]^\top,$$

gives rise to the following Veronese vector

$$\nu_n(\mathbf{r}_k) = \begin{bmatrix} x_k^2 \\ x_k\ x_{k-1} \\ x_k\ u_{k-1} \\ x_{k-1}^2 \\ x_{k-1}\ u_{k-1} \\ u_{k-1}^2 \end{bmatrix}, \tag{8}$$

whose length is $\binom{n+n_a+n_b}{n} = \binom{2+1+1}{2} = 6$. The corresponding coefficient vector $\mathbf{c}_2$ assumes the form

$$\mathbf{c}_2 = [1, -(a_1 + a_2), -(b_1 + b_2), a_1a_2, a_1b_2 + b_1a_2, b_1b_2]^\top$$

and its components can be observed to be polynomial functions of the parameters of the subsystems.

2.1 | A reformulation of the hybrid decoupling constraint

Note that the number of rows of the Veronese matrix $\mathbf{V}_n$ is equal to the number of measurements available for the regressor, that is, in the notation of our article, the number of rows is N. Therefore, very large datasets (large N) lead to computational problems that are ill conditioned or even impossible to solve. Hence, in this article, we work with an equivalent condition that is more suitable for the problem of SARX system identification from very large datasets. We now elaborate on this.

As previously mentioned, in the absence of noise, the SARX system identification is equivalent to finding a vector $\mathbf{c}_n$ satisfying

$$\mathbf{c}_n^\top \nu_n(\mathbf{r}_k) = 0 \quad \text{for all} \;\; k = 1, 2, \dots N.$$

We observe that this is in turn equivalent to finding $\mathbf{c}_n$ so that

$$\frac{1}{N}\sum_{k=1}^{N} \mathbf{c}_n^\top \nu_n(\mathbf{r}_k)\nu_n^\top(\mathbf{r}_k)\, \mathbf{c}_n = 0.$$

As a result, for the noiseless case, identifying the coefficients of the submodels of switched system is equivalent to finding the singular vector $\mathbf{c}_n$ associated with the minimum singular value of the matrix

$$\mathcal{M}_N = \frac{1}{N}\sum_{k=1}^{N} \nu_n(\mathbf{r}_k)\; \nu_n^\top(\mathbf{r}_k) \doteq \frac{1}{N}\sum_{k=1}^{N} \mathbf{M}_k. \tag{9}$$

Note that, by using this equivalent condition, we only need to consider square matrices of size $\binom{n+n_a+n_b}{n}$. In other words, the size of this matrix *does not depend* on the number of measurements. This is especially important when considering very large datasets.

Example 2. To illustrate the notation introduced, we revisit Example 1: in this case the matrix $\mathbf{M}_k$ has the form

$$\mathbf{M}_k = \nu_n(\mathbf{r}_k)\; \nu_n^\top(\mathbf{r}_k) = \begin{pmatrix} x_k^4 & x_k^3\, x_{k-1} & x_k^3\, u_{k-1} & x_k^2\, x_{k-1}^2 & x_k^2\, x_{k-1}\, u_{k-1} & x_k^2\, u_{k-1}^2 \\ * & x_k^2\, x_{k-1}^2 & x_k^2\, x_{k-1}\, u_{k-1} & x_k\, x_{k-1}^3 & x_k\, x_{k-1}^2\, u_{k-1} & x_k\, x_{k-1}\, u_{k-1}^2 \\ * & * & x_k^2\, u_{k-1}^2 & x_k\, x_{k-1}^2\, u_{k-1} & x_k\, x_{k-1}\, u_{k-1}^2 & x_k\, u_{k-1}^3 \\ * & * & * & x_{k-1}^4 & x_{k-1}^3\, u_{k-1} & x_{k-1}^2\, u_{k-1}^2 \\ * & * & * & * & x_{k-1}^2\, u_{k-1}^2 & x_{k-1}\, u_{k-1}^3 \\ * & * & * & * & * & u_{k-1}^4 \end{pmatrix} \tag{10}$$

where the $*$s are to avoid double writing of the entries and refer to the fact that $\mathbf{M}_k$ is a symmetric matrix, $\mathbf{M}_k^T = \mathbf{M}_k$. Matrix $\mathcal{M}_N$ is just the time average of $\mathbf{M}_k$ above.

3 | IDENTIFICATION OF SWITCHED AUTOREGRESSIVE MOVING AVERAGE WITH EXOGENOUS INPUT SYSTEMS

In this section, we address the problem of identification of switching autoregressive moving average with exogenous input (SARMAX) systems. More precisely, we consider SARMAX systems of the form

$$x_k = \sum_{j=1}^{n_a} a_{j\delta_k}\, x_{k-j} + \sum_{j=1}^{n_b} b_{j\delta_k}\, u_{k-j}, \tag{11}$$

$$y_k = x_k + \eta_k, \tag{12}$$

where y_k is the observed output, which is assumed to be contaminated by (possibly large) noise η_k. As before, $x_k \in \mathbb{R}$ is the noiseless system output at time k and $u_k \in \mathbb{R}$ is input at time k. Moreover, the variable $\delta_k \in \{1, \dots, n\}$ denotes the subsystem active at time k, where n is the total number of subsystems.

As a first step in the development of the proposed identification procedure, the following assumptions are made on the SARMAX system model and measurement noise.

Assumption 1. Throughout this article for SARMAX system identification, it is assumed that:

(a) Model orders n_a and n_b are available.
(b) The number of subsystems n is available, and each subsystem is "visited" infinitely often. More precisely, let $N_i(N)$ be the number of "visits" of subsystem i up until time N. Then, for all $i = 1, 2, \dots, n$

$$\lim_{N\to\infty} \frac{N_i(N)}{N} > 0.$$

(c) Noise η_k is independent from η_l for $k \neq l$ and identically distributed with probability density $f(\eta|\theta)$; where θ is a (low dimensional) vector of unknown parameters
(d) Moments of noise m_d (up to order $d = 4n$) are bounded.
(e) Input sequence u_k applied to the system is known and bounded, that is, there exists a L_u such that $|u(k)| \leq L_u$ for all k.
(f) There exists a finite constant L_x so that $|x_k| \leq L_x$ for all k.

We now provide a few comments on the assumptions made above. Assumption 1.a can be relaxed to assume only knowledge of upper bounds on n_a and n_b. In this case, on top of the approach proposed, a search over the allowable values of n_a and n_b is needed to determine the values that better fit the data collected.

In the proposed procedure we rely on the use of estimates of the matrix $\mathcal{M}_N$ described in (9) to determine the coefficients of the subsystems. In the case of large N, to be able to identify all subsystems we need Assumption 1.b so that each subsystem has a "significant impact" in the construction of $\mathcal{M}_N$. Indeed, if the condition is not satisfied for some subsystem i, then $\mathcal{M}_N$ will not depend on it for large values of N.

In Assumption 1.c, we allow for incomplete knowledge of the measurement noise. More precisely, we assume that the overall "form" of the noise is known but some of its parameters will be estimated from the data. An example of this is zero mean iid Gaussian noise where the variance is not known and needs to be estimated together with the parameters of the subsystems.

Finally, Assumptions 1.d-f, are related to "stability" of the system and are needed to enforce boundedness of mean and variance of the quantities used to estimate the parameters of the subsystems and the parameters of the noise.

3.1 | Problem statement and preliminary results

To simplify the exposition to follow, we start discussing the case when the parameters θ of the noise distribution are known and, hence, we can compute its moments. The more general case, where joint estimation of the parameters of the distribution of the noise is needed, is addressed in Section 5.

We start with the definition of the problem that we want to solve and provide some preliminary results that will allow us to develop efficient algorithms for estimation of the coefficients of the subsystems. Consider the following problem:

Problem 1. Given Assumption 1, an input sequence u_k, $k = -n_b + 1, \dots, N-1$ and noisy output measurements y_k, $k = -n_a + 1, \dots, N$, determine coefficients of the SARMAX model $a_{i,j}$, $i = 1, 2, \dots, n_a$, $j = 1, 2, \dots, n$, $b_{i,j}$, $i = 1, 2, \dots, n_b$, $j = 1, 2, \dots, n$.

As we have seen when discussing the noiseless case, the switched autoregressive exogenous system identification problem is equivalent to finding a vector in the null space of the matrix $\mathcal{M}_N$ defined in (9). Under mild conditions, the null space of this matrix has dimension one if and only if the data are compatible with the assumed model. However, if noise is present, x_k is not known and, therefore, $\mathcal{M}_N$ cannot be computed. In the remainder of this section, we make use of the available measurements as well as the a priori information on the statistics of the noise to compute approximations of the matrix $\mathcal{M}_N$ and, consequently, approximations of vectors in its null space. Let us start by establishing some properties of the entries of this matrix.

On the powers of x_k: Since we do not have access to the values of the output x_k to estimate the values of the quantities in Equation (9), we need to relate the powers of x_k to the measurements and available information of the noise, that is, its moments. Note that x_k is a (unknown) deterministic quantity. Therefore, for any integer h,

$$x_k^h = E[x_k^h]. \tag{13}$$

Since $x_k = y_k - \eta_k$ we have

$$x_k^h = E[x_k^h] = E[(y_k - \eta_k)^h]. \quad \forall k = 1, 2, \dots, N. \tag{14}$$

Assume now, for simplicity, the distribution of the noise is symmetric with respect to the origin. As a result, all odd moments are zero (in particular, the noise is zero mean, ie, $m_1 = 0$). We remark that this assumption is made only to simplify the calculations below, and that the approach can be extended to the nonsymmetric case.

