Finite Sample Analysis for a Class of Subspace Identification Methods

Jiabao He, Ingvar Ziemann, Cristian R. Rojas and Håkan Hjalmarsson

***Abstract*— While subspace identification methods (SIMs) are appealing due to their simple parameterization for MIMO systems and robust numerical realizations, a comprehensive statistical analysis of SIMs remains an open problem, especially in the non-asymptotic regime. In this work, we provide a finite sample analysis for a class of SIMs, which reveals that the convergence rates for estimating Markov parameters and system matrices are $\mathcal{O}(1/\sqrt{N})$, in line with classical asymptotic results. Based on the observation that the model format in classical SIMs becomes non-causal because of a projection step, we choose a parsimonious SIM that bypasses the projection step and strictly enforces a causal model to facilitate the analysis, where a bank of ARX models are estimated in parallel. Leveraging recent results from finite sample analysis of an individual ARX model, we obtain an overall error bound of an array of ARX models and proceed to derive error bounds for system matrices via robustness results for the singular value decomposition.**

I. Introduction

Originating from the celebrated Ho-Kalman algorithm [1], subspace identification methods (SIMs) have proven extremely useful for estimating linear state-space models. Over the past 50 years, numerous efforts have been made to develop improved algorithms and gain a deeper understanding of the family of SIMs, exemplified by the successful narrative of closed-loop identification [2]–[5]. For a comprehensive overview of SIMs, we refer to [6], [7]. Despite their tremendous success both in theory and practice, several drawbacks have been recognized, including a lower accuracy compared to prediction error methods (PEMs) and an incomplete statistical analysis. There are some significant contributions to statistical properties of SIMs in the asymptotic regime. The consistency of open-loop and closed-loop SIMs is analyzed in [8] and [5], respectively, where the former suggests that persistence of excitation (PE) of the input signals is not sufficient for consistency, and stronger conditions are required in some cases. The asymptotic variance of SIMs is discussed in [9]–[12]. Furthermore, the asymptotic equivalence of some SIMs is discussed in [13]–[15]. Regarding the optimality, although the canonical variate analysis (CVA) method is stated to be optimal when measured inputs are white [16], [17], simulation studies indicate that it is not asymptotically efficient, as it does not reach Cramér-Rao lower bound (CRLB) [14]. In short, SIMs are generally consistent, however, the question of whether there are SIMs that are asymptotically efficient remains unresolved. Besides, although some of the methods are asymptotically equivalent, their performance differs in finite sample setups. Meanwhile, it is difficult to capture their transient behaviors using the asymptotic theory. Therefore, a complete statistical analysis and comparison among various SIMs, as well as the pursuit of an efficient SIM are still open problems.

There has been a recent resurgence of interest in identifying state-space models for dynamic systems, where the focus is on the non-asymptotic regime. Finite sample analysis in the field of system identification was pioneered by [18], [19], where the performance of PEMs were analyzed. Over the last few years, a series of papers have revisited this topic and introduced many promising developments on fully observed systems [20]–[22] and partially observed systems [23]–[26]. For a broader overview of these results, we refer to [27], [28]. As pointed out in [23], finite sample analysis has been a standard tool for comparing algorithms in the non-asymptotic regime. It is expected that such analysis of SIMs will not only provide a detailed qualitative characterization of learning complexity and error bounds, but also bring more insights into the comparison of different SIMs regarding optimality, the selection of past and future horizons, and the design of controllers [27]. Broadly speaking, an analysis in the non-asymptotic regime brings a wider understanding of SIMs.

Despite the auspicious future, the path of finite sample analysis for SIMs proves to be challenging. Usually, multi-step statistical operations are involved in SIMs, including pre-estimation, projection, weighted singular value decomposition (SVD) and maximum likelihood (ML) estimation. While these steps enhance performance, they simultaneously complicate the model and pose challenges for statistical analysis. To the best of our knowledge, except for a finite sample analysis of the Ho-Kalman algorithm [24], a stochastic SIM in the absence of input signals [23] and an individual ARX model [26], [28], a finite sample analysis of classical SIMs in the presence of inputs is still unavailable. Compared with SIMs without taking account of measured inputs, the inclusion of inputs results in a higher complexity due to the presence of an unknown block matrix with a lower-triangular Toeplitz structure. This matrix is responsible for recording the impact of 'future' input on 'future' output. To manage this complexity, classical SIMs choose to remove this matrix via a projection step. Although this method is computationally efficient, it makes the model format non-causal anymore and poses challenges for statistical analysis.

Jiabao He, Cristian R. Rojas and Håkan Hjalmarsson are with the Division of Decision and Control Systems, School of Electrical Engineering and Computer Science, KTH Royal Institute of Technology, 100 44 Stockholm, Sweden. (Emails: jiabaoh, crro, hjalmars@kth.se)

Ingvar Ziemann is with the University of Pennsylvania, Philadelphia, PA 19104 USA. (Email: ingvarz@seas.upenn.edu)

This work was supported by VINNOVA Competence Center AdBIOPRO, contract [2016-05181] and by the Swedish Research Council through the research environment NewLEADS (New Directions in Learning Dynamical Systems), contract [2016-06079], and contract 2019-04956.

Therefore, we propose to use a parallel and parsimonious SIM, namely PARSIM [29], which bypasses the projection step and strictly enforces a causal model to facilitate the analysis. Except the projection step, PARSIM aligns with the unified framework of the family of SIMs in all other aspects, opting for PARSIM will not constrain our comprehension of the full panorama of SIMs.

The main contributions of this paper are:

1) We leverage recent results from finite sample analysis of an individual ARX model to obtain an error bound for the bank of ARX models featured in PARSIM. Our analysis reveals that the convergence rates for estimating Markov parameters and system matrices are $\mathcal{O}(1/\sqrt{N})$ even in the presence of inputs, which is in line with classical asymptotics.
2) Our method can be extended to the class of SIMs that estimate an array of ARX models, for instance, the included PARSIM and SIMs based on predictor identification (PBSID). Therefore, it paves the way for comprehensively understanding the broader landscape of SIMs.

The disposition of the paper is as follows: A short review of SIMs with the focus on PARSIM is given in Section II, and the problem as well as the roadmap ahead to analyze its finite sample behavior are described at the end of Section II. The finite sample analysis of an individual ARX model is summarized in Section III. An overall error bound for a bank of ARX models and the resulting error bounds of the system matrices are given in Section IV. Finally, some discussions and future work are provided in Section V.

