Gaussian Markov transition models of molecular kinetics

Hao Wu[a] and Frank Noé[b]
Department for Mathematics and Computer Science, FU Berlin, Arnimallee 6, 14195 Berlin, Germany

(Received 21 November 2014; accepted 6 February 2015; published online 24 February 2015)

The slow processes of molecular dynamics (MD) simulations—governed by dominant eigenvalues and eigenfunctions of MD propagators—contain essential information on structures of and transition rates between long-lived conformations. Existing approaches to this problem, including Markov state models and the variational approach, represent the dominant eigenfunctions as linear combinations of a set of basis functions. However the choice of the basis functions and their systematic statistical estimation are unsolved problems. Here, we propose a new class of kinetic models called Markov transition models (MTMs) that approximate the transition density of the MD propagator by a mixture of probability densities. Specifically, we use Gaussian MTMs where a Gaussian mixture model is used to approximate the symmetrized transition density. This approach allows for a direct computation of spectral components. In contrast with the other Galerkin-type approximations, our approach can automatically adjust the involved Gaussian basis functions and handle the statistical uncertainties in a Bayesian framework. We demonstrate by some simulation examples the effectiveness and accuracy of the proposed approach. © *2015 AIP Publishing LLC.* [http://dx.doi.org/10.1063/1.4913214]

I. INTRODUCTION

With increasing computational power, molecular dynamics (MD) simulation has now become one of the most important computational tools for simulating and investigating biomolecular systems.[1] MD simulation can provide atomic-level insight into biophysical processes such as protein folding, protein inhibition by ligands, and protein aggregation. It allows researchers to resolve the relationship between the function of a biomolecule and the underlying conformational transitions. However, understanding and modeling the conformational dynamics on large timescales from MD simulation data are generally a challenging task because many structural changes are governed by rare events. An effective mathematical approach is to approximate the dominant spectral components (i.e., the largest eigenvalues and the associated eigenfunctions) of the Markov propagator defined by the MD simulation algorithm. Based on the spectral components, we can decompose the conformational transition process on large timescales into a small number of slow relaxation processes. A lot of dynamical information can be extracted from the slow processes, e.g., for the computation of ensemble averages and correlation functions,[2] detection of spatial structures of metastable states,[3] choice of reaction coordinates,[4,5] and construction of low-dimensional approximate models.[6]

The most popular and successful method for the spectral estimation is the Markov state model (MSM) method.[7–12] A MSM consists of a discretization of the state space into a set of discrete bins and an estimation of a Markovian transition matrix describing the dynamics between them. Given a MSM, one can obtain a direct estimate of the dominant spectral components (eigenvalues and eigenvectors) by conducting an eigenvalue decomposition on the transition matrix. It is easy to see that the main difficulty of this method comes from the choice of the discretization, and a "bad" discretization of the state space will severely violate the Markov assumption and lead to a poor estimation.[13,7] Roughly speaking, the discretization in the MSM method can be improved in two ways. The first way is to increase the number of bins by using, e.g., subdivision[14,15] and cell-mapping[16,17] techniques, so that the state space can be finely discretized with small truncation error. But this way suffers from the "curse of dimensionality" when applied to macromolecules. The second way is to modify the shapes and locations of bins, and the most commonly used optimality criterion for constructing state space discretization of MSMs is to maximize metastability, which can be done by discovering long lived conformations[10] or free energy basins[18–20] of molecular systems. Recent theoretical investigations[7,13,21] have shown that the maximization of metastability is not exactly consistent with the minimization of dynamical approximation error, and the discretization of transition subspaces between metastable states is also important to increase the approximation accuracy. However, there is currently no systematic algorithm to perform the discretization by directly minimizing the dynamical approximation error of MSMs. Furthermore, the quality of MSMs can also be improved by performing the discretization in generalized state spaces defined by time-lagged independent components[5,22] or state sequences.[23–25]

Another spectral estimation method, called variational method,[26,27] was developed recently. It approximates eigenfunctions by linear combinations of a set of basis functions and utilizes the Ritz method (or generalized Ritz method) to achieve the best combinations and estimation in the sense of the variational principle. It can be proved that the MSM method is in fact a specific version of the variational method with basis functions being step functions.[26] Therefore,

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

compared to the MSM method, the variational method provides a more flexible and general framework for the spectral estimation and usually can get accurate estimates of eigenpairs with a small number of smooth basis functions. The major disadvantages of the variation method are (i) the choice of basis functions is still a difficult problem and (ii) it is difficult to analyze the influence of statistical noise.

In the present paper, we propose a new class of models to solve the spectral estimation problem of MD analysis, called Markov transition models (MTMs). MTMs use a parametric model in order to approximate the continuous transition density of the MD operator. Specifically, we propose the use of Gaussian mixtures, leading to Gaussian MTMs, or briefly GMTMs. After estimating the GMTM, we can easily extract the dominant spectral components from the model parameters. The GMTM method provides a flexible way for spectral estimation without discretization as the MSM method. Further, all the parameters in the GMTM, including the means and covariance matrices of Gaussian basis functions, can be adjusted based on the likelihood function, so that the uncertainties caused by statistical noise can be evaluated and handled in a Bayesian manner.

II. SPECTRAL EXPANSION OF MOLECULAR KINETICS AND SYMMETRIZED PROPAGATORS

Let us consider a molecular system with Hamiltonian H at thermal equilibrium at a constant temperature T. We suppose that a standard molecular dynamics simulation is implemented such that the temporal evolution of the molecular configuration can be viewed as a time-homogeneous, ergodic, and reversible Markov process which has a unique stationary distribution,

$$\pi(x) \propto \exp\left(-\frac{H(x)}{k_B T}\right), \tag{1}$$

where k_B denotes the Boltzmann constant. We assume that $|H(x)| < \infty$ and thus $\pi(x) > 0$ for all $x \in \mathbb{R}^d$.

We describe the state of the system by the weighted density $u(x) = \rho(x)/\sqrt{\pi(x)}$, where $\rho(x)$ is the probability density that the molecule is at configuration x. In particular, the weighted stationary density is given by $\sqrt{\pi(x)}$, i.e., the actual probability density is obtained by taking the square of our density function. In an analogy to quantum mechanics, our functions $u(x)$ take the role as wave functions.

