

Kernel Embedding based Variational Approach for Low-dimensional Approximation of Dynamical Systems

Wenchong Tian ^{*}, Hao Wu [†]

August 10, 2020

Transfer operators such as Perron-Frobenius or Koopman operator play a key role in modeling and analysis of complex dynamical systems, which allow linear representations of nonlinear dynamics by transforming the original state variables to feature spaces. However, it remains challenging to identify the optimal low-dimensional feature mappings from data. The variational approach for Markov processes (VAMP) provides a comprehensive framework for the evaluation and optimization of feature mappings based on the variational estimation of modeling errors, but it still suffers from a flawed assumption on the transfer operator and therefore sometimes fails to capture the essential structure of system dynamics. In this paper, we develop a powerful alternative to VAMP, called kernel embedding based variational approach for dynamical systems (KVAD). By using the distance measure of functions in the kernel embedding space, KVAD effectively overcomes theoretical and practical limitations of VAMP. In addition, we develop a data-driven KVAD algorithm for seeking the ideal feature mapping within a subspace spanned by given basis functions, and numerical experiments show that the proposed algorithm can significantly improve the modeling accuracy compared to VAMP.

^{*}College of Environmental Science and Engineering, Tongji University, Shanghai 200092, P. R. China, E-mail: wenchong@tongji.edu.cn

[†]Correspond Author, School of Mathematical Sciences, Tongji University, Shanghai 200092, P. R. China, E-mail: hwu@tongji.edu.cn

1. Introduction

It has been shown that complex nonlinear processes can be accurately described by linear models in many science and engineering fields, including wireless communications [19, 56], molecular dynamics [54, 4, 55], fluid dynamics [28, 43] and control theory [1], where the linear models can be expressed by a unified formula

$$\mathbb{E}[\mathbf{f}(\mathbf{x}_{t+\tau})] = \mathbf{K}^\top \mathbb{E}[\mathbf{f}(\mathbf{x}_t)], \quad (1)$$

and the expectation operator can be removed for deterministic systems. In such models, the state variable \mathbf{x} is transformed into a feature space by the transformation $\mathbf{f}(\mathbf{x}) = (f_1(\mathbf{x}), \dots, f_m(\mathbf{x}))^\top$, and the dynamics with lag time τ is characterized by a linear time-invariant system in the feature space. Then, all the dynamical properties of the system can be quantitatively analyzed after estimating the transition matrix \mathbf{K} from data via linear regression. A special case of the linear models is Markov state models [39, 35] for conformational dynamics, which is equivalent to the well-known Ulam's method [8, 14]. In a Markov state model, the feature mapping \mathbf{f} consists of indicator functions of subsets of state and $\mathbf{K} = [K_{ij}]$ represents the transition probability from subset i to subset j . Besides Markov state models and the Ulam's method, a large number of similar modeling methods, e.g., dynamic mode decomposition [38, 2, 48, 21], time-lagged independent component analysis (TICA) [30, 34, 41], extended dynamic mode decomposition (EDMD) [50, 16, 40], Markov transition models [52], variational approach of conformation dynamics (VAC) [31, 32, 33], variational Koopman models [54] and their variants based on kernel embeddings [15, 16] and tensors [14, 33], are proposed by using different feature mappings.

From the perspective of operator theory, all the models in the form of (1) can be interpreted as algebraic representations of transfer operators of systems, including Frobenius-Perron (FP) operators and Koopman operators, and some of them are universal approximators for nonlinear dynamical systems under the assumption that the dimension of feature space is large enough [20] or the infinite-dimensional kernel mappings are utilized [45]. However, due to the limitation of computational cost and requirements of dynamical analysis, the low-dimensional approximation of transfer operators is still a critical and challenging problem in applications [17].

One common way to solve this problem is to identify the dominant dynamical structures, e.g., metastable states [9, 37], cycles [6] and coherent sets [11], and achieve the corresponding low-dimensional representations via spectral clustering. But this strategy assumes that an accurate high-dimensional model is known a priori, which is often violated especially for large-scale systems.

Another way for deterministic systems is to seek the feature mapping \mathbf{f} by minimizing the regression error of (1) under the constraint that the state variable can also be accurately reconstructed from \mathbf{f} [23, 24]. Notice that the constraint is necessary, otherwise a trivial but uninformative model with $\mathbf{f}(\mathbf{x}) \equiv 1$ and $\mathbf{K} = 1$ could be found. Some similar methods are developed for stochastic systems by considering (1) as a conditional generative model, where the parameters of \mathbf{f} can be trained based on the likelihood or the other statistical criteria [51, 25]. However, these methods are applicable only if \mathbf{f} are non-negative functions and usually involves the intractable probability density estimation.

In recent years, the variational approach has led to great progress for low-dimensional dynamical modeling, which was first proposed for time-reversible processes [31, 32, 27, 54] and extended to non-reversible processes in [53]. In contrast with the other methods, this approach provides a general and unified framework for data-driven model choice, reduction and optimization of dynamical systems based on the presented variational scores related to approximation errors of transfer operators. It can be easily integrated with deep learning to effectively analyze high-dimensional time series in an end-to-end manner [26, 3]. The existing variational principle based methods suffer from two drawbacks: First, it is necessary to assume that the transfer operator is Hilbert-Schmidt (HS) or compact as an operator between two weighted \mathcal{L}^2 spaces so that the maximum values of variational scores exist. But there is no easy way to test the assumption especially when we do not have strong prior knowledge regarding the system. Specifically, it can be proved that the assumption does not hold for most deterministic systems. Second, even for stochastic systems which satisfies the assumption, the common variational scores are possibly sensitive to small modeling variations, which could affect the effectiveness of the variational approach.

In this work, we introduce a kernel embedding based variational approach for dynamical systems (KVAD) using the theory of kernel embedding of functions [44, 46, 47, 45], where the modeling error is measured by using the distance between kernel embeddings of transition densities. The kernel based variational score in KVAD provides a robust and smooth quantification of differences between transfer operators, and is proved to be bounded for general dynamical systems, including deterministic and stochastic systems. Hence, it can effectively overcome the difficulties of existing variational methods, and expands significantly the range of applications. Like the previous variational scores, the kernel based score can also be consistently estimated from trajectory data without solving any intermediate statistical problem. Therefore, we develop a data-driven KVAD algorithm by considering \mathbf{f} as a linear superposition of a given set of basis functions. Furthermore, we establish a relationship between KVAD, diffusion maps[5] and the singular components of transfer operators. Finally, the effectiveness the proposed algorithm is demonstrated by numerical experiments.

2. Problem formulation and preliminaries

For a Markovian dynamical system in the state space $\mathbb{M} \subset \mathbb{R}^D$, its dynamics can be completely characterized by the transition density

$$p_\tau(\mathbf{x}, \mathbf{y}) \triangleq \mathbb{P}(\mathbf{x}_{t+\tau} = \mathbf{y} | \mathbf{x}_t = \mathbf{x}) \quad (2)$$

and the time evolution of the system state distribution can be formulated as

$$p_{t+\tau}(\mathbf{y}) = (\mathcal{P}_\tau p_t)(\mathbf{y}) \triangleq \int p_t(\mathbf{x}) p_\tau(\mathbf{x}, \mathbf{y}) d\mathbf{x}, \quad (3)$$

Here \mathbf{x}_t denotes the state of the system at time t and p_t is the probability density of \mathbf{x}_t . The transfer operator \mathcal{P}_τ is called the *Perron-Frobenius (PF) operator*¹, which is a linear but

¹Another commonly used transfer operator for Markovian dynamics is Koopman operator [49], which describes the evolution of observables instead of probability densities, and is the dual of the PF operator. In this paper, we focus only on the PF operator for convenience of analysis.

usually infinite-dimensional operator. Notice that the deterministic dynamics in the form of $\mathbf{x}_{t+\tau} = \Theta_\tau(\mathbf{x}_t)$ is a specific case of the Markovian dynamics, where $p_\tau(\mathbf{x}, \cdot) = \delta_{\Theta_\tau(\mathbf{x})}(\cdot)$ is a Dirac function centered at $\Theta_\tau(\mathbf{x})$, and the corresponding PF operator is given by

$$\int_{\mathbb{A}} (\mathcal{P}_\tau p_t)(\mathbf{y}) d\mathbf{y} = \int_{\mathbf{x} \in \Theta_\tau^{-1}(\mathbb{A})} p_t(\mathbf{x}) d\mathbf{x}.$$

By further assuming that the conditional distribution of $\mathbf{x}_{t+\tau}$ for given $\mathbf{x}_t = \mathbf{x}$ can always be represented by a linear combination of m density basis functions $\mathbf{q} = (q_1, \dots, q_m)^\top : \mathbb{M} \rightarrow \mathbb{R}^m$, we obtain a finite-dimensional approximation of the transition density:

$$\hat{p}_\tau(\mathbf{x}, \mathbf{y}) = \mathbf{f}(\mathbf{x})^\top \mathbf{q}(\mathbf{y}). \quad (4)$$

The feature mapping $\mathbf{f}(\mathbf{x}) = (f_1(\mathbf{x}), \dots, f_m(\mathbf{x}))^\top$ are real-valued observables of the state $\mathbf{x}_t = \mathbf{x}$, and provide a sufficient statistic for predicting the future states. Based on this approximation, the time evolution equation (3) of the state distribution can then be transformed into a linear evolution model of the feature functions \mathbf{f} in the form of (1) with the transition matrix

$$\mathbf{K} = \int \mathbf{q}(\mathbf{y}) \mathbf{f}(\mathbf{y})^\top d\mathbf{y}, \quad (5)$$

and many dynamical properties of the Markov system, including spectral components, coherent sets and the stationary distribution, can be efficiently from the linear model.

It is shown in [20] that Eq. (4) provides a universal approximator of Markovian dynamics if the set of basis function is rich enough. But in this paper, we focus on a more practically problem: *Given a small m , find \mathbf{f} and \mathbf{q} with $\dim(\mathbf{f}) = \dim(\mathbf{q}) = m$ such that the modeling error of (4) is minimized.*

2.1. Variational principle for Perron-Frobenius operators

We now briefly introduce the variational principle for evaluating the approximation quality of linear models (1). The detailed analysis and derivations can be found in [53].

For simplicity of notation, we assume that the available trajectory data are organized as

$$\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_N)^\top, \quad \mathbf{Y} = (\mathbf{y}_1, \dots, \mathbf{y}_N)^\top,$$

where $\{(\mathbf{x}_n, \mathbf{y}_n)\}_{n=1}^N$ are set of all transition pairs occurring in the given trajectories, and we denote the limits of empirical distributions of \mathbf{X}, \mathbf{Y} by ρ_0 and ρ_1 .

