Projected metastable Markov processes and their estimation with observable operator models

Hao Wu, Jan-Hendrik Prinz, and Frank Noé

Citation: The Journal of Chemical Physics **143**, 144101 (2015); doi: 10.1063/1.4932406
View online: http://dx.doi.org/10.1063/1.4932406
View Table of Contents: http://scitation.aip.org/content/aip/journal/jcp/143/14?ver=pdfcov
Published by the AIP Publishing

Projected metastable Markov processes and their estimation with observable operator models

Hao Wu,[a] Jan-Hendrik Prinz,[b] and Frank Noé[c]
DFG Research Center Matheon, Free University Berlin, Arnimallee 6, 14195 Berlin, Germany

(Received 6 August 2015; accepted 23 September 2015; published online 8 October 2015)

The determination of kinetics of high-dimensional dynamical systems, such as macromolecules, polymers, or spin systems, is a difficult and generally unsolved problem — both in simulation, where the optimal reaction coordinate(s) are generally unknown and are difficult to compute, and in experimental measurements, where only specific coordinates are observable. Markov models, or Markov state models, are widely used but suffer from the fact that the dynamics on a coarsely discretized state spaced are no longer Markovian, even if the dynamics in the full phase space are. The recently proposed projected Markov models (PMMs) are a formulation that provides a description of the kinetics on a low-dimensional projection without making the Markovianity assumption. However, as yet no general way of estimating PMMs from data has been available. Here, we show that the observed dynamics of a PMM can be exactly described by an observable operator model (OOM) and derive a PMM estimator based on the OOM learning. © *2015 AIP Publishing LLC.* [http://dx.doi.org/10.1063/1.4932406]

I. INTRODUCTION

High-dimensional Markov processes are ubiquitous in physical systems. A recurring and hard problem is the task of estimating effective dynamical models from low-dimensional projections of such processes. Examples include effective modeling of protein kinetics from all-atom simulations, the estimation of transition rates from single-molecule force probe experiments, or the detection of hidden states in ion channels from patch-clamp voltage measurements.

A general formalism is that of Mori and Zwanzig,[1,2] where the projected Markov dynamics lead to a generalized Langevin equation. One may attempt to directly estimate such Langevin models from data,[3,4] but in general their parameters, in particular the memory kernel, are difficult to deal with, and require some regularization.

Many dynamical systems are metastable, i.e., the dynamics on large time scales is mainly governed by transitions between "metastable subsets" in state space while the local equilibrium within each metastable subset can be reached quickly. In the fortunate situation that the observed subspace resolves the coordinates indicative of these slow processes, the Langevin memory kernel may decay sufficiently quickly, such that the process can be described by Markov state models (MSMs). MSMs use memoryless transition or rate matrices on a set decomposition (e.g., Voronoi tessellation) of the observed subspace.[5–12] Due to the Markov assumption, the parameters of MSMs can be efficiently estimated from simulation or experimental data. Both thermodynamic and kinetic quantities of the metastable systems can be easily obtained from MSMs — including equilibrium expectations, free energy differences, the characteristic time scales of equilibrium processes, time correlation functions, reaction rates, and transition pathways. However, the Markov assumption relies heavily on an appropriately chosen state space discretization, generally requiring a large number of discrete states that resolve the transition states of metastable subsets.[13,14] Such a discretization is difficult to find in high-dimensional applications. In Refs. 10 and 15, MSMs between "core sets" have been proposed to solve this problem, which avoid the full partition of the state space by using a set of core sets to define the discrete states instead. However, the selection of core sets is still not satisfactorily solved. For the analysis of single-molecule experimental data, the direct application of MSMs is often impossible because the observable subspace (e.g., a pulling coordinate in a force probe experiment) is not Markovian, thereby inherently limiting the discretization quality.

Here we pursue an approach that does *not* enforce the assumption that the observed dynamics are (nearly) Markovian. We developed a new framework in Ref. 16, called projected Markov models (PMMs), which allows one to investigate the coarse-grained metastable dynamics without imposing the Markov assumption on the discretized state space. A PMM consists of two sub-models: (a) a *state evolution model* which is Markovian on the full state space without any discretization and is assumed to have a spectral gap due to the metastability, and (b) an *observation model* which describes the system dynamics observed on the discrete state space as a deterministic or stochastic projection of the state evolution model. PMMs were successfully applied to the problem of estimating the rates of a two-state system (see Ref. 17). But until now, it has been an open problem how to characterize the complete observed dynamics of a general PMM by a low dimensional parametric model which can be efficiently

[a]Electronic mail: hao.wu@fu-berlin.de
[b]Electronic mail: jan-hendrik.prinz@fu-berlin.de
[c]Electronic mail: frank.noe@fu-berlin.de

estimated from the observation data. In Ref. 16, we have proposed to replace the PMM by a discrete-state hidden Markov model (HMM), for which efficient estimation methods are known. However, PMMs can be accurately modeled by HMMs only under a number of nontrivial assumptions on the eigenfunctions of the dynamical system that cannot be verified in practice.

Here we propose a general and efficient approach to the modeling and estimation problems of PMMs. The approach relies on establishing the observable operator model (OOM) representation of PMMs, where OOMs are a class of dynamic algebra systems developed in the field of machine learning as an extension of HMMs.[18] We show that the observed dynamics of a strongly metastable PMM can be described by an OOM. This allows us to characterize the non-Markovian dynamics of metastable systems on the observable coarse-grained state space through building OOMs, and the statistical and spectral properties of the metastable dynamics can then be directly computed from the OOM parameters.

II. PROJECTED MARKOV MODELS

Before presenting our results, we give a formal description and some basic assumptions of the molecular dynamics, and describe PMMs.

a. *State evolution model*: The molecular dynamics in full state space is described by a stochastic process $\{x_t\}$ where $x \in \Omega$, an often high-dimensional state space that may include momenta, and t is the time index. Trajectories from a molecular dynamics implementation or a single-molecule experiment sample this process and provide realizations of $\{x_t\}$. Furthermore, we will assume that the process itself is an ergodic and reversible Markov process with a unique and everywhere positive stationary density $\mu(x)$. In practice this means: if the dynamical system is a single-molecule experiment in equilibrium, these conditions will be fulfilled. If the dynamical system is a molecular dynamics (MD) simulation, we must choose an integrator and thermostat that will sample from the correct Boltzmann distribution $\mu(x) \propto \exp(-\beta u(x))$ with inverse temperature β. Secondly, reversibility means that when we simulate the molecule in equilibrium, the absolute (unconditional) probability to travel between two points x and x' is symmetric. This property is important for physically realistic system and is actually required by the second law of thermodynamics. However, many practically used integrators and thermostats do not exactly obey reversibility. Many dynamical models, such as Langevin dynamics, obey some generalized reversibility under the inversion of momenta. In fact these models can also be dealt with, but here we will restrict the mathematical description to reversible dynamics implementations in order to keep the equations reasonably simple. See Refs. 14 and 19 for a more detailed discussion. In practice, most reasonable MD implementations are sufficiently consistent with the assumptions above for the purpose of this paper.

The operator theory of Markov processes was thoroughly discussed in Refs. 13, 14, and 19. Here, we just summarize a few aspects of the theory relevant for analyzing PMMs. The dynamics of $\{x_t\}$ can be formally described by a propagator $\mathcal{P}(\tau)$ with

$$p_{t+\tau}(x) = (\mathcal{P}(\tau)\, p_t)(x) \triangleq \int_\Omega \mathrm{d}x' p(x_{t+\tau} = x | x_t = x')\, p_t(x'), \tag{1}$$

where $p_t(x)$ denotes the probability density of the system state x_t at time t, $\mathcal{P}(\tau)$ is an integral operator, and the transition probability density $p(x_{t+\tau} = x | x_t = x')$ is the corresponding integral kernel. From the ergodicity and reversibility of $\{x_t\}$, we can conclude that $\mathcal{P}(\tau)$ is a compact and self-adjoint operator on a Hilbert space defined by the weighted inner product

$$\langle u, v \rangle = \int_\Omega \mathrm{d}x u(x)\, v(x)\, \mu^{-1}(x), \tag{2}$$

and $p_{t+\tau}$ can be decomposed by using spectral components of $\mathcal{P}(\tau)$ as

$$p_{t+\tau} = \sum_{i=1}^{\infty} \lambda_i(\tau) \langle p_t, \phi_i \rangle \phi_i, \tag{3}$$

for a given state density p_t at time t, where

$$\lambda_i(\tau) = \exp(-\kappa_i \tau) = \exp\left(-\frac{\tau}{\mathrm{ITS}_i}\right) \tag{4}$$

is the ith largest magnitude eigenvalue of $\mathcal{P}$ with eigenfunction ϕ_i. $\kappa_i \geq 0$ is the (experimentally observable) relaxation rate of the ith dynamical process, and $\mathrm{ITS}_i = \kappa_i^{-1}$ the corresponding relaxation time scale. The eigenfunctions $\{\phi_i\}$ form an orthonormal basis for the range of $\mathcal{P}$, i.e., $\langle \phi_i, \phi_j \rangle = \delta_{ij}$ for all i, j. The largest eigenvalues is $\lambda_1 = 1$ and dominant, $\lambda_1 > \lambda_2$. Consequently, the first relaxation rate is $\kappa_1 = 0$ and $\kappa_1 < \kappa_2$. The corresponding eigenfunction is $\phi_1(x) = \mu(x)$, the stationary or equilibrium distribution. Note that (4) furthermore assumes that eigenvalues are positive, this assumption is reasonable for most physical and chemical processes.

