2024 IEEE 63rd Conference on Decision and Control (CDC)
December 16-19, 2024. MiCo, Milan, Italy

System Identification for Linear Dynamics with Bilinear Observation Models: An Expectation–Maximization Approach

Diyou Liu and Mohammad Khosravi

***Abstract*—In this paper, we study the system identification problem for linear time-invariant dynamics with bilinear observation models. Accordingly, we consider a suitable parametric description for the system model and formulate the identification problem as estimating the parameters characterizing the mathematical representation of the system through input-output measurement data. To this end, we employ a probabilistic framework aiming to obtain the maximum likelihood estimates of the parameters. Accordingly, we propose utilizing the expectation-maximization approach to improve the tractability of the identification procedure. Through the numerical experiments, we verify the efficacy of the proposed scheme and demonstrate its performance.**

I. Introduction

System identification, as initially introduced in [1], is an active area of research [2] focusing on the theory and methods for data-driven modeling of dynamical systems [3]. Due to the ubiquitous nature of dynamical systems across various science and technology domains and considering the significance of models for prediction and control, system identification has received extensive attention [4]. As a result, numerous methodologies are developed for diverse categories of systems, spanning from linear dynamics to nonlinear dynamical systems [5–8]. The complexity of real systems poses a significant challenge in accurately capturing their behaviors, necessitating the utilization of various tools introduced in statistics and optimization theory, such as different parametric and nonparametric estimation techniques [9–14].

Parametric system identification is a common method that considers a rich class of models with specific structures, and subsequently, the parameters characterizing the model are estimated using measurement data. In this context, various classes of parametric models and a broad range of systems are considered, and different parameter estimation methods are employed. For example, the prediction error method (PEM) [11] is widely used in identification of linear time-invariant systems, which adjusts the parameters iteratively to minimize the norm of the losses, evaluated based on the difference between the model's predicted output and the actual output from the real system. For estimating the parameters, probabilistic frameworks are also commonly utilized. For instance, the Bayesian approach [15] treats parameters as random variables with probability distributions of specific forms. With sufficient data, the Bayesian approach obtains accurate probability distributions of the parameters of interest. As an example, maximum likelihood (ML) estimation [16] is commonly used in system identification, aiming to find parameter values that maximize the likelihood function (i.e., the probability of observing the given measurement data given the parameter values). Another widely used method is maximum a posteriori (MAP) estimation [17]. The MAP estimation method is similar to the ML estimation method, but instead of maximizing the likelihood function, it finds the parameter values that maximize the posterior distribution. MAP estimation is more robust than ML estimation when informative prior is available.

In addition to linear system identification, there is a growing interest in methods for identifying nonlinear systems [18, 19]. For example, in [20], the authors proposed an ML-based algorithm for Hammerstein-Wiener models. Also, in [21], a Bayesian approach-based method is introduced to identify Wiener-Hammerstein models. Besides the study of general nonlinear systems, bilinear systems, as a special and simple case of nonlinear systems, are also widely studied due to their technical tractability and their relevance in various fields [22]. The characteristics of bilinear systems can be exploited in designing effective identification methods. For instance, in [23, 24], subspace techniques are used to identify bilinear state space systems. Additionally, in [25] and [26], a bilinear system in the observability canonical form is considered. In the former paper, the authors proposed an approach using the Kalman filter to estimate the system states and a gradient-based iterative algorithm to identify system parameters. In the latter one, the authors use the Rauch–Tung–Striebel smoother (RTS) to estimate the state variables and the Expectation-Maximization (EM) algorithm to identify parameters. These papers consider systems with bilinear dynamics and linear observation models. Inspired by the Wiener-Hammerstein models and Hammerstein–Wiener models, in the current paper, we consider bilinear systems with linear dynamics and bilinear observation models.

In this paper, we propose a scheme for identifying systems with linear dynamics and bilinear observation models, based on the RTS smoother [27] and the EM algorithm [28]. The system matrices, together with the mean and covariance of noise distributions, are treated as parameters. The scheme consists of two steps. In the first step, state estimates are computed, and a log-likelihood function of the parameters is defined. To estimate the states, we employ a Kalman filter [29], which recursively updates the state estimates

Delft Center for Systems and Control, Delft University of Technology, Delft, Netherlands. Email: {d.liu-9, mohammad.khosravi}@tudelft.nl

by combining the information from previous measurements and new measurements. There are several extensions of Kalman filter to enhance estimation accuracy[30, 31]. In this paper, the RTS smoother is considered. Rather than utilizing only the previous and current measurement data, the RTS smoother also employs estimates of future measurements to smooth the estimates, resulting in more accurate estimation results compared to using the Kalman filter alone. After obtaining these estimates, the expected value of the log-likelihood function of the parameters can be defined. In the second step, with the estimation of states, we compute an optimal parameter that maximizes the mentioned expected value based on the partial derivatives. These two steps are executed iteratively until convergence. The contributions of this paper are summarized as follows.

- This paper formulates a tractable approach to identify systems with linear dynamics and bilinear observation models. Additionally, using RTS smoother and EM algorithm, a scheme is proposed to solve the problem.
- Suitable examples are provided to demonstrate the performance of the proposed scheme. Moreover, simulations are presented to study the performance of the proposed scheme under different SNR levels.

The rest of this paper is structured as follows. In Section II, the main notations used in this paper are listed. In Section III, a tractable identification problem is formulated. In Section IV, we present the proposed scheme and derive the EM-based approach. Finally, in Section V, numerical examples are provided to verify the performance of proposed scheme.

II. Notation

In this paper, we use $\mathbb{Z}$, $\mathbb{Z}_+$, $\mathbb{R}$, and $\mathbb{R}^{n\times m}$ to denote, the set of integers, the set of positive integers, the set of real numbers, and the set of n by m matrices with real value respectively. For a matrix $\mathrm{A}\in\mathbb{R}^{m\times n}$, $\mathrm{vec}(\mathrm{A})$ is denoted as a column vector in $\mathbb{R}^{mn}$ obtained from stacking the columns of the matrix A. Additionally, we denote $\otimes$ as the Kronecker product. The vector 2-norm of a vector $\mathrm{x}\in\mathbb{R}^n$ is denoted as $\|\mathrm{x}\|$. Finally, the conditional probability of A given B is denoted as $p(A|B)$.