Due to the above analysis, the approximation quality of (4) can be evaluated by the difference between the PF operator $\hat{\mathcal{P}}_\tau$ deduced from $\hat{p}_\tau(\mathbf{x}, \mathbf{y})$ and the actual one. In the variational principle proposed by [53], \mathcal{P}_τ is considered as a mapping from $\mathcal{L}_{\rho_0}^2 = \{q \mid \|q\|_{\rho_0}^2 = \langle q, q \rangle_{\rho_0} < \infty\}$ to $\mathcal{L}_{\rho_1}^2 = \{q \mid \|q\|_{\rho_1}^2 = \langle q, q \rangle_{\rho_1} < \infty\}$ with inner products

$$\langle q, q' \rangle_{\rho_0} \triangleq \int q(\mathbf{x}) q'(\mathbf{x}) \rho_0(\mathbf{x})^{-1} d\mathbf{x}, \quad \langle q, q' \rangle_{\rho_1} \triangleq \int q(\mathbf{x}) q'(\mathbf{x}) \rho_1(\mathbf{x})^{-1} d\mathbf{x}.$$

From this insight, the Hilbert-Schmidt (HS) norm of the modeling error can be expressed as a weighted mean square error of conditional distributions

$$\|\hat{\mathcal{P}}_\tau - \mathcal{P}_\tau\|_{\text{HS}}^2 = \int \rho_0(\mathbf{x}) \|\hat{p}_\tau(\mathbf{x}, \cdot) - p_\tau(\mathbf{x}, \cdot)\|_{\rho_1^{-1}}^2 d\mathbf{x}, \quad (6)$$

and has the decomposition

$$\|\hat{\mathcal{P}}_\tau - \mathcal{P}_\tau\|_{\text{HS}}^2 = -\mathcal{R}[\mathbf{f}, \mathbf{q}] + \|\mathcal{P}_\tau\|_{\text{HS}}^2$$

with

$$\mathcal{R}[\mathbf{f}, \mathbf{q}] = \text{tr} \left(2\mathbb{E}_n [\mathbf{f}(\mathbf{x}_n)\mathbf{g}(\mathbf{y}_n)^\top] - \mathbb{E}_n [\mathbf{f}(\mathbf{x}_n)\mathbf{f}(\mathbf{x}_n)^\top] \mathbb{E}_n [\mathbf{g}(\mathbf{y}_n)\mathbf{g}(\mathbf{y}_n)^\top] \right)$$

for $\mathbf{q}(\mathbf{y}) = \mathbf{g}(\mathbf{y})\rho_1(\mathbf{y})$. Here $\mathbb{E}_n[\cdot]$ denotes the mean value over all transition pairs $\{(\mathbf{x}_n, \mathbf{y}_n)\}_{n=1}^N$ as $N \rightarrow \infty$. Because $\|\mathcal{P}_\tau\|_{\text{HS}}^2$ is a constant independent of modeling and \mathcal{R} can be easily estimated from data via empirical averaging, we can learn parametric models of $\mathbf{f}(\mathbf{x})$ and $\mathbf{g}(\mathbf{y})$ by maximizing \mathcal{R} as a variational score, which yields the variational approach for Markov processes (VAMP) [53].

It can be seen that the variational principle is developed under the assumption that $\mathcal{P}_\tau : \mathcal{L}_{\rho_0}^2 \rightarrow \mathcal{L}_{\rho_1}^2$ is an HS operator.² However, in many practical applications, it is difficult to justify the assumption for unknown transition densities. Particularly, for deterministic systems, this assumption does not hold and the maximization of \mathcal{R} could lead to unreasonable models.

Proposition 1. *For a deterministic system $\mathbf{x}_{t+\tau} = \Theta_\tau(\mathbf{x}_t)$, if $\mathcal{L}_{\rho_1}^2$ is an infinite-dimensional Hilbert space,*

1. \mathcal{P}_τ is not a compact operator from $\mathcal{L}_{\rho_0}^2$ to $\mathcal{L}_{\rho_1}^2$ and hence not an HS operator,
2. $\mathcal{R}[\mathbf{f}, \mathbf{q}]$ can be maximized by an arbitrary density basis $\mathbf{q} = (g_1 \cdot \rho_1, \dots, g_m \cdot \rho_1)^\top$ with $\mathbb{E}_n [\mathbf{g}(\mathbf{y}_n)\mathbf{g}(\mathbf{y}_n)^\top] = \mathbf{I}$ and $f_i(\mathbf{x}) = g_i(\Theta_\tau(\mathbf{x}))$.

Proof. See Appendix A. □

2.2. Kernel embedding of functions

Moving away from dynamical systems for a moment, here we introduce the theory of kernel embedding of functions [46, 45], which will be utilized to address the difficulty of VAMP in Section 3.

A kernel function $\kappa : \mathbb{M} \times \mathbb{M} \rightarrow \mathbb{R}$ is a symmetric and positive definite function, which implicitly defines a kernel mapping φ from \mathbb{M} to a reproducing kernel Hilbert space \mathbb{H} , and the inner product of \mathbb{H} satisfies the reproducing property

$$\langle \varphi(\mathbf{x}), \varphi(\mathbf{y}) \rangle_{\mathbb{H}} = \kappa(\mathbf{x}, \mathbf{y}).$$

²This assumption can be relaxed to compactness of \mathcal{P}_τ for some variants of the variational principle, but the relaxed assumption is not satisfied by deterministic systems either (see Proposition 1).

By using the kernel mapping, we can embed a function $q : \mathbb{M} \rightarrow \mathbb{R}$ in the Hilbert space \mathbb{H} as

$$\mathcal{E}q = \int \varphi(\mathbf{x})q(\mathbf{x})d\mathbf{x} \in \mathbb{H}.$$

Here \mathcal{E} is an injective mapping for $q \in \mathcal{L}^1(\mathbb{M})$ if κ is a universal kernel [47], and we can then measure the similarity between functions q and q' by the distance between $\mathcal{E}q$ and $\mathcal{E}q'$:

$$\begin{aligned} \|q - q'\|_{\mathcal{E}}^2 &= \langle \mathcal{E}(q - q'), \mathcal{E}(q - q') \rangle_{\mathbb{H}} \\ &= \iint \kappa(\mathbf{x}, \mathbf{y}) (q(\mathbf{x}) - q'(\mathbf{x})) (q(\mathbf{y}) - q'(\mathbf{y})) d\mathbf{x}d\mathbf{y}, \end{aligned}$$

where $\|q\|_{\mathcal{E}}^2 \triangleq \langle q, q \rangle_{\mathcal{E}}$ and $\langle q, q' \rangle_{\mathcal{E}} \triangleq \langle \mathcal{E}q, \mathcal{E}q' \rangle_{\mathbb{H}}$. The most commonly used universal kernel for $\mathbb{M} \subset \mathbb{R}^D$ is the Gaussian kernel $\kappa(\mathbf{x}, \mathbf{y}) = \exp(-\|\mathbf{x} - \mathbf{y}\|^2 / \sigma^2)$, where σ denotes the bandwidth of the kernel.

In the specific case where both q and q' are probability density functions, $\|q - q'\|_{\mathcal{E}}$ is called the maximum mean discrepancy (MMD) and can be estimated from samples of q, q' [45].

3. Theory

In this section, we develop a new variational principle for Markovian dynamics based on the kernel embedding of trainstion densities.

Assuming that $\kappa : \mathbb{M} \times \mathbb{M} \rightarrow \mathbb{R}$ is a universal kernel and bounded by B , $\mathcal{E}p_{\tau}(\mathbf{x}, \cdot)$ is also bounded in \mathbb{H} with

$$\begin{aligned} \|p_{\tau}(\mathbf{x}, \cdot)\|_{\mathcal{E}}^2 &= \iint \kappa(\mathbf{y}, \mathbf{y}') p_{\tau}(\mathbf{x}, \mathbf{y}) p_{\tau}(\mathbf{x}, \mathbf{y}') d\mathbf{y}d\mathbf{y}' \\ &= \mathbb{E}_{\mathbf{y}, \mathbf{y}' \sim p_{\tau}(\mathbf{x}, \cdot)} [\kappa(\mathbf{y}, \mathbf{y}')] \leq B. \end{aligned}$$

Motivated by this conclusion, we propose a new measure for approximation errors of PF operators

$$\int \rho_0(\mathbf{x}) \|\hat{p}_{\tau}(\mathbf{x}, \cdot) - p_{\tau}(\mathbf{x}, \cdot)\|_{\mathcal{E}}^2 d\mathbf{x} \quad (7)$$

by replacing the $\mathcal{L}_{\rho_1}^2$ norm with $\|\cdot\|_{\mathcal{E}}$ in (6), which is finite for both deterministic and stochastic systems if $\|\hat{p}_{\tau}(\mathbf{x}, \cdot)\|_{\mathcal{E}} < \infty$. In contrast with Eq. (6), the new measure provides a more general way to quantify modeling errors of dynamics. Furthermore, from an application point of view, Eq. (7) provides a more smooth and effective representation of modeling errors of the conditional distributions.

Example 2. Let us consider a one-dimensional system with $\mathbb{M} = [-5, 5]$ and $\rho_1(\mathbf{x}) = 0.1 \cdot 1_{\mathbf{x} \in \mathbb{M}}$. Suppose that for a given \mathbf{x} , $p_{\tau}(\mathbf{x}, \mathbf{y})$ and $\hat{p}_{\tau}(\mathbf{x}, \mathbf{y})$ are separately uniform distributions within $[-0.1, 0.1]$ and $[c - 0.1, c + 0.1]$ as shown in Fig. 1A, where c is the model parameter. In VAMP, the approximation error between the two conditional distributions are calculated as $\|\hat{p}_{\tau}(\mathbf{x}, \cdot) - p_{\tau}(\mathbf{x}, \cdot)\|_{\rho_1}^2$, and it can be observed from Fig. 1B that this quantity is a constant independent of c except in a small range $c \in [-0.2, 0.2]$. But kernel embedding based error $\|\hat{p}_{\tau}(\mathbf{x}, \cdot) - p_{\tau}(\mathbf{x}, \cdot)\|_{\mathcal{E}}^2$ in (7) is a smooth function of c and provides a more reasonable metric for the evaluation of c .

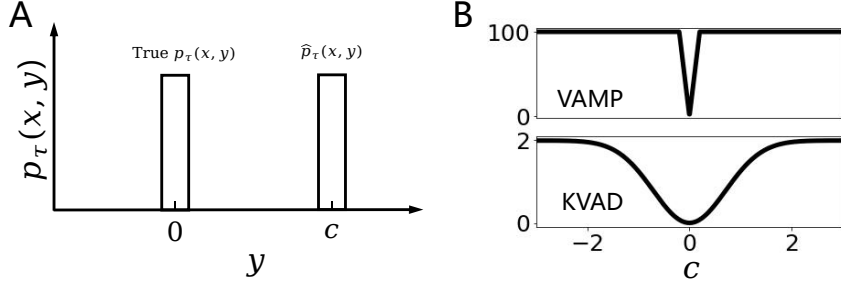


Figure 1: Illustration of distribution distances utilized in VAMP and KVAD. (A) The true conditional density is $p_\tau(x, y) = 5 \cdot 1_{|y| \leq 0.1}$ for a given x , and the approximate density is $\hat{p}_\tau(x, y) = 5 \cdot 1_{|y-c| \leq 0.1}$ for the same x . (B) The distribution distances $\|\hat{p}_\tau(\mathbf{x}, \cdot) - p_\tau(\mathbf{x}, \cdot)\|_{\rho_1}^2$ defined in VAMP and $\|\hat{p}_\tau(\mathbf{x}, \cdot) - p_\tau(\mathbf{x}, \cdot)\|_{\mathcal{E}}^2$ in KVAD with different values c , where κ is selected as the Gaussian kernel with $\sigma = 1$.

The following proposition shows that Eq. (7) can also be derived from the HS norm of operator error by treating \mathcal{P}_τ as a mapping from $\mathcal{L}_{\rho_0}^2$ to $\mathcal{L}_{\mathcal{E}}^2 = \{q \mid \|q\|_{\mathcal{E}}^2 < \infty\}$.

Proposition 3. *If κ is a universal and bounded kernel,*

1. \mathcal{P}_τ is an HS operator from $\mathcal{L}_{\rho_0}^2$ to $\mathcal{L}_{\mathcal{E}}^2$, and the corresponding $\|\hat{\mathcal{P}}_\tau - \mathcal{P}_\tau\|_{\text{HS}}^2$ is equal to (7),
2. the HS norm of $(\hat{\mathcal{P}}_\tau - \mathcal{P}_\tau)$ satisfies

$$\|\hat{\mathcal{P}}_\tau - \mathcal{P}_\tau\|_{\text{HS}}^2 = -\mathcal{R}_{\mathcal{E}}[\mathbf{f}, \mathbf{q}] + \|\mathcal{P}_\tau\|_{\text{HS}}^2,$$

with

$$\mathcal{R}_{\mathcal{E}}[\mathbf{f}, \mathbf{q}] = \text{tr}(2\mathbf{C}_{fq} - \mathbf{C}_{ff}\mathbf{C}_{qq})$$

for \hat{p}_τ defined by (4), where

$$\mathbf{C}_{ff} = \mathbb{E}_n[\mathbf{f}(\mathbf{x}_n)\mathbf{f}(\mathbf{x}_n)^\top],$$

$$\mathbf{C}_{qq} = [\langle q_i, q_j \rangle_{\mathcal{E}}],$$

$$\mathbf{C}_{fq} = [\langle \mathcal{P}_\tau(f_i \rho_0), q_j \rangle_{\mathcal{E}}],$$

are matrices of size $m \times m$, and $\mathcal{R}_{\mathcal{E}}[\mathbf{f}, \mathbf{q}]$ is called the KVAD score of \mathbf{f} and \mathbf{q} .

