Foundations of Data Science
Vol. 5, No. 3, September 2023, pp. 340-365
doi:10.3934/fods.2023002

UNSUPERVISED LEARNING OF OBSERVATION FUNCTIONS IN STATE SPACE MODELS BY NONPARAMETRIC MOMENT METHODS

Qingci An[1], Yannis Kevrekidis[2,3], Fei Lu*[1] and Mauro Maggioni[1,2]

[1] Department of Mathematics, Johns Hopkins University

[2] Department of Applied Mathematics and Statistics, Johns Hopkins University

[3] Department of Chemical and Biomolecular Engineering, Johns Hopkins University
3400 N. Charles Street, Baltimore, MD 21218, USA

(Communicated by Dimitrios Giannakis)

Abstract. We investigate the unsupervised learning of non-invertible observation functions in nonlinear state space models. Assuming abundant data of the observation process along with the distribution of the state process, we introduce a nonparametric generalized moment method to estimate the observation function via constrained regression. The major challenge comes from the non-invertibility of the observation function and the lack of data pairs between the state and observation. We address the fundamental issue of identifiability from quadratic loss functionals and show that the function space of identifiability is the closure of a RKHS that is intrinsic to the state process. Numerical results show that the first two moments and temporal correlations, along with upper and lower bounds, can identify functions ranging from piecewise polynomials to smooth functions, leading to convergent estimators. The limitations of this method, such as non-identifiability due to symmetry and stationarity, are also discussed.

1. **Introduction.** We consider the following state space model for (X_t, Y_t) processes in $\mathbb{R} \times \mathbb{R}$:

$$\text{State space model:} \quad dX_t = a(X_t)dt + b(X_t)dB_t, \quad \text{with } a, b \text{ are known;} \tag{1.1}$$

$$\text{Observation model:} \quad Y_t = f_*(X_t), \quad \text{with } f_* \text{ unknown.} \tag{1.2}$$

Here (B_t) is the standard Brownian motion, the drift function $a(x)$ and the diffusion coefficient $b(x)$ are given, satisfying the linear growth and global Lipschitz conditions. We assume that the initial distribution of X_{t_0} is given; the state space model is therefore known, in the sense that the distribution of the process (X_t) is known.

Our goal is to estimate the unknown observation function f_* from data consisting of a large ensemble of trajectories of the process Y_t, denoted by $\{Y^{(m)}_{t_0:t_L}\}_{m=1}^M$, where

2020 *Mathematics Subject Classification.* Primary: 62G05, 68Q32, 62M15.

Key words and phrases. State-space models, nonparametric regression, generalized moment method, RKHS..

MM, YGK and FL are partially supported by DE-SC0021361 and FA9550-21-1-0317. FL is partially funded by the NSF Award DMS-1913243.

*Corresponding author: Fei Lu.

m indexes trajectories, and $t_0 < \cdots < t_L$ are the times at which the observations are made. In particular, there are no pairs (X_t, Y_t) being observed, so in the language of machine learning this may be considered an unsupervised learning problem. A case of particular interest in the present work is when the observation function f_* is nonlinear *and non-invertible*, and it is within a large class of functions, including smooth functions but also, for example, piecewise regular functions. We will also emphasize the role and usefulness of many short trajectories, vs. few long trajectories, albeit both the theory and algorithms that we consider are generally applicable in a wide range of regimes.

We estimate the observation function f_* by matching generalized moments, while constraining the estimator to a suitably chosen finite-dimensional hypothesis (function) space, whose dimension depends on the number of observations, in the spirit of nonparametric statistics. We consider both first- and second-order moments, as well as temporal correlations, of the observation process. The estimator minimizes the discrepancy between the moments over an hypothesis space (e.g. spanned by B-spline functions), with upper and lower pointwise constraints estimated from data. The method we propose has several significant strengths:

- the generalized moments do not require the invertibility of the observation function f_*;
- low-order generalized moments tend to be robust to additive observation noise;
- generalize moments avoid the need of local constructions, since they depend on the entire distribution of the latent and observed processes;
- our nonparametric approach does not require *a priori* information about the observation function, and, for example, it can deal with both regular and piecewise regular functions;
- the method is computationally efficient because the moments need to be estimated only once, and the computation is easily performed in parallel.

We note that the method we propose readily extends to multivariate state space models, with the main statistical and computational bottlenecks coming from the curse of dimensionality in the representation and estimation of a higher-dimensional f_* in terms of basis functions.

The problem we are considering has been studied in many contexts, including nonlinear system identification [2, 24], filtering and data assimilation [4, 22], albeit typically only when observations are in the form of one, or a small number of, long trajectories, and in the case of an invertible or smooth observations function f_*. The estimation of the unknown observation function and of the latent dynamics from unlabeled data has been considered in [11, 15, 18, 28] and references therein. Inference for state space models (SSMs) has been widely studied; most classical approaches focus on estimating the parameters in the SSM from a single trajectory of the observation process, by expectation-maximization methods maximizing the likelihood, or Bayesian approaches [2,4,12,19,24], with the recent studies estimating the coefficients in a kernel representation [37] or the coefficients of a pre-specified set of basis functions [36]. The recent work [38] estimates a slow manifold (and effective equations on it), image under a nonlinear but invertible map of a latent space where slow and fast variables in a slow-fast system of SDEs are independent and orthogonal, using short bursts of trajectories; see discussions and references therein for motivations, applications and related works.

Our framework combines nonparametric learning [7, 14] with the generalized moments method, that is mainly studied in the setting of parametric inference

[31, 32, 34]. We study the identifiability of the observation function f_* from first-order moments, and show that the first-order generalized moments can identify the function in the L^2 closure of a reproducing kernel Hilbert space (RKHS) that is intrinsic to the state space model. As far as we know, this is the first result on the function space of identifiability for nonparametric learning of observation functions in SSMs.

When the observation function is invertible, its unsupervised regression is investigated [33] by maximizing the likelihood for high-dimensional data. However, in many applications, particularly those involving complex dynamics, the observation functions are non-invertible, for example they are projections or nonlinear non-invertible transformations (e.g., $f(x) = |x|^2$ with $x \in \mathbb{R}^d$). As a consequence, the resulting observed process may have discontinuous or singular probability densities [13, 17]. In [28], it has been shown empirically that delayed coordinates with principal component analysis may be used to estimate the dimension of the hidden process, and diffusion maps [6] may yield a diffeomorphic copy of the observation function.

The remainder of the paper is organized as follows. We present the nonparametric generalized moments method in Section 2. In Section 3 we study the identifiability of the observation function from first-order moments, and show that the function spaces of identifiability are RKHSs intrinsic to the state space model. We present numerical examples to demonstrate the effectiveness and the limitations of the proposed method in Section 4. Section 5 summarizes this study and discusses directions of future research; we review the basic elements about RKHSs in Appendix A.

2. Non-parametric regression based on generalized moments.

Throughout this work, we focus on discrete-time observations of the state space model (1.1)–(1.2), because data in practice are discrete in time, and the extension to continuous time trajectories is straightforward. We thereby suppose that the data is in the form $\{Y^{(m)}_{t_0:t_L}\}_{m=1}^M$, with m indexing multiple independent trajectories, observed at the vector $t_0 : t_L$ of discrete times $(t_0, \cdots, t_L)$.

2.1. Generalized moments method.

We estimate the observation function f_* by the generalized moment method (GMM) [31, 32, 34], searching for an observation function $\widehat{f}$, in a suitable finite-dimensional hypothesis (function) space, such that the moments of functionals of the process $(\widehat{f}(X_t))$ are close to the empirical ones (computed from data) of $f_*(X_t)$.

We consider "generalized moments" in the form $\mathbb{E}[\xi(Y_{t_0:t_L})]$, where $\xi : \mathbb{R}^{L+1} \to \mathbb{R}^K$ is a functional of the trajectory $Y_{t_0:t_L}$. For example, the functional ξ can be $\xi(Y_{t_0:t_L}) = [Y_{t_0:t_L}, Y_{t_0}Y_{t_1}, \ldots, Y_{t_{L-1}}Y_{t_L}] \in \mathbb{R}^{2L+1}$, in which case $\mathbb{E}[\xi(Y_{t_0:t_L})] = \left[\mathbb{E}[Y_{t_0:t_L}], \mathbb{E}[Y_{t_0}Y_{t_1}], \ldots, \mathbb{E}\left[Y_{t_{L-1}}Y_{t_L}\right]\right]$ is the vector of the first moments and of temporal correlations at consecutive observation times. The empirical generalized moments ξ are computed from data by Monte Carlo approximation:

$$\mathbb{E}[\xi(Y_{t_0:t_L})] \approx E_M[\xi(Y_{t_0:t_L})] := \frac{1}{M}\sum_{m=1}^{M} \xi(Y^{(m)}_{t_0:t_L}), \tag{2.1}$$

which converges at the rate $M^{-1/2}$ by the Central Limit Theorem, since the M trajectories are independent. Meanwhile, since the state space model (hence the distribution of the state process) is known, for any putative observation function f,

we approximate the moments of the process $(f(X_t)))$ by simulating M' independent trajectories of the state process (X_t):

$$\mathbb{E}\left[\xi(f(X)_{t_0:t_L})\right] \approx \frac{1}{M'} \sum_{m=1}^{M'} \xi(f(X)^{(m)}_{t_0:t_L}) \,. \tag{2.2}$$

Here, with some abuse of notation, $f(X)^{(m)}_{t_0:t_L} := (f(X^{(m)}_{t_0}), \ldots, f(X^{(m)}_{t_L}))$. The number M' can be as large as we can afford from a computational perspective; note of course that the calculations above a trivially parallelizable over trajectories. In what follows, since M' can be chosen large – only subject to computational constraints – we consider the error in this empirical approximation negligible and work with $\mathbb{E}\left[\xi(f(X)_{t_0:t_L})\right]$ directly.

We estimate the observation function f_* by minimizing a notion of discrepancy between these two empirical generalized moments:

$$\widehat{f} = \operatorname*{arg\,min}_{f\in\mathcal{H}} \mathcal{E}^M(f), \quad \text{with } \mathcal{E}^M(f) := \operatorname{dist}\left(E_M[\xi(Y_{t_0:t_L})], \mathbb{E}\left[\xi(f(X)_{t_0:t_L})\right]\right)^2, \tag{2.3}$$

where f is restricted to some suitable hypothesis space $\mathcal{H}$, and $\operatorname{dist}(\cdot,\cdot)$ is a suitable distance between the moments to be specified later. We choose $\mathcal{H}$ to be a subset of an n-dimensional function space, spanned by basis functions $\{\phi_i\}$, within which we can write $\widehat{f} = \sum_{i=1}^n \widehat{c}_i\phi_i$. By the law of large numbers, $\mathcal{E}^M(f)$ tends almost surely to $\mathcal{E}(f) := \operatorname{dist}\left(\mathbb{E}\left[\xi(Y_{t_0:t_L})\right], \mathbb{E}\left[\xi(f(X)_{t_0:t_L})\right]\right)^2$.

It is desirable to choose the generalized moment functional ξ and the hypothesis space $\mathcal{H}$ so that the minimization in (2.3) can be performed efficiently. We select the functional ξ so that the moments $\mathbb{E}\left[\xi(f(X)_{t_0:t_L})\right]$, for $f = \sum_{i=1}^n c_i\phi_i$, can be efficiently evaluated for all $(c_1, \ldots, c_n)$. To this end, we choose linear functionals or low-degree polynomials, so that we only need to compute the moments of the basis functions once, and use these moments repeatedly during the optimization process, as discussed in Section 2.2. The selection of the hypothesis space is detailed in Section 2.3.

2.2. **Loss functional and estimator.** The generalized moments we consider include the first and the second moments, and the one-step temporal correlation:

$$\xi(Y_{t_0:t_L}) := (Y_{t_0:t_L}, Y^2_{t_0:t_L}, Y_{t_0}Y_{t_1}, \ldots, Y_{t_{L-1}}Y_{t_L}) \in \mathbb{R}^{3L+2}.$$

The loss functional in (2.3) is then chosen in the following form: for weights $w_1, \ldots, w_3 > 0$,

$$\begin{aligned}\mathcal{E}(f) :=& w_1 \underbrace{\frac{1}{L}\sum_{l=1}^{L} \left|\mathbb{E}[f(X_{t_l})] - \mathbb{E}[Y_{t_l}]\right|^2}_{\mathcal{E}_1(f)} + w_2 \underbrace{\frac{1}{L}\sum_{l=1}^{L} \left|\mathbb{E}[f(X_{t_l})^2] - \mathbb{E}[Y^2_{t_l}]\right|^2}_{\mathcal{E}_2(f)} \\ &+ w_3 \underbrace{\frac{1}{L}\sum_{l=1}^{L} \left|\mathbb{E}[f(X_{t_l})f(X_{t_{l-1}})] - \mathbb{E}[Y_{t_l}Y_{t_{l-1}}]\right|^2}_{\mathcal{E}_3(f)} .\end{aligned} \tag{2.4}$$

In principle, these weights are selected to balance the contributions of these terms, and we set them according to data as detailed in (4.1).