III. Identification of Linear Dynamics with Bilinear Observation Models

Consider an *unknown* time-invariant random dynamical system $\mathcal{S}$ with linear dynamics and a bilinear observation model. More precisely, let the process model describing the dynamics of $\mathcal{S}$ be as

$$\mathrm{x}_{k+1}=\mathrm{A}\mathrm{x}_k+\mathrm{B}\mathrm{u}_k+\mathrm{w}_k,\qquad \forall k\in\mathbb{Z}_+,\tag{1}$$

where $\mathrm{x}_k\in\mathbb{R}^{n_\mathrm{x}}$, $\mathrm{u}_k\in\mathbb{R}^{n_\mathrm{u}}$, and $\mathrm{w}_k\in\mathbb{R}^{n_\mathrm{x}}$ are respectively the vectors of state variables, input, and process noise, at time instant $k\in\mathbb{Z}_+$, and, $\mathrm{A}\in\mathbb{R}^{n_\mathrm{x}\times n_\mathrm{x}}$ and $\mathrm{B}\in\mathbb{R}^{n_\mathrm{x}\times n_\mathrm{u}}$ are *unknown* matrices characterizing the dynamics of system. Also, let the observation model of the system have a bilinear form as

$$\mathrm{y}_{k+1}=\big(\mathrm{C}_0+\sum_{i=1}^{n_\mathrm{u}}\mathrm{C}_i u_{k,i}\big)\mathrm{x}_k+\mathrm{D}\mathrm{u}_k+\mathrm{v}_k,\qquad \forall k\in\mathbb{Z}_+,\tag{2}$$

where $u_{k,i}$ denotes the i^{th} entry of u_k, $\forall i=1,\dots,n_\mathrm{u}$, $k\in\mathbb{Z}_+$, $\mathrm{y}_k\in\mathbb{R}^{n_\mathrm{y}}$ and $\mathrm{v}_k\in\mathbb{R}^{n_\mathrm{y}}$ are respectively the vectors of output observations and measurement noise, and, $\mathrm{C}_0,\mathrm{C}_1,\dots,\mathrm{C}_{n_\mathrm{u}}\in\mathbb{R}^{n_\mathrm{y}\times n_\mathrm{x}}$ and $\mathrm{D}\in\mathbb{R}^{n_\mathrm{y}\times n_\mathrm{u}}$ are *unknown* matrices describing the observation model of the system. Suppose the initial state x_0, measurement noise $\big(\mathrm{v}_k\big)_{k\in\mathbb{Z}_+}$, and process noise $\big(\mathrm{w}_k\big)_{k\in\mathbb{Z}_+}$ are mutually independent Gaussian random variables as

$$\mathrm{x}_0\sim\mathcal{N}(\mu_{\mathrm{x}_0},\mathrm{S}_{\mathrm{x}_0}),\tag{3}$$
$$\mathrm{v}_k\sim\mathcal{N}(0,\mathrm{S}_\mathrm{v}),\qquad\forall k\in\mathbb{Z}_+,\tag{4}$$
$$\mathrm{w}_k\sim\mathcal{N}(0,\mathrm{S}_\mathrm{w}),\qquad\forall k\in\mathbb{Z}_+,\tag{5}$$

where $\mu_{\mathrm{x}_0}\in\mathbb{R}^{n_\mathrm{x}}$ is an *unknown* vector and $\mathrm{S}_{\mathrm{x}_0}\in\mathbb{R}^{n_\mathrm{x}\times n_\mathrm{x}}$ is an *unknown* positive definite matrix respectively denoting the mean and covariance of x_0, and, $\mathrm{S}_\mathrm{w}\in\mathbb{R}^{n_\mathrm{x}\times n_\mathrm{x}}$ and $\mathrm{S}_\mathrm{v}\in\mathbb{R}^{n_\mathrm{y}\times n_\mathrm{y}}$ are *unknown* positive definite matrices representing respectively the covariance of vector w_k and the covariance of vector v_k, for any $k\in\mathbb{Z}_+$.

Assume a measurement dataset of $n_\mathcal{D}\in\mathbb{N}$ input-output pairs is given as

$$\mathcal{D}:=\big\{(\mathrm{u}_k,\mathrm{y}_k)\,|\,k\in[0,n_\mathcal{D}-1]\big\}.\tag{6}$$

Accordingly, we present the main problem as identifying system $\mathcal{S}$ through estimating the unknown vector and matrices mentioned above using the dataset $\mathcal{D}$. For the ease of discussion, assume $\mathrm{D}=0$ throughout this paper.

Problem (Identification Problem for Linear Dynamics with Bilinear Observation Models). *Given the measurement set of data $\mathcal{D}$, estimate* A, B, C_0, C_1, …, $\mathrm{C}_{n_\mathrm{u}}$, μ_{x_0}, $\mathrm{S}_{\mathrm{x}_0}$, S_w, *and* S_v.

To address this problem, we employ an expectation-maximization (EM) approach [32] and obtain a tractable procedure. More details are discussed in the next section.

IV. An Expectation-Maximization Approach

This section presents an expectation-maximization (EM) algorithm for identifying linear time-invariant dynamics with bilinear observation models. The EM algorithm is a powerful iterative estimation scheme, particularly beneficial when deriving the solution of maximum likelihood estimation is computationally intractable. The EM approach has an iterative procedure where each of its iterations consists of two main steps: the expectation (E) step and the maximization (M) step. In the expectation (E) step, given measurement data and an initial estimation for the parameters of the system, a probability distribution for the trajectory of state variables is obtained, e.g., through Rauch-Tung-Striebel smoother [27]. Subsequently, in the maximization (M) step, using these state estimates and the resulting distributions, the system parameter estimates are updated by maximizing the expected log-likelihood function obtained from the distributions estimated in the E step.

A. Rauch–Tung–Striebel smoother

The RTS smoother is widely used in states variables estimation, which adds an additional smoothing part using the estimated states distributions computed by the Kalman filter [29].

For the ease of discussion, we define vector of parameters, denoted by θ, as

$$\begin{aligned}\theta := [&\text{vec}(\mathrm{A})^\mathsf{T},\text{vec}(\mathrm{B})^\mathsf{T},\text{vec}(\mathrm{C}_0)^\mathsf{T},...,\text{vec}(\mathrm{C}_{n_\mathrm{u}})^\mathsf{T},\\ &\text{vec}(\mu_{\mathrm{x}_0})^\mathsf{T},\text{vec}(\mathrm{S}_{\mathrm{x}_0})^\mathsf{T},\text{vec}(\mathrm{S}_\mathrm{w})^\mathsf{T},\text{vec}(\mathrm{S}_\mathrm{v})^\mathsf{T}].\end{aligned} \tag{7}$$

Given measurement data $\mathcal{D}$, we want to obtain the estimates of θ, which is denoted as $\hat{\theta}$. From the Kalman filter, we can compute the estimated mean

$$\hat{\mathrm{x}}_{t|t} = \mathbb{E}[\mathrm{x}_t|\mathrm{y}_0,...,\mathrm{y}_t,\hat{\theta}], \tag{8}$$

and, the estimated covariance

$$\mathrm{P}_{t|t} = \mathbb{E}[(\mathrm{x}_t-\hat{\mathrm{x}}_{t|t})(\mathrm{x}_t-\hat{\mathrm{x}}_{t|t})^\mathsf{T}|\mathrm{y}_0,...,\mathrm{y}_t,\hat{\theta}] \tag{9}$$

using following iterative scheme

$$\begin{aligned}\hat{\mathrm{x}}_{t|t} &= \hat{\mathrm{x}}_{t|t-1} + \mathrm{K}_t(\mathrm{y}_t - \hat{\Xi}_t\hat{\mathrm{x}}_{t|t-1}),\\ \mathrm{P}_{t|t} &= (\mathbb{I}_{n_\mathrm{x}} - \mathrm{K}_t\hat{\Xi}_t)\mathrm{P}_{t|t-1},\\ \mathrm{K}_t &= \mathrm{P}_{t|t-1}\hat{\Xi}_t^\mathsf{T}[\hat{\Xi}_t\mathrm{P}_{t|t-1}\hat{\Xi}_t^\mathsf{T} + \hat{\mathrm{S}}_\mathrm{v}]^{-1},\\ \hat{\mathrm{x}}_{t+1|t} &= \hat{\mathrm{A}}\hat{\mathrm{x}}_{t|t} + \hat{\mathrm{B}}\mathrm{u}_t,\\ \mathrm{P}_{t+1|t} &= \hat{\mathrm{A}}\mathrm{P}_{t|t}\hat{\mathrm{A}}^\mathsf{T} + \hat{\mathrm{S}}_\mathrm{w},\end{aligned} \tag{10}$$

where $\hat{\Xi}_t$ is defined as

$$\hat{\Xi}_t = \hat{\mathrm{C}}_0 + \sum_{i=1}^{n_\mathrm{u}} \hat{\mathrm{C}}_i u_{t,i}. \tag{11}$$