If $\{x_t\}$ is a metastable process, there may be a spectral gap between the decay rates of the largest few eigenvalues and the remaining ones which are separately denoted as $\{\kappa_i\}_{i=1}^m$ and $\{\kappa_i\}_{i>m}$ with $\kappa_m \ll \kappa_{m+1}$ (see, e.g., Refs. 20 and 21). The fast rates quickly decay to zero as $\tau \to \infty$, and then the dynamics of $\{x_t\}$ can be well approximated as

$$p_{t+\tau} \approx \sum_{i=1}^{m} \lambda_i(\tau) \langle p_t, \phi_i \rangle \phi_i \tag{5}$$

on sufficiently long time scales $\tau \gg 1/\kappa_{m+1}$. In this paper, we restrict our discussion to the case that $\{x_t\}$ is "m-metastable" with $\kappa_i = \infty$ for all $i > m$ and the above approximation is exact for convenience of the analysis. Note that in practice $\tau > 3/\kappa_{m+1}$ is already sufficient for this approximation to be excellent.

Using (5), we can write down the joint probability to observe the system in state x_t at time t and state $x_{t+\tau}$ at time $t + \tau$,

$$p(x_t = x', x_{t+\tau} = x) \approx \sum_{i=1}^{m} \lambda_i(\tau)\, \phi_i(x') \phi_i(x). \tag{6}$$

b. *Observation model*: So far we have assumed the dynamics to be Markovian. The strategy of Markov models is to directly

approximate the eigenfunctions ϕ_i by seeking a suitable state space approximation, which can be efficient for molecular dynamics simulations, if an appropriate low-dimensional state space can be found.[22,23] However, for large molecular systems, obtaining a good state space discretization may be difficult. Moreover, if we want to model the kinetics of an experimentally observed system, we do not have the freedom of choosing a suitable space and making the state space discretization arbitrarily good—we are forced to observe the system dynamics projected on an experimentally observable parameter.[17] Such a projection will render the dynamics non-Markovian and will deteriorate the estimation of the kinetics — in general in such a way that eigenvalues and time scales are underestimated, and rates are overestimated.[17,24,25]

Now we will explicitly assume that we discretize the Markovian dynamics $\{x_t\}$ or observe its projection onto an experimentally accessible parameter. For this, let us assume that the observation y_t at each time t is a function of x_t. In the present paper we assume that the observation space $O = \{1, \ldots, N\}$ is a finite set. For a molecular dynamics simulation this can be achieved by conducting a state space discretization using data clustering. For experimental data, it is easily achieved by binning the experimental observable(s).

Considering that the observation might be noisy, here we describe the relationship between y_t and x_t as

$$\Pr(y_t = n | x_t = x) = \chi_n(x), \tag{7}$$

where $\chi_n(x)$ is called the observation probability function for the observed value n. Note that (7) can be used to describe a Galerkin discretization with $\chi_n(x) \in \{0,1\}$, where each k represents a finite element space and $\chi_n(x)$ is the corresponding characteristic basis function. This way we can straightforwardly describe the case of fuzzy clustering or smooth membership functions, e.g., Gaussian basis functions used in Refs. 25–27.

Now we can model the dynamics on the observed state space. Using (6), we can write down the joint probability of observing the system in two states y_t at time t and $y_{t+\tau}$ at time $t + \tau$,

$$\begin{aligned} C_{ij} &\triangleq \Pr(y_t = i, y_{t+\tau} = j) \\ &\approx \int_{x_t \in \Omega} dx_t \chi_i(x_t) \int_{x_{t+\tau} \in \Omega} dx_{t+\tau} \chi_i(x_{t+\tau}) \\ &\quad \times \sum_{k=1}^{m} \lambda_k(\tau) \phi_k(x_t) \phi_k(x_{t+\tau}) \\ &= \sum_{k=1}^{m} \lambda_k(\tau) q_{ik} q_{jk}, \end{aligned} \tag{8}$$

where $\mathbf{q}_k = (q_{1k}, \ldots, q_{Nk})^\top$ is the projection of eigenfunction ϕ_k onto the observed state space with

$$q_{ik} = \int_\Omega \phi_k(x) \chi_i(x) \, \mathrm{d}x. \tag{9}$$

We can also express the correlation matrix $\mathbf{C} = [C_{ij}]$ in a compact form as

$$\mathbf{C} = \mathbf{Q} \mathbf{\Lambda}(\tau) \mathbf{Q}^\top, \tag{10}$$

where the columns of $\mathbf{Q} = (\mathbf{q}_1, \ldots, \mathbf{q}_m) \in \mathbb{R}^{N \times m}$ contain the elements of the first m eigenfunctions projected onto the observed states, and $\mathbf{\Lambda}(\tau) \in \mathbb{R}^{m \times m}$ contain the first m eigenvalues of the dynamical systems. In the rest of the paper, we write the eigenvalue matrix as $\mathbf{\Lambda}$ with omitting τ from the notation if there is no understanding.

Eq. (10) provides a spectral description of the observation process $\{y_t\}$ of a PMM as derived in Ref. 28. It is these two quantities, $\mathbf{Q}$ and $\mathbf{\Lambda}$ that are of our primary interest and that are important for spectral analysis of the system's kinetics, although they *cannot* fully describe the dynamics of the PMM on the observation space due to the non-Markovianity of $\{y_t\}$. If we have them, we can compute the temporal evolution of the observation distribution and a lot of physically interesting quantities on the PMM. See Ref. 28 for more details on a quantitative analysis of PMMs based on $(\mathbf{\Lambda}, \mathbf{Q})$.

Now we address the question: how can we efficiently approximate the dynamics of a PMM on the observation space and estimate projected spectral components $(\mathbf{\Lambda}, \mathbf{Q})$ from observation data?

III. REPRESENTATION AND ESTIMATION OF PROJECTED MARKOV MODELS BY OBSERVABLE OPERATOR MODELS

A. Observable operator models

In this subsection, we formally define the elements of an OOM and explain how the model represents a discrete-valued stochastic process. According to Refs. 18 and 29, an OOM can be characterized by the following variables.

1. m, the dimension of states in the model. The OOM state at each time t is represented by a row vector $\boldsymbol{\omega}_t \in \mathbb{R}^{1 \times m}$.
2. N, the number of distinct observable states. We denote the observation space as $O = \{1, \ldots, N\}$, and the observation at time t as y_t.
3. The initial state vector $\boldsymbol{\omega}$ with $\boldsymbol{\omega}_0 = \boldsymbol{\omega}$.
4. An observable operator $\mathbf{\Xi}^{(y)} \in \mathbb{R}^{m \times m}$ for each observation value $y \in O$. For a given observation sequence $y_\tau, y_{2\tau}, \ldots$, the evolution equation of model states can be expressed as
$$\boldsymbol{\omega}_{k\tau} = \boldsymbol{\omega}_{(k-1)\tau} \mathbf{\Xi}^{(y_{k\tau})}. \tag{11}$$
5. The evaluation vector $\boldsymbol{\sigma} \in \mathbb{R}^{m \times 1}$, which maps the model state $\boldsymbol{\omega}_t$ to the probability of the observation sequence up to time t as
$$\boldsymbol{\omega}_{k\tau} \boldsymbol{\sigma} = \Pr(y_\tau, y_{2\tau}, \ldots, y_{k\tau}). \tag{12}$$

Consequently, we can write the probability of observing an observation sequence $(y_\tau, y_{2\tau}, \ldots, y_{k\tau})$ in terms of the OOM parameters as:

$$\Pr(y_\tau, y_{2\tau}, \ldots, y_{k\tau}) = \boldsymbol{\omega} \mathbf{\Xi}^{(y_\tau)} \mathbf{\Xi}^{(y_{2\tau})} \ldots \mathbf{\Xi}^{(y_{k\tau})} \boldsymbol{\sigma} \tag{13}$$

for all $k \geq 0$ (assuming the convention that Pr(empty sequence) = 1).

OOMs are closely related to another class of non-Markovian models called HMMs. In Ref. 30, it was shown that each m-state HMM can be expressed as a m-dimensional OOM. Conversely, however, there exist finite-dimensional OOMs which cannot be formulated as HMMs with a finite

number of states. In other words, OOMs are a strict generalization of HMMs and are therefore able to model a wider class of stochastic processes.

B. Converting projected Markov models into observable operator models

Considering that efficient estimation algorithms for OOMs have been developed, we want to estimate the observed dynamics of a PMM and projected spectral components through OOM learning in this paper. To this end, it must be shown that the observation process $\{y_t\}$ produced by a PMM can also described by an OOM.

