Analyzing and Improving Supervised Nonlinear Dynamical Probabilistic Latent Variable Model for Inferential Sensors

Zhichao Chen, Hao Wang, *Student Member, IEEE*, Guofei Chen, Yiran Ma, Le Yao, *Member, IEEE*, Zhiqiang Ge, *Senior Member, IEEE*, and Zhihuan Song

***Abstract*—Nonlinear dynamical probabilistic latent variable model (NDPLVM) and its variants, essential in industrial inferential sensors, face challenges in latent space inference and deep learning (DL) backend implementation. The first issue arises from the assumption that covariates directly infer the latent variable, potentially leading to inaccuracies. The second issue involves the discrepancy between the probabilistic distribution function form of NDPLVMs and data sample-based operation of DL backends. Addressing these, this study introduces the optimal control-NDPLVM (OC-NDPLVM), a model designed to enhance performance by analyzing NDPLVMs learning and tackling these issues. For the first problem, NDPLVMs' learning is reinterpreted as an optimization problem, solved by alternating direction method of multipliers, and selecting the inference network's input via studying optimal solution's structure. To address the second issue, OC-NDPLVM adapts mean and covariance equations for compatibility with DL backends. This model's effectiveness is validated through experiments on inferential sensor datasets.**

***Index Terms*—Inferential sensor, machine learning, probabilistic latent variable model (PLVM), variational inference.**

Manuscript received 25 April 2024; revised 20 June 2024; accepted 16 July 2024. This work was supported in part by the National Natural Science Foundation of China under Grant 61933013 and Grant 92167106, in part by the National Science and Technology Major Project of China under Grant 2022ZD0120001, and in part by the Jiangsu Provincial Scientific Research Center of Applied Mathematics under Grant BK20233002. Paper no. TII-24-1963. *(Corresponding authors: Zhiqiang Ge; Zhihuan Song.)*

Zhichao Chen and Zhihuan Song are with the Guangdong Provincial Key Laboratory of Petrochemical Equipment Fault Diagnosis, Guangdong University of Petrochemical Technology, Maoming 525000, China, and also with the State Key Laboratory of Industrial Control Technology, College of Control Science and Engineering, Zhejiang University, Hangzhou 310027, China (e-mail: 12032042@zju.edu.cn; songzhihuan@zju.edu.cn).

Hao Wang, Guofei Chen, and Yiran Ma are with the State Key Laboratory of Industrial Control Technology, College of Control Science and Engineering, Zhejiang University, Hangzhou 310027, China (e-mail: 22032130@zju.edu.cn; gfchen@zju.edu.cn; mayiran@zju.edu.cn).

Le Yao is with the School of Mathematics, Hangzhou Normal University, Hangzhou 311121, China, and also with the Qinting Data and Intelligence Company, Ltd., Hangzhou 311121, China (e-mail: yaole@hznu.edu.cn).

Zhiqiang Ge is with the School of Mathematics, Southeast University, Nanjing 210096, China (e-mail: zhiqiang.ge@hotmail.com).

This article has supplementary material provided by the authors and color versions of one or more figures available at https://doi.org/10.1109/TII.2024.3435466.

Digital Object Identifier 10.1109/TII.2024.3435466

I. Introduction

IN INDUSTRIAL manufacturing, accurately inferring hard-to-measure quality variables (labels) from easy-to-measure process variables (covariates) are crucial for applications such as energy consumption estimation and product quality control [1]. Probabilistic latent variable models (PLVMs) and its variants [2] excel in this area. These models effectively compress spatial(-temporal) patterns into a low-dimensional, (Markovian latent space,) gaining popularity due to their superior ability to capture the nonlinearity, evolution, and uncertainty inherent in the data.

Initially designed for unsupervised learning tasks, PLVMs have gradually been applied to inferential sensor tasks, which are categorized as (semi-)supervised learning tasks based on reference [3] by concatenating covariates and labels as observational data. It should be pointed out that, conventional PLVMs like probabilistic principal component analysis, probabilistic factor analysis, and probabilistic independent component analysis, however, struggle to adapt to multimodal industrial data due to the reliance on the unimodal assumption of observational data. To alleviate this issue, researchers have explored the use of finite mixture model-based approaches [4], converting PLVMs into multimodal PLVMs [5]. In addition, to mitigate the heavy-tailed nature of industrial measurement data, the application of Student's-t distribution has been proposed for a "robust inferential sensor" [6]. Furthermore, considering the dynamic properties of processes, the introduction of the dynamic PLVM (DPLVM) [4] has been proposed for inferential sensor modeling. Notable adaptations include Ge et al. [7] modified the DPLVM for inferential sensor applications, Shang et al. [8] incorporated slow feature analysis (SFA) to capture the slowest varying features in industrial processes, and Ma et al. [9] combined finite mixture models with linear dynamical systems to propose the switching linear dynamical system (SLDS) that accounts for both dynamical and multimodal properties.

Recently, with the advancements in deep learning (DL) techniques, deep neural modules have increasingly been incorporated into the modeling of PLVMs, aiming to enhance the accuracy of inferential sensors. Unlike traditional PLVM structures, these deep neural modules are typically "noninvertible," a characteristic that makes the application of the conventional (variational) expectation maximization (EM) algorithm challenging (as previous works often relied on matrix

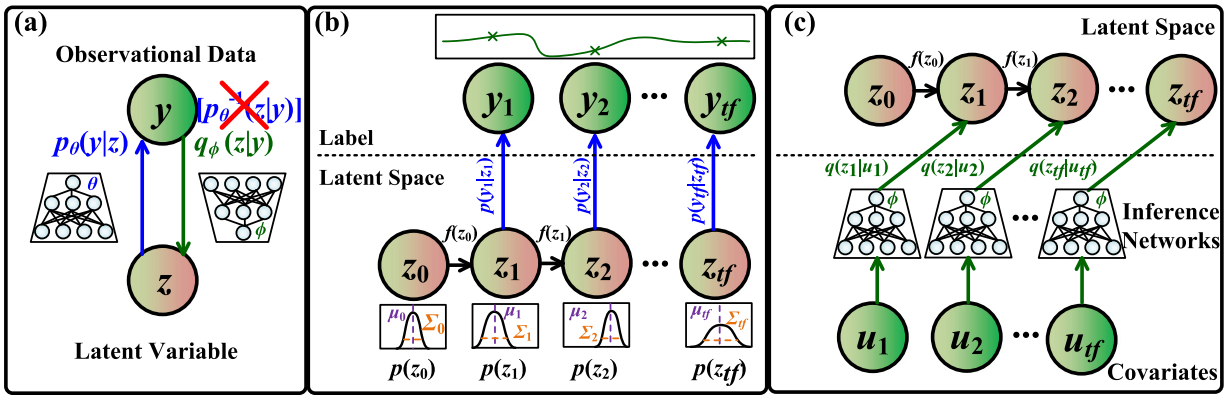


Fig. 1. Illustration of: (a) AVI, with a red cross indicating the impossibility of reversing $p_\theta(y|z)$ when modeled by a neural network. (b) The generative network (decoder) of NDPLVMs with unimodal latent space for inferential sensor modeling (we plot label y merely). (c) The inference network (encoder) of NDPLVMs with unimodal latent space for inferential sensor modeling.

(pseudo-)inverse to revert the generative model for latent variable inference in E-step). To address this challenge, as shown in Fig. 1(a), Kingma et al. [10] introduced an additional neural network, known as the inference network and proposed a novel EM algorithm called amortized variational inference (AVI). Building on this development, Shen et al. [11] proposed the nonlinear probabilistic latent variable regression (NPLVR) approach, incorporating multilayer perceptrons into the training of PLVMs.

Notably, the advancements in DL and AVI techniques have also facilitated the development of nonlinear variants of DPLVMs, namely nonlinear dynamical probabilistic latent variable models (NDPLVMs). These models incorporate various neural architectures, such as long-short term memory (LSTM) and gated-recurrent unit (GRU), to transcend the Markov assumption of the latent space, as summarized in [12]. Currently, NDPLVMs are gaining traction in inferential sensor modeling, as demonstrated by models like deep Bayesian SFA (DBPSFA) model [13] and dynamical mixture variational autoencoder regression (DMVAER) model [14].