The Markovian molecular dynamics has a Markov propagator with conjugate propagator $\mathcal{S}_\tau$,

$$(\mathcal{S}_\tau u)(y) \triangleq \int s_\tau(x,y)\, u(x)\, \mathrm{d}x \tag{2}$$

with the symmetric kernel

$$s_\tau(x,y) = \sqrt{\frac{\pi(x)}{\pi(y)}}\, p(x_{t+\tau} = y | x_t = x), \tag{3}$$

$\mathcal{S}$ is a compact operator on the Hilbert space $\mathcal{L}^2 = \{f | \langle f, f\rangle < \infty\}$ with the inner product $\langle f, g\rangle = \int f(x)\, g(x)\, \mathrm{d}x$. Using this formalism, we can write the propagation of densities by the following spectral expansion:

$$u_{t+\tau}(x) = \sum_{i=1}^{\infty} \lambda_i(\tau) \langle u_t, \phi_i\rangle \phi_i(x), \tag{4}$$

where

$$\lambda_i(\tau) = \exp(-\kappa_i \tau) = \exp\left(-\frac{\tau}{t_i}\right) \tag{5}$$

is the i-th largest eigenvalue of $\mathcal{S}_\tau$ with decay rate κ_i, $t_i = 1/\kappa_i$ is the corresponding implied timescale, ϕ_i denotes the corresponding eigenfunction which satisfies $\langle \phi_i, \phi_j\rangle = \delta_{ij}$, and $(\lambda_i(\tau), \phi_i)$ is called the i-th eigenpair of $\mathcal{S}_\tau$. Note that the first eigenpair satisfies $\lambda_1(\tau) \equiv 1 > \lambda_2(\tau)$ and $\phi_1 = \sqrt{\pi}$ due to the uniqueness of the stationary distribution. Reference 7 contains an illustration of the role of eigenvalues and eigenfunctions in molecular kinetics.

Equation (4) shows that the molecular process can be decomposed into a set of independent relaxation processes by using eigenpairs of the Markov propagator. If the molecular kinetics exhibits metastability, there may be a large spectral gap between a few slow processes and the other ones, i.e., $\kappa_n \ll \kappa_{n+1}$ for some small n. In this case, the probability density of $x_{t+\tau}$ can be well approximated by

$$u_{t+\tau}(x) \approx \sum_{i=1}^{n} \lambda_i(\tau) \langle u_t, \phi_i\rangle \phi_i(x) \tag{6}$$

for $\tau \gg 1/\kappa_n$. The importance of dominant eigenvalues and their eigenfunctions of the Markov propagator is therefore obvious: they describe the main components of the molecular kinetics and give a low-dimensional approximation of the evolution of the configurational distribution on a large timescale.

Equation (6) provides the theoretical basis for a variety of MD data analysis methods, such as Markov state models,[13] diffusion maps,[4] and dynamical fingerprints.[2] However, the direct computation of eigenvalues and eigenfunctions is generally infeasible due to the intractability of the underlying stochastic dynamics. In this paper, we will develop a Bayesian framework for estimating dominant eigenpairs of the molecular kinetics from MD simulation data.

Further, it can be shown that the following properties of $\mathcal{S}_\tau$ and its integral kernel s_τ when the dynamics are reversible and some technical assumptions hold for $\{x_t\}$ (see supplementary material,[28] Sec. B for details):

1. s_τ is a positive and symmetric function, i.e.,
$$s_\tau(x', x) > 0 \text{ and } s_\tau(x', x) = s_\tau(x, x') \tag{7}$$
for all $x, x' \in \mathbb{R}^d$.
2. $\mathcal{S}_\tau$ has the same eigenvalues as the commonly used Markov propagator $\mathcal{P}_\tau$ and transfer operator $\mathcal{T}_\tau$ (see supplementary material,[28] Sec. A for definitions). Its eigenfunctions $\{\phi_i\}$ satisfy
$$\phi_i(x) = l_i(x)/\sqrt{\pi(x)} = r_i(x)\sqrt{\pi(x)}, \tag{8}$$
where l_i is the i-th eigenfunction of $\mathcal{P}_\tau$ and r_i is the i-th eigenfunction of $\mathcal{T}_\tau$, sometimes called *left* and *right* eigenfunctions.[27] Consequently, if solutions of the eigenvalue problem $\mathcal{S}_\tau \phi = \lambda \phi$ are available in closed form,

eigenpairs of the Markov propagator and transfer operator can also be easily obtained.

3. s_τ is square integrable with $\|s_\tau\|_2 < \infty$, which implies that s_τ is a local function with the property

$$\lim_{R\to\infty} \iint_{\|(x',x)\|\geq R} s_\tau^2(x,y)\,\mathrm{d}x\,\mathrm{d}y = 0. \tag{9}$$

Here,

$$\|s_\tau\|_2 = \sqrt{\iint s_\tau^2(x,y)\,\mathrm{d}x\mathrm{d}y} \tag{10}$$

denotes the 2-norm of s_τ and $\|(x',x)\| = \sqrt{\|x'\|^2 + \|x\|^2}$ denotes the Euclidean norm of (x',x).

It can be seen from (7) that $\mathcal{S}_\tau$ is in fact a "symmetric operator" in diffusion maps,[29,30] so we call $\mathcal{S}_\tau$ defined by (2) the *symmetrized propagator* of Markov process $\{x_t\}$, and the integral kernel s_τ given by (3) the symmetrized transition density. It is natural to ask how to check whether a given integral operator describes a valid symmetrized propagator. For this problem, we have the following theorem:

Theorem 1. *Let $\mathcal{S}_\tau$ be a compact integral operator on the Hilbert space $\mathcal{L}^2$ with a continuous integral kernel s_τ. If the spectral radius of $\mathcal{S}_\tau$ is 1 and s_τ satisfies (7), then $\mathcal{S}_\tau$ is a symmetrized propagator of a reversible Markov process.*

Proof. See supplementary material,[28] Sec. C. □

In supplementary Table 1,[28] we compare the three different integral operators: Markov propagator $\mathcal{P}_\tau$, symmetrized propagator $\mathcal{S}_\tau$, and transfer operator $\mathcal{T}_\tau$. It can be observed from the table that all the three operators can equivalently describe the dynamics of a given reversible Markov process, and the eigenpairs of different operators are explicitly related to each other. However, in the three operators, only the symmetrized propagator $\mathcal{S}_\tau$ has a symmetric and square integrable kernel function s_τ, which provides the following advantages for dynamical modeling: (i) s_τ can be approximated by combining multiple local basis functions, (ii) the reversibility of the dynamical model can be simply satisfied by enforcing s_τ to be symmetric. We will exploit these advantages of the symmetric operator in order to construct a model of molecular kinetics.

III. GAUSSIAN MARKOV TRANSITION MODELS

A. Model definition

The properties of symmetrized propagators stated in Sec. II suggest a way of modeling Markovian dynamics on a continuous phase space. For a given reversible process $\{x_t\}$, we can fit some parametric distribution model to the shape of the symmetrized transition density s_τ. Then the dominant spectral components of $\{x_t\}$ can be estimated from the corresponding approximation of the symmetrized propagator $\mathcal{S}_\tau$.

