Journal of
Nonlinear
Science

Variational Approach for Learning Markov Processes from Time Series Data

Hao Wu[1,2] · Frank Noé[2,3,4]

Abstract

Inference, prediction, and control of complex dynamical systems from time series is important in many areas, including financial markets, power grid management, climate and weather modeling, or molecular dynamics. The analysis of such highly nonlinear dynamical systems is facilitated by the fact that we can often find a (generally nonlinear) transformation of the system coordinates to features in which the dynamics can be excellently approximated by a linear Markovian model. Moreover, the large number of system variables often change collectively on large time- and length-scales, facilitating a low-dimensional analysis in feature space. In this paper, we introduce a variational approach for Markov processes (VAMP) that allows us to find optimal feature mappings and optimal Markovian models of the dynamics from given time series data. The key insight is that the best linear model can be obtained from the top singular components of the Koopman operator. This leads to the definition of a family of score functions called VAMP-r which can be calculated from data, and can be employed to optimize a Markovian model. In addition, based on the relationship between the variational scores and approximation errors of Koopman operators, we propose a new VAMP-E score, which can be applied to cross-validation for hyper-parameter optimization and model selection in VAMP. VAMP is valid for both reversible and nonreversible processes and for stationary and nonstationary processes or realizations.

Keywords Koopman operator · Variational approach · Markov process · Data-driven methods

Mathematics Subject Classification 37M10 · 37L65 · 47N30 · 65K10

Communicated by Dr. Paul Newton.

Extended author information available on the last page of the article

🌀 Springer

1 Introduction

Extracting dynamical models and their main characteristics from time series data is a recurring problem in many areas of science and engineering. In the particularly popular approach of Markovian models, the future evolution of the system, e.g., state $\mathbf{x}_{t+\tau}$, only depends on the current state $\mathbf{x}_t$, where t is the time step and τ is the delay or lag time. Markovian models are easier to analyze than models with explicit memory terms. They are justified by the fact that many physical processes—including both deterministic and stochastic processes—are inherently Markovian. Even when only a subset of the variables in which the system is Markovian are observed, a variety of physics and engineering processes have been shown to be accurately modeled by Markovian models on sufficiently long observation lag times τ. Examples include molecular dynamics (Chodera and Noé 2014; Prinz et al. 2011), wireless communications (Konrad et al. 2001; Ma et al. 2001) and fluid dynamics (Mezić 2013; Froyland et al. 2016).

In the past decades, a collection of closely related Markov modeling methods were developed in different fields, including Markov state models (MSMs) (Schütte et al. 1999; Prinz et al. 2011; Bowman et al. 2014), Markov transition models (Wu and Noé 2015), Ulam's Galerkin method (Dellnitz et al. 2001; Bollt and Santitissadeekorn 2013; Froyland et al. 2014), blind-source separation (Molgedey and Schuster 1994; Ziehe and Müller 1998), the variational approach of conformation dynamics (VAC) (Noé and Nüske 2013; Nüske et al. 2014), time-lagged independent component analysis (TICA) (Perez-Hernandez et al. 2013; Schwantes and Pande 2013), dynamic mode decomposition (DMD) (Rowley et al. 2009; Schmid 2010; Tu et al. 2014), extended dynamic mode decomposition (EDMD) (Williams et al. 2015a), variational Koopman models (Hao et al. 2017), variational diffusion maps (Boninsegna et al. 2015), sparse identification of nonlinear dynamics (Brunton et al. 2016b) and corresponding kernel embeddings (Harmeling et al. 2003; Song et al. 2013; Schwantes and Pande 2015) and tensor formulations (Nüske et al. 2016; Klus and Schütte 2015). All these models approximate the Markov dynamics at a lag time τ by a linear model in the following form:

$$\mathbb{E}\left[\mathbf{g}(\mathbf{x}_{t+\tau})\right] = \mathbf{K}^{\top}\mathbb{E}\left[\mathbf{f}(\mathbf{x}_t)\right]. \tag{1}$$

Here $\mathbf{f}(\mathbf{x}) = (f_1(\mathbf{x}), f_2(\mathbf{x}), \ldots)^{\top}$ and $\mathbf{g}(\mathbf{x}) = (g_1(\mathbf{x}), g_2(\mathbf{x}), \ldots)^{\top}$ are feature transformations that transform the state variables $\mathbf{x}$ into the feature space in which the dynamics are approximately linear. $\mathbb{E}$ denotes an expectation value over time that accounts for stochasticity in the dynamics and can be omitted for deterministic dynamical systems. In some methods, such as DMD, the feature transformation is an identity transformation: $\mathbf{f}(\mathbf{x}) = \mathbf{g}(\mathbf{x}) = \mathbf{x}$—and then Eq. (1) defines a linear dynamical system in the original state variables. If $\mathbf{f}$ and $\mathbf{g}$ are indicator functions that partition Ω into substates, such that $f_i(\mathbf{x}) = g_i(\mathbf{x}) = 1$ if $\mathbf{x} \in A_i$ and 0 otherwise, Eq. (1) is the propagation law of an MSM, or equivalently of Ulam's Galerkin method, as the expectation values $\mathbb{E}[\mathbf{f}(\mathbf{x}_t)]$ and $\mathbb{E}\left[\mathbf{g}(\mathbf{x}_{t+\tau})\right]$ represent the vector of probabilities to be in any substate at times t and $t + \tau$, and K_{ij} is the probability to transition from set A_i to set A_j in time τ. In general, (1) can be interpreted as a finite-rank approximation of the so-called Koopman operator (Koopman 1931; Mezić 2005), which governs the time evolution

of observables of the system state and can fully characterize the Markovian dynamics. As shown in Korda and Mezić (2018), this approximation becomes exact in the limit of infinitely sized feature transformations with $\mathbf{f} = \mathbf{g}$, and a similar conclusion can also be obtained when $\mathbf{f}$, $\mathbf{g}$ are infinite-dimensional feature functions deduced from a characteristic kernel (Song et al. 2013).

A direct method to estimate the matrix $\mathbf{K}$ from data is to solve the linear regression problem $\mathbf{g}(\mathbf{x}_{t+\tau}) \approx \mathbf{K}^{\top}\mathbf{f}(\mathbf{x}_t)$, which facilitates the use of regularized solution methods, such as the LASSO method (Tibshirani 1996). Alternatively, feature functions $\mathbf{f}$ and $\mathbf{g}$ that allow Eq. (1) to have a probabilistic interpretation (e.g., in MSMs), $\mathbf{K}$ can be estimated by a maximum likelihood or Bayesian method (Prinz et al. 2011; Noé 2008).

However, as yet, it is still unclear *what are the optimal choices for* $\mathbf{f}$ *and* $\mathbf{g}$—*either given a fixed dimension or a fixed amount of data*. Notice that this problem cannot be solved by minimizing the regression error of Eq. (1), because a regression error of zero can be trivially achieved by choosing a completely uninformative model with $\mathbf{f}(\mathbf{x}) \equiv \mathbf{g}(\mathbf{x}) \equiv 1$ and $\mathbf{K} = 1$. An approach that can be applied to deterministic systems and for stochastic systems with additive white noise is to set $\mathbf{g}(\mathbf{x}) = \mathbf{x}$, and then choose $\mathbf{f}$ as the transformation with smallest modeling error (Brunton et al. 2016a, b).

A more general approach is to optimize the dominant spectrum of the Koopman operator. At long timescales, the dynamics of the system are usually dominated by the Koopman eigenfunctions of the Koopman operator with large eigenvalues. If the dynamics obey detailed balance, those eigenvalues are real-valued, and the variational approach for reversible Markov processes can be applied that has made great progress in the field of molecular dynamics (Noé and Nüske 2013; Nüske et al. 2014). In such processes, the smallest modeling error of (1) is achieved by setting $\mathbf{f} = \mathbf{g}$ equal to the corresponding eigenfunctions. Noé and Nüske (2013) describes a general approach to approximate the unknown eigenfunction from time series data of a reversible Markov process: Given a set of orthogonal candidate functions, $\mathbf{f}$, it can be shown that their time-autocorrelations are lower bounds to the corresponding Koopman eigenvalues, and are equal to them exactly if, and only if $\mathbf{f}$ are equal to the Koopman eigenfunctions. This approach provides a variational score, such as the sum of estimated eigenvalues (the Rayleigh trace), that can be optimized to approximate the eigenfunctions. If $\mathbf{f}$ is defined by a linear superposition of a given set of basis functions, then the optimal coefficients are found equivalently by either maximizing the variational score, or minimizing the regression error in the feature space as done in EDMD (Williams et al. 2015a)—see Hao et al. (2017). However, the regression error cannot be used to select the form and the number of basis functions themselves, whereas the variational score can. When working with a finite dataset, however, it is important to avoid overfitting, and to this end a cross-validation method has been proposed to compute variational scores that take the statistical error into account (McGibbon and Pande 2015). Such cross-validated variational scores can be used to determine the size and type of the function classes and the other hyper-parameters of the dynamical model.

While this approach is extremely powerful for stationary and data and reversible Markov processes, almost all real-world dynamical processes and time series thereof are irreversible and often even nonstationary. In this paper, we introduce a variational

approach for Markov processes (VAMP) that can be employed to optimize parameters and hyper-parameters of arbitrary Markov processes. VAMP is based on the singular value decomposition of the Koopman operator, which overcomes the limited usefulness of the eigenvalue decomposition of time-irreversible and nonstationary processes. We first show that the approximation error of the Koopman operator deduced from the linear model (1) can be minimized by setting $\mathbf{f}$ and $\mathbf{g}$ to be the top left and right singular functions of the Koopman operator. Then, by using the variational description of singular components, a class of variational scores, VAMP-r for $r = 1, 2, \ldots$, are proposed to measure the similarity between the estimated singular functions and the true ones. Maximization of any of these variational scores leads to optimal model parameters and is algorithmically identical to Canonical Correlation Analysis (CCA) between the featurized time-lagged pair of variables $\mathbf{x}_t$ and $\mathbf{x}_{t+\tau}$. This approach can also be employed to learn the feature transformations by nonlinear function approximators, such as deep neural networks. Furthermore, we establish a relationship between the VAMP-2 score and the approximation error of the dynamical model with respect to the true Koopman operator. We show that this approximation error can be practically computed up to a constant, and define its negative as the VAMP-E score. Finally, we demonstrate that optimizing the VAMP-E score in a cross-validation framework leads to an optimal choice of hyper-parameters.

2 Theory

2.1 Koopman Analysis of Dynamical Systems and Its Singular Value Decomposition

The Koopman operator $\mathcal{K}_\tau$ of a Markov process is a linear operator defined by

$$\mathcal{K}_\tau g(\mathbf{x}) \triangleq \mathbb{E}\left[g(\mathbf{x}_{t+\tau}) \mid \mathbf{x}_t = \mathbf{x}\right]. \tag{2}$$

For given $\mathbf{x}_t$, the Koopman operator can be used to compute the conditional expected value of an arbitrary observable g at time $t + \tau$. For the special choice that g is the Dirac delta function $\delta_\mathbf{y}$ centered at $\mathbf{y}$, application of the Koopman operator evaluates the transition density of the dynamics, $\mathcal{K}_\tau \delta_\mathbf{y}(\mathbf{x}) = \mathbb{P}(\mathbf{x}_{t+\tau} = \mathbf{y}|\mathbf{x}_t = \mathbf{x})$ (see Appendix A.3). Thus, the Koopman operator is a complete description of the dynamical properties of a Markovian system. For convenience of analysis, we consider here $\mathcal{K}_\tau$ as a mapping from $\mathcal{L}_{\rho_1}^2 = \left\{g| \langle g, g \rangle_{\rho_1} < \infty\right\}$ to $\mathcal{L}_{\rho_0}^2 = \left\{f| \langle f, f \rangle_{\rho_0} < \infty\right\}$, where ρ_0 and ρ_1 are empirical distributions of $\mathbf{x}_t$ and $\mathbf{x}_{t+\tau}$ of all transition pairs $\{(\mathbf{x}_t, \mathbf{x}_{t+\tau})\}$ occurring in the given time series (see Appendix A.1), and the inner products are defined by

$$\langle f, g \rangle_{\rho_0} = \int f(\mathbf{x}) g(\mathbf{x}) \rho_0(\mathbf{x}) \, d\mathbf{x}, \quad \langle f, g \rangle_{\rho_1} = \int f(\mathbf{x}) g(\mathbf{x}) \rho_1(\mathbf{x}) \, d\mathbf{x}. \tag{3}$$

How is the finite-dimensional linear model (1) related to the Koopman operator description? Let us consider $\mathbf{f}(\mathbf{x}_t)$ to be a sufficient statistics for $\mathbf{x}_t$, and let $\mathbf{g}$ be a dictionary of observables, then the value of an arbitrary observable h in the subspace

of $\mathbf{g}$, i.e., $h = \mathbf{c}^\top \mathbf{g}$, with some coefficients $\mathbf{c}$, can be predicted from $\mathbf{x}_t$ as $\mathbb{E}\left[h(\mathbf{x}_{t+\tau})|\mathbf{x}_t\right]$ $= \mathbf{c}^\top \mathbf{K}^\top \mathbf{f}(\mathbf{x}_t)$. This implies that Eq. (1) is an algebraic representation of the projection of the Koopman operator onto the subspace spanned by functions $\mathbf{f}$ and $\mathbf{g}$, and the matrix $\mathbf{K}$ is therefore called the Koopman matrix. Combining this insight with the generalized Eckart–Young Theorem (Hsing and Eubank 2015) leads to our first result, namely what is the optimal choice of functions $\mathbf{f}$ and $\mathbf{g}$:

Theorem 1 Optimal approximation of Koopman operator. *If $\mathcal{K}_\tau$ is a Hilbert–Schmidt operator between the separable Hilbert spaces $\mathcal{L}^2_{\rho_1}$ and $\mathcal{L}^2_{\rho_0}$, the linear model (1) with the smallest modeling error in Hilbert–Schmidt norm is given by $\mathbf{f} = (\psi_1, \ldots, \psi_k)^\top$, $\mathbf{g} = (\phi_1, \ldots, \phi_k)^\top$ and $\mathbf{K} = \mathrm{diag}(\sigma_1, \ldots, \sigma_k)$, i.e.,*

$$\mathbb{E}\left[\phi_i(\mathbf{x}_{t+\tau})\right] = \sigma_i \mathbb{E}\left[\psi_i(\mathbf{x}_t)\right], \quad \text{for } i = 1, \ldots, k \tag{4}$$

under the constraint $\dim(\mathbf{f}), \dim(\mathbf{g}) \le k$, and the projected Koopman operator deduced from (4) is

$$\hat{\mathcal{K}}_\tau g = \sum_{i=1}^{k} \sigma_i \langle g, \phi_i \rangle_{\rho_1} \psi_i, \tag{5}$$

where the singular value $\sigma_i > 0$ is the square root of the ith largest eigenvalue of $\mathcal{K}_\tau^* \mathcal{K}_\tau$ or $\mathcal{K}_\tau \mathcal{K}_\tau^*$, the left and right singular function ψ_i, ϕ_i are the ith eigenfunctions of $\mathcal{K}_\tau^* \mathcal{K}_\tau$ and $\mathcal{K}_\tau \mathcal{K}_\tau^*$ with

$$\left\langle \psi_i, \psi_j \right\rangle_{\rho_0} = 1_{i=j}, \quad \left\langle \phi_i, \phi_j \right\rangle_{\rho_1} = 1_{i=j}, \tag{6}$$

and the first singular component is always given by $(\sigma_1, \phi_1, \psi_1) = (1, \mathbb{1}, \mathbb{1})$ with $\mathbb{1}(\mathbf{x}) \equiv 1$.

Proof See Appendix A.2.   □

This theorem is universal for Markov processes, and the major assumption is that the Koopman operator is Hilbert–Schmidt, which is required for the existences of the singular value decomposition (SVD) of $\mathcal{K}_\tau$ and the finite Hilbert–Schmidt norm $\|\mathcal{K}_\tau\|_{\mathrm{HS}}$. Appendix A.4 provides two sufficient conditions for the assumption. However, it is worth noting that the Koopman operators of deterministic systems are not Hilbert–Schmidt or even compact in usual cases (see Appendix A.5), and thus all conclusions and methods in this paper are not applicable to deterministic systems.

In addition, we prove in Appendix A.3 that $\|\hat{\mathcal{K}}_\tau - \mathcal{K}_\tau\|_{\mathrm{HS}}$ is equal to a weighted $\mathcal{L}^2$ error of the transition density, which provides a more meaningful interpretation of the modeling error in Hilbert–Schmidt norm.

Example 1 Consider a one-dimensional dynamical system

$$x_{t+1} = \frac{x_t}{2} + \frac{7x_t}{1 + 0.12x_t^2} + 6\cos x_t + \sqrt{10}u_t \tag{7}$$

evolving in the state space $[-20, 20]$, where u_t is a standard Gaussian white noise zero mean and unit variance (see Appendix K.1 for details on the numerical simulations
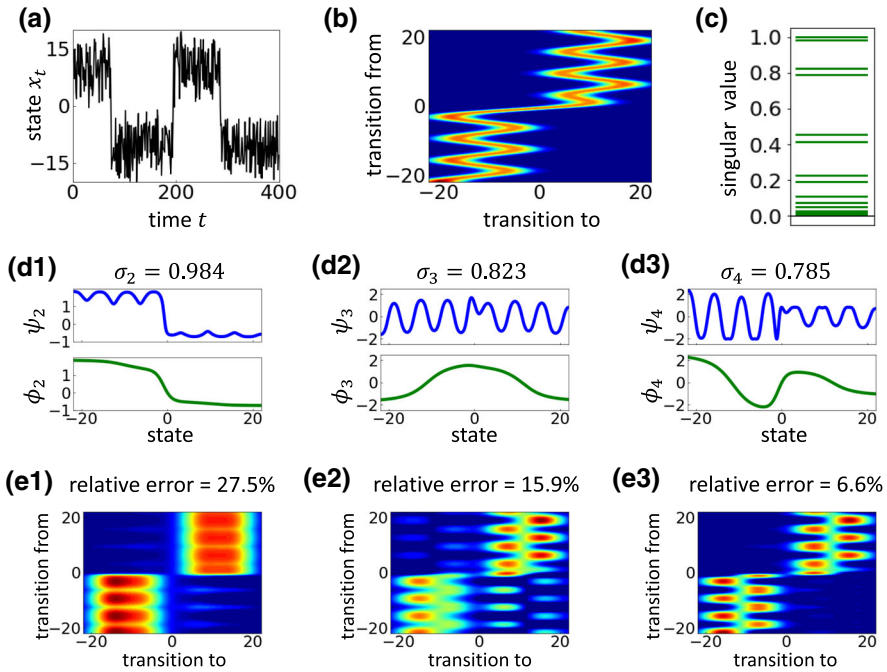
Fig. 1 Analysis results of the dynamical system (7) with lag time $\tau = 1$. **a** A typical simulation trajectory. **b** Transition density $\mathbb{P}(x_{t+1}|x_t)$. **c** The singular values. **d** The first three nontrivial left and right singular functions. [The first singular component is $(\sigma_1, \phi_1, \psi_1) = (1, \mathbb{1}, \mathbb{1})$.] **e** Approximate transition densities obtained from the projected Koopman operator $\hat{\mathcal{K}}_\tau$ consisting of first k singular components defined by (5) for $k = 2, 3, 4$, where the relative error is calculated as $\|\hat{\mathcal{K}}_\tau - \mathcal{K}_\tau\|_{\mathrm{HS}}/\|\mathcal{K}_\tau\|_{\mathrm{HS}}$

and analysis). This system has two metastable states with the boundary close to $x = 0$ as shown in Fig. 1a, and the singular components are summarized in Fig. 1c, d. As shown in the figures, the sign structures of the second left and right singular functions clearly indicate the metastable states, and the third and forth singular functions provide more detailed information on the dynamics. An accurate estimate of the transition density can be obtained by combining the first four singular components, and the corresponding relative approximation error of the Koopman operator is only 6.6% (see Fig. 1b, e). In addition, we utilize the finite-rank approximate Koopman operators to predict the time evolution of the distribution of x_t for $t = 1, \ldots, 256$ with the initial state $x_0 = 12$, and a small error can also be achieved when the rank is only 4 as displayed in Fig. 2, where

$$\text{error} = \sum_{t=1}^{256} \int \rho_1(x_t)^{-1} \left(\hat{\mathbb{P}}(x_t|x_0) - \mathbb{P}(x_t|x_0) \right)^2 \mathrm{d}x_t \qquad (8)$$

is the cumulative kinetic distance (Noé and Clementi 2015) between the transition density $\mathbb{P}(x_t|x_0)$ and its estimate $\hat{\mathbb{P}}(x_t|x_0)$, and ρ_1 is the stationary distribution.
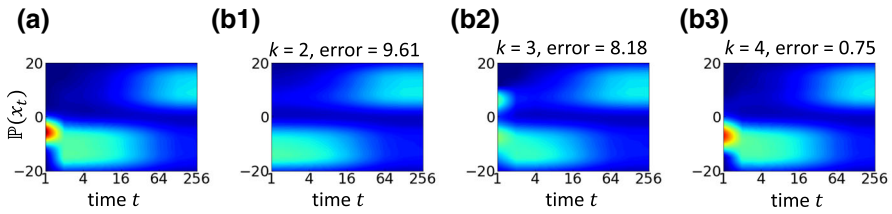
Fig. 2 Probability density of state x_t predicted by **a** the full model and **b** the projected Koopman operator $\hat{\mathcal{K}}_\tau$ with rank $k = 2, 3, 4$, where the initial state is $x_0 = -12$

There are other formalisms to describe Markovian dynamics, for example, the Markov propagator or the weighted Markov propagator, also called transfer operator (Schütte et al. 1999). These propagators are commonly used for modeling physical processes such as molecular dynamics, and describe the evolution of probability densities instead of observables. We show in Appendix B that all conclusions in this paper can be equivalently established by interpreting $(\sigma_i, \rho_1\phi_i, \rho_0\psi_i)$ as the singular components of the Markov propagator.