Notations: For a matrix X with appropriate dimensions, $X^\top$, X^{-1}, $X^{1/2}$, $X^\dagger$, $\|X\|$, $\det(X)$, $\rho(X)$, $\lambda_{\max}(X)$, $\lambda_{\min}(X)$ and $\sigma_n(X)$ denote its transpose, inverse, square root, Moore-Penrose pseudo-inverse, spectral norm, determinant, spectral radius, maximum eigenvalue, minimum eigenvalue and $n-$th largest singular value, respectively. $X \succ (\succcurlyeq)$ 0 means that X is positive (semi) definite. The matrices I and 0 are the identity and zero matrices with compatible dimensions. The multivariate normal distribution with mean μ and covariance Σ is denoted as $\mathcal{N}(\mu, \Sigma)$. The notation $\mathbb{E}x$ is the expectation of a random vector x, and $\mathbb{P}(\mathcal{E})$ is the probability of the event $\mathcal{E}$. $\mathcal{E}^c$ is the complementary event of $\mathcal{E}$, and $\mathcal{E}_1 \cup \mathcal{E}_2$ and $\mathcal{E}_1 \cap \mathcal{E}_2$ are the union and intersection of events $\mathcal{E}_1$ and $\mathcal{E}_2$, respectively.

II. Preliminaries

A. Model and Assumptions

Consider the following discrete-time linear time-invariant (LTI) system on innovations form:

$$x_{k+1} = Ax_k + Bu_k + Ke_k, \tag{1a}$$
$$y_k = Cx_k + e_k, \tag{1b}$$

where $x_t \in \mathbb{R}^{n_x}$, $u_t \in \mathbb{R}^{n_u}$, $y_t \in \mathbb{R}^{n_y}$ and $e_t \in \mathbb{R}^{n_y}$ are the state, input, output and innovations, respectively. For brevity of notation, we assume that the initial time starts at $k = 1$, and the initial state is $x_1 = 0$. It has been widely recognized that under mild conditions, the above innovations model describes the same input-output trajectories with identical statistics as a standard state-space model [6], [28]. Thus, without loss of generality, we study the innovations model. We make the following assumptions which are commonly used in SIMs:

Assumption 2.1: 1) The spectral radius of A and $A - KC$ satisfies $\rho(A) \leq 1$ and $\rho(A - KC) < 1$.

2) The system is minimal, i.e., $(A, [B, K])$ is controllable and (A, C) is observable, and the system order n_x is known to the user.
3) The innovations $\{e_k\}$ consists of independent and identically distributed (i.i.d.) Gaussian random variables, i.e., $e_k \sim \mathcal{N}(0, \sigma_e^2 I)$.[1]
4) The input sequence $\{u_k\}$ consists also of i.i.d. Gaussian random variables, i.e., $u_k \sim \mathcal{N}(0, \sigma_u^2 I)$. Moreover, it is assumed independent of $\{e_k\}$.

B. A Recap of PARSIM

Here we provide a short review of SIMs, with a focus on PARSIM. An extended state-space model for (1) can be derived as [29]

$$Y_f = \Gamma_f X_k + G_f U_f + H_f E_f, \tag{2a}$$
$$Y_p = \Gamma_p X_{k-p} + G_p U_p + H_p E_p, \tag{2b}$$

where f and p denote future and past horizons chosen by the user, respectively. The extended observability matrix is

$$\Gamma_f = \begin{bmatrix} C^\top & (CA)^\top & \cdots & (CA^{f-1})^\top \end{bmatrix}^\top, \tag{3}$$

and G_f with H_f are lower-triangular Toeplitz matrices of Markov parameters with respect to the input and innovations,

$$G_f = \begin{bmatrix} 0 & 0 & \cdots & 0 \\ CB & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ CA^{f-2}B & CA^{f-3}B & \cdots & 0 \end{bmatrix}, \tag{4a}$$

$$H_f = \begin{bmatrix} I & 0 & \cdots & 0 \\ CK & I & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ CA^{f-2}K & CA^{f-3}K & \cdots & I \end{bmatrix}. \tag{4b}$$

Past and future inputs are collected in the Hankel matrices

$$U_p = \begin{bmatrix} u_{k-p} & u_{k-p+1} & \cdots & u_{k-p+N-1} \\ u_{k-p+1} & u_{k-p+2} & \cdots & u_{k-p+N} \\ \vdots & \vdots & \ddots & \vdots \\ u_{k-1} & u_k & \cdots & u_{k+N-2} \end{bmatrix}, \tag{5a}$$

$$U_f = \begin{bmatrix} u_k & u_{k+1} & \cdots & u_{k+N-1} \\ u_{k+1} & u_{k+2} & \cdots & u_{k+N} \\ \vdots & \vdots & \ddots & \vdots \\ u_{k+f-1} & u_{k+f} & \cdots & u_{k+f+N-2} \end{bmatrix}. \tag{5b}$$

[1] We believe that our results can be extended to more general setups, for instance, sub-Gaussians.

Similar definitions are given for matrices Γ_p, G_p, H_p, Y_p, Y_f, E_p and E_f [29]. Usually, we take $k = p+1$, and then the total number of samples is denoted as $\bar{N} = p+f+N-1$. The state sequences are defined as

$$X_k = \begin{bmatrix} x_k & x_{k+1} & \cdots & x_{k+N-1} \end{bmatrix}, \tag{6a}$$
$$X_{k-p} = \begin{bmatrix} x_{k-p} & x_{k-p+1} & \cdots & x_{k-p+N-1} \end{bmatrix}. \tag{6b}$$

Furthermore, by replacing e_k with $y_k - Cx_k$ in (1a) and iterating the equation, we obtain the relation

$$X_k = L_p Z_p + \bar{A}^p X_{k-p}, \tag{7}$$

where $A_c = A - KC$, $Z_p = \begin{bmatrix} Y_p^\top & U_p^\top \end{bmatrix}^\top$, and L_p is the extended controllability matrix defined as

$$L_p = \begin{bmatrix} A_c^{p-1}K & \cdots & A_c K & K & A_c^{p-1}B & \cdots & A_c B & B \end{bmatrix}. \tag{8}$$

After substituting (7) into (2a), we have

$$Y_f = \Gamma_f L_p Z_p + G_f U_f + H_f E_f + \Gamma_f \bar{A}^p X_{k-p}. \tag{9}$$

Most SIMs use (9) to first estimate either the extended observability matrix Γ_f or system state X_k, and then obtain a realization of the system matrices up to a similarity transformation. To illustrate, since G_f is a lower-triangular Toeplitz matrix recording the effect of U_f on Y_f, and it is difficult to preserve such structure with least-squares, what classical SIMs do is to eliminate this term by projecting out U_f as

$$Y_f \Pi_{U_f}^{\perp} = \Gamma_f L_p Z_p \Pi_{U_f}^{\perp} + H_f E_f \Pi_{U_f}^{\perp} + \Gamma_f \bar{A}^p X_{k-p} \Pi_{U_f}^{\perp}, \tag{10}$$

where $\Pi_{U_f}^{\perp} = I - U_f^\top (U_f U_f^\top)^{-1} U_f$. Then the above equation is simplified based on the following observations: When p is sufficiently large, we have $\bar{A}^p \approx 0$. Also, as U_f is uncorrelated with E_f, we have $E_f \Pi_{U_f}^{\perp} \approx E_f$. Furthermore, E_f is uncorrelated with Z_p, i.e., $\frac{1}{N} E_f Z_p^\top \approx 0$. By multiplying $Z_p^\top$ on both sides of (10) we have

$$Y_f \Pi_{U_f}^{\perp} Z_p^\top \approx \Gamma_f L_p Z_p \Pi_{U_f}^{\perp} Z_p^\top. \tag{11}$$