Let the hypothesis space $\mathcal{H}$ be a subset of the span of a linearly independent set $\{\phi_i\}_{i=1}^n$, which we specify in the next section. For $f = \sum_{i=1}^n c_i\phi_i \in \mathcal{H}$, we can write

the loss functionals $\mathcal{E}_1(f)$ in (2.4) as

$$\mathcal{E}_1(f) = \frac{1}{L}\sum_{l=1}^{L}\left|\sum_{i=1}^{n} c_i \mathbb{E}\left[\phi_i(X_{t_l})\right] - \mathbb{E}\left[Y_{t_l}\right]\right|^2 = c^\top \overline{A}_1 c - 2c^\top \overline{b}_1 + \tilde{b}_1, \tag{2.5}$$

where $\tilde{b}_1 := \frac{1}{L}\sum_{l=1}^{L}\mathbb{E}\left[Y_{t_l}\right]^2$, and the matrix $\overline{A}_1$ and the vector $\overline{b}_1$ are given by

$$\overline{A}_1(i,j) := \frac{1}{L}\sum_{l=1}^{L}\underbrace{\mathbb{E}\left[\phi_i(X_{t_l})\right]\mathbb{E}\left[\phi_j(X_{t_l})\right]}_{A_{1,l}(i,j)}, \qquad \overline{b}_1(i) := \frac{1}{L}\sum_{l=1}^{L}\underbrace{\mathbb{E}\left[\phi_i(X_{t_l})\right]\mathbb{E}\left[Y_{t_l}\right]}_{b_{1,l}(i)}. \tag{2.6}$$

Similarly, we can write $\mathcal{E}_2(f)$ and $\mathcal{E}_3(f)$ in (2.4) as

$$\begin{aligned}
\mathcal{E}_2(f) &= \frac{1}{L}\sum_{l=1}^{L}\left|\sum_{i=1}^{n} c_i c_j \underbrace{\mathbb{E}\left[\phi_i(X_{t_l})\phi_j(X_{t_l})\right]}_{A_{2,l}(i,j)} - \underbrace{\mathbb{E}\left[Y_{t_l}^2\right]}_{b_{2,l}}\right|^2,\\
\mathcal{E}_3(f) &= \frac{1}{L}\sum_{l=1}^{L}\left|\sum_{i=1}^{n} c_i c_j \underbrace{\mathbb{E}\left[\phi_i(X_{t_{l-1}})\phi_j(X_{t_l})\right]}_{A_{3,l}(i,j)} - \underbrace{\mathbb{E}\left[Y_{t_{l-1}}Y_{t_l}\right]}_{b_{3,l}}\right|^2.
\end{aligned} \tag{2.7}$$

Thus, with the above notations in (2.6)-(2.7), the minimizer of the loss functional $\mathcal{E}(f)$ over $\mathcal{H}$ is

$$\begin{aligned}
&\widehat{f}_{\mathcal{H}} := \sum_{i=1}^{n}\widehat{c}_i\phi_i\,, \qquad \widehat{c} := \underset{c\in\mathbb{R}^n \text{ s.t. } \sum_{i=1}^{n} c_i\phi_i\in\mathcal{H}}{\arg\min}\ \mathcal{E}(c), \qquad \text{where}\\
&\mathcal{E}(c) := w_1[c^\top\overline{A}_1 c - 2c^\top\overline{b}_1 + \tilde{b}_1] + \sum_{k=2}^{3} w_k \frac{1}{L}\sum_{l=1}^{L}\left|c^\top A_{k,l}c - b_{k,l}\right|^2.
\end{aligned} \tag{2.8}$$

Here, with an abuse of notation, we denote $\mathcal{E}(\sum_{i=1}^{n} c_i\phi_i)$ by $\mathcal{E}(c)$.

In practice, with data $\{Y^{(m)}_{[t_1:t_N]}\}_{m=1}^{M}$, we approximate the expectations involving the observation process (Y_t) by the corresponding empirical means as in (2.1). Meanwhile, we approximate the expectations involving the state process (X_t) by Monte Carlo as in (2.2), using M' trajectories. We assume that the sampling errors in the expectations of (X_t), i.e. in the terms $\{A_{k,l}\}_{k=1}^{3}$, are negligible, since the basis $\{\phi_i\}$ can be chosen to be bounded functions (such as B-spline polynomials) and M' can be as large as we can afford. We approximate $\{b_{k,l}\}_{k=1}^{3}$ by their empirical means $\{b^M_{k,l}\}_{k=1}^{3}$:

$$b_{1,l}(i) = \mathbb{E}\left[\phi_i(X_{t_l})\right]\mathbb{E}\left[Y_{t_l}\right] \approx \mathbb{E}\left[\phi_i(X_{t_l})\right]\frac{1}{M}\sum_{m=1}^{M} Y^{(m)}_{t_l} =: b^M_{1,l}(i)\,, \tag{2.9}$$

$$b_{2,l} = \mathbb{E}\left[|Y_{t_l}|^2\right] \approx \frac{1}{M}\sum_{m=1}^{M}|Y^{(m)}_{t_l}|^2 =: b^M_{2,l}\,, \tag{2.10}$$

$$b_{3,l} = \mathbb{E}\left[Y_{t_{l-1}}Y_{t_l}\right] \approx \frac{1}{M}\sum_{m=1}^{M} Y^{(m)}_{t_{l-1}}Y^{(m)}_{t_l} =: b^M_{3,l}\,. \tag{2.11}$$

Then, with $\overline{b}_1^M = \frac{1}{L}\sum_{l=1}^{L} b^M_{1,l}$ and $\widetilde{b}_1^M = \frac{1}{LM}\sum_{l=1}^{L}\sum_{m=1}^{M}\left(Y^{(m)}_{t_l}\right)^2$, the estimator from data is

$$\widehat{f}_{\mathcal{H},M} = \sum_{i=1}^{n} \widehat{c}_i \phi_i, \qquad \widehat{c} = \underset{c\in\mathbb{R}^n \text{ s.t. } \sum_{i=1}^n c_i\phi_i \in \mathcal{H}}{\arg\min} \mathcal{E}^M(c), \text{ where}$$
$$\mathcal{E}^M(c) = w_1\big[c^\top \overline{A}_1 c - 2c^\top \overline{b}_1^M + \widetilde{b}_1^M\big] + \sum_{k=2}^{3} w_k \frac{1}{L}\sum_{l=1}^{L} \left|c^\top A_{k,l} c - b_{k,l}^M\right|^2. \tag{2.12}$$

The minimization of $\mathcal{E}^M(c)$ can be performed with iterative algorithms, with each optimization iteration, with respect to c, performed efficiently since the data-based matrices and vectors, $\overline{A}_1, \overline{b}_1^M$ and $\{A_{k,l}, b_{k,l}^M\}_{k=2}^3$, only need to be computed once. The main source of sampling error is the empirical approximation of the moments of the process (Y_t). We specify the hypothesis space in the next section and provide a detailed algorithm for the computation of the estimator in Section 2.4.

Remark 2.1 (Moments involving Itô's formula)**.** When the data trajectories are continuous in time (or when they are sampled with a high frequency in time), we can utilize additional moments from Itô's formula. Recall that for $f \in C_b^2$, applying Itô formula for the diffusion process in (1.1), we have

$$f(X_{t+\Delta t}) - f(X_t) = \int_t^{t+\Delta t} \nabla f \cdot b(X_s) dW_s + \int_t^{t+\Delta t} \mathcal{L}f(X_s) ds,$$

where the operator $\mathcal{L}$ is

$$\mathcal{L}f = \nabla f \cdot a + \frac{1}{2} Hess(f) : b^\top b. \tag{2.13}$$

Hence, $\mathbb{E}\left[\Delta Y_{t_l}\right] = \mathbb{E}\left[\mathcal{L}f_*(X_{t_{l-1}})\right]\Delta t + o(\Delta t)$, where $\Delta Y_{t_l} = Y_{t_l} - Y_{t_{l-1}}$. Thus, when Δt is small, we can consider matching the generalized moments

$$\mathcal{E}_4(f) = \frac{1}{L}\sum_{l=1}^{L} \left| \mathbb{E}\left[\mathcal{L}f(X_{t_{l-1}})\right]\Delta t - \mathbb{E}\left[\Delta Y_{t_l}\right] \right|^2. \tag{2.14}$$

Similarly, we can further consider the generalized moments $\mathbb{E}\left[Y_t \Delta Y_t\right]$ and $\mathrm{Var}(\Delta Y_t)$ and the corresponding quartic loss functionals. Since they require the moments of the first- and second-order derivatives of the observation function, they are helpful when the observation function is smooth with bounded derivatives.

2.3. **Hypothesis space and optimal dimension.** We let the hypothesis space $\mathcal{H}$ be a class of bounded functions in $\mathrm{span}\{\phi_i\}_{i=1}^n$,

$$\mathcal{H} := \{f \,:\, f = \sum_{i=1}^{n} c_i\phi_i : y_{\min} \leqslant f(x) \leqslant y_{\max} \text{ for all } x \in \mathrm{supp}(\bar{\rho}_T)\}, \tag{2.15}$$

where the basis functions $\{\phi_i\}$ are to be specified below, and the empirical bounds

$$y_{\min} := \min\{Y_{t_l}^{(m)}\}_{l,m=1}^{L,M}, \quad y_{\max} := \max\{Y_{t_l}^{(m)}\}_{l,m=1}^{L,M}$$

aim to approximate the upper and lower bounds for f_*. Here the dimension n will be selected adaptive to data to avoid under- and over-fitting, as detailed in Algorithm 1. Note that the hypothesis space $\mathcal{H}$ is a bounded convex subset of the linear space $\mathrm{span}\{\phi_i\}_{i=1}^n$. While the pointwise bound constraints are for all x, in practice, for efficient computation, we apply these constraints at representative points, for example at the mesh-grid points used when the basis functions are piecewise polynomials. One may apply stronger constraints, such as requiring time-dependent bounds to hold at all times: $y_{\min}(t) \leqslant \sum_{i=1}^n c_i f_i(x) \leqslant y_{\max}(t)$ for each time t, where $y_{\min}(t)$ and $y_{\max}(t)$ are the minimum and maximum of the data set $\{Y_t^{(m)}\}_{m=1}^M$.

Basis functions. As basis functions $\{\phi_i\}$ for the subspace containing $\mathcal{H}$ we choose B-spline basis consisting of piecewise polynomials (see Appendix B.1 for details). To specify the knots of B-spline functions, we introduce a density function $\bar{\rho}_T^L$, which is the average of the probability densities $\{p_{t_l}\}_{l=1}^L$ of $\{X_{t_l}\}_{l=1}^L$:

$$\bar{\rho}_T^L(x) = \frac{1}{L}\sum_{l=1}^{L} p_{t_l}(x) \quad \xrightarrow{L\to\infty} \quad \bar{\rho}_T(x) = \frac{1}{T}\int_0^T p_t(x)dt, \tag{2.16}$$

when $t_L = T$ and $\max_{1\leqslant l\leqslant L} |t_l - t_{l-1}| \to 0$. Here $\bar{\rho}_T^L$ (and its continuous time limit $\bar{\rho}_T(x)$) describes the intensity of visits to the regions explored by the process (X_t). The knots of the B-spline function are from a uniform partition of $[R_{min}, R_{max}]$, the smallest interval enclosing the support of $\bar{\rho}_T^L$. Thus, the basis functions $\{\phi_i\}$ are piecewise polynomials with knots adaptive to the state space model which determines $\bar{\rho}_T^L$.

Dimension of the hypothesis space. It is important to select a suitable dimension of the hypothesis space to avoid under- or over-fitting. We select the dimension in two steps. First, we introduce an algorithm, namely *Cross-validating Estimation of Dimension Range* (CEDR), to estimate the range of the dimension from the quadratic loss functional $\mathcal{E}_1$. Its main idea is to avoid the sampling error amplification due to an unsuitably large dimension. The sampling error is estimated from data by splitting the data into two sets. Then, we select the optimal dimension that minimizes the 2-Wasserstein distance between the measures of data and prediction. See Appendix B.1 for details. Here we use the 2-Wasserstein distance because it is sensitive to small changes in $\widehat{f}$ caused by overfitting, and at the same time it can be efficiently computed even for large-sample datasets.

2.4. **Algorithm.** We summarize the above method of nonparametric regression with generalized moments in Algorithm 1. It minimizes a quartic loss function with the upper and lower bound constraints, and we perform the optimization with multiple initial conditions (see Appendix B.2 for the details).

Input: The state space model and data $\{Y_{t_0:t_L}^{(m)}\}_{m=1}^M$ consisting of multiple trajectories of the observation process.
Output: Estimator $\widehat{f}$.
1: Estimate the empirical density $\bar{\rho}_T^L$ in (2.16) and find its support $[R_{min}, R_{max}]$.
2: Select a basis type, Fourier or B-spline, with an estimated dimension range $[1, N]$ (by Algorithm 2), and compute the basis functions as described in Section 2.3 using the support of $\bar{\rho}_T^L$.
3: **for** $n = 1 : N$ **do**
4: Compute the moment matrices in (2.6)-(2.7) and the vectors $b_{k,l}^M$ in (2.11).
5: Find the estimator $\widehat{c}_n$ by optimization with multiple initial conditions. Compute and record the values of the loss functional and the 2-Wasserstein distances.
6: Select the optimal dimension n (and degree if B-spline basis) that has the minimal 2-Wasserstein distance in (B.5). Return the estimator $\widehat{f} = \sum_{i=1}^n c_n^i \phi_i$.