Define $\mathrm{X} := \{\mathrm{x}_0,\mathrm{x}_1,..,\mathrm{x}_{n_\mathcal{D}}\}$, $\mathrm{Y} := \{\mathrm{y}_0,\mathrm{y}_1,..,\mathrm{y}_{n_\mathcal{D}-1}\}$ and $\mathrm{U} := \{\mathrm{u}_0,\mathrm{u}_1,..,\mathrm{u}_{n_\mathcal{D}-1}\}$. Using the RTS smoother, the estimated mean $\hat{\mathrm{x}}_{t|n_\mathcal{D}} = \mathbb{E}[\mathrm{x}_t|\mathrm{Y},\hat{\theta}]$ and estimated covariance $\mathrm{P}_{t|n_\mathcal{D}} = \mathbb{E}[(\mathrm{x}_t-\hat{\mathrm{x}}_{t|n_\mathcal{D}})(\mathrm{x}_t-\hat{\mathrm{x}}_{t|n_\mathcal{D}})^\mathsf{T}|\mathrm{Y},\hat{\theta}]$ can be recursively computed as

$$\begin{aligned}\hat{\mathrm{x}}_{t|n_\mathcal{D}} &= \hat{\mathrm{x}}_{t|t} + \mathrm{H}_t(\hat{\mathrm{x}}_{t+1|n_\mathcal{D}} - \hat{\mathrm{x}}_{t+1|t}),\\ \mathrm{P}_{t|n_\mathcal{D}} &= \mathrm{P}_{t|t} + \mathrm{H}_t(\mathrm{P}_{t+1|n_\mathcal{D}} - \mathrm{P}_{t+1|t})\mathrm{H}_t^\mathsf{T},\\ \mathrm{H}_t &= \mathrm{P}_{t|t}\hat{\mathrm{A}}^\mathsf{T}\mathrm{P}_{t+1|t}^{-1},\end{aligned} \tag{12}$$

where the initial conditions $\hat{\mathrm{x}}_{n_\mathcal{D}|n_\mathcal{D}}$ and $\mathrm{P}_{n_\mathcal{D}|n_\mathcal{D}}$ can be obtained from the results of Kalman filter. Moreover, we can further derive the following estimations [33]

$$\begin{aligned}\mathbb{E}[\mathrm{x}_t\mathrm{x}_t^\mathsf{T}|\mathrm{Y},\hat{\theta}] &= \hat{\mathrm{x}}_{t|n_\mathcal{D}}\hat{\mathrm{x}}_{t|n_\mathcal{D}}^\mathsf{T} + \mathrm{P}_{t|n_\mathcal{D}},\\ \mathbb{E}[\mathrm{x}_{t+1}\mathrm{x}_t^\mathsf{T}|\mathrm{Y},\hat{\theta}] &= \hat{\mathrm{x}}_{t+1|n_\mathcal{D}}\hat{\mathrm{x}}_{t|n_\mathcal{D}}^\mathsf{T} + \mathrm{P}_{t+1|n_\mathcal{D}}\mathrm{H}_t^\mathsf{T},\end{aligned} \tag{13}$$

which will be used in EM algorithm as discussed in the remainder of this section.

B. Parameters Estimation using EM algorithm

The EM contains two basic steps, which are discussed in the sequel. The expectation step computes a log likelihood function of parameters from the current estimates and given data. The maximization step finds the parameters that maximize the log likelihood function. In this work, we use EM algorithm with RTS smoother to estimate parameters of dynamical system $\mathcal{S}$.

Denote $\hat{\theta}_k$ as the estimated parameters at iteration k and define

$$Q(\theta|\hat{\theta}_k) = \mathbb{E}_{p(\mathrm{X}|\mathrm{Y},\hat{\theta}_k)}\Big[\ln p(\mathrm{X},\mathrm{Y}|\theta)\Big], \tag{14}$$

Instead of directly optimizing the log-likelihood $\ln p(\mathrm{Y}|\theta)$, EM algorithm iteratively improves $Q(\theta|\hat{\theta}_k)$ at each step [9], which guarantees an improvement in $\ln p(\mathrm{Y}|\theta)$ at least as much [12]. Consider the likelihood function

$$p(\mathrm{X},\mathrm{Y}|\theta) = p(\mathrm{X}|\theta)p(\mathrm{Y}|\mathrm{X},\theta). \tag{15}$$

Using Markov property, $p(\mathrm{X}|\theta)$ and $p(\mathrm{Y}|\mathrm{X},\theta)$ can be decomposed as

$$p(\mathrm{X}|\theta) = p(\mathrm{x}_0|\theta)\prod_{t=1}^{n_\mathcal{D}} p(\mathrm{x}_t|\mathrm{x}_{t-1},\theta). \tag{16}$$

and

$$p(\mathrm{Y}|\mathrm{X},\theta) = \prod_{t=0}^{n_\mathcal{D}-1} p(\mathrm{y}_t|\mathrm{X},\theta) = \prod_{t=0}^{n_\mathcal{D}-1} p(\mathrm{y}_t|\mathrm{x}_t,\theta), \tag{17}$$

Thus, the log-likelihood function of (15) can be derived as

$$\begin{aligned}\ln p(\mathrm{X},\mathrm{Y}|\theta) =& \ln p(\mathrm{x}_0|\theta) + \sum_{t=0}^{n_\mathcal{D}-1}\ln p(\mathrm{y}_t|\mathrm{x}_t,\theta)\\ &+ \sum_{t=1}^{n_\mathcal{D}}\ln p(\mathrm{x}_t|\mathrm{x}_{t-1},\theta).\end{aligned} \tag{18}$$

From the dynamics function in (2), $p(\mathrm{y}_t|\mathrm{x}_t,\theta)$ is subject to a Gaussian distribution $\mathcal{N}(\mu_{\mathrm{y}_t},\mathrm{S}_\mathrm{v})$, where

$$\begin{aligned}\mu_{\mathrm{y}_t} =& \Xi_t\mathrm{x}_t,\\ \Xi_t =& \mathrm{C}_0 + \sum_{i=1}^{n_\mathrm{u}}\mathrm{C}_i u_{t,i}.\end{aligned} \tag{19}$$