Let us consider a metastable PMM which contains only m nonzero eigenvalues as described in Section II. According to (6), the transition probability density of the state process $\{x_t\}$ in the PMM can be expressed as

$$\begin{aligned} p(x_{t+\tau}|x_t) &= \frac{1}{\mu(x_t)} p(x_t, x_{t+\tau}) \\ &= \frac{1}{\mu(x_t)} \boldsymbol{\phi}(x_t)^\top \boldsymbol{\Lambda} \boldsymbol{\phi}(x_{t+\tau}), \end{aligned} \tag{14}$$

where $\boldsymbol{\phi}(x) = (\phi_1(x), \ldots, \phi_m(x))^\top$ denotes the vector of eigenfunctions. Combining (14) and (7) yields the probability of the observation sequence,

$$\Pr(y_\tau, \ldots, y_{k\tau}) = \boldsymbol{\omega}_P \boldsymbol{\Xi}_P^{(y_\tau)} \ldots \boldsymbol{\Xi}_P^{(y_{k\tau})} \boldsymbol{\sigma}_P \tag{15}$$

(see proof in Appendix A), where the OOM parameters can be expressed in terms of the PMM spectral components as

$$\boldsymbol{\omega}_P = (\langle p_0, \phi_1 \rangle, \ldots, \langle p_0, \phi_m \rangle) \tag{16}$$

$$\boldsymbol{\Xi}_P^{(y)} = \boldsymbol{\Lambda} \begin{bmatrix} \langle \chi_y \cdot \phi_1, \phi_1 \rangle & \cdots & \langle \chi_y \cdot \phi_1, \phi_m \rangle \\ \vdots & \ddots & \vdots \\ \langle \chi_y \cdot \phi_m, \phi_1 \rangle & \cdots & \langle \chi_y \cdot \phi_m, \phi_m \rangle \end{bmatrix} \tag{17}$$

$$\boldsymbol{\sigma}_P = (1, 0, \ldots, 0)^\top \tag{18}$$

and $(\chi_y \cdot \phi_i)(x) \triangleq \chi_y(x) \phi_i(x)$. Thus, the observed process $\{y_t\}$ produced by the PMM can also be described by the m-dimensional OOM $\mathcal{M}_P = (\boldsymbol{\omega}_P, \{\boldsymbol{\Xi}_P^{(y)}\}_{y \in O}, \boldsymbol{\sigma}_P)$.

With the above discussions, we can now present our main conclusion in this paper: *The observed dynamics of a PMM can be exactly characterized by a m-dimensional OOM under the assumption of m-metastability*. This conclusion shows that we can analyze the dynamical characteristics and evaluate the path probabilities of a PMM in the observation space by a simple and small-sized linear algebra system if the PMM is strongly metastable.

C. Observable operator model learning for projected Markov models

According to the conclusion in Subsection III B, we can approximate the dynamics of a m-metastable PMM from its observations $\{y_t\}$ by an OOM learning algorithm if the following assumptions holds:

1. The underlying stochastic process $\{x_t\}$ is stationary, i.e., the initial probability density of the system state $p_0(x) = \mu(x)$. Under this assumption, we can estimate the stationary distribution vector $\boldsymbol{\pi}$, the correlation matrix $\mathbf{C}$ and two-step correlation matrices $\{\mathbf{C}^{(y)}\}_{y \in O}$ consistently and without a bias, through simple counting. Here, $\boldsymbol{\pi}$, $\mathbf{C}$ and $\mathbf{C}^{(y)}$ are defined as
$$\boldsymbol{\pi} = [\pi_i] = [\Pr(y_t = i)], \tag{19}$$
$$\mathbf{C} = [C_{ij}] = [\Pr(y_t = i, y_{t+\tau} = j)], \tag{20}$$
$$\mathbf{C}^{(y)} = [C_{ij}^{(y)}] = [\Pr(y_{t-\tau} = i, y_t = y, y_{t+\tau} = j)]. \tag{21}$$
2. The m projected eigenfunctions $\mathbf{q}_1, \ldots, \mathbf{q}_m$ are linearly independent, i.e., rank $(\mathbf{Q}) = m$. We can conclude from this assumption that rank $(\mathbf{C})$ = rank $(\mathbf{Q}\boldsymbol{\Lambda}\mathbf{Q}^\top) = m$.
3. All the m dominant eigenvalues $\lambda_1, \ldots, \lambda_m$ are positive, which implies that each λ_i is related to a valid relaxation time scale as in (4).

From these assumptions and the reversibility of the PMM, it can be shown that $\mathbf{C}$ and $\mathbf{C}^{(y)}$ have the following structural properties:

$$\begin{aligned} \mathbf{C} &\succeq 0 \\ \mathbf{C}^{(y)} &= \mathbf{C}^{(y)\top}, \quad \forall y \in O \end{aligned} \tag{22}$$

where $\mathbf{C} \succeq 0$ means that $\mathbf{C}$ is a positive semidefinite matrix, which is a direct consequence of the third assumption.

We can exploit Algorithm 1 to perform the OOM learning of the PMM observed dynamics. This algorithm is similar to the weighted spectral learning algorithm[29] for OOM estimation, and it is asymptotically correct, i.e., the OOM $\hat{\mathcal{M}}$ given by the estimation algorithm satisfies

$$\hat{\boldsymbol{\omega}} \hat{\boldsymbol{\Xi}}^{(y_\tau)} \ldots \hat{\boldsymbol{\Xi}}^{(y_{k\tau})} \hat{\boldsymbol{\sigma}} \to \Pr(y_\tau, \ldots, y_{k\tau}), \tag{23}$$

if the estimates $\hat{\boldsymbol{\pi}}$, $\hat{\mathbf{C}}$ and $\{\hat{\mathbf{C}}^{(y)}\}_{y \in O}$ of $\boldsymbol{\pi}$, $\mathbf{C}$ and $\{\mathbf{C}^{(y)}\}_{y \in O}$ converge to their true values. The only difference between Algorithm 1 and the weighted spectral learning algorithm is that the estimates of $\mathbf{C}$ and $\{\mathbf{C}^{(y)}\}$ obtained by simple counting are numerically modified according to the algebraic constraints (22) (see line 2 of Algorithm 1). This modification ensures that the sum of all estimated observable operators is diagonalizable over the real numbers, i.e., that there exist a diagonal matrix $\hat{\boldsymbol{\Lambda}} \in \mathbb{R}^{m \times m}$ and a nonsingular matrix $\hat{\mathbf{W}} \in \mathbb{R}^{m \times m}$ such that

$$\hat{\boldsymbol{\Xi}}^{(0)} \triangleq \sum_{y \in O} \hat{\boldsymbol{\Xi}}^{(y)} = \hat{\mathbf{W}} \hat{\boldsymbol{\Lambda}} \hat{\mathbf{W}}^{-1} \tag{24}$$

even in the presence of statistical noise (see Appendix B). It will be seen in Section IV that the diagonalizability is very important for the spectral analysis. In this paper, we perform the numerical modification via Algorithm 2.

Remark 1. Note that an OOM generally has multiple equivalent OOMs which have different parameters.[18] So (23) implies that $\hat{\mathcal{M}}$ is asymptotically equivalent to $\mathcal{M}_P$ constructed in Subsection III B, but we *cannot* conclude from the equivalence that

$$\hat{\boldsymbol{\omega}} \to \boldsymbol{\omega}_P, \quad \hat{\boldsymbol{\Xi}}^{(y)} \to \boldsymbol{\Xi}_P^{(y)}, \quad \hat{\boldsymbol{\sigma}} \to \boldsymbol{\sigma}_P.$$

Remark 2. An important issue for PMM estimation is determining the number m of dominant eigenvalues. In our results, we assume that m is given. For practical applications

ALGORITHM 1. OOM learning for PMMs.

1. Calculate the estimates $\hat{\boldsymbol{\pi}}$, $\hat{\mathbf{C}}$, and $\{\hat{\mathbf{C}}^{(y)}\}_{y\in O}$ of $\boldsymbol{\pi}$, $\mathbf{C}$, and $\{\mathbf{C}^{(y)}\}_{y\in O}$ by counting frequencies of occurrence of the corresponding observation values and subsequences as follows:

$$\begin{aligned}\hat{\boldsymbol{\pi}} = [\hat{\pi}_i] &= \frac{1}{Z_\pi}[|\{t|y_t = i\}|],\\ \hat{\mathbf{C}} = [\hat{c}_{ij}] &= \frac{1}{Z_C}[|\{t|(y_t, y_{t+\tau}) = (i,j)\}|],\\ \hat{\mathbf{C}}^{(y)} = [\hat{c}_{ij}^{(y)}] &= \frac{1}{Z'_C}[|\{t|(y_{t-\tau}, y_t, y_{t+\tau}) = (i,y,j)\}|],\end{aligned} \tag{25}$$

where Z_π, Z_C, and Z'_C are normalizing constants determined by

$$\begin{aligned}\sum_i \hat{\pi}_i &= 1,\\ \sum_{i,j} \hat{c}_{ij} &= 1,\\ \sum_{i,j,y} \hat{c}_{ij}^{y} &= 1.\end{aligned} \tag{26}$$

2. Modify the estimates $\hat{\mathbf{C}}$ and $\{\hat{\mathbf{C}}^{(y)}\}$ to satisfy the constraint (22).
3. Define the matrix

$$\mathbf{S} = \mathrm{diag}(\hat{\boldsymbol{\pi}})^{-\frac{1}{2}}\hat{\mathbf{C}}\mathrm{diag}(\hat{\boldsymbol{\pi}})^{-\frac{1}{2}} \tag{27}$$

and perform the eigenvalue decomposition of $\mathbf{S}$ to get the diagonal matrix $\boldsymbol{\Sigma}_m \in \mathbb{R}^{m\times m}$ which contains the m largest eigenvalues of $\mathbf{S}$, and the matrix $\mathbf{U}_m \in \mathbb{R}^{n\times m}$ consisting of the corresponding right orthonormal eigenvectors.
4. Using $\mathbf{V} := \mathrm{diag}(\hat{\boldsymbol{\pi}})^{-\frac{1}{2}}\mathbf{U}_m$, compute

$$\begin{aligned}\hat{\boldsymbol{\omega}} &= \hat{\boldsymbol{\pi}}^\top\mathbf{V}\\ \hat{\boldsymbol{\Xi}}^{(y)} &= \boldsymbol{\Sigma}_m^{-1}\mathbf{V}^\top\hat{\mathbf{C}}^{(y)}\mathbf{V}\\ \hat{\boldsymbol{\sigma}} &= \boldsymbol{\Sigma}_m^{-1}\mathbf{V}^\top\hat{\boldsymbol{\pi}}\end{aligned} \tag{28}$$

5. **return** OOM $\hat{\mathcal{M}} = (\hat{\boldsymbol{\omega}}, \{\hat{\boldsymbol{\Xi}}^{(y)}\}_{y\in O}, \hat{\boldsymbol{\sigma}})$.

where the assumption of the linear independence of projected eigenfunctions holds, m can be determined by calculating the numerical rank of the estimates of $\mathbf{C}$ since $\mathrm{rank}(\mathbf{C}) = m$.

