

## Identification of non-causal systems with random switching modes

Questa è la versione preprint della seguente opera:

*Original*

Identification of non-causal systems with random switching modes / Zhang, Y., Yu, C., Fabiani, F.. - In: AUTOMATICA. - ISSN 1873-2836. - 182:(2025). [10.1016/j.automatica.2025.112532]

*Availability:*

This version is available at: 20.500.11771/36138

*Publisher:*

*Published*

DOI:10.1016/j.automatica.2025.112532

*Terms of use:*

This publication is made accessible in accordance with the terms for deposit in the institutional repository, as defined by the IMT School for Advanced Studies Lucca's Open Access Policy. ([https://library.imtlucca.it/sites/default/files/regolamento-policy-open-access-imtlib\\_0.pdf](https://library.imtlucca.it/sites/default/files/regolamento-policy-open-access-imtlib_0.pdf)).

Si prega di consultare le pagine informative dell'editore relative alle politiche di autoarchiviazione.

(Article begins on next page)

# Identification of non-causal systems with random switching modes—EXTENDED VERSION

Yanxin Zhang<sup>†</sup>, Chengpu Yu<sup>†</sup>, and Filippo Fabiani<sup>‡</sup>

<sup>†</sup> *School of Automation, Beijing Institute of Technology, Beijing 100081, PR China*

<sup>‡</sup> *IMT School for Advanced Studies Lucca, Piazza San Francesco 19, 55100, Lucca, Italy*

---

## Abstract

We consider the identification of non-causal systems with random switching modes (NCS-RSM), a class of models essential for describing typical power load management and department store inventory dynamics. The simultaneous identification of causal-and-anticausal subsystems, along with the presence of random switching sequences, however, make the overall identification problem particularly challenging. To this end, we develop an expectation-maximization (EM) based system identification technique, where the E-step proposes a modified Kalman filter (KF) to estimate the states and switching sequences of causal-and-anticausal subsystems, while the M-step consists in a switching least-squares algorithm to estimate the parameters of individual subsystems. We establish the main convergence features of the proposed identification procedure, also providing bounds on the parameter estimation errors under mild conditions. Finally, the effectiveness of our identification method is validated through two numerical simulations.

*Keywords:* Switching systems; Non-causal systems; Expectation maximization; Kalman filter

---

## 1. Introduction

Non-causal switching dynamics arise in scenarios where actions depend on both historical and future states. In addition, these systems exhibit switching characteristics, potentially transitioning among different operational states, and thus leading to variations in the system behavior. In power systems, for instance, load management requires dedicated adjustments based on future demand [1]. A controller can thus activate different modes to reduce load if a surge is anticipated, creating a dependency on future state. In traffic signal control [2], adaptive signal timing utilizes real-time vehicle flow predictions and synchronous historical/future data to enable autonomous phase adjustments without external centralized control. Its non-causal dependencies effectively prevent congestion propagation while enhancing traffic efficiency. In robotic systems [3], collaborative robots may require anticipatory motion planning to avoid collisions, where future positions of other agents influence current decisions. In supply chain management, inventory provision decisions often depend on future demand and supplier lead times, creating non-causal dependencies between current actions and future states [4]. Financial time series data often exhibit characteristics of sharp peaks and heavy tails. For instance, stock trading volumes may exhibit abnormal fluctuations prior to the release of significant announcements, which can be interpreted as the influence of future data on current observations [5]. These systems

exhibit switching behaviors due to discrete operational mode transitions (e.g., emergency shutdowns in power grids, traffic signal phase shifts, the anticipatory motion planning in collaborative robots), and their non-causality stems from the need to model feedback loops with delayed effects or predictive decision-making. Understanding and managing the complexity of these systems is therefore crucial for enhancing efficiency, reliability, and adaptability, enabling them to better meet the demands of industrial production and operations. This essentially motivates the interest in modeling, analyzing, and controlling such type of systems.

In several identification problems for dynamical systems, the input-output data are accompanied by temporal mode sequences. As the system’s mode changes over time, each data point is associated with the active mode at that specific time. Hence, it is crucial to model the dynamics of different modes and infer transitions between modes [7]. However, obtaining direct estimates of the dynamical system from input-output data is challenging and, in practice, prior knowledge about mode transitions is often unavailable. Therefore, estimating the switching behaviors poses a challenging yet highly significant problem that has attracted attention from researchers. Existing studies indeed propose algorithms to identify individual system dynamics and mode transition sequences from observed behaviors [8].

### 1.1. Literature review

Several works consider the identification of switching models [9, 10]. Among various switching system models, jump Markov linear systems (JMLS) have emerged as a powerful framework for capturing abrupt random behavioral changes. These systems utilize a probabilistic structure in which mode transitions follow a Markov chain, effectively modeling stochastic switching dynamics through discrete state transitions with memoryless properties. In [11], a joint smoothing algorithm based on the expectation-maximization (EM) framework is proposed, with an E-step solution introduced to address exponential complexity in the JMLS. In [12], a numerically efficient two-step estimation method is developed, iteratively updating parameters and the switching sequence. The flexibility of this technique lies in its adaptability to various loss functions used in jump models, which significantly influence their shape and switching behavior. Furthermore, the identification of jump Box–Jenkins (BJ) models is investigated in [13]. These models consist of a finite collection of linear dynamical submodels in BJ form, switching over time according to a Markov chain. The system parameters are estimated iteratively using the Gauss-Newton and prediction error methods. In [14], a switching least-squares algorithm for autonomous Markov jump linear systems is proposed. Here, the authors provide a formal proof of the method’s strong consistency and establish its convergence rate as  $\mathcal{O}(\sqrt{\log(T)/T})$ , where  $T$  is the time horizon. While existing literature primarily focuses on linear systems with Markov switching, these techniques may fail when mode transitions exhibit non-Markovian randomness. To address this limitation, methods for systems with random switching behavior have been proposed. For instance, [15] employs a kernel-based approach to simultaneously solve estimation and classification problems in random switching systems. Similarly, [16] proposes a maximum-likelihood algorithm combining Kalman filtering and likelihood estimation to stabilize error convergence in general switched linear systems.

Furthermore, Gaussian mixture models (GMMs) served as a cornerstone for identifying switching systems [17], where EM-based algorithms are widely adopted to estimate latent modes and subsystem parameters [18]. However, classical GMM frameworks are inherently

limited to causal dynamics with single switching sequences, failing to address systems governed by bidirectional dependencies and dual independent switching behaviors. Our work extends GMM principles to non-causal switching dynamics, where outputs depend on both historical and future states through two distinct switching sequences. However, to the best of our knowledge, there is no literature on the system identification problem of non-causal systems with random switching modes (NCS-RSM). Although, there are some studies available on system identification for non-causal systems, such as the subspace [19] and the kernel methods [20, 21], these studies can only handle a single, non-causal system, rather than *switching non-causal systems*.

### 1.2. Summary of contribution

In this paper, we focus on the identification of NCS-RSM. The proposed method is developed under the expectation-maximization (EM) framework, which can be divided into two main parts. Specifically, in the E-step we adopt a Bayesian rule to compute the posterior estimate of the switching sequence, along with a modified Kalman filter (KF) for estimating the state of the causal and non-causal parts. In the M-step, instead, we propose a switching least-squares method to obtain the closed-form solution for the parameters and establish the convergence rate of the estimated parameters. Our main contributions can hence be summarized as follows:

1. To the best of our knowledge, this is the first work considering the identification of NCS-RSM. To contrast with causal systems [12, 13], we propose a modified Kalman filter (KF) with bidirectional estimation (forward for causal states, backward for non-causal ones) to resolve the non-causal dependency.
2. Unlike prior works [11, 14], which focus on causal systems with Markov switching, our method explicitly handles bidirectional switching dynamics (causal and non-causal subsystems) and random (non-Markov) switching sequences. Moreover, the switching sequences of the two directions is allowed to differ from each other. This enables modeling real-world scenarios where outputs depend on both past and future states.
3. We show that our method has a  $\mathcal{O}(\sqrt{\log(T)/T})$  rate of convergence under the assumptions of average stability and martingale difference noise. This result is generalized from causal Markov systems [14] to non-causal random ones.

### 1.3. Paper organization

The rest of the paper is organized as follows: in Section 2 we describe the considered system and formulate the related identification problem. In Section 3, instead, we discuss our EM method for the identification of the NCS-RSM, while in Section 4 we provide its implementation details, as well as characterize the related convergence properties. Two simulation examples are finally discussed in Section 5 to test the effectiveness of the proposed method numerically. The proofs of the technical results of the paper are all deferred to Appendix Appendix A.

*Notation:*  $\mathbb{Z}$  and  $\mathbb{R}$  denote the set of integer and real numbers, respectively. Given a matrix  $X$ ,  $\|X\|$  and  $\|X\|_\infty$  denote respectively its spectral and infinity norms, and  $\text{tr}(X)$  denotes the trace. For a real symmetric matrix  $P$ ,  $\lambda_{\max}(P)$  and  $\lambda_{\min}(P)$  are respectively its maximum and minimum eigenvalues.  $\mathbb{P}[\cdot]$  and  $\mathbb{E}[\cdot]$  respectively denote a probability distribution and the related expected value.  $\mathbb{P}_\theta[\cdot]$  denotes the probability density function with  $\theta$  as parameters.  $\mathbb{E}_\theta[\cdot]$  denotes the expectation operator with respect to the distribution  $\mathbb{P}_\theta[\cdot]$ .  $\mathbb{S}^n$  is the space of  $n \times n$  symmetric matrices and  $\mathbb{S}_{>0}^n$  ( $\mathbb{S}_{\geq 0}^n$ ) is the cone of positive

(semi-)definite matrices. Given two square matrices  $A, B$  of compatible dimension,  $A \succcurlyeq B$  means that  $A - B$  is positive semidefinite. For a random sequence  $\{x(t)\}_{t \geq 1}$ ,  $x(1:t)$  denotes  $\{x(1), \dots, x(t)\}$ ,  $\sigma(x(1:t))$  denotes the sigma field generated by random variables  $x(1:t)$ . For a sequence  $\{s_t\}_{t \in \mathbb{N}}$ ,  $s_T = \mathcal{O}(T)$  indicates that  $\limsup_{T \rightarrow \infty} s_T/T < \infty$ , while  $s_T = o(T)$  that  $\limsup_{T \rightarrow \infty} s_T/T = 0$ . Finally,  $I$  identifies a standard identity matrix.  $\mathcal{N}(\mu, \Sigma^2)$  denotes the normal distribution of a random variable with mean  $\mu$  and standard deviation  $\Sigma$ . In the remainder we will use Standing Assumption to postulate properties that hold throughout the paper.

## 2. Mathematical formulation

We now describe the system concerned in this paper, together with the main assumptions, and successively formalize the problem to be addressed.

