Unsupervised learning of observation functions in state-space models by nonparametric moment methods

Qingci An
(Joint work with Fei Lu)

Outline

Unsupervised learning

Data: Unlabeled input X, output Y.
Goal: Learn the function f that $Y = f(X)$.

Unsupervised learning

Data: Unlabeled input X, output Y.
Goal: Learn the function f that $Y = f(X)$.

Problem formulation

Consider the state-space model:

State model:	$dX_t = a(X_t)dt + b(X_t)dB_t,$	with a, b known;
Observation model:	$Y_t = f_{true}(X_t),$	with f_{true} unknown.

Data: $\{Y^{(m)}_{t_0:t_L}\}_{m=1}^M$.

Goal: Identify the observation function f_{true}.

Generalized moments method

Let $g : \mathbb{R}^{L+1} \to \mathbb{R}^K$ is a functional of the trajectory $Y_{t_0:t_L}$. We are matching the moments

$$\mathbb{E}\left[g(Y_{t_0:t_L})\right] \longleftrightarrow \mathbb{E}\left[g(f(X_{t_0:t_L}))\right].$$

$$\approx \frac{1}{M}\sum_{m=1}^{M} g(Y_{t_0:t_L}^{(m)})$$

$$=: E_M[g(Y_{t_0:t_L})]$$

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

$$\widehat{f} = \arg\min_{f\in\mathcal{H}} \mathcal{E}_M(f),$$

where $\mathcal{E}_M(f) := \mathrm{dist}\left(E_M[g(Y_{t_0:t_L})], \mathbb{E}\left[g(f(X_{t_0:t_L}))\right]\right)^2$

Proper choice of g

For efficient optimization, we select the functional g such that the moments $\mathbb{E}\left[g(f(X_{t_0:t_L}))\right]$ for $f = \sum_{i=1}^{n} c_i\phi_i$ can be efficiently evaluated for all $c = (c_1, \ldots, c_n)$.

For example,

1. $g_1(Y_{t_0:t_L}) = Y_{t_0:t_L}$;
2. $g_2(Y_{t_0:t_L}) = Y_{t_0:t_L}^2$;
3. $g_3(Y_{t_0:t_L}) = \left(Y_{t_0}Y_{t_1}, \cdots, Y_{t_{L-1}}Y_{t_L}\right)$;

Moments from Itô formula

For $f \in C_b^2$, i.e., 2nd-order differentiable with bounded derivatives, applying Itô formula:

$$df(X_t) = \nabla f \cdot b(X_t) dW_t + [\nabla f \cdot a + \frac{1}{2} Hess(f) : b^\top b](X_t) dt.$$

In integral form, it is

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

where the 2nd-order differential operator $\mathcal{L}$ is

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

Moments from Itô formula

For $Y_t = f_{true}(X_t)$ and $\Delta Y_t = Y_{t+\delta} - Y_t$, we have the following equalities for moments

$$\mathbb{E}[\Delta Y_t] = \mathbb{E}[\int_t^{t+\delta} Lf_{true}(X_s)ds] = \mathbb{E}[\mathcal{L}f_{true}(X_t)\delta] + O(\delta^2),$$

$$\mathbb{E}\left[Y_t \Delta Y_t\right] = \mathbb{E}\left[f_{true}(X_t)\int_t^{t+\delta} \mathcal{L}f_{true}(X_s)ds\right] = \mathbb{E}[f\mathcal{L}f_{true}(X_t)\delta] + O(\delta^2),$$

$$\begin{aligned}\mathbb{E}\left[(\Delta Y_t)^2\right] &= \mathbb{E}\left[|\int_t^{t+\delta} \nabla f_{true} b(X_s)|^2 ds\right] + \mathbb{E}\left[(\int_t^{t+\delta} \mathcal{L}f(X_s)ds)^2\right] \\ &+ \mathbb{E}[\int_t^{t+\delta} \mathcal{L}f(X_s)ds \int_t^{t+\delta} \nabla f b(X_s)dW_s] \\ &= \mathbb{E}[|\nabla f_{true} b(X_s)|^2\delta] + O(\delta^{1.5})\end{aligned}$$