Then the range space of the extended observability matrix Γ_f can be estimated using

$$\widehat{\Gamma_f L_p} = Y_f \Pi_{U_f}^{\perp} Z_p^\top (Z_p \Pi_{U_f}^{\perp} Z_p^\top)^{-1}. \tag{12}$$

To recover the extended observability matrix Γ_f or the state X_k, weighted SVD is often used, i.e.,

$$W_1 \widehat{\Gamma_f L_p} W_2 = \hat{U} \hat{\Lambda} \hat{V}^\top \approx \hat{U}_1 \hat{\Lambda}_1 \hat{V}_1^\top, \tag{13}$$

where $\hat{\Lambda}_1$ contains the n_x largest singular values. In this way, a balanced realization of $\hat{\Gamma}_f$ or $\hat{L}_p$ is

$$\hat{\Gamma}_f = W_1^{-1} \hat{U}_1 \hat{\Lambda}_1^{1/2}, \tag{14a}$$
$$\hat{L}_p = \hat{\Lambda}_1^{1/2} \hat{V}_1^\top W_2^{-1}. \tag{14b}$$

Different choices of weighting matrices W_1 and W_2 lead to distinct classical SIMs [6], [30]. As pointed out in [29], one of the main issues of those classical SIMs is that the model is not causal anymore due to the absence of G_f in (10). As a result, the estimated parameters have inflated variance due to the existence of unnecessary and extra terms. Furthermore, this poses a challenge in analyzing statistical properties, which we will see later in detail.

To enforce causal models, a parallel and parsimonious SIM, PARSIM, is proposed in [29]. Instead of doing the projection step in (10) once, PARSIM zooms into each row of (9) and equivalently performs f ordinary least-squares (OLS) to estimate a bank of ARX models. To illustrate this, the extended state-space model (9) can be partitioned row-wise as

$$Y_{fi} = \Gamma_{fi} L_p Z_p + G_{fi} U_i + H_{fi} E_i + \Gamma_{fi} \bar{A}^p X_{k-p}, \tag{15}$$

where $i = 1, 2, \ldots f$, and

$$\begin{aligned}
\Gamma_{fi} &= CA^{i-1} \in \mathbb{R}^{n_y \times n_x}, \\
Y_{fi} &= \begin{bmatrix} y_{k+i-1} & y_{k+i} & \cdots & y_{k+N+i-2} \end{bmatrix} \in \mathbb{R}^{n_y \times N}, \\
U_{fi} &= \begin{bmatrix} u_{k+i-1} & u_{k+i} & \cdots & u_{k+N+i-2} \end{bmatrix} \in \mathbb{R}^{n_u \times N}, \\
U_i &= \begin{bmatrix} U_{f1}^\top & U_{f2}^\top & \cdots & U_{fi}^\top \end{bmatrix}^\top \in \mathbb{R}^{in_u \times N}, \\
E_{fi} &= \begin{bmatrix} e_{k+i-1} & e_{k+i} & \cdots & e_{k+N+i-2} \end{bmatrix} \in \mathbb{R}^{n_y \times N}, \\
E_i &= \begin{bmatrix} E_{f1}^\top & E_{f2}^\top & \cdots & E_{fi}^\top \end{bmatrix}^\top \in \mathbb{R}^{in_y \times N}, \\
G_{fi} &= \begin{bmatrix} CA^{i-2}B & \cdots & CB & 0 \end{bmatrix} \\
&\triangleq \begin{bmatrix} G_{i-1} & \cdots & G_1 & G_0 \end{bmatrix} \in \mathbb{R}^{n_y \times in_u}, \\
H_{fi} &= \begin{bmatrix} CA^{i-2}K & \cdots & CK & I \end{bmatrix} \\
&\triangleq \begin{bmatrix} H_{i-1} & \cdots & H_1 & H_0 \end{bmatrix} \in \mathbb{R}^{n_y \times in_y}.
\end{aligned}$$

Then PARSIM uses OLS to estimate each $\Gamma_{fi} L_p$ and G_{fi} simultaneously from the causal model (15):

$$\hat{\theta}_i \triangleq \begin{bmatrix} \widehat{\Gamma_{fi} L_p} & \hat{G}_{fi} \end{bmatrix} = Y_{fi} \begin{bmatrix} Z_p \\ U_i \end{bmatrix}^\dagger. \tag{16}$$

At last, the whole estimate of $\Gamma_f L_p$ is obtained by stacking the f estimates together as

$$\widehat{\Gamma_f L_p} = \begin{bmatrix} \widehat{\Gamma_{f1} L_p} \\ \widehat{\Gamma_{f2} L_p} \\ \vdots \\ \widehat{\Gamma_{ff} L_p} \end{bmatrix}. \tag{17}$$

Comparing with classical SIMs that only estimate $\Gamma_f L_p$ in (12), PARSIM estimates $\Gamma_f L_p$ and Markov parameters G_{fi} simultaneously. In this way, the lower-triangular Toeplitz structure of G_f is preserved and the causality is strictly enforced. Furthermore, it has been shown that the above algorithm gives a smaller variance of $\widehat{\Gamma_f L_p}$ than classical SIMs in the asymptotic regime. For the subsequent realization step, PARSIM goes back to the weighted SVD step (13).

In this paper, we consider a simplified implementation of the original PARSIM [29], i.e., we choose the weighting matrices $W_1 = I$ and $W_2 = I$, and the system matrices are obtained in the following way:

$$\hat{C} = \hat{\Gamma}_f(1 : n_y, :), \tag{18a}$$
$$\hat{A} = (\underline{I}\hat{\Gamma}_f)^\dagger (\bar{I}\hat{\Gamma}_f), \tag{18b}$$
$$\hat{K} = \hat{L}_p(:, (p-1)n_y + 1 : pn_y), \tag{18c}$$
$$\hat{B} = \hat{L}_p(:, (2p-1)n_y + 1 : 2pn_y), \tag{18d}$$

where $\underline{I} = \begin{bmatrix} I & 0 \end{bmatrix}$, $\bar{I} = \begin{bmatrix} 0 & I \end{bmatrix}$ and the indexing of matrices follows MATLAB syntax.

C. Problem and Roadmap Ahead

Now we define the problem explicitly and sketch the path ahead to its solution. Under Assumption 2.1, given a finite number $\bar{N}$ of input-output samples and horizons f and p, we aim to provide error bounds with high probability for the realization (18). To be specific, with probability at least $1-\delta$, we wish to establish the following error bounds explicitly [23]:

$$\|\hat{A} - T^{-1}AT\| \leq \epsilon_A(\delta, N), \|\hat{C} - CT\| \leq \epsilon_C(\delta, N), \tag{19a}$$

$$\|\hat{K} - T^{-1}K\| \leq \epsilon_K(\delta, N), \|\hat{B} - T^{-1}B\| \leq \epsilon_B(\delta, N), \tag{19b}$$

where T is a non-singular matrix.