### 2.1. System model description

Consider the following discrete-time, non-causal system characterized by random switching modes:

$$x_c(t) = A_c(s_c(t))x_c(t-1) + v_c(t), \quad (1a)$$

$$x_a(t) = A_a(s_a(t))x_a(t+1) + v_a(t), \quad (1b)$$

$$y(t) = C_c(s_c(t))x_c(t) + C_a(s_a(t))x_a(t) + v_m(t), \quad (1c)$$

where  $t \in \mathbb{Z}$  is the time instant,  $x_c(t) \in \mathbb{R}^{n_{x_c}}$ ,  $x_a(t) \in \mathbb{R}^{n_{x_a}}$  are the causal and non-causal state vectors, respectively,  $y(t) \in \mathbb{R}^{n_y}$  denotes the system output, while  $s_c(t) \in \{1, 2, \dots, m_c\} \triangleq \Lambda_c$  and  $s_a(t) \in \{1, 2, \dots, m_a\} \triangleq \Lambda_a$  are two discrete variables representing the possible switching modes. In addition,  $v_c(t) \in \mathbb{R}^{n_{x_c}}$  and  $v_a(t) \in \mathbb{R}^{n_{x_a}}$  are the system noise vectors, and  $v_m(t) \in \mathbb{R}^{n_y}$  is the measurement noise vector. Finally,  $A_c : \Lambda_c \rightarrow \mathbb{R}^{n_{x_c} \times n_{x_c}}$  and  $A_a : \Lambda_a \rightarrow \mathbb{R}^{n_{x_a} \times n_{x_a}}$  denote the matrix functions associated to the causal and non-causal state dynamics, respectively, while  $C_c : \Lambda_c \rightarrow \mathbb{R}^{n_y \times n_{x_c}}$  and  $C_a : \Lambda_a \rightarrow \mathbb{R}^{n_y \times n_{x_a}}$  are those mapping the two state vectors to the measured output. Assume that the noise terms  $v_c(t)$ ,  $v_a(t)$  and  $v_m(t)$  are distributed according to a Gaussian distribution with zero mean and finite variance  $v_c(t) \sim \mathcal{N}(0, \Sigma_c(s_c(t)))$ ,  $v_a(t) \sim \mathcal{N}(0, \Sigma_a(s_a(t)))$ ,  $v_m(t) \sim \mathcal{N}(0, \Sigma_m)$ .

**Standing Assumption 1.** *The NCS-RSM (1) is stable in the average sense, which means*

$$\limsup_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^T \|x_c(t)\|^2 < \infty$$

$$\limsup_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^T \|x_a(T+1-t)\|^2 < \infty$$

where  $T$  is the sample size of the available dataset.

As a main consequence of Standing Assumption 1, we note that the underlying stability requirement implies the existence of a stationary distribution for both causal and non-causal states, thereby ensuring the ergodicity of the system dynamics.

**Remark 1.** *Stability in the average sense is widely applied in linear systems [14, 22, 23]. Compared to other commonly used notions (e.g., mean-square stability [24], which requires  $\lim_{T \rightarrow \infty} \mathbb{E}[\|x(T)\|^2] < \infty$ , and almost sure (a.s.) stability [25], ensuring  $\mathbb{P}[\lim_{T \rightarrow \infty} \|x(T)\| = 0] = 1$ ).*

0] = 1), the assumption of stability in the average sense is weaker [14]. Specifically, it only imposes convergence in a time-averaged sense, avoiding stricter pointwise or moment-based constraints.

Assume the sample size is  $T$ . We denote the sigma field as  $\mathcal{G}_t = \sigma(x_c(1:t))$  for the causal part,  $\mathcal{G}'_t = \sigma(x_a(T:T+1-t))$  for the non-causal part, and  $\mathcal{G}_t^* = \sigma(x_c(1:t), x_a(1:t))$  for the measured part.

**Standing Assumption 2.** *The sequences of noise vectors  $v_c(1:t), v_a(T:T+1-t), v_m(1:t)$  are the martingale difference sequences with respect to the filtrations  $\{\mathcal{G}_t\}_{t \geq 1}, \{\mathcal{G}'_t\}_{t \geq 1}, \{\mathcal{G}_t^*\}_{t \geq 1}$ , respectively, i.e.  $\mathbb{E}[v_c(t)|\mathcal{G}_{t-1}] = 0, \mathbb{E}[v_a(t)|\mathcal{G}'_{t-1}] = 0, \mathbb{E}[v_m(t)|\mathcal{G}_{t-1}^*] = 0$  and satisfy the following conditions:*

$$\liminf_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^T v_k(t)v_k(t)^\top \succ 0, \quad k \in \{c, m\},$$

$$\liminf_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^T v_a(T+1-t)v_a(T+1-t)^\top \succ 0.$$

**Remark 2.** *A martingale is a sequence of random variables for which, at a particular time of the realized sequence, the expectation of the next value in the sequence is equal to the present observed value even given knowledge of all prior observed values. Standing Assumption 2 denotes a common requirement for analyzing the convergence of system identification algorithms, enabling the noise process to exhibit non-stationary and heavy-tailed characteristics—see, e.g., [29, 30, 31].*

The NCS-RSM in (1) thus consists of two state equations and one output equation. Specifically, the first state equation represents the dynamics of the causal state variables, while the second one the dynamics of the non-causal state variables. The system output is determined by both the causal and non-causal states. Furthermore, both the causal and non-causal parts of the system are composed of multiple subsystems, and their corresponding switching sequences are different. Given some  $T \in \mathbb{Z}$ , which will denote the sample size of the available dataset, let the switching sequences of the causal and non-causal parts being denoted by  $\mathbf{s}_c \triangleq \{s_c(t)\}_{t=1}^T$  and  $\mathbf{s}_a \triangleq \{s_a(t)\}_{t=1}^T$ , respectively. Each of them corresponds to a set of parameters, i.e.,  $s_c(t) = i$  determines the model parameter  $\theta_i^c \triangleq \{A_c(i), C_c(i), \Sigma_c(i)\}$  that is active at the time instant  $t$ . In particular, the sequences  $\mathbf{s}_c$  and  $\mathbf{s}_a$  undergo random switches with certain (fixed) probabilities over time. Then, we have the following assumption:

**Standing Assumption 3.** *The following conditions hold true:*

1. *The switching sequences  $\mathbf{s}_c, \mathbf{s}_a$ , and the subsystem parameters  $\theta^c, \theta^a$  are all independent among them, i.e.,  $\mathbb{P}[\mathbf{s}_c|\theta^c] = \mathbb{P}[\mathbf{s}_c], \mathbb{P}[\theta^c|\mathbf{s}_c] = \mathbb{P}[\theta^c], \mathbb{P}[\mathbf{s}_a|\theta^a] = \mathbb{P}[\mathbf{s}_a], \mathbb{P}[\theta^a|\mathbf{s}_a] = \mathbb{P}[\theta^a]$ .*
2. *The switching sequence follows a multinomial distribution, i.e., for any  $t$ , we have  $\mathbb{P}[s_c(t) = i] = \pi_i^c, i = 1, \dots, m_c, \mathbb{P}[s_a(t) = i] = \pi_i^a, i = 1, \dots, m_a$ , with  $\sum_{i=1}^{m_c} \pi_i^c = 1, \sum_{i=1}^{m_a} \pi_i^a = 1$ .*

The complete set of model parameters that comprehensively describe the NCS-RSM can be conveniently encapsulated into a parameter object  $\theta$ , defined as follows:

$$\theta \triangleq \{ \{ \theta_i^c \}_{i=1}^{m_c}, \{ \theta_i^a \}_{i=1}^{m_a}, \{ \pi_i^c \}_{i=1}^{m_c}, \{ \pi_i^a \}_{i=1}^{m_a}, \Sigma_m \}.$$

## 2.2. Problem statement

Our goal is hence to estimate the *unknown* model parameters  $\theta$  characterizing the NCS-RSM (1) with the known state dimension, the initial states  $x_c(0)$ ,  $x_a(T + 1)$ , number of causal system modes  $m_c$  and non-causal system modes  $m_a$ , together with a collection of noisy output measurements  $\mathbf{y}$ :

$$\mathbf{y} \triangleq \mathbf{y}_{1:T} = \{y(1), \dots, y(T)\}.$$

The NCS-RSM system contains both causal and non-causal components with distinct switching sequences. Addressing this problem faces two main challenges. First, the output depends on both causal and non-causal states, which are unmeasurable. Simultaneously identifying parameters for all subsystems is challenging due to their continuous switching patterns. Second, the system has two independent switching sequences. For example, at time  $t$ , the causal component might activate subsystem  $i$  while the non-causal component activates subsystem  $j$ , creating  $m_a \times m_c$  possible combinations.

In addition, the switching behavior of the subsystems is random and independent across different time instants, i.e.,  $\mathbb{P}[s_c(t)|s_c(t-1), \dots, s_c(1)] = \mathbb{P}[s_c(t)]$ ,  $\mathbb{P}[s_a(t)|s_a(t+1), \dots, s_a(T)] = \mathbb{P}[s_a(t)]$ ,  $t = 1, \dots, T$ . To deal with the identification problem of the NCS-RSM (1), the EM framework is adopted, which is an iterative method that can yield an estimate of the parameters at each iteration [26]. Let us denote the parameter estimate at the  $k$ -th iteration of the underlying algorithm as  $\theta^k$ . Then, the proposed method can be (qualitatively, for the moment) described by means of the following two steps:

1. In the E-step, we develop a modified KF to estimate the states of the causal and non-causal parts. Furthermore, the Bayesian rule is used to obtain a posterior estimate of the switching sequence. Subsequently, the full-data likelihood function  $Q(\theta, \theta^k)$  can be calculated.
2. In the M-step, the likelihood function  $Q(\theta, \theta^k)$  is maximized with respect to the parameters  $\theta$ . Then, the identification of the NCS-RSM is updated, yielding  $\theta^{k+1}$ .

Next section will discuss in detail each step of the proposed technique for NCS-RSM identification.

## 3. The EM method for identifying NCS-RSM

By making use of the dataset  $\mathbf{y}$ , we aim at estimating the system parameters  $\theta$ . To this end, a standard approach is to let coincide  $\hat{\theta}$ , i.e., our estimate of the true  $\theta$ , with a maximizer of the likelihood function, namely:

$$\hat{\theta} \in \arg \max_{\theta} \ln \mathbb{P}_{\theta}(\mathbf{y}) \text{ s.t. (1),} \quad (2)$$

where we indicate with  $\mathbb{P}_{\theta}(\mathbf{y})$  the probability density function of the output  $\mathbf{y}$  given some sets of parameters  $\theta$ . Note that in switching systems, the likelihood function may exhibit multiple equivalent maxima due to the interchangeability of subsystem parameters (e.g., permuting subsystem labels can yield identical likelihood values). However, these equivalent solutions do not affect the system's physical properties. During model validation, we only need to compare whether the subsystem dynamic responses remain consistent across different permutations.