Proof. See Appendix B. □

As a result of this proposition, we can find optimal \mathbf{f} and \mathbf{q} by maximizing $\mathcal{R}_{\mathcal{E}}$.

4. Approximation scheme

In this section, we derive a data-driven algorithm to estimate the optimal low-dimensional linear models based on the variational principle stated in Proposition 3.

4.1. Approximation with fixed \mathbf{f}

We first propose a solution for the problem of finding the optimal \mathbf{q} given that \mathbf{f} is fixed.

Proposition 4. *If $\mathbf{C}_{ff} = \mathbb{E}_n [\mathbf{f}(\mathbf{x}_n)\mathbf{f}(\mathbf{x}_n)^\top]$ is a full-rank matrix, the solution to $\max_{\mathbf{q}} \mathcal{R}_{\mathcal{E}} [\mathbf{f}, \mathbf{q}]$ is*

$$\mathbf{q}(\mathbf{y}) = \mathbf{C}_{ff}^{-1} \int \rho_0(\mathbf{x}) p_{\tau}(\mathbf{x}, \mathbf{y}) \mathbf{f}(\mathbf{x}) d\mathbf{x} \quad (8)$$

and

$$\begin{aligned} \mathcal{R}_{\mathcal{E}} [\mathbf{f}] &\triangleq \max_{\mathbf{q}} \mathcal{R}_{\mathcal{E}} [\mathbf{f}, \mathbf{q}] \\ &= \text{tr} \left(\mathbf{C}_{ff}^{-1} \mathbb{E}_{n,n'} \left[\mathbf{f}(\mathbf{x}_n) \kappa(\mathbf{y}_n, \mathbf{y}_{n'}) \mathbf{f}(\mathbf{x}_{n'})^\top \right] \right), \end{aligned} \quad (9)$$

where $\mathbb{E}_{n,n'} [\cdot]$ denotes the expected value with $(\mathbf{x}_n, \mathbf{y}_n)$ and $(\mathbf{x}_{n'}, \mathbf{y}_{n'})$ independently drawn from the joint distribution of transition pairs.

Proof. See Appendix C. □

As $\rho_0(\mathbf{x}) p_{\tau}(\mathbf{x}, \mathbf{y})$ in Eq. (8) is the joint distribution of transition pairs $(\mathbf{x}_n, \mathbf{y}_n)$, we can get a nonparametric approximation of \mathbf{q}

$$\mathbf{q}(\mathbf{y}) = \frac{1}{N} \sum_n \mathbf{C}_{ff}^{-1} \mathbf{f}(\mathbf{x}_n) \delta_{\mathbf{y}_n}(\mathbf{y}) \quad (10)$$

by replacing the the transition pair distribution with its empirical estimate. This result gives us a linear model (1) with transition matrix

$$\begin{aligned} \mathbf{K} &= \frac{1}{N} \mathbf{C}_{ff}^{-1} \mathbf{f}(\mathbf{X})^\top \mathbf{f}(\mathbf{Y}) \\ &= \mathbf{f}(\mathbf{X})^+ \mathbf{f}(\mathbf{Y}) \end{aligned} \quad (11)$$

with $\mathbf{f}(\mathbf{X}) = (\mathbf{f}(\mathbf{x}_1), \dots, \mathbf{f}(\mathbf{x}_N))^\top \in \mathbb{R}^{N \times m}$ and $\mathbf{f}(\mathbf{X})^+$ denoting the Penrose-Moore pseudo-inverse of $\mathbf{f}(\mathbf{X})$, which is equal to the least square solution to the regression problem $\mathbf{f}(\mathbf{y}_n) \approx \mathbf{K}^\top \mathbf{f}(\mathbf{x}_n)$.

4.2. Approximation with unknown \mathbf{f}

We now consider the modeling problem with the normalization condition

$$\int \hat{p}_{\tau}(\mathbf{x}, \mathbf{y}) d\mathbf{y} = \int \mathbf{f}(\mathbf{x})^\top \mathbf{q}(\mathbf{y}) d\mathbf{y} \equiv 1, \quad (12)$$

where \mathbf{f} and \mathbf{q} are both unknown, and we make the Ansatz to represent \mathbf{f} as linear combinations of basis functions $\boldsymbol{\chi} = (\chi_1, \dots, \chi_M)^\top$. Furthermore, we assume without loss of generality that the whitening transformation is applied to the basis functions so that

$$\mathbb{E}_n [\boldsymbol{\chi}(\mathbf{x}_n)] = \mathbf{0}, \quad \mathbb{E}_n [\boldsymbol{\chi}(\mathbf{x}_n)\boldsymbol{\chi}(\mathbf{x}_n)^\top] = \mathbf{I}. \quad (13)$$

(See, e.g., Appendix F in [53] for the details of whitening transformation.)

It is proved in Appendix D that there must be a solution to $\max_{\mathbf{f}} \mathcal{R}_{\mathcal{E}}[\mathbf{f}]$ under constraint (12) satisfying

$$f_1(\mathbf{x}) \equiv 1, \quad \mathbf{C}_{ff} = \mathbf{I}. \quad (14)$$

Therefore, we can model \mathbf{f} in the form of

$$\mathbf{f}(\mathbf{x}) = (1, \boldsymbol{\chi}(\mathbf{x})^\top \mathbf{U})^\top, \quad \mathbf{U} \in \mathbb{R}^{m \times M}. \quad (15)$$

Substituting this Ansatz into the KVAD score, shows that \mathbf{U} can be computed as the solution to the maximization problem:

$$\begin{aligned} \max_{\mathbf{U}} \quad & \mathcal{R}_{\mathcal{E}}(\mathbf{U}) \\ \text{s.t.} \quad & \mathbf{C}_{ff} = \mathbf{U}^\top \mathbf{U} = \mathbf{I} \end{aligned} \quad (16)$$

with

$$\mathcal{R}_{\mathcal{E}}(\mathbf{U}) = \frac{1}{N^2} \text{tr} \left(\mathbf{U}^\top \boldsymbol{\chi}(\mathbf{X})^\top \mathbf{G}_{yy} \boldsymbol{\chi}(\mathbf{X}) \mathbf{U} \right) + \frac{1}{N^2} \mathbf{1}^\top \mathbf{G}_{yy} \mathbf{1} \quad (17)$$

being a matrix representation of $\mathcal{R}_{\mathcal{E}}[\mathbf{f}]$. Here

$$\mathbf{G}_{yy} = [\kappa(\mathbf{y}_i, \mathbf{y}_j)] \in \mathbb{R}^{N \times N}$$

is the Gram matrix of \mathbf{Y} , and $\boldsymbol{\chi}(\mathbf{X}) = (\boldsymbol{\chi}(\mathbf{x}_1), \dots, \boldsymbol{\chi}(\mathbf{x}_N))^\top \in \mathbb{R}^{N \times M}$. This problem has the same form as principal component analysis problem and can be effectively solved via the eigendecomposition of matrix $\boldsymbol{\chi}(\mathbf{X})^\top \mathbf{G}_{yy} \boldsymbol{\chi}(\mathbf{X})$ [13]. The resulting KVAD algorithm is as follows, and it can be verified that the normalization conditions (12) exactly holds for the estimated transition density (see Appendix E).

1. Select a set of basis function $\boldsymbol{\chi} = (\chi_1, \dots, \chi_M)^\top$ with $M \gg m$.
2. Perform the whitening transformation so that (13) holds.
3. Perform the truncated eigendecomposition

$$\boldsymbol{\chi}(\mathbf{X})^\top \mathbf{G}_{yy} \boldsymbol{\chi}(\mathbf{X}) \approx \mathbf{U} \mathbf{S}^2 \mathbf{U}^\top,$$

where $\mathbf{S} = \text{diag}(s_1, \dots, s_{m-1})$, $s_1 \geq s_2 \geq \dots \geq s_{m-1}$ are square roots of the largest m eigenvalues of $\boldsymbol{\chi}(\mathbf{X})^\top \mathbf{G}_{yy} \boldsymbol{\chi}(\mathbf{X})$, and $\mathbf{U} = (\mathbf{u}_1, \dots, \mathbf{u}_{m-1})$ consists of the corresponding dominant eigenvectors. This step is a bottleneck of the algorithm due to the large size Gram matrix \mathbf{G}_{yy} , and the computational cost can be reduced by Nyström approximation or random fourier features [10, 36].

4. Calculate \mathbf{f} , \mathbf{q} and \mathbf{K} by (15, 10, 11) with $\mathbf{C}_{ff} = \mathbf{I}$.

4.3. Component analysis

Due to the fact that f_1, q_1 are non-trainable, the approximate PF operator obtained by the KVAD algorithm can be decomposed as

$$\begin{aligned}\hat{\mathcal{P}}_\tau q &= \langle f_1, \rho_0 \rangle_{\rho_0^{-1}} q_1 + \sum_{i=2}^m \langle q, f_i \rho_0 \rangle_{\rho_0^{-1}} q_i \\ &= \langle q, \rho_0 \rangle_{\rho_0^{-1}} \rho_1 + \sum_{i=2}^m s_i \langle q, f_i \rho_0 \rangle_{\rho_0^{-1}} (s_i^{-1} q_i).\end{aligned}$$

It is worth pointing out that $s_i, s_i^{-1} q_{i+1}, f_{i+1} \rho_0$ obtained by KVAD algorithm are variational estimates of the i th singular value, left singular function and right singular function of the operator $\tilde{\mathcal{P}}_\tau$ defined by

$$\mathcal{P}_\tau q = \langle q, \rho_0 \rangle_{\rho_0^{-1}} \rho_1 + \tilde{\mathcal{P}}_\tau q, \quad (18)$$

where $(\tilde{\mathcal{P}}_\tau \rho_0)(\mathbf{y}) \equiv 0$. Thus, the KVAD algorithm indeed performs truncated singular value decomposition (SVD) of $\tilde{\mathcal{P}}_\tau$ (see Appendix F).

At the limit case where the all singular components of $\tilde{\mathcal{P}}_\tau$ are exactly estimated by KVAD, we have

$$\begin{aligned}D_\tau(\mathbf{x}, \mathbf{x}')^2 &\triangleq \|p_\tau(\mathbf{x}, \cdot) - p_\tau(\mathbf{x}', \cdot)\|_{\mathcal{E}}^2 \\ &= \sum_{i=1}^{m-1} s_i^2 (f_{i+1}(\mathbf{x}) - f_{i+1}(\mathbf{x}'))^2\end{aligned} \quad (19)$$

for all $\mathbf{x}, \mathbf{x}' \in \mathbb{M}$. The distance D_τ measures the dynamical similarity of two points in the state space, and can be approximated by the Euclidean distance derived from coordinates $(s_1 f_2(\mathbf{x}), \dots, s_{m-1} f_m(\mathbf{x}))^\top$ as shown in (19). Hence, KVAD provides an ideal low-dimensional embedding of system dynamics, and can be reinterpreted a variant of the diffusion mapping method for dynamical model reduction [5] (see Appendix G).

5. Relationship with Other Methods

5.1. EDMD and VAMP

It can be seen from (11) that the optimal linear model obtained by KVAD is consistent with the model of EDMD [49] for given feature functions \mathbf{f} . However, the optimization and the dimension reduction of the observables are not considered in the conventional EDMD.