Despite advancements in network architectures for feature extraction, applying NDPLVMs to inferential sensor tasks raises two critical yet overlooked questions from the perspective of inference network and model implementation (It is necessary to highlight that the inference network plays a crucial role in "reversing" the generative network within the context of AVI, as mentioned above). We define key elements for clarity: unobservable latent variable (z), covariates (u), label (y), inference network q, generative network p, and the transition function (f). The NDPLVMs training algorithm employs AVI, which utilizes two neural networks: the inference network (q) to deduce latent variable z from data (u), and the generative network (p) to decode label y from z. This approach aims to minimize two primary components: the regularization term, indicating the deviation between inferred and prior latent spaces, and the likelihood term, the difference between original and generated data, both quantified using probabilistic density functions (pdf). To better understand the challenges, the architecture of conventional NDPLVMs for inferential sensor modeling, which consists of a generative network (decoder) and an inference network (encoder), are delineated in Fig. 1(b) and (c), respectively. The essential challenges in this process are summarized as follows:

1) *Inaccurate Inference of Latent Space:* According to Bayes' theorem, the inference network's inputs should align with the structure of the generative network $p(y|z)$ / $p(u, y|z)$. In the context of inferential sensor tasks, where the generative network invariably involves y, the inference network should ideally be formulated as $q(z|y)$ / $q(z|u, y)$ based on the comparison of Fig. 1(a) and (b). However, most of current works have not incorporated the label information y into the inference network, as depicted in Fig. 1(c). This omission could limit model performance due to inaccurate inference of latent variable z. The key to addressing this issue is reframing the model learning problem into an optimization problem and selecting the inference network's input based on solving the optimization problem.
2) *Model Implementation Within DL Backends:* As illustrated by the top and bottom rectangles in Fig. 1(b), it is imperative to note that the foundational framework of NDPLVMs derivation is predicated on the pdf form. However, DL backends [15] are fundamentally structured around individual data samples, which are instances sampled from the pdf. This divergence between the theoretical pdf and practical data samples engenders difficulties in computational realization. To delineate further, consider f tasked with mapping a latent variable z over a time increment from t to $t+1$ ($z_{t+1} = f(z_t)$). Drawing upon the celebrated Liouville's theorem, the pdf must adhere to the equation $p(z_{t+1}) \det|\partial_z f(z_t)| = p(z_t)$, which necessitates the computation of the Jacobian matrix. This requirement, unfortunately, gives rise to a substantial realization in network implementation by DL backends. Consequently, a pivotal aspect is the derivation of moment expressions like mean and covariance, a step vital in bridging the gap between pdf and DL backends.

The key to addressing the aforementioned issues lies in conducting a fundamental *ab initio* analysis of NDPLVMs from the

perspective of learning objective derivation, learning objective optimization, and model implementation. This involves thoroughly understanding the employed assumptions and strategies when applying NDPLVMs to inferential sensors, validating the effectiveness of these assumptions and strategies based on rigorous mathematical principles, and proposing novel approaches to replace any unreasonable aspects.

To these ends, under the task of supervised learning, this article addresses the identified challenges by introducing a new NDPLVM, termed optimal control-NDPLVM (OC-NDPLVM), tailored for inferential sensor tasks. Specifically, we derive its learning objective using stochastic differential equation theory, examine its parameter learning based on the celebrated alternating direction multiplier method (ADMM), and rigorously analyze its architectural components. In this analysis procedure, our approach reveals that the inference network mirrors an OC subproblem in ADMM, providing the analysis of solution property of OC problem to guide the selection of inference network's input and therefore address issue 1). To address issue 2), we rigorously analyze the moment expressions within the neural network structure. Furthermore, we summarize the training and testing inference algorithm and discuss its convergence properties. Finally, empirical validation is provided through experiments on two industrial process datasets, demonstrating the efficacy of our approach. In summary, this article's contributions are summarized as follows:

1) *Ab Initio Analysis for Model Accuracy Improvement:* We conduct a foundational *ab initio* analysis of NDPLVMs for inferential sensor tasks, drawing on mathematical theories from stochastic differential equations and OC. This analysis seeks to enhance the accuracy of inferential sensors by rigorously examining, validating, and, when necessary, revising the assumptions involved in the derivation of learning objectives, parameter estimation, and model implementation.
2) *Novel Learning Objective and Parameter Learning Procedure:* Throughout our analysis, we rederive a novel learning objective for NDPLVMs, recasting it as an optimization problem. We then propose a novel algorithm based on the ADMM for efficient model learning.
3) *OC-NDPLVM Development:* By leveraging the learning objective and ADMM-based learning procedure, we recognize that the inference network of NDPLVM acts as a simulator for the control signal of an OC problem, select the input for the inference network based on the solution structure of the OC problem, and consequently introduce a novel NDPLVM named OC-NDPLVM.
4) *Moment Expressions for Numerical Implementation:* For practical model implementation, we derive an approximation of the moment expressions in OC-NDPLVM. This approximation is based on the analysis of moments in SDEs, facilitating numerical implementation.

Organization: The rest of this article is organized as follows: We introduce preliminary concepts adopted this study in Section II. We then derive the architecture, learning objective, learning algorithm, and the approximate moment expressions of OC-NDPLVM in Section III. We finally demonstrate the efficacy of OC-NDPLVM with two inferential sensor tasks in Section IV. Finally, Section V concludes this article.

II. Preliminaries

A. Amortized Variational Inference

Let y and z be the observed and latent variables, respectively. Variational inference tends to approximate the posterior distribution of the latent variable $p(z|y)$ with the variational distribution $q(z)$ by minimizing their Kullback–Leiber divergence (KL divergence), which can be reformulated as the maximization of the Evidence Lower BOund (ELBO) for model training [16]:

$$\begin{aligned}\mathbb{D}_{\mathrm{KL}}(q(z)\|p(z|y)) &= \int q(z)\log\frac{q(z)}{p(z|y)}\mathrm{d}z \\ &= \overbrace{\int q(z)\left[\log\frac{q(z)}{p(z)} - \log p(y|z)\right]\mathrm{d}z}^{\text{ELBO}} + \log p(y) \end{aligned} \tag{1}$$

where $\mathbb{D}_{\mathrm{KL}}(q(z)\|p(z|y))$ is the KL divergence between $q(z)$ and $p(z|y)$. From the derivation presented in (1), it becomes apparent that our learning objective evolves to be the ELBO, given that $\log p(y)$ retains its status as a constant.

Note that, the optimal $q^*(z)$ is approximate to $p(z|y)$. Built upon this, to estimate the optimal variational distribution $q(z)$, AVI employs a stochastic function $q_\phi(z|y)$ that maps the observed variable to the latent variable belonging to the variational posterior density; the parameter ϕ is learned during the optimization process [16]. Moreover, in the context of AVI, it is a common assumption that $q_\phi(z)$ can effectively model the optimal variational distribution $q^*(z)$. If we consider the optimal variational distribution $q^*(z)$ as a function, the fundamental goal of AVI is to identify the input variable of this function and approximate it using a function parameterized by ϕ. In this way, the model can infer the latent variables for new data points, without rerunning the optimization process.

B. Stochastic Differential Equation

Let $(\Omega, \mathcal{F}, \mathbb{P})$ [17] be a probabilistic space, where Ω is the sample space, σ-algebra $\mathcal{F}$ is the set of events, and $\mathbb{P}$ is probability measure: $\mathcal{F} \mapsto [0, 1]$. Let W_t be a $\mathcal{F}_t$-adapted Wiener process on this probabilistic space, $b(x, t)$ and L be two $\mathcal{F}_t$-adapted stochastic process, we have an Itô process

$$x(t) = x(0) + \int_0^t b(x,\tau)\mathrm{d}\tau + \int_0^t L\mathrm{d}W_t \tag{2}$$

which is the solution to the stochastic differential equation

$$\mathrm{d}x(t) = b(x,t)\mathrm{d}t + L\mathrm{d}W_t \tag{3}$$

where $b(x, t)$ is referred to the drift term, L is referred to the volatility term. More detailed information about these concepts are given in Section S.I of Supplementary Material.

Based on the above mentioned concepts, the likelihood ratio of two Itô processes is given as follows based on the celebrated Girsanov theorem and Radon–Nikodym theorem [17].

Theorem 1 (Likelihood ratio of Itô Process): Based on (3), we can introduce two Itô processes as follows:

$$\begin{aligned} \mathrm{d}x &= f(x,t)\mathrm{d}t + \mathrm{d}W, x(0) = x_0 \\ \mathrm{d}y &= g(y,t)\mathrm{d}t + \mathrm{d}W, y(0) = x_0 \end{aligned} \tag{4}$$

where $f(x,t)$ and $g(y,t)$ are drift terms. The Radon–Nikodym derivative along their respective path measures $\mathbb{P}$ and $\mathbb{Q}$ is given by

$$\begin{aligned} \frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}(x) = \exp\Bigg(-\frac{1}{2}\int_0^t \|g(x,\tau) - f(x,\tau)\|^2 \mathrm{d}\tau \\ + \int_0^t (g(x,\tau) - f(x,\tau))^\top \mathrm{d}W_t\Bigg). \end{aligned} \tag{5}$$

From this theorem, we can further obtain the KL divergence of two Itô processes as follow [17] (the detailed derivation is given in Section S.II.A of Supplementary Material)

$$\mathbb{D}_{\mathrm{KL}}(\mathbb{Q}_t \| \mathbb{P}_t) = \mathbb{E}_{\mathbb{Q}(z)}\left[\frac{1}{2}\int_0^t \|\nu\|^2 \mathrm{d}\tau\right] \tag{6}$$

where ν is defined as the follow equation according to (5)

$$L\nu = g(x,\tau) - f(x,\tau) \tag{7}$$

and $\mathbb{E}$ is expected operator.