Here, the symmetrized transition density s_τ is fitted by the following Gaussian mixture model (GMM):

$$s_\tau(x,y) = \frac{1}{Z}\sum_{1\leq i,j\leq m} W_{ij}\mathcal{N}(x|\mu_i,\Sigma_i)\,\mathcal{N}(y|\mu_j,\Sigma_j), \tag{11}$$

where $\mathcal{N}(\cdot|\mu,\Sigma)$ denotes a multivariate normal distribution with mean μ and covariance matrix Σ, $\mu_i \in \mathbb{R}^d$ and $\Sigma_i \in \mathbb{R}^{d\times d}$ for all i, $W_{ij} \geq 0$ denotes the weight of the (i,j)-th component with $\sum_{i,j} W_{ij} = 1$. The normalization constant Z can be determined by the fact that the spectral radius of $\mathcal{S}_\tau$ is 1.

Denoting $W = [W_{ij}]$ as the weight matrix and $\chi(x) = [\chi_i(x)]^\top = [\mathcal{N}(x|\mu_i,\Sigma_i)]^\top$ as the column vector of basis functions, (11) can be written more compactly as

$$s_\tau(x,y) = \frac{1}{Z}\chi^\top(x)\,W\chi(y). \tag{12}$$

(Note that W has to be symmetric due to the symmetry of s_τ shown in (7).)

A lot of studies have demonstrated the capacity of Gaussian mixture models to form smooth approximations to arbitrarily shaped densities,[31] and we can prove that proposed model (11) is also general enough to approximate any symmetrized transition density with arbitrary accuracy (see supplementary material,[28] Sec. D). In what follows, we call a reversible Markov process with a GMM-type symmetrized transition density as in (11) an m-th order GMTM since it uses the GMM to describe Markov transition probabilities instead of a simple static data distribution.

B. Eigenpairs of Gaussian transition models

The power of GMTM comes from the fact that their eigenpairs can be calculated from its parameters. Therefore we can estimate the GMTM in a Bayesian approach, as described below, but have direct access to its eigenpairs. According to the third property of symmetrized propagator given in Sec. II, eigenpairs of a GMTM defined by (11) can be obtained from the following eigenvalue problem:

$$\mathcal{S}_\tau\phi = \lambda\phi. \tag{13}$$

Substituting (12) into (13), we have

$$\begin{aligned}\phi(x) &= (\lambda Z)^{-1}\int \chi(x')^\top W\chi(x)\,\phi(x')\,\mathrm{d}x' \\ &= (\lambda Z)^{-1}\left(\int W\chi(x')\,\phi(x')\,\mathrm{d}x'\right)^\top\chi(x).\end{aligned} \tag{14}$$

This implies that for any solution ϕ of (13), there is a coefficient vector $b \in \mathbb{R}^m$ such that

$$\phi = b^\top\chi. \tag{15}$$

Substituting (15) into (13) and defining the inner product matrix $B \in \mathbb{R}^{m\times m}$ by

$$B = \int \chi(x)\,\chi(x)^\top\mathrm{d}x, \tag{16}$$

we arrive at an equivalent form of (13)

$$Z^{-1}b^\top BW = \lambda b^\top \tag{17}$$

which is a simple matrix eigenvalue problem. Furthermore, it can be concluded from (17) that the normalizing constant Z equals the spectral radius of the matrix BW.

Based on the above discussion, the procedure of extracting eigenpairs from a GMTM can be summarized as follows:

1. The coefficient matrix W is obtained by fitting the Gaussian mixture model to the symmetrized transition density (see Sec. IV).
2. The overlap matrix $B = [B_{ij}]$ is computed by

$$B_{ij} = \frac{\mathcal{N}(0|\mu_i, \Sigma_i) \cdot \mathcal{N}(0|\mu_j, \Sigma_j)}{\mathcal{N}(0|\mu_{(i,j)}, \Sigma_{(i,j)})} \tag{18}$$

with

$$\Sigma_{(i,j)} = \left(\Sigma_i^{-1} + \Sigma_j^{-1}\right)^{-1}, \tag{19}$$

$$\mu_{(i,j)} = \Sigma_{(i,j)}\left(\Sigma_i^{-1}\mu_i + \Sigma_j^{-1}\mu_j\right) \tag{20}$$

(see supplementary material,[28] Sec. E for detailed derivations).
3. Compute Z as the spectral radius (largest eigenvalue) of BW.
4. Solve eigenvalue problem (17) and obtain eigenpairs $\{(\lambda_i, \phi_i)\}$ of the symmetrized propagator by using

$$\phi_i = b_i^\top \chi, \tag{21}$$

where (λ_i, b_i) is the i-th solution of (17). Please note that the coefficient vector b_i needs to be normalized according to the constraint

$$\langle \phi_i, \phi_i \rangle = b_i^\top B b_i = 1. \tag{22}$$

IV. ESTIMATION METHODS

A. Likelihood of Gaussian Markov transition models

In order to conduct an estimation of GMTMs, we construct its likelihood to produce a given data set. Using (3), the transition density of a GMTM defined by (11) can be expressed in terms of the symmetrized transition density. We can then obtain the likelihood of GMTM (11) given a simulation trajectory $\{x_{k\tau}\}_{k=0}^K$ as the product of transition densities along the trajectory

$$\begin{aligned} p(\{x_{k\tau}\}|s_\tau) &= \rho_0(x_0)\prod_{k=1}^{K} p(x_{k\tau}|x_{(k-1)\tau}) \\ &= \frac{\rho_0(x_0)}{Z^K}\frac{\phi_1(x_{K\tau})}{\phi_1(x_0)} \\ &\quad \cdot \prod_{k=1}^{K} \chi(x_{(k-1)\tau})^\top W \chi(x_{k\tau}), \end{aligned} \tag{23}$$

where $\rho_0(x)$ is the probability density from which the starting point x_0 is drawn, and $\phi_1(x_0)$ and $\phi_1(x_{K\tau})$ are the stationary densities of these points. Using (23), the expectation-maximization (EM) and Gibbs sampler algorithm can be used like in Gaussian mixture models to get maximum likelihood (ML) and Bayesian estimates of parameters in GMTMs.