Both VAMP [53] and KVAD solve this problem by variational formulations of modeling errors. As analyzed in Sections 2.1 and 3, KVAD is applicable to more general systems, including deterministic systems, compared to VAMP. Moreover, VAMP needs to represent both \mathbf{f} and \mathbf{q} by parametric models for dynamical approximation, whereas KVAD can obtain the optimal \mathbf{q} from data without any parametric model for given \mathbf{f} . Our numerical experiments (see Section 6) show that KVAD can often provide more accurate low-dimensional models than VAMP when the same Ansatz basis functions are used.

5.2. Conditional mean embedding, kernel EDMD and kernel CCA

For given two random variables \mathbf{x} and \mathbf{y} , the conditional mean embedding proposed in [46] characterizes the conditional distribution of \mathbf{y} for given \mathbf{x} by the conditional embedding operator $\mathcal{C}_{\mathbf{y}|\mathbf{x}}$ with

$$\mathbb{E}[\varphi(\mathbf{y})|\mathbf{x}] = \mathcal{C}_{\mathbf{y}|\mathbf{x}}\varphi(\mathbf{x}),$$

where φ denotes the kernel mapping and $\mathcal{C}_{\mathbf{y}|\mathbf{x}}$ can be consistently estimated from data. When applied to dynamical data, this method has the same form as the kernel EDMD and its variants [42, 16, 15], and is indeed a specific case of KVAD with Ansatz functions $\chi = \varphi$ and dimension $m = N$ (see Appendix H).

In addition, for most kernel based dynamical modeling methods, the dimension reduction problem is not thoroughly investigated. In [18], a kernel method for eigendecomposition of transfer operators (including PF operators and Koopman operators) was developed. But as analyzed in [53], the dominant eigen-components may not yield an accurate low-dimensional dynamical model. Kernel canonical correlation analysis (CCA) [22] can overcome this problem as a kernelization of VAMP, but it is also applicable only if \mathcal{P}_τ is a compact operator from $\mathcal{L}_{\rho_0}^2$ to $\mathcal{L}_{\rho_1}^2$.

Compared to the previous kernel methods, KVAD has more flexibility in model choice, where the dimension and model class of \mathbf{f} can be arbitrarily selected according to practical requirements.

6. Numerical experiments

In what follows, we demonstrate the benefits of the KVAD method for studies of nonlinear dynamical systems by two examples, and compare the results from KVAD with VAMP and kernel EDMD, where the basis functions in VAMP and kernel functions in kernel EDMD are the same as those in KVAD. For kernel EDMD, the low-dimensional linear model is achieved by leading eigenvalues and eigenfunctions as in [18], which characterizes invariant subspaces of systems.

Example 5. *Van der Pol oscillator*, which is a two-dimensional system governed by

$$\begin{aligned} dx_t &= y_t dt + \xi \cdot dw_{x,t}, \\ dy_t &= \left(2(0.2 - x_t^2)y_t - x_t\right) dt + \xi \cdot dw_{y,t}, \end{aligned}$$

where $w_{x,t}$ and $w_{y,t}$ are standard Wiener processes. The flow field of this system for $\xi = 0$ is depicted in Fig. 2A. We generate $N = 2000$ transition pairs for modeling, where the lag time $\tau = 0.2$, $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_N)^\top$ are randomly drawn from $[-1.5, 1.5]^2$, and $\mathbf{Y} = (\mathbf{y}_1, \dots, \mathbf{y}_N)^\top$ are obtained by the Euler-Maruyama scheme with step size 0.01.

Example 6. *Lorenz system* defined by

$$\begin{aligned} dx_t &= 10(y_t - x_t)dt + \xi \cdot x_t \cdot dw_{x,t}, \\ dy_t &= (28x_t - y_t - x_t z_t)dt + \xi \cdot y_t \cdot dw_{y,t}, \\ dz_t &= \left(x_t y_t - \frac{8}{3}z_t\right)dt + \xi \cdot z_t \cdot dw_{z,t}, \end{aligned}$$

with $w_{x,t}, w_{y,t}, w_{z,t}$ being standard Wiener processes. Fig. 2B plots a trajectory of this system in the state space with $\xi = 0$. We sample 2000 transition pairs from a simulation trajectory with length 200 and lag time $\tau = 0.1$ as training data for each ξ , and perform simulations by the Euler-Maruyama scheme with step size 0.005.

In both examples, the feature mapping \mathbf{f} is represented by basis functions

$$\chi_i(\mathbf{x}) = \exp\left(-\left(\boldsymbol{\theta}_i^\top \mathbf{x} + b_i\right)^2\right), \text{ for } i = 1, \dots, 500$$

with all components of $\boldsymbol{\theta}_i$ and b_i randomly drawn from $[-1, 1]$, which are widely used in shallow neural networks [12, 7]. The kernel κ is selected as the Gaussian kernel with $\sigma = 1.5$ for the oscillator and 10 for the Lorenz system.

Fig. 2 shows estimates of singular values of $\tilde{\mathcal{P}}_\tau : \mathcal{L}_{\rho_0}^2 \rightarrow \mathcal{L}_{\mathcal{E}}^2$ (KVAD), singular values of $\mathcal{P}_\tau : \mathcal{L}_{\rho_0}^2 \rightarrow \mathcal{L}_{\rho_1}^2$ (VAMP) and absolute values of eigenvalues of \mathcal{P}_τ (kernel EDMD) with different noise parameters, where singular values must be nonnegative real numbers but eigenvalues could be complex or negative. We see that the singular values and eigenvalues given by VAMP and kernel EDMD decay very slowly. Hence, it is difficult to extract an accurate model from the estimation results of VAMP and kernel EDMD for a small m . Especially for VAMP, a large number of singular values are close to 1 as analyzed in Proposition 1 when the systems are deterministic with $\xi = 0$. In contrast, the singular values utilized in KVAD rapidly converges to zero, which implies one can effectively extract the essential part of dynamics from a small number of feature mappings.

The first two singular components of $\tilde{\mathcal{P}}_\tau$ for $\xi = 0$ obtained by KVAD are shown in Fig. 3 (see Section 4.3).³ It can be observed that f_1, f_2 of the oscillator characterize transitions between left-right and up-down areas separately. Those of the Lorenz systems are related to the two attractor lobes and the transition areas.

It is interesting to note that the singular values of $\tilde{\mathcal{P}}_\tau$ given by KVAD are slightly influenced by ξ as illustrated in Fig. 2. Our numerical experience also show that the right singular functions remain almost unchanged for different ξ (see Fig. 5 in Appendix I). This phenomenon can be partially explained by the fact that the variational score \mathcal{R}_ϵ estimated by (17) is not sensitive to small perturbations of \mathbf{Y} if the bandwidth of the kernel is large. More thorough investigations on this phenomenon will be performed in future.

In order to quantitatively evaluate the performance of the three methods, we define the following trajectory reconstruction error:

$$\text{error} = \sqrt{\frac{1}{L} \sum_{l=1}^L \|\mathbf{x}_{l\tau} - \mathbb{E}_{\text{model}}[\mathbf{x}_{l\tau} | \mathbf{x}_0]\|},$$

where \mathbf{x}_t is the true trajectory data and $\mathbb{E}_{\text{model}}[\mathbf{x}_t | \mathbf{x}_0]$ is the conditional mean value of \mathbf{x}_t obtained by the model. The average error over multiple replicate simulations is minimized if and only if $\mathbb{E}_{\text{model}}[\mathbf{x}_{l\tau} | \mathbf{x}_0]$ equals to the exact conditional mean value of $\mathbf{x}_{l\tau}$ for all l . For all

³The q_i is approximated by multiple delta functions and hard to visualize.

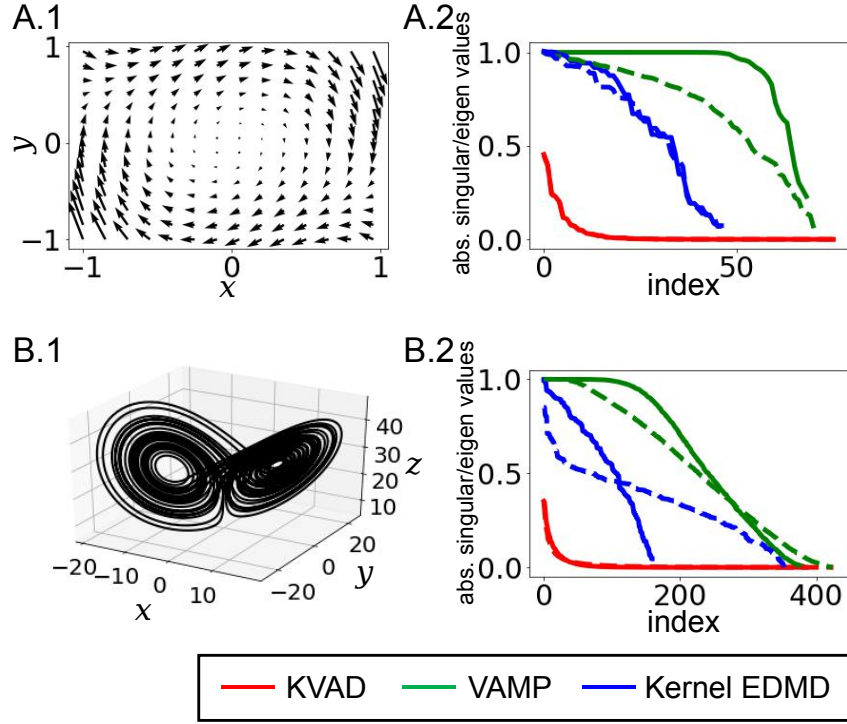


Figure 2: (A.1) Flow map of the Van der Pol oscillator, where the arrows represent directions of (dx_t, dy_t) with $\xi = 0$. (B.1) A typical trajectory of the Lorenz system with $\xi = 0$ generated by the Euler-Maruyama scheme. (A.2 and B.2) Estimated singular values and absolute values of eigenvalues of the oscillator and the Lorenz system. Red lines represent singular values of $\tilde{\mathcal{P}}_\tau$ estimated by KVAD (see (18)), green lines singular values of \mathcal{P}_τ estimated by VAMP, red lines absolute values of eigenvalues of \mathcal{P}_τ estimated by kernel EDMD, solid lines estimates with $\xi = 0$, and dashed lines those with $\xi = 0.2$ (oscillator) and 0.5 (Lorenz system). Notice the total number of spectral components changes in different cases due to the rank truncation in implementations of SVD and pseudo inverse.