Remark 3. If $\mathbf{Q}$ is rank deficient, it is still possible to determine the value of m and perform the OOM learning by using the stationary distribution, correlation matrix, and two-step correlation matrices of observation subsequences.[29] However, it is an open problem on how to achieve a diagonalizable $\hat{\boldsymbol{\Xi}}^{(0)}$ in this case.

D. Comparison with other modeling approaches

The most popular approach for estimating PMMs are Markov state models, where the observed process $\{y_t\}$ is modeled as a Markov chain. This is, of course, an oversimplified description of the non-Markovian dynamics of PMMs and can achieve an accurate approximation only in the case that the coarse-graining defined by the observation model is fine enough and can be conducted in the space spanned by the slow process eigenvectors. The HMM approach provides a non-Markovian description of PMMs by treating each

ALGORITHM 2. Modification of $\hat{\mathbf{C}}$ and $\{\hat{\mathbf{C}^{(y)}}\}$.

1. Calculate

$$\begin{aligned}\hat{\mathbf{C}}'^{(y)} &= \frac{1}{2}\left(\hat{\mathbf{C}}^{(y)} + \hat{\mathbf{C}}^{(y)\top}\right), \quad \forall y \in O\\ \hat{\mathbf{C}}' &= \frac{1}{2}\left(\hat{\mathbf{C}} + \hat{\mathbf{C}}^\top\right).\end{aligned} \tag{29}$$

2. Perform the eigenvalue decomposition of $\hat{\mathbf{C}}'$ to get

$$\hat{\mathbf{C}}' = \mathbf{U}_C\boldsymbol{\Sigma}_C\mathbf{U}_C^\top, \tag{30}$$

where $\boldsymbol{\Sigma}_C$ is a diagonal matrix consisting of eigenvalues of $\hat{\mathbf{C}}'$, and $\mathbf{U}_C$ consists of the corresponding right orthonormal eigenvectors.
3. Let $\hat{\mathbf{C}}^{(y)} := \hat{\mathbf{C}}'^{(y)}$ for all y and

$$\hat{\mathbf{C}} := \mathbf{U}_C\max\{\boldsymbol{\Sigma}_C, 0\}\mathbf{U}_C^\top. \tag{31}$$

4. **return** $\hat{\mathbf{C}}$ and $\{\hat{\mathbf{C}}^{(y)}\}$.

metastable state as a hidden state and can robustly approximate PMM dynamics even when the state discretization is very poor. In Ref. 16, it was shown that the observation process of a m-metastable PMM can only be described by a m-state HMM if the eigenfunctions of the state process can be expressed as linear combinations of a set of probability density functions with non-overlapping supports. This assumption is obviously unrealistic in real world scenarios and can only be approximately satisfied when the transition probabilities between metastable states in the state space are very close to zero in most practical cases. In contrast with MSMs and HMMs, OOMs can *exactly* describe the PMM dynamics without any assumption on the state evolution and observation models except the m-metastability. While being only slightly more complex, OOMs provide a much improved model of state-discretized molecular dynamics and are less affected by model mismatching than MSMs and HMMs.

Interestingly, both MSM and OOM parameters can be simply extracted from correlation functions (or high-order correlation functions) of observed processes $\{y_t\}$. Therefore, both MSMs and OOMs can be efficiently learned from observation data, and the statistical errors would go to zero with increasing size of observation data almost surely. Note that HMMs do not have this property: HMM parameters can only be iteratively optimized through an expectation-maximization procedure, which often suffers from slow convergence and susceptibility to local optima and is generally not asymptotically correct.

IV. OBSERVABLE OPERATOR MODEL BASED SPECTRAL ANALYSIS

We now investigate how to recover the PMM eigenvalues $\mathbf{\Lambda}$ and projected eigenfunctions $\mathbf{Q}$ defined in (10) from the corresponding OOM parameters.

Let us consider a m-metastable PMM, which can be exactly described as a m-dimensional OOM $\mathcal{M} = (\boldsymbol{\omega}, \{\mathbf{\Xi}^{(y)}\}_{y\in\mathcal{O}}, \boldsymbol{\sigma})$ and denote by $\mathbf{\Xi}^{(0)} \triangleq \sum_{y\in\mathcal{O}} \mathbf{\Xi}^{(y)}$ the sum of all observable operators in $\mathcal{M}$. Furthermore, we assume for simplicity that all assumptions listed in Subsection III C are satisfied and all the m dominant eigenvalues of the Markov propagator of the state process are distinct, i.e., $1 = \lambda_1 > \lambda_2 > \cdots > \lambda_m > 0$. Then it can be proven that $\mathbf{\Xi}^{(0)}$ is a nonsingular matrix and similar to a real diagonal matrix (see Eq. (C4) in Appendix C). According to Eq. (13), the correlation matrix $\mathbf{C}$ of the PMM can be calculated by

$$\mathbf{C} = \begin{bmatrix} \boldsymbol{\omega}\mathbf{\Xi}^{(1)} \\ \vdots \\ \boldsymbol{\omega}\mathbf{\Xi}^{(N)} \end{bmatrix} \begin{bmatrix} \mathbf{\Xi}^{(1)}\boldsymbol{\sigma} & \cdots & \mathbf{\Xi}^{(N)}\boldsymbol{\sigma} \end{bmatrix}. \tag{32}$$

After some algebraic manipulations, we can get

$$\begin{aligned} \mathbf{C} &= \begin{bmatrix} \boldsymbol{\omega}\mathbf{\Xi}^{(1)} \\ \vdots \\ \boldsymbol{\omega}\mathbf{\Xi}^{(N)} \end{bmatrix} \mathbf{W}\tilde{\mathbf{\Lambda}}\mathbf{W}^{-1}\left(\mathbf{\Xi}^{(0)}\right)^{-1} \begin{bmatrix} \mathbf{\Xi}^{(1)}\boldsymbol{\sigma} & \cdots & \mathbf{\Xi}^{(N)}\boldsymbol{\sigma} \end{bmatrix} \\ &= \tilde{\mathbf{Q}}\tilde{\mathbf{\Lambda}}\tilde{\mathbf{Q}}'^{\top}, \end{aligned} \tag{33}$$

where $\mathbf{\Xi}^{(0)} = \mathbf{W}\tilde{\mathbf{\Lambda}}\mathbf{W}^{-1}$ is an arbitrary eigenvalue decomposition of $\mathbf{\Xi}^{(0)}$, and

$$\tilde{\mathbf{Q}} = [\tilde{q}_{ij}] = \begin{bmatrix} \boldsymbol{\omega}\mathbf{\Xi}^{(1)} \\ \vdots \\ \boldsymbol{\omega}\mathbf{\Xi}^{(N)} \end{bmatrix} \mathbf{W}, \tag{34}$$

$$\tilde{\mathbf{Q}}' = [\tilde{q}'_{ij}] = \begin{bmatrix} \boldsymbol{\sigma}^{\top}\mathbf{\Xi}^{(1)\top} \\ \vdots \\ \boldsymbol{\sigma}^{\top}\mathbf{\Xi}^{(N)\top} \end{bmatrix} \left(\mathbf{\Xi}^{(0)}\right)^{-\top}\mathbf{W}^{-\top}. \tag{35}$$

Note the form of (33) is similar with that of PMM equation (10). According to the equivalence theorem of OOMs,[29] it can be proved that the $(\tilde{\mathbf{\Lambda}}, \tilde{\mathbf{Q}}, \tilde{\mathbf{Q}}')$ in (33) are related to the PMM quantities $(\mathbf{\Lambda}, \mathbf{Q})$ as

$$\mathbf{\Lambda} = \tilde{\mathbf{\Lambda}}, \tag{36}$$

$$\mathbf{Q} = [q_{ij}] = [\frac{1}{2}\left(\operatorname{sgn}\left(\tilde{q}_{ij}\right) + \operatorname{sgn}\left(\tilde{q}'_{ij}\right)\right)\sqrt{\tilde{q}_{ij}\tilde{q}'_{ij}}], \tag{37}$$

where diagonal elements of $\mathbf{\Lambda}$ and $\tilde{\mathbf{\Lambda}}$ are both sorted in the descending order, and $\operatorname{sgn}(\cdot)$ denotes the sign function with $\operatorname{sgn}(a) = 1$ for $a \geq 0$ and $\operatorname{sgn}(a) = -1$ for $a < 0$. (See Appendix C for the detailed proof.) The above two equations give a simple way to extract the PMM spectral components from the parameters of $\mathcal{M}$.