Remark 1. In applications, we can simply set the distribution of x_0 to the uniform distribution $\rho_0(x) \propto 1$ if there is no prior information on x_0. Using this prior, the likelihood of the GMTM is equivalent to the transition probability conditional on the starting state with

$$p(\{x_{k\tau}\}_{k\geq 0}|s_\tau) \propto p(\{x_{k\tau}\}_{k\geq 1}|s_\tau, x_0) \tag{24}$$

and the proposed GMTM estimators are then applicable to the case that only a set of short and non-equilibrium simulation trajectories, instead of multiple long trajectories, are available.

Both the maximum-likelihood and Bayesian estimators for GMMs can be efficiently computed by introducing a latent allocation random variable for each observation in the data set.[32] For the proposed GMTM, we can define

$$\mathcal{I} = \{(I_k, J_k)\}_{k=1}^K \tag{25}$$

as the latent variables, which associate $x_{(k-1)\tau}$ and $x_{k\tau}$ to the I_k-th and J_k-th Gaussian components in the basis function vector $\chi(\cdot)$. We can then obtain the conditional dependence relationships between the GMTM, latent variables, and the observed trajectory as

$$\begin{aligned} p(\mathcal{I}|s_\tau) &= \prod_{k=1}^{K} W_{I_k J_k} \\ p(\{x_{k\tau}\}|\mathcal{I}, s_\tau) &= \frac{\rho_0(x_0)}{Z^K}\frac{\phi_1(x_{K\tau})}{\phi_1(x_0)} \\ &\quad \cdot \prod_{k=1}^{K} \mathcal{N}(x_{(k-1)\tau}|\mu_{I_k}, \Sigma_{I_k}) \\ &\quad \cdot \prod_{k=1}^{K} \mathcal{N}(x_{k\tau}|\mu_{J_k}, \Sigma_{J_k}). \end{aligned} \tag{26}$$

In probabilistic model (26), the inference of $\mathcal{I}$ can be simply handled since they are conditionally independent given the value of weight matrix W, and the likelihood $p(\{x_{k\tau}\}|\mathcal{I}, s_\tau)$ of the m-th order GMTM s_τ with fixed latent variables $\mathcal{I}$ can be evaluated with a computational time complexity of only $O(md^2)$ after calculating some sufficient statistics of latent variables (see supplementary material,[28] Sec. F for details). Therefore, in contrast with the original observation model of the GMTM defined by marginal likelihood function (23), probabilistic model (26) is more suitable for statistical inference of GMTMs. (The computation of marginal likelihood function (23) requires time $O(Kd^2)$ for a simulation trajectory with length K.) In what follows, we will investigate the Bayesian and ML estimations of GMTMs based on probabilistic model (26).

B. Maximum likelihood estimation

We now investigate how to search the ML estimate of the GMTM which is determined by the optimization problem

$$\theta^* = \arg\max_\theta \log p(\{x_{k\tau}\}|s_\tau(\theta)) \tag{27}$$

by using the EM algorithm, where θ represents a parameter vector consisting of all independent parameters of means and covariance matrices of Gaussian components and the weight matrix (see supplementary material,[28] Sec. G1 for detailed definition of θ).

Note that the conditional distribution $p(\{x_{k\tau}\}|s_\tau(\theta))$ can be interpreted as the marginal conditional distribution given

by summing the joint distribution $p(\{x_{k\tau}\}, \mathcal{I}|s_\tau(\theta))$ over all possible values of $\mathcal{I}$, i.e.,

$$p(\{x_{k\tau}\}|s_\tau(\theta)) = \sum_{\mathcal{I}} p(\{x_{k\tau}\}, \mathcal{I}|s_\tau(\theta)). \quad (28)$$

Therefore, instead of solving (27) directly, we can utilize the EM algorithm to maximize the likelihood of θ by iteratively applying the following two steps:

E-step: Compute the functional

$$Q\left(\theta|\theta^{(\ell-1)}\right) = \mathbb{E}_{q^{(\ell)}}[\log p(\mathcal{I}, \{x_{k\tau}\}|s_\tau(\theta))], \quad (29)$$

where

$$q^{(\ell)}(\mathcal{I}) = p\left(\mathcal{I}|\{x_{k\tau}\}, s_\tau\left(\theta^{(\ell-1)}\right)\right), \quad (30)$$

and $\mathbb{E}_q[\cdot]$ denotes the expected value under the assumption that $\mathcal{I}$ follows the distribution $q(\mathcal{I})$. (Explicit expressions of $Q\left(\theta|\theta^{(\ell-1)}\right)$ are given in supplementary material,[28] Sec. H1.)

M-step: Solve

$$\theta^{(\ell)} = \arg\max_\theta Q\left(\theta|\theta^{(\ell-1)}\right). \quad (31)$$

This is a nonlinear problem and has no analytical solution, but a numerical and locally optimal solution can be obtained by using numerical optimization schemes.

For more technical and implementation details of the EM algorithm, please see supplementary material,[28] Sec. H.

C. Bayesian estimation

According to (26), the posterior distribution of the GMTM given $\{x_{k\tau}\}$ can be sampled by using the Gibbs sampler presented in Algorithm I.

ALGORITHM I. Gibbs sampler for GMTMs.

1: (**Initialization**) Specify a prior distribution $p(\theta)$ of the GMTM parameter vector θ and choose an initial value $\theta = \theta^{(0)}$ arbitrarily, where θ has the same definition as in the ML estimator (27).
2: **for** $\ell = 1$ to $M' + M$ **do**
3: (**Sampling of latent variables**) Draw $\mathcal{I}$ from its full conditional posterior

$$p(\mathcal{I}|\{x_{k\tau}\}, s_\tau(\theta)). \quad (32)$$

4: (**Sampling of** θ) Draw θ from its full conditional posterior

$$p(\theta|\mathcal{I}, \{x_{k\tau}\}). \quad (33)$$

5: Let $\theta^{(\ell)} = \theta$.
6: **end for**
7: **return** $\theta^{(M'+1)}, \ldots, \theta^{(M'+M)}$

Note that the simulated Markov chain $\{(\mathcal{I}^{(\ell)}, \theta^{(\ell)})\}$ in Algorithm I is ergodic due to the fact that $p(\mathcal{I}^{(\ell+1)}, \theta^{(\ell+1)}|\mathcal{I}^{(\ell)}, \theta^{(\ell)}) > 0$ for all $(\mathcal{I}^{(\ell+1)}, \theta^{(\ell+1)})$ and $(\mathcal{I}^{(\ell)}, \theta^{(\ell)})$. Therefore, after discarding the first M' burn-in samples, we can approximate the posterior distribution of θ as

$$p(\theta|\{x_{k\tau}\}) \approx \frac{1}{M} \sum_{\ell=M'}^{M'+M} \delta_{\theta^{(\ell)}}(\theta), \quad (34)$$

where $\delta_{\theta^{(\ell)}}(\cdot)$ denotes the Dirac measure centered on the point $\theta^{(\ell)}$.