Let us denote the collection of state variables over  $T$  as  $\mathbf{x}_c \triangleq \{x_c(t)\}_{t=1}^T$  and  $\mathbf{x}_a \triangleq \{x_a(t)\}_{t=1}^T$ . Given any collection of data  $\mathbf{y}$ , note that the likelihood function  $\ln \mathbb{P}_\theta(\mathbf{y})$ , also called marginal density function of  $\mathbf{y}$ , can be decomposed into the following form:

$$\begin{aligned} \ln \mathbb{P}_\theta(\mathbf{y}) &= \ln \sum_{\mathbf{s}_c} \sum_{\mathbf{s}_a} \int \int \mathbb{P}_\theta[\mathbf{y}|\mathbf{x}_a, \mathbf{x}_c, \mathbf{s}_a, \mathbf{s}_c] \\ &\quad \mathbb{P}_\theta[\mathbf{x}_a, \mathbf{x}_c, \mathbf{s}_a, \mathbf{s}_c] d\mathbf{x}_a d\mathbf{x}_c \\ &= \ln \sum_{\mathbf{s}_c} \sum_{\mathbf{s}_a} \int \int \mathbb{P}_\theta[\mathbf{y}|\mathbf{x}_a, \mathbf{x}_c, \mathbf{s}_a, \mathbf{s}_c] \\ &\quad \mathbb{P}_\theta[\mathbf{x}_a, \mathbf{s}_a] \mathbb{P}_\theta[\mathbf{x}_c, \mathbf{s}_c] d\mathbf{x}_a d\mathbf{x}_c \end{aligned} \quad (3)$$

where the first equality follows from the law of total probability, which expands the marginal log-likelihood  $\ln \mathbb{P}_\theta[\mathbf{y}]$  by marginalizing over all latent variables. The second equality stems from the mutual independence between the causal and non-causal subsystems. Since the states of both subsystems obey to Markovian dynamics and the outputs are conditionally independent given the switching mode, states and mode parameters, the posterior probability density function in (3) can be decomposed into the following structured forms:

$$\begin{aligned} \mathbb{P}_\theta[\mathbf{y}|\mathbf{x}_a, \mathbf{x}_c, \mathbf{s}_a, \mathbf{s}_c] &= \prod_{t=1}^T \mathbb{P}_\theta[y(t)|x_c(t), x_a(t), s_c(t), s_a(t)] \\ \mathbb{P}_\theta[\mathbf{x}_c, \mathbf{s}_c] &= \mathbb{P}_\theta[x_c(1), s_c(1)] \\ &\quad \prod_{t=2}^T \mathbb{P}_\theta[x_c(t)|x_c(t-1), s_c(t)] \mathbb{P}_\theta[s_c(t)] \\ \mathbb{P}_\theta[\mathbf{x}_a, \mathbf{s}_a] &= \mathbb{P}_\theta[x_a(T), s_a(T)] \\ &\quad \prod_{t=1}^{T-1} \mathbb{P}_\theta[x_a(t)|x_a(t+1), s_a(t)] \mathbb{P}_\theta[s_a(t)]. \end{aligned} \quad (4)$$

In the NCS-RSM (1), the state variables  $\mathbf{x}_c, \mathbf{x}_a$  are governed by the switching sequences  $\mathbf{s}_c, \mathbf{s}_a$ . Direct maximization of the marginal log-likelihood  $\ln \mathbb{P}_\theta[\mathbf{y}]$  is inherently challenging due to its nonconvexity and high-dimensional nature. Furthermore, as shown in the decomposition (3), evaluating  $\mathbb{P}_\theta[\mathbf{y}]$  requires summation over all possible realizations of  $\mathbf{s}_c, \mathbf{s}_a$ , which exponentially increases the computational complexity of solving (2).

Another way to marginalize the latent variables (such as  $\mathbf{x}_c, \mathbf{x}_a, \mathbf{s}_c, \mathbf{s}_a$ ) is by taking the expectation over these latter. Instead of maximizing the incomplete likelihood function  $\ln \mathbb{P}_\theta(\mathbf{y})$ , we can estimate the conditional density of the hidden variables given the observations  $\mathbf{y}$  and an estimate of parameter  $\hat{\theta}$ . Then, parameter estimate  $\hat{\theta}$  can be obtained by maximizing the complete likelihood function. The full-data complete likelihood function can be expressed as follows:

$$\begin{aligned} \ln \mathbb{P}_\theta[\mathbf{y}, \mathbf{x}_c, \mathbf{s}_c, \mathbf{x}_a, \mathbf{s}_a] &= \ln \mathbb{P}_\theta[\mathbf{y}] \\ &\quad + \ln \mathbb{P}_\theta[\mathbf{x}_c, \mathbf{s}_c, \mathbf{x}_a, \mathbf{s}_a|\mathbf{y}]. \end{aligned} \quad (5)$$

This relation directly links  $\mathbb{P}_\theta(\mathbf{y})$  and  $\mathbb{P}_\theta[\mathbf{y}, \mathbf{x}_c, \mathbf{s}_c, \mathbf{x}_a, \mathbf{s}_a]$ , with the latter depending on the unknown states  $\mathbf{x}_c, \mathbf{x}_a$  and switching sequences  $\mathbf{s}_c, \mathbf{s}_a$ . The key step is then to approximate  $\ln \mathbb{P}_\theta[\mathbf{y}]$  by the above relation (5), where  $\mathbf{x}_c, \mathbf{s}_c, \mathbf{x}_a$ , and  $\mathbf{s}_a$  can be approximated

by their conditional expectations based on the observed data  $\mathbf{y}$ . Therefore, at each iteration  $k$  of our EM-based algorithm, given the estimate  $\theta^k$ , the conditional expectation of  $\ln \mathbb{P}_{\theta^k}[\mathbf{x}_c, \mathbf{s}_c, \mathbf{x}_a, \mathbf{s}_a | \mathbf{y}]$  is abbreviated as:

$$\mathbb{E}_{\theta^k}[\cdot] = \int \int \sum_{\mathbf{s}_c} \sum_{\mathbf{s}_a} (\cdot) \mathbb{P}_{\theta^k}[\mathbf{x}_c, \mathbf{s}_c, \mathbf{x}_a, \mathbf{s}_a | \mathbf{y}] d(\mathbf{x}_c) d(\mathbf{x}_a). \quad (6)$$

Then, by applying the expectation operator  $\mathbb{E}_{\theta^k}[\cdot]$  to both sides of (5), one obtains:

$$\begin{aligned} \mathbb{E}_{\theta^k}[\ln \mathbb{P}_{\theta}[\mathbf{y}, \mathbf{x}_c, \mathbf{s}_c, \mathbf{x}_a, \mathbf{s}_a]] \\ &= \mathbb{E}_{\theta^k}[\ln \mathbb{P}_{\theta}(\mathbf{y})] + \mathbb{E}_{\theta^k}[\ln \mathbb{P}_{\theta}[\mathbf{x}_c, \mathbf{s}_c, \mathbf{x}_a, \mathbf{s}_a | \mathbf{y}]] \\ &= \ln \mathbb{P}_{\theta}(\mathbf{y}) + \mathbb{E}_{\theta^k}[\ln \mathbb{P}_{\theta}[\mathbf{x}_c, \mathbf{s}_c, \mathbf{x}_a, \mathbf{s}_a | \mathbf{y}]]. \end{aligned}$$

Define  $Q(\theta, \theta^k) = \mathbb{E}_{\theta^k}[\ln \mathbb{P}_{\theta}[\mathbf{y}, \mathbf{x}_c, \mathbf{s}_c, \mathbf{x}_a, \mathbf{s}_a]]$ ,  $V(\theta, \theta^k) = \mathbb{E}_{\theta^k}[\ln \mathbb{P}_{\theta}[\mathbf{x}_c, \mathbf{s}_c, \mathbf{x}_a, \mathbf{s}_a | \mathbf{y}]]$ . The EM approach iteratively estimates the parameters in the following two steps. First, we compute the expectation  $Q(\theta, \theta^k)$  based on  $\theta^k$  obtained from the previous iteration. Under Standing Assumption 3, the full-data complete likelihood function can be decomposed by using the Bayesian rule, which is given as follows:

$$\begin{aligned} \ln \mathbb{P}_{\theta}[\mathbf{y}, \mathbf{x}_c, \mathbf{s}_c, \mathbf{x}_a, \mathbf{s}_a] &= \ln \mathbb{P}_{\theta}[\mathbf{y} | \mathbf{x}_c, \mathbf{s}_c, \mathbf{x}_a, \mathbf{s}_a] \\ &\quad + \ln \mathbb{P}_{\theta}[\mathbf{x}_c, \mathbf{s}_c] + \ln \mathbb{P}_{\theta}[\mathbf{x}_a, \mathbf{s}_a] \end{aligned} \quad (7)$$

where the posterior probabilities are shown in (4). In view of the white noise assumption characterizing the disturbance affecting both state variables and measured output, note that the distribution of these variables, given the subsystem modes  $s_c(t) = j$ ,  $s_a(t) = l$ , is Gaussian too and given as follows:

$$\begin{aligned} \mathbb{P}_{\theta}[y(t) | x_c(t), x_a(t), s_c(t) = j, s_a(t) = l] &= |2\pi \Sigma_m|^{-1/2} \\ &\quad \exp\{(y(t) - \mu_1(t, j, l))^{\top} \Sigma_m^{-1} (y(t) - \mu_1(t, j, l))\}, \\ \mathbb{P}_{\theta}[x_c(t) | x_c(t-1), s_c(t) = j] &= |2\pi \Sigma_c(t)|^{-1/2} \\ &\quad \exp\{(x_c(t) - \mu_2(t, j))^{\top} \Sigma_c^{-1}(t) (x_c(t) - \mu_2(t, j))\}, \\ \mathbb{P}_{\theta}[x_a(t) | x_a(t+1), s_a(t) = l] &= |2\pi \Sigma_a(t)|^{-1/2} \\ &\quad \exp\{(x_a(t) - \mu_3(t, l))^{\top} \Sigma_a^{-1}(t) (x_a(t) - \mu_3(t, l))\}, \end{aligned} \quad (8)$$

where  $\mu_1(t, j, l) = C_c(j)x_c(t) + C_a(l)x_a(t)$ ,  $\mu_2(t, j) = A_c(j)x_c(t-1)$ ,  $\mu_3(t, l) = A_a(l)x_a(t+1)$ . Denote  $w_{ij}^c$  as the posterior probability of the switching sequence given the  $\theta^k$ , dataset  $\mathbf{y}$  and  $s_c(t) = j$  ( $w_{il}^a$  is defined similarly)

$$\begin{aligned} w_{ij}^c &= \mathbb{P}_{\theta^k}[s_c(t) = j | \mathbf{y}] = \frac{\mathbb{P}_{\theta^k}[\mathbf{y} | s_c(t) = j] \pi_j^c}{\sum_{t=1}^T \mathbb{P}_{\theta^k}[\mathbf{y} | s_c(t) = j] \pi_j^c}, \\ w_{il}^a &= \mathbb{P}_{\theta^k}[s_a(t) = l | \mathbf{y}] = \frac{\mathbb{P}_{\theta^k}[\mathbf{y} | s_a(t) = l] \pi_l^a}{\sum_{t=1}^T \mathbb{P}_{\theta^k}[\mathbf{y} | s_a(t) = l] \pi_l^a}. \end{aligned} \quad (9)$$

Then, the objective function  $Q(\theta, \theta^k)$  can be computed by using together (4), (6), (7), (8), and (9):

$$\mathbb{E}_{\theta^k}[\ln \mathbb{P}_{\theta}[\mathbf{y}, \mathbf{x}_c, \mathbf{s}_c, \mathbf{x}_a, \mathbf{s}_a]] = \sum_{i=1}^3 Q_i(\theta, \theta^k), \quad (10)$$

where

$$\begin{aligned}
Q_1(\theta, \theta^k) &= \mathbb{E}_{\theta^k}[\ln \mathbb{P}_\theta[\mathbf{y} | \mathbf{x}_c, \mathbf{s}_c, \mathbf{x}_a, \mathbf{s}_a]] \\
&= \sum_{t=1}^T \sum_{j=1}^{m_c} \sum_{l=1}^{m_a} w_{tj}^c w_{tl}^a \mathbb{E}_{\theta^k}[\ln \mathcal{N}(\mu_1(t, j, l), \Sigma_m)] \\
Q_2(\theta, \theta^k) &= \mathbb{E}_{\theta^k}[\ln \mathbb{P}_\theta[\mathbf{x}_c, \mathbf{s}_c]] \\
&= \sum_{t=1}^T \sum_{j=1}^{m_c} w_{tj}^c \mathbb{E}_{\theta^k}[\ln \mathcal{N}(\mu_2(t, j), \Sigma_c(j))] + \sum_{j=1}^{m_c} \ln \pi_j^c \sum_{t=1}^T w_{tj}^c \\
Q_3(\theta, \theta^k) &= \mathbb{E}_{\theta^k}[\ln \mathbb{P}_\theta[\mathbf{x}_a, \mathbf{s}_a]] \\
&= \sum_{t=1}^T \sum_{l=1}^{m_a} w_{tl}^a \mathbb{E}_{\theta^k}[\ln \mathcal{N}(\mu_3(t, l), \Sigma_a(l))] + \sum_{l=1}^{m_a} \ln \pi_l^a \sum_{t=1}^T w_{tl}^a. \tag{11}
\end{aligned}$$

Subsequently, the second step is to maximize the  $Q(\theta, \theta^k)$  to obtain  $\theta^{k+1}$ , formally defined as  $\theta^{k+1} = \arg \max_{\theta} Q(\theta, \theta^k)$ .