Remark 4. It is obvious that we cannot utilize (36) and (37) to compute the projected spectral components if $\hat{\mathbf{\Xi}}^{(0)}$ is not diagonalizable over the real numbers due to the influence of statistical error. That is the reason why it is required to perform the numerical modification of $\mathbf{C}$ and $\mathbf{C}^{(y)}$ as shown in Algorithm 2.

V. APPLICATIONS

A. Two-dimensional diffusion process

In order to demonstrate the usefulness of OOMs to estimate the kinetics of metastable systems, let us first study an example of two-dimensional Brownian dynamics on the domain $\Omega = [-2, 2] \times [-1.5, 2.5]$ with a three-well potential and a simple box discretization observation model, as shown in Fig. 1(a) (see Appendix D for details). The six largest eigenvalues of the propagator with lag time $\tau = 0.6$ are

λ_1	λ_2	λ_3	λ_4	λ_5	λ_6
1.0000	0.9524	0.6372	0.0956	0.0424	0.0202,

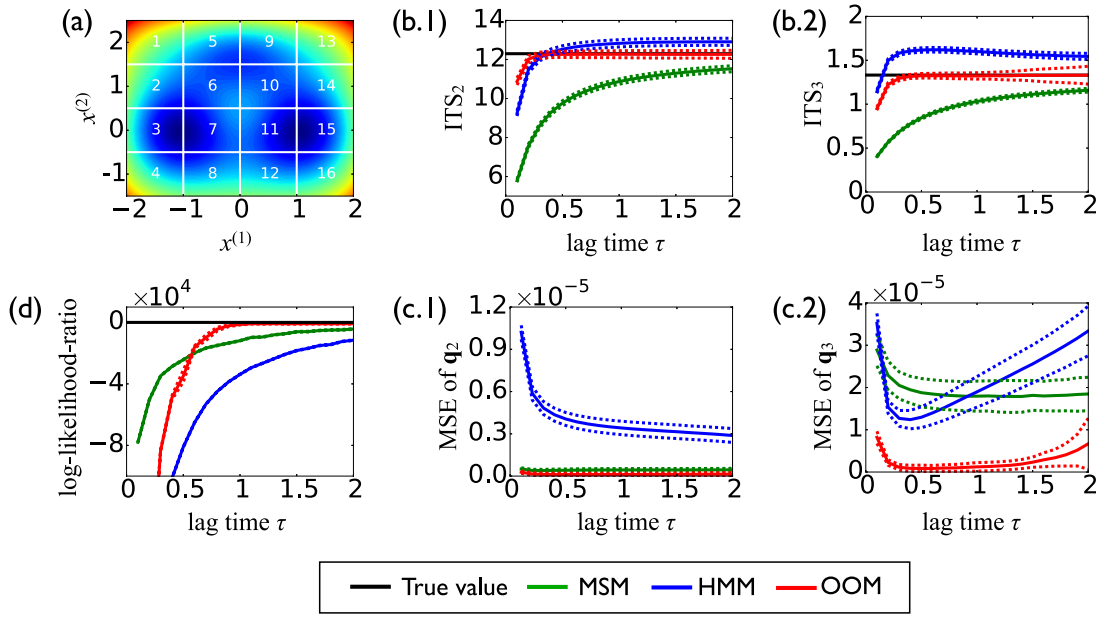

FIG. 1. Comparison of OOM, MSM, and HMM for modeling the diffusion in a three-well potential. (a) Illustration of the potential function and observation model, where each grid represents a finite element space of the observation model. (b) Estimates of implied time scales $\mathrm{ITS}_i = -\tau/\ln\lambda_i(\tau)$ for $i = 2, 3$ ($\mathrm{ITS}_1 \equiv \infty$). (c) Mean square errors between the true second and third projected eigenfunctions and their estimates. (d) Log-likelihood-ratios $\ln p(\mathbf{y}_{\mathrm{test}}(\tau)|\mathrm{estimatedmodel}) - \ln p(\mathbf{y}_{\mathrm{test}}(\tau)|\mathrm{truemodel})$ between estimated models and the true model, where $\mathbf{y}_{\mathrm{test}}(\tau)$ denotes an observation sequence of length $5\times10^6\tau$ and sampling step size τ generated from the true model. Solid lines in ((b)-(d)) represent average over 30 simulations and dash lines show the corresponding one-sigma confidence intervals.

so the dynamics are clearly dominated by three metastable states.

We have performed 30 independent simulations of the system with sampling step size 0.1 and length 10^5 and have built the following three models from the observation data for comparison in each run: a three dimensional OOM, a MSM and a three-state HMM which describes the dynamics of the state evolution as a reversible Markov chain that transitions between three metastable substates.[16] Fig. 1 summarizes the performances of the three models, where the estimation results at $\tau > 0.1$ are obtained by using the subsequences with lag time τ extracted from simulated trajectories. It is obvious that the MSM is actually a finite element approximation of the state evolution. The poor finite element mesh defined by the observation model — a mesh that cannot accurately capture boundaries between the metastable states — leads to slow convergence and large errors in the estimated relaxation time scales. The HMM overcomes this limitation by searching for a suitable metastable subspace partition of the underlying dynamics through the HMM learning process.[16] From Fig. 1(b) it can be seen that the HMM performs significantly better than the MSM on the implied time scale estimation. However, the local dynamics within each metastable subspace is ignored in the HMM, so that the estimates of projected eigenfunctions and observation likelihood obtained by the HMM are even worse than that obtained by the MSM especially when τ is small (see Figs. 1(c) and 1(d)). The OOM outperforms the other two models, yielding accurate and precise time scale and eigenvector estimates and providing the largest likelihood ratio when $\tau \geq 0.6$, because it is close to an "exact" model of the system without model mismatch for a large τ. (All the three models achieve similar estimation results on $\mathbf{q}_1$ as it can easily be estimated as the stationary distribution of $\{y_t\}$.)

B. Bovine pancreatic trypsin inhibitor

We now investigate the conformation dynamics of the bovine pancreatic trypsin inhibitor (BPTI) protein (see Fig. 2(a) for the secondary structure) by using a 1 ms molecular dynamics simulation which was generated by the Anton supercomputer.[31] It was reported in Ref. 16 that this system has three dominant spectral components which are associated with conformational transitions between three metastable states, and the free energy function in the $\mathrm{IC}^{(1)}$-$\mathrm{IC}^{(2)}$ coordinates (see Fig. 2(b)) shows the three metastable states which are centered at about $(\mathrm{IC}^{(1)}, \mathrm{IC}^{(2)}) = (-0.25, 0.7)$, $(-0.3, -1.65)$, and $(6.4, -0.6)$, where $\mathrm{IC}^{(1)}$ and $\mathrm{IC}^{(2)}$ denote the slowest independent components given by the time-lagged independent component analysis.[23,24] We utilize the regular space clustering algorithm[14] to discretize the independent component space into 8 and 87 clusters and then model the corresponding coarse-grained dynamics with OOM, MSM and HMM, where both the dimension of the OOM and the state number of the HMM are set to be three. Fig. 2(c) plots implied time scale estimation results. It can be observed that the MSM fails to give stable estimates of implied time scales with lag time $\tau \leq 1$ μs due to the non-Markovian projected dynamics. The HMM estimates of ITS_2 and ITS_3 converge to τ-independent constants about 45 μs and 22 μs at $\tau = 0.2$ μs with 87 clusters. But for the coarser discretization with 8 clusters, the HMM converges to much smaller estimates of 22 μs and 18 μs. In contrast with the MSM and HMM estimates, the OOM estimates quickly converge to around 45 μs and 22 μs for both 87- and 8-cluster partitions.

Moreover, we evaluate the mean square errors between the estimated autocorrelation functions of independent components,