ALGORITHM 1. Estimating the observation function by nonparametric generalized moment methods

Computational complexity. The computational complexity is driven by the construction of the normal matrix and vectors and the evaluation of the 2-Wasserstein distances, which have complexity of order $\mathrm{O}(n^2LM)$ and $\mathrm{O}(nLM)$, respectively, for an overall complexity $\mathrm{O}((n^2+n)LM)$.

2.5. **Tolerance to noise in the observations.** The (generalized) moment method can tolerate large additive observation noise if the distribution of the noise is known. The estimation error caused by the noise is at the scale of the sampling error, which is negligible when the sample size is large.

Specifically, suppose that we observe $\{Y^{(m)}_{t_0:t_L}\}_{m=1}^M$ from the observation model

$$Y_{t_l} = f_*(X_{t_l}) + \eta_{t_l}, \tag{2.17}$$

where $\{\eta_{t_l}\}$ is sampled from a process (η_t) that is independent of (X_t) and has moments

$$\mathbb{E}[\eta_t] = 0, \quad C(s,t) = \mathbb{E}[\eta_t \eta_s], \text{ for } s,t \geqslant 0. \tag{2.18}$$

A typical example is when η being identically distributed independent Gaussian noise $\mathcal{N}(0,\sigma^2)$, which gives $C(s,t) = \sigma^2\delta(t-s)$.

The algorithm in Section 2 applies the noisy data with only a few changes. First, note that the loss functional in (2.4) involves only the moments $\mathbb{E}[Y_t]$, $\mathbb{E}[Y_t^2]$ and $\mathbb{E}[Y_{t_l}Y_{t_{l-1}}]$, which are moments of $f_*(X_t)$. When Y_t in (2.17) has observation noise specified above, we have

$$\begin{aligned}\mathbb{E}[f_*(X_t)] &= \mathbb{E}[Y_t] - \mathbb{E}[\eta_t] = \mathbb{E}[Y_t];\\ \mathbb{E}[f_*(X_t)f_*(X_s)] &= \mathbb{E}[Y_tY_s] - \mathbb{E}[\eta_t\eta_s] = \mathbb{E}[Y_tY_s] - C(t,s)\end{aligned}$$

for all $t, s \geqslant 0$. Thus, we only need to change the loss functional to be

$$\begin{aligned}\mathcal{E}(f) =& w_1\frac{1}{L}\sum_{l=1}^L \left|\mathbb{E}[f(X_{t_l})] - \mathbb{E}[Y_{t_l}]\right|^2 + w_2\frac{1}{L}\sum_{l=1}^L \left|\mathbb{E}[f(X_{t_l})^2] - \mathbb{E}[Y_{t_l}^2] + C(t,t)\right|^2\\ &+ w_3\frac{1}{L}\sum_{l=1}^L \left|\mathbb{E}[f(X_{t_l})f(X_{t_{l-1}})] - \mathbb{E}[Y_{t_l}Y_{t_{l-1}}] + C(t,s)\right|^2.\end{aligned} \tag{2.19}$$

Similar to (2.12), the minimizer of the loss functional can be then computed as

$$\begin{aligned}&\widehat{f}_{\mathcal{H},M} = \sum_{i=1}^n \widehat{c}_i\phi_i, \quad \widehat{c} = \underset{c\in\mathbb{R}^n \text{ s.t. } \sum_{i=1}^n c_i\phi_i\in\mathcal{H}}{\arg\min} \mathcal{E}^M(c), \text{ where}\\ &\mathcal{E}^M(c) = w_1[c^\top \overline{A}_1 c - 2c^\top \overline{b}_1^M + \widetilde{b}_1^M] + w_2\frac{1}{L}\sum_{l=1}^L \left|c^\top A_{2,l}c - b_{2,l}^M + C(t_l,t_l)\right|^2\\ &\quad + w_3\frac{1}{L}\sum_{l=1}^L \left|c^\top A_{3,l}c - b_{3,l}^M + C(t_l,t_{l+1})\right|^2,\end{aligned} \tag{2.20}$$

where all the A-matrices and b-vectors are the same as before (e.g., in (2.6)–(2.7) and (2.11)).

Note that the observation noise introduces sampling errors through b_1^M, $b_{2,l}^M$ and $b_{2,l}^M$, which are at the scale $\mathrm{O}(\frac{1}{\sqrt{M}})$. Also, note the A-matrices are independent of the observation noise. Thus, the observation noise affects the estimator only through the sampling error at the scale $\mathrm{O}(\frac{1}{\sqrt{M}})$, the same as the sampling error in the estimator from noiseless data.

3. **Identifiability.** We discuss in this section the identifiability of the observation function by the loss functionals in the previous section. We show that $\mathcal{E}_1$, the quadratic loss functional based on the 1st-order moments in (2.5), can identify the observation function in the $L^2(\bar{\rho}_T^L)$-closure of a reproducing kernel Hilbert space (RKHS) that is intrinsic to the state space model. In addition, the loss functional $\mathcal{E}_4$

in (2.14), based on the Itô formula, enlarges the function space of identifiability. We also discuss, in Section 3.2, some limitations of the loss functional $\mathcal{E}$ in (2.19), that combines the quadratic and quartic loss functionals: in particular, symmetries or sampling from a stationary measure may prevent us from identifying the observation function when using only generalized moments. The starting point is a definition of identifiability, which is a generalization of the uniqueness of minimizer of a loss function in parametric inference (see e.g., [3, page 431] and [8]).

Definition 3.1 (Identifiability)**.** We say that the observation function f_* is *identifiable* by a data-based loss functional $\mathcal{E}$ on a function space H if f_* is the unique minimizer of $\mathcal{E}$ in H.

When the loss functional is quadratic (such as $\mathcal{E}_1$ or $\mathcal{E}_4$), it has a unique minimizer in a Hilbert space if and only if its Frechét derivative is invertible in the Hilbert space; thus, the main task is to find such function spaces [21,23,25]. We will specify such function spaces for $\mathcal{E}_1$ and/or $\mathcal{E}_4$ in Section 3.1. We note that these function spaces do not take into account the constraints of upper and lower bounds, which generically lead to minimizers near or at the boundary of the constrained set. This consideration applies also to the piecewise quadratic functionals $\mathcal{E}_2$ and $\mathcal{E}_3$, which can be viewed as providing additional constraints (see Section 3.2).

3.1. **Identifiability by quadratic loss functionals.** We consider the quadratic loss functionals $\mathcal{E}_1$ and $\mathcal{E}_4$, and show that they can identify the observation function in the $L^2(\bar{\rho}_T^L)$-closure of reproducing kernel Hilbert spaces (RKHSs) that are intrinsic to the state space model.

Assumption 3.2. *We make the following assumptions on the state space model.*

- *The coefficients in the state space model (1.1) satisfy a global Lipschitz condition, and therefore also a linear growth condition: there exists a constant $C > 0$ such that $|a(x) - a(y)| + |b(x) - b(y)| \leqslant C|x - y|$ for all $x, y \in \mathbb{R}$, and $|a(x)| + |b(x)| \leqslant C(1 + |x|)$. We assume that $\inf_{x\in\mathbb{R}} b(x) > 0$ for all $x \in \mathbb{R}$. Furthermore, we assume that X_0 has a bounded density.*
- *The observation function f_* satisfies $\sup_{t\in[0,t_L]} \mathbb{E}\left[|f_*(X_t)|^2\right] < \infty$.*

Theorem 3.3. *Given discrete-time data $\{Y_{t_0:t_L}^{(m)}\}_{m=1}^M$ from the state space model (1.1) satisfying Assumption 3.2, let $\mathcal{E}_1$ and $\mathcal{E}_4$ be the loss functionals defined in (2.4) and (2.14). Denote $p_t(x)$ the density of the state process X_t at time t, and recall that $\bar{\rho}_T^L$ in (2.16) is the average, in time, of these densities. Let $\mathcal{L}^*$ be the adjoint of the 2nd-order elliptic operator $\mathcal{L}$ in (2.13). Then,*

(a) $\mathcal{E}_1$ has a unique minimizer in H_1, the $L^2(\bar{\rho}_T^L)$ closure of the RKHS $\mathcal{H}_{K_1}$ with reproducing kernel

$$K_1(x, x') = \frac{1}{\bar{\rho}_T^L(x)\bar{\rho}_T^L(x')} \frac{1}{L} \sum_{l=1}^{L} p_{t_l}(x) p_{t_l}(x'), \tag{3.1}$$

for (x, x') such that $\bar{\rho}_T^L(x)\bar{\rho}_T^L(x') > 0$, and $K_1(x, x') = 0$ otherwise. When the data is continuous ($L \to \infty$), we have $K_1(x, x') = \frac{1}{\bar{\rho}_T(x)\bar{\rho}_T(x')} \frac{1}{T} \int_0^T p_t(x)p_t(x')dt$.

(b) $\mathcal{E}_4$ has a unique minimizer in H_4, the $L^2(\bar{\rho}_T^L)$ closure of the RKHS $\mathcal{H}_{K_4}$ with reproducing kernel

$$K_4(x, x') = \frac{1}{\bar{\rho}_T^L(x)\bar{\rho}_T^L(x')} \frac{1}{L} \sum_{l=1}^{L} \mathcal{L}^* p_{t_l}(x) \mathcal{L}^* p_{t_l}(x'), \tag{3.2}$$

for (x, x') *such that* $\bar{\rho}_T^L(x)\bar{\rho}_T^L(x') > 0$, *and* $K_4(x, x') = 0$ *otherwise. When the data is continuous, we have* $K_4(x, x') = \frac{1}{\bar{\rho}_T(x)\bar{\rho}_T(x')} \frac{1}{T} \int_0^T \mathcal{L}^* p_t(x) \mathcal{L}^* p_t(x') dt$.

(c) $\mathcal{E}_1 + \mathcal{E}_4$ *has a unique minimizer in* H, *the* $L^2(\bar{\rho}_T^L)$ *closure of the RKHS* $\mathcal{H}_K$ *with reproducing kernel*

$$K(x, x') = \frac{1}{\bar{\rho}_T^L(x)\bar{\rho}_T^L(x')} \frac{1}{L} \sum_{l=1}^{L} \left[p_{t_l}(x) p_{t_l}(x') + \mathcal{L}^* p_{t_l}(x) \mathcal{L}^* p_{t_l}(x') \right], \tag{3.3}$$

for (x, x') *such that* $\bar{\rho}_T^L(x)\bar{\rho}_T^L(x') > 0$, *and* $K(x, x') = 0$ *otherwise. Similarly, we have* $K(x, x') = \frac{1}{\bar{\rho}_T(x)\bar{\rho}_T(x')} \frac{1}{T} \int_0^T [p_t(x)p_t(x') + \mathcal{L}^* p_t(x) \mathcal{L}^* p_t(x')] dt$ *for continuous data.*

In particular, f_* *is the unique minimizer of these loss functionals if* f_* *is in* H_1, H_4 *or* H.

To prove this theorem, we first introduce an operator characterization of the RKHS $\mathcal{H}_{K_1}$ in the next lemma. Similar characterizations hold for the RKHSs $\mathcal{H}_{K_4}$ and $\mathcal{H}_K$.

Lemma 3.4. *The function* K_1 *in* (3.1) *is a Mercer kernel, that is, it is continuous, symmetric and positive semi-definite. Also,* K_1 *is square integrable in* $L^2(\bar{\rho}_T^L \times \bar{\rho}_T^L)$, *and it defines a compact positive integral operator* $L_{K_1} : L^2(\bar{\rho}_T^L) \to L^2(\bar{\rho}_T^L)$:

$$[L_{K_1} h](x') = \int h(x) K_1(x, x') \bar{\rho}_T^L(x) dx. \tag{3.4}$$

Also, the RKHS $\mathcal{H}_{K_1}$ *has the operator characterization:* $\mathcal{H}_{K_1} = L_{K_1}^{1/2}(L^2(\bar{\rho}_T^L))$ *and* $\{\sqrt{\lambda_i}\psi_i\}_{i=1}^{\infty}$ *is an orthonormal basis of the RKHS* $\mathcal{H}_{K_1}$, *where* $\{\lambda_i, \psi_i\}$ *are the pairs of positive eigenvalues and corresponding eigenfunctions of* L_{K_1}.

Proof. Since the densities $\{p_{t_l}\}$ are smooth, the kernel K_1 is continuous on the support of $\bar{\rho}_T^L$ and it is symmetric. It is positive semi-definite (see Appendix A for a definition) because for any $(c_1, \ldots, c_n) \in \mathbb{R}^n$ and $(x_1, \ldots, x_n)$, we have

$$\sum_{i,j=1}^{n} c_i c_j K(x_i, x_j) = \frac{1}{L} \sum_{l=1}^{L} \sum_{i,j=1}^{n} c_i c_j \frac{p_{t_l}(x_i) p_{t_l}(x_j)}{\bar{\rho}_T^L(x_i)\bar{\rho}_T^L(x_j)} = \frac{1}{L} \sum_{l=1}^{L} \left(\sum_{i=1}^{n} c_i \frac{p_{t_l}(x_i)}{\bar{\rho}_T^L(x_i)} \right)^2 \geqslant 0.$$

Thus, K_1 is a Mercer kernel.