1. $g_4(Y_{t_0:t_L}) = \left(Y_{t_1} - Y_{t_0}, \cdots, Y_{t_L} - Y_{t_{L-1}}\right)$;
2. $g_5(Y_{t_0:t_L}) = \left(Y_{t_1}(Y_{t_1} - Y_{t_0}), \cdots, Y_{t_L}(Y_{t_L} - Y_{t_{L-1}})\right)$;
3. $g_6(Y_{t_0:t_L}) = \left((Y_{t_1} - Y_{t_0})^2, \cdots, (Y_{t_L} - Y_{t_{L-1}})^2\right)$;

Loss functionals

The generalized moments we consider include the first and the second moments, as well as the one-step temporal correlation:

$$\mathcal{E}(f) := \underbrace{\lambda_1 \frac{1}{L}\sum_{l=1}^{L} \left|\mathbb{E}[f(X_{t_l})] - \mathbb{E}[Y_{t_l}]\right|^2}_{\mathcal{E}_1(f)} + \underbrace{\lambda_2 \frac{1}{L}\sum_{l=1}^{L} \left|\mathbb{E}[f(X_{t_l})^2] - \mathbb{E}[Y_{t_l}^2]\right|^2}_{\mathcal{E}_2(f)}$$

$$+ \underbrace{\lambda_3 \frac{1}{L}\sum_{l=1}^{L} \left|\mathbb{E}[f(X_{t_l})f(X_{t_{l-1}})] - \mathbb{E}[Y_{t_l}Y_{t_{l-1}}]\right|^2}_{\mathcal{E}_3(f)},$$

$\mathcal{E}^M(f) :=$ Replace $\mathbb{E}[Y_{t_l}]$, $\mathbb{E}[Y_{t_l}^2]$, $\mathbb{E}[Y_{t_l}Y_{t_{l-1}}]$ by their empirical mean.

Hypothesis space

$$\mathcal{H} = \{f = \sum_{i=1}^{n} c_i\phi_i : y_{min} \leqslant f(x) \leqslant y_{max} \text{ for all } x \in \text{supp}(\overline{\rho_T})\}.$$

While ideally y_{min} and y_{max} are bounds for f_{true}, these are typically unknown, and we use instead the empirical upper and lower bounds:

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

Estimator

The estimator from data is

$$\widehat{f}_{\mathcal{H},M} = \sum_{i=1}^{n} \widehat{c}_i \phi_i,$$

$$\widehat{c} = \underset{c\in\mathbb{R}^n \text{ s.t. } \sum_{i=1}^{n} c_i\phi_i\in\mathcal{H}}{\arg\min} \mathcal{E}^M(c).$$

Identifiability by quadratic loss functional

Definition (Identifiability)

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

Rewrite the quadratic loss function as:

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

where

- L_{K_1} is an integral operator with kernel K_1,
- $K_1(x, x') = \frac{1}{\bar{\rho}_T^L(x)\bar{\rho}_T^L(x')} \frac{1}{L} \sum_{l=1}^{L} p_{t_l}(x) p_{t_l}(x')$,
- p_{t_l} is the density of X_{t_l},
- $\bar{\rho}_T^L$ is the average density of p_{t_l}, $l = 0, \cdots, L$.

Inverse problem

$$\nabla \mathcal{E}_1(f) = L_{K_1}(f - f_{true})$$
$$f = "L_{K_1}^{-1}" L_{K_1}(f_{true})$$

1. When L_{K_1} is not strictly positive (eg, X_t stationary process), the inverse problem is *ill-defined* on $L^2(\bar{\rho}_T^L)$.
2. On the reproducing kernel hilbert space $\mathcal{H}_{K_1}$, the operator L_{K_1} is invertible but **unbounded**, the inverse problem is *ill-posed* on the RKHS.

