*International Journal of*
*Network Dynamics and Intelligence (IJNDI)*

Scilight

*Article*

# Deep Variational System Identification for Nonlinear State-Space Models via Alternating Optimization

Kaiqi Fang [1], Guijun Ma [1], Yasen Wang [2], Zuogong Yue [1], Xiaoyou Duan [1], and Maolin Wang [1,*]

[1] School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, Wuhan 430074, China
[2] School of Future Technology, South China University of Technology, Guangzhou 510641, China
* Correspondence: mlw@hust.edu.cn
Received: 31 March 2026; Revised: 4 June 2026; Accepted: 3 August 2026; Published: 17 August 2026

**Abstract:** Nonlinear state-space model identification is inherently challenging due to the need for joint latent-state inference and parameter learning. Variational inference offers a tractable framework, where structured Gaussian posteriors enable scalable inference over latent trajectories. However, existing parameterizations of the posterior mean struggle to capture complex nonlinear dynamics, while more expressive deep parameterizations tend to introduce instability in joint optimization. To address these issues, this article proposes a deep variational identification method for nonlinear state-space models based on a structured Gaussian posterior. Specifically, the posterior mean is parameterized using a non-causal dilated residual convolutional network, while a Markov structure with block-tridiagonal precision is preserved to ensure linear-time inference. Furthermore, an alternating optimization scheme is developed to separate variational smoothing from model identification. The variational parameters are updated by maximizing a differentiable approximation to the evidence lower bound, whereas the model and noise parameters are updated via analytic identification steps using samples from the variational posterior. Experiments on a nonlinear discrete-time system and a stochastic Duffing oscillator demonstrate that the proposed approach achieves stable optimization and reliable estimation of the system dynamics and noise statistics.

**Keywords:** nonlinear state-space models; system identification; Gaussian variational inference; structured Gaussian posterior; alternating optimization

## 1. Introduction

System identification aims to infer mathematical models of dynamical systems from observed input-output data [1,2]. It plays a fundamental role in analyzing complex dynamics in engineering, with critical applications in industrial processes [3], robotics [4], and biomedical systems [5]. While linear system identification is supported by mature theory and efficient algorithms [6–8], nonlinear identification remains more challenging in both theory and practice [9,10]. This paper focuses on nonlinear state-space models (SSM), a widely used framework in which a latent state evolves according to nonlinear dynamics and generates noisy measurements [11].

A central difficulty in nonlinear SSM identification is the intrinsic coupling between latent-state inference and parameter learning [12]. In particular, maximum-likelihood estimation requires marginalizing out the full latent trajectory, which leads to a high-dimensional integral that is generally intractable beyond the linear Gaussian case [13]. To address this difficulty, classical approaches typically combine parameter estimation with an auxiliary state estimator and optimize either maximum-likelihood or prediction-error objectives, often within an expectation-maximization (EM) framework [14–16]. Although Kalman filtering [17] and Rauch–Tung–Striebel smoothing [18] are analytic in the linear Gaussian setting, nonlinear models rely on approximate estimators, including extended Kalman filtering, sigma-point methods [19–21], and particle methods [22–24]. While these estimators are effective in many applications, the approximation accuracy and computational cost of the chosen estimator significantly influence the resulting objective. This often leads to pronounced sensitivity to initialization and local optima, resulting in unstable convergence and unreliable uncertainty quantification [12].

In recent years, variational inference (VI) has become a useful approach for learning with latent trajectories. It introduces a variational posterior over the latent trajectory and yields an evidence lower bound (ELBO), which provides an optimization objective for approximating the log marginal likelihood [25–27]. A useful instance of this framework is Gaussian variational inference (GVI), in which the variational posterior is restricted to a Gaussian family [28,29]. This choice is attractive in SSM as it yields closed-form entropy terms, supports efficient reparameterized sampling, and facilitates structured posterior parameterizations. In particular, when the first-order Markov structure of the latent state sequence is exploited, the Gaussian posterior can be parameterized by a block-tridiagonal precision matrix [30]. Building on this perspective, Courts et al. cast SSM identification as a VI

problem [31] and showed that joint state and parameter estimation can be addressed using generic numerical optimizers under standard assumptions on the noise and variational family. More recently, Dutra investigated practical parameterizations of Gaussian variational posteriors over the latent trajectory, including time-varying, steady-state, and convolution smoother forms [32]. In particular, the steady-state and convolution smoother parameterizations share parameters across time while preserving the structured Markov form of the posterior, thereby reducing the number of decision variables and improving scalability.

A natural extension of these structured Gaussian variational constructions is to parameterize the posterior mean with deep inference networks [33–38]. Compared with linear convolution-based smoothers, such networks can represent more flexible nonlinear mappings from input-output sequences to latent trajectory estimates. However, this additional flexibility can complicate joint learning. Specifically, combining a flexible smoother with low-dimensional model parameters creates two practical difficulties. First, optimization can become imbalanced because the inference network is typically more expressive than the model parameterization and therefore adapts more quickly during joint training. Second, this imbalance can distort uncertainty estimates and degrade parameter calibration.

To address these challenges, this paper proposes a deep variational system identification framework for nonlinear SSM. The method is built on a structured Gaussian variational posterior. Its posterior mean is parameterized by a non-causal dilated residual convolutional network that maps input-output sequences to latent-state trajectory, while its precision matrix retains a Markov-structured form to preserve scalable inference. To further improve training stability and parameter calibration, an alternating optimization scheme that separates variational smoothing from model identification is developed. In this scheme, the variational parameters are updated by maximizing a differentiable approximation to the ELBO, while the model parameters are updated through analytic identification steps computed from samples drawn from the current variational posterior. An overview of the proposed framework is shown in Figure 1. The two components address different aspects of the identification problem. The DRN-based posterior mean improves the expressiveness of the variational smoother by learning a nonlinear non-causal map from input-output sequences to latent trajectories. In contrast, the alternating optimization strategy improves parameter learning by reducing the imbalance between the high-capacity inference network and the low-dimensional system parameters. Experiments on a nonlinear discrete-time system and a stochastic Duffing oscillator demonstrate that the proposed framework improves the reliability of model identification compared with smoother-kernel baselines.

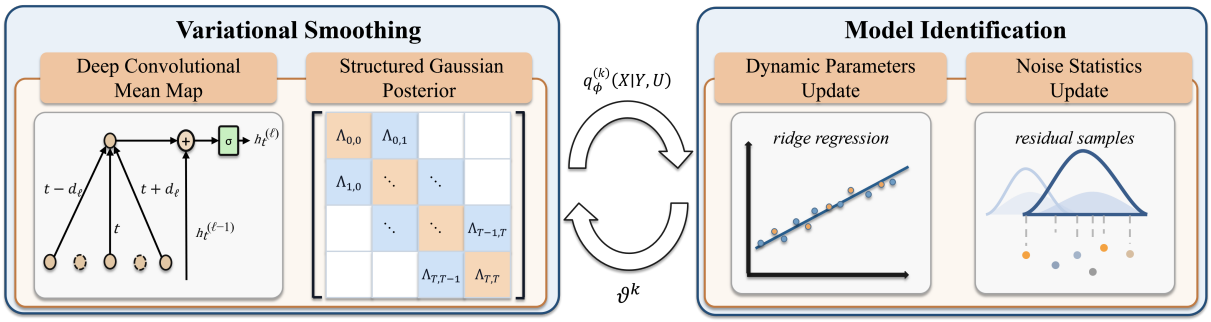


**Figure 1.** Overview of the proposed framework. The variational smoothing module combines a DRN-parameterized posterior mean with a structured Gaussian posterior, while the model identification module updates the dynamics parameters and noise statistics. The two modules are optimized in an alternating manner.

The remainder of this paper is organized as follows. Section 2 reviews GVI for nonlinear SSM and the structured Gaussian posterior parameterization used in this work. Section 3 presents the proposed deep variational system identification framework, including the deep convolutional variational smoother and the alternating optimization procedure. Section 4 reports experimental results on a nonlinear discrete-time system and a stochastic Duffing oscillator. Section 5 concludes the paper.

## 2. Gaussian Variational Inference for Nonlinear State-Space Models

This section reviews Gaussian variational inference (GVI) for nonlinear state-space models, emphasizing how the Markov structure induces a block-tridiagonal precision matrix for Gaussian posteriors over the latent state trajectory, enabling linear-time sampling and inference.