Algorithm 1 summarizes the two main steps of the proposed identification methodology for NCS-RSM. We characterize next the monotonic properties of the likelihood function in (2) when the EM algorithm is iteratively applied to estimate the system parameters  $\theta$ :

**Lemma 1.** *Given a dataset  $\mathbf{y}$ , let  $\{\theta^k\}_{k \in \mathbb{Z}}$  be the sequence generated by Algorithm 1. Then, the likelihood function in (2), evaluated along  $\{\theta^k\}_{k \in \mathbb{Z}}$ , is non-decreasing, thereby yielding  $\ln \mathbb{P}_{\theta^{k+1}}[\mathbf{y}] \geq \ln \mathbb{P}_{\theta^k}[\mathbf{y}]$  for all  $k \in \mathbb{Z}$ .*

The proof of Lemma 1 is shown in Appendix A.

---

**Algorithm 1** EM-based identification of NCS-RSM

---

**Initialization:** Collect data  $\mathbf{y}_{1:T}$ , set  $\theta^0$

**Iteration**  $k \in \mathbb{Z}$ :

1. **E-step:** Compute  $Q(\theta, \theta^k)$  using (9), (10), (11)
  2. **M-step:** Set  $\theta^{k+1} = \arg \max_{\theta} Q(\theta, \theta^k)$
- 

#### 4. Implementation details of the EM algorithm

We now delve into the details of the steps outlined in Algorithm 1, ultimately establishing our main technical result characterizing the sample complexity of the proposed identification technique for NCS-RSM.

##### 4.1. The E-step

This step requires the calculation of the objective function  $Q(\theta, \theta^k)$ . Specifically, this shall be achieved on the basis of the parameter  $\theta^k$  estimated in the previous iteration. Then, according to the expression of  $Q(\theta, \theta^k)$  in (10), the expectations of states  $\mathbf{x}_c, \mathbf{x}_a$  and the switching sequences  $\mathbf{s}_c, \mathbf{s}_a$  given the data  $\mathbf{y}$  are required.

First, we calculate the posterior estimates of the switching sequences  $\mathbf{s}_c$  and  $\mathbf{s}_a$  by leveraging the Bayesian rule, namely  $\mathbb{P}_\theta[s_c(t)|y(t)] = \mathbb{P}_\theta[s_c(t), y(t)]/\mathbb{P}_\theta[y(t)]$  and  $\mathbb{P}_\theta[s_a(t)|y(t)] =$

$\mathbb{P}_\theta[s_a(t), y(t)]/\mathbb{P}_\theta[y(t)]$ . In addition, according to the formula of total probability one obtains  $\mathbb{P}_\theta[y(t)] = \sum_{j=1}^{m_c} \mathbb{P}_\theta[y(t)|s_c(t) = j]\pi_j^c$ ,  $\mathbb{P}_\theta[y(t)] = \sum_{l=1}^{m_a} \mathbb{P}_\theta[y(t)|s_a(t) = l]\pi_l^a$ .

Then, the data point can be assigned to each subsystem at time  $i$  by solving the following optimization problem:

$$\begin{aligned}\hat{s}_c(t) &= \arg \max_{j \in \{1, \dots, m_c\}} \mathbb{P}_\theta[y(t)|s_c(t) = j]\pi_j^c, \\ \hat{s}_a(t) &= \arg \max_{l \in \{1, \dots, m_a\}} \mathbb{P}_\theta[y(t)|s_a(t) = l]\pi_l^a,\end{aligned}$$

where maximizing  $\mathbb{P}_\theta[y(t)|s_c(t) = j]\pi_j^c$  is equivalent to maximizing the posterior probability of  $\mathbb{P}_\theta[s_c(t) = j|y(t)]$  which is commonly used for data classification. In this work, we make use of a hard assignment variant of the EM algorithm as in [27]. This approach approximates the posterior distribution of the switching sequence by its maximum a posteriori (MAP) estimate, effectively collapsing the posterior probability to a Dirac delta function. Such a simplification is justified under the assumption that the posterior distribution  $\mathbb{P}_\theta[s_c(t) = j|\mathbf{y}]$  is sharply peaked around the MAP estimate  $\hat{s}_c(t)$ , which often holds when switching probabilities are highly concentrated (the same holds for the non-causal part). Similar approaches have been adopted in [28] for switching systems, where hard assignments reduce computational complexity while preserving estimation accuracy. Following the hard EM framework, the posterior probabilities  $w_{ij}^c, w_{il}^a$  are approximated by their MAP estimates, resulting in a binary assignment for any  $(t, j) \in \{1, \dots, T\} \times \{1, \dots, m_c\}$  (or  $(t, l) \in \{1, \dots, T\} \times \{1, \dots, m_a\}$ ):

$$w_{ij}^c = \begin{cases} 1 & \text{if } \hat{s}_c(t) = j \\ 0 & \text{else} \end{cases}, \quad w_{il}^a = \begin{cases} 1 & \text{if } \hat{s}_a(t) = l \\ 0 & \text{else} \end{cases}$$

Successively, we focus on the reconstruction of the state variables  $\mathbf{x}_c$  and  $\mathbf{x}_a$ , a task that is traditionally accomplished by means of a Kalman filter. Adapting the KF to our problem, however, requires few key modifications due to the dynamics in (1). When correcting the prior prediction of the state variables  $\mathbf{x}_c$  and  $\mathbf{x}_a$  using the data  $\mathbf{y}$ , special consideration must be given to their cross-correlation structure, necessitating a careful design of the KF as outlined below. To simplify notation, we omit the dependency on the switching sequence, e.g.,  $A_c = A_c(\hat{s}_c(t))$ .

First, we need to compute the prior state estimates of  $\mathbf{x}_c$  and  $\mathbf{x}_a$ , denoted as  $\hat{\mathbf{x}}_c^-$  and  $\hat{\mathbf{x}}_a^-$ . The prior estimates are derived from the first two relations in (1) as  $\hat{x}_c^-(t) = A_c \hat{x}_c^-(t-1)$  and  $\hat{x}_a^-(t) = A_a \hat{x}_a^-(t+1)$ . With this regard, note that the switching sequence for each step has already been calculated. Successively, the measurement equation in (1) allows us to perform posterior corrections  $\hat{\mathbf{x}}_c$  and  $\hat{\mathbf{x}}_a$  on the underlying prior estimates  $\hat{\mathbf{x}}_c^-$  and  $\hat{\mathbf{x}}_a^-$  as follows:

$$\begin{aligned}\hat{x}_c(t) &= \hat{x}_c^-(t) + K_c(y(t) - C_a \hat{x}_a^-(t) - C_c \hat{x}_c^-(t)), \\ \hat{x}_a(t) &= \hat{x}_a^-(t) + K_a(y(t) - C_a \hat{x}_a^-(t) - C_c \hat{x}_c^-(t)),\end{aligned}$$

where  $K_c \in \mathbb{R}^{n_{x_c} \times n_y}$  and  $K_a \in \mathbb{R}^{n_{x_a} \times n_y}$  are the Kalman gains for the causal and non-causal states, respectively, whose design is critical for the effectiveness of the KF. Before delving into the derivation of  $K_c$  and  $K_a$ , let us first calculate the error covariance matrix for the prior state estimates based on the prior estimation errors  $e_c^-(t) = x_c(t) - \hat{x}_c^-(t)$  and  $e_a^-(t) = x_a(t) - \hat{x}_a^-(t)$ , and the posterior estimation errors  $e_c(t) = x_c(t) - \hat{x}_c(t)$  and  $e_a(t) = x_a(t) - \hat{x}_a(t)$ . Denote  $P_c^- \triangleq \mathbb{E}[e_c^-(t)e_c^-(t)^\top]$ ,  $P_a^- \triangleq \mathbb{E}[e_a^-(t)e_a^-(t)^\top]$ ,  $P_c \triangleq$

$\mathbb{E}[e_c(t)e_c(t)^\top]$ ,  $P_a \triangleq \mathbb{E}[e_a(t)e_a(t)^\top]$ . The Kalman gains are calculated to minimize the error covariance matrices of the posterior state estimates. The posterior estimation error can be rewritten as:

$$\begin{aligned} e_c(t) &= (I - K_c C_c) e_c^-(t) - K_c C_a e_a^-(t) - K_c v_m(t), \\ e_a(t) &= (I - K_a C_a) e_a^-(t) - K_a C_c e_c^-(t) - K_a v_m(t), \end{aligned}$$

while the error covariance matrices of the state estimates:

$$\begin{aligned} P_c &= P_c^- - P_c^- C_c^\top K_c^\top - K_c C_c P_c^- + K_c C_c P_c^- C_c^\top K_c^\top \\ &\quad + K_c C_a P_a^- C_a^\top K_c^\top + K_c \Sigma_m K_c^\top, \end{aligned} \quad (12)$$

$$\begin{aligned} P_a &= P_a^- - P_a^- C_a^\top K_a^\top - K_a C_a P_a^- + K_a C_a P_a^- C_a^\top K_a^\top \\ &\quad + K_a C_c P_c^- C_c^\top K_a^\top + K_a \Sigma_m K_a^\top, \end{aligned} \quad (13)$$

where the second equality in each derivation is established based on the independence of  $e_c^-(t)$ ,  $e_a^-(t)$  and  $v_m(t)$ . Note that minimizing the variances of  $P_c$  and  $P_a$  is equivalent to minimizing their traces. Therefore, given the unconstrained nature of such trace minimization, the optimal Kalman gains  $K_c$  and  $K_a$  can be found as:

$$\begin{aligned} K_c &= (C_c P_c^- C_c^\top + C_a P_a^- C_a^\top + \Sigma_m)^{-1} (P_c^- C_c^\top), \\ K_a &= (C_a P_a^- C_a^\top + C_c P_c^- C_c^\top + \Sigma_m)^{-1} (P_a^- C_a^\top). \end{aligned}$$

By substituting the Kalman gains above into (12)–(13), the updated error covariance matrices can be obtained as:  $P_c = (I - K_c C_c) P_c^-$ ,  $P_a = (I - K_a C_a) P_a^-$ . By completing the steps of the modified KF, including the prediction, measurement update, and error covariance matrix update [16], one can obtain all the posterior estimates of the state vectors  $\mathbf{x}_c$  and  $\mathbf{x}_a$ , which are optimal state estimates based on the available measurements and prior knowledge. In addition, to ensure the convergence of the proposed state estimation method, we establish the following properties of the state estimates:

**Lemma 2.** *Let  $\eta_c(t) = x_c(t) - A(\hat{s}_c(t))x_c(t-1)$ ,  $\eta_a(t) = x_a(t) - A(\hat{s}_a(t))x_a(t-1)$ , and  $\eta_m(t) = y(t) - C_c(\hat{s}_c(t))x_c(t) - C_a(\hat{s}_a(t))x_a(t)$ . There exist  $\alpha_1, \alpha_2, \alpha_3 > 0$  so that  $\|\eta_c(t)\|^2 \leq \alpha_1$ ,  $\|\eta_a(t)\|^2 \leq \alpha_2$ , and  $\|\eta_m(t)\|^2 \leq \alpha_3$ , for all  $t \in \mathbb{Z}$ .*

The proof of the Lemma 2 is shown in Appendix A.