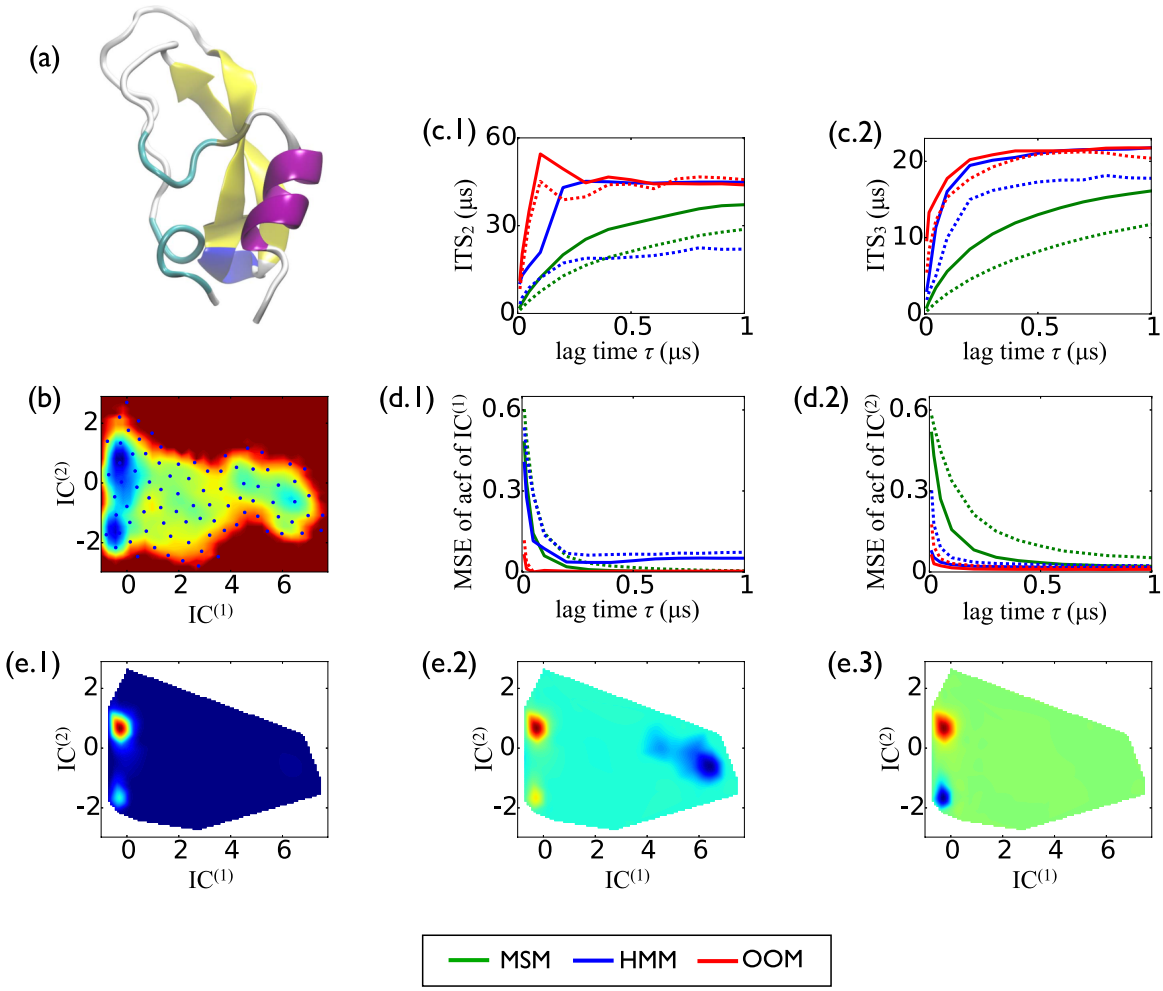

FIG. 2. Comparison of OOM, MSM, and HMM for modeling the conformation dynamics of BPTI. (a) Illustration of the structure of BPTI. (b) Free energy landscape obtained by direct population inversion of the trajectory data (in units of k_BT), where the blue dots show the centers of 87 clusters for coarse-graining. (c) Estimates of implied time scales ITS_2 and ITS_3. (d) Mean square errors of autocorrelation functions $\mathrm{acf}(\mathrm{IC}^{(1)};\Delta t)$ and $\mathrm{acf}(\mathrm{IC}^{(2)};\Delta t)$ obtained by comparing the model prediction with the Monte Carlo estimation. (e) The first three projected eigenvectors for the 87 cluster discretization extracted from the OOM with lag time $\tau = 1$ μs. Dashed and solid lines in ((c) and (d)) represent estimation results given by using 8 clusters and 87 clusters, respectively.

$$\mathrm{acf}\left(\mathrm{IC}^{(i)},\Delta t\right) = \mathbb{E}\left[\mathrm{IC}_t^{(i)}\mathrm{IC}_{t+\Delta t}^{(i)}\right], \tag{38}$$

for $\Delta t \in \{k\tau | 0 < k\tau \leq 10\ \mu\mathrm{s}\}$ given by the dynamical models and direct Monte Carlo estimation. The computation of (38) from models is described in Appendix E. (The time cross correlation of $\mathrm{IC}^{(1)}$ and $\mathrm{IC}^{(2)}$ is close to zero for the time-lagged independence between them.) It can be seen from Fig. 2(d) that the OOM achieves the smallest errors, which implies that the OOMs characterize the conformation dynamics of BPTI in the projected space more accurately than the MSM and HMM.

Fig. 2(e) display the three projected eigenvectors given by the OOM with 87 clusters. It is interesting to see that the sign patterns of the projected eigenvectors clearly indicates the metastable conformations.

VI. CONCLUSIONS

PMMs provide a general theoretical framework for describing observed (projected) Markov processes, which are especially relevant for complex dynamical systems where the optimal reaction coordinate(s) are either unknown in simulations or inaccessible in experiments. When the dynamics are metastable, the PMM can be specified by a few spectral components. In this paper, we show that OOMs are a powerful modeling tool for PMM systems. In contrast to the widely used MSMs, the assumption of the Markovianity in the observation space is not required for OOMs, and the PMM dynamics on the observation space can be exactly modeled by OOMs for any choice of the coarse graining operation. Furthermore, although OOMs are non-Markovian, they can also be asymptotically correctly estimated with very low computational complexity as do MSMs. Therefore, OOMs appear to be an efficient and effective alternative to MSMs for modeling and analyzing metastable systems. We are currently developing a new estimation code package for OOMs and related methods and this code will be released as a subpackage of PyEMMA (http://www.pyemma.org).

The focus of our future research includes OOM estimation methods for nonequilibrium and continuous-observation PMMs and exploration of other possible applications of OOMs to metastable systems such as analysis of reaction trajectories and convoluted time-series.

ACKNOWLEDGMENTS

We thank M. Thon (Jacobs University Bremen) for valuable scientific discussions on OOM theory and methods. We acknowledge funding from Nos. DFG WU 744/1-1, DFG SFB 1114 (Wu), and ERC Grant "pcCell" (Noé).

APPENDIX A: PROOF OF Eq. (15)

According to (14) and (7), we have

$$\begin{aligned} &p(x_0, x_\tau, y_\tau, \ldots, x_{k\tau}, y_{k\tau}) \\ &\quad = p_0(x_0) \prod_{l=1}^{k} p(x_{l\tau}|x_{(l-1)\tau}) \Pr(y_{l\tau}|x_{l\tau}) \\ &\quad = p_0(x_0) \prod_{l=1}^{k} \frac{\chi_{y_{l\tau}}(x_{l\tau})}{\mu(x_{(l-1)\tau})} \boldsymbol{\phi}(x_{(l-1)\tau})^\top \boldsymbol{\Lambda} \boldsymbol{\phi}(x_{l\tau}) \\ &\quad = \frac{p_0(x_0)}{\mu(x_0)} \boldsymbol{\phi}(x_0)^\top \prod_{l=1}^{k-1} \frac{\chi_{y_{l\tau}}(x_{l\tau})}{\mu(x_{l\tau})} \boldsymbol{\Lambda} \boldsymbol{\phi}(x_{l\tau}) \boldsymbol{\phi}(x_{l\tau})^\top \\ &\qquad \cdot \chi_{y_{k\tau}}(x_{k\tau}) \boldsymbol{\Lambda} \boldsymbol{\phi}(x_{k\tau}). \end{aligned} \tag{A1}$$

Noticing that $\phi_1(x) = \mu(x)$, we can rewrite the last term on the right-hand side of the above equation as

$$\chi_{y_{k\tau}}(x_{k\tau}) \boldsymbol{\Lambda} \boldsymbol{\phi}(x_{k\tau}) = \frac{\chi_{y_{k\tau}}(x_{k\tau})}{\mu(x_{k\tau})} \boldsymbol{\Lambda} \boldsymbol{\phi}(x_{k\tau}) \boldsymbol{\phi}(x_{k\tau})^\top \boldsymbol{\sigma}_P, \tag{A2}$$

with $\boldsymbol{\sigma}_P = (1, 0, \ldots, 0)^\top$. Substituting (A2) into (A1), we get

$$\begin{aligned} &p(x_0, x_\tau, y_\tau, \ldots, x_{k\tau}, y_{k\tau}) \\ &\quad = \frac{p_0(x_0)}{\mu(x_0)} \boldsymbol{\phi}(x_0)^\top \\ &\qquad \cdot \prod_{l=1}^{k} \frac{\chi_{y_{l\tau}}(x_{l\tau})}{\mu(x_{l\tau})} \boldsymbol{\Lambda} \boldsymbol{\phi}(x_{l\tau}) \boldsymbol{\phi}(x_{l\tau})^\top \boldsymbol{\sigma}_P. \end{aligned} \tag{A3}$$