### *2.1. Nonlinear State-Space Model Formulation*

Consider the following nonlinear SSM:

$$\mathbf{x}_{t+1} = f_\theta(\mathbf{x}_t, \mathbf{u}_t) + \mathbf{w}_t, \quad t = 0, \dots, T-1, \tag{1}$$

$$\mathbf{y}_t = h(\mathbf{x}_t) + \mathbf{v}_t, \qquad t = 0, \dots, T, \tag{2}$$

where $\mathbf{x}_t \in \mathbb{R}^{d_x}$ is the latent state, $\mathbf{u}_t \in \mathbb{R}^{d_u}$ is a known input, and $\mathbf{y}_t \in \mathbb{R}^{d_y}$ is the measurement. The dynamics $f_\theta$ are parameterized by unknown model parameter $\theta \in \mathbb{R}^{d_\theta}$. This article focuses on a grey-box model that is linear in parameters,

$$f_\theta(\mathbf{x}_t, \mathbf{u}_t) = \Phi(\mathbf{x}_t, \mathbf{u}_t)\,\theta, \qquad \Phi(\mathbf{x}_t, \mathbf{u}_t) \in \mathbb{R}^{d_x \times d_\theta}, \tag{3}$$

where $\Phi(\cdot)$ is a known nonlinear feature map (basis functions) specified from prior physical knowledge. The measurement map $h(\cdot)$ is assumed known and fixed, which covers the common linear case $h(\mathbf{x}) = C\mathbf{x}$ with a known observation matrix $C \in \mathbb{R}^{d_y \times d_x}$. The process and measurement noises are independent Gaussians, with $\mathbf{w}_t \sim \mathcal{N}(0, Q)$ and $\mathbf{v}_t \sim \mathcal{N}(0, R)$, where $Q \in \mathbb{R}^{d_x \times d_x}$ and $R \in \mathbb{R}^{d_y \times d_y}$ are positive definite covariances. Noise statistics are parameterized by $\eta$, and the set of learnable model parameters is denoted by $\vartheta := (\theta, \eta)$. For compactness, the input, state, and measurement sequences are denoted by $\mathbf{U} \triangleq \{\mathbf{u}_0, \dots, \mathbf{u}_{T-1}\}$, $\mathbf{X} \triangleq \{\mathbf{x}_0, \dots, \mathbf{x}_T\}$, and $\mathbf{Y} \triangleq \{\mathbf{y}_0, \dots, \mathbf{y}_T\}$. When required in the variational Gaussian representation, these sequences are understood in stacked vector form.

Under the SSM assumptions, the joint density of the latent states and measurements factorizes as

$$p_\vartheta(\mathbf{X}, \mathbf{Y} \mid \mathbf{U}) = p(\mathbf{x}_0) \prod_{t=0}^{T-1} p_\vartheta(\mathbf{x}_{t+1} \mid \mathbf{x}_t, \mathbf{u}_t) \prod_{t=0}^{T} p_\vartheta(\mathbf{y}_t \mid \mathbf{x}_t), \tag{4}$$

where $p(\mathbf{x}_0)$ is a prior for the initial state. Under (1) and (2), the conditional densities are given by

$$p_\vartheta(\mathbf{x}_{t+1} \mid \mathbf{x}_t, \mathbf{u}_t) = \mathcal{N}\big(\mathbf{x}_{t+1}; f_\theta(\mathbf{x}_t, \mathbf{u}_t), Q(\eta)\big), \tag{5}$$

$$p_\vartheta(\mathbf{y}_t \mid \mathbf{x}_t) = \mathcal{N}\big(\mathbf{y}_t; h(\mathbf{x}_t), R(\eta)\big). \tag{6}$$

The objective is to infer both the latent state trajectory $\mathbf{X}$ and the model parameters $\vartheta$ from the observed input-output data $(\mathbf{U}, \mathbf{Y})$. In the probabilistic formulation adopted here, $\vartheta$ is learned by maximizing the marginal likelihood:

$$\max_\vartheta \ \log p_\vartheta(\mathbf{Y} \mid \mathbf{U}) = \max_\vartheta \ \log \int p_\vartheta(\mathbf{X}, \mathbf{Y} \mid \mathbf{U})\, \mathrm{d}\mathbf{X}, \tag{7}$$

while $\mathbf{X}$ is inferred through the smoothing posterior:

$$p_\vartheta(\mathbf{X} \mid \mathbf{Y}, \mathbf{U}) = \frac{p_\vartheta(\mathbf{X}, \mathbf{Y} \mid \mathbf{U})}{p_\vartheta(\mathbf{Y} \mid \mathbf{U})}. \tag{8}$$

### *2.2. Variational Inference Objective*

For nonlinear SSM, direct maximization of the marginal likelihood $\log p_\vartheta(\mathbf{Y} \mid \mathbf{U})$ is intractable, since it requires integrating out the high-dimensional latent trajectory $\mathbf{X}$. Introducing a variational distribution $q_\phi(\mathbf{X} \mid \mathbf{Y}, \mathbf{U})$ yields the following equivalent representation of the log-evidence:

$$\begin{aligned} \log p_\vartheta(\mathbf{Y} \mid \mathbf{U}) &= \log \int q_\phi(\mathbf{X} \mid \mathbf{Y}, \mathbf{U}) \, \frac{p_\vartheta(\mathbf{X}, \mathbf{Y} \mid \mathbf{U})}{q_\phi(\mathbf{X} \mid \mathbf{Y}, \mathbf{U})} \, \mathrm{d}\mathbf{X} \\ &= \log \mathbb{E}_{q_\phi}\left[ \frac{p_\vartheta(\mathbf{X}, \mathbf{Y} \mid \mathbf{U})}{q_\phi(\mathbf{X} \mid \mathbf{Y}, \mathbf{U})} \right]. \end{aligned} \tag{9}$$

Applying Jensen's inequality to (9) yields the Evidence Lower Bound (ELBO):

$$\begin{aligned} \log p_\vartheta(\mathbf{Y} \mid \mathbf{U}) &\geq \mathbb{E}_{q_\phi}[\log p_\vartheta(\mathbf{X}, \mathbf{Y} \mid \mathbf{U}) - \log q_\phi(\mathbf{X} \mid \mathbf{Y}, \mathbf{U})] \\ &=: \mathcal{L}(\vartheta, \phi). \end{aligned} \tag{10}$$

The bound is tight if and only if $q_\phi(\mathbf{X} \mid \mathbf{Y}, \mathbf{U})$ coincides with the true smoothing posterior $p_\vartheta(\mathbf{X} \mid \mathbf{Y}, \mathbf{U})$ almost everywhere [26]. Equivalently, the inequality $\log p_\vartheta(\mathbf{Y} \mid \mathbf{U}) \geq \mathcal{L}(\vartheta, \phi)$ can be rewritten as the Kullback-Leibler (KL) identity:

$$\begin{aligned} &\log p_\vartheta(\mathbf{Y} \mid \mathbf{U}) \\ &= \mathcal{L}(\vartheta, \phi) + D_{\mathrm{KL}}\big(q_\phi(\mathbf{X} \mid \mathbf{Y}, \mathbf{U}) \,\big\|\, p_\vartheta(\mathbf{X} \mid \mathbf{Y}, \mathbf{U})\big). \end{aligned} \tag{11}$$

Using the Markov factorization in (4), the complete-data log-density expands as

$$\begin{aligned}\log p_\vartheta(\mathbf{X}, \mathbf{Y} \mid \mathbf{U}) = \log p(\mathbf{x}_0) + \sum_{t=0}^{T-1} \log p_\vartheta(\mathbf{x}_{t+1} \mid \mathbf{x}_t, \mathbf{u}_t) \\ + \sum_{t=0}^{T} \log p_\vartheta(\mathbf{y}_t \mid \mathbf{x}_t).\end{aligned} \tag{12}$$

Substituting (12) into the ELBO definition yields the additive decomposition

$$\begin{aligned}\mathcal{L}(\vartheta, \phi) = \mathbb{H}[q_\phi] + \mathbb{E}_{q_\phi}\left[\sum_{t=0}^{T} \log p_\vartheta(\mathbf{y}_t \mid \mathbf{x}_t)\right] \\ + \mathbb{E}_{q_\phi}\left[\log p(\mathbf{x}_0) + \sum_{t=0}^{T-1} \log p_\vartheta(\mathbf{x}_{t+1} \mid \mathbf{x}_t, \mathbf{u}_t)\right],\end{aligned} \tag{13}$$

where $\mathbb{H}[q_\phi] := -\mathbb{E}_{q_\phi}[\log q_\phi(\mathbf{X} \mid \mathbf{Y}, \mathbf{U})]$ denotes the entropy.

To obtain a more manageable variational objective, Gaussian variational approximations of the following form are considered:

$$q_\phi(\mathbf{X} \mid \mathbf{Y}, \mathbf{U}) = \mathcal{N}(\mathbf{X};\, m_\phi(\mathbf{Y}, \mathbf{U}),\, \Sigma_\phi)\,, \tag{14}$$

where $m_\phi(\mathbf{Y}, \mathbf{U})$ denotes the posterior mean trajectory, and $\Sigma_\phi$ is the covariance matrix associated with its stacked vector representation. This Gaussian parameterization yields closed-form expressions for the entropy term in (13) and supports reparameterized sampling from $q_\phi$. To enable scalable inference over long horizons, the Gaussian posterior is structured as detailed in Section 2.3.