We concentrate on computing the expected value of powers of x_k recursively and in a closed form. First, we give an example of how to compute the expected value of powers of x_k for powers $h = 1, 2$. For $h = 1$, we have

$$x_k = E[x_k] = E[y_k - \eta_k] = E[y_k] - E[\eta_k] = E[y_k] - m_1 = E[y_k], \tag{15}$$

while, for $h = 2$, we can write

$$x_k^2 = E[x_k^2] = E[(y_k - \eta_k)^2] = E[y_k^2] - 2E[y_k\eta_k] + E[\eta_k^2] = E[y_k^2] - 2E[y_k\eta_k] + m_2. \tag{16}$$

We remark again that the second moment of noise $E[\eta_k^2] = m_2$ is assumed to be known. To estimate the value of $E[y_k\eta_k]$, consider the following

$$E[y_k\eta_k] = E[(x_k + \eta_k)\eta_k] = E[x_k\eta_k] + E[\eta_k^2]. \tag{17}$$

The quantities x_k and η_k are mutually independent and, therefore, $E[x_k\eta_k] = E[x_k]E[\eta_k]$, with $E[\eta_k] = m_1 = 0$. As a consequence, we have

$$E[y_k\eta_k] = E[\eta_k^2], \tag{18}$$

and finally the value of Equation (16) is

$$\begin{aligned} E[x_k^2] &= E[y_k^2] - 2E[\eta_k^2] + E[\eta_k^2] = E[y_k^2] - E[\eta_k^2] \\ &= E[y_k^2] - m_2. \end{aligned} \tag{19}$$

The reasoning above can be generalized to any power of x_k. More precisely, we have the following result, whose proof is an immediate consequence of the derivations so far.

Lemma 1. *The expected value of the powers of x_k satisfies*

$$\begin{aligned} E[x_k^h] = E[(y_k - \eta_k)^h] &= E[y_k^h] - \sum_{d=1}^{h} \binom{h}{d} E[x_k^{h-d}]\, E[\eta_k^d] \\ &= E[y_k^h] - \sum_{d=1}^{h} \binom{h}{d} E[x_k^{h-d}]\, m_d \quad k = 1, 2, \dots, N. \end{aligned} \tag{20}$$

The result above provides a systematic way of relating the matrix $\mathbf{M}_k$ to the statistical properties of the measured output y_k and of the noise η_k. This relationship will be exploited later on to estimate $\mathcal{M}_N$ from data.

Example 3 (Construction of $\mathbf{M}_k$). To illustrate the use of the concepts above, we revisit again the example used in previous sections. Recall that, for this example, the matrix $\mathbf{M}_k$ has the form provided in Equation (10). Now, we can compute expected value of powers of x_k in terms of expected value of powers of y_k and moments of measurement noise.

More precisely, using Lemma 1, we obtain an equivalent expression for the matrix $\mathbf{M}_k$ in (10), which is provided below.

$$\mathbf{M}_k = \begin{pmatrix} E[y_k^4] - 6\, m_2\, (E[y_k^2] - m_2) - m_4 & (E[y_k^3] - 3\, m_2\, E[y_k])\, E[y_{k-1}] & (E[y_k^3] - 3\, m_2\, E[y_k])\, u_{k-1} & \\ * & (E[y_k^2] - m_2)\, (E[y_{k-1}^2] - m_2) & (E[y_k^2] - m_2)\, E[y_{k-1}]\, u_{k-1} & \\ * & * & (E[y_k^2] - m_2)\, u_{k-1}^2 & \cdots \\ * & * & * & \\ * & * & * & \\ * & * & * & \end{pmatrix}$$

$$\begin{pmatrix} (E[y_k^2] - m_2)\, (E[y_{k-1}^2] - m_2) & (E[y_k^2] - m_2)\, E[y_{k-1}]\, u_{k-1} & (E[y_k^2] - m_2)\, u_{k-1}^2 \\ (E[y_{k-1}^3] - 3\, m_2\, E[y_{k-1}])\, E[y_k] & (E[y_{k-1}^2] - m_2)\, E[y_k]\, u_{k-1} & E[y_k]\, E[y_{k-1}]\, u_{k-1}^2 \\ (E[y_{k-1}^2] - m_2)\, E[y_k]\, u_{k-1} & E[y_k]\, E[y_{k-1}]\, u_{k-1}^2 & E[y_k]\, u_{k-1}^3 \\ \cdots\ E[y_{k-1}^4] - 6\, m_2\, (y_{k-1}^2 - m_2) - m_4 & (E[y_{k-1}^3] - 3\, m_2\, E[y_{k-1}])\, u_{k-1} & (E[y_{k-1}^2] - m_2)\, u_{k-1}^2 \\ * & (E[y_{k-1}^2] - m2)\, u_{k-1}^2 & E[y_{k-1}]\, u_{k-1}^3 \\ * & * & u_{k-1}^4 \end{pmatrix} \tag{21}$$

On the structure of $\mathbf{M}_k$*:* We now provide one of the properties of the matrices $\mathbf{M}_k = \nu_n(\mathbf{r}_k)\ \nu_n^\top(\mathbf{r}_k)$ that is central to the results to follow. If we look at the example above, we see that for given moments of the noise, this new representation of $\mathbf{M}_k$ is an affine function of monomials of y_k and u_k. This is a general result which is an immediate consequence of the reasoning described above and the fact that y_k and y_l are independent random variables for $k \neq l$ and u_k is a given deterministic signal.

Lemma 2. *Assume that the noise distribution and the input signal are given and fixed. Let* $\mathrm{mon}_n(\cdot)$ *denote a function that returns a vector with all monomials up to order n of its argument. Then there exists an affine matrix function* $M(\cdot)$ *so that*

$$\begin{aligned} \mathbf{M}_k = \nu_n(\mathbf{r}_k)\ \nu_n^\top(\mathbf{r}_k) &= E\{M[\mathrm{mon}_n(y_k, \dots, y_{k-n_a}, u_{k-1}, \dots, u_{k-n_b})]\} \\ &= M\{E[\mathrm{mon}_n(y_k, \dots, y_{k-n_a}, u_{k-1}, \dots, u_{k-n_b})]\}. \end{aligned}$$

3.2 | SARMAX identification algorithm

As mentioned before, to identify the parameters of the SARMAX system, we need to be able to estimate the matrix $\mathcal{M}_N$ in Equation (9). It turns out that it can be done by using the available noisy measurements. More precisely, we have the following result.

Theorem 1. *Let* $M(\cdot)$ *and* $\mathrm{mon}_n(\cdot)$ *be the functions defined in Lemma 2. Define*

$$\widehat{\mathcal{M}}_N \doteq \frac{1}{N} \sum_{k=1}^{N} M[\mathrm{mon}_n(y_k, \dots, y_{k-n_a}, u_{k-1}, \dots, u_{k-n_b})].$$

Then, as $N \to \infty$*,*

$$\widehat{\mathcal{M}}_N - \mathcal{M}_N \to 0 \text{ a.s.}$$

and, for any (i, j) *entry and for any* $\epsilon > 0$

$$\text{Probability} \left\{ \max_{N \geq J} \left| \widehat{\mathcal{M}}_{N_{(i,j)}} - \mathcal{M}_{N_{(i,j)}} \right| \geq \epsilon \right\} \leq O(1/J).$$

Proof. See Appendix. ∎

As a result, the empirical average computed using the noisy measurements (where expected values of monomials are replaced by the measured monomial values) converges to the desired matrix in Equation (9). Therefore, we propose the following algorithm for identification of a SARMAX system.

Algorithm 1 (SARMAX identification). *Let n_a, n_b, n and moments of the noise be given.*

Step 1. Compute matrix

$$\widehat{\mathcal{M}}_N = \frac{1}{N}\sum_{k=1}^{N} M[\text{mon}_n(y_k, \dots, y_{k-n_a}, u_{k-1}, \dots, u_{k-n_b})].$$

Step 2. Let $\mathbf{c}_n$ be the singular vector associated with the minimum singular value of $\widehat{\mathcal{M}}_N$.
Step 3. Determine the coefficients of the subsystems from the vector $\mathbf{c}_n$.

In order to perform step 3 in Algorithm 1, we adopt polynomial differentiation algorithm for mixtures of hyperplanes, introduced by Vidal[25,pp69-70]. For the sake of completeness, we now review this algorithm.

Algorithm 2 (Polynomial differentiation for mixtures of hyperplanes). *Let the set of regressors $\mathbf{r}$ be given and let $\mathbf{c}_n$ be the vector computed by Algorithm 1.*

Step 1. Define polynomial $p_n(\mathbf{r}_k) = \mathbf{c}_n^\top \nu_n(\mathbf{r}_k)$
Step 2. Let $Dp(\mathbf{r}_k)$ be the gradient of a polynomial p at $\mathbf{r}_k$.
for $i = n : 1$

$$\mathbf{y}_i = \text{argmin}_{\mathbf{r}_k \in \mathbf{r},\ Dp_i(\mathbf{r}_k) \neq 0} \frac{|p_i(\mathbf{r}_k)|}{\|Dp_i(\mathbf{r}_k)\|},$$

$$\mathbf{t}_i = \frac{Dp_i(\mathbf{y}_i)}{\|Dp_i(\mathbf{y}_i)\|},$$

$$p_{i-1}(\mathbf{r}_k) = \frac{p_i(\mathbf{r}_k)}{\mathbf{t}_i^\top \mathbf{r}_k},$$

end
Step 3. Assign point $\mathbf{r}_k$ to subspace S^i if $i = \text{argmin}_{l=1,\dots,n} |\mathbf{t}_l^\top \mathbf{r}_k|$

4 | SWITCHED AUTOREGRESSIVE EXOGENOUS SYSTEM IDENTIFICATION

We now show how the approach developed in the previous section can be adapted to the problem of identification of switched autoregressive exogenous systems. Consider SARX models of the form

$$y_k = \sum_{j=1}^{n_a} a_{j\delta_k}\, y_{k-j} + \sum_{j=1}^{n_b} b_{j\delta_k}\, u_{k-j} + \epsilon_k, \tag{22}$$

where ϵ_k denotes process noise, $y_k \in \mathbb{R}$ is the output at time k, and $u_k \in \mathbb{R}$ is the input at time k. As before, the variable $\delta_k \in \{1, \dots, n\}$ denotes the subsystem active at time k, where n is the total number of subsystems. Furthermore, $a_{j\delta_k}$ and $b_{j\delta_k}$ denote unknown coefficients corresponding to mode δ_k.

The following assumptions are made on the above SARX system model and process noise.

Assumption 2. For SARX system identification, it is assumed that:

1. Model orders n_a and n_b are available.
2. The number of subsystems n is available, and each subsystem is "visited" infinitely often. See precise definition in Assumption 1.

3. Noise ϵ_k is independent from ϵ_l for $k \neq l$, and identically distributed with probability density $f(\epsilon|\theta)$, where θ is a (low dimensional) vector of unknown parameters.
4. Moments of noise m_d (up to order $d = 4n$) are bounded.
5. Input sequence u_k applied to the system is known and bounded.