Note that

$$\begin{aligned} \int_\Omega \mathrm{d}x \, \frac{p_0(x)}{\mu(x)} \boldsymbol{\phi}(x)^\top &= (\langle p_0, \phi_1 \rangle, \ldots, \langle p_0, \phi_m \rangle) \\ &= \boldsymbol{\omega}_P \end{aligned} \tag{A4}$$

and

$$\begin{aligned} &\int_\Omega \mathrm{d}x \, \frac{\chi_y(x)}{\mu(x)} \boldsymbol{\Lambda} \boldsymbol{\phi}(x) \boldsymbol{\phi}(x)^\top \\ &\quad = \boldsymbol{\Lambda} \int_\Omega \mathrm{d}x \, \frac{\chi_y(x)}{\mu(x)} \boldsymbol{\phi}(x) \boldsymbol{\phi}(x)^\top \\ &\quad = \boldsymbol{\Lambda} \begin{bmatrix} \langle \chi_y \cdot \phi_1, \phi_1 \rangle & \cdots & \langle \chi_y \cdot \phi_1, \phi_m \rangle \\ \vdots & \ddots & \vdots \\ \langle \chi_y \cdot \phi_m, \phi_1 \rangle & \cdots & \langle \chi_y \cdot \phi_m, \phi_m \rangle \end{bmatrix} \\ &\quad = \boldsymbol{\Xi}_P^{(y)}, \end{aligned} \tag{A5}$$

where $\langle \cdot, \cdot \rangle$ denotes the weighted inner product defined by (2). Thus, integrating over $x_0, \ldots, x_{k\tau}$ on both the left- and right-hand sides of (A3) leads to

$$\begin{aligned} &\Pr(y_\tau, \ldots, y_{k\tau}) \\ &\quad = \left(\int_\Omega \mathrm{d}x_0 \, \frac{p_0(x_0)}{\mu(x_0)} \boldsymbol{\phi}(x_0)^\top \right) \\ &\qquad \cdot \prod_{l=1}^{k} \left(\int_\Omega \mathrm{d}x_{l\tau} \, \frac{\chi_{y_{l\tau}}(x_{l\tau})}{\mu(x_{l\tau})} \boldsymbol{\Lambda} \boldsymbol{\phi}(x_{l\tau}) \boldsymbol{\phi}(x_{l\tau})^\top \right) \\ &\qquad \cdot \boldsymbol{\sigma}_P \\ &\quad = \boldsymbol{\omega}_P \boldsymbol{\Xi}_P^{(y_\tau)} \ldots \boldsymbol{\Xi}_P^{(y_{k\tau})} \boldsymbol{\sigma}_P. \end{aligned} \tag{A6}$$

APPENDIX B: DIAGONALIZABILITY OF $\hat{\boldsymbol{\Xi}}^{(0)}$ IN ALGORITHM 1

In this appendix, we show that the sum of all estimated observable operators given by Algorithm 1 is diagonalizable in the real sense if $\hat{\boldsymbol{\pi}}$, $\hat{\mathbf{C}}$, and $\hat{\mathbf{C}}^{(y)}$ satisfy the constraints in (22) and

$$\operatorname{rank}\left(\hat{\mathbf{C}}\right) \geq m, \tag{B1}$$

$$\hat{\boldsymbol{\pi}} > 0, \tag{B2}$$

where $\hat{\boldsymbol{\pi}} = [\hat{\pi}_i] > 0$ means all elements of $\hat{\boldsymbol{\pi}}$. Note that if $\operatorname{rank}\left(\hat{\mathbf{C}}\right) < m$, we can reduce the value of m to $\operatorname{rank}\left(\hat{\mathbf{C}}\right)$ according to the analysis in Remark 2, and if $\hat{\pi}_y = 0$ for some observation value $y \in O$, we can simply leave y out of the observation set O since it does not appear in data. So it is reasonable to assume that (B1) and (B2) are always satisfied in applications.

Since $\mathbf{S} = \operatorname{diag}(\hat{\boldsymbol{\pi}})^{-\frac{1}{2}} \hat{\mathbf{C}} \operatorname{diag}(\hat{\boldsymbol{\pi}})^{-\frac{1}{2}}$, $\mathbf{S}$ is also a positive semidefinite matrix with $\operatorname{rank}(\mathbf{S}) \geq m$. Therefore the m largest eigenvalue of $\mathbf{S}$ are all positive and the diagonal matrix $\boldsymbol{\Sigma}_m$ obtained in line 3 of the algorithm is positive definite. Considering that $\hat{\boldsymbol{\Xi}}^{(0)}$ can be written as

$$\begin{aligned} \hat{\boldsymbol{\Xi}}^{(0)} &= \sum_{y \in O} \hat{\boldsymbol{\Xi}}^{(y)} \\ &= \boldsymbol{\Sigma}_m^{-1} \cdot \left(\sum_{y \in O} \mathbf{V}^\top \hat{\mathbf{C}}^{(y)} \mathbf{V} \right), \end{aligned} \tag{B3}$$

where the first term on right hand side is a positive definite matrix and the second term is a symmetric matrix, we can conclude that $\hat{\boldsymbol{\Xi}}^{(0)}$ is diagonalizable over the real numbers according to Theorem 12.19 in Ref. 32.

APPENDIX C: PROOF OF Eqs. (36) and (37)

We first show that the m-dimensional OOM $\mathcal{M}$ is a minimal OOM, i.e., there exists no equivalent model of lower dimension. According to the assumption of $\operatorname{rank}(\mathbf{Q}) = m$ and Eq. (32), we have $\operatorname{rank}([\boldsymbol{\Xi}^{(1)\top}\boldsymbol{\omega}^\top \quad \cdots \quad \boldsymbol{\Xi}^{(N)\top}\boldsymbol{\omega}^\top]) = \operatorname{rank}([\boldsymbol{\Xi}^{(1)}\boldsymbol{\sigma} \quad \cdots \quad \boldsymbol{\Xi}^{(N)}\boldsymbol{\sigma}]) = \operatorname{rank}(\mathbf{C}) = m$. Then $\mathcal{M}$ is trimmed and therefore minimal (see Definition 5 and Corollary 8 in Ref. 29).

We now show Eq. (37). Let $\mathcal{M}_P = (\boldsymbol{\omega}_P, \{\boldsymbol{\Xi}_P^{(y)}\}_{y \in O}, \boldsymbol{\sigma}_P)$ denote the m-dimensional OOM constructed in Subsection III B and note that $\mathcal{M}$ and $\mathcal{M}_P$ are both exact m-dimensional models of the observable dynamics of the PMM and therefore equivalent. We can then conclude from the minimality of $\mathcal{M}$ and the OOM equivalence theorem (see Proposition 12 in

Ref. 29) that there is a matrix $\mathbf{R} \in \mathbb{R}^{m\times m}$ such that

$$\boldsymbol{\omega} = \boldsymbol{\omega}_P \mathbf{R}^{-1}, \tag{C1}$$

$$\boldsymbol{\Xi}^{(y)} = \mathbf{R}\boldsymbol{\Xi}_P^{(y)}\mathbf{R}^{-1}, \quad \forall k \in O, \tag{C2}$$

$$\boldsymbol{\sigma} = \mathbf{R}\boldsymbol{\sigma}_P, \tag{C3}$$

and $\boldsymbol{\Xi}^{(0)}$ can be written as

$$\boldsymbol{\Xi}^{(0)} = \mathbf{R}\boldsymbol{\Xi}_P^{(0)}\mathbf{R}^{-1} = \mathbf{R}\boldsymbol{\Lambda}\mathbf{R}^{-1}. \tag{C4}$$

This means for an arbitrary eigenvalue decomposition $\boldsymbol{\Xi}^{(0)} = \mathbf{W}\tilde{\boldsymbol{\Lambda}}\mathbf{W}^{-1}$ of $\boldsymbol{\Xi}^{(0)}$, we have

$$\tilde{\boldsymbol{\Lambda}} = \boldsymbol{\Lambda} \tag{C5}$$

and

$$\mathbf{W} = \mathbf{R}\boldsymbol{\Gamma}, \tag{C6}$$

if diagonal elements of $\boldsymbol{\Lambda}$ and $\tilde{\boldsymbol{\Lambda}}$ are both sorted in the descending order, where $\boldsymbol{\Gamma} = \mathrm{diag}\,(\gamma_1, \ldots, \gamma_m)$ is a full-rank diagonal matrix.

Next, we show Eq. (37). Using (C1)–(C3), (C5), and (C6). Note that

$$\begin{aligned}\boldsymbol{\omega}_P &= (\langle p_0, \phi_1\rangle, \ldots, \langle p_0, \phi_m\rangle)\\ &= (1, 0, \ldots, 0) = \boldsymbol{\sigma}_P^{\top}\end{aligned} \tag{C7}$$

due to the stationarity. We can write the nth row of $\tilde{\mathbf{Q}}$ and $\tilde{\mathbf{Q}}'$ as

$$\begin{aligned}\boldsymbol{\omega}\boldsymbol{\Xi}^{(n)}\mathbf{W} &= \boldsymbol{\omega}_P\boldsymbol{\Xi}_P^{(n)}\boldsymbol{\Gamma},\\ &= (q_{n1}, \ldots, q_{nm})\,\boldsymbol{\Gamma}\end{aligned} \tag{C8}$$

$$\begin{aligned}\boldsymbol{\sigma}^{\top}\boldsymbol{\Xi}^{(n)\top}\left(\boldsymbol{\Xi}^{(0)}\right)^{-\top}\mathbf{W}^{-\top} &= \boldsymbol{\sigma}_P^{\top}\boldsymbol{\Xi}_P^{(n)\top}\boldsymbol{\Lambda}^{-1}\boldsymbol{\Gamma}^{-1}\\ &= (q_{n1}, \ldots, q_{nm})\,\boldsymbol{\Lambda}\boldsymbol{\Lambda}^{-1}\boldsymbol{\Gamma}^{-1}\\ &= (q_{n1}, \ldots, q_{nm})\,\boldsymbol{\Gamma}^{-1},\end{aligned} \tag{C9}$$

which imply that

$$\tilde{\mathbf{Q}} = \mathbf{Q}\boldsymbol{\Gamma} \tag{C10}$$

$$\tilde{\mathbf{Q}}' = \mathbf{Q}\boldsymbol{\Gamma}^{-1}. \tag{C11}$$

Then we have

$$q_{ij}^2 = \tilde{q}_{ij}\tilde{q}'_{ij}, \tag{C12}$$

$$\mathrm{sgn}\left(q_{ij}\right) = \mathrm{sgn}\left(\gamma_j\right)\mathrm{sgn}\left(\tilde{q}_{ij}\right) = \mathrm{sgn}\left(\gamma_j\right)\mathrm{sgn}\left(\tilde{q}'_{ij}\right). \tag{C13}$$