Lemma 2 states that the error of state estimation is bounded in the mean square sense, regardless of how the state trajectory evolves in time.

**Remark 3.** *In the E-step, the modified Kalman filter provides posterior state estimates along with their error covariance matrices. Under the Gaussian noise assumptions, the posterior distributions of states given the observations  $\mathbf{y}$  are also Gaussian. Substituting these closed-form expectations (states, error covariance matrices) into (11), the components in  $Q(\theta, \theta^k)$  can be calculated.*

#### 4.2. The M-step

The second step in Algorithm 1 requires the maximization of  $Q(\theta, \theta^k)$  to update the parameters estimate  $\theta^k$ :

$$\theta^{k+1} = \arg \max_{\theta} Q(\theta, \theta^k).$$

Let us first focus on the elements  $\{\{\pi_i^c\}_{i=1}^{m_c}, \{\pi_i^a\}_{i=1}^{m_a}\}$ , and recall the objective function in (10). The  $(k+1)$ -th estimate of  $\{\{\pi_i^c\}_{i=1}^{m_c}, \{\pi_i^a\}_{i=1}^{m_a}\}$  can hence be obtained in closed-form by applying the first-order optimality conditions as follows:

$$\pi_j^c = \frac{\sum_{t=1}^T w_{tj}^c}{\sum_{t=1}^T \sum_{j=1}^{m_c} w_{tj}^c}, \quad \pi_l^a = \frac{\sum_{t=1}^T w_{tl}^a}{\sum_{t=1}^T \sum_{l=1}^{m_a} w_{tl}^a}.$$

Furthermore, the expression for the parameters  $\{\{\theta_i^c\}_{i=1}^{m_c}, \{\theta_i^a\}_{i=1}^{m_a}, \Sigma_m\}$  can be computed in closed-form by using the switching least-squares approach as follows:

$$\begin{aligned} A_c(j) &= \arg \min_{A_c(j)} \sum_{t=1}^T w_{tj}^c \|x_c(t) - \mu_2(t, j)\|^2, \\ A_a(l) &= \arg \min_{A_a(l)} \sum_{t=1}^T w_{tl}^a \|x_a(t) - \mu_3(t, l)\|^2, \\ (C_c(j), C_a(l)) &= \arg \min_{(C_c(j), C_a(l))} \sum_{t=1}^T w_{tj}^c w_{tl}^a \|y(t) - \mu_1(t, j, l)\|^2. \end{aligned}$$

Then, the covariance matrices related to the disturbances  $v_c$ ,  $v_a$ , and  $v_m$  can also be estimated as:

$$\begin{aligned} \Sigma_c(j) &= \sum_{t=1}^T w_{tj}^c (x_c(t) - \mu_2(t, j))(x_c(t) - \mu_2(t, j))^\top, \\ \Sigma_a(l) &= \sum_{t=1}^T w_{tl}^a (x_a(t) - \mu_3(t, l))(x_a(t) - \mu_3(t, l))^\top, \\ \Sigma_m &= \sum_{t=1}^T w_{tj}^c w_{tl}^a (y(t) - \mu_1(t, j, l))(y(t) - \mu_1(t, j, l))^\top. \end{aligned}$$

To show the convergence rate of the system matrices, we need the following definition of strong consistency of the parameter estimates. Recall that  $\hat{\theta}$  is the estimate of  $\theta$  made by exploiting  $T$  samples.

**Definition 1.** ([29]) *The estimate  $\hat{\theta}$  is strongly consistent if  $\lim_{T \rightarrow \infty} \hat{\theta} = \theta$ .*

We are now ready to establish the convergence rate for  $\hat{\theta}$ . Due to the possible different active subsystems at time  $t$ , it is convenient to define the following partition of the considered time interval  $\{1, \dots, T\}$  as  $\mathbb{T}_{j,T}^c = \{t \leq T | s_c(t) = j\}$  and  $\mathbb{T}_{l,T}^a = \{t \leq T | s_a(t) = l\}$ .

**Theorem 1.** *Under Standing Assumptions 1 and 2. Let  $W_{j,T}^c = \sum_{t \in \mathbb{T}_{j,T}^c} x_c(t) x_c^\top(t)$  and  $W_{l,T}^a = \sum_{t \in \mathbb{T}_{l,T}^a} x_a(t) x_a^\top(t)$ . Denote  $\Delta_{j,T}^c = \mathcal{O}\left(\sqrt{\frac{\log(\lambda_{\max}(W_{l,T}^a))}{\lambda_{\min}(W_{j,T}^c)}}\right)$ ,  $\Delta_{l,T}^a = \mathcal{O}\left(\sqrt{\frac{\log(\lambda_{\max}(W_{l,T}^a))}{\lambda_{\min}(W_{l,T}^a)}}\right)$ , and  $\Delta_T = \mathcal{O}\left(\frac{\log(T)}{T}\right)$ . Then, the estimate  $\hat{\theta}$  generated by Algorithm 1 is strongly consistent for any  $s_c \in \Lambda_c^T$  and  $s_a \in \Lambda_a^T$ , and the convergence rates are*

$$\begin{aligned} \|\hat{A}_c(j) - A_c(j)\|_\infty &\leq \Delta_{j,T}^c, & \|\hat{A}_a(l) - A_a(l)\|_\infty &\leq \Delta_{l,T}^a, \\ \|\hat{C}_c(j) - C_c(j)\|_\infty &\leq \Delta_{j,T}^c, & \|\hat{C}_a(l) - C_a(l)\|_\infty &\leq \Delta_{l,T}^a, \\ \|\hat{\Sigma}_c(j) - \Sigma_c(j)\|_\infty &\leq \Delta_T, & \|\hat{\Sigma}_a(l) - \Sigma_a(l)\|_\infty &\leq \Delta_T, \\ \|\hat{\Sigma}_m - \Sigma_m\|_\infty &\leq \Delta_T. \end{aligned}$$

The proof of Theorem 1 is shown in Appendix A.

**Remark 4.** *Theorem 1 gives data-dependent upper bounds for the estimation errors of the parameter matrices. In order to have a data-independent characterization of the convergence rate for adaptive control or reinforcement learning purposes, in the proof of Theorem 1, specifically equation (A.3), we provide with the corresponding convergence rate of the parameter estimate  $\hat{\theta}$ , which is equal to  $\mathcal{O}(\sqrt{\log(T)/T})$ .*

## 5. Numerical examples

We now verify the effectiveness of the proposed methodology on two simulation examples. In both cases, we note that the true switching sequences  $\mathbf{s}_c$  and  $\mathbf{s}_a$  are only used to verify the accuracy of the estimated switching sequences, i.e.,  $\hat{\mathbf{s}}_c$  and  $\hat{\mathbf{s}}_a$ . As performance index we make use of the mode match rate, defined as:

$$L_{\text{mr}} = \frac{1}{T} \sum_{t=1}^T \iota(s_c(t), \hat{s}_c(t)),$$

where  $\iota(\cdot, \cdot)$  denotes the standard indicator function, i.e.,  $\iota(s_c(t), \hat{s}_c(t)) = 1$  if  $s_c(t) = \hat{s}_c(t)$ , 0 otherwise.

### 5.1. Example 1: Academic NCS-RSM

For illustrative purposes, we start by considering a simple non-causal system described in (1) with  $m_c = m_a = 2$  modes and main parameters reported in Table 1 (refer to the ‘‘True’’ columns). The dimensions of the outputs, causal states, and non-causal states are  $n_y = 1, n_{x_c} = n_{x_a} = 2$ . The probabilities of all modes are  $\pi_1^c = 0.7, \pi_2^c = 0.3, \pi_1^a = \pi_2^a = 0.5$ . The system is excited with white noise with zero mean and finite variance, and the data length is  $T = 10^4$ .

The true and estimated parameters are reported in Table 1, which clearly shows that the parameter estimates are very close to their true values. In Fig. 1 we report the partial estimation of the switching sequences  $\mathbf{s}_c$  and  $\mathbf{s}_a$ , where the mode match rates are 97.4% and 99.2%, respectively. Note that our method achieves an accurate parameter estimate, since each data point can be accurately assigned to the corresponding mode. To better validate the accuracy of the proposed algorithm in parameter estimation, Fig. 2 illustrates the estimated states using the modified KF. The relative estimation errors, defined as  $\delta_c = \|x_c - \hat{x}_c\|^2 / \|x_c\|^2$  ( $\delta_a$  has the same structure), are  $\delta_c = 3.74\%$  and  $\delta_a = 3.14\%$ , respectively.

To validate the effectiveness of the proposed method, we compare it with the EM-based identification framework for a JMLS in [11], which was originally derived for causal switching systems. To accommodate the non-causal switching dynamics inherent in NCS-RSM, we extend the two-filter approach to a bidirectional filtering architecture consisting of a forward filter optimized for causal system components, and a backward filter specifically tailored for non-causal dynamics. The length of the data is set to  $T = 10^4$ . The transition matrix in [11] is set to  $\mathcal{T} = \begin{bmatrix} 0.5 & 0.5 \\ 0.5 & 0.5 \end{bmatrix}$ , and the probability of the switching sequence in this paper is set to  $\pi_1^c = \pi_2^c = \pi_1^a = \pi_2^a = 0.5$ . The subsystem match rates of the proposed method and [11] are compared at different noise levels by assuming  $\Sigma = \Sigma_c = \Sigma_a$ . The identification accuracy of the switching sequences are shown in Table 2.