Remark 1: It is only possible to obtain the system matrices up to a similarity transformation due to the non-uniqueness of the realization [24].

We arrive at this result by a two-step procedure. In Step 1, we derive an error bound for $\hat{\theta}_i$ in (16) for every ARX model. In other words, we define the event

$$\mathcal{E}_i \triangleq \left\{ \left\| \hat{\theta}_i - \theta_i \right\| \leq \epsilon_{\theta_i}(\delta, N) \right\}, \tag{20}$$

and require that $\mathbb{P}(\mathcal{E}_i^c) \leq \delta/f$ for $i = 1, 2, ..., f$. In Step 2, we first utilize a norm inequality between the block matrix $\widehat{\Gamma_f L_p}$ and its sub-matrices $\widehat{\Gamma_{fi} L_p}$ to obtain the total error bound of $\widehat{\Gamma_f L_p}$. This essentially requires that the intersection of f events has probability $\mathbb{P}(\mathcal{E}_1 \cap \mathcal{E}_2 \cap \cdots \cap \mathcal{E}_f) \geq 1 - \delta$, which is guaranteed due to Bonferroni's inequality: the probability $\mathbb{P}(\mathcal{E}_1^c \cup \mathcal{E}_2^c \cup \cdots \cup \mathcal{E}_f^c) \leq \sum_{i=1}^{f} \mathbb{P}(\mathcal{E}_i^c) \leq \delta$, in Step 1. At last, using recent results from SVD robustness [23], [24], error bounds in (19) are obtained.

III. Finite Sample Analysis of each ARX Model

Following our roadmap, we formalize Step 1 above in this section. We emphasize that the results presented in this section apply to $i = 1, 2, ..., f$, with the acknowledgment of their reliance on the specific value of i. This dependency is underscored through the use of the subscript i. First, we partition the following matrices column-wise:

$$\begin{bmatrix} Z_p \\ \hdashline U_i \end{bmatrix} = \begin{bmatrix} y_p(1) & y_p(2) & \cdots & y_p(N) \\ u_p(1) & u_p(2) & \cdots & u_p(N) \\ \hdashline u_i(1) & u_i(2) & \cdots & u_i(N) \end{bmatrix}, \tag{21a}$$

$$E_i = \begin{bmatrix} e_i(1) & e_i(2) & \cdots & e_i(N) \end{bmatrix}, \tag{21b}$$

where $u_p(k) = \begin{bmatrix} u_k^\top & u_{k+1}^\top & \cdots & u_{k+p-1}^\top \end{bmatrix}^\top$ and $u_i(k) = \begin{bmatrix} u_{k+p}^\top & u_{k+p+1}^\top & \cdots & u_{k+p+i-1}^\top \end{bmatrix}^\top$ are past input and future input, respectively, and similar definitions apply to $y_p(k)$ and $e_i(k)$. For brevity, we define a covariate

$$z_{p,i}(k) = \begin{bmatrix} y_p(k) \\ u_p(k) \\ u_i(k) \end{bmatrix}, \tag{22}$$

so the error of the OLS estimate (16) can be written as

$$\begin{aligned} \tilde{\theta}_i \triangleq& \hat{\theta}_i - \theta_i = H_{fi} E_i \begin{bmatrix} Z_p \\ U_i \end{bmatrix}^\dagger + \Gamma_{fi} \bar{A}^p X_{k-p} \begin{bmatrix} Z_p \\ U_i \end{bmatrix}^\dagger \\ =& \underbrace{H_{fi} \sum_{k=1}^{N} \frac{1}{N} e_i(k) z_{p,i}^\top(k) \left(\sum_{k=1}^{N} \frac{1}{N} z_{p,i}(k) z_{p,i}^\top(k) \right)^{-1}}_{\text{stochastic error } \tilde{\theta}_i^S} + \\ & \underbrace{\Gamma_{fi} \bar{A}^p \sum_{k=1}^{N} \frac{1}{N} x_k z_{p,i}^\top(k) \left(\sum_{k=1}^{N} \frac{1}{N} z_{p,i}(k) z_{p,i}^\top(k) \right)^{-1}}_{\text{truncation bias } \tilde{\theta}_i^B}. \end{aligned} \tag{23}$$

There are two types of errors, namely, the stochastic error $\tilde{\theta}_i^S$ and truncation bias $\tilde{\theta}_i^B$. The key observation is that the future innovations $e_i(k)$ are independent of the covariate $z_{p,i}(l)$ for all $l < k$, due to the fact that $z_{p,i}(l)$ consists of the past output, past input and future input. This provides a martingale structure, which is convenient to analyze. By contrast, if we revisit the classical SIMs, the stochastic error for the estimate (12) will be $E_f \Pi_{U_f}^\perp Z_p^\top (Z_p \Pi_{U_f}^\perp Z_p^\top)^{-1}$. Due to the projection matrix $\Pi_{U_f}^\perp$, the columns of E_f and Z_p are mixed together, making the above term non-causal, resulting in the loss of the martingale structure. We believe that this is one of the main barriers preventing a finite sample analysis for classical SIMs, which is also the reason why we choose PARSIM that bypasses the projection step. Before proceeding further, we have the following definitions regarding the covariance and empirical covariance of the covariant $z_{p,i}(k)$:

$$\Sigma_{p,i}(k) \triangleq \mathbb{E} z_{p,i}(k) z_{p,i}^\top(k), \tag{24a}$$

$$\hat{\Sigma}_{p,i}(N) \triangleq \frac{1}{N} \sum_{k=1}^{N} z_{p,i}(k) z_{p,i}^\top(k). \tag{24b}$$

For simplicity, with a slight abuse of notation, we use $\Sigma_{i,k} = \Sigma_{p,i}(k)$ and $\hat{\Sigma}_{i,N} = \hat{\Sigma}_{p,i}(N)$, where the dependency of covariance on the past horizon p is concealed. Also, the covariance of the state x_k is defined similarly as

$$\Sigma_{x,k} \triangleq \mathbb{E} x_k x_k^\top. \tag{25}$$

In this way, the stochastic error $\tilde{\theta}_i^S$ can be rewritten as

$$\tilde{\theta}_i^S = \left(H_{fi} \sum_{k=1}^{N} \frac{1}{N} e_i(k) z_{p,i}^\top(k) \hat{\Sigma}_{i,N}^{-1/2} \right) \hat{\Sigma}_{i,N}^{-1/2}. \tag{26}$$

To bound the above term, we use a self-normalized martingale to deal with the leftmost term in the bracket. As for $\hat{\Sigma}_{i,N}^{-1/2}$, we use recent results from the smallest eigenvalue of the empirical covariance of causal Gaussian processes to bound it [28], [31], which establish the condition of PE.

A. Persistence of Excitation

First, we make the following assumption regarding the selection of the past horizon p.