Note that, compared to conventional KL divergence between two d-dimensional Gaussian distribution [denoted as $q(z) \sim \mathcal{N}(\mu_1, \Sigma_1)$ and $p(z) \sim \mathcal{N}(\mu_2, \Sigma_2)$], which widely applied in conventional NDPLVMs derivation

$$\begin{aligned} \mathbb{D}_{\mathrm{KL}}(q(z)\|p(z)) = \frac{1}{2}\Bigg[\log\frac{\det\Sigma_1}{\det\Sigma_2} + \mathrm{Tr}\left(\Sigma_1^{-1}\Sigma_2\right) \\ + (\mu_1 - \mu_2)^\top \Sigma_1^{-1}(\mu_1 - \mu_2) - d\Bigg]. \end{aligned} \tag{8}$$

The KL divergence between two Itô processes, as defined in (8), adopts a quadratic form that eliminates the need for matrix inversion computations. This simplification greatly facilitates the process of model derivation. Moreover, the Itô process is closely linked to Brownian motion, a phenomenon frequently observed in various natural occurrences. It is worth emphasizing that industrial processes commonly employ the Kalman Filter, which can be viewed as a variant of the Itô process [17]. Taking these factors into consideration, we have chosen the Itô process as the model prior in this article.

C. ADMM Algorihtm

Consider a special case where the decision variables w and v in the objective function term are separable (assume f and h are convex)

$$\begin{aligned} \min \quad & \mathrm{Obj}(v,w) = f(w) + h(v) \\ \text{s.t.} \quad & \mathrm{A}w + \mathrm{B}v = \mathrm{c}.. \end{aligned} \tag{9}$$

According to the augmented Lagrangian multiplier method [18], we can cast (9) to an unconstrained optimization problem, with the objective function defined as follows:

$$\begin{aligned} \mathcal{L} = f(w) + h(v) + \lambda^\top(Aw + Bv - c) \\ + \frac{\rho}{2}\|Aw + Bv - c\|_2^2 \end{aligned} \tag{10}$$

where λ is Lagrangian multiplier, and ρ is quadratic penalty coefficient. The celebrated ADMM algorithm [18] is a common algorithm to solve (10), where variables w, v, and λ are optimized separately at each iteration

$$\begin{cases} w^{k+1} = \arg\min_w \mathcal{L}(w, v^k) \\ v^{k+1} = \arg\min_v \mathcal{L}(w^{k+1}, v) \\ \lambda^{k+1} = \lambda^k + \rho(Aw^{k+1} + Bv^{k+1} - c) \end{cases}. \tag{11}$$

III. Proposed Approach: Analyzing and Improving NDPLVMs

In this section, we perform an *ab initio* analysis of NDPLVMs and introduce our OC-NDPLVM to address challenges related to the "inaccurate inference of latent space" and "model integration within DL backends." On this basis, the structure of this section is organized as follows: In Section III-A, we define the problem we aim to solve and outline the assumptions underlying our approach. Section III-B derives the learning objective, following the approach of previous NDPLVMs. In Section III-C, we address the "inaccurate inference of latent space" issue by optimizing the learning objective and selecting the inference network's input throughout investigating the solution structure of OC signal. Sections III-D and III-E tackle the "model implementation within DL backend" challenge by exploring the solution of the OC problem and deriving the relevant moment expressions, respectively. The model architecture and convergence analysis are presented in Sections III-F and III-G, respectively.

A. Problem Statement

In this study, as highlighted in our title, our focus is solely on inferential sensor tasks under a *supervised learning context*. This domain represents a specialized subset of time-series analysis. Extending the concepts presented in [19], we define the forecast horizon as H and the historical sequence length as T. Based on this framework, our task is as follows: Given the historical sequence of key indices $y_{1:\mathrm{T}} \in \mathbb{R}^{\mathrm{T}}$ and the sequence of covariates $u_{1:\mathrm{T+H}} \in \mathbb{R}^{\mathrm{D}\times(\mathrm{T+H})}$, we aim to predict the key index for the next H steps, specifically $y_{\mathrm{T+1:T+H}} \in \mathbb{R}^{\mathrm{H}}$.

B. Learninig Objective Derivation

1) ELBO Acquirement: Based on Section III-A, the learning objective can be defined as follows to maximize the log-likelihood of $p(\vec{y}|\vec{u})$:

$$\arg\max \ \log p(\vec{y}|\vec{u}) = \arg\max \ \log \int p(\vec{y}, \vec{z}|\vec{u})\mathrm{d}z. \tag{12}$$

However, the right-hand-side of (12) is intractable. To solve this problem, the assumption on latent space z is introduced.

As introduced in Section II-B, compared to simpler random walks, the Itô process accommodates a higher degree of complexity and variability in modeling dynamic systems. Moreover, the Itô process, when utilized as a prior, lends itself to the direct application of established results and theorems from the theory of stochastic processes. Last, in Bayesian inference, a critical task is to update our knowledge about unknown parameters as new data becomes available.

Therefore, based on Sections II-A and II-B, we first define the prior Itô process to describe the transition between states at different time with parameter θ as follows:

$$\mathrm{d}z = f_\theta(z,u)\mathrm{d}t + L\mathrm{d}W_t \tag{13}$$

where Wiener process W_t has the spectral density matrix Q, which is set as $\mathcal{I}$ in this article.

To align with the prior Itô process, based on the concept of variational inference introduced in Section II-A, the posterior is designed as another Itô process with parameter ϕ as follows to approximate the prior Itô process:

$$\mathrm{d}z = f_\phi(z,u)\mathrm{d}t + L\mathrm{d}W_t. \tag{14}$$

Besides, we define ν as follows:

$$L\nu = f_\phi(z,u) - f_\theta(z,u). \tag{15}$$

Based on (14) and (15), the inference network q with parameter ϕ, as mentioned in Section II-A, is transformed into the control policy ν (i.e., the control policy ν is simulated by the inference network: $\nu = q_\phi(\cdot)$). It is important to note that the input to q remains unspecified and can be elucidated by examining the structural solution of the OC policy ν. *In other words, the task of determining the input to the network q is addressed by exploring the solution structure of the OC ν.*

Consolidating (12) to (15), we can obtain the following proposition of our model learning objective:

Proposition 2: Optimizing (12) is equivalent to optimize the problem defined as follows:

$$\begin{aligned}&\min_{\theta,\nu}\sum_{t=1}^{\mathrm{T}}\left\{\mathbb{E}_{\mathbb{Q}(z)}\left[-\log p_\theta(y_t|z_t,u_{1:t}) + \int_{t-1}^{t}\frac{1}{2}\|\nu\|^2\mathrm{d}\tau\right]\right\}\\ &\text{s.t. } \mathrm{dz} = \mathrm{f}_\phi(\mathrm{z},\mathrm{u})\mathrm{dt} + \mathrm{LdW_t}\\ &\qquad = f_\theta(z,u)\mathrm{d}t + L\nu\mathrm{d}t + L\mathrm{d}W_t.\end{aligned} \tag{16}$$

Proof: The proof is given in Section S.II.B. of Supplementary Material. □

2) Upper Bound of ELBO for Inferential Sensor: In Proposition 2, we reformulate the parameter learning optimization problem associated with NDPLVMs. This reformulation provides a perspective for understanding the core principles underlying parameter learning within convex optimization framework. However, the learning objective outlined in (16) proves to be *challenging to optimize* owing to the indeterminate initial points across various intervals. In addition, its computation necessitates a "backtracking" operation, an impractical approach for inferential sensor tasks given their inherent causality. To highlight the issue, we can consider the model at timestamp t: At time t, y_t is predicted using the latent variable z_t. Consequently, according to (16), z_t depends on future values $y_{t+1:\mathrm{T}}$. However, this is impossible to achieve without a "time machine." This inconsistency poses a significant challenge for practical implementation. Fortunately, this problem can be solved based on "one-step lookahead minimization" method according to [20]. And thus, we propose the following proposition to derive an upper bound for (16):

Proposition 3: The objective function defined in (12) have the following upper bound:

$$\begin{aligned}&\min_{\theta,\nu}\sum_{t=1}^{\mathrm{T}}\left\{\mathbb{E}_{\mathbb{Q}(z)}\left[-\log p_\theta(y_t|z_t,u_{1:t}) + \int_{t-1}^{t}\frac{1}{2}\|\nu\|^2\mathrm{d}\tau\right]\right\}\\ &\leq\\ &\sum_{t=1}^{\mathrm{T}}\min_{\theta,\nu}\left\{\mathbb{E}_{\mathbb{Q}(z)}\left[-\log p_\theta(y_t|z_t,u_{1:t}) + \int_{t-1}^{t}\frac{1}{2}\|\nu\|^2\mathrm{d}\tau\right]\right\}.\end{aligned} \tag{17}$$

Proof: The proof is given in Section S.II.C of Supplementary material. □

It is noteworthy that, to the best of our knowledge, the majority of NDPLVMs tailored for inferential sensor tasks, as exemplified by the works referenced in [13], [14], [21], [22], employ the right-hand side of (17) as their learning objective. This is the case even though they may not have explicitly derived the right-hand side in their respective studies [13], [14], [21], [22]. A key criterion for this determination is the absence of "backtracking operations" in solving the objective function, indicating that the right-hand side of (17) is being used as the learning objective. However, the inequality defined in (17) has not been thoroughly explored in previous works [13], [14], [21], [22].