2.2 Variational Principle for Markov Processes

In order to allow the optimal model (4) to be estimated from data, we develop a variational principle for the approximation of singular values and singular functions of Markov processes.

According to the Rayleigh variational principle of singular values, the first singular component maximizes the generalized Rayleigh quotient of $\mathcal{K}_\tau$ as

$$(\psi_1, \phi_1) = \arg\max_{f,g} \frac{\langle f, \mathcal{K}_\tau g\rangle_{\rho_0}}{\sqrt{\langle f, f\rangle_{\rho_0} \cdot \langle g, g\rangle_{\rho_1}}} \tag{9}$$

and the maximal value of the generalized Rayleigh quotient is equal to the first singular value $\sigma_1 = \langle\psi_1, \mathcal{K}_\tau\phi_1\rangle_{\rho_0}$. For the ith singular component with $i > 1$, we have

$$(\psi_i, \phi_i) = \arg\max_{f,g} \frac{\langle f, \mathcal{K}_\tau g\rangle_{\rho_0}}{\sqrt{\langle f, f\rangle_{\rho_0} \cdot \langle g, g\rangle_{\rho_1}}} \tag{10}$$

under constraints

$$\langle f, \psi_j\rangle_{\rho_0} = \langle g, \phi_j\rangle_{\rho_1} = 0, \quad \forall j = 1, \ldots, i-1 \tag{11}$$

and the maximal value is equal to $\sigma_i = \langle\psi_i, \mathcal{K}_\tau\phi_i\rangle_{\rho_0}$. These insights can be summarized by the following variational theorem for seeking all top k singular components simultaneously:

Theorem 2 *VAMP variational principle*. *The k dominant singular components of a Koopman operator are the solution of the following maximization problem:*

$$\sum_{i=1}^{k} \sigma_i^r = \max_{\mathbf{f},\mathbf{g}} \mathcal{R}_r\left[\mathbf{f},\mathbf{g}\right],$$
$$s.t. \left\langle f_i, f_j \right\rangle_{\rho_0} = 1_{i=j},$$
$$\left\langle g_i, g_j \right\rangle_{\rho_1} = 1_{i=j}, \tag{12}$$

where $r \geq 1$ can be any positive integer. The maximal value is achieved by the singular functions $f_i = \psi_i$ and $g_i = \phi_i$ and

$$\mathcal{R}_r\left[\mathbf{f},\mathbf{g}\right] = \sum_{i=1}^{k} \langle f_i, \mathcal{K}_\tau g_i \rangle_{\rho_0}^r \tag{13}$$

is called the VAMP-r score of $\mathbf{f}$ and $\mathbf{g}$.

Proof See Appendix C. □

This theorem generalizes Proposition 2 in Froyland (2013) where only the case of $k = 2$ is considered. It is important to note that this theorem has direct implications for the data-driven estimation of dynamical models. For $r = 1$, $\mathcal{R}_r\left[\mathbf{f},\mathbf{g}\right]$ is actually the time correlation between $\mathbf{f}(\mathbf{x}_t)$ and $\mathbf{g}(\mathbf{x}_{t+\tau})$ since $\langle f_i, \mathcal{K}_\tau g_i \rangle_{\rho_0} = \mathbb{E}_t[f_i(\mathbf{x}_t)g_i(\mathbf{x}_{t+\tau})]$ and $\mathbb{E}_t[\cdot]$ denotes the expectation value over all transition pairs $(x_t, x_{t+\tau})$ in the time series. Hence the maximization of VAMP-*r* is analogous to the problem of seeking orthonormal transformations of $\mathbf{x}_t$ and $\mathbf{x}_{t+\tau}$ with maximal time-correlations, and we can thus utilize the canonical correlation analysis (CCA) algorithm (Hardoon et al. 2004) in order to estimate the singular components from data.

2.3 Comparison with Related Analysis Approaches

The SVD of the Koopman operator is equivalent to the eigenvalue decomposition when the Markov process is time-reversible and stationary with $\rho_0 = \rho_1$, and therefore the variational principle presented here is a generalization of that developed for reversible conformation dynamics (Noé and Nüske 2013; Nüske et al. 2014). Specifically, VAMP-1 maximizes the *Rayleigh trace*, i.e., the sum of the estimated eigenvalues (Noé and Nüske 2013; McGibbon and Pande 2015), and VAMP-2 maximizes the *kinetic variance* introduced in Noé and Clementi (2015). See Appendix D for a detailed derivation of the reversible variational principle from the VAMP variational principle. For irreversible Markov processes, the singular functions can provide low-dimensional embeddings of kinetic distances between states like eigenfunctions of reversible processes (Paul et al. 2018). Furthermore, the coherent sets of nonstationary Markov processes, which are the generalization of metastable states, can be identified from dominant singular functions (Koltai et al. 2018).

The dynamics of an irreversible Markov process can also be analyzed through solving the eigenvalue problem $\mathcal{K}_\tau g = \lambda g$ [see, e.g., Williams et al. (2015a, b), Klus et al. (2015),Klus and Schütte (2015)], and the eigenfunctions form an invariant subspace of the Koopman operator for multiple lag times since the eigenvalue problem satisfies

$$\mathcal{K}_\tau g = \lambda g \Rightarrow \mathcal{K}_{n\tau} g = \lambda^n g, \quad \forall n \geq 1. \tag{14}$$

However, as far as we know, there is no variational principle for approximate eigenfunctions of irreversible Markov processes, and it is difficult to evaluate errors of projections of Koopman operators to the invariant subspaces. The SVD-based analysis approach overcomes the above problems and yields the optimal finite-rank approximate models. The major limitation of this approach comes from the fact that the singular functions are dependent on the choice of the lag time and the optimality of model (4) holds only for a fixed τ. The optimization and error analysis of Koopman models for multiple lag times will be studied in our future work.

3 Estimation Algorithms

We introduce algorithms to estimate optimal dynamical models from time series data. We make the Ansatz to represent the feature functions $\mathbf{f}$ and $\mathbf{g}$ as linear combinations of *basis* functions $\boldsymbol{\chi}_0 = (\chi_{0,1}, \chi_{0,2}, \ldots)^\top$ and $\boldsymbol{\chi}_1 = (\chi_{1,1}, \chi_{1,2}, \ldots)^\top$:

$$\begin{aligned} \mathbf{f} &= \mathbf{U}^\top \boldsymbol{\chi}_0, \\ \mathbf{g} &= \mathbf{V}^\top \boldsymbol{\chi}_1. \end{aligned} \tag{15}$$

Here, $\mathbf{U}$ and $\mathbf{V}$ are matrices of size $m \times k$ and $m' \times k$, i.e., we are trying to approximate k singular components by linearly combining m and m' basis functions. For the sake of generality we have assumed that $\mathbf{f}$ and $\mathbf{g}$ are represented by different basis sets. However, in practice one can justify using a single basis set the joint set $\boldsymbol{\chi}^\top = (\boldsymbol{\chi}_0^\top, \boldsymbol{\chi}_1^\top)$ as an Ansatz for both $\mathbf{f}$ and $\mathbf{g}$. Please note that despite the linear Ansatz (15), the feature functions may be strongly nonlinear in the system's state variables $\mathbf{x}$, thus we are not restricting the generality of the functions $\mathbf{f}$ and $\mathbf{g}$ that can be represented. In this section, we consider three problems: (i) optimizing $\mathbf{U}$ and $\mathbf{V}$, (ii) optimizing $\boldsymbol{\chi}_0$ and $\boldsymbol{\chi}_1$ and (iii) assessing the quality of the resulting dynamical model.

For convenience of notation, we denote by $\mathbf{C}_{00}, \mathbf{C}_{11}, \mathbf{C}_{01}$ the covariance matrices and time-lagged covariance matrices of basis functions, which can be computed from a trajectory $\{x_1, \ldots, x_T\}$ by

$$\mathbf{C}_{00} \triangleq \mathbb{E}_t \left[\boldsymbol{\chi}_0 (\mathbf{x}_t) \, \boldsymbol{\chi}_0 (\mathbf{x}_t)^\top \right] \approx \frac{1}{T - \tau} \sum_{t=1}^{T-\tau} \boldsymbol{\chi}_0 (\mathbf{x}_t) \, \boldsymbol{\chi}_0 (\mathbf{x}_t)^\top, \tag{16}$$

$$\mathbf{C}_{11} \triangleq \mathbb{E}_t \left[\boldsymbol{\chi}_1 (\mathbf{x}_{t+\tau}) \, \boldsymbol{\chi}_1 (\mathbf{x}_{t+\tau})^\top \right] \approx \frac{1}{T - \tau} \sum_{t=1+\tau}^{T} \boldsymbol{\chi}_1 (\mathbf{x}_t) \, \boldsymbol{\chi}_1 (\mathbf{x}_t)^\top, \tag{17}$$

$$\mathbf{C}_{01} \triangleq \mathbb{E}_t \left[\boldsymbol{\chi}_0 \left(\mathbf{x}_t \right) \boldsymbol{\chi}_1 \left(\mathbf{x}_{t+\tau} \right)^\top \right] \approx \frac{1}{T-\tau} \sum_{t=1}^{T-\tau} \boldsymbol{\chi}_0 \left(\mathbf{x}_t \right) \boldsymbol{\chi}_1 \left(\mathbf{x}_{t+\tau} \right)^\top . \tag{18}$$

If there are multiple trajectories, the covariance matrices can be computed in the same manner by averaging over all trajectories. Instead of the direct estimators (16–18), more elaborated estimation methods such as regularization methods (Tibshirani 1996) and reweighting estimators (Hao et al. 2017) may be used.

3.1 Feature TCCA: Finding the Best Linear Model in a Given Feature Space

We first propose a solution for the problem of finding the optimal parameter matrices $\mathbf{U}$ and $\mathbf{V}$ given that the basis functions $\boldsymbol{\chi}_0$ and $\boldsymbol{\chi}_1$ are known. Substituting the linear Ansatz (15) into the VAMP variational principle shows that $\mathbf{U}$ and $\mathbf{V}$ can be computed as the solutions of the maximization problem:

$$\max_{\mathbf{U},\mathbf{V}} \mathcal{R}_r (\mathbf{U}, \mathbf{V})$$
$$\text{s.t.} \mathbf{U}^\top \mathbf{C}_{00} \mathbf{U} = \mathbf{I}$$
$$\mathbf{V}^\top \mathbf{C}_{11} \mathbf{V} = \mathbf{I}, \tag{19}$$

where

$$\mathcal{R}_r (\mathbf{U}, \mathbf{V}) = \sum_{i=1}^{k} \left(\mathbf{u}_i^\top \mathbf{C}_{01} \mathbf{v}_i \right)^r \tag{20}$$

is a matrix representation of VAMP-r score, and $\mathbf{u}_i$ and $\mathbf{v}_i$ are the ith columns of $\mathbf{U}$ and $\mathbf{V}$. This problem can be solved by applying linear CCA (Hardoon et al. 2004) in the feature spaces defined by the basis sets $\boldsymbol{\chi}_0(\mathbf{x}_t)$ and $\boldsymbol{\chi}_1(\mathbf{x}_{t+\tau})$, and the same solution will be obtained for any other choice of r. (See Appendices E.1 and E.2 for more detailed proof and analysis.) The resulting algorithm for finding the best linear model is a CCA in feature space, applied on time-lagged data. Hence we briefly call this algorithm feature TCCA:

1. Compute covariance matrices $\mathbf{C}_{00}, \mathbf{C}_{01}, \mathbf{C}_{11}$ via (16–18).
2. Perform the truncated SVD

$$\bar{\mathbf{K}} = \mathbf{C}_{00}^{-\frac{1}{2}} \mathbf{C}_{01} \mathbf{C}_{11}^{-\frac{1}{2}} \approx \mathbf{U}' \mathbf{K} \mathbf{V}'^\top ,$$

where $\bar{\mathbf{K}}$ is the Koopman matrix for the normalized basis functions $\mathbf{C}_{00}^{-\frac{1}{2}} \boldsymbol{\chi}_0$ and $\mathbf{C}_{11}^{-\frac{1}{2}} \boldsymbol{\chi}_1, \mathbf{K} = \operatorname{diag}(K_{11}, \dots, K_{kk})$ is a diagonal matrix of the first k singular values that approximate the true singular values $\sigma_1, \dots, \sigma_k$, and $\mathbf{U}'$ and $\mathbf{V}'$ consist of the k corresponding left and right singular vectors respectively.
3. Compute $\mathbf{U} = \mathbf{C}_{00}^{-\frac{1}{2}} \mathbf{U}'$ and $\mathbf{V} = \mathbf{C}_{11}^{-\frac{1}{2}} \mathbf{V}'$.

4. Output the linear model (1) with K_{ii}, $f_i = \mathbf{u}_i^\top \boldsymbol{\chi}_0$ and $g_i = \mathbf{v}_i^\top \boldsymbol{\chi}_1$ being the estimates of the ith singular value, left singular function and right singular function of the Koopman operator.

Please note that this pseudocode is given only for illustrative purposes and cannot be executed literally if $\mathbf{C}_{00}$ and $\mathbf{C}_{11}$ do not have full rank, i.e., are not invertible. To handle this problem, we ensure that the basis functions are linearly independent by applying a decorrelation (whitening) transformation that ensures that $\mathbf{C}_{00}$ and $\mathbf{C}_{11}$ will both have full rank. We then add the constant function $\mathbb{1}(x) \equiv 1$ to the decorrelated basis sets to ensure that $\mathbb{1}$ belongs to the subspaces spanned by $\boldsymbol{\chi}_0$ and by $\boldsymbol{\chi}_1$. It can be shown that the singular values given by the feature TCCA algorithm with these numerical modifications are bounded by 1, and the first estimated singular component is exactly $(K_{11}, f_1, g_1) = (1, \mathbb{1}, \mathbb{1})$ even in the presence of statistical noise and modeling error—see Appendix F.1 for details.

In the case of $k = \dim(\boldsymbol{\chi}_0) = \dim(\boldsymbol{\chi}_1)$ and full rank $\mathbf{C}_{00}, \mathbf{C}_{11}$, the output of the feature TCCA can be equivalently written as

$$
\mathbb{E}\left[\mathbf{V}^\top \boldsymbol{\chi}_1 (\mathbf{x}_{t+\tau})\right] = \mathbf{K}^\top \mathbb{E}\left[\mathbf{U}^\top \boldsymbol{\chi}_0 (\mathbf{x}_t)\right]
$$
$$
\Rightarrow \mathbb{E}\left[\boldsymbol{\chi}_1 (\mathbf{x}_{t+\tau})\right] = \mathbf{K}_\chi^\top \mathbb{E}\left[\boldsymbol{\chi}_0 (\mathbf{x}_t)\right] \tag{21}
$$

where

$$
\mathbf{K}_\chi = \mathbf{U} \mathbf{K} \mathbf{V}^{-1}
$$
$$
= \mathbf{C}_{00}^{-1} \mathbf{C}_{01} \tag{22}
$$

is equal to the least square solution to the regression problem $\boldsymbol{\chi}_1 (\mathbf{x}_{t+\tau}) \approx \mathbf{K}_\chi^\top \boldsymbol{\chi}_0 (\mathbf{x}_t)$. Note that if we further assume that $\boldsymbol{\chi}_0 = \boldsymbol{\chi}_1$, (21) is identical to the linear model of EDMD. Thus, the feature TCCA can be seen as a generalization of EDMD that can provide approximate Markov models for different basis $\boldsymbol{\chi}_0$ and $\boldsymbol{\chi}_1$. More discussion on the relationship between the two methods is provided in Appendix G.

3.2 Nonlinear TCCA: Optimizing the Basis Functions

We now extend feature TCCA to a more flexible representation of the transformation functions $\mathbf{f}$ and $\mathbf{g}$ by optimizing the basis functions themselves:

$$
\mathbf{f}(\mathbf{x}) = \mathbf{U}^\top \boldsymbol{\chi}_0 (\mathbf{x}; \mathbf{w}),
$$
$$
\mathbf{g}(\mathbf{x}) = \mathbf{V}^\top \boldsymbol{\chi}_1 (\mathbf{x}; \mathbf{w}). \tag{23}
$$

Here, $\mathbf{w}$ represents a set of parameters that determines the form of the basis functions. As a simple example, consider $\mathbf{w}$ to represent the mean vectors and covariance matrices of a Gaussian basis set. However, $\boldsymbol{\chi}_0 (\mathbf{x}; \mathbf{w})$ and $\boldsymbol{\chi}_1 (\mathbf{x}; \mathbf{w})$ can also represent very complex and nonlinear learning structures, such as neural networks and decision trees.

The parameters $\mathbf{w}$ could conceptually be determined together with the linear expansion coefficients $\mathbf{U}$, $\mathbf{V}$ by solving (19) with $\mathbf{C}_{00}, \mathbf{C}_{11}, \mathbf{C}_{01}$ treated as functions of $\mathbf{w}$,

but this method is not practical due to the nonlinear equality constraints are involved. In practice, we can set k to be $\min\{\dim(\chi_0), \dim(\chi_1)\}$, i.e., the largest number of singular components that can be approximated given the basis set. Then the maximal VAMP-r score for a fixed $\mathbf{w}$ can be represented as

$$\max_{\mathbf{U},\mathbf{V}} \mathcal{R}_r = \left\| \mathbf{C}_{00}(\mathbf{w})^{-\frac{1}{2}} \mathbf{C}_{01}(\mathbf{w}) \mathbf{C}_{11}(\mathbf{w})^{-\frac{1}{2}} \right\|_r^r, \tag{24}$$

which can also be interpreted as the sum over the r'th power of all singular values of the projected Koopman operator on subspaces of χ_0, χ_1 (see Eq. (80) in Appendix E.1 and Eq. (84) in Appendix E.2). Here $\|\mathbf{A}\|_r$ denotes the r-Schatten norm of matrix $\mathbf{A}$, which is the ℓ^r norm of singular values of $\mathbf{A}$, and $\|\mathbf{A}\|_2$ equals the Frobenius norm of $\mathbf{A}$. The parameters $\mathbf{w}$ can be optimized without computing $\mathbf{U}$ and $\mathbf{V}$ explicitly. Using these ideas, nonlinear TCCA can be performed as follows:

1. Compute $\mathbf{w}^* = \arg\max_{\mathbf{w}} \left\| \mathbf{C}_{00}(\mathbf{w})^{-\frac{1}{2}} \mathbf{C}_{01}(\mathbf{w}) \mathbf{C}_{11}(\mathbf{w})^{-\frac{1}{2}} \right\|_r^r$ by gradient descent or other nonlinear optimization methods.
2. Approximate the Koopman singular values and singular functions using the feature TCCA algorithm with basis sets $\chi_0(\mathbf{x}; \mathbf{w}^*)$ and $\chi_1(\mathbf{x}; \mathbf{w}^*)$.

