Proceedings of Machine Learning Research vol 144:1–12, 2021

A New Objective for Identification of Partially Observed Linear Time-Invariant Dynamical Systems from Input-Output Data

Nicholas Galioto NGALIOTO@UMICH.EDU
1320 Beal Avenue, Ann Arbor, MI 48109, USA

Alex Arkady Gorodetsky GORODA@UMICH.EDU
1320 Beal Avenue, Ann Arbor, MI 48109, USA

Abstract

In this work we consider the identification of partially observed dynamical systems from a single trajectory of arbitrary input-output data. We propose a new optimization objective, derived as a MAP estimator of a certain posterior, that explicitly accounts for model, measurement, and parameter uncertainty. This algorithm identifies a linear time-invariant model on a hidden latent space of pre-specified dimension. In contrast to Markov parameter-based least squares approaches, our algorithm can be applied to systems with arbitrary forcing and initial condition, and we empirically show several magnitude improvement in prediction quality compared to state-of-the-art approaches on both linear and nonlinear systems. Furthermore, we theoretically demonstrate how these existing approaches can be derived from simplifying assumptions on our system that neglect the possibility of model errors.

Keywords: System identification, data-driven learning, single-trajectory learning, Bayesian inference

1. Introduction

Finding a linear model for a system can be highly advantageous in a number of system identification and controls applications. For applications where the system is high-dimensional or real-time prediction and/or control is required, it is also imperative that the model be inexpensive to evaluate. Linear systems have closed-form solutions that are cheap to evaluate relative to the numerical integration required for many nonlinear models, making them practical solutions for such problems. While there has been great progress in modeling nonlinear behavior, especially within nonlinear parametric approximations like neural networks Li et al. (2020); De Paula and Marques (2019), these approaches often require large amounts of data and computational power, lack robustness when the data are sparse/noisy, and are computationally expensive for post-processing. As a result, more efficient linear identification techniques, even for generating predictive models of nonlinear systems, offer promising routes to broad applications in control and dynamical systems.

For example, two of the most common approaches for modeling nonlinear flows, Proper Orthogonal Decomposition (POD) Berkooz et al. (1993) and dynamic mode decomposition (DMD) Schmid (2010), rely on finding a reduced-order linear model that can approximate the system. In fact, Koopman operator theory proves that every nonlinear system has a corresponding linear, infinite-dimensional Koopman operator that yields equivalent dynamics as the nonlinear system Mezić (2005). The Koopman modes of such an operator can reveal important dynamical information of the system such as modes of oscillation and growth rates. In Rowley et al. (2009), it was shown that a least squares linear approximation of a system is equivalent to the truncated Koopman operator and can thus provide approximations of these important dynamical characteristics.

These aforementioned algorithms also have their counterparts for systems with control inputs. In these cases, the estimation problem is often posed as first identifying the system's Markov parameters, and then splitting them into system matrices via the eigensystem realization algorithm (ERA) Juang and Pappa (1985) (a.k.a. the Ho-Kalman algorithm Ho and Kálmán (1966)) or similar methods. The canonical ERA approach requires highly structured inputs, e.g., by measuring impulse responses, and effort has gone into broadening these requirements. For instance, the works of Oymak and Ozay (2019); Sarkar et al. (2019); Fattahi (2020) have developed ERA-inspired algorithms to estimate linear time-invariant (LTI) systems from a single trajectory of input-output data provided that the initial condition is zero (free response is removed). The general realization algorithm De Callafon et al. (2008) requires only that the initial condition be zero to estimate an LTI model of the system. Unfortunately, even when the assumptions of these algorithms are satisfied, we will show their performance begins to deteriorate if the input-output data are sparse and/or noisy. DMD approaches, which are similar but applied to the perfectly observed case, have also been extended to consider control inputs Proctor et al. (2016).

In this work, we posit that because these existing approaches don't fully consider the interaction of parameter, model, and measurement uncertainties, they are not robust to sparse and/or noisy data. In the context of system identification without inputs Galioto and Gorodetsky (2020), building on the works of Khalil et al. (2015); Drovandi et al. (2019), presented an approach inspired by a first-principles probabilistic derivation of a suitable objective function that accounts for parameter uncertainty, model uncertainty (modeled as process noise), and measurement uncertainty. This approach was shown to significantly outperform popular linear (such as DMD) and nonlinear (such as SINDy Brunton et al. (2016)) data-driven methods that forgo modeling problem uncertainty.

We extend a portion of this prior work to the problem of learning LTI models of partially observed, arbitrarily forced systems where we have access to input-output data. We also remove the requirement that a specific coordinate space for the underlying LTI model is known, rather it is implicitly learned through the system matrices. Thus our contributions include

- A new optimization objective for identifying LTI systems from single trajectories of input-output data that simultaneously accounts for parameter uncertainty, model errors (through process noise), and measurement noise
- Proof that common least squares approaches for Markov parameter estimation discard model uncertainty in the objective specification, causing robustness issues for sparse/noisy data
- Empirical evidence that this method can outperform the least squares-based method at various values of measurement noise and frequency and can also recover partially observed nonlinear systems.

Our numerical results indicate that gains of several orders of magnitude in mean squared error can be achieved for recovering systems with sparse and noisy data. We also show that our approach is applicable to nonlinear systems with no requirements on removing the free response.