C. Inference Network's Input Selection by Solution Structure Investigation

Based on (17) in Section III-B, we can conduct concerning analysis: the right-hand-side of (17) has two sets of variables to be optimized, viz. the model parameter θ and the control policy ν. It should be pointed out that this problem involves an integral term $\int$ with respect to ν in time domain, which belongs to an OC problem as per reference [23]. *Based on this, we named our model OC-NDPLVMs. Moreover, it should be pointed out that the OC problem can be regarded as an infinite-dimensional optimization problem [24].*

Comparing (17) with (9), we observe that the θ in (17) corresponds to the w in (9); the ν in (17) corresponds to the v in (9). The log-likelihood term $\log p_\theta(y_t|z_t,u_{1:t})$ and control policy terms $\int_{t-1}^{t}\|\nu\|^2\mathrm{d}\tau$ are separable. Besides, since the latent space satisfies the differential equation in NDPLVMs, the Lagrangian multiplier λ and the quadratic penalty coefficient ρ are explicitly ignored in the remainder of this manuscript. *This observation motivates us to solve the parameter learning of NDPLVMs given in (17) through the lens of ADMM.* Since the model parameter θ can be optimized through automatic-differentiation-based DL backends such as PyTorch [15] by stochastic gradient descent-based optimizers, the minimization subproblem concerns with

θ will not be discussed. The rest of this section will discuss the minimization subproblem (the OC problem) concerns with ν.

Based on Proposition 3, optimizing the *global* ELBO defined in (16) can be converted to optimizing the *local* ELBO within interval $[t, t+1]$ for $t \in [0, \mathrm{T})$. And thus, the following section will focus on the optimization of *local* ELBO within interval $[t, t+1]$. We will take the OC signal derivation between interval $[t, t+1]$ as an example to illustrate this subproblem.

Based on the Gaussian assumption on the observation data, the ELBO between interval $[t, t+1]$ can be expanded as

$$\begin{aligned}&\mathbb{E}_{\mathbb{Q}(z)}\left[-\log p_\theta(y_t|z_t, u_{1:t}) + \int_{t-1}^{t} \frac{1}{2}\|\nu\|^2 \mathrm{d}\tau\right] \\ &= \mathbb{E}_{\mathbb{Q}(z)}\left[\frac{1}{2}(y_t - \mu_t^y)^\top (y_t - \mu_t^y) + \int_{t-1}^{t} \frac{1}{2}\|\nu\|^2 \mathrm{d}\tau\right] \end{aligned} \tag{18}$$

where we parameterize $p(y|z_t, u_{1:t})$ as $\mathcal{N}(\mu_t^y, \mathcal{I})$ as per [13] and [21], and μ_t^y is generative by generative network $g_\theta(\cdot)$ according to $\mu_t^y = g(\mu_t^z)$. It is noteworthy that this approximation can be understood in terms of conceptualizing $p_\theta(y|z_t, u_{1:t})$ as a Dirac delta distribution (i.e., $\delta(y|z_t, u_{1:t})$) and then approximating it with a normal distribution $\mathcal{N}(y - \mu_t^y, \mathcal{I})$. This is a common assumption in the context of kernel density estimation, when the bandwidth is set to 1 according to [25]. Consequently, we conduct the approximation in the last line. Besides, the mean value μ_t^y is obtained via neural network denoted as g_θ

$$\mu_t^y = \mathbb{E}[g_\theta(z_t)] \approx g_\theta(\mu_t^z). \tag{19}$$

Based on the approximation operations, the following OC problem can be formulated to obtain the optimal ν (denoted as ν^*)

$$\begin{aligned}&\min_{\nu} \quad (y_t - \mu_t^y)^\top (y_t - \mu_t^y) + \int_{t-1}^{t} \frac{1}{2}\|\nu\|^2 \mathrm{d}\tau \\ &\text{s.t.} \quad \mathrm{dz} = \mathrm{f}_\theta(\mathrm{z}, \mathrm{u})\mathrm{dt} + \mathrm{L}\nu\mathrm{dt}\end{aligned} \tag{20}$$

where the diffusion term is omitted since the model training mainly concentrates on the mean. And thus, the following Hamiltonian equation can be derived to solve the constrained OC problem according to the Pontryagin's maximum principle [23]:

$$\mathcal{H} = \frac{1}{2}\|\nu\|^2 + \lambda^\top (f_\theta(z, u) + L\nu) \tag{21}$$

where λ is Lagrangian multiplier. According to the sufficient condition for OC extreme value, the following equation can be obtained:

$$\frac{\partial \mathcal{H}}{\partial \nu} = \nu + L\lambda = 0. \tag{22}$$

Note that, the second-order derivative of the Hamiltonian function is a positive-definite matrix (identity matrix $\mathcal{I}$): $\nabla_\nu^2 \mathcal{H} = \mathcal{I} \succ 0$, which indicates that the generalized Legendre–Clebsch necessary condition can be satisfied [23]. As such, the extreme value obtained by the OC signal ν^* according to (21) is the minimum value.

According to the OC principle, the mean μ_t^z, and the co-state λ satisfy the following differential equations:

$$\begin{cases} \frac{\mathrm{d}\mu_t^z}{\mathrm{d}t} = \mathbb{E}(\frac{\partial \mathcal{H}}{\partial \lambda}) = f_\theta(\mu_t^z, u) + L\nu \\ \frac{\mathrm{d}\lambda}{\mathrm{d}t} = -\frac{\partial \mathcal{H}}{\partial z} = -\frac{\partial f_\theta(z,u)}{\partial z} \end{cases}. \tag{23}$$

The corresponding boundary condition for the equations are give as follows:

$$\begin{cases} \mu_{t-1}^z = \mathbb{E}(z_{t-1}) \\ \lambda_t = 2(\mu_t^y - y_t)^\top \frac{\partial g(z_t)}{\partial z_t}\big|_{\mu_t^z} \end{cases}. \tag{24}$$

By observing (21) to (24), the optimal ν mainly concerned with the observation data $y^\top$ at final time t, stochastic variable z_t (parameterized via mean μ_t^z and covariance Σ_t). It should be pointed out that, the computation of the Jacobian will result in a higher model training time. However, we know the OC is the function of y_t, μ_t^z, and Σ_t^z. On this basis, revist (15), we can simulate a neural network denoted as q with parameter ϕ as Section II-A mentioned as follows:

$$\nu^* = q_\phi(y_t, z_t) = q_\phi(y_t, \mu_t, \Sigma_t). \tag{25}$$

Drawing on the solution of the OC subproblem delineated in (25), we have effectively addressed the issue of "inaccurate inference of latent space" by carefully selecting the input for the inference network. Moreover, our analysis extends beyond the conventional scope of NDPLVMs, which traditionally consider u as the sole inference network input. *Our findings reveal that the key to achieving accurate inference of z lies in the label y, offering a significant insight that refines the conventional understanding in this domain.*

D. Measure Change and Likelihood Term Approximation

Applying the Girsanov theorem, the posterior process under measure $\mathbb{Q}$ can be derived via the prior process under measure $\mathbb{P}$ as follows:

$$\mathrm{d}z^{\mathbb{Q}} = \exp\left(\int_{t-1}^{t} -\frac{1}{2}\|\nu\|^2 \mathrm{d}\tau\right) \mathrm{d}z^{\mathbb{P}}. \tag{26}$$