Again, we assume that the order and number of subsystems are given. If only upper bounds are available, we can search among allowable values and choose the ones better fit the data collected. As for the assumption on the system and noise, these are done do that the quantities used in the identification algorithms have bounded mean and variance.

Once more, for simplicity of exposition, in the reasoning below, we assume that the distribution of the noise is known, so its moments m_d are available. As mentioned before, estimation of the parameters of the distribution of the noise is addressed in Section 5.

We start by noting that Equation (22) is equivalent to

$$\mathbf{t}_{\delta_k}^{\top}\ \mathbf{r}_k = 0, \tag{23}$$

where, for the case of ARX system with process noise, the regressor at time k takes the form

$$\mathbf{r}_k = [y_k - \epsilon_k,\ y_{k-1}, \dots, y_{k-n_a},\ u_{k-1}, \dots, u_{k-n_b}]^{\top},$$

and the vector of unknown coefficients at time k is

$$\mathbf{t}_{\delta_k} = [-1,\ a_{1\delta_k}, \dots, a_{n_a\delta_k},\ b_{1\delta_k}, \dots, b_{n_b\delta_k}]^{\top}.$$

Hence, as before, independently of which of the n submodels is active at time k, we have

$$P_n(\mathbf{r}_k) = \prod_{i=1}^{n} \mathbf{t}_i^{\top}\mathbf{r}_k = \mathbf{c}_n^{\top}\nu_n(\mathbf{r}_k) = 0, \tag{24}$$

where the vector of parameters corresponding to the ith submodel is denoted by $\mathbf{t}_i \in \mathbb{R}^{n_a+n_b+1}$, and $\nu_{n,}(.)$ is the Veronese map of degree n. As before, the number of rows in the Veronese matrix $\mathbf{V}_n$, which consists of all the Veronese maps at time $k = 1, 2, \dots, N$, is equal to N (the number of measurements available for the regressor) and, therefore, a reformulation of the results is needed to be able to address the problem of SARX identification from very large datasets.

The switched ARX system identification is equivalent to finding a vector $\mathbf{c}_n$ satisfying

$$\mathbf{c}_n^{\top}\nu_n(\mathbf{r}_k) = 0 \quad \text{for all}\ \ k = 1, 2, \dots N.$$

This is in turn equivalent to finding a vector c_n so that

$$\frac{1}{N}\sum_{k=1}^{N} \mathbf{c}_n^{\top}\nu_n(\mathbf{r}_k)\nu_n^{\top}(\mathbf{r}_k)\mathbf{c}_n = 0.$$

Consequently, identifying the coefficients of the submodels of switched ARX system is equivalent to finding a singular vector c_n associated with the minimum singular value of the noise-dependent matrix

$$\mathcal{M}_N^{\text{proc}} = \frac{1}{N}\sum_{k=1}^{N} \nu_n(\mathbf{r}_k)\ \nu_n^{\top}(\mathbf{r}_k) \doteq \frac{1}{N}\sum_{k=1}^{N} \mathbf{M}_k^{\text{proc}}. \tag{25}$$

The main difference between the SARX case and the SARMAX discussed in the previous section is the fact that the matrix $\mathcal{M}_N^{\text{proc}}$ is a function of the unmeasurable noise ϵ_k and cannot be directly computed. Therefore, we use available

information on the statistics of the noise to compute approximations of the matrix $\mathcal{M}_N^{\text{proc}}$, and, consequently, approximations of vectors in its null space. As a first step, we now relate the expected value of powers of $y_k - \epsilon_k$ to the noisy output and available information of the noise.

Lemma 3. *Consider output monomials of the form* $e_k = y_{k-1}^{h_1} \ \dots \ y_{k-n_a}^{h_{n_a}}$, *where* $\sum_{i=1}^{n_a} h_i \leq 2n$, *the expected value of the powers of multiplication of* $y_k - \epsilon_k$ *and* e_k *satisfies*

$$\begin{aligned} E[(y_k - \epsilon_k)^h \ e_k] &= E[y_k^h \ e_k] - \sum_{d=1}^{h} \binom{h}{d} \ E[(y_k - \epsilon_k)^{h-d} \ e_k] \ E[\epsilon_k^d] \\ &= E[y_k^h \ e_k] - \sum_{d=1}^{h} \binom{h}{d} \ E[(y_k - \epsilon_k)^{h-d} \ e_k] \ m_d \\ & k = 1, \ 2, \dots, N \quad \text{and} \quad \forall i = 0, 1, \dots, 2n-h. \end{aligned} \tag{26}$$

Again, we can exploit the structure of the matrix $\mathbf{M}_k^{\text{proc}}$ to determine high-fidelity estimates from collected data. We start by emphasizing the following structural result

Lemma 4. *Assume that the noise distribution and the input signal are given and fixed. Again, let* $\text{mon}_n(\cdot)$ *denote a function that returns a vector with all monomials up to order n of its argument. Then there exists an affine function* $M_{\text{proc}}(\cdot)$ *so that*

$$\begin{aligned} \mathbf{M}_k^{\text{proc}} &= E\{M_{\text{proc}}[\text{mon}_n(y_k, \dots, y_{k-n_a}, u_{k-1}, \dots, u_{k-n_b})]\} \\ &= M_{\text{proc}}\{E[\text{mon}_n(y_k, \dots, y_{k-n_a}, u_{k-1}, \dots, u_{k-n_b})]\}. \end{aligned}$$

Example 4 (Construction of $\mathbf{M}_k^{\text{proc}}$). To better illustrate the proposed approach, we provide an example of how to construct the matrix $\mathbf{M}_k^{\text{proc}}$ required for SARX identification. To this end, consider the problem of identifying a SARX system with $n = 2$ subsystems of order $n_a = n_b = 1$ of the form

$$\begin{aligned} &\text{subsystem } 1 : y_k = a_1 \ y_{k-1} + b_1 \ u_{k-1} + \epsilon_k \\ &\text{subsystem } 2 : y_k = a_2 \ y_{k-1} + b_2 \ u_{k-1} + \epsilon_k, \end{aligned} \tag{27}$$

where ϵ_k has a symmetric distribution. We can rewrite the system as in Equation (24). In particular, the regressor vector $\mathbf{r}_k$ at time k

$$\mathbf{r}_k = \begin{bmatrix} y_k - \epsilon_k & y_{k-1} & u_{k-1} \end{bmatrix}^\top$$

gives rise to the following Veronese vector

$$\nu_n(\mathbf{r}_k) = \begin{bmatrix} (y_k - \epsilon_k)^2 \\ (y_k - \epsilon_k) \ y_{k-1} \\ (y_k - \epsilon_k) \ u_{k-1} \\ y_{k-1}^2 \\ y_{k-1} \ u_{k-1} \\ u_{k-1}^2 \end{bmatrix}, \tag{28}$$

whose size is $l \times 1$, with $l = \binom{n+n_a+n_b}{n} = \binom{2+1+1}{2} = 6$. The corresponding vector $\mathbf{c}_2$ as a function of the parameters of the subsystems, assumes the form

$$\mathbf{c}_2 = [1, -(a_1 + a_2), -(b_1 + b_2), a_1 a_2, a_1 b_2 + b_1 a_2, b_1 b_2]^\top.$$

From $\mathbf{r}_k$ and $\nu_n(\mathbf{r}_k)$, we can compute matrix $\mathbf{M}_k^{\text{proc}}$ as follows

$$\mathbf{M}_k^{\text{proc}} = \nu_n(\mathbf{r}_k) \ \nu_n^\top(\mathbf{r}_k)$$

$$= \begin{pmatrix} (y_k-\epsilon_k)^4 & (y_k-\epsilon_k)^3\, y_{k-1} & (y_k-\epsilon_k)^3\, u_{k-1} & (y_k-\epsilon_k)^2\, y_{k-1}^2 & (y_k-\epsilon_k)^2\, y_{k-1}\, u_{k-1} & (y_k-\epsilon_k)^2\, u_{k-1}^2 \\ * & (y_k-\epsilon_k)^2\, y_{k-1}^2 & (y_k-\epsilon_k)^2\, y_{k-1}\, u_{k-1} & (y_k-\epsilon_k)\, y_{k-1}^3 & (y_k-\epsilon_k)\, y_{k-1}^2\, u_{k-1} & (y_k-\epsilon_k)\, y_{k-1}\, u_{k-1}^2 \\ * & * & (y_k-\epsilon_k)^2\, u_{k-1}^2 & (y_k-\epsilon_k)\, y_{k-1}^2\, u_{k-1} & (y_k-\epsilon_k)\, y_{k-1}\, u_{k-1}^2 & (y_k-\epsilon_k)\, u_{k-1}^3 \\ * & * & * & y_{k-1}^4 & y_{k-1}^3\, u_{k-1} & y_{k-1}^2\, u_{k-1}^2 \\ * & * & * & * & y_{k-1}^2\, u_{k-1}^2 & y_{k-1}\, u_{k-1}^3 \\ * & * & * & * & * & u_{k-1}^4 \end{pmatrix}.$$

Then, as we have the values of noisy output y_k, we compute expected value of powers of $y_k - \epsilon_k$ in terms of expected value of powers of y_k and moments of process noise. Following the results in Lemma 3, we obtain the matrix below.

$$\mathbf{M}_k^{\text{proc}} = \begin{pmatrix} E[y_k^4] - 6\, m_2\, (E[y_k^2] - m_2) - m_4 & E[y_k^3\, y_{k-1}] - 3\, m_2\, E[y_k\, y_{k-1}] & (E[y_k^3] - 3\, m_2\, E[y_k])\, u_{k-1} & \cdots \\ * & E[y_k^2\, y_{k-1}^2] - m_2\, E[y_{k-1}^2] & (E[y_k^2\, y_{k-1}] - m_2\, E[y_{k-1}])\, u_{k-1} & \\ * & * & (E[y_k^2] - m_2)\, u_{k-1}^2 & \\ * & * & * & \\ * & * & * & \\ * & * & * & \end{pmatrix}$$

$$\begin{pmatrix} E[y_k^2\, y_{k-1}^2] - m_2\, E[y_{k-1}^2] & (E[y_k^2\, y_{k-1}] - m_2\, E[y_{k-1}])\, u_{k-1} & (E[y_k^2] - m_2)\, u_{k-1}^2 \\ E[y_k\, y_{k-1}^3] & E[y_k\, y_{k-1}^2]\, u_{k-1} & E[y_k\, y_{k-1}]\, u_{k-1}^2 \\ E[y_k\, y_{k-1}^2]\, u_{k-1} & E[y_k\, y_{k-1}]\, u_{k-1}^2 & E[y_k]\, u_{k-1}^3 \\ E[y_{k-1}^4] & E[y_{k-1}^3]\, u_{k-1} & E[y_{k-1}^2]\, u_{k-1}^2 \\ * & E[y_{k-1}^2]\, u_{k-1}^2 & E[y_{k-1}]\, u_{k-1}^3 \\ * & * & u_{k-1}^4 \end{pmatrix}$$

4.1 | SARX identification algorithm

As mentioned before, to identify the parameters of the SARX system, we need to be able to estimate the matrix $\mathcal{M}_N^{\text{proc}}$ in Equation (25). It turns out that it can be done by exploiting its structure and using the available noisy measurements. More precisely, we have the following result.