Assumption 3.1: The past horizon is chosen as $p = \beta \log N$, where β is large enough such that

$$\|CA_c^p\| \, \|\Sigma_{x,N}\| \leq N^{-3}. \tag{27}$$

Remark 2: The above assumption ensures that the truncation bias $\Gamma_{fi} A_c^p x_k$ is small enough, allowing the model (15) to closely approximate an ARX model. To achieve this goal, the exponentially decaying term A_c^p should counteract the magnitude of the state x_k. Since A is marginally stable, x_k scales at most polynomially with k [28]. Hence, the state norm $\|\Sigma_{x,N}\|$ grows at most polynomially with N. Since $\rho(A_c) < 1$, we have $\|A_c^p\| = \mathcal{O}(\rho^p)$ for some $\rho > \rho(A_c)$. Taking $p = \beta \log N$, we have $\|A_c^p\| = \mathcal{O}(N^{-\beta/\log(1/\rho)})$. In this way, the condition (27) will be satisfied for a large β.

PE is equivalent to requiring that the empirical covariance $\hat{\Sigma}_{i,N}$ is positive definite. For this purpose, the number of samples N should exceed a certain threshold, namely, burn-in time N_{pe}.

Definition 3.1: The burn-in time N_{pe} is defined as

$$N_{pe}(\delta, \beta, i) \triangleq \min \{N : N \geq N_0(N, \delta, \beta, i)\}, \tag{28}$$

where

$$N_0(N, \delta, \beta, i) \triangleq c_0 \tau_i \max\left\{\sigma_e^2, 1\right\} (\log \frac{1}{\delta} + d_i \log C_{\text{sys}}(N, \tau_i)),$$

$$C_{\text{sys}}(N, \tau_i) \triangleq \frac{N}{3\tau_i} \frac{\|\Sigma_{i,N}\|^2}{\lambda_{\min}^2(\Sigma_{i,\tau_i})}, \tau_i = i + p = i + \beta \log N,$$

$d_i = p n_y + \tau_i n_u$ and c_0 is a universal positive constant.

Remark 3: To show that the above definition is not vacuous, we need to demonstrate that the condition $N \geq N_0(N, \delta, \beta, i)$ is feasible. For any given β, τ_i increases logarithmically with N. Also, $\|\Sigma_{i,N}\|^2$ grows polynomially with N, hence, the system theoretic term $\log C_{\text{sys}}(N, \tau_i)$ increases at most logarithmically with N [28]. As a result, $N_0(N, \delta, \beta, i)$ grows polynomially with N.

The condition of PE is summarized as follows:

Lemma 1: Fix a failure probability $0 < \delta < 1$, if $N \geq N_{pe}(\delta, \beta, i)$, then with probability at least $1 - \delta$, we have

$$\hat{\Sigma}_{i,N} \succcurlyeq \frac{1}{16} \Sigma_{i,\tau_i}, \tag{29}$$

where $\lambda_{\min}(\Sigma_{i,\tau_i}) > 0$.

Proof: To conserve space, the complete proof is omitted here, which is identical to PE of the ARX model provided in Theorem 5.2 of [28]. ∎

Remark 4: We emphasize that the above PE result is also helpful for analyzing the impact of weighting matrices W_1 and W_2 in the weighted SVD step (13), which will be investigated in detail in the future. To illustrate this, it has been shown in [29] that $\left(Z_p \Pi_{U_f}^{\perp} Z_p^\top\right)^{1/2}$ is approximately the optimal choice for W_2. A prerequisite is that $Z_p \Pi_{U_f}^{\perp} Z_p^\top$ should be positive definite, i.e., $\frac{1}{N} U_f U_f^\top \succ 0$ and $\frac{1}{N} Z_p \Pi_{U_f}^{\perp} Z_p^\top \succ 0$. According to the Schur complement, this is equivalent to

$$\frac{1}{N} \begin{bmatrix} Z_p Z_p^\top & Z_p U_f^\top \\ U_f Z_p^\top & U_f U_f^\top \end{bmatrix} = \hat{\Sigma}_{f,N} \succ 0, \tag{30}$$

which is essentially same as PE in Lemma 1 by taking $i = f$.

B. Stochastic Error

The bound of the stochastic error $\tilde{\theta}_i^S$ is based on the following three events:

$$\mathcal{E}_{i,1} \triangleq \left\{\hat{\Sigma}_{i,N} \succcurlyeq \frac{1}{16} \Sigma_{i,\tau_i}\right\}, \tag{31a}$$

$$\mathcal{E}_{i,2} \triangleq \left\{\hat{\Sigma}_{i,N} \preccurlyeq \frac{3d_i}{\delta} \Sigma_{i,N}\right\}, \tag{31b}$$

$$\mathcal{E}_{i,3} \triangleq \left\{ \left\| \sum_{k=1}^{N} e_i(k) z_{p,i}(k)^\top \left(\Sigma + N\hat{\Sigma}_{i,N}\right)^{-1/2} \right\|^2 \leq 4\sigma_e^2 \log \frac{\det(\Sigma + N\hat{\Sigma}_{i,N})}{\det(\Sigma)} + 8\sigma_e^2 \left(n_y \log 5 + \log \frac{3}{\delta}\right) \right\}, \tag{31c}$$

where $\mathbb{P}(\mathcal{E}_{i,j}^c) \leq \delta/3$ for $j = 1, 2, 3$. The event $\mathcal{E}_{i,1}$ is due to PE in Lemma 1, the event $\mathcal{E}_{i,2}$ is derived from the matrix Markov inequality [28], and the event $\mathcal{E}_{i,3}$ is based on the result of self-normalized martingales [32]. The signal-to-noise ratio (SNR) is defined as

$$\text{SNR}_{i,k} \triangleq \frac{\lambda_{\min}(\Sigma_{i,k})}{\sigma_e^2}, \tag{32}$$

and is assumed to be uniformly lower bounded for all possible p, which has been proven to be a quite general assumption [28]. The bound of the stochastic error $\tilde{\theta}_i^S$ is summarized in the following lemma:

Lemma 2: Fix a failure probability $0 < \delta < 1$, if $N \geq N_{pe}(\delta/3, \beta, i)$, then with probability at least $1 - \delta$,

$$\left\|\tilde{\theta}_i^S\right\|^2 \leq \frac{c\,\|H_{fi}\|^2}{\text{SNR}_{i,\tau_i} N} (d_i \log \frac{d_i}{\delta} + \log\det(\Sigma_{i,N} \Sigma_{i,\tau_i}^{-1})), \tag{33}$$

where c is a universal constant which is independent of the system, δ and β.