Since at start point $t-1$ ($t \in [0, \mathrm{T})$), the probabilistic density of z under measure $\mathbb{Q}$ and $\mathbb{P}$ are same. The probabilistic density of z under measure $\mathbb{Q}$ at time t can be obtained via the probabilistic density of z at measure $\mathbb{P}$. Supposed the probabilistic density of z at time t is

$$z_t \sim \mathcal{N}\left(\mu_t^{z,\mathbb{P}}, \Sigma_t^{z,\mathbb{P}}\right) \tag{27}$$

where superscript $z, \mathbb{P}$ indicates that z_t is in measure $\mathbb{P}$, μ and Σ are the mean and covariance of normal distribution, respectively. Suppose the solution of the integral of Radon–Nikodym derivative ν is

$$\nu \sim \mathcal{N}\left(\mu_t^\nu, \Sigma_t^\nu\right). \tag{28}$$

Then, the probabilistic density of z under measure $\mathbb{Q}$ can be derived as follows at time t:

$$\begin{aligned}z_t^{\mathbb{Q}} \sim \mathcal{N}\Bigg(&\left(\Sigma_t^\nu + \Sigma_t^{z,\mathbb{P}}\right)^{-1}\left(\Sigma_t^\nu \mu_t^\nu + \Sigma_t^{z,\mathbb{P}} \mu_t^{z,\mathbb{P}}\right) \\ &\left(\Sigma_t^\nu + \Sigma_t^{z,\mathbb{P}}\right)^{-1}\left(\Sigma_t^\nu \Sigma_t^{z,\mathbb{P}}\right)\Bigg)\end{aligned} \tag{29}$$

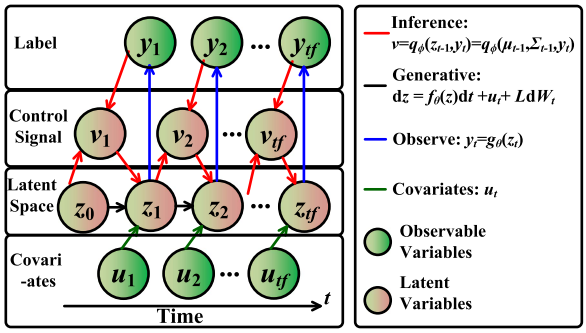


Fig. 2. Overall architecture of OC-NDPLVM.

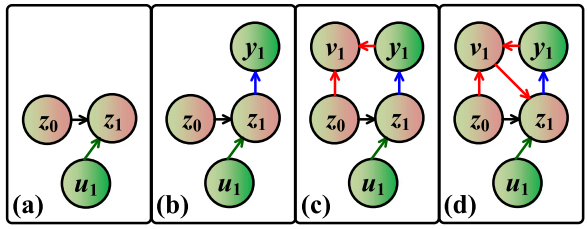


Fig. 3. Illustration of OC-NDPLVM inference between $t=0$ and $t=1$: (a) prior process transition. (b) Label prediction, (c) OC signal simulation. (d) Posterior process correction.

which can be regarded as Schur complement operation [26].

E. Derivation of Moment Expressions

In this section, the moment expressions approximation operations are proposed in detail to answer the question 'model implementation within DL backends."

1) Moment Expressions for Transition: According to [17], the distribution of an Itô process in time-axis can be represented by normal distribution denoted as $\mathcal{N}$, with a mean of μ and a covariance of Σ. Based on this, the following proposition for mean and covariance are given:

Proposition 4: The mean and covariance equations between z_t and z_{t+1} can be given in (30) and (31), respectively,

$$\frac{\mathrm{d}\mu}{\mathrm{d}t} = f(\mu, t) \tag{30}$$

$$\frac{\mathrm{d}\Sigma}{\mathrm{d}t} = \Sigma \left[\frac{\partial f(z,t)}{\partial z}\Big|_{z=\mu}\right]^\top + \Sigma^\top \left[\frac{\partial f(z,t)}{\partial z}\Big|_{z=\mu}\right] + L(\mu,t)QL^\top(\mu,t). \tag{31}$$

Proof: The proof is provided in Section S.II.D of Supplementary Material. □

2) Moment Expressions for Loss Function: Based on previous section, we further derive the moment expressions of $\mathbb{E}_{\mathbb{Q}}[\|y_t - g(z_t, u_{1:t})\|^2]$ for *inferential sensor task* by following proposition:

Proposition 5: The moment expressions of $\mathbb{E}_{\mathbb{Q}}[\|y_t - g(z_t, u_{1:t})\|^2]$ can be approximated as follows:

$$\begin{aligned}&\mathbb{E}_{\mathbb{Q}}\left[\|y_t - g(z_t, u_{1:t})\|^2\right] \\ &\approx \mathbb{E}_{\epsilon\sim\mathcal{N}(0,\mathcal{I})}\left[\mathcal{L}(\mu_t^z) + \left(\frac{\partial\mathcal{L}}{\partial z}\Big|_{z=\mu_t^z}\right)(\sigma_t^z)\left(\frac{\partial\mathcal{L}}{\partial z}\Big|_{z=\mu_t^z}\right)^\top \times\epsilon\right] \\ &:= \mathcal{L}(\mu_t^z).\end{aligned} \tag{32}$$

Proof: The proof is given in Section S.II.E of Supplementary Material. □

F. Model Overall Structure

Based on the abovementioned sections, the model architecture is summarized in Fig. 2. It can be seen that the model consists of two different colored nodes, where the green node represents the observable variables, and the orange node represents the latent variables. Note that the control policy is simulated by the inference network, where we assume that the neural network can infer the OC signal based on its input as assumed in (25).

To better understand the inference procedure of OC-NDPLVM, we visualize the inference procedure between $t=0$ and $t=1$ in Fig. 3 based on Fig. 2. According to Fig. 3, the model goes through the following steps:

1) *Prior Process Transition [Black Line, Fig. 3(a)]:* Estimate the latent state z_t as $\hat{z}_t$ via last latent state z_{t-1} and covariate u_t according to (13)
2) *Label Prediction [Blue Line, Fig. 3(b)]:* Predict the label/observation value y_t as $\hat{y}_t$ via $\hat{z}_t$ according to (19).
3) *Simulate OC Signal [Red Line, Fig. 3(c)]:* Simulate the control signal likelihood ratio (26) via z_{t-1} real label y_t/predicted label $\hat{y}_t$ according to (25)
4) *Posterior Process Correction [Green Line, Fig. 3(d)]:* Compute z_t via correcting $\hat{z}_t$ according to (29)

By repeating the steps, the model can infer the observation sequence $\vec{y}$ with the covariate sequence $\vec{u}$.

Furthermore, if we detract the control signal part in the middle of Fig. 2, the model degrades to current NDPLVMs for inferential sensor tasks (detailed comparisons are given in Section S.III.D of Supplementary Material due to page limit). On this basis, the corresponding algorithms for model inference and training & testing are summarized in Section S.III.C of Supplementary Material due to page limit.

G. Theoretical Analysis of Learning Objective Convergence

In this section, we want to analyze the convergence of the proposed ADMM-based algorithm to make our work more complete. The following proposition is given for the convergence of proposed ADMM-based algorithm:

Proposition 6: The convergence of the optimization procedure in (11) can be guaranteed, given that: 1), the inference network q_ϕ can simulate the OC signal ν^*; and 2) the learning rate in the stochastic gradient optimizer ensures the reduction of the loss function $\mathcal{L}$.

Proof: The proof is given in Section S.IV.A of Supplementary Material. □

Assumption 1) is a widely admitted setting in the context of AVI-based models as we mentioned in Section II-A, and Assumption 2) can be realized easily when the learning rate is low enough (analysis about this condition is given in Section S.IV.B of Supplementary Material due to page limit).

IV. Experimental Results and Discussions

In this section, we devise experiments on two industrial inferential sensor datasets to verify the superiority of OC-NDPLVM and answer the research questions from the perspective of theory as follows:

1) *Performance: Does OC-NDPLVMs work?* Section IV-B evaluates OC-NDPLVM's performance against a variety of baseline methods using two datasets from real industrial scenario, thereby establishing a foundational understanding of its operational efficiency.
2) *Convergence: Does it converge?* To backup our theoretical analysis in Section III-G. Section IV-C analyzes the iteration curve for two datasets along epoch, thereby illustrating the convergence trajectory at the epoch scale.
3) *Gains: Why does OC-NDPLVM work?* Section IV-D deconstructs OC-NDPLVM to discern the sources of its performance gain.

On this basis, the following research questions are also studied from the perspective of practice as follows (due to page limit, these results are provided in Supplementary Material):

1) *Complexity: What's the computational complexity?* Section S.VII.A (Supplementary Material) compares the computational complexity spatially and temporally to prove the feasibility of OC-NDPLVM from the perspective of deployment.
2) *Sensitivity: Is it sensitive to key hyperparameters?* Section S.VII.B (Supplementary Material) elucidates the impact of different hyperparameters on the prediction accuracy, analyzing the system's responsiveness to parameter alterations.

A. Experimental Settings

1) Datasets: We select two datasets namely debutanizer column (DC) and catalytic shift conversion (CSC). More details about these two datasets are given in Section S.V of Supplementary Material.