The definition of θ and the other implementation details of Algorithm I are given in the supplementary material,[28] Sec. G.

V. APPLICATIONS

To verify the validity and advantage of the developed GMTM, we apply GMTMs to spectral estimation of three test systems: a Brownian dynamics simulation in a bistable potential and MD simulations of the alanine dipeptide and the bovine pancreatic trypsin inhibitor (BPTI) protein, and compare the estimation results with that provided by traditional MSMs, where we utilize k-means algorithm to generate discrete bins for MSMs and the detailed description of Bayesian and ML inference algorithm of MSMs can be found in Refs. 7 and 33.

In addition, since a MSM with m bins can also be viewed as a transition model with m basis functions (see supplementary material,[28] Sec. I), here we use the same name "model order" and symbol m to denote the number of basis functions in either model.

A. One-dimensional diffusion process

We first consider a one-dimensional diffusion process which is driven by the following Brownian dynamics:

$$\mathrm{d}x_t = -\nabla V(x_t)\,\mathrm{d}t + \sqrt{2\beta^{-1}}\mathrm{d}W_t. \quad (35)$$

Here, W_t denotes standard Brownian motion on $\mathbb{R}$, $\beta = 1/(k_B T) = 3.125$ is used as inverse temperature, and the function $V(x) = \left(x^2 - 1\right)^2 + 0.25x$ is a smooth two-well potential as shown in Fig. 1(a).

We utilize the Euler-Maruyama algorithm to generate a simulation trajectory $\{x_{k\tau}\}$ with length 10 000 s and lagtime $\tau = 0.25$ s and identify the first two dominant spectral components by using GMTMs and MSMs with model order $m = 3, 4, \ldots, 9$. Fig. 1(b) summarizes the resulting estimates of the second implied timescale t_2, and estimates of the first two left eigenfunctions

$$\begin{aligned}(l_1(x), l_2(x)) &= (\sqrt{\pi(x)}\phi_1(x), \sqrt{\pi(x)}\phi_2(x)) \\ &= (\phi_1(x)^2, \phi_1(x)\,\phi_2(x))\end{aligned} \quad (36)$$

are displayed in Figs. 1(d) and 1(e). The GMTM estimates of the spectral components are much better than the MSM estimates especially for small model orders. The MSMs underestimate t_2 as predicted by the variational principle.[26] In contrast, the GMTMs can provide an accurate estimate of t_2 for the model order $m \geq 8$.

In order to further clarify the difference between GMTMs and MSMs in spectral estimation, we also plot the estimates

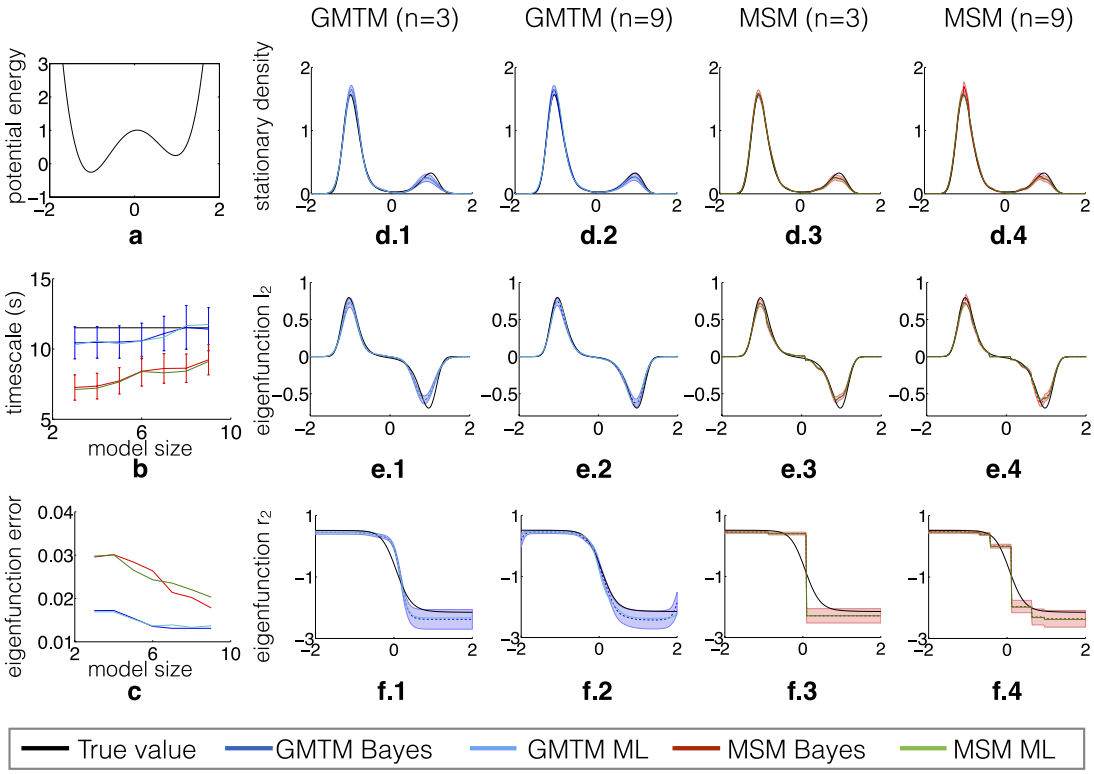

FIG. 1. Comparison of GMTM and MSM for modeling the diffusion in a two-well potential. (a) Potential function $V(x)$ in (35). (b) Estimates of the second timescale. (c) Eigenfunction approximation error δ_2 (see (38)) given by the GMTM and MSM. ((d)-(f)) Estimates of the first two left eigenfunctions (l_1, l_2) and the second right eigenfunction r_2 provided by the GMTM and MSM with model order $m = 3$ and 9. Error bars shown in (b) and ((d)-(f)) correspond to the one-sigma confidence intervals given by Bayesian inference.

of the second right eigenfunction

$$r_2(x) = \phi_2(x)/\sqrt{\pi(x)} = \phi_2(x)/\phi_1(x) \tag{37}$$

and the corresponding weighted approximation error

$$\delta_2 = \int (\hat{r}_2(x) - r_2(x))^2 \pi(x)\,\mathrm{d}x \tag{38}$$

in Figs. 1(c) and 1(f). For MSMs, weighted approximation errors of right functions are important quantities and can be used to characterize the approximation quality of a MSM.[13] It can be seen from Figs. 1(f.3) and 1(f.4) that MSMs approximate the right eigenfunction r_2 by step functions, which leads to large error between $\hat{r}_2(x)$ and $r_2(x)$ for x in the transition region between potential wells. This error decays as the model order (i.e., the number of bins) increases, but only with a low rate. In contrast with MSMs, GMTMs approximate the right eigenfunction by combining smooth basis functions, and the locations and shapes of the basis functions are optimized according to the likelihood function. Therefore, the GMTM can achieve a good approximation even with a small set of basis functions. This can also be demonstrated by Figs. 1(b) and 1(c); the estimated t_2 and r_2 given by the GMTM with 3 basis functions are more accurate than that given by the MSM with 9 bins.