Analogously from (1), $p(\mathrm{x}_t|\theta)$ is subject to a Gaussian distribution $\mathcal{N}(\mu_{\mathrm{x}_t},\mathrm{S}_\mathrm{w})$, where $\mu_{\mathrm{x}_t} = \mathrm{A}\mathrm{x}_{t-1} + \mathrm{B}\mathrm{u}_{t-1}$, $\forall t \geq 1$. Therefore, using these Gaussian distributions, $Q(\theta|\hat{\theta}_k)$ can be rewritten as

$$\begin{aligned}Q(\theta|\hat{\theta}_k) = \mathbb{E}_{p(\mathrm{X}|\mathrm{Y},\hat{\theta}_k)}\Big\{ &-\frac{n_\mathcal{D}}{2}\ln\det\mathrm{S}_\mathrm{v} - \frac{1}{2}\ln\det\mathrm{S}_{\mathrm{x}_0}\\ &-\frac{n_\mathcal{D}}{2}\ln\det\mathrm{S}_\mathrm{w} - \frac{1}{2}\sum_{t=0}^{n_\mathcal{D}-1}(\mathrm{y}_t-\Xi_t\mathrm{x}_t)^\mathsf{T}\mathrm{S}_\mathrm{v}^{-1}(\mathrm{y}_t-\Xi_t\mathrm{x}_t)\\ &-\frac{1}{2}\sum_{t=1}^{n_\mathcal{D}}(\mathrm{x}_{t+1}-\mathrm{A}\mathrm{x}_t-\mathrm{B}\mathrm{u}_t)^\mathsf{T}\mathrm{S}_\mathrm{w}^{-1}(\mathrm{x}_{t+1}-\mathrm{A}\mathrm{x}_t-\mathrm{B}\mathrm{u}_t)\\ &-\frac{1}{2}(\mathrm{x}_0-\mu_{\mathrm{x}_0})^\mathsf{T}\mathrm{S}_{\mathrm{x}_0}^{-1}(\mathrm{x}_0-\mu_{\mathrm{x}_0})\Big\}.\end{aligned} \tag{20}$$

For the convenience of notation, we define $\mathrm{M} = [\mathrm{A},\mathrm{B}]$, $\mathrm{C} = [\mathrm{C}_0,\mathrm{C}_1,...,\mathrm{C}_{n_\mathrm{u}}]$, and $\mathrm{z}_t = \begin{bmatrix}\mathrm{x}_t\\ \mathrm{u}_t\end{bmatrix}$, for all $t = 0,\ldots,n_\mathcal{D}-1$.

Using the property that $\mathrm{x}^{\mathsf{T}}\mathrm{M}\mathrm{x} = \mathrm{tr}(\mathrm{M}\mathrm{x}\mathrm{x}^{\mathsf{T}}), \forall \mathrm{M} \in \mathbb{R}^{n_{\mathrm{x}} \times n_{\mathrm{x}}}$, we can further simplify $Q(\theta|\hat{\theta}_k)$ as

$$\begin{aligned} Q(\theta|\hat{\theta}_k) = \mathbb{E}_{p(\mathrm{X}|\mathrm{Y},\hat{\theta}_k)}\Big\{ & -\frac{n_{\mathcal{D}}}{2}\ln\det \mathrm{S}_{\mathrm{v}} - \frac{1}{2}\ln\det \mathrm{S}_{\mathrm{x}_0} \\ & -\frac{n_{\mathcal{D}}}{2}\ln\det \mathrm{S}_{\mathrm{w}} - \frac{1}{2}\sum_{t=0}^{n_{\mathcal{D}}-1} \mathrm{tr}\{\mathrm{S}_{\mathrm{v}}^{-1}[\mathrm{y}_t - \mathrm{C}(\begin{bmatrix}1\\ \mathrm{u}_t\end{bmatrix} \otimes \mathrm{x}_t)] \\ & [\mathrm{y}_t - \mathrm{C}(\begin{bmatrix}1\\ \mathrm{u}_t\end{bmatrix} \otimes \mathrm{x}_t)]^{\mathsf{T}}\} - \frac{1}{2}\sum_{t=1}^{n_{\mathcal{D}}} \mathrm{tr}\{\mathrm{S}_{\mathrm{w}}^{-1}(\mathrm{x}_{t+1} - \mathrm{M}\mathrm{z}_t) \\ & (\mathrm{x}_{t+1} - \mathrm{M}\mathrm{z}_t)^{\mathsf{T}}\} - \frac{1}{2}\mathrm{tr}\{\mathrm{S}_{\mathrm{x}_0}^{-1}(\mathrm{x}_0 - \mu_{\mathrm{x}_0})(\mathrm{x}_0 - \mu_{\mathrm{x}_0})^{\mathsf{T}}\}\Big\}. \end{aligned} \tag{21}$$

To obtain the optimal θ which maximizes $Q(\theta|\hat{\theta}_k)$, we take the partial derivative of $Q(\theta|\hat{\theta}_k)$ with respect to each parameter and set it equal to zero. Based on the calculation in Appendix A, the optimal parameters can be found as