The rest of this paper is organized as follows. In Section 2, we provide a probabilistic formulation of the system identification problem that leads to our proposed objective function. In Section 3, we discuss determinisitc methods for learning partially observed LTI systems and show how these methods can be viewed as special cases of the probabilistic formulation. Then in Section 4, we compare the performance of the MAP estimate to a least squares method on a variety of data sets from a forced linear pendulum. We also show our approach's applicability to nonlinear systems. Lastly, in Section 5 we summarize our findings.

2. Problem setting and methodology

In this section, we describe the probabilistic model and optimization objective.

We are interested in learning partially observed, LTI models for systems given input-output data. Our partially observed LTI model is given by the following system of equations

$$\begin{aligned} X_{k+1} &= \mathbf{A}X_k + \mathbf{B}u_k + \xi_k, & \xi_k &\sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}), \\ Y_k &= \mathbf{C}X_k + \eta_k, & \eta_k &\sim \mathcal{N}(\mathbf{0}, \mathbf{\Gamma}), \end{aligned} \tag{1}$$

where $\mathbf{A} \in \mathbb{R}^{d_x \times d_x}$ is the state transition matrix, $\mathbf{B} \in \mathbb{R}^{d_x \times 1}$ is the control input matrix, and $\mathbf{C} \in \mathbb{R}^{d_y \times d_x}$ is the observation matrix. The hidden state at time k is $X_k \in \mathbb{R}^{d_x}$ and the observation at time k is $Y_k \in \mathbb{R}^{d_y}$. We use the uppercase X and Y to denote random variables, and later we will use lowercase $\mathbf{x}$ and $\mathbf{y}$ to denote realizations of these random variables. The control input at time k is $u_k \in \mathbb{R}$. We assume a Gaussian process noise $\xi_k \in \mathbb{R}^{d_x}$ with covariance $\mathbf{\Sigma} \in \mathbb{R}^{d_x \times d_x}$. Furthermore, we assume a Gaussian measurement noise η_k with covariance matrix $\mathbf{\Gamma} \in \mathbb{R}^{d_y \times d_y}$.

Next we outline our learning goals and methods. The combined unknowns defining the system identification problem are denoted by $\Theta = (\mathbf{x}_0, \mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{\Sigma}, \mathbf{\Gamma})$. The general target of learning is the Bayesian posterior distribution $p(\Theta \mid \mathcal{Y}_n)$ over the parameters given the observed data $\mathcal{Y}_n := (\mathbf{y}_1, \ldots, \mathbf{y}_n)$. The maximum *a-posteriori* (MAP) of the Bayesian posterior is defined as

$$\Theta^{MAP} = \arg\max_{\Theta} \log p(\Theta \mid \mathcal{Y}_n). \tag{2}$$

In this paper, the MAP estimate is viewed as the solution to our system ID problem. This Bayesian approach to system identification has been around for decades Peterka (1981); Ninness and Henriksen (2010) and can be viewed as a reinterpration of the general system ID framework Ljung (1999). Oftentimes, efforts to improve the quality of the MAP estimate are focused on the selection of the prior distribution/regularization techniques Chen (2013); Pillonetto et al. (2016). The novelty of our approach, described in detail in Galioto and Gorodetsky (2020), is that we introduce regularization through the likelihood by including process noise in our model, even when the system itself is known to be deterministic. In an overfit model, slight changes to the initial condition can yield vastly different trajectories. By including process noise, we ensure that our estimate is not overly sensitive to perturbations in the trajectory and are therefore able to avoid these overfit models.

Bayes' rule allows specifying the posterior in terms of computable quantities

$$p(\Theta \mid \mathcal{Y}_n) = \frac{\mathcal{L}(\Theta \mid \mathcal{Y}_n) p(\Theta)}{p(\mathcal{Y}_n)}, \tag{3}$$

where $\mathcal{L}(\Theta \mid \mathcal{Y}_n) := p(\mathcal{Y}_n \mid \Theta)$ is the marginal likelihood, $p(\Theta)$ is the prior, and $p(\mathcal{Y}_n)$ is the evidence. The evidence does not affect the MAP estimate and so is not required for learning. The prior serves to regularize the parameters. The primary challenge is the evaluation of the marginal likelihood.

The process noise induces uncertainty in the states and the marginal likelihood must be obtained by integrating out all state uncertainty from the joint likelihood $\mathcal{L}(\Theta, \mathcal{X}_n \mid \mathcal{Y}_n)$

$$\mathcal{L}(\Theta \mid \mathcal{Y}_n) = \int \mathcal{L}(\Theta, \mathcal{X}_n \mid \mathcal{Y}_n) d\mathcal{X}_n, \quad \mathcal{X}_n = (X_0, \ldots, X_n), \tag{4}$$

where the joint likelihood is provided in the following theorem.