Numerical examples: double-well potential

$$dX_t = (X_t - X_t^3)dt + dB_t$$

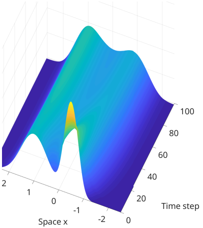
Three observation function $f(x)$ representing typical challenges: nearly invertible, non-invertible, non-invertible discontinuous:

Sine function: $f_1(x) = \sin(x);$

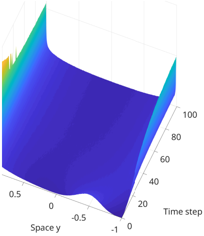
Sine-Cosine function: $f_2(x) = 2\sin(x) + \cos(6x);$

Arch function: $f_3(x) = \left(-2(1-x)^3 + 1.5(1-x) + 0.5\right) \mathbf{1}_{x\in[0,1]}.$

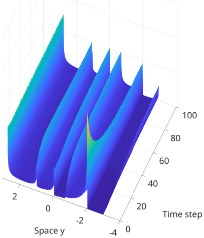
Numerical examples: double-well potential



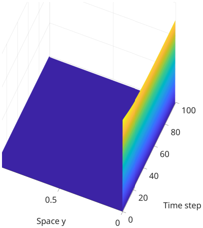
(a) Process (X_t), unobserved



(b) Process (Y_t) for f_{true} = sine function

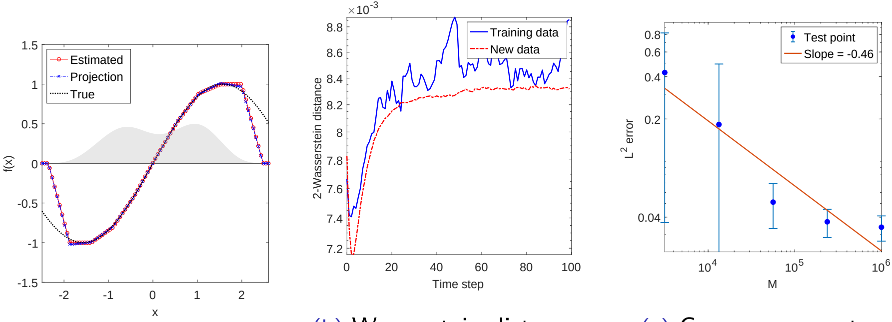


(c) Process (Y_t) for f_{true} = sine-cosine



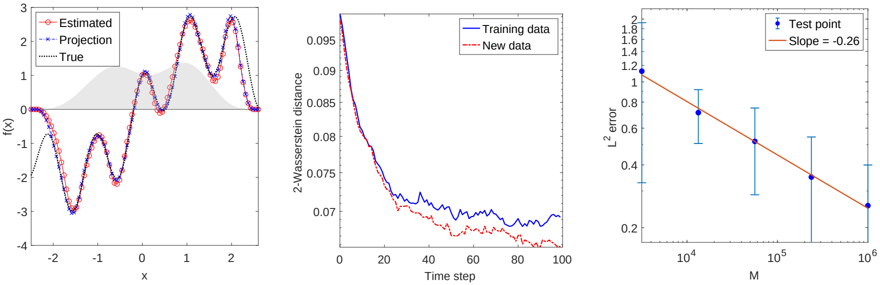
(d) Process (Y_t) for f_{true} = arch function

Estimation results



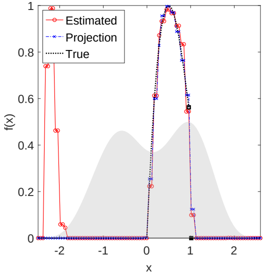
(a) Estimator (b) Wasserstein distance (c) Convergence rate

Figure: Learning results of Sine function $f_1(x) = \sin(x)$

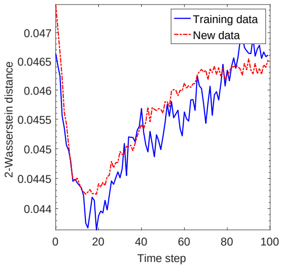


(a) Estimator (b) Wasserstein distance (c) Convergence rate

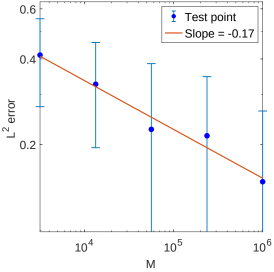
Figure: Learning results of Sine-Cosine function $f_2(x) = 2\sin(x) + \cos(6x)$



(a) Estimator



(b) Wasserstein distance



(c) Convergence rate

Figure: Learning results of Arch function f_3