Calculation of the distribution of eigenvalues and eigenvectors in Markovian state models for molecular dynamics

Nina Singhal Hinrichs
Department of Computer Science, Stanford University, Stanford, California 94305

Vijay S. Pande
Department of Chemistry, Stanford University, Stanford, California 94305

(Received 26 December 2006; accepted 20 April 2007; published online 22 June 2007)

Markovian state models (MSMs) are a convenient and efficient means to compactly describe the kinetics of a molecular system as well as a formalism for using many short simulations to predict long time scale behavior. Building a MSM consists of grouping the conformations into states and estimating the transition probabilities between these states. In a previous paper, we described an efficient method for calculating the uncertainty due to finite sampling in the mean first passage time between two states. In this paper, we extend the uncertainty analysis to derive similar closed-form solutions for the distributions of the eigenvalues and eigenvectors of the transition matrix, quantities that have numerous applications when using the model. We demonstrate the accuracy of the distributions on a six-state model of the terminally blocked alanine peptide. We also show how to significantly reduce the total number of simulations necessary to build a model with a given precision using these uncertainty estimates for the blocked alanine system and for a 2454-state MSM for the dynamics of the villin headpiece. © *2007 American Institute of Physics*. [DOI: 10.1063/1.2740261]

I. INTRODUCTION

Many important processes in biology occur at the molecular scale. A detailed understanding of these processes can lead to significant advances in the medical and life sciences—for example, many diseases are caused by protein aggregation or misfolding and potential drug molecules can be designed by understanding their binding properties and conformational changes. These processes have typically been studied through experiments. While such experiments can yield a wealth of insight, they are often insufficient to describe the system dynamics on an atomic scale. An alternative approach is to use physically based computational simulations to model the interactions and movement of the molecules. While molecular simulations are computationally expensive, it is now possible to simulate many independent molecular dynamics trajectories in a parallel fashion by using distributed computing methods such as Folding@Home.[1]

After generating large ensembles of molecular dynamics simulations, we wish to analyze these trajectories to find thermodynamic properties such as the equilibrium conformational distribution of the protein and kinetic properties such as the rate and mechanism of folding. A recent approach for such forms of analysis involves graph-based models of protein kinetics that divide the conformation space into discrete states and calculate transition probabilities or rates between the states based on molecular dynamics trajectories.[2–9] These Markovian state models (MSMs) allow one to easily combine and analyze simulation data that started from arbitrary conformations and naturally handle the existence of intermediate states and traps. This approach has been applied to small protein systems,[4,6,10] a nonbiological polymer,[11,12] and vesicle fusion[13] with good agreement with experimental rates. The MSM uses discrete states, and we expect the transition probabilities to be insensitive to the exact state boundaries after sufficient transition time.[14–16] An alternative approach uses fuzzy partitions and partial membership of conformations into states, and may be able to better characterize transition regions and describe dynamics at shorter time scales.[17,18]

For any quantities which can be calculated from the MSM, such as the mean first passage times between states,[4] probability of folding from a given conformation,[4] or rates,[5] it is also important to determine the uncertainty in these values, so that one can form an idea about the confidence of the results. One main source of error is caused by grouping conformations into states and assuming that transitions between these states are Markovian. It has been shown that if the conformations are grouped incorrectly or if the transition probabilities are calculated from a time step which is too short, the transitions are no longer history independent, and any analysis that assumes a Markovian process may produce incorrect results.[5] Even if the states are defined such that the transitions between them are Markovian, the results could still be in error. This second source of error results from the finite sampling of transitions between states, which gives uncertainties in the transition probability estimates and, in turn, leads to uncertainties in the values we calculate.