Theorem 2. *Let* $M_{\text{proc}}(\cdot)$ *and* $\text{mon}_n(\cdot)$ *be the functions defined in Lemma 4. Define*

$$\widehat{\mathcal{M}}_N^{\text{proc}} \doteq \frac{1}{N}\sum_{k=1}^{N} M_{\text{proc}}[\text{mon}_n(y_k, \dots, y_{k-n_a}, u_{k-1}, \dots, u_{k-n_b})].$$

Take any monomial

$$z_k = y_k^{h_0}\, y_{k-1}^{h_1}\, \cdots\, y_{k-n_a}^{h_{n_a}} \tag{29}$$

where $\sum_{i=0}^{n_a} h_i \le 2n$, $h_i = 0, 1, 2, \dots, 2n$. *If for any* $h_1, h_2, \dots, h_{n_a}$, *the sequence* $\{z_k,\ k \ge 1\}$ *satisfies*

- $\sum_{k=1}^{\infty} \frac{(\text{Var}\ z_k)(\log\ k)^2}{k^2} < \infty$
- $\sum_{l=1}^{\infty} \frac{\rho_l}{l^q} < \infty \quad$ for some $0 \le q < 1$,

where $\{\rho_l, l \ge 1\}$ *is a sequence of constants such that* $\sup_{k\ge 1}|\text{Cov}\ (z_k,\ z_{k+l})| \le \rho_l \quad l \ge 1$. *then, as* $N \to \infty$,

$$\widehat{\mathcal{M}}_N^{\text{proc}} - \mathcal{M}_N^{\text{proc}} \to 0 \quad \text{a.s.}$$

Proof. Direct application of the results in Reference 26 with $b_n = n$. For completeness this result is stated as Theorem 3 in Appendix. ∎

Remark 1 (On rate of convergence). Again from the results in Reference 26, it is possible to show that, for any $\epsilon > 0$, the (i, j) entry of the true and estimated matrices satisfy

$$\text{Probability}\left\{\left|\widehat{\mathcal{M}_N^{\text{proc}}}_{(i,j)} - \mathcal{M}_N^{\text{proc}}{}_{(i,j)}\right| \geq \epsilon\right\} \leq \frac{1}{\epsilon^2}\left(\sum_{k=N}^{\infty} \frac{\text{Var}M_{\text{proc}_{(i,j)}}[\text{mon}_n(y_k, \dots, y_{k-n_a}, u_{k-1}, \dots, u_{k-n_b})]}{k^2} + \frac{C}{N^{(1-q)}}\right),$$

and, hence, the rate of convergence is directly connected to the cross moments of the output of the system. Note that if these moments are bounded the rate of convergence is of the order $O(N^{(1-q)})$.

The conditions of the theorem above are rather general and state that, if the output of the system is "well behaved" then empirical averages of functions of the collected data can be used of estimate the matrix $\mathcal{M}_N$ and, hence, the coefficients of the subsystems.

Although these condition are rather abstract, it turns out that there is a an important special case where Theorem 2 can be applied, namely, the case when the SARX system is uniformly exponentially stable and the noise is normally distributed.

Corollary 1. *Let the SARX system in (22) be uniformly exponentially stable, and the noise distribution is zero mean normal, that is, $\epsilon_k \sim N(0, \sigma^2)$. Assume moreover that the dynamics of switching δ_k at time k are independent from input u_k and output y_k. Then the conditions of Theorem 2 are satisfied, and therefore as $N \to \infty$,*

$$\widehat{\mathcal{M}}_N^{\text{proc}} - \mathcal{M}_N^{\text{proc}} \to 0 \quad \text{a.s.}$$

Proof. See Appendix. ∎

As a result, the empirical average computed using the noisy measurements (where expected values of monomials are replaced by the averages of the measured monomial values) converges to the desired matrix in Equation (25). Therefore we propose the following algorithm for identification of a SARX system.

Algorithm 3 (SARX system identification). *Let n_a, n_b, n and some parameters of the noise be given.*

Step 1. Compute matrix

$$\widehat{\mathcal{M}}_N^{\text{proc}} = \frac{1}{N}\sum_{k=1}^{N} M_{proc}[\text{mon}_n(y_k, \dots, y_{k-n_a}, u_{k-1}, \dots, u_{k-n_b})].$$

Step 2. Let $\mathbf{c}_n$ be the singular vector associated with the minimum singular value of $\widehat{\mathcal{M}}_N^{\text{proc}}$.
Step 3. Determine the coefficients of the subsystems from the vector $\mathbf{c}_n$.

5 | ESTIMATING UNKNOWN NOISE PARAMETERS

We now address the case where the distribution of the noise is not completely known. In particular, as previously mentioned, in this article it is assumed that the distribution of the noise is known except for a few parameters. For simplicity of exposition, we consider the case where the noise has one scalar unknown parameter θ. The reasoning can be extended to any case where the set of allowable parameters can be efficiently gridded/searched.

Recall that throughout this article we assume that the noise density $f(x|\theta)$ and the corresponding moments $m_i(\theta)$ are continuous functions of θ. The following fact is an immediate consequence of this assumption.

Fact 1. Consider the matrices $\mathcal{M}_N(\theta)$ defined in (9) (whose dependence on the moments of noise is described in Lemma 1) and $\mathcal{M}_N^{\text{proc}}(\theta)$ defined in (25) (whose dependence on the moments of noise is described in Lemma 4). Moreover, consider their estimates $\widehat{\mathcal{M}}_N(\theta)$ and $\widehat{\mathcal{M}}_N^{\text{proc}}(\theta)$. Then, $\mathcal{M}_N^{\text{proc}}(\theta)$, $\widehat{\mathcal{M}}_N(\theta)$, and $\widehat{\mathcal{M}}_N^{\text{proc}}(\theta)$ and the respective singular values are continuous functions of θ.

Furthermore, we have the following result that is a consequence of the definition of the matrices $\mathcal{M}_N(\theta)$ and $\mathcal{M}_N^{\text{proc}}(\theta)$.

Lemma 5. *Let θ^* be the true value of the parameter θ. Then $\mathcal{M}_N(\theta^*)$ (respectively $\mathcal{M}_N^{\text{proc}}(\theta^*)$) is rank deficient.*

Finally, we have the following result which is an immediate consequence of the convergence results in Theorems 1 and 2.

Lemma 6. *Let $\sigma_{\min}(.)$ denote the minimum singular value. Then, for the cause of measurement noise, we have*

$$\lim_{N\to 0} \sigma_{\min}[\mathcal{M}_N(\theta)] - \sigma_{\min}[\widehat{\mathcal{M}}_N(\theta)] = 0 \quad \text{for all admissible } \theta$$

and, for the case of process noise, we have

$$\lim_{N\to 0} \sigma_{\min}[\mathcal{M}_N^{\text{proc}}(\theta)] - \sigma_{\min}[\widehat{\mathcal{M}}_N^{\text{proc}}(\theta)] = 0 \quad \text{for all admissible } \theta.$$

Given the convergence results above, the true value of θ will result in matrices $\widehat{\mathcal{M}}_N(\theta)$ ($\widehat{\mathcal{M}}_N^{\text{proc}}(\theta)$ for process noise) with a "very small" minimum singular value, especially for large values of N. For this reason, estimation of the noise parameter θ can be performed by minimizing the minimum singular value of $\widehat{\mathcal{M}}_N(\theta)$ (for process noise $\widehat{\mathcal{M}}_N^{\text{proc}}(\theta)$) over the allowable values. More precisely, we propose the following algorithm

Algorithm 4 (Joint SARMAX system and noise parameter identification). *Let n_a, n_b, n, $\theta_{\min}$, and $\theta_{\max}$ be given.*

Step 1. *Compute matrix $\widehat{\mathcal{M}}_N(\theta)$ (or $\widehat{\mathcal{M}}_N^{\text{proc}}(\theta)$ for process noise) as a function of the noise parameter θ.*
Step 2. *Find the value $\theta^* \in [\theta_{\min}, \theta_{\max}]$ that minimizes the minimum singular value of $\widehat{\mathcal{M}}_N(\theta)$ (or $\widehat{\mathcal{M}}_N^{\text{proc}}(\theta)$ for process noise).*
Step 3. *Let $\mathbf{c}_n$ be associated singular vector.*
Step 4. *Determine the coefficients of the subsystems from the vector $\mathbf{c}_n$.*

Remark 2. Note that the optimization in step 2 is in general nonconvex, but it can be solved via an easily implementable line search, given the continuity of the minimum singular value with respect to θ. However, the solution θ^* might not be unique, that is, there might exist several values of θ that lead to a minimum singular value very close to zero. In practice, our experience has been that, for sufficiently large N, the above algorithm provides both a good estimate of the systems coefficients, and noise parameters; especially if we take θ^* to be the smallest value of θ for which the minimum singular value of $\widehat{\mathcal{M}}_N$ ($\widehat{\mathcal{M}}_N^{\text{proc}}$) is below a given threshold.

Remark 3. In the algorithm above, minimizing the minimum singular value can be replaced by minimizing the determinant since both are zero for rank deficient matrices. A case where this might be preferable is when the noise is Gaussian with zero mean and unknown variance. In this case, the determinant of $\widehat{\mathcal{M}}_N(\theta)$ and $\widehat{\mathcal{M}}_N^{\text{proc}}(\theta)$ are polynomial function of the variance and, hence, minimization can be performed by looking at the real zeros of the derivative of the determinant with respect to θ.

6 | NUMERICAL RESULTS

In this section, we present some numerical examples that illustrate the effectiveness of the proposed approach.