To show that K_1 is square integrable, note first that $p_{t_l}(x) \leqslant \max_{1 \leqslant k \leqslant L} p_{t_k}(x) \leqslant L\bar{\rho}_T^L(x)$ for any x. Thus for each x, x', we have

$$\frac{1}{L} \sum_{l=1}^{L} p_{t_l}(x) p_{t_l}(x') \leqslant L^2 \bar{\rho}_T^L(x) \bar{\rho}_T^L(x')$$

and $K_1(x, x') \leqslant L$. It follows that K_1 is in $L^2(\bar{\rho}_T^L \times \bar{\rho}_T^L)$.

Since K_1 is positive definite and square integrable, the integral operator L_{K_1} is compact and positive. The operator characterization follows from Theorem A.3. □

Remark 3.5. The above lemma is only applicable to discrete-time observations because it uses the bound $K_1(x, x') \leqslant L$. When the data is continuous in time on $[0, T]$, we have $K_1 \in L^2(\bar{\rho}_T \times \bar{\rho}_T)$ if the support of $\bar{\rho}_T$ is compact, since p_t is uniformly bounded above, i.e. $p_t(x) \leqslant \max_{y \in \mathbb{R}, s \in [0,T]} p_s(y) < \infty$, since it is a regular

solution of a Fokker-Planck equation which is uniformly elliptic by Assumption 3.1 (see e.g., [10, Chapter 6]). Thus for each x, x', we have

$$\frac{1}{T}\int_0^T p_t(x)p_t(x')dt \leqslant \left|\frac{1}{T}\int_0^T p_t(x)^2\ dt\right|^{1/2}\left|\frac{1}{T}\int_0^T p_t(x')^2\ dt\right|^{1/2}$$
$$= \bar\rho_T(x)^{1/2}\bar\rho_T(x')^{1/2}\max_{y\in\mathbb{R},s\in[0,T]} p_s(y)$$

by Cauchy-Schwartz for the first inequality. Then,

$$K_1(x,x') = \frac{1}{\bar\rho_T(x)\bar\rho_T(x')}\frac{1}{T}\int_0^T p_t(x)p_t(x')dt \leqslant \bar\rho_T(x)^{-1/2}\bar\rho_T(x')^{-1/2}\max_{y\in\mathbb{R},s\in[0,T]} p_s(y).$$

It follows that K_1 is in $L^2(\bar\rho_T\times\bar\rho_T)$:

$$\int\int K_1^2(x,x')\bar\rho_T(x)\bar\rho_T(x')dxdx' \leqslant |\mathrm{supp}(\bar\rho_T)|\max_{y\in\mathbb{R},s\in[0,T]} p_s(y)^2 < \infty.$$

When $\bar\rho_T$ has non-compact support, it remains to be proved that $K_1 \in L^2(\bar\rho_T\times\bar\rho_T)$.

Proof of Theorem 3.3. The proof for (a)–(c) are similar, so we focus on (a) and only sketch the proof for (b)–(c).

To prove (a), we only need to show the uniqueness of the minimizer, because Lemma 3.4 has shown that K_1 is a Mercer kernel. Furthermore, note that by Lemma 3.4, the $L^2(\bar\rho_T^L)$ closure of the RKHS $\mathcal{H}_{K_1}$ is $H_1 = \overline{\mathrm{span}\{\psi_i\}_{i=1}^\infty}$, the closure in $L^2(\bar\rho_T^L)$ of the eigenspace of L_{K_1} with non-zero eigenvalues, where L_{K_1} is the operator defined in (3.4).

For any $f\in H_1$, denoting $h = f - f_*$, we have $\mathbb{E}[f(X_t)] - \mathbb{E}[Y_t] = \mathbb{E}[h(X_t)]$ for each t (recall that $Y_t = f_*(X_t)$). Hence, we can write the loss functional as

$$\begin{aligned}\mathcal{E}_1(f) =& \frac{1}{L}\sum_{l=1}^L |\mathbb{E}[f(X_{t_l})] - \mathbb{E}[Y_{t_l}]|^2\\ =& \frac{1}{L}\sum_{l=1}^L |\mathbb{E}[h(X_{t_l})]|^2 = \int\int h(x)h(x')\frac{1}{L}\sum_{l=1}^L p_{t_l}(x)p_{t_l}(x')dxdx' \qquad (3.5)\\ =& \int\int h(x)h(x')K_1(x,x')\bar\rho_T^L(x)\bar\rho_T^L(x')dxdx' \geqslant 0.\end{aligned}$$

Thus, $\mathcal{E}_1$ attains its unique minimizer in H_1 at f_* if and only if $\mathcal{E}_1(f_*+h) = 0$ with $h\in H_1$ implies that $h = 0$. Note that the second equality in (3.5) implies that $\mathcal{E}_1(f_*+h) = 0$ if and only if $\mathbb{E}[h(X_{t_l})] = 0$, i.e. $\int h(x)p_{t_l}(x)dx = 0$, for all t_l. Then, $\int h(x)p_{t_l}(x)\frac{p_{t_l}(x')}{\bar\rho_T^L(x')}dx = 0$ for each t_l and x'. Thus, the sum of them is also zero:

$$0 = \int h(x)\frac{1}{L}\sum_{l=1}^L \frac{p_{t_l}(x)p_{t_l}(x')}{\bar\rho_T^L(x')\bar\rho_T^L(x)}\bar\rho_T^L(x)dx = \int h(x)K_1(x,x')\bar\rho_T^L(x)dx$$

for each x'. By the definition of the operator L_{K_1}, this implies that $L_{K_1}h = 0$. Thus, $h = 0$ because $h\in H_1$.

The above arguments hold true when the kernel K_1 is from continuous-time data: one only has to replace $\frac{1}{L}\sum_{l=1}^L$ by the averaged integral in time. This completes the proof for (a).

The proofs of (b) and (c) are the same as above except the appearance of the operator $\mathcal{L}^*$. Note that $\mathcal{E}_4$ in (2.14) reads $\mathcal{E}_4(f) = \frac{1}{L}\sum_{l=1}^L |\mathbb{E}\left[\mathcal{L}f(X_{t_l})\right] - \mathbb{E}\left[\Delta Y_{t_l}\right]|^2$,

thus, it differs from $\mathcal{E}_1$ only at the expectation $\mathbb{E}[\mathcal{L}f(X_{t_l})]$. By integration by parts, we have

$$\mathbb{E}[\mathcal{L}f(X_s)] = \int \mathcal{L}f(x)p_s(x)dx = \int f(x)\mathcal{L}^* p_s(x)dx$$

for any $f \in C_b^2$. Then, the rest of the proof for Part (b) follows exactly as above with K_1 and L_{K_1} replaced by K_4 and L_{K_4}. □

The following remarks highlight the implications of the above theorem. We consider only $\mathcal{E}_1$, but all the remarks apply also to $\mathcal{E}_4$ and $\mathcal{E}_1 + \mathcal{E}_4$.

Remark 3.6 (An operator view of identifiability)**.** The unique minimizer of $\mathcal{E}_1$ in H_1 defined in Theorem 3.3 is the zero of its Frechét derivative: $\widehat{f} = L_{K_1}^{-1} L_{K_1} f_*$, which is f_* if $f_* \in H_1$. In fact, note that with the integral operator L_{K_1}, we can write the loss functional $\mathcal{E}_1$ as

$$\mathcal{E}_1(f) = \langle f - f_*, L_{K_1}(f - f_*)\rangle_{L^2(\bar{\rho}_T^L)}.$$

Thus, the Frechét derivative of $\mathcal{E}_1$ over $L^2(\bar{\rho}_T^L)$ is $\nabla\mathcal{E}_1(f) = L_{K_1}(f - f_*)$ and we obtain the unique minimizer. Furthermore, this operator representation of the minimizer conveys two important messages about the inverse problem of finding the minimizer of $\mathcal{E}_1$: (1) it is *ill-defined* beyond H_1, and in particular, it is ill-defined on $L^2(\bar{\rho}_T^L)$ when L_{K_1} is not strictly positive; (2) the inverse problem is ill-posed on H_1, because the operator L_{K_1} is compact and its inverse $L_{K_1}^{-1}$ is unbounded.

Remark 3.7 (Identifiability and normal matrix in regression)**.** Suppose $\mathcal{H}_n = \text{span}\{\phi_i\}_{i=1}^n$ and denote $f = \sum_{i=1}^n c_i\phi_i$ with ϕ_i being basis functions such as B-splines. As shown in (2.5)-(2.6), the loss functional $\mathcal{E}_1$ becomes a quadratic function with normal matrix $\overline{A}_1 = \frac{1}{L}\sum_{l=1}^L A_{1,l}$ with $A_{1,l} = \mathbf{u}_l^\top \mathbf{u}_l$, where the vector $\mathbf{u}_l = (\mathbb{E}[\phi_1(X_{t_l})], \ldots, \mathbb{E}[\phi_n(X_{t_l})]) \in \mathbb{R}^n$. Thus, the rank of the matrix $\overline{A}_1$ is no larger than $\min\{n, L\}$. Note that $\overline{A}_1$ is the matrix approximation of L_{K_1} on the basis $\{\phi_i\}_{i=1}^n$ in the sense that

$$\overline{A}_1(i,j) = \langle L_{K_1}\phi_i, \phi_j\rangle_{L^2(\bar{\rho}_T^L)},$$

for each $1 \leqslant i, j \leqslant n$. Thus, the minimum eigenvalue of $\overline{A}_1$ approximates the minimal eigenvalue of L_{K_1} restricted in $\mathcal{H}_n$. In particular, if $\mathcal{H}_n$ contains a nonzero element in the null space of L_{K_1}, then the normal matrix will be singular; if $\mathcal{H}_n$ is a subspace of the $L^2(\bar{\rho}_T^L)$ closure of $\mathcal{H}_{K_1}$, then the normal matrix is invertible and we can find a unique minimizer.

Remark 3.8 (Convergence of estimator)**.** For a fixed hypothesis space, the estimator converges to the projection of f_* in $\mathcal{H} \cap H_1$ as the data size M increases, at the rate $O(M^{-1/2})$, with the error coming from the Monte Carlo estimation of the moments of observations. With data-adaptive hypothesis spaces, we are unable to prove the minimax rate of convergence as in classical nonparametric regression, due to the lack of a coercivity condition [23, 26], since the eigenvalues of the compact operator L_{K_1} converge to zero. A minimax rate would require an estimate on the spectral decay of L_{K_1}, which we leave for future research.

Remark 3.9 (Regularization using the RKHS)**.** The RKHS H_{K_1} provides a data-adaptive regularization norm in the Tikhonov regularization (see [25]).

Examples of the RKHS. We emphasize that the reproducing kernel and the RKHS are intrinsic to the state space model (including the initial distribution). We demonstrate the kernels by analytically computing them when the process (X_t) is either the Brownian motion or the Ornstein-Uhlenbeck (OU) process. For simplicity, we consider continuous-time data. Recall that when the diffusion coefficient in the state space model (1.1) is a constant, the second-order elliptic operators $\mathcal{L}$ is $\mathcal{L}f = \nabla f \cdot a + \frac{1}{2}b^2\Delta f$, and its adjoint operator $\mathcal{L}^*$ is

$$\mathcal{L}^* p_s = -\nabla \cdot (a p_s) + \frac{1}{2} b^2 \Delta p_s,$$

where p_s denotes the probability density of X_s.

Example 3.10 (1D Brownian motion)**.** Let $a = 0$ and $b = 1$. Assume $p_0(x) = \delta_{x_0}$, i.e., $X_0 = x_0$. Then, X_t is the Brownian motion starting from x_0 and $p_t(x) = \frac{1}{\sqrt{2\pi t}} e^{-\frac{(x-x_0)^2}{2t}}$. We have $\bar{\rho}_T(x) = \frac{1}{T}\int_0^T p_t(x)dt = \frac{x-x_0}{T\sqrt{\pi}}\Gamma(-\frac{1}{2}, \frac{(x-x_0)^2}{2T})$ and

$$\begin{aligned} K_1(x,x') =& \frac{1}{\bar{\rho}_T(x)\bar{\rho}_T(x')}\frac{1}{T}\int_0^T p_s(x)p_s(x')ds \\ =& \frac{T\Gamma(0, \frac{(x-x_0)^2+(x'-x_0)^2}{2T})}{2(x-x_0)(x'-x_0)\Gamma(-\frac{1}{2}, \frac{(x-x_0)^2}{2T})\Gamma(-\frac{1}{2}, \frac{(x'-x_0)^2}{2T})}, \end{aligned}$$

where $\Gamma(s,x) := \int_x^\infty t^{s-1}e^{-t}dt$ is the upper incomplete Gamma function. Also, we have

$$\mathcal{L}^* p_s(x) = \phi_2(s,x)p_s(x), \text{ with } \phi_2(s,x) = \left(\frac{1}{s^2}(x-x_0)^2 - \frac{1}{s}\right).$$

Thus, the kernels K_4 in (3.2) and K in (3.3) from continuous-time data are

$$\begin{aligned} K_4(x,x') &= \frac{1}{\bar{\rho}_T(x)\bar{\rho}_T(x')}\frac{1}{T}\int_0^T \phi_2(s,x)\phi_2(s,x')p_s(x)p_s(x')ds; \\ K(x,x') &= \frac{1}{\bar{\rho}_T(x)\bar{\rho}_T(x')}\frac{1}{T}\int_0^T (1+\phi_2(s,x)\phi_2(s,x'))p_s(x)p_s(x')ds. \end{aligned}$$