In a previous paper,[19] we focused on the uncertainties caused by finite sampling and showed how to efficiently calculate the resulting uncertainty in the mean first passage time between two states. Those methods can easily be applied to calculate the uncertainty in any quantity that can be expressed as the solution of a set of linear equations of the

transition probabilities. However, many interesting collective properties of the system are described using the eigenvalues and eigenvectors of the transition matrix. For example, the eigenvalues correspond to the aggregate time scales of the system, and thus can be compared with experiments to validate the model.[2,6] Additionally, they are used in some tests for determining the time at which the system becomes Markovian.[5] The eigenvectors are useful in determining the states which participate in the relaxation process corresponding to a given eigenvalue, and can be used to group kinetically similar states.[20–24]

In this paper, we will extend the uncertainty analysis methods presented previously[19] to estimate the uncertainties in the eigenvalues and eigenvectors of a transition matrix caused by finite sampling. These error estimates can again be calculated in efficient closed-form solutions. Moreover, these error estimates can be used to adaptively direct further simulations to reduce the uncertainties of functions of the eigenvalues or eigenvectors. The validity of the error estimates is demonstrated on a small system, the terminally blocked alanine peptide, and the power of adaptive sampling is demonstrated on the alanine peptide and a model of the villin peptide.

II. METHODS

Molecular dynamics simulations are a popular tool for understanding molecular motion. Analyzing these trajectories to extract kinetic information is a difficult task. Recent works[2–9] have involved modeling the system as a Markovian state model, where the conformation space of the molecule is divided into discrete regions, or states, and transition probabilities are calculated between the states. If the transitions between the states are Markovian, or history independent on some time scale, it is possible to model the long time scale behavior of the system as a Markov chain on the Markovian state model graph.

Determining a state space over which transitions are Markovian is a difficult task and there has been much work on determining appropriate decompositions.[25,26] Even if an appropriate decomposition can be found for which the dynamics are Markovian at some lag time, the kinetic properties calculated from the model still have uncertainties. Since we can only sample a finite number of transitions between states, the estimated transition probabilities between states will have statistical uncertainty. Therefore, any value calculated from the transition probabilities will also have an uncertainty associated with it. In a previous paper,[19] we mapped the uncertainty in the transition probabilities to uncertainties in the mean first passage time between two states or other similar quantities that are solutions of linear equations in the transition probabilities.

In the following section, we calculate efficient closed-form expressions for the uncertainties in the eigenvalues and eigenvectors of the Markovian state model, which describe the full kinetics of the system. The basis for the derivation and many of the equations are similar to those for the mean first passage time.[19] However, we reproduce them here for clarity.

A. Eigenvalue and eigenvector equations

In a Markovian state model, we represent the conformation space by K discrete states, each of which corresponds to some distinct group of molecular conformations. Let us define the probability of transitioning from state i to state j at a time step of Δt as p_{ij}.

An eigenvalue λ of a matrix $\mathbf{P}$ is defined as

$$\mathbf{P}\mathbf{v}_\lambda = \lambda \mathbf{v}_\lambda, \tag{1}$$

where $\mathbf{v}_\lambda$ is the eigenvector corresponding to eigenvalue λ. We define matrix $\mathbf{A}$ with rows $\mathbf{a_i}$ as

$$\mathbf{A} = \mathbf{P} - \lambda \mathbf{I} = \begin{bmatrix} p_{11}-\lambda & p_{12} & \cdots & p_{1K} \\ p_{21} & p_{22}-\lambda & \cdots & p_{2K} \\ & \ddots & & \\ p_{K1} & \cdots & p_{(K-1)K} & p_{KK}-\lambda \end{bmatrix}, \tag{2}$$

where $\mathbf{I}$ is the identity matrix. Eq. (1) is then equivalent to

$$\mathbf{A}\mathbf{v}_\lambda = \mathbf{0}, \tag{3}$$

and has nontrivial solution $\mathbf{v}_\lambda$ when the determinant of the matrix is zero,

$$\det(\mathbf{A}) = 0. \tag{4}$$

B. Transition probability distribution

Finite sampling causes uncertainties in the estimates of the transition probabilities between states. A derivation and a complete explanation of the distribution over transition probability vectors have been given before.[19] Here, we summarize the main results.