$$\begin{aligned} \hat{\mathrm{C}}_{k+1} = & \sum_{t=0}^{n_{\mathcal{D}}-1} \mathrm{y}_t(\begin{bmatrix}1\\ \mathrm{u}_t\end{bmatrix} \otimes \hat{\mathrm{x}}_{t|n_{\mathcal{D}}})^{\mathsf{T}} \\ & \Big(\sum_{t=0}^{n_{\mathcal{D}}-1} \begin{bmatrix}1\\ \mathrm{u}_t\end{bmatrix}\begin{bmatrix}1\\ \mathrm{u}_t\end{bmatrix}^{\mathsf{T}} \otimes (\hat{\mathrm{x}}_{t|n_{\mathcal{D}}}\hat{\mathrm{x}}_{t|n_{\mathcal{D}}}^{\mathsf{T}} + \mathrm{P}_{t|n_{\mathcal{D}}})\Big)^{-1}, \\ \hat{\mathrm{S}}_{\mathrm{v},k+1} = & \frac{1}{n_{\mathcal{D}}}\sum_{t=0}^{n_{\mathcal{D}}-1} \mathrm{y}_t\mathrm{y}_t^{\mathsf{T}} - \frac{1}{n_{\mathcal{D}}}\sum_{t=0}^{n_{\mathcal{D}}-1} \mathrm{y}_t(\begin{bmatrix}1\\ \mathrm{u}_t\end{bmatrix} \otimes \hat{\mathrm{x}}_{t|n_{\mathcal{D}}})^{\mathsf{T}} \\ & \Big(\sum_{t=0}^{n_{\mathcal{D}}-1} \begin{bmatrix}1\\ \mathrm{u}_t\end{bmatrix}\begin{bmatrix}1\\ \mathrm{u}_t\end{bmatrix}^{\mathsf{T}} \otimes (\hat{\mathrm{x}}_{t|n_{\mathcal{D}}}\hat{\mathrm{x}}_{t|n_{\mathcal{D}}}^{\mathsf{T}} + \mathrm{P}_{t|n_{\mathcal{D}}})\Big)^{-1} \\ & \sum_{t=0}^{n_{\mathcal{D}}-1} (\begin{bmatrix}1\\ \mathrm{u}_t\end{bmatrix} \otimes \hat{\mathrm{x}}_{t|n_{\mathcal{D}}})\mathrm{y}_t^{\mathsf{T}}, \\ \hat{\mathrm{M}}_{k+1} = & \sum_{t=1}^{n_{\mathcal{D}}} (\begin{bmatrix}\hat{\mathrm{x}}_{t+1|n_{\mathcal{D}}}\hat{\mathrm{x}}_{t|n_{\mathcal{D}}}^{\mathsf{T}} + \mathrm{P}_{t+1|n_{\mathcal{D}}}\mathrm{H}_t^{\mathsf{T}} & \hat{\mathrm{x}}_{t+1|n_{\mathcal{D}}}\mathrm{u}_t^{\mathsf{T}}\end{bmatrix} \\ & \begin{bmatrix}\hat{\mathrm{x}}_{t|n_{\mathcal{D}}}\hat{\mathrm{x}}_{t|n_{\mathcal{D}}}^{\mathsf{T}} + \mathrm{P}_{t|n_{\mathcal{D}}} & \hat{\mathrm{x}}_{t|n_{\mathcal{D}}}\mathrm{u}_t^{\mathsf{T}} \\ \mathrm{u}_t\hat{\mathrm{x}}_{t|n_{\mathcal{D}}}^{\mathsf{T}} & \mathrm{u}_t\mathrm{u}_t^{\mathsf{T}}\end{bmatrix}^{-1}, \\ \hat{\mathrm{S}}_{\mathrm{w},k+1} = & \frac{1}{n_{\mathcal{D}}}\sum_{t=0}^{n_{\mathcal{D}}-1} (\hat{\mathrm{x}}_{t+1|n_{\mathcal{D}}}\hat{\mathrm{x}}_{t+1|n_{\mathcal{D}}}^{\mathsf{T}} + \mathrm{P}_{t+1|n_{\mathcal{D}}}) \\ & -\frac{1}{n_{\mathcal{D}}}\sum_{t=0}^{n_{\mathcal{D}}-1} \begin{bmatrix}\hat{\mathrm{x}}_{t+1|n_{\mathcal{D}}}\hat{\mathrm{x}}_{t|n_{\mathcal{D}}}^{\mathsf{T}} + \mathrm{P}_{t+1|n_{\mathcal{D}}}\mathrm{H}_t^{\mathsf{T}} & \hat{\mathrm{x}}_{t+1|N}\mathrm{u}_t^{\mathsf{T}}\end{bmatrix} \\ & \Big(\sum_{t=0}^{n_{\mathcal{D}}-1} \begin{bmatrix}\hat{\mathrm{x}}_{t|n_{\mathcal{D}}}\hat{\mathrm{x}}_{t|n_{\mathcal{D}}}^{\mathsf{T}} + \mathrm{P}_{t|n_{\mathcal{D}}} & \hat{\mathrm{x}}_{t|n_{\mathcal{D}}}\mathrm{u}_t^{\mathsf{T}} \\ \mathrm{u}_t\hat{\mathrm{x}}_{t|n_{\mathcal{D}}}^{\mathsf{T}} & \mathrm{u}_t\mathrm{u}_t^{\mathsf{T}}\end{bmatrix}\Big)^{-1} \\ & \sum_{t=0}^{n_{\mathcal{D}}-1} \begin{bmatrix}\hat{\mathrm{x}}_{t+1|n_{\mathcal{D}}}\hat{\mathrm{x}}_{t|n_{\mathcal{D}}}^{\mathsf{T}} + \mathrm{P}_{t+1|n_{\mathcal{D}}}\mathrm{H}_t^{\mathsf{T}} & \hat{\mathrm{x}}_{t+1|n_{\mathcal{D}}}\mathrm{u}_t^{\mathsf{T}}\end{bmatrix}^{\mathsf{T}}, \\ \hat{\mu}_{x_0,k+1} = & \hat{\mathrm{x}}_{0|n_{\mathcal{D}}}, \\ \hat{\mathrm{S}}_{\mathrm{x}_0,k+1} = & \mathrm{P}_{0|n_{\mathcal{D}}}. \end{aligned} \tag{22}$$

Remark 1. Equation (22) implies there is one optimal parameter θ which let the partial derivatives to be zero. Intuitively, if we only consider S_{w} and M and assume $n_{\mathrm{x}} = 1$, $Q(\theta|\hat{\theta}_k)$ can be simplified as $\mathbb{E}_{p(\mathrm{X}|\mathrm{Y},\hat{\theta}_k)} - \frac{n_{\mathcal{D}}}{2}\ln \mathrm{S}_{\mathrm{w}} - \frac{1}{2}\sum_{t=1}^{n_{\mathcal{D}}} \frac{(\mathrm{x}_{t+1} - \mathrm{M}\mathrm{z}_t)^2}{\mathrm{S}_{\mathrm{w}}}$. As $f(x) = a\ln x + \frac{b}{x}, \forall a, b \in \mathbb{R}^-$ only has one maximum at $x = \frac{b}{a}$, there is only one optimal S_{w}. In addition, for any given S_{w}, there is also only one optimal M. For other parameters, using the similar method, there is one local maximum which is also global maximum for $Q(\theta|\hat{\theta}_k)$.

Algorithm 1 An EM Approach for Identification of Linear Dynamics with Bilinear Observation Models

Input: $\mathcal{D}$.
Output: θ.
Initial guess: $\hat{\theta}_0$
$k \leftarrow 0$
while 1 do
  Current parameters estimates: $\hat{\theta} \leftarrow \hat{\theta}_k$
  for $t \leftarrow 0$ to $n_{\mathcal{D}}$ **do**
    Kalman Filter: (10)
  for $t \leftarrow n_{\mathcal{D}}$ to 0 **do**
    RTS smoother: (12)
  EM approach: compute (22) to find a new parameters estimates. $\hat{\theta}_{k+1} \leftarrow \arg\max_{\theta} Q(\theta|\hat{\theta}_k)$.
  if $\|\hat{\theta}_{k+1} - \hat{\theta}_k\| < \epsilon$ **then**
    break
  else
    $k \leftarrow k + 1$
$\theta \leftarrow \hat{\theta}_k$

Algorithm 1 summarizes the introduced EM based approach for identification of linear dynamics with bilinear observation models.