Theorem 1 (Joint likelihood (Th. 12.3 Särkkä (2013))) *The joint likelihood of system* (1)

$$\mathcal{L}(\Theta, \mathcal{X}_n \mid \mathcal{Y}_n) = \prod_{k=1}^{n} \frac{\exp\left(-\frac{1}{2}\left\|X_k - \mathbf{A}X_{k-1} - \mathbf{B}u_{k-1}\right\|_{\mathbf{\Sigma}}^2\right)}{(2\pi)^{\frac{d_x}{2}}\left|\mathbf{\Sigma}\right|^{\frac{1}{2}}} \frac{\exp\left(-\frac{1}{2}\left\|\mathbf{y}_k - \mathbf{C}X_k\right\|_{\mathbf{\Gamma}}^2\right)}{(2\pi)^{\frac{d_y}{2}}\left|\mathbf{\Gamma}\right|^{\frac{1}{2}}}, \tag{5}$$

where $\|\cdot\|_{\mathbf{\Sigma}}^2 = (\cdot)^T\mathbf{\Sigma}^{-1}(\cdot)$ *refers to the weighted inner product.*

Evaluating the integral in Equation (4) can be expensive without exploiting the Markovian structure of the dynamics. This structure can be exploited through recursion by using a filtering technique for computing a sequence of marginal likelihoods as data are processed. In the context of learning LTI models, this filter is the Kalman filter. This approach is formally given by the following theorem.

Theorem 2 (Marginal likelihood (Th. 12.1 Särkkä (2013))) *Let* $\mathcal{Y}_k \equiv \{y_i; i \leq k\}$ *denote the set of all observations up to time* k*. Let the initial condition be uncertain with distribution* $p(X_0 \mid \Theta)$*. Then the marginal likelihood* (4) *is defined as* $\mathcal{L}(\Theta \mid \mathcal{Y}_n) = \prod_{k=1}^{n} \mathcal{L}_k(\Theta \mid \mathcal{Y}_k)$*, where* $\mathcal{L}_k(\Theta \mid \mathcal{Y}_k)$ *is computed recursively in three stages for* $k = 1, 2, \ldots$*: prediction*

$$p(X_{k+1} \mid \Theta, \mathcal{Y}_k) = \int \frac{\exp\left(-\frac{1}{2}\|X_{k+1} - \mathbf{A}X_k - \mathbf{B}u_k\|_{\mathbf{\Sigma}}^2\right)}{\sqrt{2\pi}^{d_x}|\mathbf{\Sigma}|^{\frac{1}{2}}} p(X_k \mid \Theta, \mathcal{Y}_k) dX_k \tag{6}$$

update,

$$p\left(X_{k+1} \mid \Theta, \mathcal{Y}_{k+1}\right) = p(X_{k+1} \mid \Theta, \mathcal{Y}_k) \frac{\exp\left(-\frac{1}{2}\|\mathbf{y}_{k+1} - \mathbf{C}X_{k+1}\|_{\mathbf{\Gamma}}^2\right)}{\sqrt{2\pi}^{d_y}|\mathbf{\Gamma}|^{\frac{1}{2}} p(Y_{k+1} \mid \Theta, \mathcal{Y}_k)} \tag{7}$$

and marginalization,

$$\mathcal{L}_{k+1}(\Theta \mid \mathcal{Y}_{k+1}) = \int p(X_{k+1} \mid \Theta, \mathcal{Y}_k) \frac{\exp\left(-\frac{1}{2}\|\mathbf{y}_{k+1} - \mathbf{C}X_{k+1}\|_{\mathbf{\Gamma}}^2\right)}{\sqrt{2\pi}^{d_y}|\mathbf{\Gamma}|^{\frac{1}{2}}} dX_{k+1}. \tag{8}$$

Given this decomposition into the likelihood and prior, the MAP optimization objective (2) becomes

$$\Theta^{MAP} = \arg\max_{\Theta} \log \mathcal{L}(\Theta \mid \mathcal{Y}_n) + \log p(\Theta). \tag{9}$$

3. Comparison to Markov parameter estimation approaches

In this section we review approaches based on Markov parameter estimation and their relation to the proposed approach. When the system is free of process and measurement noise, the model becomes

$$\mathbf{x}_{k+1} = \mathbf{A}\mathbf{x}_k + \mathbf{B}u_k, \quad \mathbf{y}_k = \mathbf{C}\mathbf{x}_k. \tag{10}$$

In this case it is possible to explicitly write the input-output relationship using Markov parameters $\mathbf{g}_i = \mathbf{C}\mathbf{A}^{i-1}\mathbf{B}$. When the input is an impulse signal, the Markov parameters are equivalent to the outputs; otherwise, the outputs are related to the inputs through the following expression

$$\mathbf{y}_k = \sum_{i=1}^{k} \mathbf{g}_i u_{k-i}. \tag{11}$$

This expression can also be written in matrix form as $\mathbf{Y} = \mathbf{G}\mathbf{U}$ where

$$\mathbf{Y} = \begin{bmatrix} \mathbf{y}_1 & \mathbf{y}_2 & \cdots & \mathbf{y}_n \end{bmatrix}, \quad \mathbf{U} = \begin{bmatrix} u_0 & u_1 & \cdots & u_{n-1} \\ 0 & u_0 & \cdots & u_{n-2} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & u_0 \end{bmatrix}, \quad \mathbf{G} = \begin{bmatrix} \mathbf{g}_1 & \mathbf{g}_2 & \cdots & \mathbf{g}_n \end{bmatrix}. \quad (12)$$