### *2.3. Structured Gaussian Posteriors and Markov Sparsity*

The variational posterior is restricted to follow the Markov chain topology of the SSM:

$$q_\phi(\mathbf{X} \mid \mathbf{Y}, \mathbf{U}) = q_\phi(\mathbf{x}_0 \mid \mathbf{Y}, \mathbf{U}) \prod_{t=1}^{T} q_\phi(\mathbf{x}_t \mid \mathbf{x}_{t-1}, \mathbf{Y}, \mathbf{U}). \tag{15}$$

Under GVI, the optimal information matrix inherits the same sparsity pattern as the complete-data factor graph [30]. This structure has been shown to yield accurate and scalable variational smoothers in practice [31].

For first-order Markov dynamics, the states $\mathbf{x}_{t-1}$ and $\mathbf{x}_{t+1}$ are conditionally independent given $\mathbf{x}_t$. Equivalently, the state-path precision matrix $\Lambda_\phi := \Sigma_\phi^{-1}$ is block tridiagonal:

$$\Lambda_\phi = \begin{bmatrix} \Lambda_{0,0} & \Lambda_{0,1} & & & & 0 \\ \Lambda_{1,0} & \Lambda_{1,1} & \Lambda_{1,2} & & & \\ & \ddots & \ddots & \ddots & & \\ & & \Lambda_{T-1,T-2} & \Lambda_{T-1,T-1} & \Lambda_{T-1,T} & \\ 0 & & & \Lambda_{T,T-1} & \Lambda_{T,T} & \end{bmatrix}, \tag{16}$$

where each block $\Lambda_{t,t'} \in \mathbb{R}^{d_x \times d_x}$. Such a structure enables efficient banded linear-algebra operations for inference under $q_\phi$, so that posterior sampling, entropy evaluation, and related computations scale linearly with the trajectory length $T$ [30].

To reduce the number of free covariance parameters, ref. [32] parameterizes the posterior covariance by a steady-state Gaussian Markov construction, while the posterior mean is represented by a linear convolution smoother. The joint covariance between two consecutive states is parameterized by a conditional covariance factor $S_{\text{cond}} \in \mathbb{R}^{d_x \times d_x}$ and a cross-covariance factor $S_{\text{cross}} \in \mathbb{R}^{d_x \times d_x}$. Intuitively, $S_{\text{cond}}$ controls the one-step prediction uncertainty of $\mathbf{x}_{t+1}$ given $\mathbf{x}_t$, while $S_{\text{cross}}$ captures the steady-state correlation between $\mathbf{x}_t$ and $\mathbf{x}_{t+1}$. The conditional covariance of $\mathbf{x}_{t+1}$ given $\mathbf{x}_t$ is defined as

$$\Sigma_{t+1|t} := S_{\text{cond}} S_{\text{cond}}^\top. \tag{17}$$

Under the steady-state assumption, the marginal covariance is represented by a stationary block

$$\Sigma_{\text{ss}} := \Sigma_{t,t} = \Sigma_{t+1,t+1} = S_{\text{cond}} S_{\text{cond}}^\top + S_{\text{cross}} S_{\text{cross}}^\top, \tag{18}$$

and the covariance blocks between consecutive states are

$$\Sigma_{t+1,t} = S_{\text{cross}} S_{\text{ss}}^{\top}, \qquad \Sigma_{t,t+1} = \Sigma_{t+1,t}^{\top} = S_{\text{ss}} S_{\text{cross}}^{\top}, \tag{19}$$

where $S_{\text{ss}}$ denotes the lower-triangular Cholesky factor of the steady-state marginal covariance, satisfying $\Sigma_{\text{ss}} = S_{\text{ss}} S_{\text{ss}}^{\top}$. The corresponding precision matrix $\Lambda_\phi = \Sigma_\phi^{-1}$ is block tridiagonal as in (16). Consequently, the entire covariance structure is encoded by the two matrix factors $S_{\text{cond}}$ and $S_{\text{cross}}$, involving only $\mathcal{O}(d_x^2)$ parameters and remaining independent of the horizon length $T$. In the proposed framework, this steady-state Gaussian family is used to parameterize the covariance of the variational posterior $q_\phi$, while the posterior mean is modeled by a data-dependent deep convolutional map.

## 3. Deep Variational System Identification

Building on the structured Gaussian posterior described in Section 2.3, the proposed deep variational system identification framework is introduced. The method combines a deep posterior mean map with a structured Gaussian posterior and an alternating optimization procedure. Its objective is to jointly learn (i) a variational smoother $q_\phi(\mathbf{X} \mid \mathbf{Y}, \mathbf{U})$ for latent trajectories (ii) model parameters $\vartheta$.

### *3.1. Deep Convolutional Gaussian Variational Smoother and Approximate ELBO*

The Gaussian variational posterior in (14) is equipped with a deep posterior mean function $m_\phi(\mathbf{Y}, \mathbf{U})$, while its covariance $\Sigma_\phi$ is parameterized through the steady-state Gaussian Markov construction described in Section 2.3.

For the posterior mean function $m_\phi : (\mathbf{Y}, \mathbf{U}) \mapsto \mathbb{R}^{(T+1)d_x}$, it is implemented by a non-causal dilated residual convolutional network (DRN). Owing to the non-causal architecture, each $\boldsymbol{\mu}_t$ can depend on both past and future observations and inputs, which is consistent with the smoothing setting. The detailed architecture of the DRN is specified as follows.

Let $\mathbf{z}_t := \begin{bmatrix} \mathbf{y}_t \\ \mathbf{u}_t \end{bmatrix} \in \mathbb{R}^{d_y + d_u}$, $t = 0, \ldots, T$, where $\mathbf{u}_T := \mathbf{0}$ is introduced only for sequence-length alignment. The aligned input sequence is denoted by $\mathbf{Z} := [\mathbf{z}_0, \ldots, \mathbf{z}_T]$. The DRN constructs a hierarchy of hidden sequences $H^{(\ell)} := [h_0^{(\ell)}, \ldots, h_T^{(\ell)}], \ell = 0, \ldots, L$, where $h_t^{(\ell)}$ denotes the feature vector at time $t$ in layer $\ell$. For $\ell \geq 1$, the hidden features satisfy $h_t^{(\ell)} \in \mathbb{R}^{d_h}$. The input layer is defined by $H^{(0)} := \mathbf{Z}$, i.e., $h_t^{(0)} = \mathbf{z}_t$.

The first layer applies a non-causal 1D convolution along the temporal dimension to project the input sequence into a hidden feature space of dimension $d_h$:

$$h_t^{(1)} = \sigma\big((W^{(1)} * H^{(0)})_t + b^{(1)}\big), \tag{20}$$

where $(W^{(1)} * H^{(0)})_t \in \mathbb{R}^{d_h}$ denotes the convolution output at time $t$, and $\sigma(\cdot)$ is an elementwise activation mapping $\mathbb{R}^{d_h}$ to $\mathbb{R}^{d_h}$. For $\ell = 2, \ldots, L$, the network updates the hidden sequence through residual dilated convolutional blocks of constant width $d_h$:

$$h_t^{(\ell)} = \sigma\Big(h_t^{(\ell-1)} + (W^{(\ell)} * H^{(\ell-1)})_t + b^{(\ell)}\Big), \tag{21}$$

where "$*$" denotes a non-causal dilated 1D convolution with same padding, applied along the temporal dimension. For a kernel of size $k = 2r + 1$ and dilation factor $d_\ell$, this operation can be written as

$$(W^{(\ell)} * H^{(\ell-1)})_t = \sum_{j=-r}^{r} W_j^{(\ell)} h_{t+jd_\ell}^{(\ell-1)}, \tag{22}$$

where same padding is used, with $h_s^{(\ell-1)}$ interpreted as zero for out-of-range indices $s < 0$ or $s > T$. In the first layer, each kernel slice satisfies $W_j^{(1)} \in \mathbb{R}^{d_h \times (d_y + d_u)}$ and $b^{(1)} \in \mathbb{R}^{d_h}$. For $\ell = 2, \ldots, L$, the residual blocks use $W_j^{(\ell)} \in \mathbb{R}^{d_h \times d_h}$ and $b^{(\ell)} \in \mathbb{R}^{d_h}$. In this work, we use kernel size $k = 3$, GELU activations, and the dilation schedule $d_1 = 1$ and $d_\ell = \ell - 1$ for $\ell = 2, \ldots, L$. GELU is used as a smooth activation function for gradient-based optimization of the variational objective, and the linear dilation schedule gradually enlarges the non-causal temporal receptive field without making the convolution overly sparse. With kernel size $k = 3$, the receptive field length is $W_{\text{rf}} = 1 + 2\sum_{\ell=1}^{L} d_\ell$. For the setting $L = 5$, this gives $W_{\text{rf}} = 23$.