Example 3.11 (Ornstein-Uhlenbeck process)**.** Let $a(x) = \theta x$ and $b = 1$ with $\theta > 0$. Assume $p_0(x) = \delta_{x_0}$, i.e., $X_0 = x_0$. Then, $X_t = e^{-\theta t}x_0 + \int_0^t e^{-\theta(t-s)}dB_s$. It has a distribution $\mathcal{N}(e^{-\theta t}x_0, \frac{1}{2\theta}(1-e^{-2\theta t}))$, thus $p_t(x) = \frac{1}{\sqrt{2\pi}\sigma_t}\exp(-\frac{(x-x_0^t)^2}{2\sigma_t^2})$, where $x_0^t := e^{-\theta t}x_0$ and $\sigma_t^2 := \frac{1}{2\theta}(1-e^{-2\theta t})$. Computing the spatial derivatives, we have $\mathcal{L}^* p_s(x) = \frac{1}{2}\left[\frac{(x-x_0^s)^2}{\sigma_s^4} - \frac{1}{\sigma_s^2}\right]p_s(x) - (\theta x p_s(x))' = \phi_2(s,x)p_s(x)$, where

$$\phi_2(s,x) := \left[\frac{(x-x_0)^2}{2\sigma_s^4} - \frac{1}{2\sigma_s^2} - \theta + \frac{\theta}{\sigma_s^2}x(x-x_0^s)\right].$$

The reproducing kernels K_1 in (3.1), K_4 in (3.2) and K in (3.3) are

$$\begin{aligned} K_1(x,x') &= \frac{1}{\bar{\rho}_T(x)\bar{\rho}_T(x')}\frac{1}{T}\int_0^T p_s(x)p_s(x')ds; \\ K_4(x,x') &= \frac{1}{\bar{\rho}_T(x)\bar{\rho}_T(x')}\frac{1}{T}\int_0^T \phi_2(s,x)\phi_2(s,x')p_s(x)p_s(x')ds; \\ K(x,x') &= \frac{1}{\bar{\rho}_T(x)\bar{\rho}_T(x')}\frac{1}{T}\int_0^T (1+\phi_2(s,x)\phi_2(s,x'))p_s(x)p_s(x')ds. \end{aligned}$$

In particular, when the process is stationary, we have $K_1(x,x') \equiv 1$ and $K_4(x,x') = 0$ because $\mathcal{L}^* p_s = 0$ when $p_s(x) = \frac{2\theta}{\sqrt{2\pi}} \exp(-\theta x^2)$ is the stationary density.

3.2. **Non-identifiability due to stationarity and symmetry.** When the hypothesis space $\mathcal{H}$ has a dimension larger than the RKHS's, the quadratic loss functional $\mathcal{E}_1$ may have multiple minimizers. The constraints of upper and lower bounds, as well as the loss functionals $\mathcal{E}_2$ and $\mathcal{E}_3$, can help identifying the observation function. However, as we show next, identifiability may still not hold due to symmetry and/or stationarity.

Stationary processes. When the process (X_t) is stationary, we have limited information from the moments in our loss functionals. We have $\mathcal{E}_1(f) = |\mathbb{E}[Y_{t_1}] - \mathbb{E}[f(X_{t_1})]|^2$ with $K_1(x,x') \equiv 1$, so $\mathcal{E}_1$ can only identify a constant function. Also, the loss functional $\mathcal{E}_4$ is identically 0 because

$$\mathcal{L}^* p_s = \partial_s p_s = 0 \quad \Leftrightarrow \quad \mathbb{E}[\mathcal{L}h(X_s)] = 0 \text{ for any } h \in C_b^2.$$

In other words, the function space of identifiability with $\mathcal{E}_1 + \mathcal{E}_4$ is the space of constant functions. Meanwhile, the quartic loss functionals $\mathcal{E}_2$ and $\mathcal{E}_3$ also provide limited information: they become $\mathcal{E}_2 = \left|\mathbb{E}[f(X_{t_1})^2] - \mathbb{E}[Y_{t_1}^2]\right|^2$ and $\mathcal{E}_3 = |\mathbb{E}[f(X_{t_2})f(X_{t_1})] - \mathbb{E}[Y_{t_2}Y_{t_1}]|^2$, the second-order moment and the temporal correlation at a single pair of times.

To see the ensuing limitations, consider the finite-dimensional hypothesis space $\mathcal{H}$ in (2.15). As in (2.12), with $f = \sum_{i=1}^n c_i \phi_i$, the loss functional becomes

$$\mathcal{E}(f) = c^\top \overline{A}_1 c - 2c^\top \overline{b}_1^M + |\mathbb{E}[Y_{t_1}]|^2 + \sum_{k=2}^{3} \left|c^\top A_{k,1} c - b_{k,1}^M\right|^2,$$

where $\overline{A}_1$ is a rank-one matrix, and $\sum_{k=2}^3 |c^\top A_{k,1} c - b_{k,1}^M|^2$ only adds two additional constraints. Thus, $\mathcal{E}$ has multiple minimizers in a linear space with dimension greater than 3. One has to resort to the upper and lower bounds in (2.15) for additional constraints, which lead to minimizers on the boundary of the resulting convex set.

Symmetry. When the distribution of the state process X_t is symmetric, a moment-based loss functional may not distinguish the true observation function from its symmetric counterpart. More specifically, if a transformation $R : \mathbb{R} \to \mathbb{R}$ preserves the distribution, i.e., $(X_t, t \geqslant 0)$ and $(R(X_t), t \geqslant 0)$ have the same distribution, then $\mathbb{E}[f(X_t)] = \mathbb{E}[f \circ R(X_t)]$ and $\mathbb{E}[f(X_t)f(X_s)] = \mathbb{E}[f \circ R(X_t) f \circ R(X_s)]$. Thus, our loss functional will not distinguish f from $f \circ R$. This is of course reasonable: the two functions yield the same observation process (in terms of the distribution), thus the observation data does not provide the information necessary for distinguishing f from $f \circ R$.

Example 3.12 (Brownian motion)**.** Consider the standard Brownian motion X_t, whose distribution is symmetric about $x = 0$ (because the two processes $(X_t, t \geqslant 0)$ and $(-X_t, t \geqslant 0)$ have the same distribution). Let the transformation R be $R(x) = -x$. Then, the two functions $f(x)$ and $f(-x)$ lead to the same observation process, thus they cannot be distinguished from the observations.

4. **Numerical examples.** We demonstrate the effectiveness and limitations of our algorithm using synthetic data in representative examples. The algorithm works well when the state space model's densities vary appreciably in time to yield a function space of identifiability whose distance to the true observation function is

small. In this case, our algorithm leads to a convergent estimator as the sample size increases. We also demonstrate that when the state process (i.e., the Ornstein-Uhlenbeck process) is stationary or symmetric in distribution (i.e., the Brownian motion), the loss functional can have multiple minimizers in the hypothesis space, preventing us from identifying the observation functions (see Section 4.3).

4.1. **Numerical setup.** The synthetic data $\{Y^{(m)}_{t_0:t_L}\}_{m=1}^M$ with $t_l = l\Delta t$ are generated from the state space model, which is solved by the Euler-Maruyama scheme with a time-step $\Delta t = 0.01$ for $L = 100$ steps. We consider sample sizes $M \in \{\lfloor 10^{3.5+j\Delta}\rfloor : j = 0, 1, 2, 3, 4,\ \Delta = 0.0625\}$ to test the convergence of the estimator.

To estimate the moments in the A-matrices and b-vectors in (2.6)–(2.7) by Monte Carlo, we generate a new set of independent trajectories $\{X^{(m)}_{t_l}\}_{m=1}^{M'}$ with $M' = 10^6$. We emphasize that these samples of X are independent of the data $\{Y^{(m)}_{t_0:t_L}\}_{m=1}^M$.

Inference algorithm. We follow Algorithm 1 to search for the global minimum of the loss functionals in (2.12). The weights for the $\mathcal{E}_k$'s are $w_k = L\sqrt{M}/\|m_k^Y\|$, where $\|\cdot\|$ is the Euclidean norm on $\mathbb{R}^L$, and for $l = 0, 1, \cdots, L-1$,

$$m_k^Y(l) = \frac{1}{M}\sum_{m=1}^{M}(Y^{(m)}_{t_l})^k \text{ for } k = 1, 2 \quad \text{and} \quad m_3^Y(l) = \frac{1}{M}\sum_{m=1}^{M} Y^{(m)}_{t_l} Y^{(m)}_{t_{l+1}}. \tag{4.1}$$

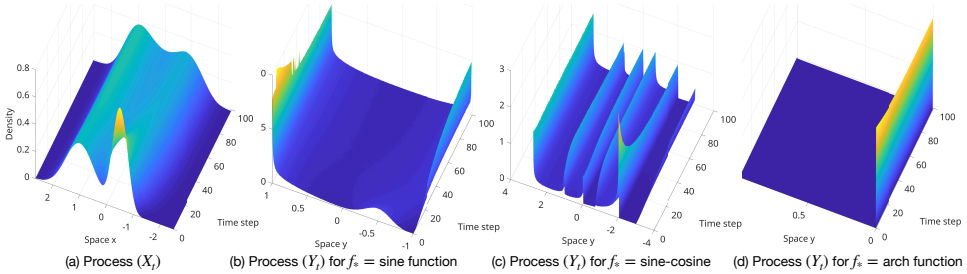
For each example, we test hypothesis spaces, spanned by B-splines with degree in $\{0, 1, 2, 3\}$, with a dimension selected by Algorithm 2 in the range $[1, N]$. We select the optimal dimension and degree with the minimal 2-Wasserstein distance between the predicted and true distribution of Y. The details are presented in Section C.

Results assessment and presentation. We present three aspects of the estimator $\widehat{f}$:

- **Estimated and true functions.** We compare the estimator with the true function f_*, along with the $L^2(\bar{\rho}_T^L)$ projection of f_* to the linear space expanded by the elements of $\mathcal{H}$.
- **2-Wasserstein distance.** We present the 2-Wasserstein distance (see (B.5)) between the distributions of $Y_{t_l} = f_*(X_{t_l})$ and $\widehat{f}(X_{t_l})$ for each time with training data and a new set of randomly generated data of size 10^6. The new (test) data has $Y^{(m)}_{t_l} = f_*(X^{(m)}_{t_l})$, i.e., the X's and Y's are generated in pairs, while in the training data the X's and Y's are generated independently. This pairing can lead to an effect on the 2-Wasserstein distance, which depends only on the empirical distribution of the samples, but such effect is negligible in our experiments due to the large sample size.
- **Convergence of $L^2(\bar{\rho}_T^L)$ error.** We test the convergence of the estimator in $L^2(\bar{\rho}_T^L)$ as the sample size M increases. The $L^2(\bar{\rho}_T^L)$ error is computed by the Riemann sum approximation. We present the mean and standard deviation of $L^2(\bar{\rho}_T^L)$ errors from 20 independent simulations. The convergence rate is also highlighted, and we compare it with the minimax convergence rate in classical nonparametric regression (see e.g., [14, 26]), which is $\frac{s}{2s+1}$ with $s-1$ being the degree of the B-spline basis. This minimax rate is not available yet for our method, see Remark 3.8.

4.2. **Examples.** The state space model we consider is a stochastic differential equation with the double-well potential

$$dX_t = (X_t - X_t^3)dt + dB_t, X_{t_0} \sim p_{t_0} \tag{4.2}$$



(a) Process (X_t) (b) Process (Y_t) for f_* = sine function (c) Process (Y_t) for f_* = sine-cosine (d) Process (Y_t) for f_* = arch function

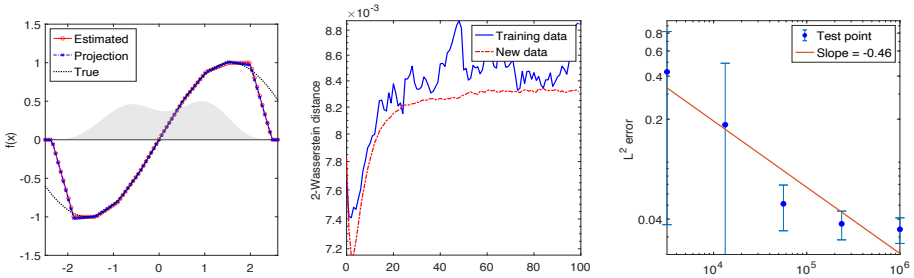
FIGURE 1. Empirical densities from the data trajectories of the process (X_{t_l}) in (4.2) and the observation processes (Y_{t_l}) with $f_* = f_i$, where f_i's are the three observation functions in (4.3). Since we do not have data pairs between $(X_{t_l}^{(m)}, Y_{t_l}^{(m)})$, these empirical densities are the available information from data. Our goal is to find the function f_* in the operator that maps the densities of $\{X_{t_l}\}$ to the densities of $\{Y_{t_l}\}$.

where the density of X_{t_0} is the average of $\mathcal{N}(-0.5, 0.2)$ and $\mathcal{N}(1, 0.5)$. The distribution of $X_{t_0:t_L}$ is non-symmetric and far from stationary (see Figure 1(a)); we therefore expect that the quadratic loss functional $\mathcal{E}_1$ provides a rich RKHS space for learning.