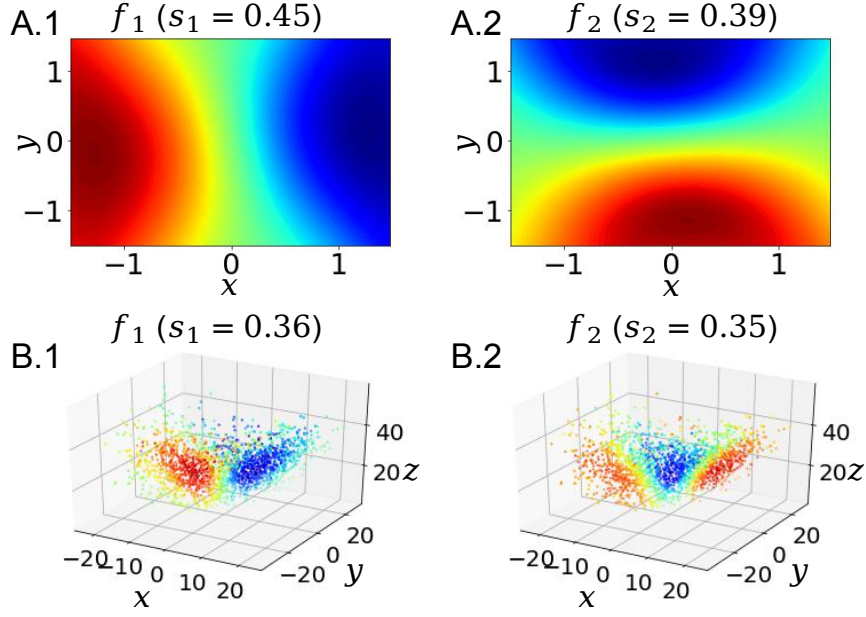


Figure 3: The first two singular components computed by KVAD, where $\xi = 0$, s_i is the estimate of the i th singular value of $\tilde{\mathcal{P}}_\tau$ and $f_i \cdot \rho_0$ the estimate of the corresponding right singular function (see Section 4.3). (A) The oscillator. (B) The Lorenz system.

the three methods,

$$\begin{aligned} \mathbb{E}_{\text{model}}[\mathbf{x}_{l\tau} | \mathbf{x}_0] &= \mathbf{G}^\top \mathbb{E}_{\text{model}}[\mathbf{f}(\mathbf{x}_{(l-1)\tau}) | \mathbf{x}_0] \\ &= (\mathbf{K}^{l-1} \mathbf{G})^\top \mathbf{f}(\mathbf{x}_0), \end{aligned}$$

where \mathbf{G} is the least square solution to the regression problem $\mathbf{x}_{t+\tau} \approx \mathbf{G}^\top \mathbf{f}(\mathbf{x}_t)$ [49]. Fig. 4 summarizes of reconstruction errors of the two systems obtained with different choices of the model dimension m and noise parameter ξ , and the superiority of our KVAD is clearly shown.

7. Conclusion

In this paper, we combine the kernel embedding technique with the variational principle for transfer operators. This provides a powerful and flexible tool for low-dimensional approximation of dynamical system, and effectively addresses the shortcomings and limitations of the existing variational approach. In the proposed KVAD framework, a bounded and well defined distance measure of transfer operators is developed based on kernel embedding of transition densities, and the corresponding variational optimization approach can be applied to a broader range of dynamical systems than the existing variational approaches.

Our future work includes the convergence analysis of KVAD and optimization of kernel functions. From the algorithmic point of view, the main remaining question is how to efficiently

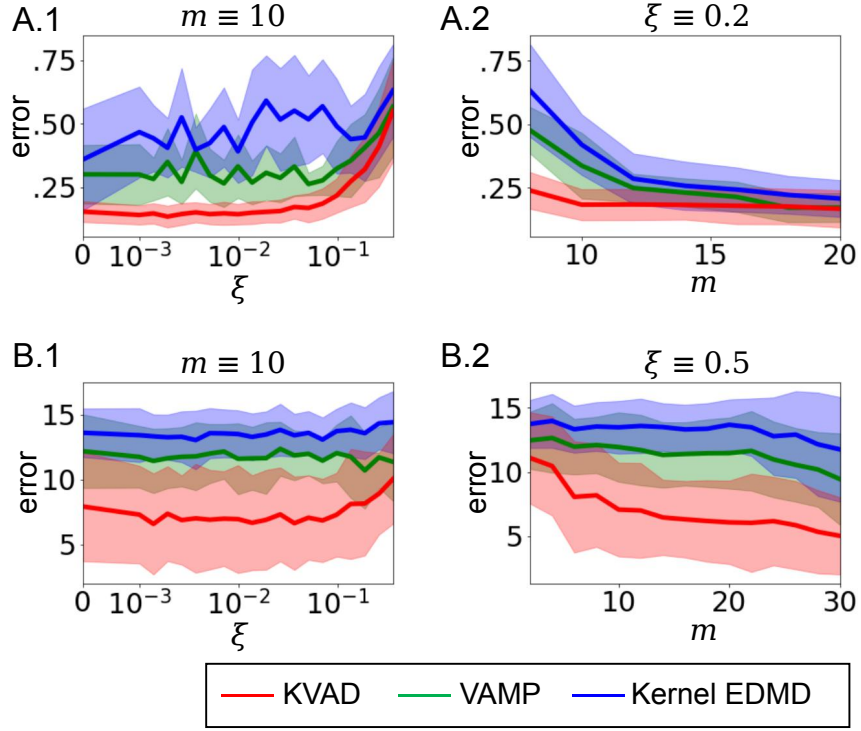


Figure 4: (A) Reconstruction errors of the oscillator, where $L = 50$ and \mathbf{x}_0 is randomly drawn in $[-1.5, 1.5]^2$. (B) Reconstruction errors of the Lorenz. Here $L = 8$, \mathbf{x}_0 is randomly sampled from a long simulation trajectory, and the trajectory is independent of the training data. Error bars represent standard deviations calculated from 100 bootstrapping replicates of simulations.

perform KVAD learning from big data with deep neural networks. It will be also interesting to apply KVAD to multi-ensemble Markov models [55] for data analysis of enhanced sampling.

Appendix

For convenience of notation, we define

$$\langle \mathbf{a}, \mathbf{b}^\top \rangle = [\langle a_i, b_j \rangle] \in \mathbb{R}^{n_1 \times n_2}$$

and

$$\mathcal{P}_\tau \mathbf{a} = (\mathcal{P}_\tau a_1, \dots, \mathcal{P}_\tau a_{n_1})^\top$$

for $\mathbf{a} = (a_1, \dots, a_{n_1})^\top$, $\mathbf{b} = (b_1, \dots, b_{n_1})^\top$ and an inner product $\langle \cdot, \cdot \rangle$.

A. Proof of Proposition 1

The proof of first conclusion is given in Appendices A.5 and B of [53], and we prove here the second conclusion.

We first show $\mathcal{R}[\mathbf{f}, \mathbf{q}] \leq m$. Because $\mathbb{E}_n [\mathbf{f}(\mathbf{x}_n) \mathbf{f}(\mathbf{x}_n)^\top]$ is a positive semi-definite matrices, it can be decomposed as

$$\mathbb{E}_n [\mathbf{f}(\mathbf{x}_n) \mathbf{f}(\mathbf{x}_n)^\top] = \mathbf{Q} \mathbf{D} \mathbf{Q}^\top, \quad (20)$$

where \mathbf{Q} is a orthogonal matrix and \mathbf{D} is a diagonal matrix. Let $\mathbf{f}' = (f'_1, \dots, f'_m)^\top = \mathbf{Q}^\top \mathbf{f}$ and $\mathbf{g}' = (g'_1, \dots, g'_m)^\top = \mathbf{Q}^\top \mathbf{g}$, we have

$$\begin{aligned} \mathcal{R}[\mathbf{f}, \mathbf{q}] &= \text{tr} \left(2\mathbb{E}_n [\mathbf{Q}^\top \mathbf{f}(\mathbf{x}_n) \mathbf{g}(\mathbf{y}_n)^\top \mathbf{Q}] - \mathbb{E}_n [\mathbf{Q}^\top \mathbf{f}(\mathbf{x}_n) \mathbf{f}(\mathbf{x}_n)^\top \mathbf{Q}] \mathbb{E}_n [\mathbf{Q}^\top \mathbf{g}(\mathbf{y}_n) \mathbf{g}(\mathbf{y}_n)^\top \mathbf{Q}] \right) \\ &= \text{tr} \left(2\mathbb{E}_n [\mathbf{f}'(\mathbf{x}_n) \mathbf{g}'(\mathbf{y}_n)^\top] - \mathbb{E}_n [\mathbf{f}'(\mathbf{x}_n) \mathbf{f}'(\mathbf{x}_n)^\top] \mathbb{E}_n [\mathbf{g}'(\mathbf{y}_n) \mathbf{g}'(\mathbf{y}_n)^\top] \right) \\ &= \sum_{i=1}^m 2\mathbb{E}_n [f'_i(\mathbf{x}_n) g'_i(\mathbf{y}_n)] - \mathbb{E}_n [f'_i(\mathbf{x}_n)^2] \mathbb{E}_n [g'_i(\mathbf{y}_n)^2] \\ &\leq \sum_{i=1}^m 2\mathbb{E}_n [f'_i(\mathbf{x}_n) g'_i(\mathbf{y}_n)] - \mathbb{E}_n [f'_i(\mathbf{x}_n) g'_i(\mathbf{y}_n)]^2 \\ &\leq m. \end{aligned}$$

Under the assumption of $\mathbb{E}_n [\mathbf{g}(\mathbf{y}_n) \mathbf{g}(\mathbf{y}_n)^\top] = \mathbf{I}$ and $f_i(\mathbf{x}) = g_i(\Theta_\tau(\mathbf{x}))$, we have

$$\begin{aligned} \langle g_i \rho_1, g_j \rho_1 \rangle_{\rho_1^{-1}} &= \int \rho_1(\mathbf{y}) g_i(\mathbf{y}) g_j(\mathbf{y}) d\mathbf{y} \\ &= \mathbb{E}_n [g_i(\mathbf{y}_n) g_j(\mathbf{y}_n)] \\ &= 1_{i=j} \end{aligned}$$

and

$$\begin{aligned}\mathbb{E}_n [\mathbf{f}(\mathbf{x}_n)\mathbf{f}(\mathbf{x}_n)^\top] &= \mathbb{E}_n [\mathbf{g}(\Theta(\mathbf{x}_n))\mathbf{g}(\Theta(\mathbf{x}_n))^\top] \\ &= \mathbb{E}_n [\mathbf{g}(\mathbf{y}_n)\mathbf{g}(\mathbf{y}_n)^\top] = \mathbf{I}\end{aligned}$$

with $\mathbf{g} = (g_1, \dots, g_m)^\top$ by considering that $\mathbf{y}_n = \Theta_\tau(\mathbf{x}_n)$. Consequently,

$$\begin{aligned}\mathcal{R}[\mathbf{f}, \mathbf{q}] &= \text{tr} \left(2\mathbb{E}_n [\mathbf{f}(\mathbf{x}_n)\mathbf{g}(\mathbf{y}_n)^\top] - \mathbf{I} \right) \\ &= \text{tr} \left(2\mathbb{E}_n [\mathbf{g}(\mathbf{y}_n)\mathbf{g}(\mathbf{y}_n)^\top] - \mathbf{I} \right) \\ &= m,\end{aligned}$$

which yields the second conclusion of this proposition.