Then, a linear readout is applied at each time step:

$$\boldsymbol{\mu}_t = W_{\text{out}} h_t^{(L)} + b_{\text{out}}, \qquad t = 0, \ldots, T, \tag{23}$$

where $W_{\text{out}} \in \mathbb{R}^{d_x \times d_h}$, and $b_{\text{out}} \in \mathbb{R}^{d_x}$. Stacking these outputs yields

$$m_\phi(\mathbf{Y},\mathbf{U}) = \begin{bmatrix} \boldsymbol{\mu}_0 \\ \vdots \\ \boldsymbol{\mu}_T \end{bmatrix} \in \mathbb{R}^{(T+1)d_x}. \tag{24}$$

Consequently, each $\boldsymbol{\mu}_t$ depends on a finite non-causal neighborhood of $(\mathbf{y},\mathbf{u})$, yielding a nonlinear and data-adaptive posterior mean map. In contrast to linear time-invariant convolutional mean parameterizations, the proposed DRN can represent state-dependent smoothing behavior while remaining fully parallelizable across time steps.

To keep the optimization unconstrained while ensuring a valid conditional covariance, a matrix-exponential parameterization of a lower-triangular factor is adopted [32]. Let $m := \frac{1}{2}d_x(d_x+1)$ and introduce free parameters $\zeta_{\text{cond}} \in \mathbb{R}^m$, mapped to a lower-triangular matrix by $\text{matl}(\cdot)$:

$$A_{\text{cond}} := \text{matl}(\zeta_{\text{cond}}) \in \mathbb{R}^{d_x \times d_x}. \tag{25}$$

The conditional factor is then defined through the matrix exponential

$$S_{\text{cond}} := \exp\big(A_{\text{cond}}\big), \tag{26}$$

which preserves the lower-triangular structure and yields strictly positive diagonal entries. As a result,

$$\Sigma_{t+1|t} = S_{\text{cond}} S_{\text{cond}}^\top \succ 0. \tag{27}$$

The cross-covariance factor $S_{\text{cross}} \in \mathbb{R}^{d_x \times d_x}$ is left unconstrained and optimized jointly with $\zeta_{\text{cond}}$ as part of $\phi$.

The trajectory covariance $\Sigma_\phi$ is kept implicit through the corresponding Markov-structured precision $\Lambda_\phi := \Sigma_\phi^{-1}$. Under the steady-state construction in Section 2.3, the pairwise marginal $q_\phi(\mathbf{x}_t, \mathbf{x}_{t+1} \mid \mathbf{Y},\mathbf{U})$ is Gaussian with covariance blocks determined by $(S_{\text{cond}}, S_{\text{cross}})$ via (18). Hence, the complete second-order structure of $q_\phi$ is encoded by the two matrix factors $S_{\text{cond}}$ and $S_{\text{cross}}$. Overall, the proposed posterior uses the DRN to model the posterior mean, while the covariance is represented through a structured steady-state Gaussian Markov parameterization.

Given this variational parameterization, the ELBO is evaluated through a differentiable approximation in which the entropy term is available in closed form and the remaining expectations are approximated numerically. Nonlinear transition and measurement expectations are approximated using tensor-product Gauss–Hermite quadrature rules defined on the corresponding pairwise and marginal Gaussian distributions. In the reported experiments, a fifth-order tensor-product Gauss–Hermite rule is used, corresponding to five nodes per Gaussian dimension. For the transition term, the quadrature rule is defined on the correlated pairwise marginal $q_\phi(\mathbf{x}_t, \mathbf{x}_{t+1} \mid \mathbf{Y},\mathbf{U})$.

Let $\{(\boldsymbol{\nu}_t^{(i)}, \boldsymbol{\nu}_{t+1}^{(i)}, \psi^{(i)})\}_{i=1}^{n_\nu}$ denote quadrature nodes and weights for the transition expectation, and define correlated pair samples by

$$\tilde{\mathbf{x}}_{t|t,t+1}^{(i)} := \boldsymbol{\mu}_t + S_{\text{ss}}\,\boldsymbol{\nu}_t^{(i)}, \tag{28}$$

$$\tilde{\mathbf{x}}_{t+1|t,t+1}^{(i)} := \boldsymbol{\mu}_{t+1} + S_{\text{cross}}\,\boldsymbol{\nu}_t^{(i)} + S_{\text{cond}}\,\boldsymbol{\nu}_{t+1}^{(i)}, \tag{29}$$

where $S_{\text{ss}}$, $S_{\text{cross}}$, and $S_{\text{cond}}$ are the covariance factors defined in the steady-state parameterization of Section 2.3. This construction matches the pairwise second-order moments of $q_\phi(\mathbf{x}_t, \mathbf{x}_{t+1} \mid \mathbf{Y},\mathbf{U})$. These quadrature nodes are local deterministic representatives of the pairwise marginal at time $t$. They are used only to approximate the transition expectation at that time step and are not intended to form a globally consistent trajectory sample across different values of $t$. Therefore, a node associated with $\mathbf{x}_{t+1}$ in the pair $(\mathbf{x}_t, \mathbf{x}_{t+1})$ need not coincide with a node associated with $\mathbf{x}_{t+1}$ in the next pair $(\mathbf{x}_{t+1}, \mathbf{x}_{t+2})$. Similarly, marginal terms such as $\mathbb{E}_{q_\phi(\mathbf{x}_t \mid \mathbf{Y},\mathbf{U})}[\log p_\vartheta(\mathbf{y}_t \mid \mathbf{x}_t)]$ are approximated using quadrature nodes $\{(\boldsymbol{\xi}^{(i)}, \omega^{(i)})\}_{i=1}^{n_\xi}$, with $\mathbf{x}_t^{(i)} := \boldsymbol{\mu}_t + S_{\text{ss}}\,\boldsymbol{\xi}^{(i)}$. Substituting these approximations into (13) yields the training objective:

$$\begin{aligned} \widehat{\mathcal{L}} = &\sum_{i=1}^{n_\xi} \omega^{(i)} \log p\Big(\mathbf{x}_0^{(i)}\Big) + \sum_{t=0}^{T}\sum_{i=1}^{n_\xi} \omega^{(i)} \log p_\vartheta\Big(\mathbf{y}_t \mid \mathbf{x}_t^{(i)}\Big) \\ &+ \sum_{t=0}^{T-1}\sum_{i=1}^{n_\nu} \psi^{(i)} \log p_\vartheta\Big(\tilde{\mathbf{x}}_{t+1|t,t+1}^{(i)} \mid \tilde{\mathbf{x}}_{t|t,t+1}^{(i)}, \mathbf{u}_t\Big) + \mathbb{H}[q_\phi], \end{aligned} \tag{30}$$

where $\omega^{(i)}$ and $\psi^{(i)}$ are the weights associated with the chosen numerical integration rules. For fixed $(n_\xi, n_\nu)$ and state dimension $d_x$, evaluating (30) has linear complexity in the trajectory length $T$, namely $\mathcal{O}((n_\xi + n_\nu)T)$ per trajectory.

Combining the closed-form entropy with the quadrature approximations yields the approximate ELBO $\widehat{\mathcal{L}}(\vartheta, \phi)$. When the model parameters are learned jointly with the variational smoother, optimization becomes imbalanced and part of the uncertainty can be misattributed to the variational posterior rather than to the noise variables. This motivates the alternating optimization procedure introduced next.

### *3.2. Alternating Optimization for Smoothing and Identification*

To address these difficulties, an alternating optimization procedure is adopted to separate variational smoothing from model identification. The ELBO admits a natural variational-EM interpretation. For fixed $\vartheta$, maximizing the variational objective with respect to $\phi$ corresponds to variational smoothing, whereas for fixed $q_\phi(\mathbf{X} \mid \mathbf{Y}, \mathbf{U})$, maximizing the expected complete-data log-likelihood with respect to $\vartheta$ corresponds to identification. In the idealized case, iteration $k$ would proceed as

$$\text{E-step:} \quad \phi^{(k+1)} \in \arg\max_{\phi} \mathcal{L}(\vartheta^{(k)}, \phi), \tag{31}$$

$$\text{M-step:} \quad \vartheta^{(k+1)} \in \arg\max_{\vartheta} \mathcal{L}(\vartheta, \phi^{(k+1)}). \tag{32}$$

However, the E-step does not admit a closed-form solution in the present nonlinear setting. Because the variational posterior is parameterized by a neural network and the system dynamics are nonlinear, the maximization with respect to $\phi$ is not analytically tractable. Instead, a generalized E-step using stochastic gradient optimization is performed. Specifically, it maximizes the approximate ELBO $\widehat{\mathcal{L}}(\vartheta, \phi)$ defined in (30). This approximation provides a stable differentiable objective for updating $\phi$.