2) Baseline Models: We compare the proposed OC-NDPLVM with the following baseline models:

1) *Recurrent Network-Based methods:* auto-regressive temporal convolution network (AR-TCN) [27] and dual-attention LSTM (DA-LSTM) [28].
2) *PLVM-Based Methods:* NPLVR [11], DPLVM [7], probabilistic discriminative time-series model (PDTM) [22] and DBPSFA [13].
3) *Self-Attentive-Based Methods (nonauto-regressive structure):* LogSparse Transformer (LogTrans) [29] and Informer (state-of-the-art, 2021) [30].
4) *Mixture Model-Based Methods:* Dirichlet process mixture model (DPMM) and DMVAER [14].

The reasons for choosing these models, computational resource, and other experimental details for experiments are also provided in Section S.VI of Supplementary Material.

3) Evaluation Metrics: The root mean squared error (RMSE) and mean absolute error (MAE) are adopted as evaluation metrics. Their expressions are given as follows:

$$\mathrm{RMSE} = \frac{1}{\mathrm{N}}\frac{1}{\mathrm{H}}\sum_{n=1}^{\mathrm{N}}\sum_{h=1}^{\mathrm{N}}\sqrt{(\hat{y}_{h,n}-y_{h,n})^2} \tag{33}$$

$$\mathrm{MAE} = \frac{1}{\mathrm{N}}\frac{1}{\mathrm{H}}\sum_{n=1}^{\mathrm{N}}\sum_{h=1}^{\mathrm{H}}|\hat{y}_{h,n}-y_{h,n}| \tag{34}$$

where $\hat{y}$ represents the predicted values, y denotes the actual values, H is the length of the prediction horizon, and N signifies the total number of evaluation instances. For metrics such as RMSE and MAE, a lower value correlates with a more accurate model. Our evaluation metrics operate on a rolling basis along the time axis with a length-H prediction horizon. Consequently, in line with references [19], [30], we do not provide the conventional "prediction-real results" comparison graphs in our experimental results.

B. Overall Performance

In this section, question "*Does OC-NDPLVMs work?*" about performance comparison is answered. The comparison results for the OC-NDPLVM and other baseline models on the DC and CSC datasets are reported in Table I. Notably, all experiments are repeated at least three times under six different random seeds. The following observations can be obtained from Table I.

1) For the DC dataset, the RMSE for $H = 2, 3, 4, 5$ are $39.73\% \sim 92.54\%$, $33.02\% \sim 89.46\%$, $17.97\% \sim 86.29\%$, and $8.14\% \sim 83.08\%$ lower than those of the baseline models, respectively; the MAE for $H = 2, 3, 4, 5$ are $42.68\% \sim 93.00\%$, $37.96\% \sim 90.40\%$, $23.14\% \sim 87.71\%$, and $13.72\% \sim 84.97\%$ lower than those of the baseline models, respectively.
2) For the CSC dataset, the RMSE for $H = 2, 3, 4, 5$ are $0.45\% \sim 80.49\%$, $3.43\% \sim 79.90\%$, $0.79\% \sim 78.86\%$, and $0.17\% \sim 77.92\%$ lower than those of the baseline models, respectively; the MAE for $H = 2, 3, 4, 5$ are $0.65\% \sim 81.55\%$, $4.06\% \sim 81.32\%$, $1.21\% \sim 80.45\%$, and $0.28\% \sim 79.21\%$ lower than those of the baseline models, respectively.
3) The performance gains of OC-NDPLVM against most baselines are significant, as is evidenced by the p-value$<$ 0.05 over the paired samples t-test.
4) The recurrent network-based methods and self-attentive-based methods have better performance than PLVM-based methods and mixture model-based methods.
5) Linear PLVMs like DPLVM and DPMM have worse performance than the nonlinear version like NPLVR and DMVAER.
6) When the prediction window increases, the performance degradation of self-attention models is smaller than that of NDPLVMs and Recurrent models.

TABLE I
MODEL PERFORMANCE ON INFERENTIAL SENSOR TASK

Dataset	Model	H = 2		H = 3		H = 4		H = 5	
		RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE
DC	AR-TCN	0.0234†	0.0230†	0.0298†	0.0292†	0.0367†	0.0356†	0.0442†	0.0427†
	DA-LSTM	0.0341†	0.0323†	0.0393†	0.0367†	0.0379†	0.0357†	0.0502†	0.0480†
	NPLVR	0.1393†	0.1390†	0.1399†	0.1392†	0.1401†	0.1389†	0.1405†	0.1389†
	DPLVM	0.1722†	0.1714†	0.1732†	0.1718†	0.1741†	0.1721†	0.175†	0.1723†
	PDTM	0.1512†	0.1499†	0.1433†	0.1419†	0.1427†	0.1411†	0.1413†	0.1393†
	DBPSFA	0.1424†	0.1412†	0.1413†	0.1394†	0.1420†	0.1400†	0.1445†	0.1415†
	LogTrans	0.0326†	0.0319†	0.0344†	0.0333†	0.0346†	0.0328†	0.0409†	0.0387†
	Informer	0.0243†	0.0238†	0.0314†	0.0305†	0.0318†	0.0302†	0.0351	0.0329†
	DPMM	0.1896†	0.1893†	0.1900†	0.1894†	0.1904†	0.1894†	0.1909†	0.1895†
	DMVAER	0.1394†	0.1389†	0.1395†	0.1387†	0.1401†	0.1388†	0.1407†	0.1388†
	OC-NDPLVM	**0.0141**	**0.0132**	**0.0200**	**0.0181**	**0.0260**	**0.0232**	**0.0322**	**0.0284**
CSC	AR-TCN	0.1036	0.0945	0.1119	0.0991†	0.1157	0.1007	0.1201	0.1037
	DA-LSTM	0.1793†	0.1631†	0.1965†	0.1713†	0.2153†	0.1836†	0.2268†	0.1910†
	NPLVR	0.5340†	0.5150†	0.5411†	0.5129†	0.5472†	0.5136†	0.5525†	0.5152†
	DPLVM	0.6046†	0.5843†	0.6162†	0.5856†	0.6230†	0.5861†	0.6277†	0.5864†
	PDTM	0.1309†	0.1204†	0.1460†	0.1298†	0.1595†	0.1401†	0.1563†	0.135†
	DBPSFA	0.1450†	0.1329†	0.1648†	0.1441†	0.1818†	0.1545†	0.1948†	0.1612†
	LogTrans	0.1065	0.0974	0.1136	0.1008	0.1186	0.1036	**0.1197**	**0.1033**
	Informer	0.1071†	0.0978†	0.1152	0.1023	0.1184	0.1033	0.1218	0.1054
	DPMM	0.6033†	0.5899†	0.6095†	0.5892†	0.6132†	0.5884†	0.6158†	0.5876†
	DMVAER	0.5296†	0.5103†	0.5381†	0.5098†	0.5440†	0.5102†	0.5472†	0.5098†
	OC-NDPLVM	**0.1031**	**0.0939**	**0.1081**	**0.0951**	**0.1148**	**0.0995**	0.1199	0.1034

† marks the variants that OC-NDPLVM outperforms significantly at p-value < 0.05 over paired samples t-test. *Bolded* results indicate the best in each metric. Underlined results indicate the second best in each metric.

Observations 1) and 2) indicate that the proposed OC-NDPLVM outperforms other baseline models. Observation 3) indicates that it is sufficient to say the OC-NDPLVM is better than other baseline models for most of the scenarios. Interestingly, Observation 4) reflects that NPLVMs still has a lot of room for improvement in the inferential sensor modeling task, and further demonstrates the necessity of implementing the inference network input selection and moment expression in this article from practice. Observation 5) emphasizes the importance of introducing DL architectures for NDPLVMs performance improvement. Observation 6) indicates that the models with auto-regressive structure may suffer from gradient vanishing with the increase of forecasting horizon, while the self-attention models can alleviate this issue thanks to their nonauto-regressive structure.

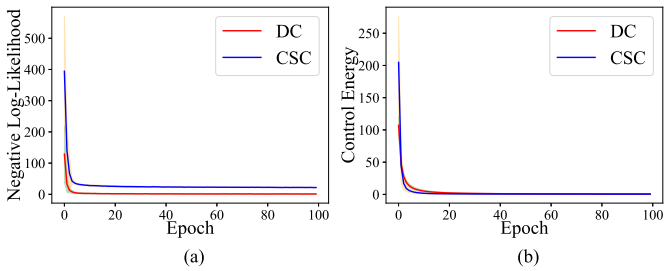


Fig. 4. Convergence analysis of the DC and CSC datasets for H = 4. (a) Negative log-likelihood ($-\frac{1}{N}\sum_{n=1}^{N}\sum_{t=1}^{T+H}\mathbb{E}_{\mathbb{Q}}[\log p(y_{t,n}|z_{t,n},u_{t,n})]$). (b) Control energy term ($\frac{1}{N}\sum_{t=1}^{T+H}\mathbb{E}_{\mathbb{Q}}[\int_{t-1}^{t}\|\nu_\tau\|^2 d\tau]$) along iteration process. The shaded area indicates the ± 1.5 standard deviation uncertainty interval.