B. Alanine dipeptide

Alanine dipeptide (acetyl-alanine-methylamide) is a small molecule whose structural and dynamical properties of this molecule have been thoroughly studied. It is known that its configuration space can be conveniently described by two backbone dihedral angles φ and ψ (see Fig. 2(a)).

We perform 20 independent MD simulations of the alanine dipeptide with length 200 ns and save configurations every 10 ps. A GMTM is estimated with $m = 6$, and MSMs are estimated with $m = 6, 12, 18$ to estimating the first three spectral components from the MD data using Bayesian inference and the maximum-likelihood estimator. The detailed settings of MD simulations are described in Ref. 27. Because the eigenpairs of alanine dipeptide cannot be directly calculated from the simulation model, here we evaluate the estimation results by the estimated implied timescales as a function of the lag time. According to (5), an implied timescale should be a constant independent of lag time τ. But in practice, the estimated implied timescales

$$\hat{t}_i(\tau) = -\frac{\tau}{\hat{\lambda}_i(\tau)} \tag{39}$$

are usually influenced by fast and un-modeled processes contained in the systems. Therefore, $\hat{t}_i$ tend to be smaller than the true values t_i when the lag time τ is too small. As τ increases, the fast processes decay to zero and we can expect that $\hat{t}_i(\tau)$ approaches to the true value. Thus, we can compare the GMTM estimates and MSM estimates based on the convergence rates of $\hat{t}_i(\tau)$. Fig. 2(b) shows that MSM estimates $\hat{t}_2(\tau)$ converge very slowly with model order 6 and 12 and converges to an almost constant value at $\tau = 10$ ps for $m = 18$. The GMTM with only 6 basis functions achieves the similar convergence rate of $\hat{t}_2(\tau)$ as

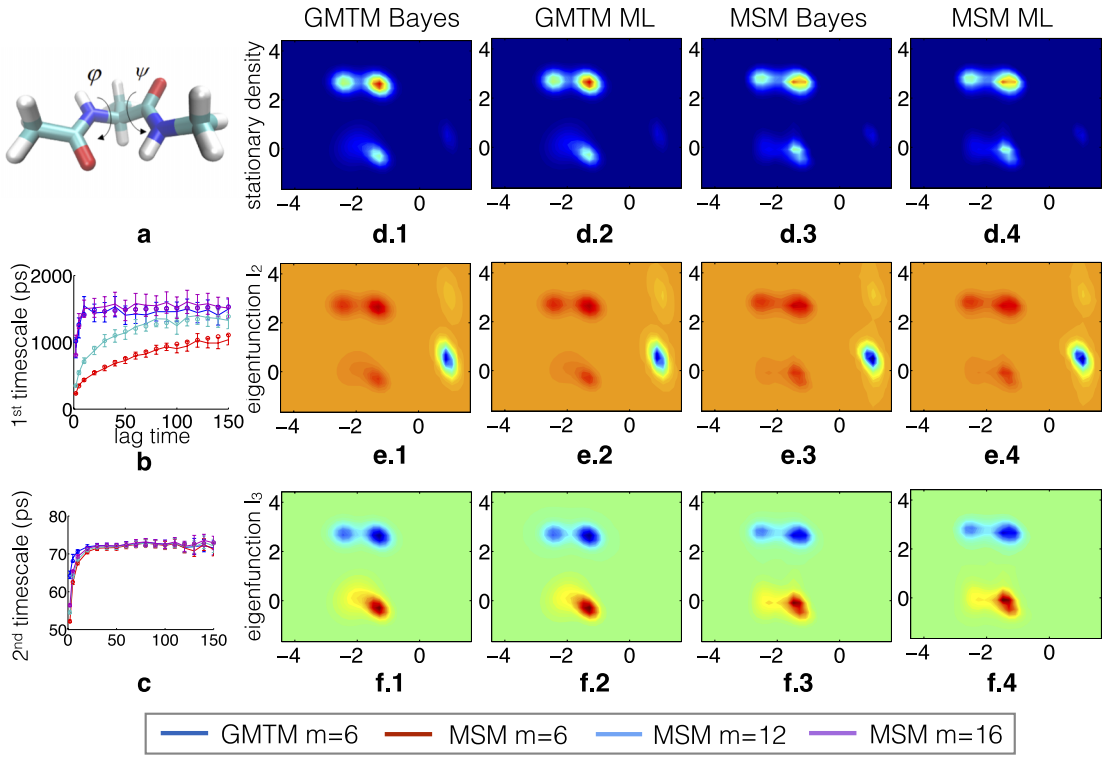

FIG. 2. Comparison of GMTM and MSM for modeling the conformation dynamics of alanine dipeptide. (a) Illustration of the structure of alanine dipeptide. ((b),(c)) Estimates of the second and third timescales, where solid lines represent the Bayesian estimation results with standard deviation error bars and the ML estimates are shown by circles. ((d)-(f)) Estimates of the first three left eigenfunctions (l_1, l_2, l_3) provided by the GMTM with model order $m = 6$ and the MSM with model order $m = 18$.

the MSM with $m = 18$, which demonstrates the superior performance of the proposed GMTMs. Note that the third spectral component of the alanine dipeptide is related to the transition between the two metastable states in the area $\varphi < 0$ (see Fig. 2(f)). This area contains most inter-metastable-state transitions in simulation data and tends to be finely coarse-grained by k-means algorithm, so the third implied timescale is relatively easy to identify for MSMs. As can be seen from Fig. 2(c), the three MSMs provide the similar $\widehat{t}_2(\tau)$ for $\tau \geq 10$ ps as the GMTM. But for lag times smaller than 10 ps, the GMTM gives better estimates of t_2 than the MSMs.

Figs. 2(d)-2(f) summarize estimates of the first three eigenfunctions given by the GMTM and the MSM with $m = 18$ at lag time $\tau = 100$ ps. As shown in the figures, the estimated dominant eigenfunctions of the GMTM are almost identical to those of the finely discretized MSM, and the three metastable states of alanine dipeptide are clearly indicated by the different sign structures of the eigenfunctions.