At iteration $k$, with $\vartheta^{(k)}$ fixed, the generalized E-step updates $\phi$ by maximizing $\widehat{\mathcal{L}}(\vartheta^{(k)}, \phi)$. Starting from $\phi^{(k,0)} := \phi^{(k)}$, the inner updates take the form

$$\phi^{(k,j+1)} = \phi^{(k,j)} + \rho_{k,j}\, \widehat{\nabla}_\phi\, \widehat{\mathcal{L}}\big(\vartheta^{(k)}, \phi^{(k,j)}\big), \tag{33}$$

for $j = 0, \ldots, N_{\text{inner}} - 1$. Then, $\phi^{(k+1)} := \phi^{(k,N_{\text{inner}})}$. Gradients are taken only with respect to $\phi$, and Adam [39] is used for the inner-loop optimization.

Given the updated variational posterior $q_{\phi^{(k+1)}}$, the M-step updates $\vartheta$ by maximizing the expected complete-data log-likelihood under $q_{\phi^{(k+1)}}$:

$$\vartheta^{(k+1)} := \arg\max_{\vartheta} \mathbb{E}_{q_{\phi^{(k+1)}}} \big[ \log p_\vartheta(\mathbf{X}, \mathbf{Y} \mid \mathbf{U}) \big]. \tag{34}$$

This is the ideal M-step objective in the variational-EM interpretation. However, in the present nonlinear setting, the expectations in (34) are generally not available in closed form. Therefore, the M-step is implemented approximately using reparameterized Monte Carlo samples drawn from $q_{\phi^{(k+1)}}(\mathbf{X} \mid \mathbf{Y}, \mathbf{U})$. At iteration $k$, let $\mathbf{X}^{(s)} \sim q_{\phi^{(k+1)}}(\mathbf{X} \mid \mathbf{Y}, \mathbf{U})$, $s = 1, \ldots, N_{\text{samp}}$, denote the sampled latent trajectories. These samples provide marginal states $\{\mathbf{x}_t^{(s)}\}_{t=0}^{T}$ and consecutive pairs $\{(\mathbf{x}_t^{(s)}, \mathbf{x}_{t+1}^{(s)})\}_{t=0}^{T-1}$, which are used to construct Monte Carlo estimates of the sufficient statistics.

The sampled trajectories are first used to update the dynamics parameter $\theta$. Since the measurement map $h(\cdot)$ is fixed, $\theta$ appears only in the transition model. The update of $\theta$ is obtained by maximizing the expected transition log-likelihood:

$$\begin{aligned} &\sum_{t=0}^{T-1} \mathbb{E}_{q_{\phi^{(k+1)}}} \left[ \log p_\vartheta(\mathbf{x}_{t+1} \mid \mathbf{x}_t, \mathbf{u}_t) \right] \\ &= -\tfrac{T}{2} \log\det Q - \tfrac{1}{2} \sum_{t=0}^{T-1} \mathbb{E}_{q_{\phi^{(k+1)}}} \left[ \left\| \mathbf{x}_{t+1} - \Phi(\mathbf{x}_t, \mathbf{u}_t)\theta \right\|_{Q^{-1}}^2 \right], \end{aligned} \tag{35}$$

where $\|\mathbf{z}\|_{Q^{-1}}^2 := \mathbf{z}^\top Q^{-1} \mathbf{z}$. When the model is linear in $\theta$ under the feature map $\Phi$ in (3), maximizing (35) with respect to $\theta$ yields the normal equations

$$\begin{aligned} &\left( \sum_{t=0}^{T-1} \mathbb{E}_{q_{\phi^{(k+1)}}} \left[ \Phi(\mathbf{x}_t, \mathbf{u}_t)^\top Q^{-1} \Phi(\mathbf{x}_t, \mathbf{u}_t) \right] \right) \theta = \\ &\qquad \sum_{t=0}^{T-1} \mathbb{E}_{q_{\phi^{(k+1)}}} \left[ \Phi(\mathbf{x}_t, \mathbf{u}_t)^\top Q^{-1} \mathbf{x}_{t+1} \right]. \end{aligned} \tag{36}$$

Using the sampled trajectories $\{\mathbf{X}^{(s)}\}_{s=1}^{N_{\text{samp}}}$, the required Monte Carlo statistics are defined as:

$$\hat{A} := \frac{1}{N_{\text{samp}}} \sum_{s=1}^{N_{\text{samp}}} \sum_{t=0}^{T-1} \Phi(\mathbf{x}_t^{(s)}, \mathbf{u}_t)^\top Q^{-1} \Phi(\mathbf{x}_t^{(s)}, \mathbf{u}_t), \tag{37}$$

$$\hat{b} := \frac{1}{N_{\text{samp}}} \sum_{s=1}^{N_{\text{samp}}} \sum_{t=0}^{T-1} \Phi(\mathbf{x}_t^{(s)}, \mathbf{u}_t)^\top Q^{-1} \mathbf{x}_{t+1}^{(s)}. \tag{38}$$

This yields the ridge-regularized weighted least-squares update

$$\theta^{(k+1)} = (\hat{A} + \lambda I)^{-1} \hat{b}, \tag{39}$$

where $\lambda \geq 0$ is a ridge parameter used to improve conditioning.

The same samples are then used to update the noise statistics. Define the transition residuals and measurement residuals as $\mathbf{r}_t^{(s)} = \mathbf{x}_{t+1}^{(s)} - f_{\theta^{(k+1)}}(\mathbf{x}_t^{(s)}, \mathbf{u}_t)$, and $\mathbf{e}_t^{(s)} = \mathbf{y}_t - h(\mathbf{x}_t^{(s)})$. Their second moments are approximated by

$$\mathbb{E}[\mathbf{r}_t \mathbf{r}_t^\top] \approx \frac{1}{N_{\text{samp}}} \sum_{s=1}^{N_{\text{samp}}} \mathbf{r}_t^{(s)} (\mathbf{r}_t^{(s)})^\top, \tag{40}$$

$$\mathbb{E}[\mathbf{e}_t \mathbf{e}_t^\top] \approx \frac{1}{N_{\text{samp}}} \sum_{s=1}^{N_{\text{samp}}} \mathbf{e}_t^{(s)} (\mathbf{e}_t^{(s)})^\top. \tag{41}$$

In the experiments, the process and measurement noise covariances are parameterized as isotropic matrices, with $Q = \sigma_w^2 \mathbf{I}_{d_x}$ and $R = \sigma_v^2 \mathbf{I}_{d_y}$. Under this parameterization, the updates reduce to

$$\sigma_w^{2,(k+1)} = \frac{1}{T d_x} \sum_{t=0}^{T-1} \mathbb{E}[\|\mathbf{r}_t\|_2^2], \tag{42}$$

$$\sigma_v^{2,(k+1)} = \frac{1}{(T+1) d_y} \sum_{t=0}^{T} \mathbb{E}[\|\mathbf{e}_t\|_2^2]. \tag{43}$$

In practice, the M-step updates may be damped to improve stability. Let $\vartheta_{\text{upd}}^{(k+1)}$ denote the undamped analytic parameter update obtained from the current Monte Carlo sufficient statistics. The generalized E-step and M-step are alternated until convergence. The complete procedure is summarized in Algorithm 1.

**Algorithm 1:** Alternating optimization for variational system identification

**Input:** Initial parameters $\vartheta^{(0)}, \phi^{(0)}$, outer iterations $K$, inner E-step steps $N_{\text{inner}}$, Monte Carlo sample size $N_{\text{samp}}$, damping factor $\rho_{\text{M}}$

**for** $k = 0, 1, \ldots, K-1$ **do**

  **E-step: variational smoothing**;

  **for** $j = 0, 1, \ldots, N_{\text{inner}} - 1$ **do**

    Update $\phi$ by stochastic gradient ascent on $\widehat{\mathcal{L}}(\vartheta^{(k)}, \phi)$;

  **end**

  Set $\phi^{(k+1)} \leftarrow \phi$;

  **M-step: parameter identification**;

  Draw $N_{\text{samp}}$ samples $\mathbf{X}^{(s)} \sim q_{\phi^{(k+1)}}(\mathbf{X} \mid \mathbf{Y}, \mathbf{U})$;

  Compute Monte Carlo estimates of the sufficient statistics from $\{\mathbf{X}^{(s)}\}_{s=1}^{N_{\text{samp}}}$;

  Compute the undamped analytic update $\vartheta_{\text{upd}}^{(k+1)}$;

  Apply damping:

$$\vartheta^{(k+1)} \leftarrow (1 - \rho_{\text{M}}) \vartheta^{(k)} + \rho_{\text{M}} \, \vartheta_{\text{upd}}^{(k+1)}$$

**end**

## 4. Experiments

The proposed framework is evaluated on two simulated systems to assess both identification accuracy and optimization behavior under unknown noise statistics. The first experiment uses a nonlinear discrete-time system as a diagnostic study of noise misattribution. The second experiment considers the stochastic Duffing oscillator and adopts the experimental setting of [32], which enables direct comparison with the smooth-kernel variational identification baseline.