C. Convergence of the ADMM Framework

In this section, question "*Does it converge?*" about convergence analysis is answered to practically demonstrate the ADMM optimization framework's convergence. Fig. 4(a) and (b) proposes the likelihood term and the control energy term along the iteration process. From Fig. 4, it can be seen that both the likelihood term and energy term decrease with the increase of the iteration epoch. The decrease of the likelihood term along the training epoch in Fig. 4(a) indicates that the latent state can represent the label pattern in the training process. Consequently, the decrease of the control energy with the increase of the training epoch in Fig. 4(b) indicates that the dependence on the label information decreases in the training process. Specifically, the likelihood and control energy terms tend to be unchanged after 10 epochs. This phenomenon reflects that the optimization strategy built upon ADMM framework has a fast convergence rate. Furthermore, it can be observed that the energy of the control signal tends to approach zero after a few epochs. This suggests that the approximation utilized in the assumption in the proof of Proposition 5 (Section S.II.E, Supplementary Material) is reasonable.

D. Ablation Study

In this section, we conduct an ablation study focusing on two key components: the OC-based inference network structure (abbreviated as "OC," where the network does not incorporate label information y if OC is disabled), and the model implementation based on moment expressions (abbreviated as "ME," where the model follows the implementation approach of previous works [13], [21], [22] with the reparameterization trick and

TABLE II
ABLATION STUDY RESULTS

H	OC	ME	DC		CSC	
			RMSE	MAE	RMSE	MAE
2	✗	✓	0.1668(1083%)†	0.1589(1103%)†	0.3495(239.0%)†	0.3201(240.9%)†
	✓	✗	0.0438(210.9%)†	0.0420(218.3%)†	0.1287(24.90%)†	0.1186(26.30%)†
	✗	✗	0.1416(903.9%)†	0.1397(957.8%)†	0.4293(316.4%)†	0.4008(326.8%)†
	✓	✓	0.0141(-)	0.0132(-)	0.1031(-)	0.0939(-)
3	✗	✓	0.1895(848.4%)†	0.1716(846.7%)†	0.4068(276.5%)†	0.3631(281.8%)†
	✓	✗	0.0498(149.0%)†	0.0475(162.2%)†	0.1456(34.80%)†	0.1310(37.80%)†
	✗	✗	0.1428(615.1%)†	0.1402(673.5%)†	0.4674(332.6%)†	0.4257(347.6%)†
	✓	✓	0.0200(-)	0.0181(-)	0.1081(-)	0.0951(-)
4	✗	✓	0.1932(641.7%)†	0.1724(642.5%)†	0.4547(296.2%)†	0.4016(303.8%)†
	✓	✗	0.0756(190.2%)†	0.0724(211.9%)†	0.1364(18.80%)†	0.1199(20.60%)†
	✗	✗	0.1442(453.8%)†	0.1411(507.7%)†	0.5225(355.3%)†	0.4757(378.3%)†
	✓	✓	0.0260(-)	0.0232(-)	0.1148(-)	0.0995(-)
5	✗	✓	0.2139(563.8%)†	0.1834(545.6%)†	0.5164(330.8%)†	0.4530(338.2%)†
	✓	✗	0.0753(133.7%)†	0.0715(151.8%)†	0.1421(18.50%)†	0.1235(19.40%)†
	✗	✗	0.1449(349.7%)†	0.1413(397.5%)†	0.5553(363.3%)†	0.4991(382.7%)†
	✓	✓	0.0322(-)	0.0284(-)	0.1199(-)	0.1034(-)

We present the results in the form of: mean value (increasing percentage), † marks the variants that OC-NDPLVM outperforms significantly at p-value < 0.05 over paired samples t-test.

conducts sampling between different time intervals if ME is disabled).

The following observations are summarized from Table II.

1) *Incorporating OC Improves Performance:* The results reveal a significant drop in model performance when the inference network's input does not incorporate OC-based structuring, particularly noted in the first and third rows of each horizon. This aligns with observations from Table I, where NDPLVMs generally underperform compared to recurrent and self-attentive models. This underscores the critical role of the OC-based approach for input selection, as proposed in Section III-C.
2) *Importance of ME:* Disabling ME, as shown in the second and third rows of each horizon, also leads to reduced performance. This emphasizes that proper model implementation, following the ME framework, is crucial when deploying NDPLVMs for inferential sensor tasks.

These points affirm the necessity of including label information in the inference network's input and adhering to the model implementation strategy provided by moment expressions to ensure robust performance of NDPLVMs in inferential sensors.

V. CONCLUSION

In this work, to answer two fundamental but essential problems, namely "inaccurate inference of latent space" and "model implementation within DL backends" in the NDPLVMs, we first proposed our OC-NDPLVM and its loss function from SDE theory, conducted detailed analysis of model training algorithm, derived moment expressions, and summarized the model overall architecture. In this procedure, we redesigned the input of inference network by solving the OC subproblem and simplified the model implementation by obtaining the moment expressions. Finally, to empirically validate the proposed method's effectiveness, we conduct various inferential sensor experiments on two industrial datasets.

Future directions can be focused on the introducing other integrators to alleviate the stiffness issue of ODEs [31] and improve model training efficiency with other algorithms like adjoint sensitivity method [23]. Besides, other distribution transformation methods like the sequential Monte-Carlo method [32] can also be adopted to estimate the likelihood function more accurately. Furthermore, this task primarily focuses on supervised learning. It is important to note that semisupervised scenarios, where some covariates may not have labels, can also arise in inferential sensor tasks. To address this, exploring how to extend the proposed approach to semisupervised context by leveraging transfer learning technique [33] or redesigning network architecture [34], and enhancing inferential sensor modeling using data with limited or no labels, is an intriguing direction for future research. Finally, industrial processes may continue to be affected by various types of data noise [35], making the investigation of NDPLVM structures that are robust to such noise a significant research direction.

REFERENCES

[1] F. Qian, Y. Jin, S. J. Qin, and K. Sundmacher, "Guest editorial special issue on deep integration of artificial intelligence and data science for process manufacturing," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 8, pp. 3294–3295, Aug. 2021.
[2] X. Kong, X. Jiang, B. Zhang, J. Yuan, and Z. Ge, "Latent variable models in the era of industrial Big Data: Extension and beyond," *Annu. Rev. Control*, vol. 54, pp. 167–199, 2022.
[3] S. Yu, K. Yu, V. Tresp, H.-P. Kriegel, and M. Wu, "Supervised probabilistic principal component analysis," in *Proc. 12th ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2006, pp. 464–473.
[4] M. J. Beal, "Variational algorithms for approximate Bayesian inference," Ph.D. dissertation, Univ. College London, London, U.K., May 2003.
[5] L. Yao and Z. Ge, "Big data quality prediction in the process industry: A distributed parallel modeling framework," *J. Process Control*, vol. 68, pp. 1–13, 2018.
[6] J. Wang, W. Shao, X. Zhang, and Z. Song, "Dynamic variational Bayesian student's t mixture regression with hidden variables propagation for industrial inferential sensor development," *IEEE Trans. Ind. Informat.*, vol. 17, no. 8, pp. 5314–5324, Aug. 2021.
[7] Z. Ge and X. Chen, "Dynamic probabilistic latent variable model for process data modeling and regression application," *IEEE Trans. Control Syst. Technol.*, vol. 27, no. 1, pp. 323–331, Jan. 2019.
[8] C. Shang, B. Huang, F. Yang, and D. Huang, "Probabilistic slow feature analysis-based representation learning from massive process data for soft sensor modeling," *AIChE J.*, vol. 61, no. 12, pp. 4126–4139, 2015.
[9] Y. Ma and B. Huang, "Extracting dynamic features with switching models for process data analytics and application in soft sensing," *AIChE J.*, vol. 64, no. 6, pp. 2037–2051, 2018.
[10] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," in *Proc. Int. Conf. Learn. Representations*, 2014, pp. 1–8.
[11] B. Shen, L. Yao, and Z. Ge, "Nonlinear probabilistic latent variable regression models for soft sensor application: From shallow to deep structure," *Control Eng. Pract.*, vol. 94, 2020, Art. no. 104198.
[12] L. Girin, S. Leglaive, X. Bie, J. Diard, T. Hueber, and X. Alameda-Pineda, *Dynamical Variational Autoencoders: A Comprehensive Review*. Boston, MA, USA: Now, 2021.
[13] C. Jiang et al., "Deep Bayesian slow feature extraction with application to industrial inferential modeling," *IEEE Trans. Ind. Informat.*, vol. 19, no. 1, pp. 40–51, Jan. 2023.
[14] L. Yao, B. Shen, L. Cui, J. Zheng, and Z. Ge, "Semi-supervised deep dynamic probabilistic latent variable model for multimode process soft sensor application," *IEEE Trans. Ind. Informat.*, vol. 19, no. 4, pp. 6056–6068, Apr. 2023.
[15] A. Paszke et al., "Pytorch: An imperative style, high-performance deep learning library," in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, 2019, vol. 32, pp. 1–12.
[16] A. Ganguly, S. Jain, and U. Watchareeruetai, "Amortized variational inference: A systematic review," *J. Artif. Intell. Res.*, vol. 78, pp. 167–215, 2023.