V. Numerical Experiments

In this section, we evaluate the performance of Algorithm 1 through numerical examples. Accordingly, we consider a time-invariant system as introduced in (1)-(2) with matrices A, B, C_0 and C_1 as

$$\begin{aligned} \mathrm{A} &= \begin{bmatrix}0.85 & -0.4\\ 0.35 & 0.65\end{bmatrix}, & \mathrm{B} &= \begin{bmatrix}0.5\\ -0.5\end{bmatrix}, \\ \mathrm{C}_0 &= \begin{bmatrix}0.5 & -0.2\end{bmatrix}, & \mathrm{C}_1 &= \begin{bmatrix}0.15 & 0.1\end{bmatrix}. \end{aligned} \tag{23}$$

Additionally, for process and measurement noise covariance, and prior of the initial state, we assume that

$$\begin{aligned} \mathrm{S}_{\mathrm{w}} &= \begin{bmatrix}0.05^2 & 0\\ 0 & 0.05^2\end{bmatrix}, & \mathrm{S}_{\mathrm{x}_0} &= \begin{bmatrix}0.05^2 & 0\\ 0 & 0.05^2\end{bmatrix}, \\ \mathrm{S}_{\mathrm{v}} &= 0.05^2, & \mu_{\mathrm{x}_0} &= \begin{bmatrix}1\\ 1\end{bmatrix}. \end{aligned} \tag{24}$$

A random binary signal is generated as the control input U for the system, with a length of $n_{\mathcal{D}} = 1000$. Following the problem settings introduced in Section III, the process noise and measurement noise are generated from Gaussian distributions $\mathcal{N}(\mathbf{0}, \mathrm{S}_{\mathrm{w}})$ and $\mathcal{N}(\mathbf{0}, \mathrm{S}_{\mathrm{v}})$, respectively. In this experiment, we set a stopping condition as $\|\hat{\theta}_{k+1} - \hat{\theta}_k\| < \epsilon$, where $\epsilon = 10^{-4}$.

Figure 1 illustrates the performance of Algorithm 1. In this figure, we plot the trajectories of relative errors $\|\hat{\mathrm{C}} - \mathrm{C})\|/\|\mathrm{C}\|$ and $\|\hat{\mathrm{M}} - \mathrm{M}\|/\|\mathrm{M}\|$, where $\mathrm{C} = \begin{bmatrix}\mathrm{C}_0 & \mathrm{C}_1\end{bmatrix}$ and $\mathrm{M} = \begin{bmatrix}\mathrm{A} & \mathrm{B}\end{bmatrix}$ as defined in (21). It can be seen that,

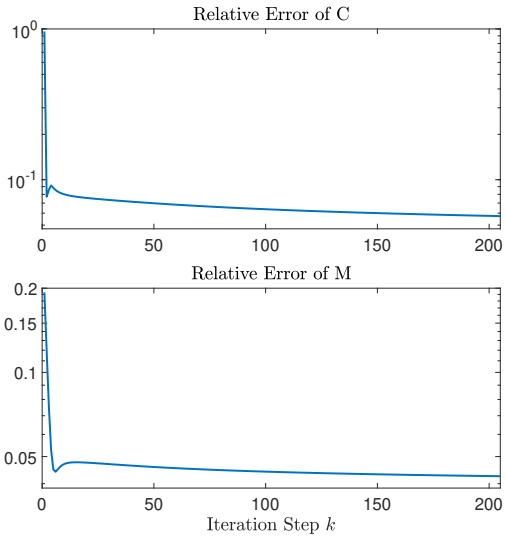

Figure 1. The relative error of system matrices estimates with respect to iteration steps.

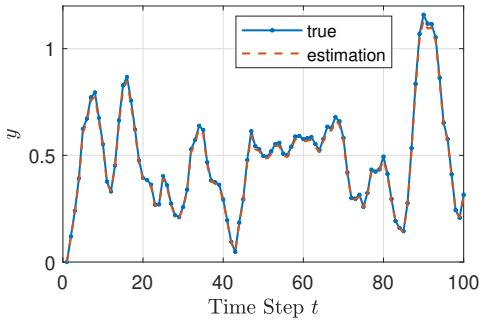

Figure 2. Comparison of real system outputs and identified system outputs.

initially, the relative error is large due to the gap between the initial guess and the true parameters. However, the relative error diminishes with each iteration. The convergence rate is fast in the initial iterations, however, it gets slower later. In Figure 1, for the sake of clarity, we have shown the first 200 iterations. Nonetheless, for $\epsilon = 10^{-4}$, the algorithm stops after about 650 EM steps. Figure 2 shows the performance of the proposed algorithm. To this end, another random input sequence with a length $T = 100$ is generated to validate the identified system. It can be seen that the predicted trajectory of outputs for the identified system is close to that of the real system. The relatively error, computed as $\sum_{t=0}^{T} \frac{\|\mathrm{y}_t - \hat{\mathrm{y}}_t\|}{\|\mathrm{y}_t\|}$, is approximately 0.02.

Remark 2. As shown in Figure 1, the convergence speed for EM algorithm is relatively slow. As discussed after (14), with each iteration step, improving $Q(\theta|\hat{\theta}_k)$ will improve the log-likelihood $\ln p(\mathrm{Y}|\theta)$. However, the rate of improvement can be slow. Furthermore, there is no assurance that the log-likelihood will reach the global optimum. More precisely, the EM algorithm may converge to a local optimum. Consequently, in simulations, it may be necessary to restart the algorithm with different initial guesses to obtain accurate parameter estimates.

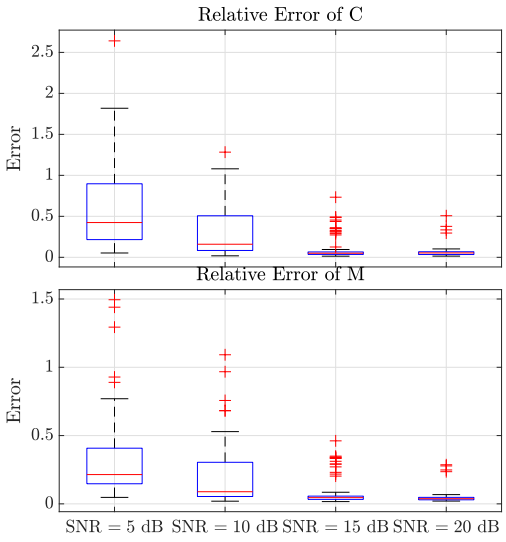

Figure 3. Relative error of system matrices under four different SNRs.

To further investigate the performance of Algorithm 1, we analyze it under four different SNR levels and perform a Monte Carlo experiment with respect to each of these levels. Given the same system (23), prior knowledge (24) and the same input sequences, 100 different noise sequences are generated under four different SNR levels: 5 dB, 10 dB, 15 dB, and 20 dB. For each realization, the proposed algorithm is used to identify the system. Figure 3 demonstrates the performance for estimating system matrices. As shown in Figure 3, in general, the mean relative errors of both C and M are lower for scenarios with higher SNRs, and consequently, the accuracy of identified parameters is higher. When SNR $= 5$ dB, the estimated system matrices are the worst. In this case, the performance depends considerably on the initial guess. Indeed, for some realizations, the relative errors are low; meanwhile, for other initial guesses, the relative errors may be higher, resulting in higher covariance. Moreover, for SNR $= 15$ dB or 20 dB, the covariance of error is relatively small, indicating that the algorithm is less dependent on the initial guess.

VI. Conclusion

In this paper, we have studied the identification problem for unknown dynamical systems with linear dynamics and bilinear observation models, and proposed a tractable identification procedure based on EM procedure. In the proposed scheme, the RTS smoother is used to find the estimates of the states, while the EM-based approach is introduced to estimate the unknown system parameters. The performance of the introduced scheme is evaluated through a numerical example. Furthermore, we also compare the performance of the proposed scheme under different SNRs through a Monte Carlo numerical experiment.