Proof: We sketch the proof here, which is similar to the proof of Theorem 5.1 in [28]. The main idea is to prove that the event in Lemma 2 is subsumed by the intersection of three events $\mathcal{E}_{i,1}$, $\mathcal{E}_{i,2}$ and $\mathcal{E}_{i,3}$, with $\mathbb{P}(\mathcal{E}_{i,1} \cap \mathcal{E}_{i,2} \cap \mathcal{E}_{i,3}) \geq 1 - \delta$, due to the fact that $\mathbb{P}(\mathcal{E}_{i,1}^c \cup \mathcal{E}_{i,2}^c \cup \mathcal{E}_{i,3}^c) \leq \delta$. According to (26), we have

$$\left\|\tilde{\theta}_i^S\right\|^2 \leq \underbrace{\left\| H_{fi} \sum_{k=1}^{N} \frac{1}{N} e_i(k) z_{p,i}^\top(k) \hat{\Sigma}_{i,N}^{-1/2} \right\|^2}_{\text{noise term}} \underbrace{\left\|\hat{\Sigma}_{i,N}^{-1}\right\|}_{\text{excitation term}}. \tag{34}$$

The excitation term is bounded based on the event $\mathcal{E}_{i,1}$, i.e.,

$$\left\|\hat{\Sigma}_{i,N}^{-1}\right\| \leq 16 \lambda_{\min}^{-1}(\Sigma_{i,\tau_i}). \tag{35}$$

For the noise term, we bound it by mimicking the form of the event $\mathcal{E}_{i,3}$. Due to $\mathcal{E}_{i,1}$, we have $2N\hat{\Sigma}_{i,N} \succcurlyeq N\hat{\Sigma}_{i,N} + \Sigma_i$, where $\Sigma_i = \frac{N}{16} \Sigma_{i,\tau_i}$. Hence, the noise term can be formulated as

$$\left\|\frac{H_{fi}}{\sqrt{N}}\sum_{k=1}^{N}e_i(k)z_{p,i}(k)^{\top}(N\hat{\Sigma}_{i,N})^{-1/2}\right\|^2 \leq \frac{2\left\|H_{fi}\right\|^2}{N}\left\|\sum_{k=1}^{N}e_i(k)z_{p,i}(k)^{\top}(\Sigma_i+N\hat{\Sigma}_{i,N})^{-1/2}\right\|^2. \tag{36}$$

Furthermore, according to $\mathcal{E}_{i,3}$ and $\mathcal{E}_{i,2}$, we have

$$\begin{aligned}&\left\|\sum_{k=1}^{N}e_i(k)z_{p,i}(k)^{\top}\left(\Sigma_i+N\hat{\Sigma}_{i,N}\right)^{-1/2}\right\|^2 \leq \\ &4\sigma_e^2\log\frac{\det(\Sigma_i+N\hat{\Sigma}_{i,N})}{\det\left(\Sigma_i\right)}+8\sigma_e^2\left(n_y\log 5+\log\frac{3}{\delta}\right) \leq \\ &4\sigma_e^2\log\det(I+48\frac{d_i}{\delta}\Sigma_{i,N}\Sigma_{i,\tau_i}^{-1})+8\sigma_e^2(n_y\log 5+\log\frac{3}{\delta}).\end{aligned} \tag{37}$$

Combining (36) and (37) with (35), and simplifying them properly, the result (33) is obtained. ∎

C. Truncation Bias Term

As for the bias term $\tilde{\theta}_i^B$, we will see that it is dominated by the stochastic error $\tilde{\theta}_i^S$, given a proper selection of p and the fact that the system is stable [28]. The bias term is similarly decomposed as

$$\tilde{\theta}_i^B=\left(\Gamma_{fi}\bar{A}^p\sum_{k=1}^{N}x_k z_{p,i}^{\top}(k)(N\hat{\Sigma}_{i,N})^{-1/2}\right)(N\hat{\Sigma}_{i,N})^{-1/2}. \tag{38}$$

Note that $N\hat{\Sigma}_{i,N}=\sum_{k=1}^{N}z_{p,i}(k)z_{p,i}^{\top}(k)\succcurlyeq z_{p,i}(k)z_{p,i}^{\top}(k)$, hence we have

$$\left\|z_{p,i}^{\top}(k)\left(N\hat{\Sigma}_{i,N}\right)^{-1}z_{p,i}(k)\right\|\leq 1. \tag{39}$$

Using the triangle inequality,

$$\left\|\sum_{k=1}^{N}x_k z_{p,i}^{\top}(k)\left(N\hat{\Sigma}_{i,N}\right)^{-1/2}\right\| \leq \sum_{k=1}^{N}\left\|x_k\right\|\left\|z_{p,i}^{\top}(k)\left(N\hat{\Sigma}_{i,N}\right)^{-1/2}\right\|. \tag{40}$$

Then based on (39) and the Cauchy-Schwarz inequality, we have

$$\left\|\sum_{k=1}^{N}x_k z_{p,i}^{\top}(k)\left(N\hat{\Sigma}_{i,N}\right)^{-1/2}\right\|\leq\sqrt{N\sum_{k=1}^{N}\left\|x_k\right\|^2}. \tag{41}$$

Now we introduce the following lemma:

Lemma 3 (Lemma E.5 in [28]): Fix a failure probability δ, there exists a universal constant c such that with probability at least $1-\delta$,

$$\sum_{k=1}^{N}\left\|x_k\right\|^2\leq c\sigma_e^2 n_x N\left\|\Sigma_{x,N}\right\|\log\frac{1}{\delta}. \tag{42}$$

Using this lemma, inequality (41) can be further simplified as

$$\left\|\sum_{k=1}^{N}x_k z_{p,i}^{\top}(k)\left(N\hat{\Sigma}_{i,N}\right)^{-1/2}\right\|\leq\sqrt{c\sigma_e^2 n_x\left\|\Sigma_{x,N}\right\|N\log\frac{1}{\delta}}. \tag{43}$$

Combining with the event $\mathcal{E}_{i,1}$, now we bound the bias as

$$\left\|\tilde{\theta}_i^B\right\|^2\leq\frac{16cn_x}{\text{SNR}_{i,\tau_i}}\left\|\Gamma_{fi}\bar{A}^p\right\|\left\|\Sigma_{x,N}\right\|N\log\frac{1}{\delta}. \tag{44}$$

According to Assumption 3.1, by taking $p=\beta\log N$ such that

$$\left\|\Gamma_{fi}\bar{A}^p\right\|\left\|\Sigma_{x,N}\right\|\leq N^{-3}, \tag{45}$$

the bias error can be further bounded as

$$\left\|\tilde{\theta}_i^B\right\|^2\leq\frac{16cn_x}{N^2\text{SNR}_{i,\tau_i}}\log\frac{1}{\delta}. \tag{46}$$

As we can see, the truncation bias $\tilde{\theta}_i^B$ decays as $\mathcal{O}(1/N)$ and is dominated by the stochastic error $\tilde{\theta}_i^S$. By merging the two errors $\tilde{\theta}_i^S$ and $\tilde{\theta}_i^B$ together, and absorbing high order terms into the dominating term by inflating the constants accordingly, we obtain the following theorem.

Theorem 3.1: Fix a failure probability δ, if $N\geq N_{pe}(\frac{\delta}{3},\beta,i)$, then with probability at least $1-2\delta$,

$$\left\|\tilde{\theta}_i\right\|^2\leq\frac{c\left\|H_{fi}\right\|^2}{\text{SNR}_{i,\tau_i}N}\left(d_i\log\frac{d_i}{\delta}+\log\det(\Sigma_{i,N}\Sigma_{i,\tau_i}^{-1})\right), \tag{47}$$

where c is a universal constant which is independent of the system, δ and β.