Unlike the estimated singular components generated by the feature TCCA, the estimation results of the nonlinear TCCA do generally depend on the value of r. (An example is given in Appendix E.3, where the VAMP scores can be analytically computed.) We suggest to set $r = 2$ in applications for the direct relationship between the VAMP-2 score and the approximation error of Koopman operators and the convenience of cross-validation (see below). The details of the nonlinear TCCA, including the optimization algorithm and regularization, are beyond the scope of this paper. Appendix F.2 provides a brief description of the implementation, and related work based on kernel methods and deep networks can be found in Andrew et al. (2013) and Mardt et al. (2018).

Example 2 Let us consider the stochastic system described in Example 1 again. We generate 10 simulation trajectories of length 500 and approximate the dominant singular components by the feature TCCA. Here, the basis functions are

$$\chi_{0,i}(x) = \chi_{1,i}(x) = 1_{\frac{40 \cdot (i-1)}{m} - 20 \leq x \leq \frac{40 \cdot i}{m} - 20}, \quad \text{for } i = 1, \ldots, m, \tag{25}$$

which define a partition of the domain $[-20, 20]$ into $m = 33$ disjoint intervals. In other words, the approximation is performed based on an MSM with 33 discrete states. Estimation results are given in Fig. 3a, where the discretization errors arising from indicator basis functions are clearly shown. For comparison, we also implement the nonlinear TCCA algorithm with radial basis functions

$$\chi_{0,i}(x; w) = \chi_{1,i}(x; w) = \frac{\exp\left(-w(x - c_i)^2\right)}{\sum_{j=1}^{m} \exp\left(-w(x - c_j)^2\right)} \tag{26}$$

with smoothing parameter $w \geq 0$, where $c_i = \frac{40 \cdot (i - 0.5)}{m} - 20$ for $i = 1, \ldots, m$ are uniformly distributed in $[-20, 20]$. Notice that the basis functions given in (25) are
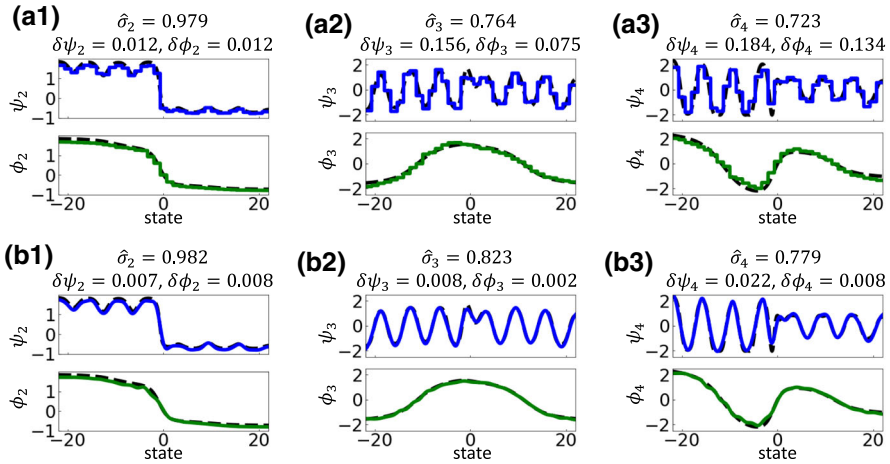
Fig. 3 Estimated singular components of the system in Example 1, where dash lines represent true singular functions, and the estimation errors of singular functions are defined as $\delta\psi_i = \int (f_i(x) - \psi_i(x))^2 \rho_0(x)\mathrm{d}x$, $\delta\phi_i = \int (g_i(x) - \phi_i(x))^2 \rho_1(x)\mathrm{d}x$ with $\rho_0 = \rho_1$ being the stationary distribution. **a** Estimates provided by feature TCCA with basis functions (25). **b** Estimates provided by nonlinear TCCA with basis functions (26)

a specific case of the radial basis functions with $w = \infty$, and it is therefore possible to achieve better approximation by optimizing w. As can be seen from Fig. 3b, the nonlinear TCCA provides more accurate estimates of singular functions and singular values (see Appendix K.1 for more details). In addition, both feature TCCA and nonlinear TCCA underestimate the dominant singular values as stated by the variational principle.

The nonlinear TCCA is similar to the EDMD with dictionary learning (EDMD-DL) (Li et al. 2017), where the feature transformations are optimized by minimizing the regression error of (1). The major advantages of the nonlinear TCCA over EDMD-DL are: First, the uninformative model with zero regression error can be systematically excluded without any extra constraints on features. Second, the optimization objective is directly related to the approximation error of the Koopman operator (see Sect. 3.3). Some recent methods extend EDMD-DL for modeling Koopman operators of deterministic systems (Takeishi et al. 2017; Lusch et al. 2018; Otto and Rowley 2019), and solve the first problem by using the prediction error between the observed x_t and that predicted by the low-dimensional model. But they cannot be applied to stochastic Koopman operators of Markov processes directly.

3.3 Error Analysis

According to (5), both feature TCCA and nonlinear TCCA lead to a rank k approximation

$$\hat{\mathcal{K}}_\tau g = \sum_{i=1}^{k} K_{ii} \langle g, g_i\rangle_{\rho_1} f_i = \sum_{i=1}^{k} K_{ii} \left\langle g, \mathbf{v}_i^\top \boldsymbol{\chi}_1 \right\rangle_{\rho_1} \mathbf{u}_i^\top \boldsymbol{\chi}_0 \qquad (27)$$

to $\mathcal{K}_\tau$. We consider here the approximation error of (27) in a general case where $\mathbf{f} = \mathbf{U}^\top \boldsymbol{\chi}_0$ and $\mathbf{g} = \mathbf{V}^\top \boldsymbol{\chi}_1$ may not satisfy the orthonormal constraints due to statistical noise and numerical errors. After a few steps of derivation, the approximation error can be expressed as

$$\left\| \hat{\mathcal{K}}_\tau - \mathcal{K}_\tau \right\|_{\mathrm{HS}}^2 = -\mathcal{R}_E[\mathbf{K}, \mathbf{f}, \mathbf{g}] + \|\mathcal{K}_\tau\|_{\mathrm{HS}}^2 \tag{28}$$

with

$$\mathcal{R}_E[\mathbf{K}, \mathbf{f}, \mathbf{g}] = 2 \sum_i K_{ii} \langle f_i, \mathcal{K}_\tau g_i \rangle_{\rho_0} - \sum_{i,j} K_{ii} K_{jj} \langle f_i, f_j \rangle_{\rho_0} \langle g_i, g_j \rangle_{\rho_1}. \tag{29}$$

Remarkably, this error decomposes into a unknown constant part (the square of Hilbert–Schmidt norm of $\mathcal{K}_\tau$), and a model-dependent part $\mathcal{R}_E$ that can be entirely estimated from data by its matrix representation:

$$\mathcal{R}_E(\mathbf{K}, \mathbf{U}, \mathbf{V}) = \mathrm{tr} \left[2\mathbf{K}\mathbf{U}^\top \mathbf{C}_{01} \mathbf{V} - \mathbf{K}\mathbf{U}^\top \mathbf{C}_{00} \mathbf{U}\mathbf{K}\mathbf{V}^\top \mathbf{C}_{11} \mathbf{V} \right]. \tag{30}$$

$\mathcal{R}_E$, is thus a score that can be used alternatively to the VAMP-r scores, and we call $\mathcal{R}_E$ VAMP-E score. It can be proved that the maximization of $\mathcal{R}_E$ is equivalent to maximization of $\mathcal{R}_2$ in feature TCCA or nonlinear TCCA. However, these scores will behave differently in terms of hyper-parameter optimization (see Sect. 4.1). Proofs and analysis are given in Appendix H.

4 Model Validation

4.1 Cross-Validation for Hyper-parameter Optimization

For a data-driven estimation of dynamical models, either using feature TCCA or nonlinear TCCA, we have to strike a balance between the modeling or discretization error and the statistical or overfitting error. The choice of number and type of basis functions is critical for both. If basis sets are very small and not flexible enough to capture singular functions, the approximation results may be inaccurate with large biases. We can improve the variational score and reduce the modeling error by larger and more flexible basis sets. But too complicated basis sets will produce unstable estimates with large statistical variances, and in particular poor predictions on data that has not been used in the estimation process—this problem is known as *overfitting* in the machine learning community. A popular way to achieve the balance between the statistical bias and variance are resampling methods, including bootstrap and cross-validation (Friedman et al. 2001). They iteratively fit a model in a training set, which are sampled from the data with or without replacement, and validate the model in the complementary dataset. Alternatively, there are also Bayesian hyper-parameter optimization methods. See Arlot and Celisse (2010) and Snoek et al. (2012) for an

overview. Here, we will focus on cross-validation and describe how to use the VAMP scores in this and similar resampling frameworks.

Let $\boldsymbol{\theta}$ be hyper-parameters in feature TCCA or nonlinear TCCA that need to be specified. For example, $\boldsymbol{\theta}$ includes the number and functional form of basis functions used in feature TCCA, or the architecture and connectivity of a neural network used for nonlinear TCCA. Generally speaking, different values of $\boldsymbol{\theta}$ correspond to different dynamical models that we want to rank, and these models may be of completely different types. The cross-validation of $\boldsymbol{\theta}$ can be performed as follows:

1. Separate the available trajectories into J disjoint folds $\mathcal{D}_1, \ldots, \mathcal{D}_J$ with approximately equal size. If there are only a small number of long trajectories, we can divide each trajectory into blocks of length L with $\tau < L \ll T$ and create folds based on the blocks. This defines a number of J training sets, with training set j consisting of all data except the jth fold, $\mathcal{D}_j^{\text{train}} = \cup_{l \neq j} \mathcal{D}_l$, and the jth fold used as test set $\mathcal{D}_j^{\text{test}} = \mathcal{D}_j$.

2. For each hyper-parameter set $\boldsymbol{\theta}$:

   (a) For $j = 1, \ldots, J$:

       i. *Train on $\mathcal{D}_j^{\text{train}}$*: training set $\mathcal{D}_j^{\text{train}}$, construct the best k-dimensional linear model consisting of $(\mathbf{K}, \mathbf{U}^\top \boldsymbol{\chi}_0, \mathbf{V}^\top \boldsymbol{\chi}_1)$ by applying the feature TCCA or nonlinear TCCA with hyper-parameters $\boldsymbol{\theta}$

       ii. *Validate on $\mathcal{D}_{\text{test}}$*: measure the performance of the estimated singular components by a score

$$\text{CV}_j(\boldsymbol{\theta}) = \text{CV}(\mathbf{K}, \mathbf{U}, \mathbf{V} | \mathcal{D}_{\text{test}}) \tag{31}$$

   (b) Compute cross-validation score

$$\text{MCV}(\boldsymbol{\theta}) = \frac{1}{J} \sum_{j=1}^{J} \text{CV}_j(\boldsymbol{\theta}) \tag{32}$$

3. Select model/hyper-parameter set with maximal MCV $(\boldsymbol{\theta})$.

The key to the above procedure is how to evaluate the estimated singular components for given test set. It is worth pointing out that we *cannot* simply define the validation score directly as the VAMP-r score of estimated singular functions for the test data, because the singular functions obtained from training data are usually not orthonormal with respect to the test data.

A feasible way is to utilize the subspace variational score as proposed for reversible Markov processes in McGibbon and Pande (2015). For VAMP-r this score becomes:

$$\text{CV}(\mathbf{K}, \mathbf{U}, \mathbf{V} | \mathcal{D}_{\text{test}}) = \mathcal{R}_r^{\text{space}}(\mathbf{U}, \mathbf{V} | \mathcal{D}_{\text{test}})$$
$$= \left\| \left(\mathbf{U}^\top \mathbf{C}_{00}^{\text{test}} \mathbf{U}\right)^{-\frac{1}{2}} \left(\mathbf{U}^\top \mathbf{C}_{01}^{\text{test}} \mathbf{V}\right) \left(\mathbf{V}^\top \mathbf{C}_{11}^{\text{test}} \mathbf{V}\right)^{-\frac{1}{2}} \right\|_r^r, \quad (33)$$

where $\mathcal{R}_r^{\text{space}}$ measures the consistency between the singular subspace and the estimated one without the constraint of orthonormality. However, this scheme suffers

from the following limitations in practical applications: Firstly, the value of k must be chosen a priori and kept fixed during the cross-validation procedure, which implies that models with a different number of singular components cannot be compared by the validation scores. Secondly, computation of the validation score possibly suffers from numerical instability. (See Appendix I for detailed analysis.)

We suggest in this paper to perform the cross-validation based on the approximation error of Koopman operators. According to conclusions in Sect. 3.3, feature TCCA and VAMP-2 base nonlinear TCCA both maximize the VAMP-E score $\mathcal{R}_E (\mathbf{K}, \mathbf{U}, \mathbf{V}|\mathcal{D}_{\text{train}})$ for a given training set $\mathcal{D}_{\text{train}}$.

Therefore, we can score the performance of estimated singular components on the test set by

$$\text{CV} (\mathbf{K}, \mathbf{U}, \mathbf{V}|\mathcal{D}_{\text{test}}) = \mathcal{R}_E (\mathbf{K}, \mathbf{U}, \mathbf{V}|\mathcal{D}_{\text{test}}) . \tag{34}$$

In contrast with the validation score (33) deduced from the subspace VAMP-r score, the validation score defined by (34) allows us to choose k according to practical requirements: If we are only interested in a small number of dominant singular components, we can select a fixed value of k. If we want to evaluate the statistical performance of the approximate model consisting of all available estimated singular components as in the EDMD method, we can set $k = \min\{\dim(\boldsymbol{\chi}_0), \dim(\boldsymbol{\chi}_1)\}$. We can even view k as a hyper-parameter and select a suitable rank of the model via cross-validation. Another advantage of the VAMP-E based validation score is that it does not involve any inverse operation of matrices and can be stably computed.

It is worth pointing out that a validation score is proposed Kurebayashi et al. (2016) for cross-validation of kernel DMD based on the analysis of approximation error of transition densities, which has a similar form to that of VAMP-E. The theoretical and empirical comparisons between the two scores will be performed in our future work.

Example 3 We consider here the choice of the basis function number m for the nonlinear TCCA in Example 2. We use five fold cross-validation with the VAMP-E score to compare different values of m. While the average score computed by training sets keeps increasing with m, both the cross-validation score and the exact VAMP-E score achieve their maximum value at $m = 33$ as in Example 2 (see Fig. 4a). The optimality can also be demonstrated by comparing Figs. 3b and 4b. A much smaller basis set with $m = 13$ yields large errors in the approximation of singular functions. When $m = 250$, the estimation of singular functions suffers from overfitting and the estimated singular value is even larger than the true value due to the statistical noise.

4.2 Chapman–Kolmogorov Test for Choice of Lag Times

Besides hyper-parameters mentioned in above, the lag time τ is also an essential parameter especially for time-continuous Markov processes. If $\tau \to 0$, the $\mathcal{K}_\tau$ is usually close to the identity operator and cannot be accurately approximated by a low-rank model, whereas a too high value of τ can cause the loss of kinetic information in data since $\mathbb{P}(\mathbf{x}_{t+\tau}|\mathbf{x}_t)$ is approximately independent of $\mathbf{x}_t$ in the case of ergodic processes. However, the variational approach presented in this paper is based on analysis of the approximation error of the Koopman operator for a fixed τ, so we cannot compare models with different lag times and choose τ by the VAMP scores.
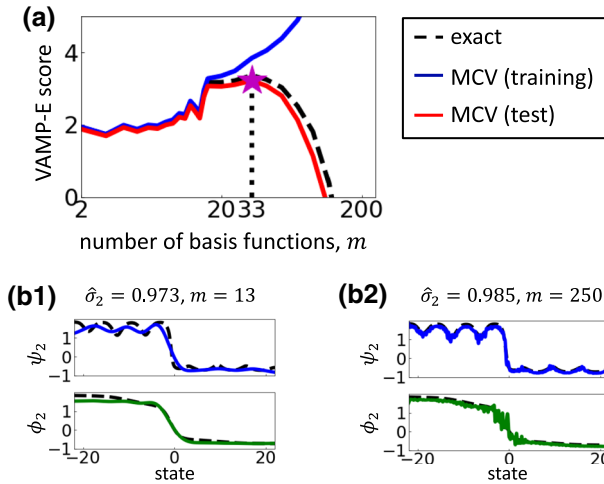
Fig. 4 Cross-validation for modeling the system in Example 1. **a** Cross-validated VAMP-E scores for the choice of the number of basis functions m. The black line indicates the exact VAMP-E score calculated according to the true model. Using cross-validation we compute the average VAMP-E scores computed from the training sets (blue) and the test sets (red). **b** Estimated ψ_2 and ϕ_2 obtained by the nonlinear TCCA with $m = 13$ and 250 (Color figure online)

In order to address this problem, the Chapman–Kolmogorov test can be used, which is common in building Markov state models (Prinz et al. 2011). Let us consider the covariance

$$
\begin{aligned}
\mathrm{cov}(f, g; n\tau) &\triangleq \langle f, \mathcal{K}_{n\tau} g \rangle_{\rho_0(n\tau)} \\
&= \mathbb{E}_{\mathbf{x}_t \sim \rho_0(n\tau)} \left[f(\mathbf{x}_t) g(\mathbf{x}_{t+n\tau}) \right]
\end{aligned} \tag{35}
$$

between observables f and g of lag time $n\tau$, which can be estimated from data as

$$
\mathrm{cov}^{\mathrm{emp}}(f, g; n\tau) = \frac{1}{T - n\tau} \sum_{t=1}^{T-n\tau} f(\mathbf{x}_t) g(\mathbf{x}_{t+n\tau})^\top, \tag{36}
$$

where $\rho_0(n\tau)$ is the empirical distribution of the simulation data excluding $\{x_t | t > T - n\tau\}$. If our methods provide an ideal Markov model of lag time τ, the Koopman operator $\mathcal{K}_{n\tau}$ can be approximated by $\hat{\mathcal{K}}_\tau^n$, and the covariance can also be predicted as

$$
\begin{aligned}
\mathrm{cov}^{\mathrm{pred}}(f, g; n\tau) &= \left\langle f, \hat{\mathcal{K}}_\tau^n g \right\rangle_{\rho_0(n\tau)} \\
&= \mathbb{E}_{\mathbf{x}_t \sim \rho_0(n\tau)} \left[f(\mathbf{x}_t) \chi_0(\mathbf{x}_t)^\top \right] \\
&\quad \cdot \mathbf{U} \mathbf{R}^{n-1} \mathbf{K} \mathbf{V}^\top \\
&\quad \cdot \mathbb{E}_{\mathbf{x}_t \sim \rho_1} \left[\chi_1(\mathbf{x}_t) g(\mathbf{x}_t) \right]
\end{aligned} \tag{37}
$$

(see Appendix J), where

$$\mathbf{R} = \mathbf{K} \cdot \mathbb{E}_{\mathbf{x}_t \sim \rho_1} \left[\mathbf{g}(\mathbf{x}_t)^\top \mathbf{f}(\mathbf{x}_t) \right]. \tag{38}$$

Therefore, the lag time τ can be selected according to the following criteria in applications: (i) The lag time is smaller than the timescale that we are interested in. (ii) The equation

$$\mathrm{cov}^{\mathrm{pred}}(f, g; n\tau) = \mathrm{cov}^{\mathrm{emp}}(f, g; n\tau) \tag{39}$$

holds approximately for multiple observables f, g and lag times $n\tau$. In this paper, we simply set f, g to be the estimated leading singular functions since they dominate the dynamics of the Markov process.

5 Numerical Examples

5.1 Double-Gyre System

Let's consider a stochastic double-gyre system defined by:

$$\begin{aligned}
\mathrm{d}x_t &= -\pi \, A \, \sin(\pi \, x_t) \, \cos(\pi \, y_t) \, \mathrm{d}t + \varepsilon \sqrt{x_t/4 + 1} \, \mathrm{d}\mathbf{W}_{t,1}, \\
\mathrm{d}y_t &= \quad \pi \, A \, \cos(\pi \, x_t) \, \sin(\pi \, y_t) \, \mathrm{d}t + \varepsilon \, \mathrm{d}\mathbf{W}_{t,2},
\end{aligned} \tag{40}$$

where $\mathbf{W}_{t,1}$ and $\mathbf{W}_{t,2}$ are two independent standard Wiener processes. The dynamics are defined on the domain $[0, 2] \times [0, 1]$ with reflecting boundary. For $\varepsilon = 0$, it can be seen from the flow field depicted in Fig. 5a that there is no transport between the left half and the right half of the domain and both subdomains are invariant sets with measure $\frac{1}{2}$ (Froyland and Padberg 2009; Froyland and Padberg-Gehle 2014). For $\varepsilon > 0$, there is a small amount of transport due to diffusion and the subdomains are almost invariant. Here we used the parameters $A = 0.25$, $\epsilon = 0.1$, and lag time $\tau = 2$ in analysis and simulations. The first two nontrivial singular components are shown in Fig. 5c, where the two almost invariant sets are clearly visible in ψ_2, ϕ_2 and ψ_3, ϕ_3 are associated with the rotational kinetics within the almost invariant sets.