Appendix

To obtain the optimal solution, we use first order necessary condition. Accordingly, we compute the partial deriva-

tives with respect to each parameter. Additionally, from Kalman filter and RTS smoother in (12) and (13), we can obtain $\mathbb{E}_{p(\mathrm{X}|\mathrm{Y},\hat{\theta}_k)}[\mathrm{x}_t|\mathrm{Y},\hat{\theta}_k]$, $\mathbb{E}_{p(\mathrm{X}|\mathrm{Y},\hat{\theta}_k)}[\mathrm{x}_t\mathrm{x}_t^\mathsf{T}|\mathrm{Y}\hat{\theta}_k]$, and $\mathbb{E}_{p(\mathrm{X}|\mathrm{Y},\hat{\theta}_k)}[\mathrm{x}_{t+1}\mathrm{x}_t^\mathsf{T}|\mathrm{Y},\hat{\theta}_k]$. Thus, the partial derivatives are

$$\begin{aligned}
\frac{\partial Q(\theta|\hat{\theta}_k)}{\partial \mathrm{C}} =& \mathrm{S}_\mathrm{v}^{-1}\Big\{\sum_{t=0}^{n_\mathcal{D}-1}\mathrm{y}_t\Big(\begin{bmatrix}1\\ \mathrm{u}_t\end{bmatrix}\otimes\hat{\mathrm{x}}_{t|n_\mathcal{D}}\Big)^\mathsf{T}-\mathrm{C}\\
&\Big(\sum_{t=0}^{n_\mathcal{D}-1}\Big(\begin{bmatrix}1\\ \mathrm{u}_t\end{bmatrix}\begin{bmatrix}1\\ \mathrm{u}_t\end{bmatrix}^\mathsf{T}\Big)\otimes\big(\hat{\mathrm{x}}_{t|n_\mathcal{D}}\hat{\mathrm{x}}_{t|n_\mathcal{D}}^\mathsf{T}+\mathrm{P}_{t|n_\mathcal{D}}\big)\Big)\Big\},\\
\frac{\partial Q(\theta|\hat{\theta}_k)}{\partial \mathrm{S}_\mathrm{v}} =& \frac{1}{2}\sum_{t=0}^{n_\mathcal{D}-1}\Big\{-n_\mathcal{D}\mathrm{S}_\mathrm{v}^{-1}+\mathrm{S}_\mathrm{v}^{-1}\Big[\mathrm{y}_t\mathrm{y}_t^\mathsf{T}\\
&-\mathrm{C}\Big(\begin{bmatrix}1\\ \mathrm{u}_t\end{bmatrix}\otimes\hat{\mathrm{x}}_{t|n_\mathcal{D}}\Big)\mathrm{y}_t^\mathsf{T}-\mathrm{y}_t\Big(\begin{bmatrix}1\\ \mathrm{u}_t\end{bmatrix}\otimes\hat{\mathrm{x}}_{t|n_\mathcal{D}}\Big)^\mathsf{T}\mathrm{C}^\mathsf{T}\\
&+\mathrm{C}\Big(\sum_{t=0}^{n_\mathcal{D}-1}\begin{bmatrix}1\\ \mathrm{u}_t\end{bmatrix}\begin{bmatrix}1\\ \mathrm{u}_t\end{bmatrix}^\mathsf{T}\otimes\big(\hat{\mathrm{x}}_{t|n_\mathcal{D}}\hat{\mathrm{x}}_{t|n_\mathcal{D}}^\mathsf{T}+\mathrm{P}_{t|n_\mathcal{D}}\big)\Big)\\
&\mathrm{C}^\mathsf{T}\Big]\mathrm{S}_\mathrm{v}^{-1}\Big\},\\
\frac{\partial Q(\theta|\hat{\theta}_k)}{\partial \mathrm{M}} =& \mathrm{S}_\mathrm{w}^{-1}\Big[\sum_{t=1}^{n_\mathcal{D}}\Big(\begin{bmatrix}\hat{\mathrm{x}}_{t+1|n_\mathcal{D}}\hat{\mathrm{x}}_{t|n_\mathcal{D}}^\mathsf{T}+\mathrm{P}_{t+1|n_\mathcal{D}}\mathrm{H}_t^\mathsf{T} & \hat{\mathrm{x}}_{t+1|n_\mathcal{D}}\mathrm{u}_t^\mathsf{T}\end{bmatrix}\\
&-\mathrm{M}\begin{bmatrix}\hat{\mathrm{x}}_{t|n_\mathcal{D}}\hat{\mathrm{x}}_{t|n_\mathcal{D}}^\mathsf{T}+\mathrm{P}_{t|n_\mathcal{D}} & \hat{\mathrm{x}}_{t|n_\mathcal{D}}\mathrm{u}_t^\mathsf{T}\\ \mathrm{u}_t\hat{\mathrm{x}}_{t|n_\mathcal{D}}^\mathsf{T} & \mathrm{u}_t\mathrm{u}_t^\mathsf{T}\end{bmatrix}\Big)\Big],\\
\frac{\partial Q(\theta|\hat{\theta}_k)}{\partial \mathrm{S}_\mathrm{w}} =& \frac{1}{2}\sum_{t=1}^{n_\mathcal{D}}\Big\{-n_\mathcal{D}\mathrm{S}_\mathrm{w}^{-1}+\mathrm{S}_\mathrm{w}^{-1}\Big(\big(\hat{\mathrm{x}}_{t+1|n_\mathcal{D}}\hat{\mathrm{x}}_{t+1|n_\mathcal{D}}^\mathsf{T}+\mathrm{P}_{t+1|n_\mathcal{D}}\big)\\
&-\begin{bmatrix}\hat{\mathrm{x}}_{t+1|n_\mathcal{D}}\hat{\mathrm{x}}_{t|n_\mathcal{D}}^\mathsf{T}+\mathrm{P}_{t+1|n_\mathcal{D}}\mathrm{H}_t^\mathsf{T} & \hat{\mathrm{x}}_{t+1|n_\mathcal{D}}\mathrm{u}_t^\mathsf{T}\end{bmatrix}\mathrm{M}^\mathsf{T}\\
&-\mathrm{M}\begin{bmatrix}\hat{\mathrm{x}}_{t+1|n_\mathcal{D}}\hat{\mathrm{x}}_{t|n_\mathcal{D}}^\mathsf{T}+\mathrm{P}_{t+1|n_\mathcal{D}}\mathrm{H}_t^\mathsf{T} & \hat{\mathrm{x}}_{t+1|N}\mathrm{u}_t^\mathsf{T}\end{bmatrix}^\mathsf{T}\\
&\mathrm{M}\begin{bmatrix}\hat{\mathrm{x}}_{t|n_\mathcal{D}}\hat{\mathrm{x}}_{t|n_\mathcal{D}}^\mathsf{T}+\mathrm{P}_{t|n_\mathcal{D}} & \hat{\mathrm{x}}_{t|n_\mathcal{D}}\mathrm{u}_t^\mathsf{T}\\ \mathrm{u}_t\hat{\mathrm{x}}_{t|n_\mathcal{D}}^\mathsf{T} & \mathrm{u}_t\mathrm{u}_t^\mathsf{T}\end{bmatrix}\mathrm{M}^\mathsf{T}\Big)\mathrm{S}_\mathrm{w}^{-1},\\
\frac{\partial Q(\theta|\hat{\theta}_k)}{\partial \mu_{\mathrm{x}_0}} =& \mathrm{S}_{\mathrm{x}_0}^{-1}\{(\hat{\mathrm{x}}_{0|n_\mathcal{D}}-\mu_{\mathrm{x}_0})\mu_{\mathrm{x}_0}^\mathsf{T}\},\\
\frac{\partial Q(\theta|\hat{\theta}_k)}{\partial \mathrm{S}_{\mathrm{x}_0}} =& \frac{1}{2}\{-\mathrm{S}_{\mathrm{x}_0}^{-1}+\mathrm{S}_{\mathrm{x}_0}^{-1}(\hat{\mathrm{x}}_{0|n_\mathcal{D}}\hat{\mathrm{x}}_{0|n_\mathcal{D}}^\mathsf{T}+\mathrm{P}_{0|n_\mathcal{D}}\\
&-\hat{\mathrm{x}}_{0|n_\mathcal{D}}\mu_{\mathrm{x}_0}^\mathsf{T}-\mu_{\mathrm{x}_0}\hat{\mathrm{x}}_{0|n_\mathcal{D}}^\mathsf{T}+\mu_{\mathrm{x}_0}\mu_{\mathrm{x}_0}^\mathsf{T})\mathrm{S}_{\mathrm{x}_0}^{-1}\}.
\end{aligned}$$