Given the Markov parameters, the system can be recovered by computing a Hankel matrix and then decomposing it into observability and controllability matrices. The Hankel matrix defined for integers M, N satisfying $M + N \leq n + 1$ is given by

$$\mathbf{H} = \begin{bmatrix} \mathbf{g}_{1:M} & \mathbf{g}_{2:M+1} & \cdots & \mathbf{g}_{N:M+N-1} \end{bmatrix}, \quad (13)$$

where $\mathbf{g}_{i:j}$ is a column vector of Markov parameters i through j for $i < j$. The Hankel matrix $\mathbf{H}$ can then be decomposed into observability and controllability matrices using the singular value decomposition (SVD) Juang and Pappa (1985) or other low-rank factorization Kramer and Gorodetsky (2018) by following the ERA. This process determines an order r set of equations, where r is obtained by truncating the SVD. Let the recovered system matrices be $\hat{\mathbf{A}} \in \mathbb{R}^{r \times r}$, $\hat{\mathbf{B}} \in \mathbb{R}^{r}$, and $\hat{\mathbf{C}} \in \mathbb{R}^{d_y \times r}$. This recovery is nonunique: given a nonsingular transformation $\mathbf{T} \in \mathbb{R}^{d_x \times d_x}$, the systems $(\hat{\mathbf{A}}, \hat{\mathbf{B}}, \hat{\mathbf{C}})$ and $(\mathbf{T}^{-1}\hat{\mathbf{A}}\mathbf{T}, \mathbf{T}^{-1}\hat{\mathbf{B}}, \hat{\mathbf{C}}\mathbf{T})$ possess identical Markov parameters.

3.1. Least squares

A vast majority of approaches that estimate Markov parameters use least squares estimation. The least squares estimate of $\mathbf{G}$ can be written as the minimizer of the following objective function

$$\hat{\mathbf{G}} = \underset{\mathbf{g}_1,\ldots,\mathbf{g}_n}{\arg\min} \sum_{k=1}^{n} \left\| \mathbf{y}_k - \sum_{i=1}^{k} \mathbf{g}_i u_{k-i} \right\|^2. \quad (14)$$

Because of the upper triangular form of the $\mathbf{U}$ matrix, Markov parameters for which the value of the objective function is identically zero can be found with forward substitution. When process and measurement noise are absent from the system, the least squares estimate recovers the true Markov parameters. This approach is used, for instance, in the general realization algorithm (GRA) De Callafon et al. (2008) that uses the least squares estimate of a small number of Markov parameters to build a large Hankel matrix to feed into the ERA.

This least squares approach is also extended to the case where process and measurement noise is present. In this case, the input-output equation derived from (1) is

$$\mathbf{y}_k = \mathbf{C}(\mathbf{A}\mathbf{x}_k + \mathbf{B}u_k + \xi_k) + \eta_k = \mathbf{C}\mathbf{A}^k\mathbf{x}_0 + \sum_{i=1}^{k} \mathbf{C}\mathbf{A}^{i-1}\mathbf{B}u_{k-i} + \sum_{i=1}^{k} \mathbf{C}\mathbf{A}^{i-1}\xi_{k-i} + \eta_k. \quad (15)$$

Recent works such as Oymak and Ozay (2019); Sarkar et al. (2019); Fattahi (2020) consider the problem where the inputs are distributed according to a standard normal distribution. Under this condition, the response of the system is zero-mean and the full time history of the inputs is not required (assuming zero initial condition) to write an expression connecting the outputs to the

Markov parameters. Instead, if we are interested in finding the first T Markov parameters, the input-output expression becomes

$$\mathbf{y}_k = \mathbf{C}\mathbf{A}^T\mathbf{x}_{k-T} + \sum_{i=1}^{k} \mathbf{C}\mathbf{A}^{i-1}\mathbf{B}u_{k-i} + \sum_{i=1}^{k} \mathbf{C}\mathbf{A}^{i-1}\xi_{k-i} + \eta_k. \tag{16}$$

Then, the first and last terms are treated as zero-mean additive noise, and the objective becomes

$$\hat{\mathbf{G}} = \arg\min_{\mathbf{G}} \sum_{i=T}^{n} \|\mathbf{y}_i - \mathbf{G}\bar{\mathbf{u}}_i\|_2^2, \quad \text{where } \bar{\mathbf{u}}_i = \begin{bmatrix} u_{i-1} & u_{i-2} & \dots & u_{i-T} \end{bmatrix}^*, \tag{17}$$

where * denotes the transpose. Equation (16) shows the variance of the outputs $\mathbf{y}_k$ depends on $\mathbf{A}^{i-1}$, and therefore time. However the approach described above ignores this effect and assumes worst-case noise, yielding a conservative estimate proving convergence for finite sample sizes.

In this work, we do not place any assumptions on the inputs of our system and must use the full time history of our inputs in each term of our objective. We therefore alter (17) slightly such that the sum starts at $i = 1$ and we define our inputs at time $t < t_0$ to be zero. With these modifications, the objective (17) is equivalent to the previously mentioned objective (14), as we show next.

3.2. Probabilistic interpretation

In this section we introduce assumptions that will map our probabilistic model of Section 2 and the resulting objective (9) to match the objectives of the discussed Markov parameter estimation approaches. Through this process, we will demonstrate that these other approaches effectively underestimate all sources of uncertainty and provide an explanation for the poor robustness that is demonstrated in the empirical results of Section 4.

In the following, we assume that the process noise and the measurement noise are fixed and so the uncertain parameters only consist of the system matrices $\Theta = (\mathbf{A}, \mathbf{B}, \mathbf{C})$. By discarding the process noise in our model, the joint log likelihood becomes a delta distribution with respect to the states and a least squares estimate with respect to parameters.