We consider three observation functions f representing typical challenges: nearly invertible, non-invertible, and non-invertible discontinuous, in the set $\text{supp}(\bar{\rho}_T)$:

$$\begin{aligned} &\text{Sine function:} && f_1(x) = \sin(x); \\ &\text{Sine-Cosine function:} && f_2(x) = 2\sin(x) + \cos(6x); \\ &\text{Arch function:} && f_3(x) = \left(-2(1-x)^3 + 1.5(1-x) + 0.5\right) \mathbf{1}_{x\in[0,1]}. \end{aligned} \tag{4.3}$$

These functions are shown in 2(a)–4(a). They lead to observation processes with dramatically different distributions, as shown in Fig.1(b-d).



(a) Estimator with n=9, deg=1 (b) Wasserstein distance (c) Convergence rate

FIGURE 2. Learning results of Sine function $f_1(x) = \sin(x)$ with model (4.2).

The learning results for these three functions are shown in Figure 2–4. For each of these three observation functions, we present the estimator with the optimal hypothesis space, the 2-Wasserstein distance in prediction and the convergence of the estimator in $L^2(\bar{\rho}_T^L)$ (see Section 4.1 for details).

Sine function: Fig. 2(a) shows the estimator with degree-1 B-spline basis with dimension $n = 9$ for $M = 10^6$. The $L^2(\bar{\rho}_T^L)$ error is 0.0245 and the relative error is 3.47%. Fig. 2(b) shows that the Wasserstein distances are small at the scale 10^{-3}, in

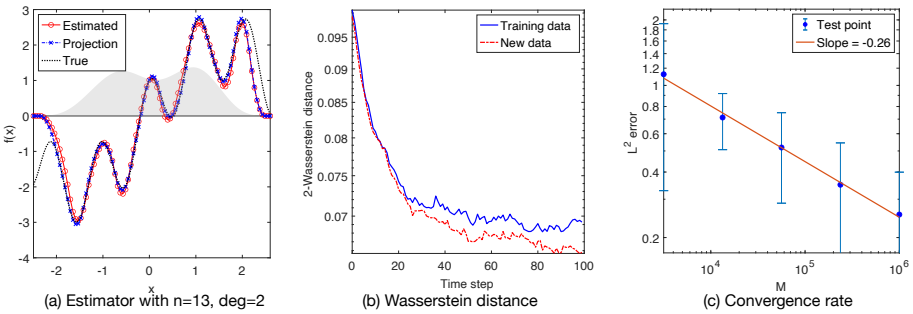


FIGURE 3. Learning results of Sine-Cosine function $f_2(x) = 2\sin(x) + \cos(6x)$ with model (4.2).

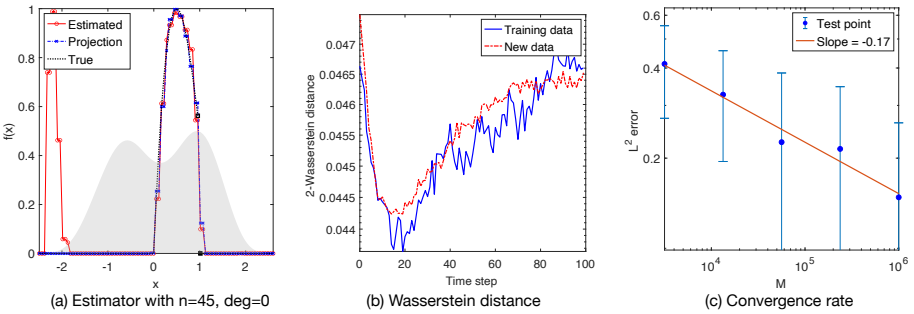


FIGURE 4. Learning results of Arch function f_3 with model (4.2).

agreement with the sampling error since we used 10^6 samples. Fig. 2(c) shows that the convergence rate of the $L^2(\bar{\rho}_T^L)$ error is 0.46. This rate is close to the minimax rate $\frac{2}{5} = 0.4$.

Sine-Cosine function: Fig. 3(a) shows the estimator with degree-2 B-spline basis with dimension $n = 13$. The $L^2(\bar{\rho}_T^L)$ error is 0.1596 and the relative error is 9.90%. Fig. 3(b) shows that the Wasserstein distances are at the scale of 10^{-2}. Fig. 3(c) shows that the convergence rate of the $L^2(\bar{\rho}_T^L)$ error is 0.26, less than the classical minimax rate $\frac{3}{7} \approx 0.42$. Note also that the variance of the L^2 error does not decrease as M increases. In comparison with the results for f_1 in Fig.2(a), we attribute this relatively low convergence rate and the large variance to the high-frequency component $\cos(6x)$, which is harder to identify from moments than the low frequency component $\sin(x)$.

Arch function: Fig. 4(a) shows the estimator with degree-0 B-spline basis with dimension $n = 45$. The $L^2(\bar{\rho}_T^L)$ error is 0.0645 and the relative error is 14.44%. Fig. 4(b) shows that the Wasserstein distances are small, at the scale 10^{-2}. Fig. 4(c) shows that the convergence rate of the $L^2(\bar{\rho}_T^L)$ error is 0.17, less than the would-be minimax rate $\frac{1}{3} \approx 0.33$.

Arch function with observation noise: To demonstrate that our method can tolerate large observation noise, we present the estimation results from noisy observations of the Arch function, which is the most difficult among the three examples. Suppose that the observation noise ξ in (2.17) is iid $\mathcal{N}(0, 0.25)$. Note that the average of $\mathbb{E}\left[|Y_t|^2\right]$ is about 0.2, so the signal-to-noise ratio is rather small, at $\mathbb{E}[|Y|^2]/\mathbb{E}[\xi^2] \approx 0.8$. Nevertheless, our method can identify the function using the moments of the noise as discussed in Section 2.5. Fig. 5(a) shows the estimator with

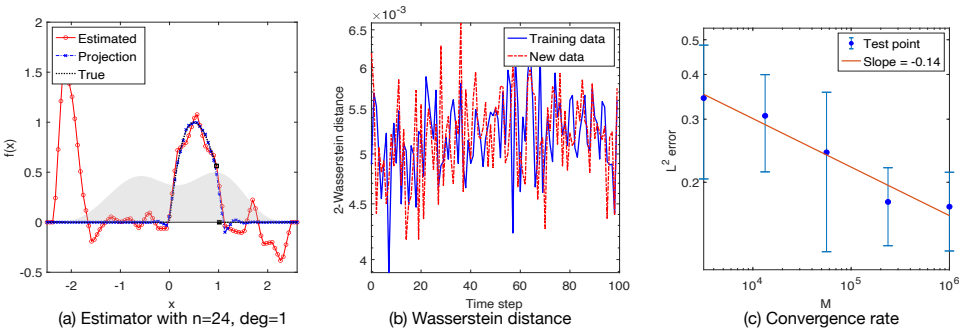


FIGURE 5. Learning results of Arch function f_3 with model (4.2) and i.i.d Gaussian observation noise.

degree-1 B-spline basis with dimension $n = 24$. The $L^2(\bar{\rho}_T^L)$ error is 0.1220 and the relative error is 27.32%. Fig. 5(b) shows that the Wasserstein distances are small, of order 10^{-3}. The Wasserstein distances are approximated from samples of the noisy data $Y = f_{true}(X) + \xi$ and of the noisy prediction $\widehat{Y} = \widehat{f}(X) + \xi$. Fig. 5(c) shows that the convergence rate of the $L^2(\bar{\rho}_T^L)$ error is 0.14. The estimation is not as good as the noise-free case, also because the noisy observation data lead to slightly lower and upper bound constraints in (2.15).

We consider this tolerance and robustness to noise to be quite surprising for such an ill-posed inverse problem, and the main reason for it is the use of moments, which average the noise so that the error occurs at scale $O(1/\sqrt{M})$.

We have also tested piecewise constant observation functions. Our method has difficulty in identifying such functions, due to two issues: (i) the uniform partition often misses the jump discontinuities (even the projection of f_* has a large error); and (ii) the moments we considered depend on the observation function non-locally, thus, they provide limited information to identify the true function from its local perturbations. We leave it for future research to overcome these difficulties by searching the jump discontinuities and by introducing moments detecting local information.

4.3. **Limitations.** We demonstrate by examples the non-identifiability due to symmetry and stationarity.

Symmetric distribution. Let the state space model be the Brownian motion with initial distribution Unif$(0, 1)$. The state process (X_t) has a distribution that is symmetric with respect to the line $x = \frac{1}{2}$, i.e., the processes (X_t) and $(1 - X_t)$ have the same distribution. Thus, with the reflection function $R(x) = 1 - x$, the processes $f(X_t)$ and $f \circ R(X_t)$ have the same distribution, and the observation data does not provide information for distinguishing f from $f \circ R$. The loss functional (2.4) has at least two minima.

Figure 6 shows that our algorithm finds the reflection of the true function $f_* = \sin(x)$. The hypothesis space $\mathcal{H}$ has B-spline basis functions with degree 2 and dimension 58. Our estimator is close to $f_* \circ R(x) = \sin(1 - x)$. Its $L^2(\bar{\rho}_T^L)$ error is 1.1244 and its reflection's $L^2(\bar{\rho}_T^L)$ error is 0.0790. Both the estimator and its reflection correctly predict the distribution of the observation process (Y_t).

Stationary process. When the diffusion process (X_t) is stationary, the loss functional (2.4) provides limited information about the observation function. As discussed in Section 3.2, the matrix $\overline{A}_1$ has rank 1, and $\mathcal{E}_2 = 0$ and $\mathcal{E}_3 = 0$ lead to only two

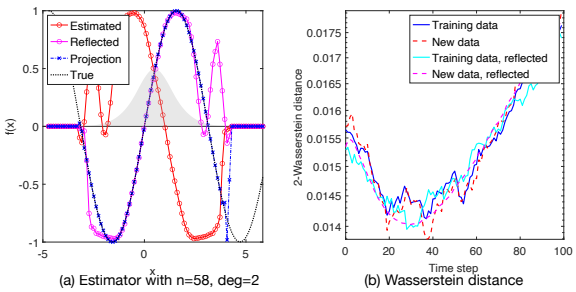


FIGURE 6. Learning results of $f_*(x) = \sin(x)$ with the state space model being $X_t = B_t + X_0$ where $X_0 \sim \text{Unif}(0,1)$. Due to the symmetry with respect to the line $x = \frac{1}{2}$, the estimator $\widehat{f}(x)$ and its reflection $\widehat{f}(1-x)$ are indistinguishable by the loss functional and they lead to similar prediction of the distribution of $\{Y_{t_l}\}$.

more constraints. The constraints from the upper and lower bounds in (2.15) play a major role in leading to a minimizer at the boundary of the convex set $\mathcal{H}$.

Figure 7 shows the learning results with the stationary Ornstein-Uhlenbeck process $dX_t = -X_t dt + dB_t$ and with the observation function $f_*(x) = \sin(x)$. The stationary density of (X_t) is $\mathcal{N}(0, \frac{1}{2})$. Due to the limited information, the estimator has a large $L^2(\bar{\rho}_T^L)$ error, which is 0.2656 and its prediction has large 2-Wasserstein distances oscillating near 0.1290.

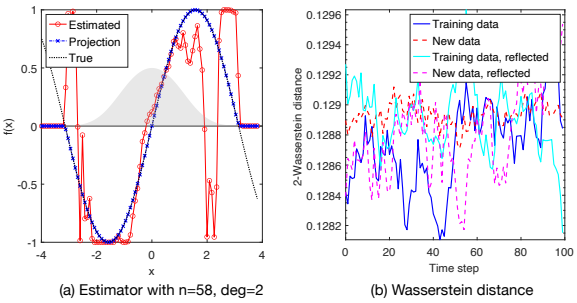


FIGURE 7. Learning results of $f_*(x) = \sin(x)$ with stationary Ornstein-Uhlenbeck process. Due to limited information from the moments, the estimator is inaccurate.

5. **Discussions and conclusion.** We have proposed a nonparametric learning method to estimate the observation functions in nonlinear state space models. It matches the generalized moments via constrained regression. The algorithm is suitable for large sets of unlabeled data. Moreover, it can deal with challenging cases when the observation function is non-invertible. We address the fundamental issue of identifiability from first-order moments. We show that the function spaces of identifiability are the closure of RKHS spaces intrinsic to the state space model. Numerical examples show that the first two moments and temporal correlations, along with upper and lower bounds, can identify functions ranging from piecewise polynomials to smooth functions and tolerate considerable observation noise. The limitations of this method, such as non-identifiability due to symmetry and stationarity, are also discussed.

This study provides a first step in the unsupervised learning of latent dynamics from abundant unlabeled data. Several directions are calling for further exploration: (i) a mixture of unsupervised and supervised learning that combines unlabeled data with limited labeled data, particularly for high-dimensional functions; (ii) enlarging the function space of learning, either by construction of more first-order generalized moments or by designing experiments to collect more informative data; (iii) joint estimation of the observation function and the state space model.

Appendix A. **A review of RKHS.** We review the definitions and properties of the positive definite functions, the Mercer kernel, the reproducing kernel Hilbert space (RKHS), and the related integral operator, see e.g., [7] for them on a compact domain [35] for them on a non-compact domain.

Positive definite functions. The following is a real-variable version of the definition of positive definite functions in [1, p.67].