We generate 10 trajectories of length 4 with step size 0.02, and perform modeling by nonlinear TCCA with basis functions

$$\chi_{0,i}(x, y; w) = \chi_{1,i}(x, y; w) = \frac{\exp\left(-w \left\| (x, y)^\top - \mathbf{c}_i \right\|^2\right)}{\sum_{j=1}^{m} \exp\left(-w \left\| (x, y)^\top - \mathbf{c}_j \right\|^2\right)}, \quad \text{for } i = 1, \dots, m \tag{41}$$

where $\mathbf{c}_1, \dots, \mathbf{c}_m$ are cluster centers given by k-means algorithm, and the smoothing parameter w is determined via maximizing the VAMP-2 score given in (24) (see Appendix K.2 for more details of numerical computations). The size of basis set $m = 37$ is selected by the VAMP-E based cross-validation proposed in 4.1 with 5 folds (see Fig. 5b), and it can be observed from Fig. 5c–e that the leading singular components are accurately estimated. In contrast, as shown in Fig. 5f, g, a much small value of m leads to significant approximation errors of singular components, while for a much larger value, the estimates are obviously influenced by statistical noise. Figure 6 illustrates that the
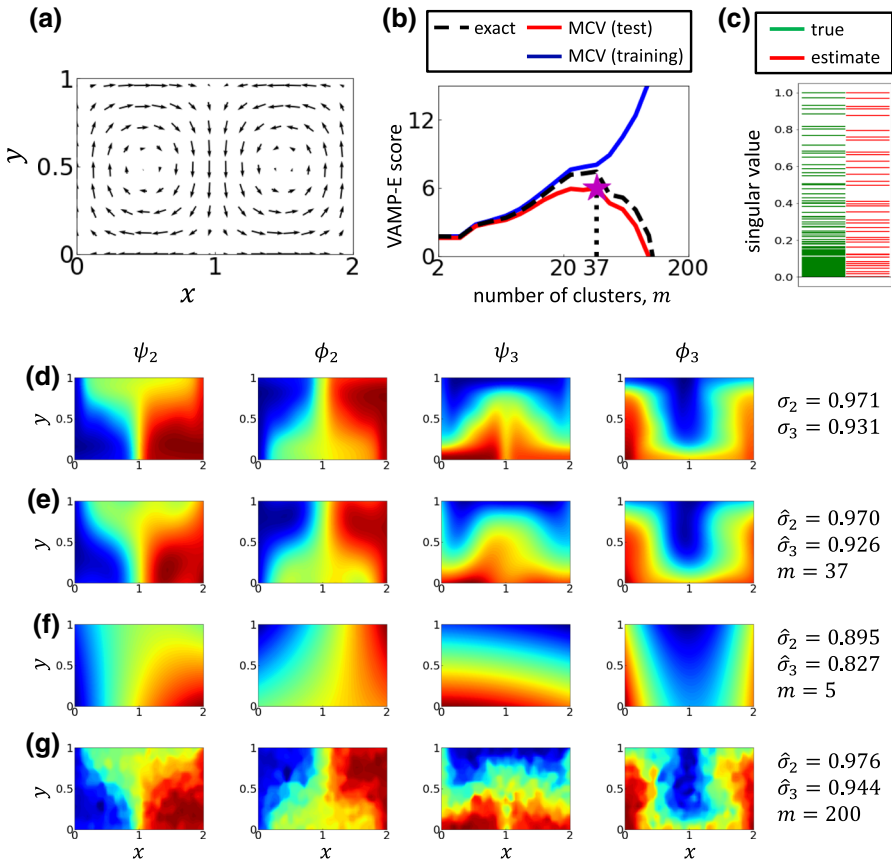
Fig. 5 Modeling of the double-gyre system (40). **a** Flow field of the system, where the arrows represent directions and magnitudes of $(\mathrm{d}x_t, \mathrm{d}y_t)$ with $\epsilon = 0$. **b** VAMP-E scores of estimated models obtained from the train sets, test sets and true model respectively. The largest MCV on test sets and exact VAMP-E score are both achieved with $m = 37$ basis functions. **c** The true singular values and estimated ones given by the nonlinear TCCA with $m = 37$. **d** The first two nontrivial singular components. **e–g** The estimated singular components obtained by the nonlinear TCCA with $m = 37$, 5 and 200

Koopman operator estimated by nonlinear TCCA can successfully predict the long-time evolution of the distribution of the state. The Chapman–Kolmogorov test results are displayed in Fig. 7, which confirm that $\tau = 2$ is a suitable choice of the lag time.

5.2 Stochastic Lorenz System

As the last example, we investigate the stochastic Lorenz system which obeys the following stochastic differential equation:

$$
\begin{aligned}
\mathrm{d}x_t &= s(y - x)\,\mathrm{d}t + \epsilon x_t\,\mathrm{d}\mathbf{W}_{t,1}, \\
\mathrm{d}y_t &= (rx_t - y_t - x_t z_t)\,\mathrm{d}t + \varepsilon y_t\,\mathrm{d}\mathbf{W}_{t,2}, \\
\mathrm{d}z_t &= (-bz_t + x_t y_t)\,\mathrm{d}t + \varepsilon z_t\,\mathrm{d}\mathbf{W}_{t,2},
\end{aligned}
\tag{42}
$$
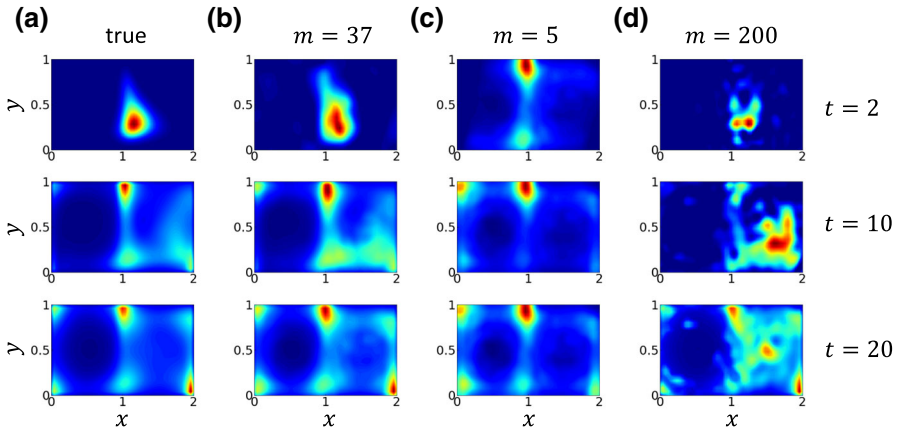
Fig. 6 Probability density of state (x_t, y_t) of the double-gyre system predicted by **a** the full simulation model and **b** the estimated Koopman operator obtained by the nonlinear TCCA with $m = 37, 5$ and 200, where the initial state is $(x_0, y_0) = (1.48, 0.8)$
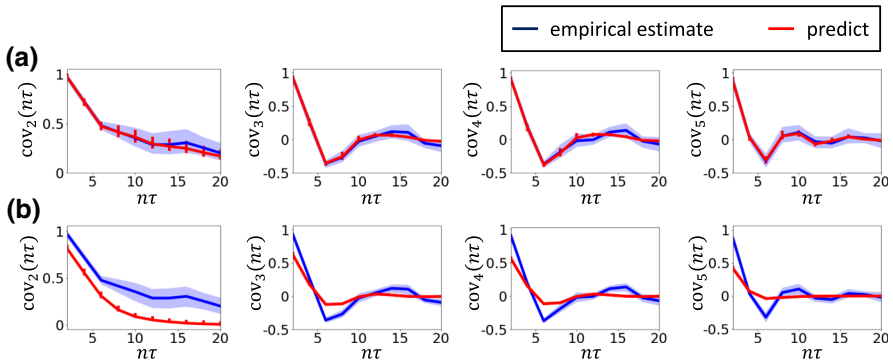
Fig. 7 Chapman–Kolmogorov test for modeling the double-gyre system by nonlinear TCCA with **a** $\tau = 2$ and **b** $\tau = 0.1$. The number of basis functions $m = 37$ for both cases, $\mathrm{cov}_i(n\tau) = \mathrm{cov}(\hat{\psi}_i, \hat{\phi}_i; n\tau)$ is the time-lagged covariance between $\hat{\psi}_i, \hat{\phi}_i$ as defined in (35), and $\hat{\psi}_i, \hat{\phi}_i$ are the ith singular functions estimated with $\tau = 2$. Blue lines indicate empirical covariances directly calculated from data, red lines indicate the predicted values given by $\hat{\mathcal{K}}_\tau$ as in (37), and error bars represent standard deviations calculated from 100 bootstrapping replicates of simulation data (Color figure online)

with parameters $s = 10$, $r = 28$ and $b = 8/3$. The deterministic Lorenz system with $\epsilon = 0$ is known to exhibit chaotic behavior (Sparrow 1982) with a strange attractor characterized by two lobes as illustrated in Fig. 8a. We generate 20 trajectories of length 25 with $\epsilon = 0.3$ by using the Euler–Maruyama scheme with step size 0.005, and one of them is shown in Fig. 8b. As stated in Chekroun et al. (2011), all the trajectories move around the deterministic attractor with small random perturbations and switch between the two lobes.

The leading singular components computed from the simulation data by the nonlinear TCCA are summarized in Fig. 8c, where the lag time $\tau = 0.75$ is determined via the Chapman–Kolmogorov test (see Fig. 8d), $\chi_0 = \chi_1$ consist of m normalized radial basis functions similar to those used in Sect. 5.1, and the selection of m is
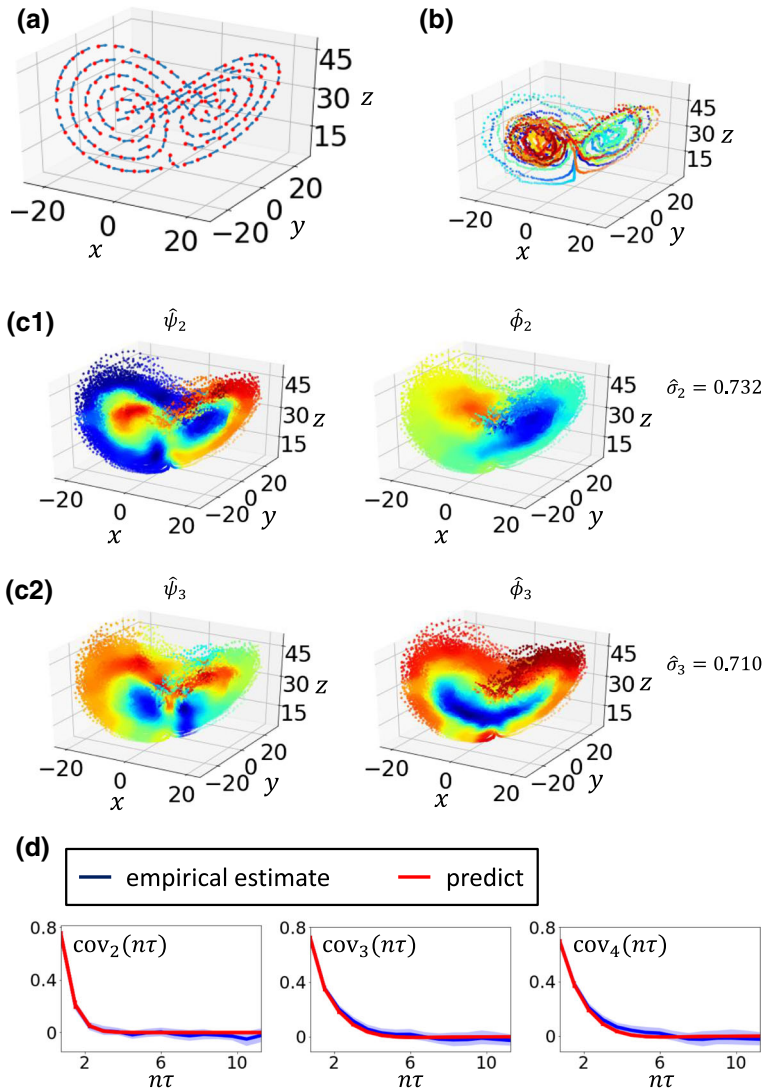
Fig. 8 Modeling of the stochastic Lorenz system (42). **a** Flow field of the system, where the arrows represent the mean directions of (dx_t, dy_t, dz_t). (b) A typical trajectory with $\epsilon = 0.3$ generated by the Euler–Maruyama scheme, which is colored according to time (from blue to red). **c** The first two nontrivial singular components computed by nonlinear TCCA. **d** Chapman–Kolmogorov test results for $\tau = 0.75$, where $\text{cov}_i(n\tau) = \text{cov}(\hat{\psi}_i, \hat{\phi}_i; n\tau)$, where error bars represent standard deviations calculated from 100 bootstrapping replicates of simulation data (Color figure online)

also implemented by 5-fold cross-validation. According to the patterns of the singular functions, the stochastic Lorenz system can be coarse-grained into a simplified model which transitions between four macrostates corresponding to inner and outer basins of the two attractor lobes. In particular, the sign boundary of ψ_1 closely matches that between the almost invariant sets of the Lorenz flow (Froyland and Padberg 2009).
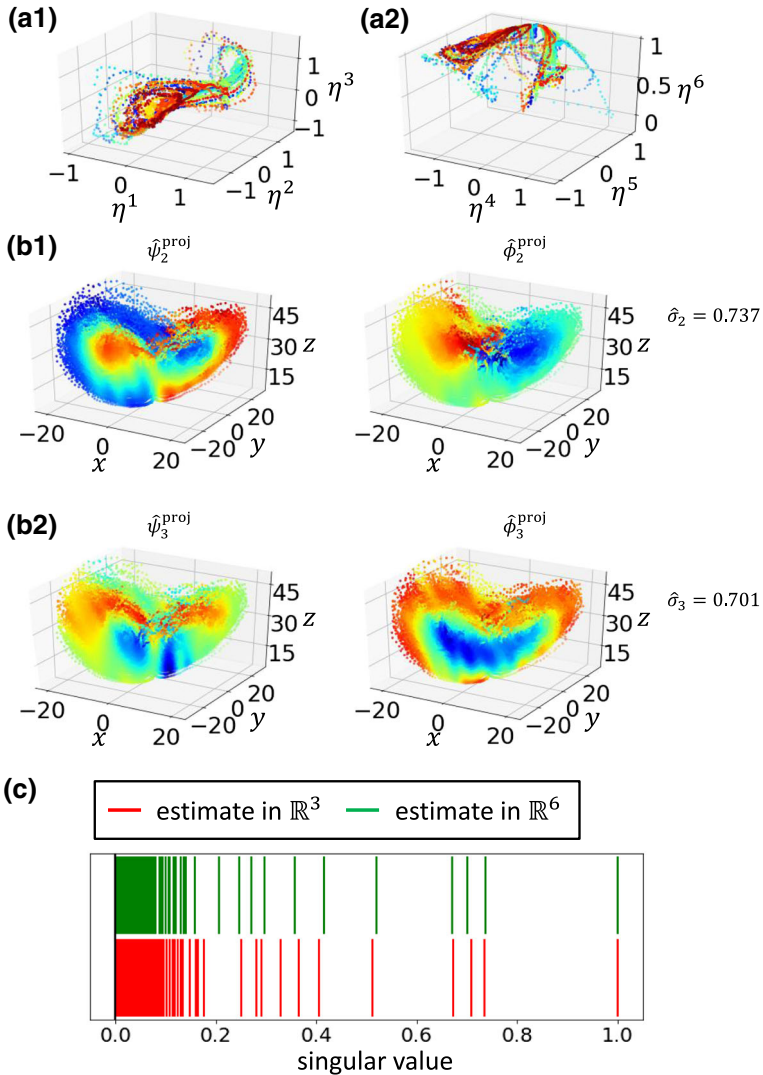
Fig. 9 Modeling of the stochastic Lorenz system (42) in the space of $\boldsymbol{\eta}_t$. **a** Plots of a typical trajectory in spaces of $(\eta_t^1, \eta_t^2, \eta_t^3)$ and $(\eta_t^4, \eta_t^5, \eta_t^6)$, which are colored according to time (from blue to red). Force field of the system. **b** The projected singular functions in the space of (x_t, y_t, z_t) computed by nonlinear TCCA. **c** The singular values estimated from trajectories of $(x_t, y_t, z_t) \in \mathbb{R}^3$ and $(\eta_t^1, \ldots, \eta_t^6) \in \mathbb{R}^6$ (Color figure online)

Next, we map the simulation data to a higher dimensional space via the nonlinear transformation $\boldsymbol{\eta}_t = \boldsymbol{\eta}(x_t, y_t, z_t)$ defined by

$$
\begin{aligned}
\eta_t^1 &= \left(\tfrac{z_t}{50} + \tfrac{1}{2}\right) \cos\left(\tfrac{\pi x_t}{30} + \tfrac{z_t}{50} - 1\right), & \eta_t^2 &= \left(\tfrac{z_t}{50} + \tfrac{1}{2}\right) \sin\left(\tfrac{\pi x_t}{30} + \tfrac{z_t}{50} - 1\right), \\
\eta_t^3 &= \left(\tfrac{z_t}{50} + \tfrac{1}{2}\right) \cos\left(\tfrac{\pi y_t}{30} + \tfrac{z_t}{50} - 1\right), & \eta_t^4 &= \left(\tfrac{z_t}{50} + \tfrac{1}{2}\right) \sin\left(\tfrac{\pi y_t}{30} + \tfrac{z_t}{50} - 1\right), \quad (43) \\
\eta_t^5 &= \cos\tfrac{\pi(x_t + y_t)}{40}, & \eta_t^6 &= \cos\tfrac{\pi(x_t - y_t)}{40}.
\end{aligned}
$$

Figure 9a plots the transformed points of the illustrative trajectory in Fig. 8b. We utilize the nonlinear TCCA to compute the singular components in the space of $\boldsymbol{\eta}_t = (\eta_t^1, \ldots, \eta_t^6)$ by assuming that the available observable is $\boldsymbol{\eta}_t$ instead of (x_t, y_t, z_t), show in Fig. 9b the projections of the singular functions back on the three-dimensional space

$$\psi_i^{\text{proj}}(x_t, y_t, z_t) = \psi_i(\boldsymbol{\eta}(x_t, y_t, z_t)), \quad \phi_i^{\text{proj}}(x_t, y_t, z_t) = \phi_i(\boldsymbol{\eta}(x_t, y_t, z_t)), \quad (44)$$

and compares the singular values estimated from trajectories of (x_t, y_t, z_t) and $(\eta_t^1, \ldots, \eta_t^6)$. It can be seen the projected leading singular components are almost the same as those directly computed from the three-dimensional data, which illustrates the transformation invariance of VAMP. Notice it is straightforward to prove that the exact ψ_i^{proj} and ϕ_i^{proj} are the solution to the variational problem (12) in the space of $(x_t, y_t, z_t)^\top$ if there is an inverse mapping $\boldsymbol{\eta}^{-1}$ with $\boldsymbol{\eta}^{-1}(\boldsymbol{\eta}(x_t, y_t, z_t)) \equiv (x_t, y_t, z_t)$.

6 Conclusion

The linearized coarse-grained models of Markov systems are commonly used in a broad range of fields, such as power systems, fluid mechanics and molecular dynamics. Although the models were developed independently in different communities, the VAMP proposed in this paper provides a general framework for analysis of them, and the modeling accuracy can be quantitatively evaluated by the VAMP-r and VAMP-E scores. Moreover, a set of data-driven methods, including feature TCCA, nonlinear TCCA and VAMP-E based cross-validation, are developed to achieve optimal modeling for given finite model dimensions and finite data sets.

The major challenge in real-world applications of VAMP is how to overcome the curse of dimensionality and solve the variational problem effectively and efficiently for high-dimensional systems. One feasible way of addressing this challenge is to approximate singular components by deep neural networks, which yields the concept of *VAMPnet* (Mardt et al. 2018). The optimal models can therefore be obtained by deep learning techniques. Another possible way is to utilize tensor decomposition based approximation approaches. Some tensor analysis methods have been presented based on the reversible variational principle and EDMD (Nüske et al. 2016; Klus and Schütte 2015; Klus et al. 2018), and it is worth studying more general variational tensor method within the framework of VAMP in future.