To verify the robust performance of the proposed method against several noise levels, we run 100 Monte Carlo experiments under four different noise conditions, i.e.,  $\Sigma \in$

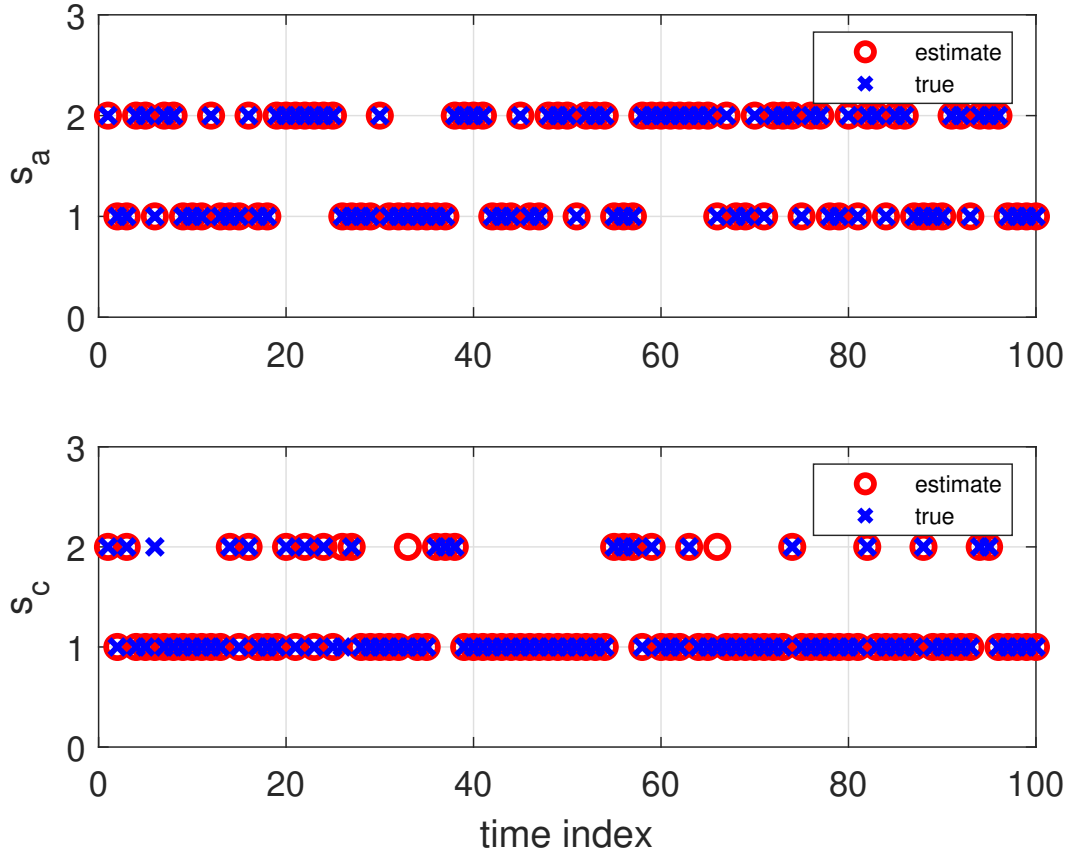


Figure 1: The true (blue cross) and estimated (red circle) mode sequences over a certain time window of length 100.

Table 1: The true and estimated system parameters

	True	Estimate		True	Estimate
$A_a(1)$	$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$	$\begin{bmatrix} 0.9681 & 0.0120 \\ 0.0142 & 0.9868 \end{bmatrix}$	$A_a(2)$	$\begin{bmatrix} 0.6 & 0.2 \\ 0.3 & 0.8 \end{bmatrix}$	$\begin{bmatrix} 0.6242 & 0.1992 \\ 0.3283 & 0.7738 \end{bmatrix}$
$A_c(1)$	$\begin{bmatrix} 1 & 0.2 \\ 0.3 & 0.8 \end{bmatrix}$	$\begin{bmatrix} 1.0131 & 0.2130 \\ 0.2849 & 0.8333 \end{bmatrix}$	$A_c(2)$	$\begin{bmatrix} 0.8 & 0.2 \\ 0.3 & 0.5 \end{bmatrix}$	$\begin{bmatrix} 0.8118 & 0.1899 \\ 0.3291 & 0.4784 \end{bmatrix}$
$C_a(1)$	$\begin{bmatrix} 0.2 & 0.6 \end{bmatrix}$	$\begin{bmatrix} 0.2011 & 0.5962 \end{bmatrix}$	$C_a(2)$	$\begin{bmatrix} 0.3 & 0.76 \end{bmatrix}$	$\begin{bmatrix} 0.2850 & 0.7677 \end{bmatrix}$
$C_c(1)$	$\begin{bmatrix} 0.3 & 0.7 \end{bmatrix}$	$\begin{bmatrix} 0.2983 & 0.6979 \end{bmatrix}$	$C_c(2)$	$\begin{bmatrix} 0.7 & 0.2 \end{bmatrix}$	$\begin{bmatrix} 0.7023 & 0.2029 \end{bmatrix}$
$\pi_1^c$	0.7	0.6963	$\pi_2^c$	0.3	0.3037
$\pi_1^a$	0.5	0.493	$\pi_2^a$	0.5	0.507
$\Sigma_a(1)$	$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$	$\begin{bmatrix} 1.1111 & -0.0711 \\ -0.0711 & 0.9865 \end{bmatrix}$	$\Sigma_a(2)$	$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$	$\begin{bmatrix} 0.9307 & 0.0567 \\ 0.0567 & 1.0386 \end{bmatrix}$
$\Sigma_c(1)$	$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$	$\begin{bmatrix} 0.9773 & -0.0067 \\ -0.0067 & 0.9763 \end{bmatrix}$	$\Sigma_c(2)$	$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$	$\begin{bmatrix} 1.0134 & -0.0001 \\ -0.0001 & 0.9850 \end{bmatrix}$
$\Sigma_m$	1	1.0049			

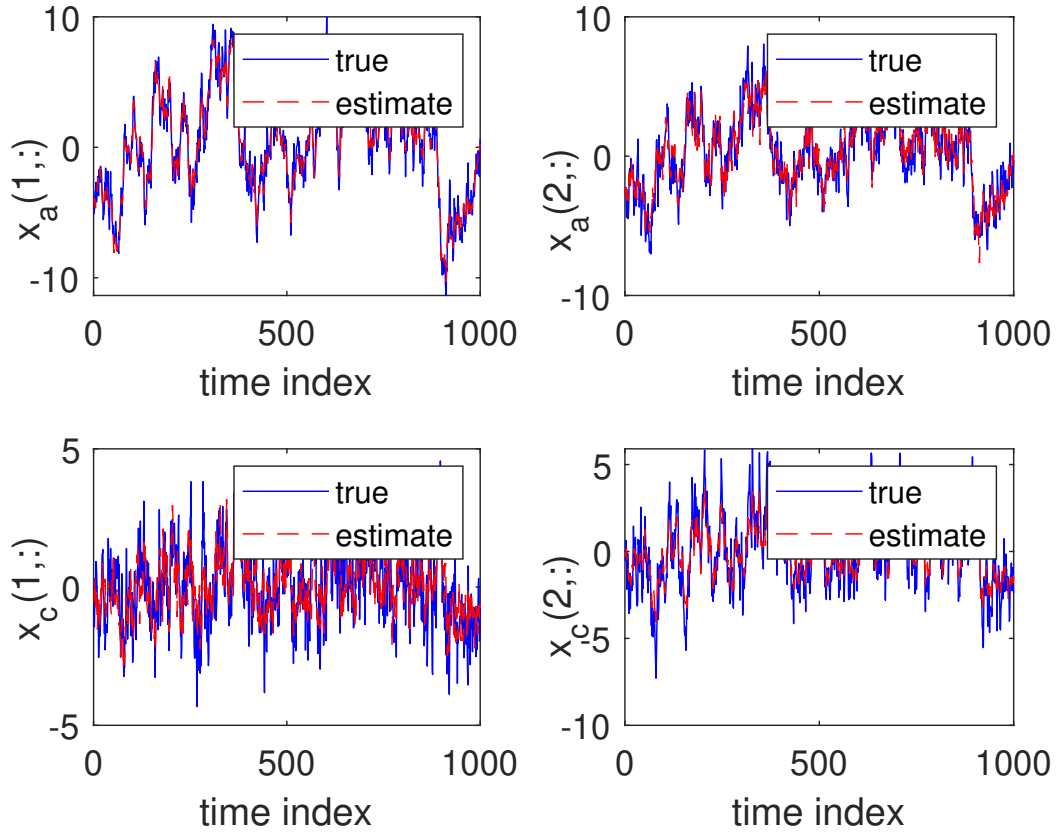


Figure 2: Dynamical evolution of the true state variables  $\mathbf{x}_c$  and  $\mathbf{x}_a$  (solid blue line), and of the estimated ones  $\hat{\mathbf{x}}_c$ , and  $\hat{\mathbf{x}}_a$  (dashed red lines).

Table 2: The mode match rates achieved by the EM algorithm in [11] and by the proposed method.

	$L_{\text{mr}}(\mathbf{s})$ [11]	$L_{\text{mr}}(\mathbf{s}_c)$	$L_{\text{mr}}(\mathbf{s}_a)$
$\Sigma = 0.00$	100%	100%	100%
$\Sigma = 0.01$	99.5%	98.5%	99.3%
$\Sigma = 0.1$	96.5%	97.6%	99.1%
$\Sigma = 1$	89.2%	97.4%	99.2%

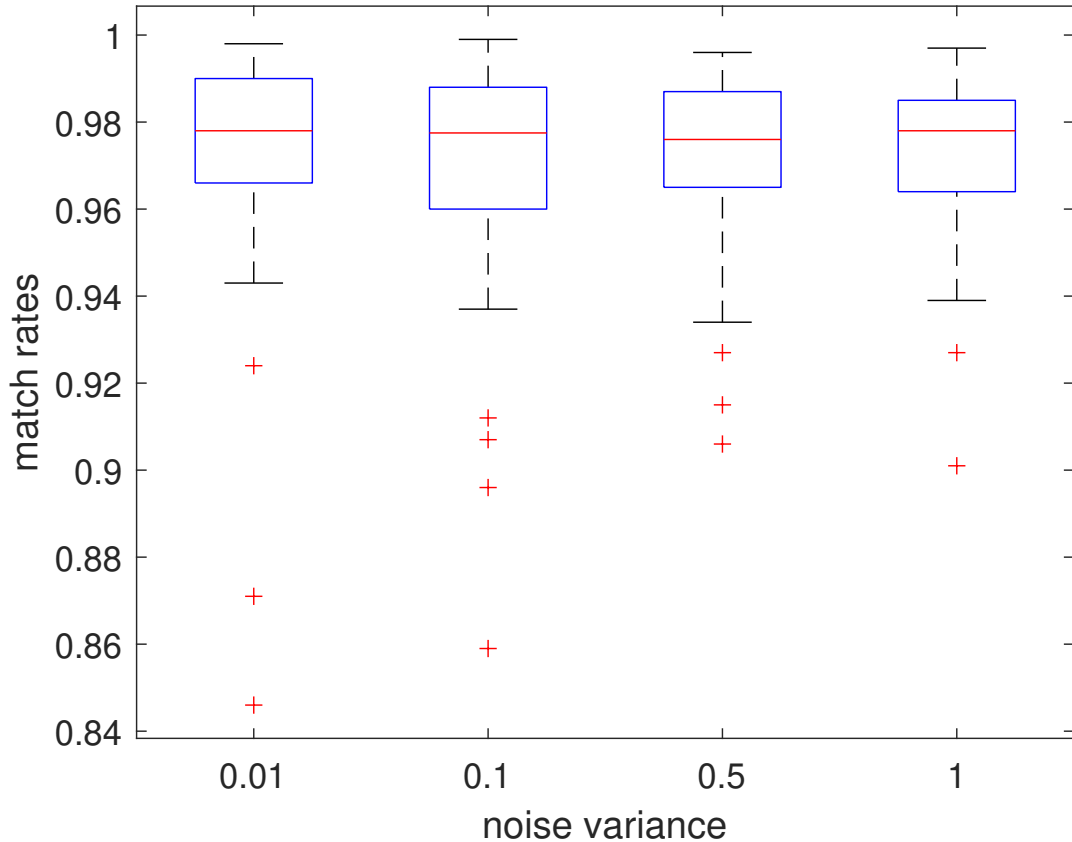


Figure 3: Match rates obtained by the proposed algorithm for different noise levels.