Theorem 3 *Assume zero process noise and zero initial condition. Furthermore assume diagonal measurement noise with a constant variance for each element of the measurement. Then the negative log of the marginal likelihood* (4) *is equivalent to the least squares objective* (14)

Proof We begin by taking the negative log of the joint likelihood provided in Theorem 1

$$\log \mathcal{L}(\Theta, \mathcal{X}_n \mid \mathcal{Y}_n) \propto \sum_{k=1}^{n} \|X_k - \mathbf{A}X_{k-1} - \mathbf{B}u_{k-1}\|_{\mathbf{\Sigma}}^2 + \|\mathbf{y}_k - \mathbf{C}X_k\|_{\mathbf{\Gamma}}^2. \tag{18}$$

Under the assumption of zero process noise, the state dynamics become deterministic and the distribution $p(X_k \mid \Theta)$ becomes the Dirac delta distribution $\delta_{X_k}(\mathbf{A}\mathbf{x}_{k-1} + \mathbf{B}u_{k-1})$ that takes the value one if X_k satisfies the difference equation and zero otherwise. The nonzero case is given by

$$\log \mathcal{L}(\Theta, \mathcal{X}_n \mid \mathcal{Y}_n) \propto \sum_{k=1}^{n} \|\mathbf{y}_k - \mathbf{C}X_k\|_{\mathbf{\Gamma}}^2 \quad \text{subject to} \quad X_{k+1} = \mathbf{A}X_k + \mathbf{B}u_k. \tag{19}$$

Furthermore, deterministic dynamics implies that the state X_k can be written in terms of the initial condition $\mathbf{x}_0$. Therefore, the state equation becomes

$$\mathbf{x}_k = \mathbf{A}^{k-1}\mathbf{x}_0 + \sum_{i=1}^{k} \mathbf{A}^{i-1}\mathbf{B}u_{k-i}. \tag{20}$$

Setting $\mathbf{x}_0 = \mathbf{0}$ and using this expression within the log likelihood leads to

$$\log \mathcal{L}(\Theta \mid \mathcal{Y}_n) \propto \sum_{k=1}^{n} \left\| \mathbf{y}_k - \mathbf{C}\sum_{i=1}^{k} \mathbf{A}^{i-1}\mathbf{B}u_{k-i} \right\|_{\mathbf{\Gamma}}^2 = \sum_{k=1}^{n} \left\| \mathbf{y}_k - \sum_{i=1}^{k} \mathbf{g}_i u_{k-i} \right\|_{\mathbf{\Gamma}}. \tag{21}$$

where the log likelihood no longer depends on the states and is thus equivalent to the marginal log likelihood. The second equality uses the definition of the Markov parameters. Assuming a diagonal covariance with identical measurement noise for each component of $\mathbf{y}_k$ yields our stated result. ■

Here we have recovered a least squares problem equivalent to existing state of the art procedures. In other words, the common least squares objective (14) is derived from the assumption of an exact dynamical model. However, the standard ERA approaches typically recover reduced order models — ones that are not exact. Dropping the possibility of model error is both inconsistent with practical usage of the ERA and also leads to issues in the case of sparse and noisy data. Furthermore, these techniques typically require separate approximation of the initial condition, whereas our approach accounts for the initial condition by incorporating $\mathbf{x}_0$ within the same objective as the system matrices. Dealing with this initial condition becomes non-trivial in many problems of interest where an experiment cannot be closely controlled to ensure no free-response is exhibited. For these reasons, we included the process noise, measurement noise, and the initial condition. In the results of the next section, we show that this approach has significantly improved robustness compared to approaches that make the simplifying assumptions outlined above.

4. Numerical experiments

In this section, we provide a comparison of the performance of our objective (9) to the method of least squares (14) coupled with ERA used in Oymak and Ozay (2019) and henceforth referred to as LS+ERA on a set of simulated data. Our method not only learns a realization of the system matrices like the LS+ERA method, but also learns the initial condition, process noise, and measurement noise of the system. We measure performance via mean squared error (MSE) on both training and testing data. In our examples, we split a time-series of input-output data into a training and testing set. The training set occurs within the first T seconds and the testing data occurs beyond T seconds. The testing data set therefore tests predictive performance.

The first example provides an exhaustive comparison for LTI reconstruction of a linear system across a wide range of data sparsity and measurement noise. For each combination of sparsity level and noise variance, we compute the mean squared error of the LS+ERA approach and the proposed MAP estimator. We find that our MAP estimate is up to 1.8×10^6 times better in the low noise case and up to 1.5×10^2 times better in the high noise case. The improved performance of the MSE in the low noise case occurs because the absolute MSE of both methods is very small, and the proposed approach has greater precision. The second example considers identification of a nonlinear system. Our test cases use $T = 20$ for the first example and $T = 100$ for the second.

The MAP estimate is obtained using MATLAB's `fmincon` function on the negative log posterior. We use as a prior a standard half-normal distribution on our variance parameters as suggested in Gelman (2006) and an improper uniform distribution on the system matrices. This prior is largely uninformative in order to emphasize the regularization benefits of including process noise. The marginal likelihood is computed with a Kalman filter, and we pair the optimizer with a multistart algorithm in order to avoid local minima. Any subsequent use of the word 'learning' refers to this process of finding the MAP.