One drawback of the methods developed in this paper is that the resulting models are possibly not valid probabilistic models with nonnegative transition densities if only the operator error is considered, and the probability-preserving modeling method requires further investigations. Moreover, the applications of VAMP to detection of metastable states (Deuflhard and Weber 2005), coherent sets (Froyland and Padberg-Gehle 2014) and dominant cycles (Conrad et al. 2016) will also be explored in next steps.

Appendix

For convenience of notation, we denote by $p_\tau(\mathbf{x}, \mathbf{y}) = \mathbb{P}(\mathbf{x}_{t+\tau} = \mathbf{y}|\mathbf{x}_t = \mathbf{x})$ the transition density which satisfies

$$\int_A p_\tau(\mathbf{x}, \mathbf{y})\mathrm{d}\mathbf{y} = \mathbb{P}(\mathbf{x}_{t+\tau} \in A|\mathbf{x}_t = \mathbf{x}) \tag{45}$$

for every measurable set A, and define the matrix of scalar products:

$$\left\langle \mathbf{a}, \mathbf{b}^\top \right\rangle_\rho = \left[\langle a_i, b_j \rangle_\rho \right] \in \mathbb{R}^{m \times n} \tag{46}$$

$$\mathcal{K}\mathbf{g} = (\mathcal{K}g_1, \mathcal{K}g_2, \ldots)^\top \tag{47}$$

for $\mathbf{a} = (a_1, a_2, \ldots, a_m)^\top, \mathbf{b} = (b_1, b_2, \ldots, b_n)^\top$ and $\mathbf{g} = (g_1, g_2, \ldots)^\top$. In addition, $\mathcal{N}(\cdot|c, \sigma^2)$ denotes the probability density function of the normal distribution with mean c and variance σ^2.

A Analysis of Koopman Operators

A.1 Definition of Empirical Distributions

We first consider the case where the simulation data consist of S independent trajectories $\{\mathbf{x}_t^1\}_{t=1}^T, \ldots, \{\mathbf{x}_t^S\}_{t=1}^T$ of length T and the initial state $x_0^s \overset{\text{iid}}{\sim} p_0(\mathbf{x})$. In this case, ρ_0 and ρ_1 can be defined by

$$\rho_0 = \frac{1}{T-\tau} \sum_{t=1}^{T-\tau} \mathcal{P}_t p_0, \quad \rho_1 = \frac{1}{T-\tau} \sum_{t=1}^{T-\tau} \mathcal{P}_{t+\tau} p_0, \tag{48}$$

and they satisfy

$$\rho_1 = \mathcal{P}_\tau \rho_0, \tag{49}$$

where $\mathcal{P}_t$ denotes the Markov propagator defined in (63). We can then conclude that the estimates of $\mathbf{C}_{00}, \mathbf{C}_{11}, \mathbf{C}_{01}$ given by (16–18) are unbiased and consistent as $S \to \infty$.

In more general cases where trajectories $\{\mathbf{x}_t^1\}_{t=1}^{T_1}, \ldots, \{\mathbf{x}_t^S\}_{t=1}^{T_S}$ are generated with different initial conditions and different lengths, the similar conclusions can be obtained by defining ρ_0, ρ_1 as the averages of marginal distributions of $\{\mathbf{x}_t^s|1 \le t \le T_s - \tau, 1 \le s \le S\}$ and $\{\mathbf{x}_t^s|1 + \tau \le t \le T_s, 1 \le s \le S\}$ respectively.

A.2 Proof of Theorem 1

Because $\mathcal{K}_\tau$ is a Hilbert–Schmidt operator from $\mathcal{L}^2_{\rho_1}$ to $\mathcal{L}^2_{\rho_0}$, there exists the following SVD of $\mathcal{K}_\tau$:

$$\mathcal{K}_\tau g = \sum_{i=1}^{\infty} \sigma_i \langle g, \phi_i \rangle_{\rho_1} \psi_i. \tag{50}$$

Due to the orthonormality of right singular functions, the projection of any function $g \in \mathcal{L}^2_{\rho_1}$ onto the space spanned by $\{\phi_1, \ldots, \phi_k\}$ can be written as $\sum_{i=1}^{k} \langle g, \phi_i \rangle_{\rho_1} \phi_i$. Then $\hat{\mathcal{K}}_\tau$ defined by (5) is the approximate Koopman operator deduced from model (4), and it is the best rank k approximation to $\mathcal{K}_\tau$ in Hilbert–Schmidt norm according to the generalized Eckart–Young Theorem (see Theorem 4.4.7 in Hsing and Eubank (2015)).

Since the adjoint operator $\mathcal{K}^*_\tau$ of $\mathcal{K}_\tau$ satisfies

$$\begin{aligned}
\langle f, \mathcal{K}^*_\tau \mathbb{1} \rangle_{\rho_1} &= \langle \mathcal{K}_\tau f, \mathbb{1} \rangle_{\rho_0} \\
&= \int \mathbb{E}[f(\mathbf{x}_{t+\tau})|\mathbf{x}_t = \mathbf{x}]\rho_0(\mathbf{x})\mathrm{d}\mathbf{x} \\
&= \int \mathbb{E}[f(\mathbf{x})]\rho_1(\mathbf{x})\mathrm{d}\mathbf{x} \\
&= \langle f, \mathbb{1} \rangle_{\rho_1}
\end{aligned}$$

for all f, we can obtain

$$\mathcal{K}^*_\tau \mathbb{1} = \mathcal{K}_\tau \mathbb{1} = \mathbb{1}, \tag{51}$$

and conclude from Proposition 2 in Froyland (2013) that $(\sigma_1, \phi_1, \psi_1) = (1, \mathbb{1}, \mathbb{1})$.

A.3 Transition Densities Deduced from Koopman Operators

The Koopman operator can also be written as

$$\mathcal{K}_\tau g(\mathbf{x}) = \int p_\tau(\mathbf{x}, \mathbf{y})g(\mathbf{y})\mathrm{d}\mathbf{y} \tag{52}$$

if the transition density is given, which implies that

$$\mathcal{K}_\tau \delta_{\mathbf{y}}(\mathbf{x}) = p_\tau(\mathbf{x}, \mathbf{y}). \tag{53}$$

Then the transition density deduced from the approximate Koopman operator $\hat{\mathcal{K}}_\tau$ defined by (5) is

$$\begin{aligned}
\hat{p}_\tau(\mathbf{x}, \mathbf{y}) &= \hat{\mathcal{K}}_\tau \delta_{\mathbf{y}}(\mathbf{x}) \\
&= \sum_{i=1}^{k} \sigma_i \psi_i(\mathbf{x})\phi_i(\mathbf{y})\rho_1(\mathbf{y}). \tag{54}
\end{aligned}$$

From (52), we can show that

$$
\begin{aligned}
\|\mathcal{K}_\tau\|_{\mathrm{HS}}^2 &= \sum_i \langle \mathcal{K}_\tau \phi_i, \mathcal{K}_\tau \phi_i \rangle_{\rho_0} \\
&= \int \sum_i \left(\int p(\mathbf{x}, \mathbf{y}) \phi_i(\mathbf{y}) \mathrm{d}\mathbf{y} \right)^2 \rho_0(\mathbf{x}) \mathrm{d}\mathbf{x} \\
&= \int \sum_i \left(\int \frac{p(\mathbf{x}, \mathbf{y})}{\rho_1(\mathbf{y})} \cdot \phi_i(\mathbf{y}) \cdot \rho_1(\mathbf{y}) \mathrm{d}\mathbf{y} \right)^2 \rho_0(\mathbf{x}) \mathrm{d}\mathbf{x} \\
&= \int \left(\int \left(\frac{p(\mathbf{x}, \mathbf{y})}{\rho_1(\mathbf{y})} \right)^2 \cdot \rho_1(\mathbf{y}) \mathrm{d}\mathbf{y} \right) \rho_0(\mathbf{x}) \mathrm{d}\mathbf{x} \\
&= \iint \frac{\rho_0(\mathbf{x})}{\rho_1(\mathbf{y})} p(\mathbf{x}, \mathbf{y})^2 \mathrm{d}\mathbf{x}\mathrm{d}\mathbf{y},
\end{aligned}
\tag{55}
$$

and

$$
\left\| \hat{\mathcal{K}}_\tau - \mathcal{K}_\tau \right\|_{\mathrm{HS}}^2 = \iint \frac{\rho_0(\mathbf{x})}{\rho_1(\mathbf{y})} \left(\hat{p}(\mathbf{x}, \mathbf{y}) - p(\mathbf{x}, \mathbf{y}) \right)^2 \mathrm{d}\mathbf{x}\mathrm{d}\mathbf{y},
\tag{56}
$$

i.e., the operator error between $\hat{\mathcal{K}}_\tau$ and $\mathcal{K}_\tau$ can be represented by the error between $\hat{p}_\tau$ and p_τ.

It is worth pointing out that the approximate transition density in (54) satisfies the normalization constraint with

$$
\begin{aligned}
\int \hat{p}_\tau(\mathbf{x}, \mathbf{y}) \mathrm{d}\mathbf{y} &= \sum_{i=1}^k \sigma_i \psi_i(\mathbf{x}) \left\langle \phi_j, \mathbb{1} \right\rangle_{\rho_1} \\
&= \sigma_1 \psi_1(\mathbf{x}) \\
&\equiv 1,
\end{aligned}
\tag{57}
$$

but $\hat{p}_\tau(\mathbf{x}, \mathbf{y})$ is possibly negative for some $\mathbf{x}, \mathbf{y}$. Thus, the approximate Koopman operators and transition densities are not guaranteed to yield valid probabilistic models, although they can still be utilized to quantitative analysis of Markov processes.

A.4 Sufficient Conditions for Theorem 1

We show here $\mathcal{L}_{\rho_0}^2$, $\mathcal{L}_{\rho_1}^2$ are separable Hilbert spaces and $\mathcal{K}_\tau : \mathcal{L}_{\rho_1}^2 \mapsto \mathcal{L}_{\rho_0}^2$ is Hilbert–Schmidt if one of the following conditions is satisfied:

Condition 1 *The state space of the Markov process is a finite set.*

Proof The proof is trivial by considering $\mathcal{K}_\tau$ is a linear operator between finite-dimensional spaces, and thus omitted.   □

Condition 2 *The state space of the Markov process is $\mathbb{R}^d$, $\rho_0(\mathbf{x})$, $\rho_1(\mathbf{y})$ are positive for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, and there exists a constant M so that*

$$
p_\tau(\mathbf{x}, \mathbf{y}) \le M\rho_1(\mathbf{y}), \quad \forall \mathbf{x}, \mathbf{y}
\tag{58}
$$

Proof Let $\{e_1, e_2, \ldots\}$ be a orthonormal basis of $\mathcal{L}^2(\mathbb{R}^d)$. Then $\mathcal{L}^2_{\rho_0}, \mathcal{L}^2_{\rho_1}$ are separable because they have the countable orthonormal bases $\{\rho_0^{-\frac{1}{2}} e_1, \rho_0^{-\frac{1}{2}} e_2, \ldots\}$ and $\{\rho_1^{-\frac{1}{2}} e_1, \rho_1^{-\frac{1}{2}} e_2, \ldots\}$.

Now we prove that $\|\mathcal{K}_\tau\|_{\mathrm{HS}} < \infty$. Because

$$\iint \frac{\rho_0(\mathbf{x})}{\rho_1(\mathbf{y})} p_\tau(\mathbf{x}, \mathbf{y})^2 \mathrm{d}\mathbf{x} \mathrm{d}\mathbf{y} \leq \iint M \rho_0(\mathbf{x}) p_\tau(\mathbf{x}, \mathbf{y}) \mathrm{d}\mathbf{x} \mathrm{d}\mathbf{y}$$
$$= M, \tag{59}$$

the operator $\mathcal{S}$ defined by

$$\mathcal{S} f(\mathbf{x}) = \int \sqrt{\frac{\rho_0(\mathbf{x})}{\rho_1(\mathbf{y})}} p_\tau(\mathbf{x}, \mathbf{y}) f(\mathbf{y}) \mathrm{d}\mathbf{y} \tag{60}$$

is a Hilbert–Schmidt integral operator from $\mathcal{L}^2(\mathbb{R}^d)$ to $\mathcal{L}^2(\mathbb{R}^d)$ with $\|\mathcal{S}\|^2_{\mathrm{HS}} \leq M$ (Renardy and Rogers 2004). Therefore,

$$\|\mathcal{K}_\tau\|^2_{\mathrm{HS}} = \sum_i \left\langle \mathcal{K}_\tau \rho_1^{-\frac{1}{2}} e_i, \mathcal{K}_\tau \rho_1^{-\frac{1}{2}} e_i \right\rangle_{\rho_0}$$
$$= \sum_i \langle \mathcal{S} e_i, \mathcal{S} e_i \rangle$$
$$= \|\mathcal{S}\|^2_{\mathrm{HS}} \leq M, \tag{61}$$

where $\langle f, g \rangle = \int f(\mathbf{x}) g(\mathbf{x}) \mathrm{d}\mathbf{x}$. □

A.5 Koopman Operators of Deterministic Systems

For the completeness of paper, we prove here the following proposition by contradiction: *The Koopman operator $\mathcal{K}_\tau$ of the deterministic system $\mathbf{x}_{t+\tau} = F(\mathbf{x}_t)$ defined by*

$$\mathcal{K}_\tau g(\mathbf{x}) = g(F(\mathbf{x})) \tag{62}$$

is not a compact operator from $\mathcal{L}^2_{\rho_1}$ to $\mathcal{L}^2_{\rho_0}$ if $\mathcal{L}^2_{\rho_1}$ is infinite-dimensional.

Assume that $\mathcal{K}_\tau$ is compact. Then, the SVD (50) of $\mathcal{K}_\tau$ exists with $\sigma_i \to 0$ as $i \to \infty$, and there is j so that $0 \leq \sigma_j < 1$. This implies $\langle \mathcal{K}_\tau \psi_j, \mathcal{K}_\tau \psi_j \rangle_{\rho_0} = \sigma_j^2 < 1$. However, according to the definition of the Koopman operator, $\langle \mathcal{K}_\tau \psi_j, \mathcal{K}_\tau \psi_j \rangle_{\rho_0} = \langle \psi_j, \psi_j \rangle_{\rho_1} = 1$, which leads to a contradiction. We can conclude that $\mathcal{K}_\tau$ is not compact and hence not Hilbert–Schmidt.

B Markov Propagators

The Markov propagator $\mathcal{P}_\tau$ is defined by

$$
\begin{aligned}
p_{t+\tau}(\mathbf{x}) &= \mathcal{P}_\tau p_t(\mathbf{x}) \\
&\triangleq \int p_\tau(\mathbf{y}, \mathbf{x})\, p_t(\mathbf{y})\, d\mathbf{y},
\end{aligned}
\tag{63}
$$

with $p_t(\mathbf{x}) = \mathbb{P}(\mathbf{x}_t = \mathbf{x})$ being the probability density of $\mathbf{x}_t$. According to the SVD of the Koopman operator given in (50), we have

$$
p_\tau(\mathbf{x}, \mathbf{y}) = \mathcal{K}_\tau \delta_{\mathbf{y}}(\mathbf{x}) = \sum_{i=1}^{\infty} \sigma_i \psi_i(\mathbf{x}) \phi_i(\mathbf{y}) \rho_1(\mathbf{y}).
\tag{64}
$$

Then

$$
\begin{aligned}
\mathcal{P}_\tau p_t(\mathbf{x}) &= \int p_\tau(\mathbf{y}, \mathbf{x})\, p_t(\mathbf{y})\, d\mathbf{y} \\
&= \sum_{i=1}^{\infty} \sigma_i \langle p_t, \rho_0 \psi_i \rangle_{\rho_0^{-1}} \rho_1(\mathbf{x}) \phi_i(\mathbf{x}).
\end{aligned}
\tag{65}
$$

Where the following normalizations were used:

$$
\langle \rho_0 \psi_i, \rho_0 \psi_j \rangle_{\rho_0^{-1}} = \langle \psi_i, \psi_j \rangle_{\rho_0} = 1_{i=j}
\tag{66}
$$

$$
\langle \rho_1 \phi_i, \rho_1 \phi_j \rangle_{\rho_1^{-1}} = \langle \phi_i, \phi_j \rangle_{\rho_1} = 1_{i=j},
\tag{67}
$$

The SVD of $\mathcal{P}_\tau$ can be written as

$$
\mathcal{P}_\tau p_t = \sum_{i=1}^{\infty} \sigma_i \langle p_t, \rho_0 \psi_i \rangle_{\rho_0^{-1}} \rho_1 \phi_i.
\tag{68}
$$

C Proof of the Variational Principle

Notice that $\mathbf{f}$ and $\mathbf{g}$ can be expressed as

$$
\mathbf{f} = \mathbf{D}_0^\top \boldsymbol{\psi}, \quad \mathbf{g} = \mathbf{D}_1^\top \boldsymbol{\phi}
\tag{69}
$$

where $\boldsymbol{\psi} = (\psi_1, \psi_2, \ldots)^\top$, $\boldsymbol{\phi} = (\phi_1, \phi_2, \ldots)^\top$ and $\mathbf{D}_0, \mathbf{D}_1 \in \mathbb{R}^{\infty \times k}$.

Since

$$
\left\langle \mathbf{f}, \mathbf{f}^\top \right\rangle_{\rho_0} = \mathbf{D}_0^\top \mathbf{D}_0
\tag{70}
$$

$$
\left\langle \mathbf{g}, \mathbf{g}^\top \right\rangle_{\rho_1} = \mathbf{D}_1^\top \mathbf{D}_1
\tag{71}
$$

and

$$
\begin{aligned}
\left\langle \mathbf{f}, \mathcal{K}_\tau \mathbf{g}^\top \right\rangle_{\rho_0} &= \mathbf{D}_0^\top \left\langle \boldsymbol{\psi}, \mathcal{K}_\tau \boldsymbol{\phi}^\top \right\rangle_{\rho_0} \mathbf{D}_1 \\
&= \mathbf{D}_0^\top \left\langle \boldsymbol{\psi}, \boldsymbol{\psi}^\top \right\rangle_{\rho_0} \boldsymbol{\Sigma} \mathbf{D}_1 \\
&= \mathbf{D}_0^\top \boldsymbol{\Sigma} \mathbf{D}_1 ,
\end{aligned}
$$

the optimization problem can be equivalently written as

$$
\max_{\mathbf{D}_0^\top \mathbf{D}_0 = \mathbf{I}, \mathbf{D}_1^\top \mathbf{D}_1 = \mathbf{I}} \sum_{i=1}^k \left(\sigma_i \mathbf{d}_{0,i}^\top \mathbf{d}_{1,i} \right)^r , \tag{72}
$$

where $\boldsymbol{\Sigma} = \mathrm{diag}(\sigma_1, \sigma_2, \ldots)$. According to the Cauchy–Schwarz inequality and the conclusion in Section I.3.C of Marshall et al. (1979), we have

$$
\sum_{i=1}^k \left| \sigma_i \mathbf{d}_{0,i}^\top \mathbf{d}_{1,i} \right| \le \sum_{i=1}^k \sigma_i \tag{73}
$$

and

$$
\sum_{i=1}^k \left(\sigma_i \mathbf{d}_{0,i}^\top \mathbf{d}_{1,i} \right)^r \le \sum_{i=1}^k \left| \sigma_i \mathbf{d}_{0,i}^\top \mathbf{d}_{1,i} \right|^r \le \sum_{i=1}^k \sigma_i^r \tag{74}
$$

under the constraint $\mathbf{D}_0^\top \mathbf{D}_0 = \mathbf{I}, \mathbf{D}_1^\top \mathbf{D}_1 = \mathbf{I}$. The variational principle can then be proven by considering

$$
\sum_{i=1}^k \left(\sigma_i \mathbf{d}_{0,i}^\top \mathbf{d}_{1,i} \right)^r = \sum_{i=1}^k \sigma_i^r \tag{75}
$$

when the first k rows of $\mathbf{D}_0$ and $\mathbf{D}_1$ are identity matrix.