6.1 | SARMAX system identification

In the following example, we address the problem of identifying a two-mode switched system of the form (11)-(12), whose true coefficients are $a_1 = 0.3, b_1 = 1, a_2 = -0.5$, and $b_2 = -1$. Measurement noise is assumed to be zero mean with normal distribution. In the numerical examples presented, $N = 10^6$ input-output data are given. True and identified coefficients for different variances of noise are presented in Table 1. Variance of noise and noise to output ratio for each experiment are also shown in this table. The provided noise to output ratio (γ) is defined as

$$\gamma = \frac{\max |\eta|}{\max |y|}. \tag{30}$$

TABLE 1 Identifying polynomial coefficients for different values of noise variance and different SARMAX system run

Experiment #	Value 1 1	Value 2 $-(a_1+a_2)$	Value 3 $-(b_1+b_2)$	Value 4 $a_1 a_2$	Value 5 $a_1 b_2 + b_1 a_2$	Value 6 $b_1 b_2$	Value 7 γ	Value 8 σ^2	Value 9 $\hat{\sigma}^2$
True parameters	1	0.2	0	−0.15	−0.8	−1	—	—	—
Identification 1	1	0.2002	0.0001	−0.1503	−0.7989	−0.9996	0.2410	0.1	0.1000
Identification 2	1	0.2002	0.0011	−0.1510	−0.7974	−1.0004	0.5187	0.5	0.4980
Identification 3	1	0.1977	0.0046	−0.1548	−0.7997	−0.9966	0.6494	1	1.0010
Identification 4	1	0.2120	0.0003	−0.1485	−0.8006	−1.0017	0.8516	2	1.9950

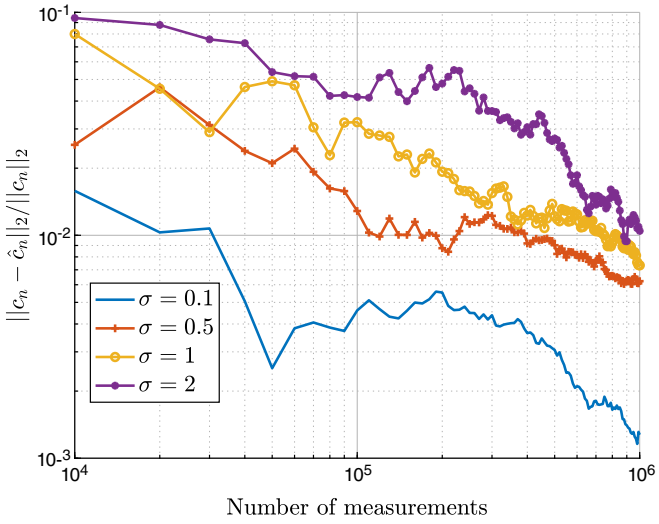


FIGURE 1 Estimation error of system coefficients [Colour figure can be viewed at wileyonlinelibrary.com]

Results are as expected even for high values of noise in comparison to output. As it is illustrated in Table 1, the identified parameters are very close to true values, which demonstrates the convergence of proposed algorithm even for small signal to noise ratio. Moreover, the algorithm requires a very small computational effort. For the case of 10^6 measurements and using an off-the-shelf core i5 laptop with 8 Gigs of RAM, the running time is between 7 and 8 seconds, which shows the effectiveness of approach for very large datasets.

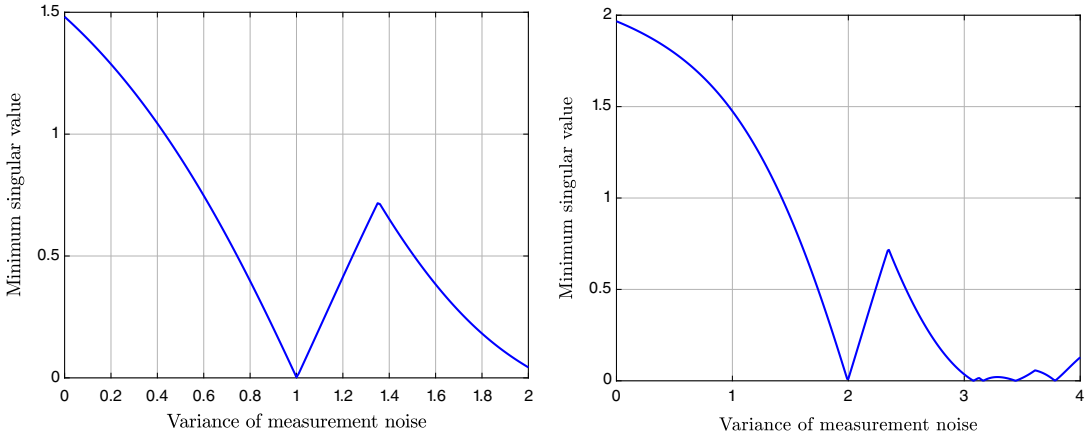
The error between true coefficients of system and estimated coefficients, $\|\mathbf{c}_n - \hat{\mathbf{c}}_n\|_2/\|\mathbf{c}_n\|_2$, as a function of number of measurements, N, is depicted in Figure 1 for different values of noise variance. As it can be seen from Figure 1, the error decreases as the number of measurements increases. Rate of convergence is fast, despite the fact that, in some of the experiments, a large amount of noise is used. It should be noted that these results are for one experiment and given that this is a realization of a random process, error is not always decreasing. For all values of noise variance, error will eventually decrease and the estimated values of coefficients converge to the true values.

Now we consider estimation of the individual subsystems. For the above mentioned example, Table 2 shows the values of subsystems coefficients for different experiments related to different values of noise variance. As we see in this table the values of coefficients are very close to the true values, even when the noise variance is high with noise magnitude in average around 85% of the signal magnitude.

Several experiments are done for the problem of identifying a four-mode switched system of the form (11)-(12), where each subsystem is of order 2. The systems have been created randomly. Measurement noise is assumed to be zero mean with normal distribution and $N = 5 \times 10^6$ input-output data are given. Experiments have been done for different systems and different variances of noise. As two examples, for randomly generated systems with variance of noise $\sigma^2 = 0.1$ and $\sigma^2 = 0.3$, the error between true and estimated coefficients of the system, $\|\mathbf{c}_n - \hat{\mathbf{c}}_n\|_2/\|\mathbf{c}_n\|_2$, is 0.03 and 0.16, respectively. Note that the size of the vector $\mathbf{c}_n$ is, in this case, 35, indicating that the identification problem here is more complex than previous ones. Hence, a slightly larger error in estimation is expected.

TABLE 2 Identifying submodels' coefficients for different values of noise variance in SARMAX systems

Submodels' coefficients	True values	Variance $\sigma^2 = 0.1$	Variance $\sigma^2 = 0.5$	Variance $\sigma^2 = 1$	Variance $\sigma^2 = 2$
a_1	0.3	0.3002	0.2981	0.3006	0.2938
b_1	1	0.9988	1.0007	0.9412	1.0031
a_2	−0.5	−0.4996	−0.5000	−0.5006	−0.5059
b_2	−1	−0.9999	−0.9991	−1.0004	−1.0011



(A) Case 1: true noise variance $\sigma^2 = 1$.

(B) Case 2: true noise variance $\sigma^2 = 2$.

FIGURE 2 Estimation of noise variance using Algorithm 3, for ARMAX system consisting two subsystems of order 1 [Colour figure can be viewed at wileyonlinelibrary.com]

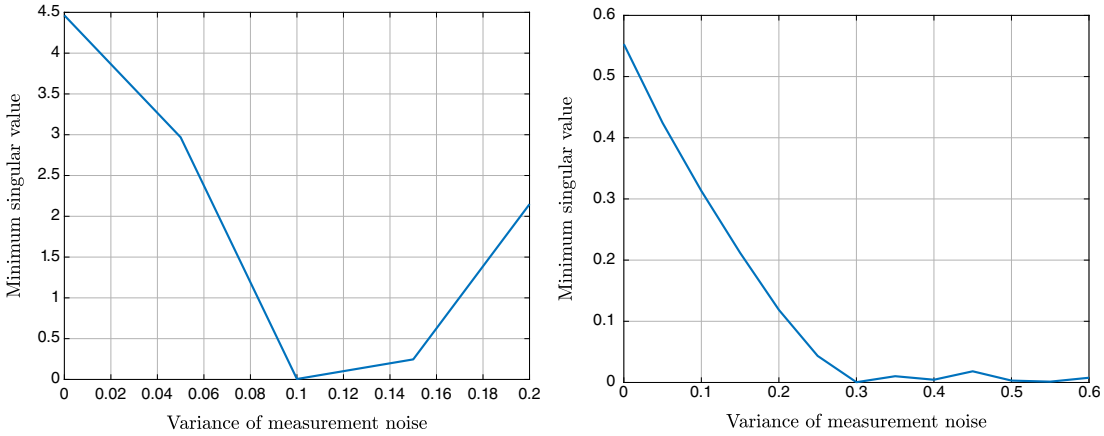
The estimation of noise variance based on the structure of matrix $\mathbf{M}_k$ is shown in Table 1 as $\hat{\sigma}^2$. The estimates of noise variance are very close to the true values of variance. By knowing the structure of matrix $\mathbf{M}_k$, the dependence of every entry on the moments of noise, and the relation in between these moments and the unknown variance (see Section 1.2), we are able to estimate the noise parameter (in this case, noise variance). This illustrates the capability of the proposed algorithm to estimate both system and noise parameters even for large values of noise.

Several examples of the process of estimating the unknown variance of noise are shown in Figures 2 and 3. Figure 2 shows the results for ARMAX system consisting two subsystems of order 1, where Figure 2A is for the case of given data contaminated with noise of variance 1, and Figure 2B is for data with measurement noise of variance 2. By taking σ^* as the smallest local minimum, the estimated variance for both cases in Figure 2A,B is very close to the true values.

Figure 3 demonstrates the estimation of unknown variance of noise for ARMAX system consisting four subsystems of order 2, where Figure 3A is for the case of given data contaminated with noise of variance 0.1, Figure 3B is for data with measurement noise of variance 0.3. By taking σ^* as the smallest local minimum, the estimated variances for both cases in Figure 3 are very close to the true values.