$\{0.01I, 0.1I, 0.5I, I\}$ . In Fig. 3 we report the mean and the variance of the match rates in all the considered cases. We observe that the estimation accuracy of the switching sequence is not significantly affected by the noise variance, since even for high noise levels the estimation accuracy can still reach 98% due to the excellent performance of the modified KF.

### 5.2. Example 2: The Department Store Inventory Price Index

In this subsection we adopt “The Department Store Inventory Price Index” (DSIP) dataset from The Bureau of Labor Statistics (BLS). These data come from inventory weighted price indices of goods carried by department stores.

The dynamics of the DSIP are shaped by the interplay of both causal and non-causal factors. Traditional causal models focus on predicting future prices using historical data (e.g., supply-demand fluctuations, production costs), while the core value of the proposed NCS-ASM lies in historical data smoothing and dynamic interpretation. For instance, future expectations (e.g., pre-holiday inventory adjustments) influence historical price smoothing estimates through non-causal subsystems, correcting fluctuations caused by short-term market noise or measurement errors. Additionally, price dynamics often exhibit bidirectional feedback (e.g., interactions between current inventory and future restocking plans) and mode switching (e.g., seasonal patterns). Therefore, a NCS-RSM model (1) is suitable for describing the DSIP.

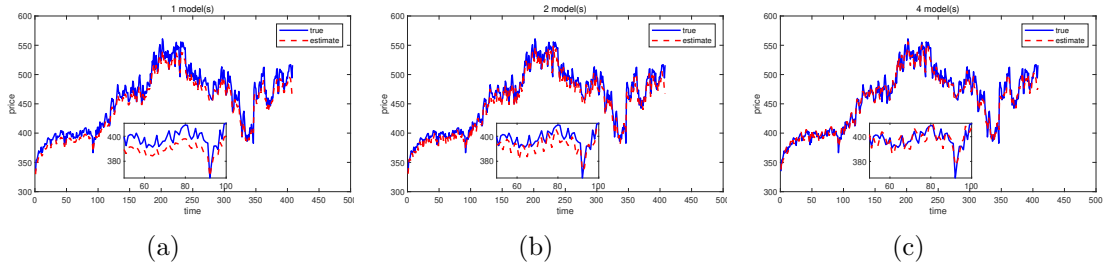


Figure 4: The smoothing estimated prices and the true prices with different numbers of the subsystem. (a)  $\mathbf{s}_c = \mathbf{s}_a$  and  $m_c = m_a = 1$ ; (b)  $\mathbf{s}_c \neq \mathbf{s}_a$  and  $m_c = m_a = 1$ ; (c)  $\mathbf{s}_c \neq \mathbf{s}_a$  and  $m_c = m_a = 2$

In Fig. 4 we show the true prices and the smoothing estimated prices with different number of subsystems. The smoothing estimation errors  $\delta = \|\mathbf{y} - \hat{\mathbf{y}}\|/\|\mathbf{y}\|$  with different number of subsystems are shown in Table 3. Specifically, we can infer that the larger the number of subsystems, the more accuracy the smoothing estimation becomes.

Table 3: The smoothing estimation errors against different number of subsystems.

switching sequence	# of $\mathbf{s}_c$	# of $\mathbf{s}_a$	$\delta$
$\mathbf{s}_c = \mathbf{s}_a$	$m_c = 1$	$m_a = 1$	0.0249
$\mathbf{s}_c \neq \mathbf{s}_a$	$m_c = 1$	$m_a = 1$	0.0195
$\mathbf{s}_c \neq \mathbf{s}_a$	$m_c = 2$	$m_a = 2$	0.0188

In conclusion, from Fig. 4 and Table 3 we note that switching systems with a larger number of modes can obtain the more accurate smoothing estimate of inventory levels, which can offer a guide restocking decisions.

## 6. Conclusion

We have proposed an expectation-maximization framework for identifying non-causal systems with random switching modes. In the E-step, we have embedded the reconstructed switching sequence into the modified Kalman filter so that the proposed algorithm can handle the joint state variable estimation for the causal and non-causal parts. Furthermore, in M-step we have developed a switching least-squares algorithm that can get the parameter estimates in closed-form. From a technical perspective, we have established the convergence of our identification methodology, also deriving an upper bound  $\mathcal{O}(\sqrt{\log(T)/T})$  for the parameter errors.

Note that the identification algorithm proposed in this paper can be adapted to the identification of switching linear descriptor systems with minor modifications, since a descriptor state-space model can be represented in the mixed causal and non-causal form. When the subsystems are nonlinear, however, the identification task becomes more challenging, thus posing greater difficulties. This aspect will be further investigated in our future work. In addition, addressing the joint identification of structured subsystems and piecewise constant switching sequences is an interesting future research direction.

## References

- [1] K. Tan, W. J. Parquette, & M. Tao. (2023). A predictive algorithm for maximum power point tracking in solar photovoltaic systems through load management. *Solar Energy*, 265, 112127.

- [2] S. Liao, Y. Wu, K. Ma, & Y. Niu, (2024). Ant Colony Optimization With Look-Ahead Mechanism for Dynamic Traffic Signal Control of IoV Systems. *IEEE Internet of Things Journal*, 11(1), 366-377.
- [3] R. Carloni, R. G. Sanfelice, A. R. Teel, & C. Melchiorri. (2007). A hybrid control strategy for robust contact detection and force regulation. In *Proc. American Control Conf.*, New York City, USA, 1461-1466.
- [4] C. Liu, J. Li, J. Wang, & Y. Tian. (2016). Supply Chain Simulation with Switching Adaptive Model Predictive Control Methodology. *Proceedings of the 2016 International Conference on Sensor Network and Computer Engineering* (pp. 639-646).
- [5] Andrews, B., Calder, M., & Davis, R.A. (2007). Maximum likelihood estimation for  $\alpha$ -stable autoregressive processes. *Ann. Statist.* 37, 1946–1982.
- [6] T. Schlegl, M. Buss, & G. Schmidt. (2003). A hybrid systems approach toward modeling and dynamical simulation of dextrous manipulation. *IEEE/ASME Trans. on Mechatronics*, 8(3), 352-361.
- [7] Chan, A. B., & Vasconcelos, N. (2008). Modeling, clustering, and segmenting video with mixtures of dynamic textures. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(5), 909–926.
- [8] Ferrari-Trecate, G., Muselli, M., Liberati, D., & Morari, M. (2003). A clustering technique for the identification of piecewise affine systems. *Automatica*, 39(2), 205–217.
- [9] Garulli, Andrea, Paoletti, Simone, & Vicino, Antonio (2012). A survey on switched and piecewise affine system identification. In *16th IFAC symposium on system identification*, Brussels, Belgium (pp. 344–355).
- [10] Bianchi, Federico, Breschi, Valentina, Piga, Dario, & Piroddi, Luigi (2021). Model structure selection for switched NARX system identification: A randomized approach. *Automatica*, 125, Article 109415.
- [11] Mark P. Balenzuela, Adrian G. Wills, Christopher Renton, & Brett Ninness. (2022). Parameter estimation for Jump Markov Linear Systems. *Automatica*, 135 109949.
- [12] Alberto Bemporad, Valentina Breschi, Dario Piga, & Stephen P. Boyd. (2018). Fitting jump models. *Automatica*, 96, 11-21.
- [13] Dario Piga, Valentina Breschi, & Alberto Bemporad. (2020). Estimation of jump Box–Jenkins models. *Automatica*, 120 109126.
- [14] Borna Sayedana, Mohammad Afshari, Peter E. Caines, & Aditya Mahajan. (2024). Strong Consistency and Rate of Convergence of Switched Least Squares System Identification for Autonomous Markov Jump Linear Systems. *IEEE transactions on Automatic Control*, 1-8.
- [15] Anna Scampicchio, Alberto Giaretta, & Gianluigi Pillonetto. (2018). Nonlinear Hybrid Systems Identification using Kernel-Based Techniques. In *IFAC-PapersOnline*, 51(15), 269-274.
- [16] Angelo Alessandri, Marco Baglietto, & Giorgio Battistelli. (2010). A maximum-likelihood Kalman filter for switching discrete-time linear systems. *Automatica*, 46, 1870-1876.
- [17] X. Cao, Bruce Stephen, Ibrahim F. Abdulhadi, C. D. Booth, & G. M. Burt. (2016). Switching Markov Gaussian Models for Dynamic Power System Inertia Estimation. *IEEE Transactions on Power Systems*, 31(5), 3394-3403.
- [18] Miin-Shen Yang, Chien-Yo Lai, & Chih-Ying Lin. (2012), A robust EM clustering algorithm for Gaussian mixture models. *Pattern Recognition*, 45(11), 3950-3961.
- [19] Verhaegen, M. (1996). A subspace model identification solution to the identification of mixed causal, anti-causal LTI systems. *SIAM Journal on Matrix Analysis and Applications*, 17(2), 332–347.
- [20] X. Fang, & T. Chen. (2024). On kernel design for regularized non-causal system identification. *Automatica*, 159, 111335.
- [21] Blanken, L., & Oomen, T. (2020). Kernel-based identification of non-causal systems with application to inverse model control. *Automatica*, 114.
- [22] T. E. Duncan, & B. Pasik-Duncan. (1990). Adaptive control of continuous-time linear stochastic systems. *Math. Control signals systems*, 3(1), 45–60.
- [23] M. K. S. Faradonbeh, A. Tewari, & G. Michailidis. (2020). On adaptive linear–quadratic regulators. *Automatica*, 117, 108982.
- [24] Jiamei Long, Yuqian Guo, & Weihua Gui. (2023). Mean square stability of discrete-time linear systems with random impulsive disturbances. *Science China Information Sciences*, 66(6), 169203.
- [25] Shen Cong. (2024). Almost sure stability criteria for linear Markovian switching systems. *ISA Transactions*, 146, 285-290.
- [26] Dempster, Arthur P., Laird, Nan M., & Rubin, Donald B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, 1–38.
- [27] Andrea Ruggieri, Francesco Stranieri, Fabio Stella, & Marco Scutari. (2020). Hard and Soft EM in Bayesian Network Learning from Incomplete Data. *Algorithms*, 13(12), 329.
- [28] Pal, S., & Heumann, C. (2024). Gaussian mixture model with modified hard EM algorithm in clustering problems. In *Statistical Modeling and Applications on Real-Time Problems* (pp. 153-179). CRC Press.

- [29] T. L. Lai, & C. Z. Wei. (1982). Least squares estimates in stochastic regression models with applications to identification and control of dynamic systems. *Ann. Statist.*, 10(1), 154–166.
- [30] P. E. Caines. (2018). *Linear stochastic systems*. SIAM.
- [31] H. F. Chen, & L. Guo. (1986). Convergence rate of least-squares identification and adaptive control for stochastic systems. *Int J Control*, 44(5), 1459–1476.