D Variational Principle of Reversible Markov Processes

The variational principle of reversible Markov processes can be summarized as follows: If the Markov process $\{\mathbf{x}_t\}$ is time-reversible with respect to stationary distribution μ and all eigenvalues of $\mathcal{K}_\tau$ is nonnegative, then

$$
\begin{aligned}
\sum_{i=1}^k \lambda_i^r = \max \sum_{i=1}^k &\langle f_i, \mathcal{K}_\tau f_i \rangle_\mu^r \\
s.t. &\langle f_i, f_j \rangle_\mu = 1_{i=j}
\end{aligned} \tag{76}
$$

for $r \ge 1$ and the maximal value is achieved with $f_i = \psi_i$, where ψ_i denotes the eigenfunction with the ith largest eigenvalue λ_i. The proof is trivial by using variational

principle of general Markov processes and considering that the eigendecomposition of $\mathcal{K}_\tau$ is equivalent to its SVD if $\{\mathbf{x}_t\}$ is time-reversible and $\rho_0 = \rho_1 = \mu$.

E Analysis of Estimation Algorithms

E.1 Correctness of Feature TCCA

We show in this appendix that the feature TCCA algorithm described in Sect. 3.1 solves the optimization problem (19).

Let $\mathbf{U}' = \mathbf{C}_{00}^{\frac{1}{2}}\mathbf{U} = (\mathbf{u}'_1, \ldots, \mathbf{u}'_k)$ and $\mathbf{V}' = \mathbf{C}_{00}^{\frac{1}{2}}\mathbf{V} = (\mathbf{v}'_1, \ldots, \mathbf{v}'_k)$, (19) can be equivalently expressed as

$$\max_{\mathbf{U}',\mathbf{V}'} \sum_{i=1}^{k} \left(\mathbf{u}_i'^{\top} \mathbf{C}_{00}^{-\frac{1}{2}} \mathbf{C}_{01} \mathbf{C}_{11}^{-\frac{1}{2}} \mathbf{v}_i' \right)^r$$
$$\text{s.t.} \mathbf{U}'^{\top}\mathbf{U}' = \mathbf{I}$$
$$\mathbf{V}'^{\top}\mathbf{V}' = \mathbf{I}. \tag{77}$$

According to the Cauchy–Schwarz inequality and the conclusion in Section I.3.C of Marshall et al. (1979), we have

$$\sum_{i=1}^{k} \left(\mathbf{u}_i'^{\top} \mathbf{C}_{00}^{-\frac{1}{2}} \mathbf{C}_{01} \mathbf{C}_{11}^{-\frac{1}{2}} \mathbf{v}_i' \right)^r \leq \sum_{i=1}^{k} \left| \mathbf{u}_i'^{\top} \mathbf{C}_{00}^{-\frac{1}{2}} \mathbf{C}_{01} \mathbf{C}_{11}^{-\frac{1}{2}} \mathbf{v}_i' \right|^r$$
$$\leq \sum_{i=1}^{k} s_i^r \tag{78}$$

under the constraints $\mathbf{U}'^{\top}\mathbf{U}' = \mathbf{I}$, $\mathbf{V}'^{\top}\mathbf{V}' = \mathbf{I}$, where s_i is the ith largest singular value of $\mathbf{C}_{00}^{-\frac{1}{2}} \mathbf{C}_{01} \mathbf{C}_{11}^{-\frac{1}{2}}$. Considering the equalities hold in the above when $\mathbf{U}'$, $\mathbf{V}'$ are the first k left and right singular vectors of $\mathbf{C}_{00}^{-\frac{1}{2}} \mathbf{C}_{01} \mathbf{C}_{11}^{-\frac{1}{2}}$, we get

$$\max_{\mathbf{U},\mathbf{V}} \mathcal{R}_r(\mathbf{U}, \mathbf{V}) = \sum_{i=1}^{k} s_i^r$$
$$\text{s.t.} \mathbf{U}^{\top}\mathbf{C}_{00}\mathbf{U} = \mathbf{I}$$
$$\mathbf{V}^{\top}\mathbf{C}_{11}\mathbf{V} = \mathbf{I}, \tag{79}$$

and the correctness of the feature TCCA can then be proved.

Furthermore, if $k = \min\{\dim(\chi_0), \dim(\chi_1)\}$, we can get

$$\max_{\mathbf{U},\mathbf{V}} \mathcal{R}_r(\mathbf{U}, \mathbf{V}) = \left\| \mathbf{C}_{00}^{-\frac{1}{2}} \mathbf{C}_{01} \mathbf{C}_{11}^{-\frac{1}{2}} \right\|_r^r \tag{80}$$

under the orthonormality constraints.

E.2 Feature TCCA of Projected Koopman Operators

Define projection operators

$$
\begin{aligned}
\mathcal{Q}_{\chi_0} f &\triangleq \underset{f' \in \mathrm{span}\{\chi_{0,1}, \chi_{0,2}, \dots\}}{\arg\min} \left\langle f' - f, f' - f \right\rangle_{\rho_0} \\
&= \left\langle f, \chi_0^\top \right\rangle_{\rho_0} \mathbf{C}_{00}^{-1} \chi_0,
\end{aligned}
\tag{81}
$$

$$
\begin{aligned}
\mathcal{Q}_{\chi_1} g &\triangleq \underset{g' \in \mathrm{span}\{\chi_{1,1}, \chi_{1,2}, \dots\}}{\arg\min} \left\langle g' - g, g' - g \right\rangle_{\rho_1} \\
&= \left\langle g, \chi_1^\top \right\rangle_{\rho_1} \mathbf{C}_{11}^{-1} \chi_1,
\end{aligned}
\tag{82}
$$

and let $\mathcal{K}_\tau^{\mathrm{proj}} = \mathcal{Q}_{\chi_0} \mathcal{K}_\tau \mathcal{Q}_{\chi_1}$ be the projection of the Koopman operator $\mathcal{K}_\tau$ onto the subspaces of χ_0, χ_1. Then for any $f = \mathbf{u}^\top \chi_0 \in \mathrm{span}\{\chi_{0,1}, \chi_{0,2}, \dots\}$ and $g = \mathbf{v}^\top \chi_1 \in \mathrm{span}\{\chi_{1,1}, \chi_{1,2}, \dots\}$,

$$
\begin{aligned}
\left\langle f, \mathcal{K}_\tau^{\mathrm{proj}} g \right\rangle_{\rho_0} &= \left\langle g, \chi_1^\top \right\rangle_{\rho_1} \mathbf{C}_{11}^{-1} \mathbf{C}_{01}^\top \mathbf{C}_{00}^{-1} \left\langle \chi_0, f \right\rangle_{\rho_0} \\
&= \mathbf{u}^\top \mathbf{C}_{01} \mathbf{v} \\
&= \langle f, \mathcal{K}_\tau g \rangle_{\rho_0},
\end{aligned}
\tag{83}
$$

which implies that Eq. (19) can also be interpreted as the variational problem for the feature TCCA of $\mathcal{K}_\tau^{\mathrm{proj}}$.

By ignoring the statistical noise, we can conclude from Theorem 2 that the $\{(s_i, f_i, g_i)\}$ provided by the feature TCCA are exactly the singular components of $\mathcal{K}_\tau^{\mathrm{proj}}$, and the optimality of the estimation result is therefore invariant for any choice of $r \geq 1$. In addition, the sum over the r'th power of all singular values of $\mathcal{K}_\tau^{\mathrm{proj}}$ is

$$
\sum_i s_i^r = \left\| \mathbf{C}_{00}^{-\frac{1}{2}} \mathbf{C}_{01} \mathbf{C}_{11}^{-\frac{1}{2}} \right\|_r^r.
\tag{84}
$$

E.3 An Example of Nonlinear TCCA

Consider a stochastic system

$$
x_{t+1} = \frac{1}{2} x_t + u_t,
\tag{85}
$$

where u_t is Gaussian white noise with mean zero and variance 1. By setting

$$
\rho_0(x) = \rho_1(x) = \mathcal{N}\left(x \mid 0, \frac{4}{3} \right)
\tag{86}
$$

to be the stationary distribution and basis functions

$$\boldsymbol{\chi}_0(x) = \boldsymbol{\chi}_1(x) = \left(1, \exp(-wx^2) - \sqrt{\frac{3}{8w+3}}, x \exp(-(1 - w^{\frac{1}{10}})x^2) \right)^\top \tag{87}$$

with parameter $w \in [0.01, 1]$, we can obtain

$$\mathbf{C}_{00} = \mathbf{C}_{11} = \mathrm{diag}\left(1, \left(\frac{16}{3}w + 1\right)^{-\frac{1}{2}} - \frac{3}{8w+3}, 4\sqrt{3}\left(-16w^{\frac{1}{10}} + 19\right)^{-\frac{3}{2}} \right),$$

$$\mathbf{C}_{01} = \mathrm{diag}\left(1, \left(\frac{16}{3}w^2 + \frac{16}{3}w + 1\right)^{-\frac{1}{2}} - \frac{3}{8w+3}, \right.$$
$$\left. 2\sqrt{3}\left(16(1 - w^{\frac{1}{10}})^2 - 16w^{\frac{1}{10}} + 19\right)^{-\frac{3}{2}} \right) \tag{88}$$

The maximal VAMP-r score for a given w can then be analytically computed by

$$\mathcal{R}_r(w) = \mathrm{tr}\left[\left(\mathbf{C}_{00}(w)^{-1} \mathbf{C}_{01}(w) \right)^r \right] \tag{89}$$

according to (24). We evaluate $\mathcal{R}_r(w)$ at 9901 equally spaced points of w in the interval [0.01, 1] for $r = 1, 2$, and the maximal values of $\mathcal{R}_1, \mathcal{R}_2$ are achieved at $w = 0.3157$ and $w = 0.7069$ respectively.

F Implementation of Estimation Algorithms

F.1 Decorrelation of Basis Functions

For convenience of notation, here we define

$$\mathbf{X} = \left(\boldsymbol{\chi}_0(\mathbf{x}_1), \ldots, \boldsymbol{\chi}_0(\mathbf{x}_{T-\tau}) \right)^\top \tag{90}$$
$$\mathbf{Y} = \left(\boldsymbol{\chi}_1(\mathbf{x}_{1+\tau}), \ldots, \boldsymbol{\chi}_0(\mathbf{x}_T) \right)^\top. \tag{91}$$

In this paper, we utilize principal component analysis (PCA) to explicitly reduce correlations between basis functions as follows: First, we compute the empirical means of basis functions and the covariance matrices of mean-centered basis functions:

$$\boldsymbol{\pi}_0 = \frac{1}{T-\tau}\mathbf{X}^\top \mathbf{1} \tag{92}$$

$$\boldsymbol{\pi}_1 = \frac{1}{T-\tau}\mathbf{Y}^\top \mathbf{1} \tag{93}$$

$$\mathrm{COV}_0 = \frac{1}{T-\tau}\mathbf{X}^\top \mathbf{X} - \boldsymbol{\pi}_0 \boldsymbol{\pi}_0^\top \tag{94}$$

$$\mathrm{COV}_1 = \frac{1}{T-\tau}\mathbf{Y}^\top \mathbf{Y} - \boldsymbol{\pi}_1 \boldsymbol{\pi}_1^\top. \tag{95}$$

Next, perform the truncated eigen decomposition of the covariance matrices as

$$\text{COV}_0 \approx \mathbf{Q}_{0,d}^\top \mathbf{S}_{0,d} \mathbf{Q}_{0,d} \tag{96}$$

$$\text{COV}_1 \approx \mathbf{Q}_{1,d}^\top \mathbf{S}_{1,d} \mathbf{Q}_{1,d}, \tag{97}$$

where the diagonal of matrices $\mathbf{S}_{0,d}, \mathbf{S}_{1,d}$ contain all positive eigenvalues that are larger than ϵ_0 and absolute values of all negative eigenvalues ($\epsilon_0 = 10^{-10}$ in our applications). Last, the new basis functions are given by

$$\boldsymbol{\chi}_0^{\text{new}} = \begin{bmatrix} \mathbf{Q}_{0,d}^\top \mathbf{S}_{0,d}^{\frac{1}{2}} \left(\boldsymbol{\chi}_0 - \boldsymbol{\pi}_0 \right) \\ \mathbb{1} \end{bmatrix}, \quad \boldsymbol{\chi}_1^{\text{new}} = \begin{bmatrix} \mathbf{Q}_{1,d}^\top \mathbf{S}_{1,d}^{\frac{1}{2}} \left(\boldsymbol{\chi}_1 - \boldsymbol{\pi}_1 \right) \\ \mathbb{1} \end{bmatrix} \tag{98}$$

We denote the transformation (98) by

$$\boldsymbol{\chi}_0^{\text{new}}, \boldsymbol{\chi}_1^{\text{new}} = \text{DC} \left[\boldsymbol{\chi}_0, \boldsymbol{\chi}_1 | \boldsymbol{\pi}_0, \boldsymbol{\pi}_1, \text{COV}_0, \text{COV}_1 \right] \tag{99}$$

Then the feature TCCA algorithm with decorrelation of basis functions can be summarized as:

1. Compute $\boldsymbol{\pi}_0, \boldsymbol{\pi}_1$ and $\text{COV}_0, \text{COV}_1$ by (92–95).
2. Let $\boldsymbol{\chi}_0, \boldsymbol{\chi}_1 := \text{DC} \left[\boldsymbol{\chi}_0, \boldsymbol{\chi}_1 | \boldsymbol{\pi}_0, \boldsymbol{\pi}_1, \text{COV}_0, \text{COV}_1 \right]$, and recalculate $\mathbf{X}$ and $\mathbf{Y}$ according to the new basis functions.
3. Compute covariance matrices $\mathbf{C}_{00}, \mathbf{C}_{01}, \mathbf{C}_{11}$ by

$$\mathbf{C}_{00} = \frac{1}{T - \tau} \mathbf{X}^\top \mathbf{X}$$

$$\mathbf{C}_{01} = \frac{1}{T - \tau} \mathbf{X}^\top \mathbf{Y}$$

$$\mathbf{C}_{11} = \frac{1}{T - \tau} \mathbf{Y}^\top \mathbf{Y}$$

4. Perform the truncated SVD $\mathbf{C}_{00}^{-\frac{1}{2}} \mathbf{C}_{01} \mathbf{C}_{11}^{-\frac{1}{2}} = \mathbf{U}_k' \hat{\boldsymbol{\Sigma}}_k \mathbf{V}_k'^\top$.
5. Output estimated singular components $\hat{\boldsymbol{\Sigma}}_k = \text{diag}(\hat{\sigma}_1, \ldots, \hat{\sigma}_k)$, $\mathbf{U}_k^\top \boldsymbol{\chi}_0 = (\hat{\psi}_1, \ldots, \hat{\psi}_k)^\top$ and $\mathbf{V}_k^\top \boldsymbol{\chi}_1 = (\hat{\phi}_1, \ldots, \hat{\phi}_k)^\top$ with $\mathbf{U}_k = \mathbf{C}_{00}^{-\frac{1}{2}} \mathbf{U}_k'$ and $\mathbf{V}_k = \mathbf{C}_{11}^{-\frac{1}{2}} \mathbf{V}_k'$.

Notice that the estimated $\mathbf{C}_{00}, \mathbf{C}_{01}$ and $\mathbf{C}_{11}$ in the above algorithm satisfy

$$\begin{bmatrix} \mathbf{C}_{00} & \mathbf{C}_{01} \\ \mathbf{C}_{01}^\top & \mathbf{C}_{11} \end{bmatrix} = \frac{1}{T - \tau} \begin{bmatrix} \mathbf{X}^\top \mathbf{X} & \mathbf{X}^\top \mathbf{Y} \\ \mathbf{Y}^\top \mathbf{X} & \mathbf{Y}^\top \mathbf{Y} \end{bmatrix}$$

$$= \frac{1}{T - \tau} (\mathbf{X}, \mathbf{Y})^\top (\mathbf{X}, \mathbf{Y})$$

$$\succeq 0 \tag{100}$$

where $\mathbf{C} \succeq 0$ means $\mathbf{C}$ is a positive semi-definite matrix. According to the Schur complement lemma, we have

$$\mathbf{C}_{01}\mathbf{C}_{11}^{-1}\mathbf{C}_{01}^{\top} \preceq \mathbf{C}_{00}$$

$$\Rightarrow \left(\mathbf{C}_{00}^{-\frac{1}{2}}\mathbf{C}_{01}\mathbf{C}_{11}^{-\frac{1}{2}}\right)\left(\mathbf{C}_{00}^{-\frac{1}{2}}\mathbf{C}_{01}\mathbf{C}_{11}^{-\frac{1}{2}}\right)^{\top} \preceq \mathbf{I} \tag{101}$$

where $\mathbf{I}$ denotes an identity matrix of appropriate size. So the estimated $\sigma_1 \leq 1$.

Furthermore, since $\mathbf{v}_0^{\top}\boldsymbol{\chi}_0 = \mathbf{v}_1^{\top}\boldsymbol{\chi}_1 = \mathbb{1}$ for $\mathbf{v}_0 = (0, \ldots, 0, 1)^{\top}$ and $\mathbf{v}_1 = (0, \ldots, 0, 1)^{\top}$,

$$\left(\mathbf{C}_{00}^{-\frac{1}{2}}\mathbf{C}_{01}\mathbf{C}_{11}^{-\frac{1}{2}}\right)\left(\mathbf{C}_{00}^{-\frac{1}{2}}\mathbf{C}_{01}\mathbf{C}_{11}^{-\frac{1}{2}}\right)^{\top}\mathbf{C}_{00}^{\frac{1}{2}}\mathbf{v}_0 = \mathbf{C}_{00}^{\frac{1}{2}}\left(\mathbf{X}^{\top}\mathbf{X}\right)^{-1}\mathbf{X}^{\top}\mathbf{Y}\left(\mathbf{Y}^{\top}\mathbf{Y}\right)^{-1}\mathbf{Y}^{\top}\mathbf{X}\mathbf{v}_0$$

$$= \mathbf{C}_{00}^{\frac{1}{2}}\mathbf{X}^{+}\mathbf{Y}\mathbf{Y}^{+}\mathbf{1}$$

$$= \mathbf{C}_{00}^{\frac{1}{2}}\mathbf{v}_0 \tag{102}$$

which implies that 1 is the largest singular value of $\mathbf{C}_{00}^{-\frac{1}{2}}\mathbf{C}_{01}\mathbf{C}_{11}^{-\frac{1}{2}}$.

F.2 Parameter Optimization in Nonlinear TCCA

The optimization problem

$$\max_{\mathbf{w}} \mathcal{R}_r(\mathbf{w}) = \left\|\mathbf{C}_{00}\left(\mathbf{w}\right)^{-\frac{1}{2}}\mathbf{C}_{01}\left(\mathbf{w}\right)\mathbf{C}_{11}\left(\mathbf{w}\right)^{-\frac{1}{2}}\right\|_r^r \tag{103}$$

can be solved by direct search as in our examples (see Appendix K.1). But for a high-dimensional parameter vector $\mathbf{w}$, it is more efficient to perform the optimization by the gradient descent method in the form of

$$\mathbf{w} \leftarrow \mathbf{w} + \eta\frac{\partial\mathcal{R}_r(\mathbf{w})}{\partial\mathbf{w}}, \tag{104}$$

where η is the step size. When $r = 2$, the gradient of $\mathcal{R}_r$ with respect to an element w_i in $\mathbf{w}$ can be written as

$$\frac{\partial\mathcal{R}_r}{\partial w_i} = \frac{2}{T-\tau}\mathrm{tr}\left[\mathbf{C}_{00}^{-1}\mathbf{C}_{01}\mathbf{C}_{11}^{-1}\left(\mathbf{Y} - \mathbf{C}_{01}^{\top}\mathbf{C}_{00}^{-1}\mathbf{X}\right)\left(\frac{\partial\mathbf{X}}{\partial w_i}\right)^{\top}\right]$$

$$+ \frac{2}{T-\tau}\mathrm{tr}\left[\mathbf{C}_{11}^{-1}\mathbf{C}_{01}^{\top}\mathbf{C}_{00}^{-1}\left(\mathbf{X} - \mathbf{C}_{01}\mathbf{C}_{11}^{-1}\mathbf{Y}\right)\left(\frac{\partial\mathbf{Y}}{\partial w_i}\right)^{\top}\right], \tag{105}$$

where $\mathbf{X}$, $\mathbf{Y}$ have the same definitions as in Appendix F.1. If the data size is too large, we can approximate the gradient based on a random subset of data in each iteration, and update $\mathbf{w}$ in a stochastic gradient descent manner (Andrew et al. 2013; Mardt et al. 2018).