Definition A.1 (Positive definite function)**.** Let X be a nonempty set. A function $G : X \times X \to \mathbb{R}$ is positive definite if and only if it is symmetric (i.e. $G(x, y) = G(y, x)$) and $\sum_{j,k=1}^{n} c_j c_k G(x_j, x_k) \geqslant 0$ for all $n \in \mathbb{N}$, $\{x_1, \ldots, x_n\} \subset X$ and $\mathbf{c} = (c_1, \ldots, c_n) \in \mathbb{R}^n$. The function ϕ is strictly positive definite if the equality hold only when $\mathbf{c} = \mathbf{0} \in \mathbb{R}^n$.

Theorem A.2 (Properties of positive definite kernels)**.** *Suppose that $k, k_1, k_2 : X \times X \subset \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ are positive definite kernels. Then*

(a) $k_1 k_2$ is positive definite. ([1, p.69])
(b) Inner product $\langle u, v\rangle = \sum_{j=1}^{d} u_j v_j$ is positive definite ([1, p.73])
(c) $f(u)f(v)$ is positive definite for any function $f : X \to \mathbb{R}$ ([1, p.69]).

RKHS and positive integral operators. Let (X, d) be a metric space and $G : X \times X \to \mathbb{R}$ be continuous and symmetric. We say that G is a Mercer kernel if it is positive definite (as in Definition A.1). The RKHS $\mathcal{H}_G$ associated with G is defined to be closure of $\text{span}\{G(x, \cdot) : x \in X\}$ with the inner product

$$\langle f, g\rangle_{\mathcal{H}_G} = \sum_{i=1,j=1}^{n,m} c_i d_j G(x_i, y_j)$$

for any $f = \sum_{i=1}^{n} c_i G(x_i, \cdot)$ and $g = \sum_{j=1}^{m} d_j G(y_j, \cdot)$. It is the unique Hilbert space such that: (1) the linear space $\text{span}\{G(\cdot, y), y \in X\}$ is dense in it; (2) it has the reproducing kernel property in the sense that for all $f \in \mathcal{H}_G$ and $x \in X$, $f(x) = \langle G(x, \cdot), f\rangle_G$ (see [7, Theorem 2.9]).

By means of the Mercer Theorem, we can characterize the RKHS $\mathcal{H}_G$ through the integral operator associated with the kernel. Let μ be a nondegenerate Borel measure on (X, d) (that is, $\mu(U) > 0$ for every open set $U \subset X$). Define the integral operator L_G on $L^2(X, \mu)$ by

$$L_G f(x) = \int_X G(x, y) f(y) d\mu(y).$$

The RKHS has the operator characterization (see e.g., [7, Section 4.4] and [35]):

Theorem A.3. *Assume that G is a Mercer kernel and $G \in L^2(X \times X, \mu \otimes \mu)$. Then*

1. *L_G is a compact positive self-adjoint operator. It has countably many positive eigenvalues $\{\lambda_i\}_{i=1}^{\infty}$ and corresponding orthonormal eigenfunctions $\{\phi_i\}_{i=1}^{\infty}$.*

Note that when zero is an eigenvalue of L_G, the linear space $H = \text{span}\{\phi_i\}_{i=1}^{\infty}$ is a proper subspace of $L^2(\mu)$.

2. *$\{\sqrt{\lambda_i}\phi_i\}_{i=1}^{\infty}$ is an orthonormal basis of the RKHS $\mathcal{H}_G$.*
3. *The RKHS is the image of the square root of the integral operator, i.e., $\mathcal{H}_G = L_G^{1/2}(L^2(X,\mu))$.*

Appendix B. Algorithm details.

B.1. **B-spline basis and dimension of the hypothesis space.** The choice of hypothesis space is important for the nonparametric regression. One can use global basis functions such as polynomials or Fourier basis when the observation function is known in prior to be smooth. On the other hand, when the observation function may be discontinuous, local basis functions such as B-splines or wavelets improve the estimation. In all our numerical experiments we choose the basis functions to be the B-splines, as we assume only limited information about the observation function. To select an optimal dimension of the hypothesis space, we introduce a new algorithm to estimate the range for the dimension and then we select the optimal dimension that minimizes the 2-Wasserstein distance between the measures of data and prediction.

B-Spline basis functions. We briefly review the definition of B-spline basis functions and we refer to [30, Chapter 2] and [27] for details. Given a nondecreasing sequence of real numbers, called knots, $(r_0, r_1, \ldots, r_m)$, the B-spline basis functions of degree p, denoted by $\{N_{i,p}\}_{i=0}^{m-p-1}$, are defined recursively as

$$N_{i,0}(r) = \begin{cases} 1, & r_i \leqslant r < r_{i+1} \\ 0, & \text{otherwise} \end{cases},$$

$$N_{i,p}(r) = \frac{r - r_i}{r_{i+p} - r_i} N_{i,p-1}(r) + \frac{r_{i+p+1} - r}{r_{i+p+1} - r_{i+1}} N_{i+1,p-1}(r).$$

Each function $N_{i,p}$ is a nonnegative local polynomial of degree p, supported on $[r_i, r_{i+p+1}]$. At a knot with multiplicity k, it is $p-k$ times continuously differentiable. Hence, the differentiability increases with the degree but decreases when the knot multiplicity increases. The basis satisfies a partition unity property: for each $r \in [r_i, r_{i+1}]$, $\sum_j N_{j,p}(r) = \sum_{j=i-p}^{i} N_{j,p}(r) = 1$.

We set the knots of the spline functions to be a uniform partition of $[R_{min}, R_{max}]$ (the support of the measure $\bar{\rho}_T^L$ in (2.16)) $R_{min} = r_0 \leqslant r_1 \leqslant \cdots \leqslant r_m = R_{min}$. For any choice of degree p, we set the basis functions of the hypothesis space $\mathcal{H}$, contained in a subspace with dimension $n = m - p$, to be

$$\phi_i(r) = N_{i,p}(r), \ i = 0, \ldots, m-p-1.$$

Thus, the basis functions $\{\phi_i\}$ are piecewise degree-p polynomials with knots adaptive to $\bar{\rho}_T^L$.

Dimension of the hypothesis space. The choice of dimension n of $\mathcal{H}$ is important to avoid under- and over-fitting: we choose it by minimizing the 2-Wasserstein distance between the empirical distributions of observed process (Y_t) and that predicted by our estimated observation function. To reduce the computational burden, we proceed in 2 steps: first we determine a rough range for n, and then within this range we select the dimension with the minimal Wasserstein distance.

Step 1: we introduce an algorithm, called *Cross-validating Estimation of Dimension Range (CEDR)*, to estimate the range $[1, N]$ for the dimension of the hypothesis space, based on the quadratic loss functional $\mathcal{E}_1$. Its main idea is to restrict N to

avoid overly amplifying the estimator's sampling error, which is estimated by splitting the data into two sets. It incorporates the function space of identifiability in Section 3.1 into the SVD analysis [9, 16] of the normal matrix and vector from $\mathcal{E}_1$.

The CEDR algorithm estimates the sampling error in the minimizer of loss functional $\mathcal{E}_1$ through SVD analysis in three steps. First, we compute the normal matrix $\overline{A}_1$ and vector $\overline{b}_1$ in (2.6) by Monte Carlo; to estimate the sampling error in $\overline{b}_1$, we compute two copies, b and b', of $\overline{b}_1$ from two halves of the data:

$$\begin{aligned} b(i) &= \frac{1}{L}\sum_{l=1}^{L}\mathbb{E}\left[\phi_i(X_{t_l})\right]\frac{2}{M}\sum_{m=1}^{\lfloor\frac{M}{2}\rfloor}Y_{t_l}^{(m)}, \\ b'(i) &= \frac{1}{L}\sum_{l=1}^{L}\mathbb{E}\left[\phi_i(X_{t_l})\right]\frac{2}{M}\sum_{m=\lfloor\frac{M}{2}\rfloor+1}^{M}Y_{t_l}^{(m)}. \end{aligned} \tag{B.1}$$

Second, we implement an eigen-decomposition to find an orthonormal basis of $L^2(\bar{\rho}_T^L)$, the default function space of learning. The matrix $\overline{A}_1$ is a representation of the integral operator L_{K_1} in Lemma 3.4 on $\mathcal{H} = \operatorname{span}\{\phi_i\}_{i=1}^n$, and L_{K_1}'s eigenvalues are solved by the generalized eigenvalue problem

$$\overline{A}_1 u = \lambda B u, \quad \text{where } B = (\langle\phi_i,\phi_j\rangle_{L^2(\bar{\rho}_T^L)}) \tag{B.2}$$

(see [21, Theorem 5.1]). Denote the eigen-pairs by $\{\sigma_i, u_i\}$, where the eigenvalues $\{\sigma_i\}$ are non-increasingly ordered and the eigenvectors are subject to normalization $u_i^\top B u_j = \delta_{i,j}$. Thus, we have $\overline{A}_1 = \sum_{i=1}^n \sigma_i u_i u_i^\top$ (assuming that all σ_i's are positive; otherwise, we drop those zero eigenvalues). The least-squares estimators from b and b' are $c = \sum_{i=1}^n \frac{u_i^\top b}{\sigma_i}u_i$ and $c' = \sum_{i=1}^n \frac{u_i^\top b'}{\sigma_i}u_i$, respectively. Third, the difference between their function estimators represents the sampling error (with $\Delta c = c - c'$)

$$\begin{aligned} g(n) :=& \|\widehat{f} - \widehat{f}'\|^2_{L^2(\bar{\rho}_T^L)} = \|\sum_{k=1}^n \Delta c_k \phi_k\|^2_{L^2(\bar{\rho}_T^L)} = \sum_{i,j=1}^n \Delta c_i \langle\phi_i,\phi_j\rangle_{L^2(\bar{\rho}_T^L)}\Delta c_j = \Delta c^\top B \Delta c \\ &= \sum_{i,j=1}^n \frac{u_i^\top(b-b')}{\sigma_i} u_i^\top B u_j \frac{u_j^\top(b-b')}{\sigma_j} = \sum_{i=1}^n r_i^2, \end{aligned} \tag{B.3}$$

where $r_i = \frac{|u_i^\top(b-b')|}{\sigma_i}$. The ratio r_i is in the same spirit as the *Picard projection ratio* $\frac{|u_i^\top b|}{\sigma_i}$ in [16], which is used to detect overfitting. Note that the eigenvalues σ_i will vanish as n increases because the operator L_{K_1} is compact. Clearly, the sampling error $g(n)$ should be less than $\|f_*\|^2_{L^2(\bar{\rho}_T^L)}$, which is the average of the second moments. Thus, we set N to be

$$N = \max\{k \geqslant 1 : g(k) \leqslant \tau\}, \text{ where } \tau = \frac{1}{LM}\sum_{l=1,m=1}^{L,M}|Y_{t_l}^{(m)}|^2. \tag{B.4}$$

We note that this threshold is relatively large, neglecting the rich information in g, a subject worthy of further investigation.

Algorithm 2 summarizes the above procedure.

Input: The state space model and data $\{Y^{(m)}_{t_0:t_L}\}_{m=1}^M$.
Output: A range $[1, N]$ for the dimension of the hypothesis space for further selection.
1: Estimate the empirical density $\bar{\rho}_T$ in (2.16) and find its support $[R_{min}, R_{max}]$.
2: Set $n = 1$ and $g(n) = 0$. Estimate the threshold τ in (B.4).
3: **while** $g(n) \leqslant \tau$ **do**
4: Set $n \leftarrow n + 1$. Update the basis functions, Fourier or B-spline, as in Section 2.3.
5: Compute normal matrix $\overline{A}_1$ in (2.6) by Monte Carlo. Also, compute b and b' in (B.1).
6: Eigen-decomposition of $\overline{A}_1$ as in (B.2); return $\overline{A}_1 = \sum_{i=1}^n u_i \sigma_i u_i^T$ with $u_i^\top B u_j = \delta_{i,j}$.
7: Compute the Picard projection ratios: $r_i = \frac{|u_i^\top (b-b')|}{\sigma_i}$ for $i = 1, \ldots, n$ and $g(n) = \sum_{i=1}^n r_i^2$.
8: Return $N = n$.

ALGORITHM 2. Cross-validating Estimation of Dimension Range (CEDR) for hypothesis space

Step 2: We select the dimension n and degree for B-spline basis functions to be the one with the smallest 2-Wasserstein distance between the distribution of the data and that of the predictions. More precisely, let $\mu^f_{t_l}$ and $\mu^{\hat{f}}_{t_l}$ denote the distributions of $Y_{t_l} = f(X_{t_l})$ and of $\widehat{f}(X_{t_l})$, respectively. Let F_{t_l} and $\widehat{F}_{t_l}$ denote their cumulative distribution functions (CDF), with $F^{-1}_{t_l}$ and $\widehat{F}^{-1}_{t_l}$ being their inverses. We compute F_{t_l} from the data and $\widehat{F}_{t_l}$ from independent simulations, approximate their inverses using quantiles, and consider the root mean squared 2-Wasserstein distance

$$\left(\frac{1}{L}\sum_{l=1}^{L} W_2(\mu^f_{t_l}, \mu^{\hat{f}}_{t_l})^2\right)^{1/2}, \text{ with } W_2(\mu^f_{t_l}, \mu^{\hat{f}}_{t_l})^2 = \int_0^1 (F^{-1}_{t_l}(r) - \widehat{F}^{-1}_{t_l}(r))^2 dr\,. \quad \text{(B.5)}$$

This method of computing the Wasserstein distance is based on an observation in [5], and it has been used in [20, 29]. Recall that the 2-Wasserstein distance $W_2(\mu, \nu)$ of two probability measures μ and ν over Ω with finite second order moments is defined as $W_2(\mu,\nu) := \inf_{\gamma\in\Gamma(\mu,\nu)} \left(\int_{\Omega\times\Omega} |x-y|^2 d\gamma(x,y)\right)^{1/2}$, where $\Gamma(\mu,\nu)$ denotes the set of all measures on $\Omega\times\Omega$ with μ and ν as marginals. Let F and G be the CDFs of μ and ν respectively, and let F^{-1} and G^{-1} be their quantile functions. Then the L^2 distance of the quantile functions $d_2(\mu,\nu) := \left(\int_0^1 |F^{-1}(r) - G^{-1}(r) dr|^2\right)^{1/2}$ is equal to the 2-Wasserstein distance $W_2(\mu,\nu)$.