## Appendix A. Technical proofs

*Proof of Lemma 1:* The log likelihood difference between the  $\theta$  and  $\theta^k$  can be expressed as

$$\begin{aligned} \ln \mathbb{P}_\theta(\mathbf{y}) - \ln \mathbb{P}_{\theta^k}[\mathbf{y}] &= Q(\theta, \theta^k) - Q(\theta^k, \theta^k) \\ &\quad + V(\theta, \theta^k) - V(\theta^k, \theta^k), \end{aligned}$$

where the difference  $V(\theta, \theta^k) - V(\theta^k, \theta^k)$  coincides with the Kullback–Leibler distance that possess an important property, i.e., being non-negative. Therefore, the maximization of  $Q(\theta, \theta^k)$  can yield an increase in the log-likelihood function  $\ln \mathbb{P}_\theta(\mathbf{y})$ , namely

$$Q(\theta, \theta^{k+1}) \geq Q(\theta, \theta^k) \Rightarrow \ln \mathbb{P}_{\theta^{k+1}}[\mathbf{y}] \geq \ln \mathbb{P}_{\theta^k}[\mathbf{y}],$$

thus concluding the proof. ■

*Proof of Lemma 2:* Only the boundedness of  $\eta_c(t)$  will be proven in detail, since that of  $\eta_a(t)$  and  $\eta_m(t)$  can be derived in a similar way.

First, we note that  $x_c(t-1)$  can be equivalently expressed as follows:

$$x_c(t-1) = \varphi_1(\mathbf{s}_c)x_c(1) + \varphi_2(\mathbf{s}_c)\mathbf{v}_c(1:t-1),$$

where  $\mathbf{v}_c(1:t-1) \triangleq [v_c(1), \dots, v_c(t-1)]$ ,  $\varphi_1(\mathbf{s}_c)$  and  $\varphi_2(\mathbf{s}_c)$  are shown as follows:

$$\varphi_1(\mathbf{s}_c) = A_c(s_c(2)) + A_c(s_c(3))A_c(s_c(2)) + \dots + A_c(s_c(t-1))A_c(s_{t-2}) \dots A_c(s_c(2)) \tag{A.1}$$

$$\varphi_2(\mathbf{s}_c) = \begin{bmatrix} 1 + A_c(s_c(2)) + A_c(s_c(3))A_c(s_c(2)) + \dots + A_c(s_c(t-1)) \dots A_c(s_c(2)) \\ 1 + A_c(s_c(3)) + \dots + A_c(s_c(t-1)) \dots A_c(s_c(3)) \\ \vdots \\ 1 + A_c(s_c(t-1)) \\ 1 \end{bmatrix}^\top \tag{A.2}$$

Both matrices are uniquely determined by the switching sequence  $\mathbf{s}_c$  and system matrices  $A_c$ . Then, one obtains that:

$$\begin{aligned} \eta_c(t) &= x_c(t) - A(\hat{s}_c(t))x_c(t-1) \\ &= (A_c(s_c(t)) - A(\hat{s}_c(t)))x_c(t-1) + v_c(t) \\ &= (A_c(s_c(t)) - A(\hat{s}_c(t)))\varphi_1(\mathbf{s}_c)x_c(1) \\ &\quad + \varphi_3(\mathbf{s}_c)\mathbf{v}_c(1:t), \end{aligned}$$

where  $\varphi_3(\mathbf{s}_c) = [\varphi_2(\mathbf{s}_c), 1]$ . Passing to the (squared) norm in the expression above we note that, in view of the fact that the noise  $\mathbf{v}_c$  has a bounded covariance, the last term is bounded too. For what concerns the first term, instead, we have:

$$\| [A_c(\mathbf{s}_c(t)) - A(\hat{\mathbf{s}}_c(t))] \varphi_1(\mathbf{s}) x_c(1) \|^2 \leq \lambda_1 \|x_c(1)\|^2,$$

where

$$\lambda_1 \triangleq \lambda_{\max}(\varphi_1^\top(\mathbf{s})(A_c(\mathbf{s}_c(t)) - A(\hat{\mathbf{s}}_c(t)))^\top (A_c(\mathbf{s}_c(t)) - A(\hat{\mathbf{s}}_c(t)))\varphi_1(\mathbf{s})),$$

which concludes the proof. ■

*Proof of Theorem 1:* In the interest of space, we establish the convergence rate for  $\hat{A}_a(l)$  only, since the other bounds on the system matrices can be derived similarly. We start by introducing two necessary lemmas. In particular, the following result holds true by virtue of Standing Assumptions 1 and 2:

**Lemma 3.** ([14, Lemma 3]) *The following asymptotic relations hold true almost surely (a.s.):*

$$\begin{aligned} \left\| \sum_{i=1}^T A(\mathbf{s}_c(i)) x_c(i) v_c^\top(i) + v_c(i) x_c^\top(i) A(\mathbf{s}_c(i)) \right\| &= o(T), \\ \left\| \sum_{i=1}^T A(\mathbf{s}_a(i)) x_a(i) v_a^\top(i) + v_a(i) x_a^\top(i) A(\mathbf{s}_a(i)) \right\| &= o(T). \end{aligned}$$

The proof extends [14, Lemma 3] by treating the non-causal dynamics (1) as a time-reversed causal process. By Standing Assumption 1, the reversed process preserves stability in the average sense. The martingale property of  $v_a(t)$  (Standing Assumption 2) ensures the applicability of the covariance analysis in [14].

Next, we report a lemma whose validity follows from Standing Assumption 2:

**Lemma 4.** ([29]) *The standard least-squares solution can be expressed as  $\hat{A}_a(l) = \arg \min_{A_a(l)} \|x_a(t) - A_a(l)x_a(t+1)\|^2$ ,  $t \in \mathbb{T}_{l,T}^a$ , for all  $l = 1, \dots, m_a$ . If*

(C1)  $\lambda_{\min}(W_{l,T}^a) \rightarrow \infty$  a.s., and

(C2)  $\log \lambda_{\max}(W_{l,T}^a) = o(\lambda_{\min}(W_{l,T}^a))$  a.s.,

then the least-squares estimate  $\hat{A}_a(l)$  is strongly consistent with convergence rate

$$\|\hat{A}_a(l) - A_a(l)\|_\infty = \mathcal{O} \left( \sqrt{\frac{\log(\lambda_{\max}(W_{l,T}^a))}{\lambda_{\min}(W_{l,T}^a)}} \right) \text{ a.s.}$$

In view of Lemma 4, sufficient conditions for establishing the convergence rate of  $\hat{A}_a(l)$  are (C1)  $\lambda_{\min}(W_{l,T}^a) \rightarrow \infty$ , a.s., and (C2)  $\log \lambda_{\max}(W_{l,T}^a) = o(\lambda_{\min}(W_{l,T}^a))$ , a.s.. We therefore have to show that these two conditions are verified in our case. Then, for what concerns

(C1), one has:

$$\begin{aligned}
x_a(t)x_a(t)^\top &= (\hat{A}_a(l)x_a(t+1) + v_a(t)) \\
&\quad (\hat{A}_a(l)x_a(t+1) + v_a(t))^\top \\
&= \hat{A}_a(l)x_a(t+1)x_a^\top(t+1)\hat{A}_a^\top(l) \\
&\quad + 2v_a(t)x_a^\top(t+1)\hat{A}_a^\top(l) + v_a(t)v_a^\top(t).
\end{aligned}$$

Since  $\hat{A}_a(l)x_a(t+1)x_a^\top(t+1)\hat{A}_a^\top(l)$  is a positive semidefinite matrix, by relying on Lemma 3 we can infer that

$$\begin{aligned}
W_{l,T}^a &= \sum_{t \in \mathbb{T}_{l,T}^a} x_a(t)x_a^\top(t) \\
&\succcurlyeq \sum_{t \in \mathbb{T}_{l,T}^a} v_a(t)v_a^\top(t) + x_a(T)x_a^\top(T) \\
&\quad + \sum_{t \in \mathbb{T}_{l,T}^a} (\hat{A}_a(l)x_a(t+1)v_a^\top(t) + v_a(t)x_a^\top(t+1)\hat{A}_a^\top(l)) \\
&\succcurlyeq \sum_{t \in \mathbb{T}_{l,T}^a} v_a(t)v_a^\top(t) + o(T).
\end{aligned}$$

Then, we readily obtain:

$$\begin{aligned}
\lim_{|\mathbb{T}_{l,T}^a| \rightarrow \infty} \inf \frac{\sum_{t \in \mathbb{T}_{l,T}^a} x_a(t)x_a^\top(t)}{|\mathbb{T}_{l,T}^a|} \\
&\succcurlyeq \lim_{|\mathbb{T}_{l,T}^a| \rightarrow \infty} \inf \frac{\sum_{t \in \mathbb{T}_{l,T}^a} v_a(t)v_a^\top(t)}{|\mathbb{T}_{l,T}^a|} \succ 0.
\end{aligned}$$

Therefore, we can conclude that  $\lambda_{\min}(W_{l,T}^a) \rightarrow \infty$  a.s..

To prove (C2) we note that:

$$\begin{aligned}
\lambda_{\max}(\sum_{t \in \mathbb{T}_{l,T}^a} x_a(t)x_a^\top(t)) &\leq \text{tr}(\sum_{t \in \mathbb{T}_{l,T}^a} x_a(t)x_a^\top(t)) \\
&\leq \sum_{i=1}^T \|x_a(t)\|^2 = \mathcal{O}(N),
\end{aligned}$$

where the last equality follows in view of the stability, in average sense, of the NCS-RSM in (1). Then, one can readily obtain that

$$\lim_{T \rightarrow \infty} \frac{\log(\lambda_{\max}(W_{l,T}^a))}{\lambda_{\min}(W_{l,T}^a)} \leq \lim_{T \rightarrow \infty} \frac{\log(T)}{|\mathbb{T}_{l,T}^a|} = \frac{\log(T)}{\mathcal{O}(T)} = 0. \tag{A.3}$$

We are now able to establish the convergence rate for the covariance matrices. Specifically, we will give the detailed proof for  $\hat{\Sigma}_c(j)$  only, since the remaining ones follow similarly.

From the NCA-ASM in (1), the true covariance matrix for  $v_c$  can be expressed as:

$$\begin{aligned}
\Sigma_c(j) &= \frac{1}{|\mathbb{T}_{j,T}^c|} \sum_{t \in \mathbb{T}_{j,T}^c} (x_c(t) - A_c(j)x_c(t-1)) \\
&\quad (x_c(t) - A_c(j)x_c(t-1))^\top.
\end{aligned}$$

Then, the estimation error can take the following form:

$$\hat{\Sigma}_c(j) - \Sigma_c(j) = \frac{1}{|\mathbb{T}_{j,T}^c|} \sum_{t \in \mathbb{T}_{j,T}^c} ((A_c(j) - \hat{A}_c(j))x_c(t-1)) \\ ((A_c(j) - \hat{A}_c(j))x_c(t-1))^\top.$$

Therefore, the convergence rate for  $\hat{\Sigma}_c(j)$  reads as:

$$\|\hat{\Sigma}_c(j) - \Sigma_c(j)\|_\infty \leq \frac{\sum_{t \in \mathbb{T}_{j,T}^c} x_c(t-1)x_c(t-1)^\top}{|\mathbb{T}_{j,T}^c|} \\ \|(A_c(j) - \hat{A}_c(j))(A_c(j) - \hat{A}_c(j))^\top\|_\infty \\ \leq \mathcal{O}\left(\frac{\log(T)}{T}\right),$$

which completes the proof. ■