### *4.1. Nonlinear Discrete-Time Identification*

A nonlinear discrete-time SSM is considered:

$$x_{t+1} = ax_t + b\frac{x_t}{1+x_t^2} + cu_t + w_t,\ w_t \sim \mathcal{N}(0, \sigma_w^2),$$
$$\mathbf{y}_t = \mathbf{C}x_t + \sigma_v \boldsymbol{\varepsilon}_t,\ \boldsymbol{\varepsilon}_t \sim \mathcal{N}(0, I_M),$$

where $x_t \in \mathbb{R}$ is the latent state, $u_t \in \mathbb{R}$ is a known input, and $\mathbf{y}_t \in \mathbb{R}^M$ collects $M$ independent sensor measurements of the same state. The observation matrix is $\mathbf{C} := [1, \ldots, 1]^\top \in \mathbb{R}^{M\times 1}$. The noise sequences $\{w_t\}_{t=0}^{T-1}$ and $\{\boldsymbol{\varepsilon}_t\}_{t=0}^{T}$ are independent. In all trials, $M = 3$ sensors are used and a single trajectory of length $T = 2000$ is simulated. The initial state is drawn as $x_0 \sim \mathcal{N}(0, 1)$. The ground-truth parameters are set to $\vartheta^\star = (a^\star, b^\star, c^\star, \sigma_w^\star, \sigma_v^\star) = (0.5, 25, 5, 0.5, 0.1)$, and the input is generated as $u_t = \cos(1.2t)$ for $t = 0, \ldots, T-1$. This setting serves as a diagnostic example highlighting noise misattribution under joint end-to-end optimization.

The variational mean network receives $z_t := [\mathbf{y}_t^\top, u_t]^\top \in \mathbb{R}^{M+1}$ for $t = 0, \ldots, T$ and outputs the posterior mean trajectory. Two DRN-based identification schemes are compared in this experiment. The proposed method, denoted DRN (alt), optimizes the variational parameters using the alternating procedure in Section 3.2, with one identification update for $(a, b, c, \sigma_w, \sigma_v)$ performed every $N_{\text{inner}} = 300$ Adam updates of $\phi$. As an ablation, DRN (joint) uses the same DRN-based variational smoother but trains all parameters jointly by end-to-end optimization. For DRN (alt), the identification step uses $N_{\text{samp}} = 64$ posterior samples, ridge coefficient $\lambda = 10^{-6}$, and damping factor $\rho_{\text{M}} = 0.5$. The DRN uses depth $L = 5$, 32 hidden channels, kernel size $k = 3$, and Adam with learning rate $5 \times 10^{-3}$ and default moment parameters.

Table 1 reports results over $N_{\text{trial}} = 20$ random seeds. DRN (alt) accurately recovers the dynamics parameters $(a, b, c)$ and the parameter $\sigma_w$. DRN (joint) attains a similar state estimation RMSE, but produces substantially biased parameter estimates and a markedly inflated estimate of $\sigma_w$. The state estimation RMSE is defined as $\text{RMSE} = \sqrt{\frac{1}{(T+1)d_x}\sum_{t=0}^{T}\|\hat{\mathbf{x}}_t - \mathbf{x}_t^\star\|_2^2}$, where $\hat{\mathbf{x}}_t$ denotes the posterior mean estimate under the learned variational posterior, i.e., $\hat{\mathbf{x}}_t = \boldsymbol{\mu}_t$, and $\mathbf{x}_t^\star$ denotes the ground-truth state. These results suggest that joint optimization can achieve accurate state reconstruction while attributing part of the model mismatch to the process-noise term, thereby producing inflated estimates of $\sigma_w$ and biased estimates of the dynamics parameters. By decoupling smoothing from analytic identification updates based on Monte Carlo sufficient-statistics estimates, the alternating procedure mitigates this effect and improves parameter reliability.

**Table 1.** System identification results on the nonlinear discrete-time benchmark. The results are averaged over $N_{\text{trial}} = 20$ independent simulations (mean ± standard deviation).

| Metric | True Value | DRN (alt) [1] | DRN (joint) [2] |
|---|---|---|---|
| $a$ | 0.500 | $0.501 \pm 0.002$ | $0.752 \pm 0.001$ |
| $b$ | 25.000 | $24.956 \pm 0.077$ | $6.599 \pm 0.010$ |
| $c$ | 5.000 | $4.993 \pm 0.016$ | $3.430 \pm 0.016$ |
| $\sigma_w$ | 0.500 | $0.501 \pm 0.009$ | $2.678 \pm 0.023$ |
| $\sigma_v$ | 0.100 | $0.086 \pm 0.001$ | $0.085 \pm 0.001$ |
| State RMSE | – | $0.0577 \pm 0.0007$ | $0.0577 \pm 0.0007$ |

Note: [1] DRN-based variational smoother trained with the proposed alternating optimization scheme. [2] DRN-based variational smoother trained by joint end-to-end optimization.

### *4.2. Stochastic Duffing Oscillator*

The stochastic Duffing oscillator is next considered as a representative nonlinear benchmark with continuous-time stochastic dynamics. The system is governed by the stochastic differential equation

$$\begin{aligned}\mathrm{d}x_1(t) &= x_2(t)\,\mathrm{d}t,\\ \mathrm{d}x_2(t) &= \big[-\alpha x_1(t) - \beta x_1^3(t) - \delta x_2(t) + \gamma\cos(t)\big]\mathrm{d}t\\ &\quad + e^{\zeta_w}\,\mathrm{d}W(t),\end{aligned}$$

where $W(t)$ is a scalar Wiener process and $(\alpha, \beta, \delta, \gamma, \zeta_w)$ are unknown parameters. Discrete-time measurements are scalar observations of the position $y_k \sim \mathcal{N}\big(x_1(kT_s),\, e^{2\zeta_v}\big), k = 0, \ldots, N$. The SDE is simulated using the same numerical scheme as in [32], and trajectories are subsampled with period $T_s$ to obtain discrete-time data.

Following [32], the ground-truth parameters are set to $(\alpha^\star, \beta^\star, \delta^\star, \gamma^\star, \zeta_w^\star, \zeta_v^\star) = (-1, 1, 0.2, 0.3, -2.3, -3.0)$, and the initial state is fixed at $\mathbf{x}_0 = [1, -1]^\top$. The total simulation duration is varied over $t_f \in \{50, 100, 200, 400, 800\}$ while keeping $T_s = 0.1$ fixed, which yields $N = t_f/T_s$ measurements per trajectory. For each duration, $N_{\text{trial}} = 20$ independent simulations are performed. All methods are evaluated under the same data-generation and discretization settings, and identification is performed on the discretized model used by the simulator.

We compare three structured Gaussian variational identification methods that share the same structured Gaussian posterior and differ only in the posterior mean parameterization and training strategy. As a reference baseline, SmoothKernel (joint) uses the smooth-kernel mean parameterization in [32] with window length $n_{\text{win}} = 51$ and jointly optimizes all variables with Adam. As an architectural ablation, DRN (joint) replaces the smooth-kernel mean with our DRN-based posterior mean map (depth $L = 5$ with 32 hidden channels) while retaining joint end-to-end training. The proposed method, denoted DRN (alt), replaces joint training with the proposed alternating optimization scheme. These comparisons are designed to disentangle two factors. First, the comparison between SmoothKernel (joint) and DRN (joint) isolates the effect of replacing the smooth-kernel posterior mean with the proposed DRN-based posterior mean under the same joint-training setting. Second, the comparison between DRN (joint) and DRN (alt) isolates the effect of the proposed alternating optimization strategy, since both methods use the same deep posterior family and differ only in how the model and variational parameters are updated. All methods use 120,000 variational gradient updates in total. For DRN (alt), variational updates are performed with stochastic optimization, and one analytic identification update for $(\alpha, \beta, \delta, \gamma, \zeta_w, \zeta_v)$ is inserted every $N_{\text{inner}} = 200$ gradient steps. The identification update uses ridge coefficient $\lambda = 10^{-6}$, damping factor $\rho_{\text{M}} = 0.9$, and trapezoidal feature construction in the parameter regression. Unless otherwise stated, Adam is used with learning rate $5 \times 10^{-3}$ and its default moment parameters; the same optimizer settings are used for all trajectory durations and random trials.

We report absolute estimation errors $|\hat{\vartheta} - \vartheta^\star|$ for $(\alpha, \beta, \delta, \gamma, \zeta_w, \zeta_v)$, aggregated across trials. Figure 2 summarizes parameter and noise identification performance as a function of $t_f$. Across the tested durations, DRN (alt) generally yields lower errors than the joint baselines, with the most pronounced improvement on the noise parameters, especially the process noise $\zeta_w$. This pattern is consistent with reduced noise misattribution in the presence of strong nonlinear stochasticity.