IV. Robustness of Balanced Realization

Following our roadmap, having obtained error bounds for each ARX model in Step 1, now we move to Step 2 and provide the overall bound for (17) and error bounds for the balanced realization (18).

A. Overall Bound

First, we introduce the following lemma about the block matrix norm:

Lemma 4 (Lemma A.1 in [23]): Let M be a block-column matrix defined as $M=\begin{bmatrix}M_1^{\top} & M_2^{\top} & \cdots & M_f^{\top}\end{bmatrix}^{\top}$, where all the M_i's have the same dimension. Then, the block matrix M satisfies

$$\left\|M\right\|\leq\sqrt{f}\max_{1\leq i\leq f}\left\|M_i\right\|. \tag{48}$$

Based on this lemma, it is straightforward to obtain the following theorem regarding the error bound of the range space of the extended observability matrix $\widehat{\Gamma_f L_p}$.

Theorem 4.1: For a fixed probability δ, if

$$N\geq\max\left\{N_{pe}(\frac{\delta}{3f},\beta,i)\right\}, \text{for } i=1,2,...f, \tag{49}$$

then with probability at least $1-2\delta$, we have

$$\begin{aligned}&\left\|\widehat{\Gamma_f L_p}-\Gamma_f L_p\right\|\leq\sqrt{f}\max_{1\leq i\leq f}\left\|\tilde{\theta}_i\right\|\leq \\ &\sqrt{\frac{f}{N}}\max_{1\leq i\leq f}\sqrt{\frac{c\left\|H_{fi}\right\|}{\text{SNR}_{i,\tau_i}}(d_i\log\frac{d_i}{\delta}+\log\det(\Sigma_{i,N}\Sigma_{i,\tau_i}^{-1}))}.\end{aligned} \tag{50}$$

B. Bounds on System Matrices

Assume that we know the true value of $\Gamma_f L_p$, and the SVD of $\Gamma_f L_p$ is

$$\Gamma_f L_p = \begin{bmatrix} U_1 & U_0 \end{bmatrix} \begin{bmatrix} \Lambda_1 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} V_1 & V_0 \end{bmatrix}^\top, \tag{51}$$

so a balanced realization for Γ_f and L_p is

$$\bar{\Gamma}_f = U_1 \Lambda_1^{1/2}, \bar{L}_p = \Lambda_1^{1/2} V_1^\top. \tag{52}$$

Moreover, the system matrices with respect to a similarity transform are obtained according to (18), which are denoted as $\{\bar{A}, \bar{B}, \bar{C}, \bar{K}\}$. We now have the following result regarding the error bounds of the system matrices:

Theorem 4.2: If the following condition is satisfied:

$$\left\| \widehat{\Gamma_f L_p} - \Gamma_f L_p \right\| \leq \frac{\sigma_n(\Gamma_f L_p)}{4}, \tag{53}$$

then there exists a unitary matrix T such that, for a fixed probability δ, if

$$N \geq \max\left\{ N_{pe}(\frac{\delta}{3f}, \beta, i) \right\}, \text{for } i = 1, 2, ...f, \tag{54}$$

then with probability at least $1 - 2\delta$, we have

$$\max\left\{ \left\| \hat{\Gamma}_f - \bar{\Gamma}_f T \right\|, \left\| \hat{L}_p - T^\top \bar{L}_p \right\| \right\} \leq$$
$$2\sqrt{\frac{10n}{\sigma_n(\Gamma_f L_p)}} \left\| \widehat{\Gamma_f L_p} - \Gamma_f L_p \right\|, \tag{55a}$$
$$\max\left\{ \left\| \hat{C} - \bar{C}T \right\|, \left\| \hat{K} - T^\top \bar{K} \right\|, \left\| \hat{B} - T^\top \bar{B} \right\| \right\} \leq$$
$$\left\| \hat{\Gamma}_f - \bar{\Gamma}_f T \right\|, \tag{55b}$$
$$\left\| \hat{A} - T^\top \bar{A} T \right\| \leq \frac{\sqrt{\|\Gamma_f L_p\|} + \sigma_o}{\sigma_o^2} \left\| \hat{\Gamma}_f - \bar{\Gamma}_f T \right\|, \tag{55c}$$

where $\sigma_o = \min\left(\sigma_n(\hat{\Gamma}_f \underline{I}), \sigma_n(\bar{\Gamma}_f \underline{I}) \right)$.

Proof: The proof of the above SVD step is similar to [23], [24], thus, it is omitted here. ■

Now we discuss the implications of our main results Theorems 4.1 and 4.2.

1) The dimension of the system affects PE and error bounds: As shown in (28) and (50), the number of samples N and the error bound scale with the dimension $d_i = p(n_u + n_y) + in_u$, which corresponds to the intuition that to estimate $\theta_i \in \mathbb{R}^{n_y \times d_i}$, we need at least as many number of independent equations coming from measurements.
2) The error bounds scale linearly with $\log\frac{1}{\delta}$, which is consistent with the result in the asymptotic regime [28].
3) According to Theorem 4.1, the total error bound of $\left\| \widehat{\Gamma_f L_p} - \Gamma_f L_p \right\|$ decays as $\mathcal{O}(1/\sqrt{N})$. This conclusion is drawn based on the following observations: The error bound scales linearly with the logarithm of $\det(\Sigma_{i,N} \Sigma_{i,\tau_i}^{-1})$ which increases at most polynomially with N, the SNR is uniformly lower bounded, $\|H_{fi}\|$ is bounded and the future horizon f is fixed.
4) According to Theorem 4.2, the error bounds of the system matrices $\{A, B, C, K\}$ depend linearly on the error bound of $\left\| \widehat{\Gamma_f L_p} - \Gamma_f L_p \right\|$, thus, they also decay as $\mathcal{O}(1/\sqrt{N})$. This indicates that as long as the identification error of $\Gamma_f L_p$ is small, the results from the subsequent SVD step are robust.

V. Discussion and Future Work

This paper presents a finite sample analysis of a parsimonious SIM that estimates a bank of ARX models using OLS in parallel. It reveals that the convergence rates for estimating Markov parameters and system matrices are $\mathcal{O}(1/\sqrt{N})$, consistent with classical asymptotic results. Besides, PE and the role of past horizon are also discussed. Although the algorithm studied in this paper streamlines the weighted SVD and realization steps, we believe that the findings herein pave the way for comprehensively grasping the broader landscape of the family of SIMs. Here are a few aspects we would like to discuss:

1) Our bound is not tight: Just like the existing bounds for partially observed systems, our bound is not tight, either. In this paper, PE in the f ARX models are dealt with separately, which is convenient but somewhat conservative, given that the past output and input Z_p is reused for every estimation. Our bound can be further optimized; however, akin to the challenge of finding an efficient SIM in the asymptotic regime, the pursuit of a lower bound in the non-saymptotic regime for partially observed systems is more challenging.
2) Selection of past and future horizons: A similar suggestion as other work is given on the selection of the past horizon p, however, the guidance for the future horizon f is not given. While it is evident that f influences the error bound, a comprehensive analysis of its impact is not given. Using asymptotic theory [17], it has been shown that the variance of the estimates of the CVA method improves as f increases, thus, it is interesting to study the impact of f in the non-asymptotic regime.
3) Comparison between different weighting matrices: We consider a simplified SIM, i.e., taking the weighting matrices $W_1 = W_2 = I$. It is generally believed that the optimal choices usually depend on specific data sets, like the CVA method. In the future, we will analyze the influence of different weighting matrices on the realization step. As we remark in the persistence of excitation, our work provides some preliminary insights on the validity of existing weighting matrices.