4.1. Linear pendulum with forcing

We first examine the performance of the two algorithms on a wide range of data sets acquired from an undamped linear pendulum with forcing. We begin by first providing the truth system used to simulate the data and the parameterization of our model used for learning. The equations of motion for the forced linear pendulum are given as follows

$$\begin{bmatrix} \dot{\phi} \\ \ddot{\phi} \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -g/L & 0 \end{bmatrix} \begin{bmatrix} \phi \\ \dot{\phi} \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \end{bmatrix} \cos(t), \qquad y_k = \begin{bmatrix} 1 & 0 \end{bmatrix} \begin{bmatrix} \phi_k \\ \dot{\phi}_k \end{bmatrix}, \tag{22}$$

where ϕ is the angular displacement of the pendulum with respect to the vertical, $g = 9.81$ is the gravitational constant, and $L = 1$ is the length of the pendulum. Only ϕ is observed.

Our learning model assumes knowledge of the state dimension and is parameterized as

$$\begin{aligned} &\begin{bmatrix} x_0^1 \\ x_0^2 \end{bmatrix} = \begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix}, \quad \begin{bmatrix} x_{k+1}^1 \\ x_{k+1}^2 \end{bmatrix} = \begin{bmatrix} \theta_3 & \theta_4 \\ \theta_5 & \theta_6 \end{bmatrix} \begin{bmatrix} x_k^1 \\ x_k^2 \end{bmatrix} + \begin{bmatrix} \theta_7 \\ \theta_8 \end{bmatrix} \cos(t_k), \qquad \Sigma(\theta) = \begin{bmatrix} \theta_{11} & 0 \\ 0 & \theta_{12} \end{bmatrix}, \\ &\qquad\qquad y_k = \begin{bmatrix} \theta_9 & \theta_{10} \end{bmatrix} \begin{bmatrix} x_k^1 \\ x_k^2 \end{bmatrix}, \qquad\qquad \Gamma(\theta) = \begin{bmatrix} \theta_{13} \end{bmatrix}. \end{aligned} \tag{23}$$

Our approximation is in discrete time, which introduces model uncertainty into our problem since the discrete-time model cannot capture the effects of a continuous input. Our analysis examines how well the proposed approach compares to least squares methods under varying noise levels and observation frequencies of the data. To test these methods, we consider collecting data at timesteps $0.10, 0.15, \ldots, 0.50$s and noise ratios $0.00, 0.05, \ldots, 0.20$. This noise ratio represents the standard deviation of the measurement noise divided by the maximum value of the signal. At each value of noise ratio, we perform 100 experiments, each with a different realization of sampled data. The true simulation is performed with an adaptive Runge-Kutta 4-5 scheme.

Lastly, we average these MSE values across the 100 data sets and plot the log of the ratio of the MSE of the proposed optimization approach to the MSE of the least squares approach across different data conditions in Figure 1. An expression for the value at each point on this figure is

$$\log\left(\frac{\sum_{i=1}^{100} MSE_i^{MAP}}{\sum_{i=1}^{100} MSE_i^{lsq}}\right), \quad \text{where } MSE_i = \frac{1}{n}\sum_{k=1}^{n}(\phi_k - \hat{\mathbf{y}}_k)^2, \tag{24}$$

where $\hat{\mathbf{y}}_k$ is the output resulting from the estimated system. Figure 1 demonstrates that the gain in performance achieved by using the MAP estimate increases as the sampling timestep Δt increases. We also observe that the LS+ERA has degraded performance when the data are noisy and frequent.

To demonstrate in more detail the qualitative differences between how the two methods perform, we provide examples of the output of the estimated system in Figure 2 for two cases of (noise ratio,

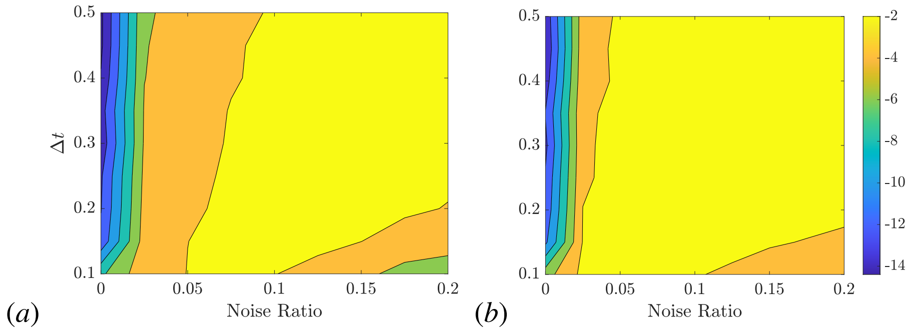

Figure 1: Contour plots of the log of the ratio of the MSE from the MAP estimate to the MSE from the LS+ERA estimate of system (22). The left figure is the training MSE and the right figure is the testing MSE. The more negative values correspond to greater magnitudes of improvement achieved with the MAP estimate. The MAP estimate outpeforms the least squares approach at all points.