To evaluate state estimation performance, Figure 3 visualizes the state reconstruction on a representative test trajectory. Across all trajectory durations, the average state estimation RMSE is 0.0220 for DRN (alt), compared with 0.0229 for DRN (joint) and 0.0419 for SmoothKernel (joint). These results indicate that replacing the smooth-kernel mean parameterization with a DRN-based posterior mean map improves posterior expressiveness and yields more accurate state estimates.

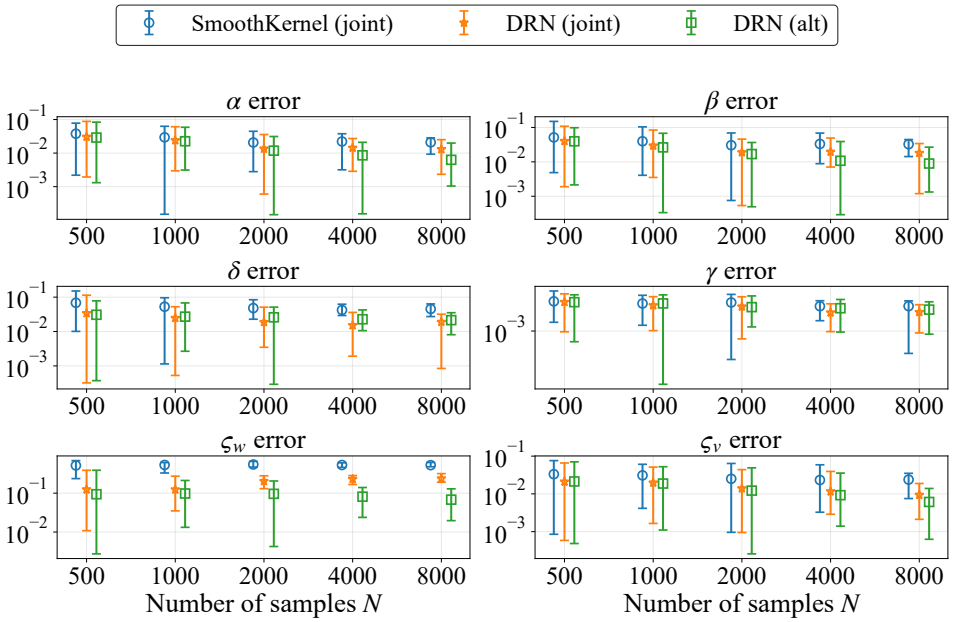


**Figure 2.** Duffing oscillator identification errors versus trajectory duration $t_f \in \{50, 100, 200, 400, 800\}$. Markers denote the mean absolute estimation error of $(\alpha, \beta, \delta, \gamma, \zeta_w, \zeta_v)$ over $N_{\text{trial}} = 20$ independent trials, and error bars indicate the minimum-to-maximum range across trials. Compared with the joint end-to-end optimization, DRN (alt) achieves improved identification accuracy.

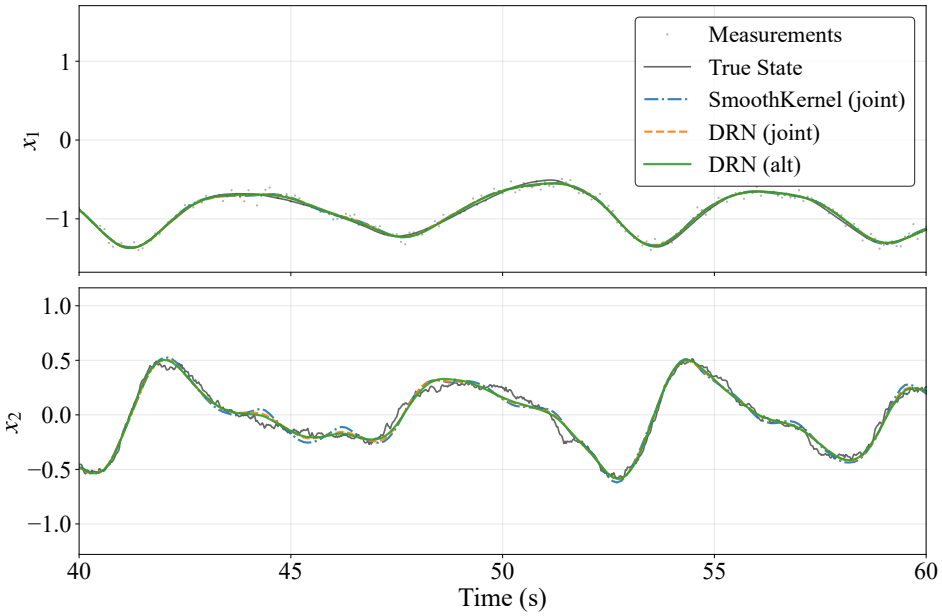


**Figure 3.** State estimation on a representative Duffing oscillator trajectory. The top and bottom panels show the first and second state components, respectively.

To examine optimization behavior, Figure 4 reports parameter estimation errors as a function of training steps. Here, "training steps" denotes the number of gradient updates of the variational parameters, with M-step updates inserted every $N_{\text{inner}} = 200$ gradient steps for DRN (alt). DRN (alt) exhibits faster convergence than the joint end-to-end baselines, benefiting from the sample-based analytic identification updates that allow the parameters to move rapidly toward values consistent with the current variational approximation. While the discrete M-steps may induce minor fluctuations, DRN (alt) avoids the optimization plateaus often encountered in fully gradient-based joint training and reaches a lower error floor more quickly.

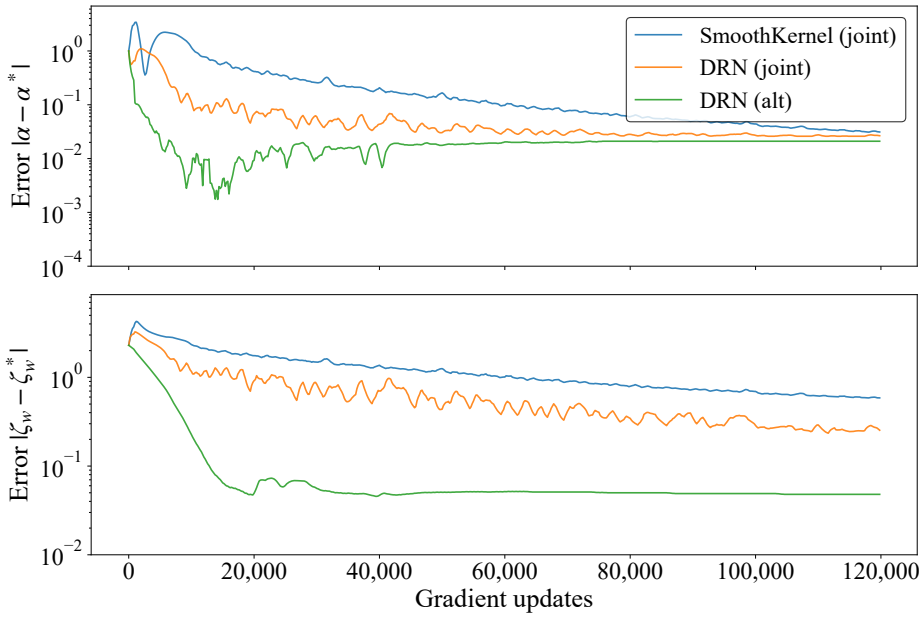


**Figure 4.** Convergence of parameter estimation errors for the Duffing oscillator ($t_f = 50$). DRN (alt) converges more rapidly than the joint gradient-based baselines, indicating that the analytic identification updates are effective.

## 5. Conclusions

This paper presented a deep variational system identification method for nonlinear SSM within a structured Gaussian variational framework. The proposed approach places latent-state smoothing and parameter identification in a unified probabilistic setting, where the posterior mean is parameterized by a deep convolutional network, posterior uncertainty is represented by a Markov-structured Gaussian family, and the two tasks are coordinated

through alternating optimization. More broadly, the framework can be viewed as an attempt to combine deep neural parameterization with classical system identification in a principled and structured manner.

A main limitation of the current approach is that the adopted steady-state covariance parameterization trades posterior expressiveness for scalability. Because its covariance factors are shared across time, it is less flexible than fully time-varying posterior parameterizations in capturing time-local posterior variability. Future work will consider richer posterior families, such as time-varying or input-dependent covariance parameterizations, extensions to nonlinear or non-Gaussian observation models, and sparse or low-rank approximations for higher-dimensional systems.

**Author Contributions:** K.F.: methodology, software, investigation, and writing—original draft preparation; G.M.: methodology, formal analysis, supervision, and funding acquisition; Y.W.: conceptualization, methodology, and supervision; Z.Y.: conceptualization, methodology, and supervision; X.D.: visualization and validation; M.W.: project administration, supervision, and writing—review and editing.