C. BPTI

BPTI is globular protein containing 58 amino acid residues and has a molecular mass of 6512 Dalton.[34] Its secondary structure is illustrated in Figs. 3(a). In this section, the spectral components of BPTI are analyzed based on a MD simulation with length ~1 ms generated by the special-purpose supercomputer Anton.[35] In order to eliminate the redundant atomistic degrees of freedom, we employ the time-lagged independent component analysis algorithm[5,22] to extract slow independent components of the molecular configuration and then perform the spectral identification in the reduced feature space of the 5 slowest independent components.

Estimated timescales of the second and third spectral components of BPTI are plotted in Figs. 3(b) and 3(c). The kinetics of BPTI is much more complicated than that of alanine dipeptide, and it is difficult for MSMs with small numbers of states to accurately capture the slow processes. As shown in Fig. 3(b), MSM estimates of t_2 converge very slowly towards around 30 μs and 40 μs for $m = 10$ and 50, and the 100-state MSM achieves a nearly τ-constant estimate around $43 \sim 45$ μs at lag time $\tau = 1.3$ μs. In contrast with MSMs, the GMTM provides larger estimates of t_2 with better convergence rate by using only 10 basis functions, which gives the estimate $\widehat{t}_2 = 44$ μs at $\tau = 0.9$ μs. From Fig. 3(c), we can also see that the GMTM has better convergence speed and stability in estimation of t_3. Within statistical error, our estimates agree with hidden Markov model estimates in Ref. 36.

Figs. 3(d)-3(f) show estimation results of projected eigenfunctions $l_1^{\text{proj}}, l_2^{\text{proj}}, l_3^{\text{proj}}$ at lag time $\tau = 20$ μs. For convenience of illustration, we display projections of eigenfunctions onto the first and second independent components. It can be observed that the GMTM with $m = 10$ and the MSM with $m = 100$ achieve similar estimates of projected eigenfunctions. Moreover, from the sign changes in projected eigenfunctions, it is interesting to see that the second and third eigenfunctions characterize the metastable state transitions along the first and second independent components separately.

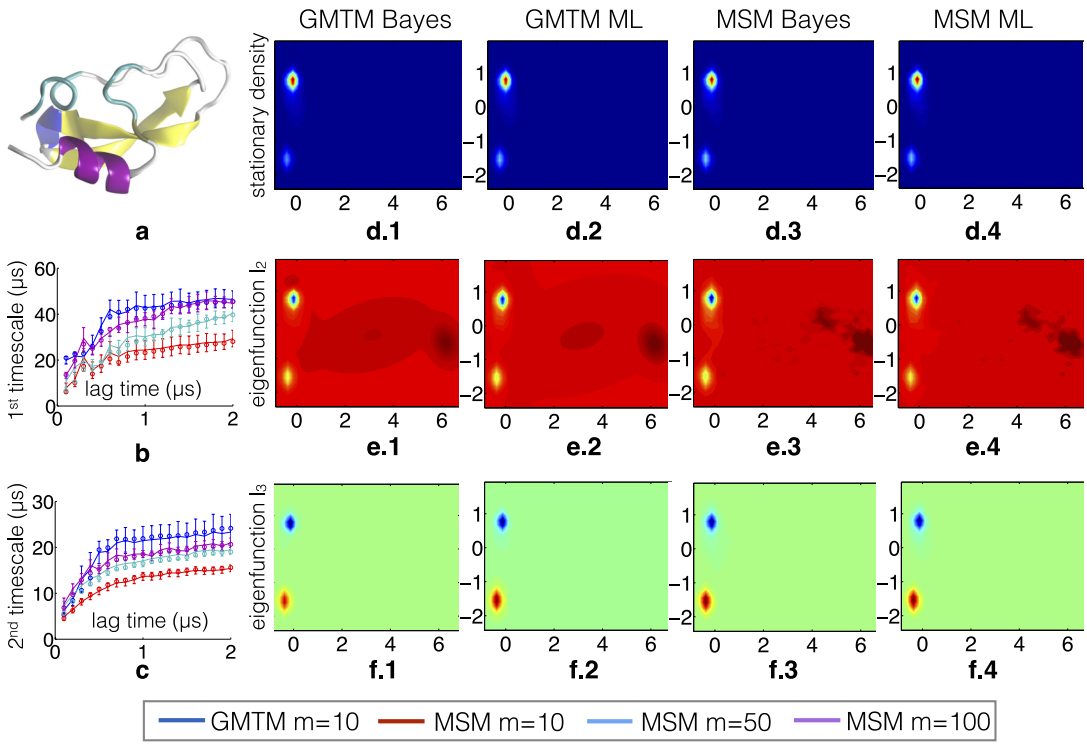

FIG. 3. Comparison of GMTM and MSM for modeling the conformation dynamics of BPTI. (a) Illustration of the structure of BPTI. ((b),(c)) Estimates of the second and third timescale, where solid lines show the Bayesian estimates with standard deviation error bars and circles represent ML estimation results. ((d)-(f)) Estimates of the first three projected eigenfunctions $(l_1^{\text{proj}}, l_2^{\text{proj}}, l_3^{\text{proj}})$ provided by the GMTM with model order $m = 10$ and the MSM with model order $m = 100$.

VI. CONCLUSIONS

We have developed a parametric model based statistical approach for extracting slow processes of molecular kinetics from MD simulations. The framework of this approach can be sketched as follows:

The GMTM presented in this paper plays an essential role in the approach. The concept of "transition model" is a continuous extension of MSM, and we can prove that the MSM is in fact a specific transition model with piecewise basis functions. The advantages of GMTMs over MSMs are (i) GMTMs can provide smooth approximations to eigenfunctions by using Gaussian basis functions. (ii) All parameters of Gaussian basis functions in GMTMs can be estimated in a Bayesian fashion. Numerical examples show that much less basis functions are required with GMTMs to obtain estimates of equal quality when MSMs are used. In comparison with the other parametric modeling tools for Markov processes in continuous spaces (e.g., parametric models of Itō processes), the eigenpairs of GMTMs can be simply and analytically computed from model parameters, and it can be shown that a GMTM is able to approximate (arbitrarily closely) the symmetric operator of any ergodic and reversible Markov process under some general assumptions. Moreover, likelihood functions of GMTMs can be written in a form like likelihood of Gaussian mixture models, then the EM algorithm and blocked Gibbs sampling algorithm can be used to achieve ML and Bayesian estimates of GMTMs as in learning Gaussian mixture models.