[17] S. Särkkä and A. Solin, *Applied Stochastic Differential Equations*, vol. 10. Cambridge, U.K.: Cambridge Univ. Press, 2019.
[18] S. Boyd et al. *Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers*, vol. 3, no. 1. Boston, MA, USA: Now, 2011.
[19] Q. Wen et al., "Transformers in time series: A survey," in *Proc. 32nd Int. Joint Conf. Artif. Intell.*, 2023, pp. 6778–6786, doi: 10.24963/ijcai.2023/759.
[20] D. Bertsekas, *A Course in Reinforcement Learning*. Belmont, MA, USA: Athena Scientific, 2023.
[21] B. Shen and Z. Ge, "Supervised nonlinear dynamic system for soft sensor application aided by variational auto-encoder," *IEEE Trans. Instrum. Meas.*, vol. 69, no. 9, pp. 6132–6142, Sep. 2020.
[22] Y. Lu, X. Peng, D. Yang, C. Jiang, and W. Zhong, "The probabilistic discriminative time-series model with latent variables and its application to industrial chemical process modeling," *Chem. Eng. J.*, vol. 423, 2021, Art. no. 130298.
[23] L. S. Pontryagin, E. Mishchenko, V. Boltyanskii, and R. Gamkrelidze, *The Mathematical Theory of Optimal Processes*. Evanston, IL, USA: Routledge, 1962.
[24] H. O. Fattorini, *Infinite Dimensional Optimization and Control Theory*, vol. 54. Cambridge, U.K.: Cambridge Univ. Press, 1999.
[25] Y.-C. Chen, "A tutorial on kernel density estimation and recent advances," *Biostatist. Epidemiol.*, vol. 1, no. 1, pp. 161–187, 2017.
[26] C. E. Rasmussen, "Gaussian processes in machine learning," in *Summer School on Machine Learning*. Berlin, Germany: Springer, 2003, pp. 63–71.
[27] X. Yuan, S. Qi, Y. Wang, K. Wang, C. Yang, and L. Ye, "Quality variable prediction for nonlinear dynamic industrial processes based on temporal convolutional networks," *IEEE Sensors J.*, vol. 21, no. 18, pp. 20493–20503, Sep. 2021.
[28] L. Feng, C. Zhao, and Y. Sun, "Dual attention-based encoder–decoder: A customized sequence-to-sequence learning for soft sensor development," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 8, pp. 3306–3317, Aug. 2021.
[29] S. Li et al., "Enhancing the locality and breaking the memory bottleneck of transformer on time series forecasting," in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, 2019, vol. 32, pp. 5243–5253.
[30] H. Zhou et al., "Informer: Beyond efficient transformer for long sequence time-series forecasting," in *Proc. 35th AAAI Conf. Artif. Intell., AAAI 2021, Virtual Conf.*, 2021, vol. 35, pp. 11106–11115.
[31] J. C. Butcher, *Numerical Methods for Ordinary Differential Equations*. Hoboken, NJ, USA: Wiley, 2016.
[32] W. Sun, W. Xiong, H. Chen, R. Chiplunkar, and B. Huang, "A novel CVAE-based sequential monte carlo framework for dynamic soft sensor applications," *IEEE Trans. Ind. Informat.*, vol. 20, no. 3, pp. 3789–3800, Mar. 2024.
[33] D. Yang, X. Peng, C. Jiang, X. Wu, S. X. Ding, and W. Zhong, "Transferable deep slow feature network with target feature attention for few-shot time-series prediction," *IEEE Trans. Ind. Informat.*, vol. 20, no. 5, pp. 7292–7302, May 2024.
[34] R. Xie, N. M. Jan, K. Hao, L. Chen, and B. Huang, "Supervised variational autoencoders for soft sensor modeling with missing data," *IEEE Trans. Ind. Informat.*, vol. 16, no. 4, pp. 2820–2828, Apr. 2020.
[35] C. Xu, S. Zhao, Y. Ma, B. Huang, F. Liu, and X. Luan, "Sensor fault estimation in a probabilistic framework for industrial processes and its applications," *IEEE Trans. Ind. Informat.*, vol. 18, no. 1, pp. 387–396, Jan. 2022.

Zhichao Chen received the B.Eng. degree in chemical engineering and technology from School of Chemical Engineering and Technology, Sun Yat-sen University, Zhuhai, China, in 2020. He is currently working toward the Ph.D. degree in control science and engineering with the State Key Laboratory of Industrial Control Technology, College of Control Science and Engineering, Zhejiang University, Hangzhou, China.

His research interests include process data analytics, both linear and nonlinear optimization, and variational methods.

Hao Wang (Student Member, IEEE) received the B.Eng. degree in detection, guidance, and control technology from the College of Aeronautics and Astronautics, Central South University, Changsha, China, in 2020. He is currently working toward the Ph.D. degree in electronic engineering with the College of Control Science and Engineering, Zhejiang University, Hangzhou, China.

His research interests include process monitoring and time-series analysis.

Guofei Chen received the B.Eng. degree in automation from Zhejiang University, Hangzhou, China, in 2023. He is currently working toward the Master of Science degree in robotics with Robotics Institute, Carnegie Mellon University, Pittsburgh, USA.

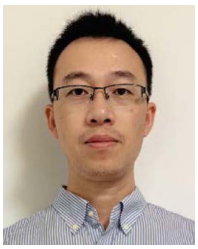
His research interests include optimization, localization, and planning with application in mobile robots.

Yiran Ma received the B.Eng. degree in automation from the School of Automation Science and Engineering, South China University of Technology, Guangzhou, China, in 2021. He is currently working toward the Ph.D. degree in control science and engineering with the State Key Laboratory of Industrial Control Technology, College of Control Science and Engineering, Zhejiang University, Hangzhou, China.

His current research interests include machine learning, Bayesian methods, and their applications in industrial data-driven modeling.

Le Yao (Member, IEEE) received the B.Eng. and M.Eng. degrees in automation from the Department of Control Science and Engineering, Jiangnan University, Wuxi, China, in 2012 and 2015, respectively, and the Ph.D. degree in automation from the Department of Control Science and Engineering, Zhejiang University, Hangzhou, China, in 2019.

He was a Post-Doctoral Research Fellow with the State Key Laboratory of Industrial Control Technology, College of Control Science and Engineering, Zhejiang University, from 2019 to 2022. From July 2023 to September 2023, he was a Visiting Scholar with the Hong Kong University of Science and Technology, Hong Kong. He is currently an Associate Professor with the School of Mathematics, Hangzhou Normal University, Hangzhou. His research interests include industrial Big Data, process monitoring, soft sensor, data-driven modeling, distributed computing, process data analysis, and their industrial applications.



Zhiqiang Ge (Senior Member, IEEE) received the B.Eng. and Ph.D. degrees in automation from Department of Control Science and Engineering, Zhejiang University, Hangzhou, China, in 2004 and 2009, respectively.

He is currently with the School of Mathematics, Southeast University, Nanjing, China. He was with Department of Chemical and Biomolecular Engineering, Hong Kong University of Science Technology, Department of Control Science and Engineering, Zhejiang University, and Peng Cheng Laboratory from 2009 to 2024. His research interests include industrial Big Data, process monitoring, soft sensor, data-driven modeling, machine intelligence, and knowledge automation.

Dr. Ge was an Alexander von Humboldt Research Fellow with University of Duisburg-Essen during 2014 to 2017, and also a JSPS invitation Fellow with Kyoto University during Jun. 2018 to Aug. 2018.

Zhihuan Song received the B.Eng. and M.Eng. degrees in industrial automation from the Hefei University of Technology, Anhui, China, in 1983 and 1986, respectively, and the Ph.D. degree in industrial automation from Zhejiang University, Hangzhou, China, in 1997.

His research interests include the modeling and fault diagnosis of industrial processes, analytics and applications of industrial Big Data, and advanced process control technologies.

Since 1997, he has been with the Department of Control Science and Engineering, Zhejiang University, where he was first a Postdoctoral Research Fellow, then an Associate Professor, and is currently a Professor. He has authored or coauthored more than 200 papers in journals and conference proceedings.