**Funding:** This work was supported by the National Natural Science Foundation of China under Grant 62503185.

**Data Availability Statement:** The code is available from the corresponding author upon reasonable request.

**Conflicts of Interest:** The authors declare no conflict of interest.

**Use of AI and AI-Assisted Technologies**: During the preparation of this work, the authors used DeepSeek for language polishing and proofreading. The authors subsequently reviewed and edited the manuscript and take full responsibility for its content.

## References

1. Ljung, L. System identification. In *Signal Analysis and Prediction*; Birkhäuser: Boston, MA, USA, 1998; pp. 163–173.
2. Yu, C.; Ljung, L.; Verhaegen, M. Identification of structured state-space models. *Automatica* **2018**, *90*, 54–61. https://doi.org/10.1016/j.automatica.2017.12.023.
3. Shen, T.; Dong, Y.; He, D.; et al. Online identification of time-varying dynamical systems for industrial robots based on sparse Bayesian learning. *Sci. China Technol. Sci.* **2022**, *65*, 386–395. https://doi.org/10.1007/s11431-021-1916-0.
4. Chen, S.; Werling, K.; Wu, A.; et al. Real-time model predictive control and system identification using differentiable simulation. *IEEE Robot. Autom. Lett.* **2023**, *8*, 312–319. https://doi.org/10.1109/LRA.2022.3213009.
5. Marmarelis, V.Z. *Nonlinear Dynamic Modeling of Physiological Systems*; John Wiley & Sons: Hoboken, NJ, USA, 2004.
6. Katayama, T. *Subspace Methods for System Identification*; Springer: London, UK, 2005.
7. Verhaegen, M.; Verdult, V. *Filtering and System Identification: A Least Squares Approach*; Cambridge University Press: Cambridge, UK, 2007.
8. Wang, Y.; Fang, K.; Ma, G.; et al. Learning linear state-space models with sparse system matrices. In Proceedings of the International Conference on Learning Representations, Rio de Janeiro, Brazil, 23–27 April 2026.
9. Ljung, L. Perspectives on system identification. *Annu. Rev. Control* **2010**, *34*, 1–12. https://doi.org/10.1016/j.arcontrol.2009.12.001.
10. Ninness, B. Some system identification challenges and approaches. *IFAC Proc. Vol.* **2009**, *42*, 1–20. https://doi.org/10.3182/20090706-3-FR-2004.00001.
11. Särkkä, S. *Bayesian Filtering and Smoothing*; Cambridge University Press: Cambridge, UK, 2013.
12. Schön, T.B.; Lindsten, F.; Dahlin, J.; et al. Sequential Monte Carlo methods for system identification. *IFAC-PapersOnLine* **2015**, *48*, 775–786.
13. Dutra, D.A.A. Uncertainty estimation in equality-constrained MAP and maximum likelihood estimation with applications to system identification and state estimation. *Automatica* **2020**, *116*, 108935. https://doi.org/10.1016/j.automatica.2020.108935.
14. Shumway, R.H.; Stoffer, D.S. An approach to time series smoothing and forecasting using the EM algorithm. *J. Time Ser. Anal.* **1982**, *3*, 253–264.
15. Dempster, A.P.; Laird, N.M.; Rubin, D.B. Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. Ser. B* **1977**, *39*, 1–22. https://doi.org/10.1111/j.2517-6161.1977.tb01600.x.
16. Schön, T.B.; Wills, A.; Ninness, B. System identification of nonlinear state-space models. *Automatica* **2011**, *47*, 39–49. https://doi.org/10.1016/j.automatica.2010.10.013.
17. Kalman, R.E. A New Approach to Linear Filtering and Prediction Problems. *J. Basic Eng.* **1960**, *82*, 35–45. https://doi.org/10.1115/1.3662552.
18. Rauch, H.E.; Tung, F.; Striebel, C.T. Maximum Likelihood Estimates of Linear Dynamic Systems. *AIAA J.* **1965**, *3*, 1445–1450. https://doi.org/10.2514/3.3166.
19. Julier, S.J.; Uhlmann, J.K. New Extension of the Kalman Filter to Nonlinear Systems. In Proceedings of the Signal Processing, Sensor Fusion, and Target Recognition VI, Orlando, FL, USA, 21–24 April 1997; Volume 3068, pp. 182–193.
20. Arasaratnam, I.; Haykin, S. Cubature Kalman Filters. *IEEE Trans. Autom. Control* **2009**, *54*, 1254–1269. https://doi.org/10.1109/TAC.2009.2019800.

21. Kokkala, J.; Solin, A.; Särkkä, S. Sigma-Point Filtering and Smoothing Based Parameter Estimation in Nonlinear Dynamic Systems. *J. Adv. Inf. Fusion* **2016**, *11*, 15–30.
22. Arulampalam, M.S.; Maskell, S.; Gordon, N.; et al. A Tutorial on Particle Filters for Online Nonlinear/Non-Gaussian Bayesian Tracking. *IEEE Trans. Signal Process.* **2002**, *50*, 174–188. https://doi.org/10.1109/78.978374.
23. Doucet, A.; de Freitas, N.; Gordon, N.J. (Eds.) *Sequential Monte Carlo Methods in Practice*; Springer: New York, NY, USA, 2001.
24. Andrieu, C.; Doucet, A.; Singh, S.S.; et al. Particle methods for change detection, system identification, and control. *Proc. IEEE* **2004**, *92*, 423–438.
25. Beal, M.J. Variational Algorithms for Approximate Bayesian Inference. Ph.D. Thesis, University of London, London, UK, 2003.
26. Blei, D.M.; Kucukelbir, A.; McAuliffe, J.D. Variational Inference: A Review for Statisticians. *J. Am. Stat. Assoc.* **2017**, *112*, 859–877. https://doi.org/10.1080/01621459.2017.1285773.
27. Lindfors, M.; Chen, T. Regularized LTI system identification in the presence of outliers: A variational EM approach. *Automatica* **2020**, *121*, 109152. https://doi.org/10.1016/j.automatica.2020.109152.
28. Tan, L.S.L.; Nott, D.J. Gaussian variational approximation with sparse precision matrices. *Stat. Comput.* **2018**, *28*, 259–275. https://doi.org/10.1007/s11222-017-9722-9.
29. Quiroz, M.; Nott, D.J.; Kohn, R. Gaussian variational approximations for high-dimensional state space models. *Bayesian Anal.* **2023**, *18*, 989–1016.
30. Barfoot, T.D.; Forbes, J.R.; Yoon, D.J. Exactly Sparse Gaussian Variational Inference with Application to Derivative-Free Batch Nonlinear State Estimation. *Int. J. Robot. Res.* **2020**, *39*, 1473–1502. https://doi.org/10.1177/0278364920937608.
31. Courts, J.; Wills, A.G.; Schön, T.B.; et al. Variational system identification for nonlinear state-space models. *Automatica* **2023**, *147*, 110687. https://doi.org/10.1016/j.automatica.2022.110687.
32. Dutra, D.A.A. Parameterizations for large-scale variational system identification using unconstrained optimization. *Automatica* **2025**, *173*, 112086. https://doi.org/10.1016/j.automatica.2024.112086.
33. Chung, J.; Kastner, K.; Dinh, L.; et al. A Recurrent Latent Variable Model for Sequential Data. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 7–12 December 2015; Volume 28.
34. Fraccaro, M.; Sønderby, S.R.K.; Paquet, U.; et al. Sequential Neural Models with Stochastic Layers. In Proceedings of the Advances in Neural Information Processing Systems, Barcelona, Spain, 5–10 December 2016; Volume 29.
35. Krishnan, R.G.; Shalit, U.; Sontag, D. Structured Inference Networks for Nonlinear State Space Models. In Proceedings of the AAAI Conference on Artificial Intelligence, San Francisco, CA, USA, 4–9 February 2017; Volume 31, pp. 2101–2109.
36. Karl, M.; Soelch, M.; Bayer, J.; et al. Deep Variational Bayes Filters: Unsupervised Learning of State Space Models from Raw Data. In Proceedings of the International Conference on Learning Representations, Toulon, France, 24–26 April 2017.
37. Andersson, C.; Ribeiro, A.H.; Tiels, K.; et al. Deep Convolutional Networks in System Identification. In Proceedings of the 2019 IEEE 58th Conference on Decision and Control (CDC), Nice, France, 11–13 December 2019; pp. 3670–3676.
38. Pillonetto, G.; Aravkin, A.; Gedon, D.; et al. Deep networks for system identification: A survey. *Automatica* **2025**, *171*, 111907. https://doi.org/10.1016/j.automatica.2024.111907.
39. Kingma, D.P.; Ba, J. Adam: A Method for Stochastic Optimization. *arXiv* **2014**, arXiv:1412.6980.