6.2 | SARX system identification

In this section's examples, we address the problem of identifying a two-mode switched system of the form of Equation (27), whose true coefficients are, again, $a_1 = 0.3, b_1 = 1, a_2 = -0.5$, and $b_2 = -1$. Process noise is assumed to be zero mean with normal distribution. A total number of $N = 10^6$ input-output data are given for each experiment. True and identified coefficients for different variances of noise are presented in Table 3. Noise to output ratio and estimate of noise variance for each experiment are also shown in this table.



(A) True noise variance $\sigma^2 = 0.1$. (B) True noise variance $\sigma^2 = 0.3$.

FIGURE 3 Estimation of noise variance using Algorithm 3, for ARMAX system consisting four subsystems of order 2 [Colour figure can be viewed at wileyonlinelibrary.com]

TABLE 3 Identifying polynomial coefficients for different values of noise variance and different SARX system run

Experiment #	Value 1 1	Value 2 $-(a_1+a_2)$	Value 3 $-(b_1+b_2)$	Value 4 $a_1 a_2$	Value 5 $a_1 b_2 + b_1 a_2$	Value 6 $b_1 b_2$	Value 7 γ	Value 8 σ^2	Value 9 $\hat{\sigma}^2$
True parameters	1	0.2	0	−0.15	−0.8	−1	—	—	—
Identification 1	1	0.2006	−0.0011	−0.1504	−0.8011	−1.0001	0.2657	0.1	0.1
Identification 2	1	0.1991	−0.0001	−0.1506	−0.8029	−1.0007	0.5044	0.5	0.5
Identification 3	1	0.1960	−0.0008	−0.1499	−0.7963	−1.0032	0.5656	1	1
Identification 4	1	0.2052	0.0084	−0.1493	−0.8050	−1.0036	0.7649	2	2

Once again we see that the proposed approach is very effective. As depicted in Table 3, the error in the identification of the system's parameters is very small, which demonstrates the convergence of proposed algorithm even for small signal to noise ratio. Again, the algorithm requires a very small computational effort. For the case of 10^6 measurements and using the same off-shelf computer as before, the running time is between 2 and 9 seconds. Again, this shows how well the proposed approach scales with the number of measurements.

The estimation error, $\|\mathbf{c}_n - \hat{\mathbf{c}}_n\|_2 / \|\mathbf{c}_n\|_2$, as a function of number of measurements, N, is depicted in Figure 4 for different values of noise variance. As it can be seen from Figure 4, the error again decreases as the number of measurements increases. For all values of noise variance, error will eventually decrease and the estimated values of coefficients converge to the true values.

For the above mentioned example, Table 4 shows the values of subsystems coefficients for different experiments using different values of noise variance. The values of coefficients are very close to the true values, even when the noise variance is high with noise magnitude in average around 76% of the signal magnitude.

The estimation of noise variance based on the structure of matrix $\mathbf{M}_k^{\text{proc}}$ is shown in Table 3 as $\hat{\sigma}^2$. As it can be seen from the results obtained, we can efficiently and simultaneously estimate the system's coefficients and the noise variance. This illustrates the capability of the proposed algorithm to estimate both system and noise parameters even for large values of noise.

Two examples of the process of estimating the unknown variance of process noise are shown in Figure 5; where Figure 5A is for the case of given data with process noise of variance 1, and Figure 5B is for data with process noise of variance 2. By taking σ^* as the smallest local minimum, the estimated variance for both cases in Figure 5A,B is the true values.

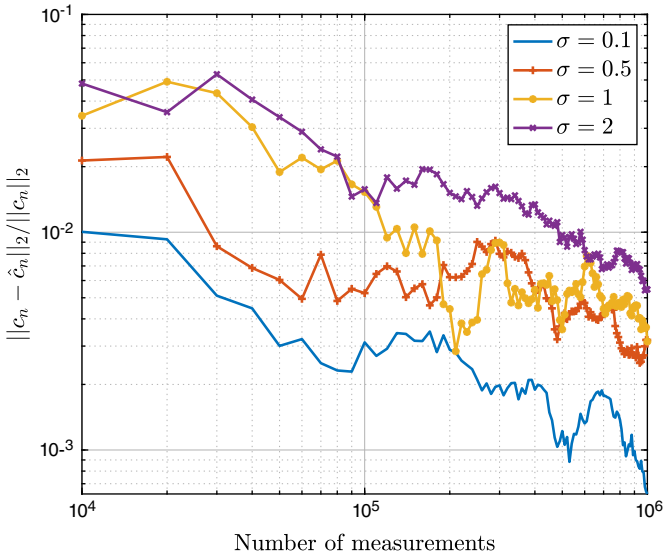


FIGURE 4 Estimation error of SARX system coefficients [Colour figure can be viewed at wileyonlinelibrary.com]

TABLE 4 Identifying submodels' coefficients for different values of noise variance in **SARX** systems

Submodels' coefficients	True values	Variance $\sigma^2 = 0.1$	Variance $\sigma^2 = 0.5$	Variance $\sigma^2 = 1$	Variance $\sigma^2 = 2$
a_1	0.3	0.3001	0.3011	0.2782	0.2987
b_1	1	1.0006	1.0022	0.9724	0.9978
a_2	−0.5	−0.5008	−0.5004	−0.4959	−0.5100
b_2	−1	−0.9995	−1.0007	−1.0012	−1.0096

6.2.1 | Average behavior of the algorithm

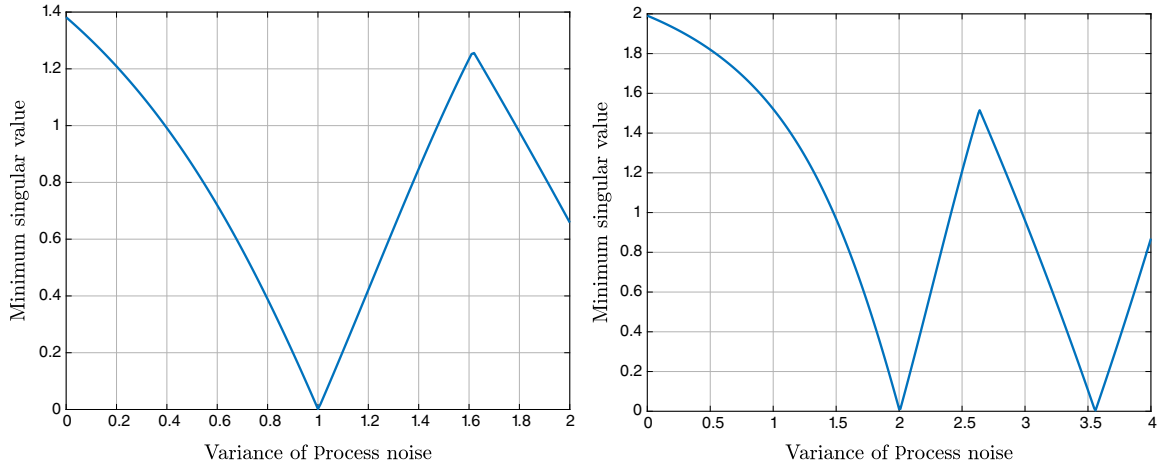
We examine the average behavior of the algorithm proposed in this article for randomly generated stable discrete ARX systems. Systems are randomly generated using the `drss` command of MATLAB, which ensures system poles are random and stable with possible exception of poles at 1. Randomly selected systems are considered to be of the form of Equation (27) with order 1, that is, $n_a = 1$ and $n_b = 1$, and switched system considered to include two submodels, that is, $n = 2$. The average behavior of system is tested for different values of noise variance. In each case, 100 random experiments were run for the total number of measurements $N = 10^6$.

The average behavior of the system is shown in Table 5. Normalized error is shown by β and computed as

$$\beta = \frac{\|\mathbf{c}_n - \hat{\mathbf{c}}_n\|}{\|\mathbf{c}_n\|}.$$

For different values of variance of noise $\sigma^2 = 0.1, 0.3, 0.5, 0.7$, the average mean and variance of normalized error are computed and shown in Table 5. In each experiment consisting of 100 runs of system, the average of noise to output ratio (γ) is computed and shown in Table 5. Note that for some of the randomly generated systems the value of noise to output ratio is close to 1. Also average of elapsed time for running the algorithm is shown in Table 5.

As we see in this table, for different values of noise variance, the average of difference in identified coefficients in comparison with the true values is really small. This happens even in the case of large noise to output ratio. For example, in the case of $\sigma^2 = 0.3$, the average of normalized error is just 0.73% and this is with having approximately 48% noise to output ratio in average. Hence, the algorithm can recover the original system efficiently, with very low estimation error and in a short period of time.



(A) Case 1: true noise variance $\sigma^2 = 1$. (B) Case 2: true noise variance = $\sigma^2 = 2$.

FIGURE 5 Estimation of noise variance [Colour figure can be viewed at wileyonlinelibrary.com]

TABLE 5 Average behavior of algorithm for different values of noise variance for randomly generated 100 SARX systems

Noise variance	Mean of β	variance of β	Mean of γ	Mean of elapsed time
0.1	0.0025	1.3564e−05	0.4259	2.9393
0.3	0.0073	3.3339e−04	0.4799	2.8286
0.5	0.0083	2.8769e−04	0.5373	2.7683
0.7	0.0111	5.1764e−04	0.5452	2.8888

7 | CONCLUDING REMARKS

In this article, we propose methodologies to identify the coefficients of SARMAX and SARX processes and unknown noise parameters, starting from partial information of the noise and given input-output data. The approach is shown to be particularly efficient in the case of large amount of data. The approach only requires the computation of singular value decomposition of a specially constructed input-output Veronese matrix. The ensuing singular vector is then related to the switched system parameters to be identified. We prove that the estimated parameters converge to the true ones as the number of measurements grows. Numerical simulations show a low estimation error, even in the case of large measurement and process noise. Also, in cases that noise distribution is not completely known, simulation results show very efficient estimation of unknown noise parameters.

ACKNOWLEDGEMENT

This work was partially supported by National Institutes of Health (NIH) Grant R01 HL142732, National Science Foundation (NSF) Grant #1808266, and the International Bilateral Joint CNR-JST Lab COOPS.