Combining the above two equations yields

$$q_{ij} = \frac{\mathrm{sgn}\left(\gamma_j\right)}{2}\left(\mathrm{sgn}\left(\tilde{q}_{ij}\right) + \mathrm{sgn}\left(\tilde{q}'_{ij}\right)\right)\sqrt{\tilde{q}_{ij}\tilde{q}'_{ij}}. \tag{C14}$$

Note that for the PMM, both $\mathbf{q}_j = (q_{1j}, \ldots, q_{Nj})^{\top}$ and $(-\mathbf{q}_j)$ can be viewed as the jth projected eigenfunction, where $(-\mathbf{q}_j)$ is equal to the projection of the eigenfunction $(-\phi_i)$ on the observation space. Hence we can neglect the term $\mathrm{sgn}\left(\gamma_j\right)$ in (C14) and get

$$q_{ij} = \frac{1}{2}\left(\mathrm{sgn}\left(\tilde{q}_{ij}\right) + \mathrm{sgn}\left(\tilde{q}'_{ij}\right)\right)\sqrt{\tilde{q}_{ij}\tilde{q}'_{ij}}. \tag{C15}$$

APPENDIX D: DIFFUSION PROCESS MODEL DESCRIPTION AND SIMULATION

The diffusion process model used for the numerical study in Subsection V A is driven by the following stochastic differential equation:

$$\mathrm{d}x_t = -\nabla V\left(x_t\right)\mathrm{d}t + \sqrt{2\beta^{-1}}\mathrm{d}W_t, \tag{D1}$$

with $x_t = \left(x_t^{(1)}, x_t^{(2)}\right) \in \Omega = [-2, 2] \times [-1.5, 2.5]$, and the observation model is defined by

$$y_t\left(x_t\right) = k, \quad \text{if } x_t \in \text{ the } i\text{-th finite element space} \tag{D2}$$

with observation space $O = \{1, \ldots, 16\}$, where W_t denotes standard Brownian motion, $\beta = 1.67$ denotes inverse temperature,

$$\begin{aligned}V\left(x^{(1)}, x^{(2)}\right) = {} & 3\exp\left(-\left(x^{(1)}\right)^2 - \left(x^{(2)} - \frac{1}{3}\right)^2\right)\\ & - 3\exp\left(-\left(x^{(1)}\right)^2 - \left(x^{(2)} - \frac{5}{3}\right)^2\right)\\ & - 5\exp\left(-\left(x^{(1)} - 1\right)^2 - \left(x^{(2)}\right)^2\right)\\ & - 5\exp\left(-\left(x^{(1)} + 1\right)^2 - \left(x^{(2)}\right)^2\right)\\ & + \frac{1}{5}\left(x^{(1)}\right)^4 + \frac{1}{5}\left(x^{(2)} - \frac{1}{3}\right)^4,\end{aligned} \tag{D3}$$

and the finite elements for observation are obtained by the uniform rectangular partition of Ω.

For convenience of simulation and analysis, here we utilize a reversibility preserving numerical discretization scheme proposed in Ref. 33 with spatial mesh size 0.2×0.2 to generate the simulation trajectories and calculate the eigenvalues and eigenfunctions of the system.

APPENDIX E: CALCULATION OF AUTOCORRELATION FUNCTIONS

For a MSM, HMM, or OOM of the molecular kinetics of the BPTI protein investigated in Subsection V B, if we can assume that the simulation within each cluster achieves local equilibrium, then the autocorrelation function of $\mathrm{IC}^{(k)}$ can be calculated by

$$\begin{aligned}\mathrm{acf}\left(\mathrm{IC}^{(k)}, \Delta t\right) &= \mathbb{E}\left[\mathrm{IC}_t^{(k)}\mathrm{IC}_{t+\Delta}^{(k)}\right]\\ &= \sum_{i,j}\mathbb{E}\left[\mathrm{IC}_t^{(k)}|\, y_t = i\right]\mathbb{E}\left[\mathrm{IC}_t^{(k)}|\, y_t = j\right]\\ &\quad \cdot \Pr\left(y_t = i, y_{t+\Delta t} = j\right),\end{aligned} \tag{E1}$$

where $\mathbb{E}\left[\mathrm{IC}_t^{(k)}|\, y_t = i\right]$ represents the mean value of $\mathrm{IC}^{(k)}$ of the ith cluster and can be simply estimated by the empirical mean of simulation data, and the joint distribution $\Pr\left(y_t, y_{t+\Delta t}\right)$ can be computed according to the model parameters.

[1]H. Mori, Prog. Theor. Chem. Phys. **33**, 423 (1965).
[2]R. Zwanzig, J. Stat. Phys. **9**, 215 (1973).
[3]O. F. Lange and H. Grubmüller, J. Chem. Phys. **124**, 214903 (2006).
[4]R. Hegger and G. Stock, J. Chem. Phys. **130**, 034106 (2009).

[5]C. Schütte, A. Fischer, W. Huisinga, and P. Deuflhard, J. Comput. Phys. **151**, 146 (1999).
[6]W. C. Swope, J. W. Pitera, and F. Suits, J. Phys. Chem. B **108**, 6571 (2004).
[7]N. Singhal and V. S. Pande, J. Chem. Phys. **123**, 204909 (2005).
[8]J. D. Chodera, K. A. Dill, N. Singhal, V. S. Pande, W. C. Swope, and J. W. Pitera, J. Chem. Phys. **126**, 155101 (2007).
[9]F. Noé, I. Horenko, C. Schütte, and J. C. Smith, J. Chem. Phys. **126**, 155102 (2007).
[10]N. V. Buchete and G. Hummer, J. Phys. Chem. B **112**, 6057 (2008).
[11]A. C. Pan and B. Roux, J. Chem. Phys. **129**, 064107 (2008).
[12]F. Noé, C. Schütte, E. Vanden-Eijnden, L. Reich, and T. R. Weikl, Proc. Natl. Acad. Sci. U. S. A. **106**, 19011 (2009).
[13]M. Sarich, F. Noé, and C. Schütte, SIAM Multiscale Model. Simul. **8**, 1154 (2010).
[14]J.-H. Prinz, H. Wu, M. Sarich, B. Keller, M. Senne, M. Held, J. D. Chodera, C. Schütte, and F. Noé, J. Chem. Phys. **134**, 174105 (2011).
[15]C. Schütte, F. Noé, J. Lu, M. Sarich, and E. Vanden-Eijnden, J. Chem. Phys. **134**, 204105 (2011).
[16]F. Noé, H. Wu, J.-H. Prinz, and N. Plattner, J. Chem. Phys. **139**, 184114 (2013).
[17]J.-H. Prinz, J. D. Chodera, and F. Noé, Phys. Rev. X **4**, 011020 (2014).
[18]H. Jaeger, "Discrete-time, discrete-valued observable operator models: A tutorial," Technical Report GMD-42, German National Research Center for Information Technology (GMD), 1998.
[19]*An Introduction to Markov State Models and Their Application to Long Timescale Molecular Simulation*, Advances in Experimental Medicine and Biology, edited by G. R. Bowman, V. S. Pande, and F. Noé (Springer, Heidelberg, 2014), Vol. 797.
[20]C. Schütte and W. Huisinga, *Handbook of Numerical Analysis* (Elsevier, 2003), pp. 699–744.
[21]A. Bovier, V. Gayrard, and M. Klein, J. Eur. Math. Soc. **7**, 69 (2005).
[22]G. Perez-Hernandez, F. Paul, T. Giorgino, G. de Fabritiis, and F. Noé, J. Chem. Phys. **139**, 015102 (2013).
[23]C. R. Schwantes and V. S. Pande, J. Chem. Theory Comput. **9**, 2000 (2013).
[24]F. Noé and F. Nüske, SIAM Multiscale Model. Simul. **11**, 635 (2013).
[25]F. Nüske, B. Keller, G. Pérez-Hernández, A. S. J. S. Mey, and F. Noé, J. Chem. Theory Comput. **10**, 1739 (2014).
[26]R. T. McGibbon, B. Ramsundar, M. M. Sultan, G. Kiss, and V. S. Pande, Proc. Int. Conf. Mach. Learn. **31**, 1197–1205 (2014).
[27]M. Weber, "Meshless methods in conformation dynamics," Ph.D. thesis, Verlag, 2006.
[28]F. Noé, H. Wu, J.-H. Prinz, and N. Plattner, J. Chem. Phys. **139**, 184114 (2013).
[29]M. Thon and H. Jaeger, J. Mach. Learn. Res. **16**, 103 (2015).
[30]H. Jaeger, Neural Comput. **12**, 1371 (2000).
[31]D. E. Shaw, P. Maragakis, K. Lindorff-Larsen, S. Piana, R. Dror, M. Eastwood, J. Bank, J. Jumper, J. Salmon, Y. Shan, and W. Wriggers, Science **330**, 341 (2010).
[32]A. J. Laub, *Matrix Analysis for Scientists and Engineers* (SIAM, 2005).
[33]J. C. Latorre, P. Metzner, C. Hartmann, and C. Schütte, Commun. Math. Sci. **9**, 1051 (2011).