B. Proof of Proposition 3

Let $\{e_1, e_2, \dots\}$ be an orthonormal basis of $\mathcal{L}_{\rho_0^{-1}}^2$. We have

$$\begin{aligned}\sum_k \langle \hat{\mathcal{P}}_\tau e_k, \mathcal{P}_\tau e_k \rangle_\kappa &= \sum_k \iiint p_\tau(\mathbf{x}, \mathbf{y}) e_k(\mathbf{x}) \langle \varphi(\mathbf{y}), \varphi(\mathbf{y}') \rangle_{\mathbb{H}} \hat{p}_\tau(\mathbf{x}', \mathbf{y}') e_k(\mathbf{x}') d\mathbf{x} d\mathbf{x}' d\mathbf{y} d\mathbf{y}' \\ &= \iiint \kappa(\mathbf{y}, \mathbf{y}') \cdot \sum_k p_\tau(\mathbf{x}, \mathbf{y}) e_k(\mathbf{x}) \hat{p}_\tau(\mathbf{x}', \mathbf{y}') e_k(\mathbf{x}') d\mathbf{x} d\mathbf{x}' d\mathbf{y} d\mathbf{y}' \\ &= \iint \kappa(\mathbf{y}, \mathbf{y}') \cdot \left(\sum_k \iint p_\tau(\mathbf{x}, \mathbf{y}) e_k(\mathbf{x}) \hat{p}_\tau(\mathbf{x}', \mathbf{y}') e_k(\mathbf{x}') d\mathbf{x} d\mathbf{x}' \right) d\mathbf{y} d\mathbf{y}' \\ &= \iint \kappa(\mathbf{y}, \mathbf{y}') \cdot \sum_k \left(\int p_\tau(\mathbf{x}, \mathbf{y}) e_k(\mathbf{x}) d\mathbf{x} \right) \\ &\quad \cdot \left(\int \hat{p}_\tau(\mathbf{x}', \mathbf{y}') e_k(\mathbf{x}') d\mathbf{x}' \right) d\mathbf{y} d\mathbf{y}' \\ &= \iint \kappa(\mathbf{y}, \mathbf{y}') \cdot \langle p_\tau(\cdot, \mathbf{y}) \rho_0(\cdot), \hat{p}_\tau(\cdot, \mathbf{y}') \rho_0(\cdot) \rangle_{\rho_0^{-1}} d\mathbf{y} d\mathbf{y}' \\ &= \iint \int \rho_0(\mathbf{x}) \cdot p_\tau(\mathbf{x}, \mathbf{y}) \hat{p}_\tau(\mathbf{x}, \mathbf{y}') \cdot \kappa(\mathbf{y}, \mathbf{y}') d\mathbf{x} d\mathbf{y} d\mathbf{y}' \\ &= \int \rho_0(\mathbf{x}) \langle p_\tau(\mathbf{x}, \cdot), \hat{p}_\tau(\mathbf{x}, \cdot) \rangle_{\mathcal{E}} d\mathbf{x}.\end{aligned}$$

Therefore, \mathcal{P}_τ is an HS operator with

$$\begin{aligned}\|\mathcal{P}_\tau\|_{\text{HS}}^2 &= \sum_k \langle \mathcal{P}_\tau e_k, \mathcal{P}_\tau e_k \rangle_{\mathcal{E}} \\ &= \int \rho_0(\mathbf{x}) \|p_\tau(\mathbf{x}, \cdot)\|_{\mathcal{E}}^2 d\mathbf{x} \leq B\end{aligned}$$

if κ is bounded by B , and

$$\begin{aligned}\|\mathcal{P}_\tau - \hat{\mathcal{P}}_\tau\|_{\text{HS}}^2 &= \sum_k \left\langle (\mathcal{P}_\tau - \hat{\mathcal{P}}_\tau) e_k, (\mathcal{P}_\tau - \hat{\mathcal{P}}_\tau) e_k \right\rangle_\kappa \\ &= \mathcal{D}(\hat{\mathcal{P}}_\tau, \mathcal{P}_\tau)^2.\end{aligned}$$

If $\hat{p}_\tau(\mathbf{x}, \mathbf{y}) = \mathbf{f}(\mathbf{x})^\top \mathbf{q}(\mathbf{y})$, we get

$$\begin{aligned}\|\hat{\mathcal{P}}_\tau\|_{\text{HS}}^2 &= \iiint \rho_0(\mathbf{x}) \mathbf{f}(\mathbf{x})^\top \mathbf{q}(\mathbf{y}) \kappa(\mathbf{y}, \mathbf{y}') \mathbf{q}(\mathbf{y}')^\top \mathbf{f}(\mathbf{x}) d\mathbf{x} d\mathbf{y} d\mathbf{y}' \\ &= \text{tr}(\mathbf{C}_{ff} \mathbf{C}_{qq}),\end{aligned}$$

$$\begin{aligned}\sum_k \left\langle \hat{\mathcal{P}}_\tau e_k, \mathcal{P}_\tau e_k \right\rangle_\kappa &= \int \int \int \rho_0(\mathbf{x}) \cdot p_\tau(\mathbf{x}, \mathbf{y}) \mathbf{q}(\mathbf{y}')^\top \mathbf{f}(\mathbf{x}) \cdot \kappa(\mathbf{y}, \mathbf{y}') d\mathbf{x} d\mathbf{y} d\mathbf{y}' \\ &= \text{tr}(\mathbf{C}_{fq}),\end{aligned}$$

and

$$\|\mathcal{P}_\tau - \hat{\mathcal{P}}_\tau\|_{\text{HS}}^2 = -\mathcal{R}_\mathcal{E}[\mathbf{f}, \mathbf{q}] + \|\mathcal{P}_\tau\|_{\text{HS}}^2.$$

C. Proof of Proposition 4

Let us first consider the case where $\mathbf{C}_{ff} = \mathbf{I}$. Then

$$\begin{aligned}\mathcal{R}_\mathcal{E}[\mathbf{f}, \mathbf{q}] &= -\text{tr} \left(\left\langle \mathbf{q}, \mathbf{q}^\top \right\rangle_\mathcal{E} - 2 \left\langle \mathbf{q}, \mathcal{P}_\tau(\mathbf{f}\rho_0)^\top \right\rangle_\mathcal{E} \right) \\ &= -\text{tr} \left(\left\langle \mathbf{q} - \mathcal{P}_\tau(\mathbf{f}\rho_0), (\mathbf{q} - \mathcal{P}_\tau(\mathbf{f}\rho_0))^\top \right\rangle_\mathcal{E} - \left\langle \mathcal{P}_\tau(\mathbf{f}\rho_0), \mathcal{P}_\tau(\mathbf{f}\rho_0)^\top \right\rangle_\mathcal{E} \right) \\ &= -\sum_i \|q_i - \mathcal{P}_\tau(f_i \rho_0)\|_\mathcal{E}^2 + \sum_i \|\mathcal{P}_\tau(f_i \rho_0)\|_\mathcal{E}^2,\end{aligned}$$

which leads to

$$\begin{aligned}\arg \max_{\mathbf{q}} \mathcal{R}_\mathcal{E}[\mathbf{f}, \mathbf{q}] &= \mathcal{P}_\tau(\mathbf{f}\rho_0) \\ &= \int \rho_0(\mathbf{x}) p_\tau(\mathbf{x}, \cdot) \mathbf{f}(\mathbf{x}) d\mathbf{x}\end{aligned}$$

and

$$\begin{aligned}\mathcal{R}_\mathcal{E}[\mathbf{f}] &= \|\mathcal{P}_\tau(\mathbf{f}\rho_0)\|_\mathcal{E}^2 \\ &= \text{tr} \left(\mathbb{E}_{n, n'} \left[\mathbf{f}(\mathbf{x}_n) \kappa(\mathbf{y}_n, \mathbf{y}_{n'}) \mathbf{f}(\mathbf{x}_{n'})^\top \right] \right).\end{aligned}$$

We now suppose that $\mathbf{C}_{ff} \neq \mathbf{I}$ and let

$$\begin{aligned}\mathbf{f}' &= \mathbf{C}_{ff}^{-\frac{1}{2}} \mathbf{f}, \\ \mathbf{q}' &= \mathbf{C}_{ff}^{\frac{1}{2}} \mathbf{q}.\end{aligned}$$

Because $\langle \mathbf{f}', \mathbf{f}'^\top \rangle_{\rho_0} = \mathbf{I}$, we have

$$\arg \max_{\mathbf{q}'} \mathcal{R}_{\mathcal{E}}[\mathbf{f}', \mathbf{q}'] = \int \rho_0(\mathbf{x}) p_{\tau}(\mathbf{x}, \cdot) \mathbf{f}'(\mathbf{x}) d\mathbf{x}$$

and

$$\mathcal{R}_{\mathcal{E}}[\mathbf{f}'] = \text{tr} \left(\mathbb{E}_{n,n'} \left[\mathbf{f}'(\mathbf{x}_n) \kappa(\mathbf{y}_n, \mathbf{y}_{n'}) \mathbf{f}'(\mathbf{x}_{n'})^\top \right] \right).$$

Considering that the transition density defined by $(\mathbf{f}', \mathbf{q}')$ is equivalent to that by (\mathbf{f}, \mathbf{q}) as

$$\mathbf{f}(\mathbf{x})^\top \mathbf{q}(\mathbf{y}) = \mathbf{f}'(\mathbf{x})^\top \mathbf{q}'(\mathbf{y}),$$

we can obtain

$$\begin{aligned} \arg \max_{\mathbf{q}} \mathcal{R}_{\mathcal{E}}[\mathbf{f}, \mathbf{q}] &= \mathbf{C}_{ff}^{-\frac{1}{2}} \mathcal{P}_{\tau}(\mathbf{f}' \rho_0) \\ &= \mathbf{C}_{ff}^{-1} \int \rho_0(\mathbf{x}) p_{\tau}(\mathbf{x}, \cdot) \mathbf{f}(\mathbf{x}) d\mathbf{x} \end{aligned}$$

and

$$\begin{aligned} \mathcal{R}_{\mathcal{E}}[\mathbf{f}] &= \mathcal{R}_{\mathcal{E}}[\mathbf{f}'] \\ &= \text{tr} \left(\mathbb{E}_{n,n'} \left[\mathbf{C}_{ff}^{-\frac{1}{2}} \mathbf{f}(\mathbf{x}_n) \kappa(\mathbf{y}_n, \mathbf{y}_{n'}) \mathbf{f}(\mathbf{x}_{n'})^\top \mathbf{C}_{ff}^{-\frac{1}{2}} \right] \right) \\ &= \text{tr} \left(\mathbb{E}_{n,n'} \left[\mathbf{C}_{ff}^{-1} \mathbf{f}(\mathbf{x}_n) \kappa(\mathbf{y}_n, \mathbf{y}_{n'}) \mathbf{f}(\mathbf{x}_{n'})^\top \right] \right). \end{aligned}$$

D. Proof of (14)

Suppose that (\mathbf{f}, \mathbf{q}) is a solution to $\max \mathcal{R}_{\mathcal{E}}[\mathbf{f}, \mathbf{q}]$ with dimension m under constraint (12). From (12), we have

$$\mathbf{f}(\mathbf{x})^\top \left(\int \mathbf{q}(\mathbf{y}) d\mathbf{y} \right) \equiv 1,$$

which implies that the constant function belongs to the subspace spanned by \mathbf{f} . Thus we can obtain an matrix \mathbf{R} so that $\mathbf{f}' = (f'_1, \dots, f'_m)^\top = \mathbf{R}\mathbf{f}$ satisfies (14) by Gram-Schmidt orthogonalization, and \mathbf{f}' and $\mathbf{q}' = \mathbf{R}^{-\top} \mathbf{q}$ also maximizes $\mathcal{R}_{\mathcal{E}}$.

E. Normalization property of estimated transition density

For the transition density $\hat{p}_{\tau}(\mathbf{x}, \mathbf{y}) = \mathbf{f}(\mathbf{x})^\top \mathbf{q}(\mathbf{y})$ obtained by the KVAD algorithm, we have

$$\int q_i(\mathbf{y}) d\mathbf{y} = \frac{1}{N} \sum_n f_i(\mathbf{x}_n) = 0$$

for $i > 1$. Therefore,

$$\begin{aligned} \int \hat{p}_{\tau}(\mathbf{x}, \mathbf{y}) d\mathbf{y} &= \mathbf{f}(\mathbf{x})^\top \left(\int \mathbf{q}(\mathbf{y}) d\mathbf{y} \right) \\ &= \mathbf{f}(\mathbf{x})^\top (1, 0, \dots, 0)^\top \\ &= 1. \end{aligned}$$

F. Singular value decomposition of $\tilde{\mathcal{P}}_\tau$

Because $\tilde{\mathcal{P}}_\tau$ is also an HS operator from $\mathcal{L}_{\rho_0}^2$ to $\mathcal{L}_\mathcal{E}^2$, there exists the following SVD:

$$\tilde{\mathcal{P}}_\tau q = \sum_{i=1}^{\infty} \sigma_i \langle q, \psi_i \rangle_{\rho_0^{-1}} \phi_i. \quad (21)$$

Here σ_i denotes the i th largest singular value, and ϕ_i, ψ_i are the corresponding left and right singular functions. According to the Rayleigh variational principle, for the i th singular component, we have

$$\sigma_i^2 = \max_q \langle \tilde{\mathcal{P}}_\tau q, \tilde{\mathcal{P}}_\tau q \rangle_\mathcal{E} \quad (22)$$

under constraints

$$\langle q, q \rangle_{\rho_0^{-1}} = 1, \quad \langle q, \psi_j \rangle_{\rho_0^{-1}} = 0, \quad \forall j = 1, \dots, i-1 \quad (23)$$

and the solution is $q = \psi_i$.