Future work on GMTMs will focus on

1. More efficient statistical inference algorithms. The computation time of GMTM estimation can be expressed as [(update time of sufficient statistics + optimization time of model parameters) × iteration number], where the computation time of sufficient statistics of latent variables $\mathcal{I}$ is linear in data size and comparable to that of count matrices in MSMs. In order to improve the efficiency of applying GMTMs to complex molecular systems, we will investigate how to reduce the number of undetermined parameters of Gaussian basis functions and accelerate the convergence rate by some heuristic method based on clustering or time lagged independent component analysis. Moreover, in this paper, parameter optimization steps in both Bayesian and ML inference methods are implemented in a random-walk fashion. It is natural to expect that the efficiency of inference algorithms of GMTMs can be further enhanced by using some advanced parameter optimization algorithms, e.g., the Newton algorithm.
2. Sparse priors for GMTMs. In the present paper, the Bayesian inference of GMTMs is performed with uniform prior on model parameters, which may lead to numerical instability or over-fitting of estimates especially for small-sized simulation data. (The ML estimator can be viewed as a specific maximum a posteriori estimator with uniform prior.) In order to improve the robustness of GMTMs, as a subject of future investigations, a sparse prior on weight matrices of GMTMs will be developed to encourage the

sparsity of weight matrices so that the weights of redundant Gaussian components are enforced to be close to zero in the case of insufficient data.

ACKNOWLEDGMENTS

We thank C. Clementi (Rice University), J. D. Chodera (MSKCC, New York), and the entire CMB group at FU Berlin for valuable scientific discussions. We acknowledge funding from DFG WU 744/1-1, DFG SFB 1114 (Wu), and ERC Grant "pcCell" (Noé).

[1]T. Hansson, C. Oostenbrink, and W. van Gunsteren, Curr. Opin. Struct. Biol. **12**, 190 (2002).
[2]F. Noé, S. Doose, I. Daidone, M. Löllmann, M. Sauer, J. D. Chodera, and J. C. Smith, Proc. Natl. Acad. Sci. U.S.A. **108**, 4822 (2011).
[3]P. Deuflhard and M. Weber, Linear Algebra Appl. **398**, 161 (2005).
[4]M. A. Rohrdanz, W. Zheng, M. Maggioni, and C. Clementi, J. Chem. Phys. **134**, 124116 (2011).
[5]G. Perez-Hernandez, F. Paul, T. Giorgino, G. De Fabritiis, and F. Noé, J. Chem. Phys. **139**, 015102 (2013).
[6]S. Kube and M. Weber, J. Chem. Phys. **126**, 024103 (2007).
[7]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).
[8]N. Djurdjevac, M. Sarich, and C. Schütte, in *Proceedings of the International Congress of Mathematicians* (World Scientific, Hyderabad, 2010), pp. 3105–3131.
[9]F. Noé, I. Horenko, C. Schütte, and J. C. Smith, J. Chem. Phys. **126**, 155102 (2007).
[10]J. D. Chodera, K. A. Dill, N. Singhal, V. S. Pande, W. C. Swope, and J. W. Pitera, J. Chem. Phys. **126**, 155101 (2007).
[11]N. Singhal and V. S. Pande, J. Chem. Phys. **123**, 204909 (2005).
[12]N. V. Buchete and G. Hummer, J. Phys. Chem. B **112**, 6057 (2008).
[13]M. Sarich, F. Noé, and C. Schütte, SIAM Multiscale Model. Simul. **8**, 1154 (2010).
[14]M. Dellnitz and A. Hohmann, Numer. Math. **75**, 293 (1997).
[15]M. Dellnitz and O. Junge, SIAM J. Numer. Anal. **36**, 491 (1999).
[16]C. Hsu, Int. J. Bifurcation Chaos **2**, 727 (1992).
[17]J. A. W. van der Spek, "Cell mapping methods: Modifications and extensions," Ph.D. thesis (Technische Universiteit Eindhoven, 1994).
[18]A. Baba and T. Komatsuzaki, Proc. Natl. Acad. Sci. U.S.A. **104**, 19297 (2007).
[19]P. Schuetz, R. Wuttke, B. Schuler, and A. Caflisch, J. Phys. Chem. B **114**, 15227 (2010).
[20]G. Berezovska, D. Prada-Gracia, S. Mostarda, and F. Rao, J. Chem. Phys. **137**, 194101 (2012).
[21]C. Schütte and M. Sarich, *Metastability and Markov State Models in Molecular Dynamics: Modeling, Analysis, Algorithmic Approaches* (AMS, 2013), Vol. 24.
[22]C. R. Schwantes and V. S. Pande, J. Chem. Theory Comput. **9**, 2000 (2013).
[23]C. R. Shalizi and J. P. Crutchfield, J. Stat. Phys. **104**, 817 (2001).
[24]C.-B. Li, H. Yang, and T. Komatsuzaki, Proc. Natl. Acad. Sci. U.S.A. **105**, 536 (2008).
[25]J. P. Crutchfield, Nat. Phys. **8**, 17 (2012).
[26]F. Noé and F. Nüske, SIAM Multiscale Model. Simul. **11**, 635 (2013).
[27]F. Nüske, B. G. Keller, G. Pérez-Hernández, A. S. Mey, and F. Noé, J. Chem. Theory Comput. **10**, 1739 (2014).
[28]See supplementary material at http://dx.doi.org/10.1063/1.4913214 for the comparison of different operator descriptions of Markov processes, proofs of propositions and implementation details of estimation methods.
[29]S. S. Lafon, "Diffusion maps and geometric harmonics," Ph.D. thesis (Yale University, 2004).
[30]R. R. Coifman and S. Lafon, Appl. Comput. Harmonic Anal. **21**, 5 (2006).
[31]G. McLachlan and D. Peel, *Finite Mixture Models* (Wiley, New York, 2000).
[32]H. Snoussi and A. Mohammad-Djafari, *Bayesian Inference and Maximum Entropy Methods in Science and Engineering*, AIP Conference Series Vol. 617, edited by R. Fry (AIP Publishing, 2002), pp. 36–46; arXiv:physics/0111007.
[33]F. Noé, J. Chem. Phys. **128**, 244103 (2008).
[34]P. Ascenzi, A. Bocedi, M. Bolognesi, A. Spallarossa, M. Coletta, R. Cristofaro, and E. Menegatti, Curr. Protein Pept. Sci. **4**, 231 (2003).
[35]D. E. Shaw, P. Maragakis, K. Lindorff-Larsen, S. Piana, R. O. Dror, M. P. Eastwood, J. A. Bank, J. M. Jumper, J. K. Salmon, Y. Shan *et al.*, Science **330**, 341 (2010).
[36]F. Noé, H. Wu, J.-H. Prinz, and N. Plattner, J. Chem. Phys. **139**, 184114 (2013).