We set the partial derivatives to be zero, and obtain (22).

REFERENCES

[1] L. Zadeh, "On the identification problem," *IRE Transactions on Circuit Theory*, vol. 3, no. 4, pp. 277–281, 1956.

[2] L. Ljung, "Perspectives on system identification," *Annual Reviews in Control*, vol. 34, no. 1, pp. 1–12, 2010.

[3] D. G. Luenberger, *Dynamic Systems*. J. Wiley Sons, 1979.

[4] L. Ljung, *System identification: Theory for the user*. Prentice Hall, 1999.

[5] J. Schoukens and L. Ljung, "Nonlinear system identification: A user-oriented road map," *IEEE Control Systems Magazine*, vol. 39, no. 6, pp. 28–99, 2019.

[6] M. Khosravi, "Representer theorem for learning Koopman operators," *IEEE Transactions on Automatic Control*, 2023.

[7] M. Khosravi and R. S. Smith, "Kernel-based identification with frequency domain side-information," *Automatica*, vol. 150, p. 110813, 2023.

[8] D. Liu and M. Khosravi, "Learning stable evolutionary PDE dynamics: A scalable system identification approach," in *IEEE Conference on Control Technology and Applications*, 2024, pp. 79–84.

[9] R. J. Little and D. B. Rubin, *Statistical analysis with missing data*. John Wiley & Sons, 2019, vol. 793.

[10] M. Khosravi and R. S. Smith, "The existence and uniqueness of solutions for kernel-based system identification," *Automatica*, vol. 148, p. 110728, 2023.

[11] K. Åstrom, "Maximum likelihood and prediction error methods," *IFAC Proceedings Volumes*, vol. 12, no. 8, pp. 551–574, 1979.

[12] P. Brémaud, *An introduction to probabilistic modeling*. Springer Science & Business Media, 2012.

[13] M. Khosravi and R. S. Smith, "Kernel-based impulse response identification with side-information on steady-state gain," *IEEE Transactions on Automatic Control*, 2023.

[14] M. Zorzi, "Nonparametric identification of kronecker networks," *Automatica*, vol. 145, p. 110518, 2022.

[15] J. M. Bernardo and A. F. Smith, *Bayesian theory*. John Wiley & Sons, 2009, vol. 405.

[16] R. J. Rossi, *Mathematical statistics: an introduction to likelihood based inference*. John Wiley & Sons, 2018.

[17] J.-L. Gauvain and C.-H. Lee, "Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains," *IEEE transactions on speech and audio processing*, vol. 2, no. 2, pp. 291–298, 1994.

[18] M. Khosravi and R. S. Smith, "Nonlinear system identification with prior knowledge on the region of attraction," *IEEE Control Systems Letters*, vol. 5, no. 3, pp. 1091–1096, 2021.

[19] M. Khosravi and R. S. Smith, "Convex nonparametric formulation for identification of gradient flows," *IEEE Control Systems Letters*, vol. 5, no. 3, pp. 1097–1102, 2021.

[20] A. Wills, T. B. Schön, L. Ljung, and B. Ninness, "Identification of hammerstein–wiener models," *Automatica*, vol. 49, no. 1, pp. 70–81, 2013.

[21] Q. Liu, X. Tang, J. Li, J. Zeng, K. Zhang, and Y. Chai, "Identification of wiener–hammerstein models based on variational bayesian approach in the presence of process noise," *Journal of the Franklin Institute*, vol. 358, no. 10, pp. 5623–5638, 2021.

[22] P. M. Pardalos and V. A. Yatsenko, *Optimization and control of bilinear systems: theory, algorithms, and applications*. Springer Science & Business Media, 2010, vol. 11.

[23] V. Verdult and M. Verhaegen, "Identification of multivariable bilinear state space systems based on subspace techniques and separable least squares optimization," *International Journal of Control*, vol. 74, no. 18, pp. 1824–1836, 2001.

[24] V. Verdult and M. Verhaegen, "Kernel methods for subspace identification of multivariable lpv and bilinear systems," *Automatica*, vol. 41, no. 9, pp. 1557–1565, 2005.

[25] S. Liu, F. Ding, and T. Hayat, "Moving data window gradient-based iterative algorithm of combined parameter and state estimation for bilinear systems," *International Journal of Robust and Nonlinear Control*, vol. 30, no. 6, pp. 2413–2429, 2020.

[26] S. Liu, X. Zhang, L. Xu, and F. Ding, "Expectation–maximization algorithm for bilinear systems by using the rauch–tung–striebel smoother," *Automatica*, vol. 142, p. 110365, 2022.

[27] H. E. Rauch, F. Tung, and C. T. Striebel, "Maximum likelihood estimates of linear dynamic systems," *AIAA journal*, vol. 3, no. 8, pp. 1445–1450, 1965.

[28] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the royal statistical society: series B (methodological)*, vol. 39, no. 1, pp. 1–22, 1977.

[29] R. E. Kalman, "A new approach to linear filtering and prediction problems," 1960.

[30] S. J. Julier and J. K. Uhlmann, "Unscented filtering and nonlinear estimation," *Proceedings of the IEEE*, vol. 92, no. 3, pp. 401–422, 2004.

[31] S. J. Julier, J. K. Uhlmann, and H. F. Durrant-Whyte, "A new approach for filtering nonlinear systems," in *Proceedings of 1995 American Control Conference-ACC'95*, vol. 3. IEEE, 1995, pp. 1628–1632.

[32] S. Theodoridis, *Pattern recognition*. Academic press, 2006.

[33] S. Särkkä and L. Svensson, *Bayesian filtering and smoothing*. Cambridge university press, 2023, vol. 17.