From the above variational formulation of SVD, we can obtain the following proposition:

Proposition 7. *The singular functions ψ_i of $\tilde{\mathcal{P}}_\tau$ satisfies*

$$\langle \rho_0, \psi_i \rangle_{\rho_0^{-1}} = 0$$

if $\sigma_i > 0$.

Proof. We first show that $\langle \rho_0, \psi_1 \rangle_{\rho_0^{-1}} = 0$ by contradiction. If $\langle \rho_0, \psi_1 \rangle_{\rho_0^{-1}} = c_1 \neq 0$, ψ_1 can be decomposed as

$$\psi_1 = c_1 \rho_0 + \tilde{\psi}_1.$$

Because

$$\begin{aligned} \langle \tilde{\mathcal{P}}_\tau \tilde{\psi}_1, \tilde{\mathcal{P}}_\tau \tilde{\psi}_1 \rangle_\mathcal{E} &= \langle \tilde{\mathcal{P}}_\tau \psi_1, \tilde{\mathcal{P}}_\tau \psi_1 \rangle_\mathcal{E}, \\ \langle \tilde{\psi}_1, \tilde{\psi}_1 \rangle_{\rho_0^{-1}} &= 1 - c_1^2, \end{aligned}$$

we can get that $\langle \tilde{\psi}_1', \tilde{\psi}_1' \rangle_{\rho_0^{-1}} = 1$ and

$$\langle \tilde{\mathcal{P}}_\tau \tilde{\psi}_1', \tilde{\mathcal{P}}_\tau \tilde{\psi}_1' \rangle_\mathcal{E} = \frac{1}{1 - c_1^2} \langle \tilde{\mathcal{P}}_\tau \psi_1, \tilde{\mathcal{P}}_\tau \psi_1 \rangle_\mathcal{E} > \sigma^2$$

with

$$\tilde{\psi}_1' = (1 - c_1^2)^{-\frac{1}{2}} \tilde{\psi}_1,$$

which leads to a contradiction. Therefore, $\langle \rho_0, \psi_1 \rangle_{\rho_0^{-1}} = 0$.

For ψ_2 , we can also show that

$$\tilde{\psi}_2' = (1 - c_2^2)^{-\frac{1}{2}} (\psi_2 - c_2 \rho_0)$$

with $c_2 = \langle \rho_0, \psi_2 \rangle_{\rho_0^{-1}}$ and $\langle \psi_1, \tilde{\psi}_2 \rangle_{\rho_0^{-1}} = 0$ is a better solution than ψ_2 for the variational optimization problem (22, 23) if $c_2 \neq 0$, and thus $\langle \rho_0, \psi_2 \rangle_{\rho_0^{-1}} = 0$. By mathematical induction, $\langle \rho_0, \psi_i \rangle_{\rho_0^{-1}} = 0$ for all $\sigma_i > 0$. \square

Based on this proposition, ψ_i can be approximated by

$$\psi_i = \mathbf{u}_i^\top \boldsymbol{\chi} \quad (24)$$

with (13) being satisfied. Substituting the Ansatz (24) into (22, 23) and replacing expected values with empirical estimates yields

$$\begin{aligned} \mathbf{u}_i &= \arg \max_{\mathbf{u}} \frac{1}{N^2} \text{tr} \left(\mathbf{u}^\top \boldsymbol{\chi}(\mathbf{X})^\top \mathbf{G}_{yy} \boldsymbol{\chi}(\mathbf{X}) \mathbf{u} \right) \\ \text{s.t. } &\mathbf{u}^\top \mathbf{u} = 1, \quad \mathbf{u}^\top \mathbf{u}_j = 0 \text{ for } j = 1, \dots, i-1. \end{aligned}$$

This problem for all i can be equivalently solved by the KVAD algorithm in Section 4.2. Consequently, $s_i, s_i^{-1} q_{i+1}, f_{i+1} \rho_0$ are variational estimates of the i th singular value, left singular function and right singular function of the operator $\tilde{\mathcal{P}}_\tau$.

G. Proof of (19)

From (21) and the orthonormality of ϕ_i , we have

$$\begin{aligned} D_\tau(\mathbf{x}, \mathbf{x}')^2 &= \|\mathcal{P}_\tau \delta_{\mathbf{x}} - \mathcal{P}_\tau \delta_{\mathbf{x}'}\|_{\mathcal{E}}^2 \\ &= \|\tilde{\mathcal{P}}_\tau \delta_{\mathbf{x}} - \tilde{\mathcal{P}}_\tau \delta_{\mathbf{x}'}\|_{\mathcal{E}}^2 \\ &= \left\| \sum_i \sigma_i \left(\psi_i(\mathbf{x}) \rho_0(\mathbf{x})^{-1} - \psi_i(\mathbf{x}') \rho_0(\mathbf{x}')^{-1} \right) \phi_i \right\|_{\mathcal{E}}^2 \\ &= \sum_{i,j} \sigma_i \sigma_j \left(\psi_i(\mathbf{x}) \rho_0(\mathbf{x})^{-1} - \psi_i(\mathbf{x}') \rho_0(\mathbf{x}')^{-1} \right) \\ &\quad \cdot \left(\psi_j(\mathbf{x}) \rho_0(\mathbf{x})^{-1} - \psi_j(\mathbf{x}') \rho_0(\mathbf{x}')^{-1} \right) \langle \phi_i, \phi_j \rangle_{\mathcal{E}} \\ &= \sum_{i=1}^{\infty} \sigma_i^2 \left(\psi_i(\mathbf{x}) \rho_0(\mathbf{x})^{-1} - \psi_i(\mathbf{x}') \rho_0(\mathbf{x}')^{-1} \right)^2. \end{aligned}$$

If the KVAD algorithm gives the exact approximation of singular components and $\sigma_i = 0$ for $i > m-1$, we can get

$$D_\tau(\mathbf{x}, \mathbf{x}')^2 = \sum_{i=1}^{m-1} s_i^2 (f_{i+1}(\mathbf{x}) - f_{i+1}(\mathbf{x}'))^2.$$

H. Comparison between KVAD and conditional mean embedding

We consider

$$f_i(\mathbf{x}) = \mathbf{u}_i^\top \varphi(\mathbf{x}),$$

for $i = 1, \dots, N$ in KVAD, and assume that $\mathbf{G}_{xx} = [\kappa(\mathbf{x}_i, \mathbf{x}_j)] \in \mathbb{R}^{N \times N}$ is invertible. Here \mathbf{u}_i is the i th column of \mathbf{U} , and $\varphi(\mathbf{x})$ is the kernel mapping and can be explicitly represented as a function from \mathbb{M} to a (possibly infinite-dimensional) Euclidean space with $\kappa(\mathbf{x}, \mathbf{y}) = \varphi(\mathbf{x})^\top \varphi(\mathbf{y})$ [29]. For a given data set $\{(\mathbf{x}_n, \mathbf{y}_n)\}_{n=1}^N$, an arbitrary \mathbf{u}_i can be decomposed as

$$\mathbf{u}_i = \varphi(\mathbf{X})^\top \mathbf{a}_i + \mathbf{u}_i^\perp$$

with $\varphi(\mathbf{X}) = (\varphi(\mathbf{x}_1), \dots, \varphi(\mathbf{x}_N))^\top$ and $\varphi(\mathbf{X})^\top \mathbf{u}_i^\perp = \mathbf{0}$, and

$$\mathbf{u}_i^\top \varphi(\mathbf{x}_n) = \left(\varphi(\mathbf{X})^\top \mathbf{a}_i \right)^\top \varphi(\mathbf{x}_n), \quad \forall n.$$

So, we can assume without loss of generality that each \mathbf{u}_i can be represented as a linear combination of $\{\varphi(\mathbf{x}_1), \dots, \varphi(\mathbf{x}_N)\}$, and therefore all invertible $\mathbf{A} = (\mathbf{a}_1, \dots, \mathbf{a}_N)$ can generate the equivalent model. For convenience of analysis, we set $\mathbf{A} = \mathbf{I}$ and

$$\mathbf{f}(\mathbf{x}) = \varphi(\mathbf{X})\varphi(\mathbf{x}).$$

$$N\mathbf{C}_{ff} = \sum_{n=1}^N \varphi(\mathbf{X})\varphi(\mathbf{x}_n)\varphi(\mathbf{x}_n)^\top \varphi(\mathbf{X})^\top = \mathbf{G}_{xx}^2$$

Then

$$\mathbf{q}(\mathbf{y}) = \sum_n \mathbf{G}_{xx}^{-2} \varphi(\mathbf{X})\varphi(\mathbf{x}_n)\delta_{\mathbf{y}_n}(\mathbf{y}),$$

$$\begin{aligned} \hat{p}_\tau(\mathbf{x}, \mathbf{y}) &= \mathbf{f}(\mathbf{x})^\top \mathbf{q}(\mathbf{y}) \\ &= \varphi(\mathbf{x})^\top \varphi(\mathbf{X})^\top \sum_n \mathbf{G}_{xx}^{-2} \varphi(\mathbf{X})\varphi(\mathbf{x}_n)\delta_{\mathbf{y}_n}(\mathbf{y}) \end{aligned}$$

and we can obtain

$$\begin{aligned} \mathbb{E}[\varphi(\mathbf{y})|\mathbf{x}] &= \sum_n \varphi(\mathbf{y}_n)\varphi(\mathbf{x}_n)^\top \varphi(\mathbf{X})^\top \mathbf{G}_{xx}^{-2} \varphi(\mathbf{X})\varphi(\mathbf{x}) \\ &= \varphi(\mathbf{Y})^\top \varphi(\mathbf{X})\varphi(\mathbf{X})^\top \mathbf{G}_{xx}^{-2} \varphi(\mathbf{X})\varphi(\mathbf{x}) \\ &= \varphi(\mathbf{Y})^\top \mathbf{G}_{xx}^{-1} (\kappa(\mathbf{x}_1, \mathbf{x}), \dots, \kappa(\mathbf{x}_N, \mathbf{x}))^\top, \end{aligned}$$

which is equivalent to the result of conditional mean embedding [45].

I. Estimated singular components of Examples 5 and 6 with $\xi \neq 0$

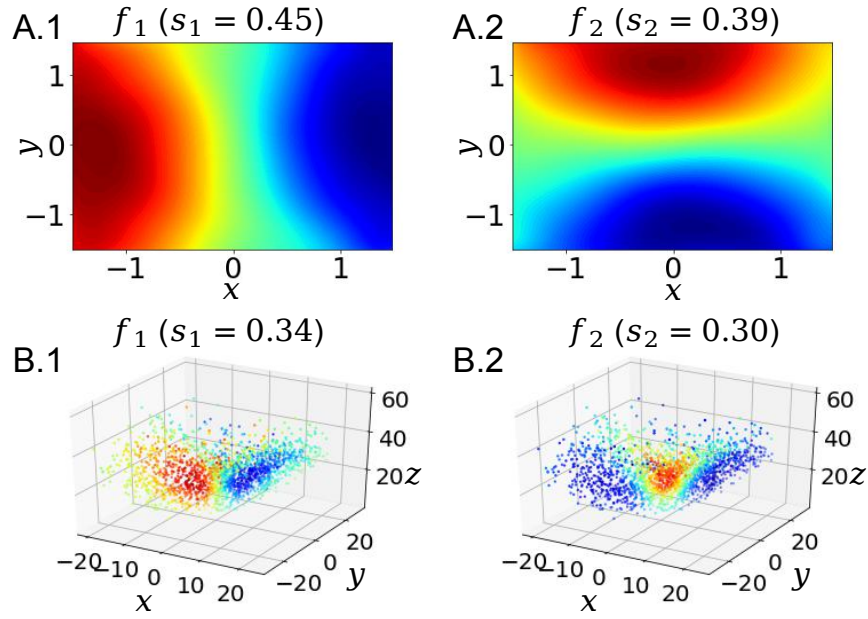


Figure 5: The first two singular components computed by KVAD. (A) The oscillator with $\xi = 0.2$. (B) The Lorenz system with $\xi = 0.5$.