Like feature TCCA, the nonlinear TCCA also suffers from the numerical singularity when $\mathbf{C}_{00}$ or $\mathbf{C}_{11}$ is not full rank. This problem can be addressed by the decorrelation

of basis functions when performing direct search. For the gradient descent method (or stochastic gradient descent method), we can replace the objective function $\mathcal{R}_r(\mathbf{w})$ by a regularized one

$$\mathcal{R}_r(\mathbf{w};\epsilon) = \left\| (\mathbf{C}_{00}(\mathbf{w}) + \epsilon\mathbf{I})^{-\frac{1}{2}} \mathbf{C}_{01}(\mathbf{w}) (\mathbf{C}_{11}(\mathbf{w}) + \epsilon\mathbf{I})^{-\frac{1}{2}} \right\|_r^r, \qquad (106)$$

where $\epsilon > 0$ is a hyper-parameter and can be selected by the cross-validation.

G Relationship Between VAMP and EDMD

The proof of (21) is trivial. Here, we only show that the eigenvalue problem of $\hat{\mathcal{K}}_\tau$ given by the feature TCCA is equivalent to that of matrix $\mathbf{K}_\chi$ as

$$\hat{\mathcal{K}}_\tau g = \lambda g \iff \mathbf{K}_\chi \mathbf{b} = \lambda \mathbf{b} \text{ with } g = \mathbf{b}^\top \chi \qquad (107)$$

under the assumption that $\chi_0 = \chi_1 = \chi$ and $\mathbf{C}_{00}$ is invertible, which is consistent with the spectral approximation theory in EDMD. First, if g and λ satisfy $\mathcal{K}_\tau g = \lambda g$, there must exist vector $\mathbf{b}$ so that $g = \mathbf{b}^\top \chi$. Then

$$\hat{\mathcal{K}}_\tau g = \lambda g$$
$$\Rightarrow \mathbf{b}^\top \mathbf{K}_\chi^\top \chi = \lambda \mathbf{b}^\top \chi$$
$$\Rightarrow \mathbf{b}^\top \mathbf{K}_\chi^\top \mathbf{C}_{00} = \lambda \mathbf{b}^\top \mathbf{C}_{00}$$
$$\Rightarrow \mathbf{K}_\chi \mathbf{b} = \lambda \mathbf{b}. \qquad (108)$$

Second, if $\mathbf{K}_\chi \mathbf{b} = \lambda \mathbf{b}$,

$$\hat{\mathcal{K}}_\tau \mathbf{b}^\top \chi = \mathbf{b}^\top \mathbf{K}_\chi^\top \chi$$
$$= \lambda \mathbf{b}^\top \chi. \qquad (109)$$

H Analysis of the VAMP-E Score

H.1 Proof of (28)

Here we define

$$\mathbf{C}_{ff} = \left\langle \mathbf{f}, \mathbf{f}^\top \right\rangle_{\rho_0} = \mathbf{U}^\top \mathbf{C}_{00} \mathbf{U}, \qquad (110)$$

$$\mathbf{C}_{gg} = \left\langle \mathbf{g}, \mathbf{g}^\top \right\rangle_{\rho_1} = \mathbf{V}^\top \mathbf{C}_{11} \mathbf{V}, \qquad (111)$$

$$\mathbf{C}_{fg} = \left\langle \mathbf{f}, \mathcal{K}_\tau \mathbf{g}^\top \right\rangle_{\rho_1} = \mathbf{U}^\top \mathbf{C}_{01} \mathbf{V}. \qquad (112)$$

Considering $\{\phi_i\}$ is an orthonormal basis of $\mathcal{L}_{\rho_1}^2$, we have

$$\left\|\hat{\mathcal{K}}_\tau\right\|_{\mathrm{HS}}^2 = \sum_j \left\langle \hat{\mathcal{K}}_\tau \phi_j, \hat{\mathcal{K}}_\tau \phi_j \right\rangle_{\rho_0}$$

$$= \sum_j \left\langle \left\langle \phi_j, \mathbf{g}^\top \right\rangle_{\rho_1} \mathbf{Kf}, \mathbf{f}^\top \mathbf{K} \left\langle \mathbf{g}, \phi_j \right\rangle_{\rho_1} \right\rangle_{\rho_0}$$

$$= \sum_j \left\langle \phi_j, \mathbf{g}^\top \right\rangle_{\rho_1} \mathbf{K} \left\langle \mathbf{f}, \mathbf{f}^\top \right\rangle_{\rho_0} \mathbf{K} \left\langle \mathbf{g}, \phi_j \right\rangle_{\rho_1}$$

$$= \mathrm{tr} \left[\mathbf{K} \left\langle \mathbf{f}, \mathbf{f}^\top \right\rangle_{\rho_0} \mathbf{K} \sum_j \left\langle \mathbf{g}, \phi_j \right\rangle_{\rho_1} \left\langle \phi_j, \mathbf{g}^\top \right\rangle_{\rho_1} \right]$$

$$= \mathrm{tr} \left[\mathbf{K} \left\langle \mathbf{f}, \mathbf{f}^\top \right\rangle_{\rho_0} \mathbf{K} \left\langle \sum_j \left\langle \mathbf{g}, \phi_j \right\rangle_{\rho_1} \phi_j, \mathbf{g}^\top \right\rangle_{\rho_1} \right]$$

$$= \mathrm{tr} \left[\mathbf{K} \left\langle \mathbf{f}, \mathbf{f}^\top \right\rangle_{\rho_0} \mathbf{K} \left\langle \mathbf{g}, \mathbf{g}^\top \right\rangle_{\rho_1} \right]$$

$$= \mathrm{tr} \left[\mathbf{K} \mathbf{C}_{ff} \mathbf{K} \mathbf{C}_{gg} \right] \tag{113}$$

and

$$\left\langle \hat{\mathcal{K}}_\tau, \mathcal{K}_\tau \right\rangle_{\mathrm{HS}} = \sum_j \left\langle \hat{\mathcal{K}}_\tau \phi_j, \mathcal{K}_\tau \phi_j \right\rangle_{\rho_0}$$

$$= \sum_j \left\langle \left\langle \phi_j, \mathbf{g}^\top \right\rangle_{\rho_1} \mathbf{Sf}, \sigma_j \psi_j \right\rangle_{\rho_0}$$

$$= \sum_j \sigma_j \left\langle \phi_j, \mathbf{g}^\top \right\rangle_{\rho_1} \mathbf{S} \left\langle \mathbf{f}, \psi_j \right\rangle_{\rho_0}$$

$$= \mathrm{tr} \left[\mathbf{K} \sum_j \sigma_j \left\langle \mathbf{f}, \psi_j \right\rangle_{\rho_0} \left\langle \phi_j, \mathbf{g}^\top \right\rangle_{\rho_1} \right]$$

$$= \mathrm{tr} \left[\mathbf{K} \left\langle \mathbf{f}, \sum_j \sigma_j \psi_j \left\langle \phi_j, \mathbf{g}^\top \right\rangle_{\rho_1} \right\rangle_{\rho_0} \right]$$

$$= \mathrm{tr} \left[\mathbf{K} \left\langle \mathbf{f}, \mathcal{K}_\tau \mathbf{g}^\top \right\rangle_{\rho_0} \right]$$

$$= \mathrm{tr} \left[\mathbf{K} \mathbf{C}_{fg} \right], \tag{114}$$

where $\langle \cdot, \cdot \rangle_{\mathrm{HS}}$ denotes the Hilbert–Schmidt inner product of operators. Then, according to the definition of Hilbert–Schmidt norm,

$$\left\| \hat{\mathcal{K}}_\tau - \mathcal{K}_\tau \right\|_{\mathrm{HS}}^2 = \left\| \hat{\mathcal{K}}_\tau \right\|_{\mathrm{HS}}^2 - 2 \sum_j \left\langle \hat{\mathcal{K}}_\tau, \mathcal{K}_\tau \right\rangle_{\mathrm{HS}} + \left\| \mathcal{K}_\tau \right\|_{\mathrm{HS}}^2$$

$$= \mathrm{tr} \left[\mathbf{K} \mathbf{C}_{ff} \mathbf{K} \mathbf{C}_{gg} - 2 \mathbf{K} \mathbf{C}_{fg} \right] + \left\| \mathcal{K}_\tau \right\|_{\mathrm{HS}}^2 \tag{115}$$

H.2 Relationship Between VAMP-2 and VAMP-E

We first show that the feature TCCA algorithm maximizes VAMP-E. Notice that

$$
\begin{aligned}
\mathcal{R}_E(\mathbf{K}, \mathbf{U}, \mathbf{V}) &= \mathrm{tr}\left[2\left(\mathbf{C}_{00}^{\frac{1}{2}} \mathbf{U}\mathbf{K}\mathbf{V}^\top \mathbf{C}_{11}^{\frac{1}{2}} \right)^\top \left(\mathbf{C}_{00}^{-\frac{1}{2}} \mathbf{C}_{01} \mathbf{C}_{11}^{-\frac{1}{2}} \right) \right.\\
&\quad \left. - \left(\mathbf{C}_{00}^{\frac{1}{2}} \mathbf{U}\mathbf{K}\mathbf{V}^\top \mathbf{C}_{11}^{\frac{1}{2}} \right)^\top \left(\mathbf{C}_{00}^{\frac{1}{2}} \mathbf{U}\mathbf{K}\mathbf{V}^\top \mathbf{C}_{11}^{\frac{1}{2}} \right) \right]\\
&= -\left\| \mathbf{C}_{00}^{\frac{1}{2}} \mathbf{U}\mathbf{K}\mathbf{V}^\top \mathbf{C}_{11}^{\frac{1}{2}} - \mathbf{C}_{00}^{-\frac{1}{2}} \mathbf{C}_{01} \mathbf{C}_{11}^{-\frac{1}{2}} \right\|_F^2 + \left\| \mathbf{C}_{00}^{-\frac{1}{2}} \mathbf{C}_{01} \mathbf{C}_{11}^{-\frac{1}{2}} \right\|_F^2\\
&= -\left\| \mathbf{U}'\mathbf{K}\mathbf{V}'^\top - \mathbf{C}_{00}^{-\frac{1}{2}} \mathbf{C}_{01} \mathbf{C}_{11}^{-\frac{1}{2}} \right\|_F^2 + \left\| \mathbf{C}_{00}^{-\frac{1}{2}} \mathbf{C}_{01} \mathbf{C}_{11}^{-\frac{1}{2}} \right\|_F^2, \qquad (116)
\end{aligned}
$$

where $\|\cdot\|_F$ denotes the Frobenius norm and $\mathbf{U}' = \mathbf{C}_{00}^{\frac{1}{2}} \mathbf{U}$, $\mathbf{V}' = \mathbf{C}_{11}^{\frac{1}{2}} \mathbf{V}$. It can be seen that the feature TCCA algorithm maximizes the first term on the right-hand side of (116) and therefore maximizes VAMP-E.

For the optimal model generated by the nonlinear TCCA, the first term on the right-hand side of (116) is equal to zero, and the second term is maximized as a function of $\mathbf{w}$. Thus, the nonlinear TCCA also maximizes VAMP-E.

In addition, for $\mathbf{K}, \mathbf{U}, \mathbf{V}$ provided by both feature TCCA and nonlinear TCCA,

$$
\begin{aligned}
\mathcal{R}_E(\mathbf{K}, \mathbf{U}, \mathbf{V}) &= -\left\| \mathbf{U}'\mathbf{K}\mathbf{V}'^\top - \mathbf{C}_{00}^{-\frac{1}{2}} \mathbf{C}_{01} \mathbf{C}_{11}^{-\frac{1}{2}} \right\|_F^2 + \left\| \mathbf{C}_{00}^{-\frac{1}{2}} \mathbf{C}_{01} \mathbf{C}_{11}^{-\frac{1}{2}} \right\|_F^2\\
&= -\sum_{i=k+1}^{\min\{m,n\}} K_{ii}^2 + \sum_{i=1}^{\min\{m,n\}} K_{ii}^2\\
&= \sum_{i=1}^{k} K_{ii}^2\\
&= \mathcal{R}_2(\mathbf{U}, \mathbf{V}). \qquad (117)
\end{aligned}
$$

I Subspace Variational Principle

The variational principle proposed in Sect. 2.2 can be further extended to singular subspaces of the Koopman operator as follows:

$$
\sum_{i=1}^{k} \sigma_i^r \geq \mathcal{R}_r^{\mathrm{space}}[\mathbf{f}, \mathbf{g}] = \left\| \mathbf{C}_{ff}^{-\frac{1}{2}} \mathbf{C}_{fg} \mathbf{C}_{gg}^{-\frac{1}{2}} \right\|_r^r \qquad (118)
$$

for $r \geq 1$, and the equality holds if $\mathrm{span}\{\psi_1, \ldots, \psi_k\} = \mathrm{span}\{f_1, \ldots, f_k\}$ and $\mathrm{span}\{\phi_1, \ldots, \phi_k\} = \mathrm{span}\{g_1, \ldots, g_k\}$, where $\mathbf{C}_{ff} = \langle \mathbf{f}, \mathbf{f}^\top \rangle_{\rho_0}$, $\mathbf{C}_{fg} = \langle \mathbf{f}, \mathcal{K}_\tau \mathbf{g}^\top \rangle_{\rho_0}$

and $\mathbf{C}_{gg} = \langle \mathbf{g}, \mathbf{g}^\top \rangle_{\rho_1}$. This statement can be proven by implementing the feature TCCA algorithm with feature functions $\mathbf{f}$ and $\mathbf{g}$.

The $\mathcal{R}_r^{\text{space}}[\mathbf{f}, \mathbf{g}]$ is a relaxation of VAMP-r, which measures the consistency between the subspaces spanned by $\mathbf{f}, \mathbf{g}$ and the dominant singular spaces, and we call it the subspace VAMP-r score. $\mathcal{R}_r^{\text{space}}[\mathbf{f}, \mathbf{g}]$ is invariant with respect to the invertible linear transformations of $\mathbf{f}$ and $\mathbf{g}$, i.e., $\mathcal{R}_r^{\text{space}}[\mathbf{f}, \mathbf{g}] = \mathcal{R}_r^{\text{space}}[\mathbf{A}_f \mathbf{f}, \mathbf{A}_g \mathbf{g}]$ for any invertible matrices $\mathbf{A}_f, \mathbf{A}_g$.

In the cross-validation for feature TCCA, we can utilize $\mathcal{R}_r^{\text{space}}$ to calculate the validation score by

$$
\begin{aligned}
\text{CV}(\mathbf{K}, \mathbf{U}, \mathbf{V}|\mathcal{D}_{\text{test}}) &= \mathcal{R}_r^{\text{space}}(\mathbf{U}, \mathbf{V}|\mathcal{D}_{\text{test}}) \\
&= \mathcal{R}_r^{\text{space}}\left[\mathbf{U}^\top \boldsymbol{\chi}_0, \mathbf{V}^\top \boldsymbol{\chi}_1|\mathcal{D}_{\text{test}}\right] \\
&= \left\| \left(\mathbf{U}^\top \mathbf{C}_{00}^{\text{test}} \mathbf{U}\right)^{-\frac{1}{2}} \left(\mathbf{U}^\top \mathbf{C}_{01}^{\text{test}} \mathbf{V}\right) \left(\mathbf{V}^\top \mathbf{C}_{11}^{\text{test}} \mathbf{V}\right)^{-\frac{1}{2}} \right\|_r^r.
\end{aligned}
$$
(119)

We now analyze the difficulties of applying $\mathcal{R}_r^{\text{space}}$ to the cross-validation. First, for given basis functions $\boldsymbol{\chi}_0, \boldsymbol{\chi}_1, \mathcal{R}_r^{\text{space}}(\mathbf{U}, \mathbf{V}|\mathcal{D}_{\text{test}})$ is monotonically increasing with respect to k and

$$
\mathcal{R}_r^{\text{space}}(\mathbf{U}_k, \mathbf{V}_k|\mathcal{D}_{\text{test}}) = \left\| \left(\mathbf{C}_{00}^{\text{test}}\right)^{-\frac{1}{2}} \mathbf{C}_{01}^{\text{test}} \left(\mathbf{C}_{11}^{\text{test}}\right)^{-\frac{1}{2}} \right\|_r^r
$$
(120)

is independent of the estimated singular components if $k = \max\{\dim(\boldsymbol{\chi}_0), \dim(\boldsymbol{\chi}_1)\}$. Therefore, k is a new hyper-parameter that cannot be determined by the cross-validation. Second, for training set, $\mathbf{U}_k^\top \mathbf{C}_{00}^{\text{train}} \mathbf{U}_k = \mathbf{V}_k^\top \mathbf{C}_{11}^{\text{train}} \mathbf{V}_k = \mathbf{I}$. But for test set, $\mathbf{U}_k^\top \mathbf{C}_{00}^{\text{test}} \mathbf{U}_k$ and $\mathbf{V}_k^\top \mathbf{C}_{11}^{\text{test}} \mathbf{V}_k$ are possibly singular and the validation score cannot be reliably computed.

J Computation of $\hat{\mathcal{K}}_\tau^n$

The approximate Koopman operator in the form of (27) can also be written as

$$
\hat{\mathcal{K}}_\tau g = \langle g, \mathbf{g}^\top \rangle_{\rho_1} \mathbf{K} \mathbf{f}.
$$
(121)

Hence,

$$
\hat{\mathcal{K}}_\tau^n g = \langle g, \mathbf{g}^\top \rangle_{\rho_1} \mathbf{K} \left(\mathbf{R}^{n-1}\right)^\top \mathbf{f},
$$
(122)

and we have

$$
\langle f, \hat{\mathcal{K}}_\tau^n g \rangle_{\rho_0(n\tau)} = \langle f, \mathbf{f}^\top \rangle_{\rho_0(n\tau)} \mathbf{R}^{n-1} \mathbf{K} \langle \mathbf{g}, g \rangle_{\rho_1}
$$
(123)

and

$$\hat{p}_{n\tau}(\mathbf{x}, \mathbf{y}) = \hat{\mathcal{K}}_\tau^n \delta_{\mathbf{y}}(\mathbf{x})$$
$$= \mathbf{f}(\mathbf{x})^\top \mathbf{R}^{n-1} \mathbf{K} \mathbf{g}(\mathbf{y}) \rho_1(\mathbf{y}), \qquad (124)$$

where

$$\mathbf{R} = \mathbf{K} \left\langle \mathbf{g}, \mathbf{f}^\top \right\rangle_{\rho_1}. \qquad (125)$$

Notice that substituting $\mathbf{f} = \mathbf{U}^\top \boldsymbol{\chi}_0$, $\mathbf{g} = \mathbf{V}^\top \boldsymbol{\chi}_1$ into (123) yields (37).

K Details of Numerical Examples

K.1 One-Dimensional System

For convenience of analysis and computation, we partition the state space $[-20, 20]$ into 2000 bins $S_1, \ldots, S_{2000}$ uniformly, and discretize the one-dimensional dynamical system described in Example 1 as

$$\mathbb{P}(x_{t+1} \in S_j | x_t \in S_i) \propto \mathcal{N}\left(s_j \Big| \frac{s_i}{2} + \frac{7s_i}{1 + 0.12s_i^2} + 6\cos s_i, 10\right), \qquad (126)$$

where s_i is the center of the bin S_i, and the local distribution of x_t within any bin is always uniform distribution. All numerical computations and simulations in Examples 1, 2 and 3 are based on (126), and the initial state x_0 is distributed according to the stationary distribution $\rho_0 = \rho_1 = \mu$.

In Example 1, the stationary distribution and singular components of the Koopman operator are analytically computed by the feature TCCA with basis functions $\chi_{0,i}(x) = \chi_{1,i}(x) = 1_{x \in S_i}$ as follows:

1. Compute the transition matrix $\mathbf{P} = [P_{ij}] = [\mathbb{P}(x_{t+1} \in S_j | x_t \in S_i)]$ and the stationary vector $\boldsymbol{\pi} = [\pi_i]$ satisfying

$$\boldsymbol{\pi}^\top \mathbf{P} = \boldsymbol{\pi}^\top, \quad \sum_i \pi_i = 1.$$

2. Compute covariance matrices $\mathbf{C}_{00} = \mathbf{C}_{11} = \mathrm{diag}(\boldsymbol{\pi})$ and $\mathbf{C}_{01} = \mathrm{diag}(\boldsymbol{\pi})\mathbf{P}$.
3. Perform the SVD

$$\bar{\mathbf{K}} = \mathbf{C}_{00}^{-\frac{1}{2}} \mathbf{C}_{01} \mathbf{C}_{11}^{-\frac{1}{2}} = \mathbf{U}' \mathbf{K} \mathbf{V}'^\top$$

with $\mathbf{K} = \mathrm{diag}(\sigma_1, \ldots, \sigma_{2000})$ and $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_{2000}$.
4. Compute $\mathbf{U} = [U_{ij}] = \mathbf{C}_{00}^{-\frac{1}{2}} \mathbf{U}'$ and $\mathbf{V} = [V_{ij}] = \mathbf{C}_{11}^{-\frac{1}{2}} \mathbf{V}'$.
5. Output the stationary distribution $\mu(x) = \sum_i 50\pi_i \cdot 1_{x \in S_i}$ and singular components $(\sigma_i, \psi_i(x), \phi_i(x)) = (\sigma_i, \sum_j U_{ji} \cdot 1_{x \in S_j}, \sum_j V_{ji} \cdot 1_{x \in S_j})$.