B.2. **Optimization with multiple initial conditions.** With the convex hypothesis space in (2.15), the minimization in (2.12) is a constrained optimization problem and it may have multiple local minima. Note that the loss functional $\mathcal{E}^M(c)$ in (2.12) consists of a quadratic term and two quartic terms. The quadratic term, which represents $\mathcal{E}^M_1$ in (2.5), has a Hessian matrix $\overline{A}_1$ which is often not full rank because it is the average of rank-one matrices by its definition (2.6), in which case the quadratic term has a valley of minima in the kernel of $\overline{A}_1$. The two quartic terms have valleys of minima at the intersections of the ellipse-shaped manifolds $\{c \in \mathbb{R}^n : c^\top A_{k,l} c = b^M_{k,l}\}_{l=1}^L$ for $k = 2, 3$. Symmetry in the distribution of the state process will also lead to multiple minima (see Section 3.2 for more discussions, and the numerical examples).

To reduce the possibility of obtaining a local minimum, we search for a minimizer from multiple initial conditions. We consider the following initial conditions: (1) the least squares estimator for the quadratic term; (2) the minimizer of the quadratic term in the hypothesis space, which is solved by least squares with linear constraints using ©MATLAB function lsqlin, starting from the LSE estimator; (3) the minimizers of the quartic terms over the hypothesis space, which is found by constrained optimization through ©MATLAB fmincon with the interior-point search. Among the minimizers obtained from these initial conditions, we finally take the one leading to the smallest 2-Wasserstein distance.

Appendix C. **Selection of dimension and degree of the B-spline basis.** We demonstrate the selection of the dimension and degree of the B-spline basis functions of the hypothesis space. As described in Section 2.3, we select the dimension and degree in two steps: we first select a rough range for the dimension by the Cross-validating Estimation of Dimension Range (CEDR) algorithm; then we pick the dimension and degree to be the ones with minimal 2-Wasserstein distance between the true and estimated distribution of the observation processes.

The CEDR algorithm helps to reduce the computational cost by estimating the dimension range for the hypothesis space. It is based on an SVD analysis of the normal matrix $\overline{A}_1$ and vector $\overline{b}_1$ from the quadratic loss functional $\mathcal{E}_1$. The key idea is to control the sampling error's effect on the estimator in the metric of the function space of learning. The sampling error is estimated by computing two copies of the normal vector through splitting the data into two halves. The function space of learning plays an important role here: it directs us to use a generalized eigenvalue problem for the SVD analysis. This is different from the classical SVD analysis in [16], where the information of the function space is neglected.

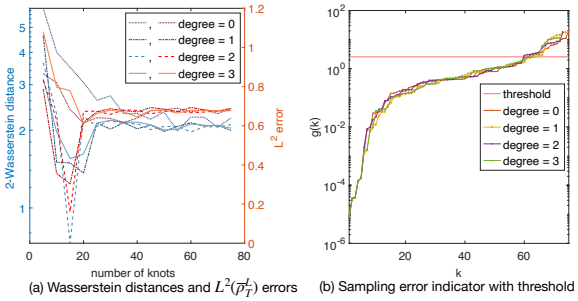


FIGURE 8. The selection of the dimension and the degree of B-spline basis functions in the case of Sine-Cosine function. In (a), the 2-Wasserstein distance reaches minimum among all cases when the degree is 2 and the knot number is 15, at the same time as the $L^2(\bar{\rho}_T^L)$ error reaches the minimum. Figure (b) shows the cross-validating error indicator g (defined in (B.3)) for selecting the dimension range N, suggesting an upper bound $N = 60$ with the threshold.

Figure 8 shows the dimension selection by 2-Wasserstein distances and by the CEDR algorithm for the example of sine-cosine function. To confirm the effectiveness of our CEDR algorithm, we compute the 2-Wasserstein distances for all dimensions in (a), side-by-side with the CEDR sampling error indicator g in (b) with relatively large dimensions $\{n = 75 - deg|$ for $deg \in \{0, 1, 2, 3\}$. First, the left figure suggests that the optimal dimension and degree are $n = 13$ and $deg = 2$, where

the 2-Wasserstein distance reaches minimum among all cases, and at the same time as the $L^2(\bar{\rho}_T^L)$ error. For the other degrees, the minimum 2-Wasserstein distances are either reached before of after the $L^2(\bar{\rho}_T^L)$ error. Thus, the 2-Wasserstein distance correctly selects the optimal dimension and degree for the hypothesis space. Second, (a) shows that the CEDR algorithm can effectively select the dimension range. With the threshold in (B.4) being $\tau = 1.60$, which is relatively large (representing a tolerance of 100% relative error), the dimension upper bounds are around $N = 60$ for all degrees, and the ranges encloses the optimal dimensions selected by the 2-Wasserstein distance in (b).

Here we used a relatively large threshold for a rough estimation of the range of dimension. Clearly, our cross-validating error indicator $g(k)$ in (B.3) provides rich information about the increase of sampling error as the dimension increases. A future direction is to extract the information, along with the decay of the integral operator, to control, both in theory and algorithmically, the trade-off between sampling error and approximation error.

Acknowledgments. The authors would like to thank the editor and the two anonymous reviewers for the helpful and constructive comments.

REFERENCES

[1] C. Berg, J. P. R. Christensen and P. Ressel, *Harmonic Analysis on Semigroups: Theory of Positive Definite and Related Functions*, volume 100. New York: Springer, 1984.
[2] S. A. Billings, *Nonlinear System Identification*, John Wiley & Sons, Ltd, Chichester, UK, 2013.
[3] P. J. Brockwell and R. A. Davis, *Time Series: Theory and Methods*, Springer, New York, 2nd edition, 1991.
[4] O. Cappé, E. Moulines and T. Rydén, *Inference in Hidden Markov Models*, Springer Series in Statistics. Springer, New York ; London, 2005.
[5] J. A. Carrillo and G. Toscani, Wasserstein metric and large–time asymptotics of nonlinear diffusion equations, In *New Trends in Mathematical Physics: In Honour of the Salvatore Rionero 70th Birthday*, 234-244. World Scientific, 2004.
[6] R. R. Coifman, S. Lafon, A. B. Lee, M. Maggioni, B. Nadler, F. Warner and S. W. Zucker, Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps, *Proceedings of the National Academy of Sciences of the United States of America*, **102** (2005), 7426-7431.
[7] F. Cucker and D.-X. Zhou, *Learning Theory: An Approximation Theory Viewpoint*, volume 24. Cambridge University Press, 2007.
[8] J. Fan and Q. Yao, *Nonlinear Time Series: Nonparametric and Parametric Methods*, Springer, New York, NY, 2003.
[9] R. D. Fierro, G. H. Golub, P. C. Hansen and D. P. O'Leary, Regularization by truncated total least squares, *SIAM J. Sci. Comput.*, **18** (1997), 1223-1241.
[10] A. Friedman, Stochastic differential equations and applications, In *Stochastic Differential Equations*, 75-148. Springer, 2010.
[11] C. Gelada, S. Kumar, J. Buckman, O. Nachum and M. G. Bellemare, DeepMDP: Learning continuous latent space models for representation learning, arXiv:1906.2736, Cs Stat, 2019.
[12] A. Ghosh, S. Mukhopadhyay, S. Roy and S. Bhattacharya, Bayesian inference in nonparametric dynamic state space models, *Statistical Methodology*, **21** (2014), 35-48.
[13] N. Guglielmi and E. Hairer, Classification of hidden dynamics in discontinuous dynamical systems, *SIAM J. Appl. Dyn. Syst.*, **14** (2015), 1454-1477.
[14] L. Györfi, M. Kohler, A. Krzyzak and H. Walk, *A Distribution-Free Theory of Nonparametric Regression*, Springer Science & Business Media, 2006.
[15] D. Hafner, T. Lillicrap, I. Fischer, R. Villegas, D. Ha, H. Lee and J. Davidson, Learning Latent dynamics for planning from pixels, arXiv:1811.4551, Cs Stat, 2019.

[16] P. C. Hansen, The L-curve and its use in the numerical treatment of inverse problems, In *in Computational Inverse Problems in Electrocardiology, ed. P. Johnston, Advances in Computational Bioengineering*, 119-142. WIT Press, 2000.
[17] M. R. Jeffrey, *Hidden Dynamics: The Mathematics of Switches, Decisions and Other Discontinuous Behaviour*, Springer International Publishing, Cham, 2018.
[18] L. Kaiser, M. Babaeizadeh, P. Milos, B. Osinski, R. H. Campbell, K. Czechowski, D. Erhan, C. Finn, P. Kozakowski, S. Levine, A. Mohiuddin, R. Sepassi, G. Tucker and H. Michalewski, Model-Based Reinforcement Learning for Atari, arXiv:1903.0374, Cs Stat, 2020.
[19] N. Kantas, A. Doucet, S. S. Singh and J. M. Maciejowski, An overview of sequential Monte Carlo methods for parameter estimation in general state-space models, *IFAC Proc. Vol.*, **42** (2009), 774-785.
[20] N. Kolbe, Wasserstein distance, https://github.com/nklb/wasserstein-distance, 2020.
[21] Q. Lang and F. Lu, Identifiability of interaction kernels in mean-field equations of interacting particles, arXiv preprint, arXiv:2106.05565, 2021.
[22] K. Law, A. Stuart and K. Zygalakis, *Data Assimilation: A Mathematical Introduction*, Springer, 2015.
[23] Z. Li, F. Lu, M. Maggioni, S. Tang and C. Zhang, On the identifiability of interaction functions in systems of interacting particles, *Stochastic Processes and their Applications*, **132** (2021), 135-163.
[24] L. Ljung, System identification, In *Signal Analysis and Prediction*, 163-173. Springer, 1998.
[25] F. Lu, Q. Lang and Q. An, Data adaptive RKHS Tikhonov regularization for learning kernels in operators, arXiv preprint, arXiv:2203.03791, 2022.
[26] F. Lu, M. Zhong, S. Tang and M. Maggioni, Nonparametric inference of interaction laws in systems of agents from trajectory data, *Proc. Natl. Acad. Sci. USA*, **116** (2019), 14424-14433.
[27] T. Lyche, C. Manni and H. Speleers, Foundations of spline theory: B-splines, spline approximation, and hierarchical refinement, *Splines and PDEs: From Approximation Theory to Numerical Linear Algebra*, volume 2219, Springer International Publishing, Cham, 2018, 1-76.
[28] C. Moosmüller, F. Dietrich and I. G. Kevrekidis, A geometric approach to the transport of discontinuous densities, arXiv:1907.8260, Phys. Stat, 2019.
[29] V. M. Panaretos and Y. Zemel, Statistical aspects of wasserstein distances, *Annual Review of Statistics and its Application*, **6** (2019), 405-431.
[30] L. Piegl and W. Tiller, *The NURBS Book*, Monographs in Visual Communication, Springer Berlin Heidelberg, Berlin, Heidelberg, 1997.
[31] Y. Pokern, A. M. Stuart and P. Wiberg, Parameter estimation for partially observed hypoelliptic diffusions, *J. R. Stat. Soc. Ser. B Stat. Methodol.*, **71** (2009), 49-73.
[32] B. L. S. Prakasa Rao, Statistical inference from sampled data for stochastic processes, In N. U. Prabhu, editor, *Contemporary Mathematics*, volume 80, 249-284. American Mathematical Society, Providence, Rhode Island, 1988.
[33] A. Rahimi and B. Recht, Unsupervised regression with applications to nonlinear system identification, In *Advances in Neural Information Processing Systems*, (2007), 1113-1120.
[34] M. Sørensen, Estimating functions for diffusion-type processes, In *Statistical Methods for Stochastic Differential Equations*, volume 124, 1-107. Monogr. Statist. Appl. Probab, 2012.
[35] H. Sun, Mercer theorem for RKHS on noncompact sets, *Journal of Complexity*, **21** (2005), 337-349.
[36] A. Svensson and T. B. Schön, A flexible state–space model for learning nonlinear dynamical systems, *Automatica*, **80** (2017), 189-199.
[37] F. Tobar, P. M. Djuric and D. P. Mandic, Unsupervised state-space modeling using reproducing kernels, *IEEE Trans. Signal Process.*, **63** (2015), 5210-5221.
[38] F. X. F. Ye, S. Yang and M. Maggioni, Nonlinear model reduction for slow-fast stochastic systems near manifolds, 2021.

Received August 2022; revised December 2022; early access February 2023.