ORCID

Constantino M. Lagoa https://orcid.org/0000-0001-6871-3240

REFERENCES

1. Ozay N, Sznaier M, Lagoa C. Convex certificates for model (in) validation of switched affine systems with unknown switches. *IEEE Trans Automat Control*. 2014;59(11):2921-2932.
2. Sznaier M, Camps O, Ozay N, Lagoa C. Surviving the upcoming data deluge: a systems and control perspective. Paper presented at: Proceedings of the 53rd IEEE Conference on Decision and Control; 2014:1488-1498.
3. Hojjatinia S, Shoorehdeli MA, Fatahi Z, Hojjatinia Z, Haghparast A. Improvement of the Izhikevich model based on the rat basolateral amygdala and hippocampus neurons, and recognition of their possible firing patterns. *Basic Clin Neurosci*. 2019.

4. Lin RS, Liu CB, Yang MH, Ahuja N, Levinson S. Learning nonlinear manifolds from time series. Paper presented at: Proceedings of the European Conference on Computer Vision; 2006:245-256.
5. Hojjatinia S, Bekiroglu K, Lagoa CM. Parsimonious volterra system identification. Paper presented at: Proceedings of the 2018 Annual American Control Conference (ACC); 2018:1933-1938.
6. Saul LK, Roweis ST. Think globally, fit locally: unsupervised learning of low dimensional manifolds. *J Mach Learn Res*. 2003;4(Jun):119-155.
7. Lunze J, Lamnabhi-Lagarrigue F. *Handbook of Hybrid Systems Control: Theory, Tools, Applications*. Cambridge, MA: Cambridge University Press; 2009.
8. Paoletti S, Juloski AL, Ferrari-Trecate G, Vidal R. Identification of hybrid systems a tutorial. *Eur J Control*. 2007;13(2-3):242-260.
9. Garulli A, Paoletti S, Vicino A. A survey on switched and piecewise affine system identification. *IFAC Proc Vol*. 2012;45(16):344-355.
10. Lauer F, Bloch G, Vidal R. A continuous optimization framework for hybrid system identification. *Automatica*. 2011;47(3):608-613.
11. Roll J, Bemporad A, Ljung L. Identification of piecewise affine systems via mixed-integer programming. *Automatica*. 2004;40(1):37-50.
12. Ferrari-Trecate G, Muselli M, Liberati D, Morari M. A clustering technique for the identification of piecewise affine systems. *Automatica*. 2003;39(2):205-217.
13. Juloski AL, Weiland S, Heemels WPH. A Bayesian approach to identification of hybrid systems. *IEEE Trans Autom Control*. 2005;50(10):1520-1533.
14. Saxén JE, Saxén H, Toivonen HT. Identification of switching linear systems using self-organizing models with application to silicon prediction in hot metal. *Appl Soft Comput J*. 2016;47:271-280.
15. Bako L. Identification of switched linear systems via sparse optimization. *Automatica*. 2011;47(4):668-677.
16. Ozay N, Lagoa C, Sznaier M. Set membership identification of switched linear systems with known number of subsystems. *Automatica*. 2015;51:180-191.
17. Hojjatinia S, Lagoa CM, Dabbene F. A method for identification of markovian jump arx processes. *IFAC-Papers OnLine*. 2017;50(1):14088-14093.
18. Lagoa CM, Conroy DE, Hojjatinia S, Yang CH. Modeling subject response to interventions aimed at increasing physical activity: a control systems approach. Poster Presented at 5th International Conference on Ambulatory Monitoring of Physical Activity and Movement; 2017.
19. Conroy DE, Hojjatinia S, Lagoa CM, Yang CH, Lanza ST, Smyth JM. Personalized models of physical activity responses to text message micro-interventions: a proof-of-concept application of control systems engineering methods. *Psychol Sport Exerc*. 2019;41:172-180.
20. Hojjatinia S, Lagoa CM, Dabbene F. Identification of switched autoregressive systems from large noisy data sets. Paper presented at: Proceedings of the 2019 American Control Conference (ACC); 2019:4313-4319.
21. Ma Y, Vidal R. *Identification of Deterministic Switched ARX Systems via Identification of Algebraic Varieties*. Berlin, Heidelberg / Germany: Springer; 2005:449-465.
22. Vidal R, Soatto S, Ma Y, Sastry S. An algebraic geometric approach to the identification of a class of linear hybrid systems. Paper presented at: Proceedings of the 42nd IEEE Conf. Decision and Control; 2003:167-172.
23. Harris J. *Algebraic Geometry: A First Course*. Vol 133. Berlin, Heidelberg / Germany: Springer Science & Business Media; 2013.
24. Vidal R, Ma Y, Sastry S. Generalized principal component analysis (GPCA). *IEEE Trans Pattern Anal Mach Intell*. 2005;27(12):1945-1959.
25. Vidal RE, Sastry S. *Generalized Principal Component Analysis (gpca): An Algebraic Geometric Approach to Subspace Clustering and Motion Segmentation*. Los Angeles, CA: Electronics Research Laboratory, College of Engineering, University of California; 2003.
26. Hu TC, Rosalsky A, Volodin A. On convergence properties of sums of dependent random variables under second moment and covariance restrictions. *Stat Probab Lett*. 2008;78(14):1999-2005.
27. Sen PK, Singer JM. *Large Sample Methods in Statistics: An Introduction with Applications*. New York, NY: Chapman & Hall; 1993.
28. Isserlis L. On a formula for the product-moment coefficient of any order of a normal frequency distribution in any number of variables. *Biometrika*. 1918;12(1/2):134-139.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Hojjatinia S, Lagoa CM, Dabbene F. Identification of switched autoregressive exogenous systems from large noisy datasets. *Int J Robust Nonlinear Control*. 2020;1–25.

https://doi.org/10.1002/rnc.4968

APPENDIX

A1. Proof of Theorem 1

For simplicity of presentation, let

$$\hat{M}_k \doteq M[\text{mon}_n(y_k, \dots, y_{k-n_a}, u_{k-1}, \dots, u_{k-n_b})].$$

We first note that, given the assumptions made on the noise, u_k and x_k, the entries of $\hat{M}_k$ have a variance uniformly bounded for all k. Moreover

$$k > l + n_a \Rightarrow \hat{M}_k \quad \text{and} \quad \hat{M}_l \ \text{ are independent.}$$

Hence, by Kolmogorov's strong law of large numbers[27] we have

$$\frac{1}{L}\sum_{l=1}^{L} \hat{M}_{k+l(n_a+1)} - \frac{1}{L}\sum_{l=1}^{L} E[\hat{M}_{k+l(n_a+1)}] \to 0 \quad \text{a.s.}$$

as $L \to \infty$. To establish speed of convergence, let us look at the (i,j) entry of the matrices above. As mentioned, there exists a $\overline{\sigma}_{i,j}$ so that

$$\text{Var}\left[\hat{M}_k(i,j)\right] \le \overline{\sigma}_{i,j}^2 \quad \text{for } \ k = 1, 2, \dots .$$

Then, the results in the proof of theorem 2.3.10 in Reference 27 imply that, for any $\epsilon > 0$

$$\text{Prob}\left\{ \max_{L \ge J} \left| \frac{1}{L}\sum_{l=1}^{L} \hat{M}_{k+l(n_a+1)}(i,j) - \frac{1}{L}\sum_{l=1}^{L} E[\hat{M}_{k+l(n_a+1)}(i,j)] \right| > \epsilon \right\} \le \frac{\overline{\sigma}_{i,j}^2}{\epsilon^2}\left(\frac{1}{J} + \sum_{k \ge J+1} k^{-2} \right).$$

In other words, for large J

$$\text{Prob}\left\{ \max_{L \ge J} \left| \frac{1}{L}\sum_{l=1}^{L} \hat{M}_{k+l(n_a+1)}(i,j) - \frac{1}{L}\sum_{l=1}^{L} E[\hat{M}_{k+l(n_a+1)}(i,j)] \right| > \epsilon \right\} \le O(1/J).$$

Since

$$E[\hat{M}_k] = M_k \ \text{ for all positive integer } \ k$$

and applying the results above for $k = 1, 2, \dots, n_a + 1$, we conclude that

$$\frac{1}{N}\sum_{j=1}^{N} \hat{M}_j - \frac{1}{N}\sum_{j=1}^{N} M_j \to 0 \quad \text{a.s.}$$

as $N \to \infty$ and the convergence rate is $O(1/N)$.

A2. Convergence properties of sums of dependent random variables

Theorem 3. *[26] Let $\{X_n, n \ge 1\}$ be a sequence of square-integrable random variables and suppose that there exists a sequence of constants $\{\rho_k, k \ge 1\}$ such that*

$$\sup_{n \ge 1} |\text{Cov}\ (X_n,\ X_{n+k})| \le \rho_k \quad k \ge 1 \tag{A1}$$

holds. Let $\{b_n, n \ge 1\}$ be a sequence of positive constants satisfying

$$n = O(b_n).$$

Suppose that

$$\sum_{n=1}^{\infty} \frac{(\mathrm{Var}\ X_n)(\log\ n)^2}{b_n^2} < \infty \tag{A2}$$

and

$$\sum_{k=1}^{\infty} \frac{\rho_k}{k^q} < \infty \quad \text{for some } 0 \le q < 1. \tag{A3}$$

Then

$$\sum_{i=1}^{n} \frac{X_i - E[X_i]}{b_i} \quad \text{converges a.s. as} \quad n \to \infty, \tag{A4}$$

and if $b_n \uparrow$, *the strong law of large number holds, i.e.*

$$\lim_{n\to\infty} \frac{\sum_{i=1}^{n}(X_i - E[X_i])}{b_n} = 0 \quad \text{a.s.} \tag{A5}$$