The transition density of the projected Koopman operator $\hat{\mathcal{K}}_\tau = \sum_{i=1}^{k} \sigma_i \langle \cdot, \phi_i \rangle_{\rho_1} \psi_i$ is obtained by

$$
\hat{p}_\tau(x, y) = \hat{\mathcal{K}}_\tau \delta_y(x)
$$
$$
= \sum_{i=1}^{k} \sigma_i \psi_i(x) \phi_i(y) \mu(y) \tag{127}
$$

(see Appendix A.3) and the corresponding approximate transition matrix is

$$
\hat{\mathbf{P}} = \mathbf{U}_k^\top \mathbf{K}_k \mathbf{V}_k \mathrm{diag}(\boldsymbol{\pi}), \tag{128}
$$

where $\mathbf{U}_k$, $\mathbf{V}_k$ consist of the first k columns of $\mathbf{U}$, $\mathbf{V}$, and $\mathbf{K}_k = \mathrm{diag}(\sigma_1, \dots, \sigma_k)$. Then the relative error of $\hat{\mathcal{K}}_\tau$ in Fig. 1e can be calculated by

$$
\frac{\|\hat{\mathcal{K}}_\tau - \mathcal{K}_\tau\|_{\mathrm{HS}}}{\|\mathcal{K}_\tau\|_{\mathrm{HS}}} = \frac{\sqrt{\sum_{i=k+1}^{2000} \sigma_i^2}}{\sqrt{\sum_{i=1}^{2000} \sigma_i^2}}, \tag{129}
$$

the long-time transition density in Fig. 2 is given by

$$
\hat{p}_{n\tau}(x, y) = 50 \sum_{j} \left[\hat{\mathbf{P}}^n \right]_{ij} \cdot 1_{y \in S_j}, \tag{130}
$$

and the cumulative error of $\hat{p}_{n\tau}(x, y)$ is

$$
\mathrm{error} = \sum_{n=1}^{256} \int \mu(y)^{-1} \left(\hat{p}_{n\tau}(x, y) - p_{n\tau}(x, y) \right)^2 \mathrm{d}y
$$
$$
= \sum_{n=1}^{256} \sum_{j=1}^{2000} \pi_j^{-1} \left(\left[\hat{\mathbf{P}}^n \right]_{ij} - \left[\mathbf{P}^n \right]_{ij} \right)^2 \tag{131}
$$

for $x \in S_i$.

In Examples 2 and 3 , the smoothing parameter w are optimized by the golden-section search algorithm (Press et al. 2007) as follows for nonlinear TCCA:

1. Let $a = -6, b = 6, c = 0.618a + 0.382b, d = 0.382a + 0.618b$.
2. Compute $\mathcal{R}_2(\exp a)$, $\mathcal{R}_2(\exp b)$, $\mathcal{R}_2(\exp c)$ and $\mathcal{R}_2(\exp d)$, where $\mathcal{R}_2(w) = \left\| \mathbf{C}_{00}(w)^{-\frac{1}{2}} \mathbf{C}_{01}(w) \mathbf{C}_{11}(w)^{-\frac{1}{2}} \right\|_F^2$ and $\| \cdot \|_F$ denotes the Frobenius norm.
3. If $\max\{\mathcal{R}_2(\exp a), \mathcal{R}_2(\exp b), \mathcal{R}_2(\exp c)\} > \max\{\mathcal{R}_2(\exp b), \mathcal{R}_2(\exp c), \mathcal{R}_2(\exp d)\}$, let $(a, b, c, d) := (a, d, 0.618a + 0.382d, c)$. Otherwise, let $(a, b, c, d) := (c, b, d, 0.618b + 0.382c)$.
4. If $|a - b| < 10^{-3}$, output $\log w \in \{a, b, c, d\}$ with the largest value of $\mathcal{R}_2(w)$. Otherwise, go back to Step 2.

Furthermore, w is computed in the same way when perform nonlinear TCCA in Sects. 5.1 and 5.2 .

K.2 Double-Gyre System

For the double-gyre system in Sect. 5.1, we first perform the temporal discretization by the Euler–Maruyama scheme as

$$
\begin{aligned}
\mathbb{P}(x_{t+\Delta}|\mathbf{x}_t) &= \mathcal{N}(x_{t+\Delta}|x_t - \pi A \sin(\pi x_t)\cos(\pi y_t)\Delta, \epsilon^2(x_t/4+1)), \\
\mathbb{P}(y_{t+\Delta}|\mathbf{x}_t) &= \mathcal{N}(y_{t+\Delta}|y_t + \pi A \cos(\pi x_t)\sin(\pi y_t)\Delta, \epsilon^2),
\end{aligned}
\tag{132}
$$

where $\mathbf{x}_t = (x_t, y_t)^\top$ and $\Delta = 0.02$ is the step size. Then perform the spatial discretization as

$$
\begin{aligned}
\mathbb{P}(\mathbf{x}_{t+\Delta} \in S_j|\mathbf{x}_t \in S_i) \propto{}& \mathcal{N}(s_{j,x}|s_{i,x} - \pi A \sin(\pi s_{i,x})\cos(\pi s_{i,y})\Delta, \epsilon^2(s_{i,x}/4+1)) \\
&\cdot\mathcal{N}(s_{j,y}|s_{i,y} + \pi A \cos(\pi s_{i,x})\sin(\pi s_{i,y})\Delta, \epsilon^2).
\end{aligned}
\tag{133}
$$

Here $S_1, \ldots, S_{1250}$ are 50×25 bins which form a uniform partition of the state space $[0, 2] \times [0, 1]$ and $(s_{i,x}, s_{i,y})$ represents the center of S_i. Simulation data and the "true" singular components are all computed by using (133) with the initial distribution of (x_0, y_0) being the stationary one.

In Fig. 6, the transition density of lag time $n\tau$ is computed from the estimated singular components $(\mathbf{K}, \mathbf{U}^\top \boldsymbol{\chi}_0, \mathbf{V}^\top \boldsymbol{\chi}_1)$ as

$$
\hat{p}_{n\tau}(\mathbf{x}, \mathbf{y}) = 625 \sum_j \left[\hat{\mathbf{P}}^n\right]_{ij} \cdot \mathbf{1}_{\mathbf{y}\in S_j}, \quad \text{for } x \in S_i
\tag{134}
$$

where

$$
\hat{\mathbf{P}} = \mathbf{U}^\top \mathbf{K}\mathbf{V}\mathrm{diag}(\boldsymbol{\rho}_1)
\tag{135}
$$

is the approximate transition matrix, and $\boldsymbol{\rho}_1 = [\boldsymbol{\rho}_{1i}]$ with

$$
\boldsymbol{\rho}_{1i} = \frac{1}{T-\tau} \sum_{t=1}^{T-\tau} \mathbf{1}_{\mathbf{x}_{t+\tau} \in S_i}.
\tag{136}
$$

References

Andrew, G., Arora, R., Bilmes, J., Livescu, K.: Deep canonical correlation analysis. In: International Conference on Machine Learning, pp. 1247–1255 (2013)

Arlot, S., Celisse, A.: A survey of cross-validation procedures for model selection. Stat. Surv. **4**, 40–79 (2010)

Bollt, E.M., Santitissadeekorn, N.: Applied and Computational Measurable Dynamics. SIAM (2013)

Boninsegna, L., Gobbo, G., Noé, F., Clementi, C.: Investigating molecular kinetics by variationally optimized diffusion maps. J. Chem. Theory Comput. **11**, 5947–5960 (2015)

Bowman, G.R., Pande, V.S., Noé, F. (eds.): An Introduction to Markov State Models and Their Application to Long Timescale Molecular Simulation. Volume 797 of Advances in Experimental Medicine and Biology. Springer, Heidelberg (2014)

Brunton, S.L., Brunton, B.W., Proctor, J.L., Kutz, J.N.: Koopman invariant subspaces and finite linear representations of nonlinear dynamical systems for control. PLoS ONE **11**(2), e0150171 (2016a)

Brunton, S.L., Proctor, J.L., Kutz, J.N.: Discovering governing equations from data by sparse identification of nonlinear dynamical systems. Proc. Natl. Acad. Sci. **113**(15), 3932–3937 (2016b)

Chekroun, M.D., Simonnet, E., Ghil, M.: Stochastic climate dynamics: random attractors and time-dependent invariant measures. Physica D Nonlinear Phenom. **240**(21), 1685–1700 (2011)

Chodera, J.D., Noé, F.: Markov state models of biomolecular conformational dynamics. Curr. Opin. Struct. Biol. **25**, 135–144 (2014)

Conrad, N.D., Weber, M., Schütte, C.: Finding dominant structures of nonreversible Markov processes. Multiscale Model. Simul. **14**(4), 1319–1340 (2016)

Dellnitz, M., Froyland, G., Junge, O.: The algorithms behind gaio–set oriented numerical methods for dynamical systems. In: Fiedler, B. (ed.) Ergodic Theory, Analysis, and Efficient Simulation of Dynamical Systems, pp. 145–174. Springer, Berlin (2001)

Deuflhard, P., Weber, M.: Robust perron cluster analysis in conformation dynamics. In: Dellnitz, M., Kirkland, S., Neumann, M., Schütte, C. (eds.) Linear Algebra Application, vol. 398C, pp. 161–184. Elsevier, New York (2005)

Friedman, J., Hastie, T., Tibshirani, R.: The Elements of Statistical Learning. Springer, New York (2001)

Froyland, G.: An analytic framework for identifying finite-time coherent sets in time-dependent dynamical systems. Physica D Nonlinear Phenom. **250**, 1–19 (2013)

Froyland, G., Padberg, K.: Almost-invariant sets and invariant manifolds—connecting probabilistic and geometric descriptions of coherent structures in flows. Physica D Nonlinear Phenom. **238**(16), 1507–1523 (2009)

Froyland, G., Padberg-Gehle, K.: Almost-invariant and finite-time coherent sets: directionality, duration, and diffusion. In: Bahsoun, W., Bose, C., Froyland, G. (eds.) Ergodic Theory, Open Dynamics, and Coherent Structures, pp. 171–216. Springer, Berlin (2014)

Froyland, G., Gottwald, G.A., Hammerlindl, A.: A computational method to extract macroscopic variables and their dynamics in multiscale systems. SIAM J. Appl. Dyn. Syst. **13**(4), 1816–1846 (2014)

Froyland, G., González-Tokman, C., Watson, T.M.: Optimal mixing enhancement by local perturbation. SIAM Rev. **58**(3), 494–513 (2016)

Hardoon, D.R., Szedmak, S., Shawe-Taylor, J.: Canonical correlation analysis: an overview with application to learning methods. Neural Comput. **16**(12), 2639–2664 (2004)

Harmeling, S., Ziehe, A., Kawanabe, M., Müller, K.-R.: Kernel-based nonlinear blind source separation. Neural Comput. **15**(5), 1089–1124 (2003)

Hsing, T., Eubank, R.: Theoretical Foundations of Functional Data Analysis, with an Introduction to Linear Operators. Wiley, Amsterdam (2015)

Klus, S., Schütte, C.: Towards tensor-based methods for the numerical approximation of the perron-frobenius and koopman operator (2015). arXiv:1512.06527

Klus, S., Koltai, P., Schütte, C.: On the numerical approximation of the perron-frobenius and koopman operator (2015). arXiv:1512.05997

Klus, S., Gelß, P., Peitz, S., Schütte, C.: Tensor-based dynamic mode decomposition. Nonlinearity **31**(7), 3359 (2018)

Koltai, P., Wu, H., Noe, F., Schütte, C.: Optimal data-driven estimation of generalized Markov state models for non-equilibrium dynamics. Computation **6**(1), 22 (2018)

Konrad, A., Zhao, B.Y., Joseph, A.D., Ludwig, R.: A Markov-based channel model algorithm for wireless networks. In: Proceedings of the 4th ACM International Workshop on Modeling, Analysis and Simulation of Wireless and Mobile Systems, pp. 28–36. ACM (2001)

Koopman, B.O.: Hamiltonian systems and transformations in hilbert space. Proc. Natl. Acad. Sci. U.S.A. **17**, 315–318 (1931)

Korda, M., Mezić, I.: On convergence of extended dynamic mode decomposition to the Koopman operator. J. Nonlinear Sci. **28**(2), 687–710 (2018)

Kurebayashi, W., Shirasaka, S., Nakao, H.: Optimal parameter selection for kernel dynamic mode decomposition. In: Proceedings of the International Symposium NOLTA, volume 370, p. 373 (2016)

Li, Q., Dietrich, F., Bollt, E.M., Kevrekidis, I.G.: Extended dynamic mode decomposition with dictionary learning: a data-driven adaptive spectral decomposition of the Koopman operator. Chaos **27**(10), 103111 (2017)

Lusch, B., Kutz, J.N., Brunton, S.L.: Deep learning for universal linear embeddings of nonlinear dynamics. Nat. Commun. **9**(1), 4950 (2018)

Ma, Y., Han, J.J., Trivedi, K.S.: Composite performance and availability analysis of wireless communication networks. IEEE Trans. Veh. Technol. **50**(5), 1216–1223 (2001)

Mardt, A., Pasquali, L., Wu, H., Noé, F.: Vampnets for deep learning of molecular kinetics. Nat. Commun. **9**(1), 5 (2018)

Marshall, A.W., Olkin, I., Arnold, B.C.: Inequalities: Theory of Majorization and Its Applications, vol. 143. Springer, Berlin (1979)

McGibbon, R.T., Pande, V.S.: Variational cross-validation of slow dynamical modes in molecular kinetics. J. Chem. Phys. **142**, 124105 (2015)

Mezić, I.: Spectral properties of dynamical systems, model reduction and decompositions. Nonlinear Dyn. **41**, 309–325 (2005)

Mezić, I.: Analysis of fluid flows via spectral properties of the Koopman operator. Annu. Rev. Fluid Mech. **45**, 357–378 (2013)

Molgedey, L., Schuster, H.G.: Separation of a mixture of independent signals using time delayed correlations. Phys. Rev. Lett. **72**, 3634–3637 (1994)

Noé, F.: Probability distributions of molecular observables computed from Markov models. J. Chem. Phys. **128**, 244103 (2008)

Noé, F., Clementi, C.: Kinetic distance and kinetic maps from molecular dynamics simulation. J. Chem. Theory Comput. **11**, 5002–5011 (2015)

Noé, F., Nüske, F.: A variational approach to modeling slow processes in stochastic dynamical systems. Multiscale Model. Simul. **11**, 635–655 (2013)

Nüske, F., Keller, B.G., Pérez-Hernández, G., Mey, A.S.J.S., Noé, F.: Variational approach to molecular kinetics. J. Chem. Theory Comput. **10**, 1739–1752 (2014)

Nüske, F., Schneider, R., Vitalini, F., Noé, F.: Variational tensor approach for approximating the rare-event kinetics of macromolecular systems. J. Chem. Phys. **144**, 054105 (2016)

Otto, S.E., Rowley, C.W.: Linearly recurrent autoencoder networks for learning dynamics. SIAM J. Appl. Dyn. Syst. **18**(1), 558–593 (2019)

Paul, F., Wu, H., Vossel, M., Groot, B., Noe, F.: Identification of kinetic order parameters for non-equilibrium dynamics. J. Chem. Phys. **150**, 164120 (2018)

Perez-Hernandez, G., Paul, F., Giorgino, T., Fabritiis, G.D., Noé, F.: Identification of slow molecular order parameters for Markov model construction. J. Chem. Phys. **139**, 015102 (2013)

Press, W.H., Teukolsky, S.A., Vetterling, W.T., Flannery, B.P.: Numerical Recipes: The Art of Scientific Computing. Cambridge University Press, Cambridge (2007)

Prinz, J.-H., Wu, H., Sarich, M., Keller, B.G., Senne, M., Held, M., Chodera, J.D., Schütte, C., Noé, F.: Markov models of molecular kinetics: generation and validation. J. Chem. Phys. **134**, 174105 (2011)

Renardy, M., Rogers, R.C.: An Introduction to Partial Differential Equations. Springer, New York (2004)

Rowley, C.W., Mezić, I., Bagheri, S., Schlatter, P., Henningson, D.S.: Spectral analysis of nonlinear flows. J. Fluid Mech. **641**, 115 (2009)

Schmid, P.J.: Dynamic mode decomposition of numerical and experimental data. J. Fluid Mech. **656**, 5–28 (2010)

Schütte, C., Fischer, A., Huisinga, W., Deuflhard, P.: A direct approach to conformational dynamics based on hybrid Monte Carlo. J. Comput. Phys. **151**, 146–168 (1999)

Schwantes, C.R., Pande, V.S.: Improvements in Markov state model construction reveal many non-native interactions in the folding of NTL9. J. Chem. Theory Comput. **9**, 2000–2009 (2013)

Schwantes, C.R., Pande, V.S.: Modeling molecular kinetics with tica and the kernel trick. J. Chem. Theory Comput. **11**, 600–608 (2015)

Snoek, J., Larochelle, H., Adams, R.P.: Practical bayesian optimization of machine learning algorithms. In: Pereira, F., Burges, C.J.C., Bottou, L., Weinberger, K.Q. (eds.) Advances in Neural Information Processing Systems, pp. 2951–2959 (2012)

Song, L., Fukumizu, K., Gretton, A.: Kernel embeddings of conditional distributions: a unified kernel framework for nonparametric inference in graphical models. IEEE Signal Process. Mag. **30**(4), 98–111 (2013)

Sparrow, C.: The Lorenz Equations: Bifurcations, Chaos, and Strange Attractors. Springer, New York (1982)

Takeishi, N., Kawahara, Y., Yairi, T.: Learning Koopman invariant subspaces for dynamic mode decomposition. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems, pp. 1130–1140 (2017)

Tibshirani, R.: Regression shrinkage and selection via the lasso. J. R. Stat. Soc. Ser. B (Methodol.) **58**, 267–288 (1996)

Tu, J.H., Rowley, C.W., Luchtenburg, D.M., Brunton, S.L., Kutz, J.N.: On dynamic mode decomposition: theory and applications. J. Comput. Dyn. **1**(2), 391–421 (2014)

Williams, M.O., Kevrekidis, I.G., Rowley, C.W.: A data-driven approximation of the Koopman operator: extending dynamic mode decomposition. J. Nonlinear Sci. **25**, 1307–1346 (2015a)

Williams, M.O., Rowley, C.W., Kevrekidis, I.G.: A kernel-based method for data-driven Koopman spectral analysis. J. Comput. Dyn. **2**(2), 247–265 (2015b)

Wu, H., Noé, F.: Gaussian Markov transition models of molecular kinetics. J. Chem. Phys. **142**, 084104 (2015)

Wu, H., Nüske, F., Paul, F., Klus, S., Koltai, P., Noé, F.: Variational Koopman models: slow collective variables and molecular kinetics from short off-equilibrium simulations. J. Chem. Phys. **146**, 154104 (2017)

Ziehe, A., Müller, K.-R.: TDSEP —an efficient algorithm for blind separation using time structure. In: ICANN 98, pp. 675–680. Springer (1998)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Affiliations

Hao Wu[1,2] [ID] · **Frank Noé[2,3,4]**

✉ Hao Wu
    hwu@tongji.edu.cn

✉ Frank Noé
    frank.noe@fu-berlin.de

[1] School of Mathematical Sciences, Tongji University, 1239 Siping Road, Shanghai 200092, China

[2] Department of Mathematics and Computer Science, Freie Universität Berlin, Arnimallee 6, 14195 Berlin, Germany

[3] Department of Physics, Freie Universität Berlin, Arnimallee 14, 14195 Berlin, Germany

[4] Department of Chemistry, Rice University, Houston, TX 77005, USA