References

- [1] S. L. Brunton, B. W. Brunton, J. L. Proctor, and J. N. Kutz, *Koopman invariant subspaces and finite linear representations of nonlinear dynamical systems for control*, PloS one, 11 (2016), p. e0150171.
- [2] K. K. Chen, J. H. Tu, and C. W. Rowley, *Variants of dynamic mode decomposition: Boundary condition, koopman, and fourier analyses*, Journal of Nonlinear Ence, 22 (2012), pp. 887–915.
- [3] W. Chen, H. Sidky, and A. L. Ferguson, *Nonlinear discovery of slow molecular modes using state-free reversible vampnets*, Journal of chemical physics, 150 (2019), p. 214114.
- [4] J. D. Chodera and F. Noé, *Markov state models of biomolecular conformational dynamics*, Curr Opin Struct Biol, 25 (2014), pp. 135–144.
- [5] R. R. Coifman and S. Lafon, *Diffusion maps*, Applied and computational harmonic analysis, 21 (2006), pp. 5–30.
- [6] N. D. Conrad, M. Weber, and C. Schütte, *Finding dominant structures of nonreversible markov processes*, Multiscale Modeling & Simulation, 14 (2016), pp. 1319–1340.
- [7] S. Dash, B. Tripathy, et al., *Handbook of Research on Modeling, Analysis, and Application of Nature-Inspired Metaheuristic Algorithms*, IGI Global, 2017.

- [8] G. Debdipta, T. Emma, and P. D. A., *Constrained ulam dynamic mode decomposition: Approximation of perron-frobenius operator for deterministic and stochastic systems*, IEEE Control Systems Letters, (2018), pp. 1–1.
- [9] P. Deuffhard and M. Weber, *Robust perron cluster analysis in conformation dynamics*, Linear algebra and its applications, 398 (2005), pp. 161–184.
- [10] P. Drineas and M. W. Mahoney, *On the nystrom method for approximating a gram matrix for improved kernel-based learning*, journal of machine learning research, 6 (2005), pp. 2153–2175.
- [11] K. Fackeldey, P. Koltai, P. N  vir, H. Rust, A. Schild, and M. Weber, *From metastable to coherent sets—time-discretization schemes*, Chaos: An Interdisciplinary Journal of Nonlinear Science, 29 (2019), p. 012101.
- [12] G.-B. Huang, Q.-Y. Zhu, and C.-K. Siew, *Extreme learning machine: theory and applications*, Neurocomputing, 70 (2006), pp. 489–501.
- [13] I. T. Jolliffe and J. Cadima, *Principal component analysis: a review and recent developments*, Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 374 (2016), p. 20150202.
- [14] O. Junge and P. Koltai, *Discretization of the frobenius-perron operator using a sparse haar tensor basis: The sparse ulam method*, Siam Journal on Numerical Analysis, 47 (2009), pp. 3464–3485.
- [15] Y. Kawahara, *Dynamic mode decomposition with reproducing kernels for koopman spectral analysis*, in Advances in neural information processing systems, 2016, pp. 911–919.
- [16] I. G. Kevrekidis, C. W. Rowley, and M. O. Williams, *A kernel-based method for data-driven koopman spectral analysis*, Journal of Computational Dynamics, 2 (2016), pp. 247–265.
- [17] S. Klus, F. N  ske, P. Koltai, H. Wu, I. Kevrekidis, C. Sch  tte, and F. No  , *Data-driven model reduction and transfer operator approximation*, Journal of Nonlinear Science, 28 (2018), pp. 985–1010.
- [18] S. Klus, I. Schuster, and K. Muandet, *Eigendecompositions of transfer operators in reproducing kernel hilbert spaces*, Journal of Nonlinear Science, 30 (2020), pp. 283–315.
- [19] A. Konrad, B. Y. Zhao, A. D. Joseph, and R. Ludwig, *A markov-based channel model algorithm for wireless networks*, Wireless Networks, 9 (2003), pp. 189–199.
- [20] M. Korda and I. Mezi  , *On convergence of extended dynamic mode decomposition to the koopman operator*, Journal of Nonlinear Science, 28 (2018), pp. 687–710.
- [21] J. N. Kutz, S. L. Brunton, B. W. Brunton, and J. L. Proctor, *Dynamic Mode Decomposition: Data Driven Modeling of Complex Systems*, 2016.

- [22] P. L. Lai and C. Fyfe, *Kernel and nonlinear canonical correlation analysis*, International Journal of Neural Systems, 10 (2000), pp. 365–377.
- [23] Q. Li, F. Dietrich, E. M. Bollt, and I. G. Kevrekidis, *Extended dynamic mode decomposition with dictionary learning: A data-driven adaptive spectral decomposition of the koopman operator*, Chaos: An Interdisciplinary Journal of Nonlinear Science, 27 (2017), p. 103111.
- [24] B. Lusch, J. N. Kutz, and S. L. Brunton, *Deep learning for universal linear embeddings of nonlinear dynamics*, Nature communications, 9 (2018), pp. 1–10.
- [25] A. Mardt, L. Pasquali, F. Noé, and H. Wu, *Deep learning markov and koopman models with physical constraints*, arXiv: Computational Physics, (2019).
- [26] A. Mardt, L. Pasquali, H. Wu, and F. Noé, *Vampnets for deep learning of molecular kinetics*, Nature communications, 9 (2018), pp. 1–11.
- [27] R. T. McGibbon and V. S. Pande, *Variational cross-validation of slow dynamical modes in molecular kinetics*, Journal of Chemical Physics, 142 (2015), p. 03B621_1.
- [28] I. Mezić, *Analysis of fluid flows via spectral properties of the koopman operator*, Annual Review of Fluid Mechanics, 45 (2013), pp. 357–378.
- [29] H. Q. Minh, P. Niyogi, and Y. Yao, *Mercer’s theorem, feature maps, and smoothing*, in International Conference on Computational Learning Theory, Springer, 2006, pp. 154–168.
- [30] L. Molgedey and H. G. Schuster, *Separation of a mixture of independent signals using time delayed correlations*, Physical Review Letters, 72 (1994), pp. 3634–3637.
- [31] F. Noe and F. Nuske, *A variational approach to modeling slow processes in stochastic dynamical systems*, Multiscale Modeling and Simulation, 11 (2013), pp. 635–655.
- [32] F. Nuske, B. G. Keller, G. Perezhernandez, A. S. J. S. Mey, and F. Noe, *Variational approach to molecular kinetics*, Journal of Chemical Theory and Computation, 10 (2014), pp. 1739–1752.
- [33] F. Núske, R. Schneider, F. Vitalini, and F. Noé, *Variational tensor approach for approximating the rare event kinetics of macromolecular systems*, Journal of Chemical Physics, 144 (2016), pp. 149–153.
- [34] G. Perezhernandez, F. Paul, T. Giorgino, G. De Fabritiis, and F. Noe, *Identification of slow molecular order parameters for markov model construction*, Journal of Chemical Physics, 139 (2013), p. 015102.
- [35] J.-H. Prinz, H. Wu, M. Sarich, B. Keller, M. Senne, M. Held, J. D. Chodera, C. Schütte, and F. Noé, *Markov models of molecular kinetics: Generation and validation*, Journal of chemical physics, 134 (2011), p. 174105.

- [36] A. Rahimi and B. Recht, *Random features for large-scale kernel machines*, in Advances in neural information processing systems, 2008, pp. 1177–1184.
- [37] S. Röblitz and M. Weber, *Fuzzy spectral clustering by pcca+: application to markov state models and data classification*, Advances in Data Analysis and Classification, 7 (2013), pp. 147–179.
- [38] P. J. Schmid and J. Sesterhenn, *Dynamic mode decomposition of numerical and experimental data*, Journal of Fluid Mechanics, 656 (2010), pp. 5–28.
- [39] C. Schütte, A. Fischer, W. Huisinga, and P. Deuflhard, *A direct approach to conformational dynamics based on hybrid monte carlo*, Journal of Computational Physics, 151 (1999), pp. 146–168.
- [40] C. Schütte, P. Koltai, and S. Klus, *On the numerical approximation of the perron frobenius and koopman operator*, Journal of Computational Dynamics, 3 (2016), pp. 1–12.
- [41] C. R. Schwantes and V. S. Pande, *Improvements in markov state model construction reveal many non native interactions in the folding of ntl9*, Journal of Chemical Theory and Computation, 9 (2013), pp. 2000–2009.
- [42] ———, *Modeling molecular kinetics with tica and the kernel trick*, Journal of chemical theory and computation, 11 (2015), pp. 600–608.
- [43] A. S. Sharma, I. Mezi, and B. J. Mckeen, *Correspondence between koopman mode decomposition, resolvent mode decomposition, and invariant solutions of the navier-stokes equations*, Phys.rev.fluids, 1 (2016).
- [44] A. Smola, A. Gretton, L. Song, and B. Schölkopf, *A hilbert space embedding for distributions*, in International Conference on Algorithmic Learning Theory, Springer, 2007, pp. 13–31.
- [45] L. Song, K. Fukumizu, and A. Gretton, *Kernel embeddings of conditional distributions: A unified kernel framework for nonparametric inference in graphical models*, IEEE Signal Processing Magazine, 30 (2013), pp. 98–111.
- [46] L. Song, J. Huang, A. Smola, and K. Fukumizu, *Hilbert space embeddings of conditional distributions with applications to dynamical systems*, in Proceedings of the 26th Annual International Conference on Machine Learning, 2009, pp. 961–968.
- [47] B. K. Sriperumbudur, K. Fukumizu, and G. R. Lanckriet, *Universality, characteristic kernels and rkhs embedding of measures.*, J. Mach. Learn. Res., 12 (2011).
- [48] J. H. Tu, C. W. Rowley, D. M. Luchtenburg, S. L. Brunton, and J. N. Kutz, *On dynamic mode decomposition: Theory and applications*, ACM Journal of Computer Documentation, 1 (2014), pp. 391–421.

- [49] M. O. Williams, I. G. Kevrekidis, and C. W. Rowley, *A data-driven approximation of the koopman operator: Extending dynamic mode decomposition*, Journal of Nonlinear Science, 25 (2015), pp. 1307–1346.
- [50] M. O. Williams, I. G. Kevrekidis, and C. W. Rowley, *A data driven approximation of the koopman operator: Extending dynamic mode decomposition*, Journal of Nonlinear Science, (2015).
- [51] H. Wu, A. Mardt, L. Pasquali, and F. Noe, *Deep generative markov state models*, in Advances in Neural Information Processing Systems, 2018, pp. 3975–3984.
- [52] H. Wu and F. Noé, *Gaussian markov transition models of molecular kinetics*, Journal of Chemical Physics, 142 (2015), p. 084104.
- [53] ———, *Variational approach for learning markov processes from time series data*, J. Nonlinear Sci., 30 (2020), pp. 23–66.
- [54] H. Wu, F. Núske, F. Paul, S. Klus, P. Koltai, and F. Noé, *Variational koopman models: Slow collective variables and molecular kinetics from short off-equilibrium simulations*, The Journal of Chemical Physics, 146 (2017), p. 154104.
- [55] H. Wu, F. Paul, C. Wehmeyer, and F. Noé, *Multiensemble markov models of molecular thermodynamics and kinetics*, Proceedings of the National Academy of Sciences, 113 (2016), pp. E3221–E3230.
- [56] M. Yue, J. Han, and K. Trivedi, *Composite performance and availability analysis of wirelesscommunication networks*, IEEE Transactions on Vehicular Technology, 50 (2001), pp. p.1216–1223.