A3. Proof of Corollary 1

If assumptions of Corollary 1 hold, the SARX system behaves like linear time varying (LTV) system. In general, the impulse response of the discrete LTV system at time k is described by

$$y_k = \sum_{m=0}^{k} g(k,m)\ \epsilon_m + R_k(u) + R_k(ic), \tag{A6}$$

where $R_k(u)$ is the response to the system input u and $R_k(ic)$ is the response to initial condition. Since the SARX system is uniformly exponentially stable and moments of input and noise are bounded, the responses $R_k(u)$ and $R_k(ic)$ are bounded. On the other hand, the computation of expected value of output monomials is a linear combination of the expected values of three responses above. Since the responses $R_k(u)$ and $R_k(ic)$ are bounded, in the following reasoning, we concentrate on the response to noise. Hence, consider the impulse response of SARX system to be of the form

$$y_k = \sum_{m=0}^{k} g(k,m)\ \epsilon_m. \tag{A7}$$

The discrete time LTV system introduced in Equation (A7) is exponentially stable if and only if there exists a constant M and $0 < a < 1$ such that

$$|g(k,m)| \le M\ a^{(k-m)} \quad \forall k \ge m. \tag{A8}$$

If we consider the vector matrix format of Equation (A7), that is,

$$\begin{pmatrix} y_0 \\ y_1 \\ \vdots \\ y_N \end{pmatrix} = \begin{pmatrix} g(0,0) & 0 & \dots & 0 \\ g(1,0) & g(1,1) & \dots & 0 \\ \dots & \dots & \dots & \dots \\ g(N,0) & g(N,1) & \dots & g(N,N) \end{pmatrix} \begin{pmatrix} \epsilon_0 \\ \epsilon_1 \\ \vdots \\ \epsilon_N \end{pmatrix}$$

or equivalently

$$\mathbf{y} = A\ \boldsymbol{\epsilon}, \tag{A9}$$

where $\mathbf{y}$ is the vector of output measurement and $\boldsymbol{\epsilon}$ is the vector of noise measurement for all the time. Note that covariance of $\mathbf{y}$ is computed as

$$\text{Cov}(\mathbf{y}) = E\,[\mathbf{y}\,\mathbf{y}^\top] \;-\; E\,[\mathbf{y}]\;\; E\,[\mathbf{y}^\top]. \tag{A10}$$

Since we consider noise to have zero mean normal distribution, output has also normal distribution and is mean is zero. Hence, the covariance of $\mathbf{y}$ in Equation (A10) is

$$\text{Cov}(\mathbf{y}) = E\,[\mathbf{y}\,\mathbf{y}^\top] = A\;E\,[\boldsymbol{\epsilon}\;\boldsymbol{\epsilon}^\top]\;A^\top = m_2\;A\;A^\top, \tag{A11}$$

where m_2 is the second moment or variance of the noise.

Considering the case where assumptions of Theorem 1 hold, we use several steps of reasoning to show the conditions of Theorem 2 are satisfied and that the algorithm in this article converges.

1. We assumed that SARX system is uniformly exponentially stable, input is bounded, and moments of noise up to order $4n$ are bounded.
2. Step (1) leads to having the expected value of output monomial up to order $2n$ bounded and therefore the variance of output monomials is bounded.
3. Steps (1) and (2) lead to the conditions of Theorem 2 being satisfied.
4. As a result, the strong law of large numbers holds for the monomials of system output up to order $2n$ and the average of the results obtained from a large number of experiments converges to the desired value in Equation (25) almost surely. In other words, Theorem 1 holds and $\widehat{\mathcal{M}}_N^{\text{proc}} - \mathcal{M}_N^{\text{proc}} \to 0$ a.s., as $N \to \infty$.

Now, we prove every step from reasoning above:

1. Input and moments of noise are bounded:

Based on Assumption 2, input is given and bounded, and moments of noise up to order $4n$ are bounded.

2. Expected value of output monomial up to order $2n$ *are bounded:*

Switched system is uniformly exponentially stable and input is bounded. Moreover, noise is assumed to have zero mean normal distribution with bounded moments, so output moments are also bounded. Therefore, the output monomial z_k as in Equation (29) is a monomial of normal random variables, so its expectation is bounded. As a result, the variance of output monomials is also bounded.

3. Conditions of Theorem 2 *are satisfied.*

Let us consider z_k as a monomial of output up to order $2n$ as defined in Equation (29), then

$$\text{Cov}\;(z_k,\; z_{k+l}) = E\,[z_k\;z_{k+l}] \;-\; E\,[z_k]\;E\,[z_{k+l}]. \tag{A12}$$

Since the noise is considered to be zero mean normal, output monomials have multivariate normal distribution. Hence, we are able to compute higher order moments of the multivariate normal distribution in terms of its covariance matrix based on Isserlis' theorem,[28] which is as follows:

Isserlis' theorem[28] If $(X_1, X_2, \dots, X_{2n+1}), \forall n = 1, 2, \dots$ are zero mean multivariate Normal random variables, then

$$E\,[X_1\;X_2\;\cdots\;X_{2n}] = \sum\;\prod\;E\,[X_iX_j] \tag{A13}$$

and

$$E\,[X_1\;X_2\;\cdots\;X_{2n+1}] = 0, \tag{A14}$$

where the notation $\sum\;\prod$ means summing over all distinct ways of partitioning $X_1, X_2, \dots, X_{2n}$ into pairs X_i, X_j, which yields to $(2n)!/(2^n n!)$ terms in the sum.

By using the results of Isserlis Theorem for computing the value $E\,[z_k\,z_{k+l}]$ in Equation (A12), we have

$$\text{Cov}\;(z_k\;z_{k+l}) = \sum_{h=1}^{w} q_h \sigma_{i_h\;j_h}, \tag{A15}$$

where $|i_h - j_h| \geq l$ and $w \leq (4n)!/(2^{2n}(2n)!)$ considering that the maximum order of monomial z_k and z_{k+l} can each be $2n$. Distance $|i_h - j_h|$ is the distance from the diagonal of covariance matrix of output, which is introduced by Equation (A11), and σ_{ij} is the ijth entry of the covariance matrix of output. Note that $\sigma_{i_h\, j_h}$ is the part of σ_{ij} where $|i_h - j_h| \geq l$, and q_h is the remaining part. Hence, in Eq. (A15) we consider the elements of covariance matrix of y with largest distance as $\sigma_{i_h\, j_h}$, and put the rest as q_h.

Since the system is uniformly exponentially stable, the system impulse response decays exponentially, therefore by going farther from diagonals of the covariance matrix, the entries of covariance matrix of output σ_{ij}s decrease exponentially and distance $|i_h - j_h|$ decays proportionally with the distance from the diagonal.

First, we prove that for $Cov\,(z_k,\ z_{k+l}) = \sum_{h=1}^{w} q_h \sigma_{i_h\, j_h}$ in Equation (A15), we always have $|i_h - j_h| \geq l$. For computing $E\,[z_k\, z_{k+l}]$ in Equation (A12), there are two cases that might happen:

1. The case that time indices of $\sigma_{i_h\, j_h}$ involved in computing $E\,[z_k\, z_{k+l}]$, are always with the interval $|i_h - j_h| \geq l$.
2. The case that time some indices of $\sigma_{i_h\, j_h}$ involved in computing $E\,[z_k\, z_{k+l}]$, are in the interval $|i_h - j_h| < l$.

First case lines with the fact that in computing the expected value of each pair based on Isserlis' theorem, there exists at least one entry of covariance matrix called $\sigma_{i_h\, j_h}$ that the distance $|i_h - j_h| \geq l$. For the second case, if there is no entry with the distance $|i_h - j_h| \geq l$, then that means the entry is separated into the multiplication of terms. In other words, there is one term that is related to the first monomial z_k, and the other term is related to the second monomial z_{k+l}. So that the multiplication of these terms is cancelled by the $E\,[z_k]\,E\,[z_{k+l}]$ term in Equation (A12).

Now that we have proved there is always distance $|i_h - j_h| \geq l$ in computing the $Cov\,(z_k,\ z_{k+l})$ in Equation (A15), it is time to find an upper bound for the $Cov\,(z_k,\ z_{k+l})$ and call it ρ_l.

As we have shown in Equation (A8), $|g(k, m)| \leq M\, a^{(k-m)} \forall k \geq m$, where M is a constant and $0 < a < 1$. Because the distance $|i_h - j_h| \geq l$, then

$$\sigma_{i_h\, j_h} \leq \tilde{M}\ a^l,$$

for some constant $\tilde{M}$. Hence,

$$\text{Cov}\ (z_k,\ z_{k+l}) = \sum_{h=1}^{w} q_h \sigma_{i_h\, j_h}\ \leq C\ a^l, \tag{A16}$$

where C is a constant. Therefore, we pick

$$\rho_l(C, a) = C\ a^l, \tag{A17}$$

where $0 < a < 1$.

Now, we prove

$$\sum_{k=1}^{\infty} \frac{(\text{Var}\ z_k)(\log\ k)^2}{k^2} < \infty \tag{A18}$$

holds. In step (2), we have proved that variance of output monomials $(\text{Var}\, z_k)$ is bounded. Moreover, $\sum_{k=1}^{\infty} \frac{(\log\ k)^2}{k^2}$ is known to be bounded and converges, so Equation (A18) holds.

The last condition of Theorem 2 that we need to prove is

$$\sum_{l=1}^{\infty} \frac{\rho_l}{l^q} < \infty \quad \text{for some}\ \ 0 \leq q < 1. \tag{A19}$$

By considering $\rho_l = C\, a^l$ and $q = 0$, Equation (A19) becomes

$$\sum_{l=1}^{\infty} \frac{\rho_l}{l^q} = \sum_{l=1}^{\infty} \rho_l = \sum_{l=1}^{\infty} C\ a^l < \infty \quad 0 \leq a < 1. \tag{A20}$$

Hence, we have shown that Equation (A19) holds.

4. Strong law of large numbers holds for system output monomials

As it has shown in previous steps, the conditions of Theorem 3 are satisfied for SARX system identification in this article, so

$$\lim_{N\to\infty} \frac{\sum_{k=1}^{N}(X_k - E[X_k])}{k} = 0 \quad \text{a.s.} \tag{A21}$$

In other words, strong law of large numbers holds for the system output and its monomials.

Now that we have proved strong law of large numbers holds for monomials of system output, the direct result is that strong law of large numbers holds for the $M[\text{mon}_n(y_k, \dots, y_{k-n_a}, u_{k-1}, \dots, u_{k-n_b})]$ and accordingly as $N \to \infty$,

$$\frac{1}{N}\sum_{k=1}^{N} M[\text{mon}_n(y_k, \dots, y_{k-n_a}, u_{k-1}, \dots, u_{k-n_b})] - \frac{1}{N}\sum_{k=1}^{N} E[M[\text{mon}_n(y_k, \dots, y_{k-n_a}, u_{k-1}, \dots, u_{k-n_b})]] \to 0 \quad \text{a.s.},$$

which based on notations in Theorem 1, Lemma 4, and Equation (25), it is: as $N \to \infty$,

$$\widehat{\mathcal{M}}_N^{\text{proc}} - \mathcal{M}_N^{\text{proc}} \to 0 \quad \text{a.s.}$$