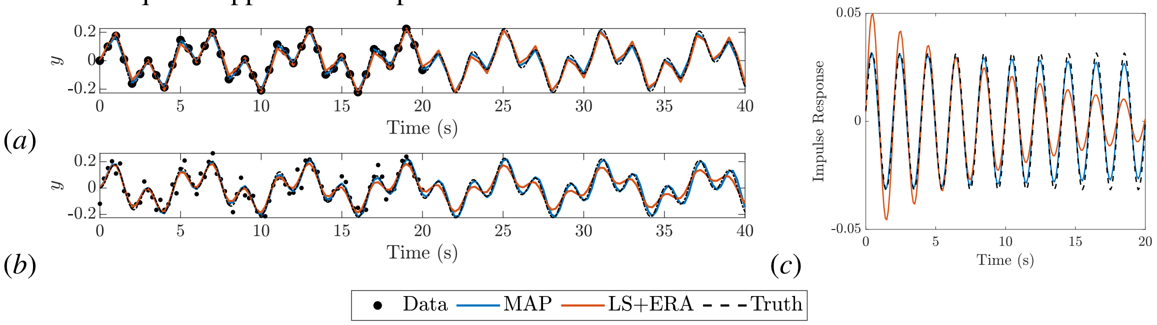

Figure 2: The left column shows trajectory estimation from two sets of training data taken over 20s from system (22). Figure 2(a) uses data sampled at $\Delta t = 0.5$ with noise ratio 0.00, and Figure 2(b) uses data sampled at $\Delta t = 0.1$ with noise ratio 0.20. Figure 2(c) shows the impulse response (Markov parameters) of the estimates corresponding to the high-noise case 2(b). The blue line represents the MAP estimate, and the orange line represents LS+ERA. Both cases illustrate improved performance of our MAP estimator. In the high-noise case, the mode of the LS+ERA model decays quickly, which is easily seen in 2(c).

timestep). These cases are $(0.00, 0.50)$ and $(0.20, 0.10)$. Figure 2(a) shows that even when the data are noiseless, LS+ERA and the MAP perform similarly from a visual inspection. However we note three things: LS+ERA is unable to match the training data points to the same precision as the MAP, the amplitudes of LS+ERA do not quite reach the truth in the testing time, and we numerically find the MSE of LS+ERA to be 3.6×10^{-4} and for the MAP to be 3.8×10^{-10}. Since measurement uncertainty does not enter this problem, this figure shows that LS+ERA cannot handle the model uncertainty corresponding to the coarse time discretization as well as the proposed optimization method. Figure 2(b) shows that when the data are noisy, LS+ERA decays more quickly. The MAP estimate remains robust to both types of uncertainty since these uncertainties are accounted for in its objective. Furthermore, the impulse response shown in Figure 2(c) demonstrates that the MAP estimate accurately characterizes the linear system. Despite the problem's overparameterization, the model is not overfit due to the regularization delivered by including a process noise term.

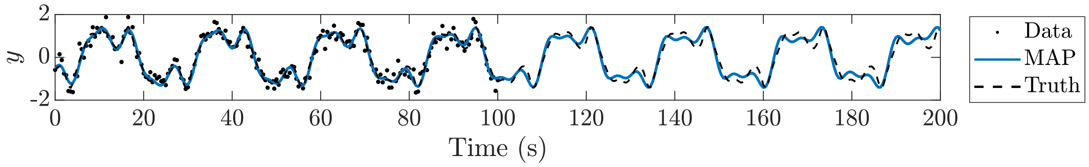

Figure 3: MAP approach for learning a partially observed nonlinear oscillator. The blue line represents the MAP estimate from 100s of training data with $\Delta t = 0.50$ and $\sigma = 0.3$. Our approach closely reconstructs the nonlinear system.

4.2. Duffing oscillator with forcing

Next, we demonstrate the applicability of our approach to learning nonlinear systems. In this case, we cannot use data starting from the initial time when the system is at rest because this would include the highly nonlinear transience of the solution. We instead collect data after the point at which the system has reached a steady, periodic solution. Therefore, the assumption used by the least squares approach that data collection starts when the system is at equilibrium no longer holds, and we cannot apply the least squares method to this system. We can, however, still use the proposed optimization approach since it learns the initial condition.

The system we consider is the periodic forced Duffing oscillator Duffing (1918)

$$\begin{bmatrix}\dot{x}\\ \ddot{x}\end{bmatrix} = \begin{bmatrix}0 & 1\\ \alpha & \delta\end{bmatrix}\begin{bmatrix}x\\ \dot{x}\end{bmatrix} + \beta\begin{bmatrix}0\\ x^3\end{bmatrix} + \begin{bmatrix}0\\ 1\end{bmatrix}\gamma\cos(\omega t), \qquad y_k = \begin{bmatrix}1 & 0\end{bmatrix}\begin{bmatrix}x_k\\ \dot{x}_k\end{bmatrix}. \tag{25}$$

For this experiment, we chose $\alpha = 1$, $\delta = -0.3$, $\beta = -1$, $\gamma = 0.37$, and $\omega = 1.2$ following an example in Jordan and Smith (2007). We observed only the position x of the system at $\Delta t = 0.50$s intervals over 100s and added zero-mean Gaussian noise with standard deviation $\sigma = 0.3$ to the data. To handle the nonlinearities of the system, we increased our state space dimension to four and learned a different process noise variance for each component of the state. We emphasize this method does not learn the dynamics of the nonlinear system, but rather a higher-dimensional linear approximation of a periodic limit cycle of the system. The results of this experiment are shown in Figure 3. Even with the noisiness of the data and the nonlinearity of the system, the MAP estimate learns a model that can closely reconstruct the output.

5. Conclusion

We have proposed a new optimization objective for learning partially observed LTI models from input-output data. The propsed approach is derived from the MAP estimate of the Bayesian posterior of the system model that accounts for parameter, model, and measurement uncertainty. Our approach also allows us to learn the initial condition of the system and the process noise of our model, which can provide us information on how well the model captures the system's dynamics. We empirically showed that our proposed estimation approach is more precise than LS+ERA, yielding several order of magnitude gains in predictive quality when data are sparse and noisy. We also demonstrated our method's applicability to partially observed, nonlinear systems. Future work will continue to explore its ability in identification of nonlinear systems.