References

[1] B. L. Ho and R. E. Kálmán, "Effective construction of linear state-variable models from input/output functions," *Automatisierungstechnik*, vol. 14, no. 1-12, pp. 545–548, 1966.

[2] S. J. Qin and L. Ljung, "Closed-loop subspace identification with innovation estimation," in *Proc. 13th IFAC Symp. Syst. Identification*, Netherlands, 2003.

[3] M. Jansson, "Subspace identification and ARX modeling," in *Proc. 13th IFAC Symp. Syst. Identification*, Netherlands, 2003.

[4] L. Ljung and T. McKelvey, "Subspace identification from closed loop data," *Signal Process.*, vol. 52, no. 2, pp. 209–215, 1996.

[5] A. Chiuso and G. Picci, "Consistency analysis of some closed-loop subspace identification methods," *Automatica*, vol. 41, no. 3, pp. 377–391, 2005.

[6] S. J. Qin, "An overview of subspace identification," *Comput. Chem. Eng.*, vol. 30, no. 10-12, pp. 1502–1513, 2006.

[7] G. Van der Veen, J.-W. van Wingerden, M. Bergamasco, M. Lovera, and M. Verhaegen, "Closed-loop subspace identification methods: An overview," *IET Control Theory Appl.*, vol. 7, no. 10, pp. 1339–1358, 2013.

[8] M. Jansson and B. Wahlberg, "On consistency of subspace methods for system identification," *Automatica*, vol. 34, no. 12, pp. 1507–1519, 1998.

[9] T. Gustafsson, "Subspace-based system identification: weighting and pre-filtering of instruments," *Automatica*, vol. 38, no. 3, pp. 433–443, 2002.

[10] M. Jansson, "Asymptotic variance analysis of subspace identification methods," *Proc. 12th IFAC Symp. Syst. Identification*, pp. 91–96, 2000.

[11] D. Bauer, "Asymptotic properties of subspace estimators," *Automatica*, vol. 41, no. 3, pp. 359–376, 2005.

[12] A. Chiuso and G. Picci, "The asymptotic variance of subspace estimates," *J. Econom.*, vol. 118, no. 1-2, pp. 257–291, 2004.

[13] A. Chiuso, "On the relation between CCA and predictor-based subspace identification," *IEEE Trans. Autom. Control*, vol. 52, no. 10, pp. 1795–1812, 2007.

[14] A. Chiuso, "The role of vector autoregressive modeling in predictor-based subspace identification," *Automatica*, vol. 43, no. 6, pp. 1034–1048, 2007.

[15] A. Chiuso, "Some insights on the choice of the future horizon in closed-loop CCA-type subspace algorithms," in *2007 Am. Control Conf.*, 2007, pp. 840–845.

[16] W. E. Larimore, "Statistical optimality and canonical variate analysis system identification," *Signal Process.*, vol. 52, no. 2, pp. 131–144, 1996.

[17] D. Bauer and L. Ljung, "Some facts about the choice of the weighting matrices in Larimore type of subspace algorithms," *Automatica*, vol. 38, no. 5, pp. 763–773, 2002.

[18] M. C. Campi and E. Weyer, "Finite sample properties of system identification methods," *IEEE Trans. Autom. Control*, vol. 47, no. 8, pp. 1329–1334, 2002.

[19] E. Weyer, R. C. Williamson, and I. M. Mareels, "Finite sample properties of linear model identification," *IEEE Trans. Autom. Control*, vol. 44, no. 7, pp. 1370–1383, 1999.

[20] T. Sarkar and A. Rakhlin, "Near optimal finite time identification of arbitrary linear dynamical systems," in *Int. Conf. on Machine Learn.*, 2019, pp. 5610–5618.

[21] M. Simchowitz, H. Mania, S. Tu, M. I. Jordan, and B. Recht, "Learning without mixing: Towards a sharp analysis of linear system identification," in *Conf. On Learn. Theory*, 2018.

[22] Y. Jedra and A. Proutiere, "Finite-time identification of linear systems: Fundamental limits and optimal algorithms," *IEEE Trans. Autom. Control*, vol. 68, no. 5, pp. 2805–2820, 2022.

[23] A. Tsiamis and G. J. Pappas, "Finite sample analysis of stochastic system identification," in *IEEE Conf. Decis. Control*, 2019, pp. 3648–3654.

[24] S. Oymak and N. Ozay, "Non-asymptotic identification of LTI systems from a single trajectory," in *Am. Control Conf.*, 2019, pp. 5655–5661.

[25] T. Sarkar, A. Rakhlin, and M. A. Dahleh, "Finite time LTI system identification," *J. Mach. Learn. Res.*, vol. 22, no. 1, pp. 1186–1246, 2021.

[26] S. Lale, K. Azizzadenesheli, B. Hassibi, and A. Anandkumar, "Finite-time system identification and adaptive control in autoregressive exogenous systems," in *Learn. Dyn. Control*. PMLR, 2021, pp. 967–979.

[27] A. Tsiamis, I. Ziemann, N. Matni, and G. J. Pappas, "Statistical learning theory for control: A finite-sample perspective," *IEEE Control Syst.*, vol. 43, no. 6, pp. 67–97, 2023.

[28] I. Ziemann, A. Tsiamis, B. Lee, Y. Jedra, N. Matni, and G. J. Pappas, "A tutorial on the non-asymptotic theory of system identification," in *IEEE Conf. Decis. and Control*, 2023, pp. 8921–8939.

[29] S. J. Qin, W. Lin, and L. Ljung, "A novel subspace identification approach with enforced causal models," *Automatica*, vol. 41, no. 12, pp. 2043–2053, 2005.

[30] P. Van Overschee and B. De Moor, "A unifying theorem for three subspace system identification algorithms," *Automatica*, vol. 31, no. 12, pp. 1853–1864, 1995.

[31] I. Ziemann, "A note on the smallest eigenvalue of the empirical covariance of causal gaussian processes," *IEEE Trans. Autom. Control*, vol. 69, no. 2, pp. 1372–1376, 2023.

[32] Y. Abbasi-Yadkori, D. Pál, and C. Szepesvári, "Online least squares estimation with self-normalized processes: An application to bandit problems," *arXiv preprint arXiv:1102.2670*, 2011.