Acknowledgments

This work is funded by the AFOSR Program in Computational Mathematics

References

Gal Berkooz, Philip Holmes, and John L Lumley. The proper orthogonal decomposition in the analysis of turbulent flows. *Annual review of fluid mechanics*, 25(1):539–575, 1993.

Steven L. Brunton, Joshua L. Proctor, and J. Nathan Kutz. Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proceedings of the National Academy of Sciences*, 113(15):3932–3937, 2016.

Lennart Ljung Tianshi Chen. What can regularization offer for estimation of dynamical systems? *IFAC Proceedings Volumes*, 46(11):1–8, 2013.

Raymond A De Callafon, Babak Moaveni, Joel P Conte, Xianfei He, and Eric Udd. General realization algorithm for modal identification of linear dynamic systems. *Journal of engineering mechanics*, 134(9):712–722, 2008.

NCG De Paula and FD Marques. Multi-variable volterra kernels identification using time-delay neural networks: application to unsteady aerodynamic loading. *Nonlinear Dynamics*, 97(1):767–780, 2019.

Christopher Drovandi, Richard G Everitt, Andrew Golightly, and Dennis Prangle. Ensemble mcmc: Accelerating pseudo-marginal mcmc for state space models using the ensemble kalman filter. *arXiv preprint arXiv:1906.02014*, 2019.

Georg Duffing. *Forced vibrations with ä changeable natural frequency and their technical meaning*. Number 41-42. F. Vieweg & son, 1918.

Salar Fattahi. Learning partially observed linear dynamical systems from logarithmic number of samples. *arXiv preprint arXiv:2010.04015*, 2020.

Nicholas Galioto and Alex Arkady Gorodetsky. Bayesian system id: optimal management of parameter, model, and measurement uncertainty. *Nonlinear Dynamics*, 102:241–267, Sep 2020. ISSN 1573-269X. doi: 10.1007/s11071-020-05925-8.

Andrew Gelman. Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, 1, 09 2006.

BL Ho and Rudolf E Kálmán. Effective construction of linear state-variable models from input/output functions. *at-Automatisierungstechnik*, 14(1-12):545–548, 1966.

Dominic Jordan and Peter Smith. *Nonlinear ordinary differential equations: an introduction for scientists and engineers*, volume 10. Oxford University Press on Demand, 2007.

Jer-Nan Juang and Richard S Pappa. An eigensystem realization algorithm for modal parameter identification and model reduction. *Journal of guidance, control, and dynamics*, 8(5):620–627, 1985.

Mohammad Khalil, Abhijit Sarkar, Sondipon Adhikari, and Dominique Poirel. The estimation of time-invariant parameters of noisy nonlinear oscillatory systems. *Journal of Sound and Vibration*, 344:81 – 100, 2015. ISSN 0022-460X.

Boris Kramer and Alex A Gorodetsky. System identification via cur-factored hankel approximation. *SIAM Journal on Scientific Computing*, 40(2):A848–A866, 2018.

Wenjie Li, Shujin Laima, Xiaowei Jin, Wenyong Yuan, and Hui Li. A novel long short-term memory neural-network-based self-excited force model of limit cycle oscillations of nonlinear flutter for various aerodynamic configurations. *NONLINEAR DYNAMICS*, 2020.

Lennart Ljung. System identification. *Wiley encyclopedia of electrical and electronics engineering*, pages 1–19, 1999.

Igor Mezić. Spectral properties of dynamical systems, model reduction and decompositions. *Nonlinear Dynamics*, 41(1):309–325, 2005.

Brett Ninness and Soren Henriksen. Bayesian system identification via markov chain monte carlo techniques. *Automatica*, 46(1):40–51, 2010.

Samet Oymak and Necmiye Ozay. Non-asymptotic identification of lti systems from a single trajectory. In *2019 American Control Conference (ACC)*, pages 5655–5661. IEEE, 2019.

Václav Peterka. Bayesian approach to system identification. In *Trends and Progress in System identification*, pages 239–304. Elsevier, 1981.

Gianluigi Pillonetto, Tianshi Chen, Alessandro Chiuso, Giuseppe De Nicolao, and Lennart Ljung. Regularized linear system identification using atomic, nuclear and kernel-based norms: The role of the stability constraint. *Automatica*, 69:137–149, 2016.

Joshua L Proctor, Steven L Brunton, and J Nathan Kutz. Dynamic mode decomposition with control. *SIAM Journal on Applied Dynamical Systems*, 15(1):142–161, 2016.

Clarence W Rowley, Igor Mezić, Shervin Bagheri, Philipp Schlatter, Dans Henningson, et al. Spectral analysis of nonlinear flows. *Journal of fluid mechanics*, 641(1):115–127, 2009.

Tuhin Sarkar, Alexander Rakhlin, and Munther A Dahleh. Finite-time system identification for partially observed lti systems of unknown order. *arXiv preprint arXiv:1902.01848*, 2019.

Simo Särkkä. *Bayesian filtering and smoothing*, volume 3. Cambridge University Press, 2013.

Peter J Schmid. Dynamic mode decomposition of numerical and experimental data. *Journal of fluid mechanics*, 656:5–28, 2010.