We define p_{ij}^* as the actual transition probability from state i to j at a time step of Δt, where the sum of the transition probabilities from state i is equal to one. We can estimate these transition probabilities by independently sampling transitions between states i and j, either through independent simulations or by only including transitions separated by the lag time at which the transitions are Markovian. We generate counts z_{ij} which are the total number of transition samples from state i to state j. We define n_i as the total number of samples originating from state i,

$$n_i = \sum_{j=1}^{K} z_{ij}. \tag{5}$$

The distribution of the z_{ij} variables follows the multinomial distribution with parameters $n_i, p_{i1}^*, p_{i2}^*, \ldots, p_{iK}^*$.[27] Using Bayesian analysis, we can compute the distribution over all possible vectors of transition probabilities. The probability of a particular column vector $\mathbf{p_i}$ being the true transition probability vector, given the observed transition counts, is, from Bayes' rule,

$$P(\mathbf{p_i}|\mathbf{z_i}) \propto P(\mathbf{z_i}|\mathbf{p_i})P(\mathbf{p_i}) = p_{i1}^{z_{i1}} p_{i2}^{z_{i2}} \cdots p_{iK}^{z_{iK}} P(\mathbf{p_i}), \tag{6}$$

where $P(\mathbf{p}_i)$ is the prior probability over the transition probability vectors, i.e., the distribution representing the state of knowledge of transition probability vectors before observing any data.

A convenient choice for the prior is the Dirichlet distribution, the conjugate prior of the multinomial distribution. The Dirichlet distribution with variables $\mathbf{p}$ and parameters $\mathbf{u}$ is defined as

$$\text{Dirichlet}(\mathbf{p};\mathbf{u}) = \frac{1}{Z(\mathbf{u})}\prod_{i=1}^{K} p_i^{u_i-1}, \tag{7}$$

$$Z(\mathbf{u}) = \frac{\prod_{i=1}^{K}\Gamma(u_i)}{\Gamma\left(\sum_{i=1}^{K}u_i\right)}, \tag{8}$$

where $Z(\mathbf{u})$ is a normalizing constant and Γ is the gamma function. If we define the prior of the transition probabilities as a Dirichlet distribution with parameters $\alpha_{i1},\alpha_{i2},\ldots,\alpha_{iK}$ and we observe transition counts $z_{i1},z_{i2},\ldots,z_{iK}$, the posterior of the transition probabilities is a Dirichlet distribution with parameters $\alpha_{i1}+z_{i1},\alpha_{i2}+z_{i2},\ldots,\alpha_{iK}+z_{iK}$. For notational convenience, we define the Dirichlet counts as

$$u_{ij} = \alpha_{ij} + z_{ij}. \tag{9}$$

Therefore, assuming a Dirichlet prior, the distribution of the transition probabilities $\mathbf{p}_i$ given the observed data counts is Dirichlet$(\mathbf{p}_i;\mathbf{u}_i)$. In the limit, as the sampling (and therefore transition counts) increases, the distribution of the transition probabilities will not depend on the choice of the prior distribution.

It will be useful to state the expected values of the posterior distribution of the transition probabilities for future reference, where w_i are normalizing weight variables,[28]

$$\bar{p}_{ij} = E(p_{ij}) = \frac{u_{ij}}{w_i},$$

$$w_i = \sum_{j=1}^{K} u_{ij}. \tag{10}$$

C. Distribution of eigenvalues and eigenvectors

It is possible to repeatedly sample from the transition probability posterior distribution and find the eigenvalues and eigenvectors for each sample to determine the posterior distributions of these quantities. However, this method is very expensive, both because many samples are required to accurately describe the distribution and the solution of the eigenvalue system is expensive [$O(K^3)$ plus some small number of iterations[29]] for each sample. For these reasons, we will make two approximations that will yield efficient closed-form solutions for the distributions of the eigenvalue λ and the corresponding eigenvector $\mathbf{v}_\lambda$. If the distributions of multiple eigenvalue/eigenvector pairs are desired, this procedure would need to be repeated independently for each pair.

1. Taylor series approximation

First, we will approximate the eigenvalue and eigenvector of interest with a Taylor series expansion about these values calculated at the mean values of the transition probabilities. We define the mean matrix $\bar{\mathbf{A}}$ as

$$\bar{\mathbf{A}} = \begin{bmatrix} \bar{p}_{11}-\lambda & \bar{p}_{12} & \cdots & \bar{p}_{1K} \\ \bar{p}_{21} & \bar{p}_{22}-\lambda & \cdots & \bar{p}_{2K} \\ & & \ddots & \\ \bar{p}_{K1} & \cdots & \bar{p}_{(K-1)K} & \bar{p}_{KK}-\lambda \end{bmatrix}, \tag{11}$$

where the variables $\bar{p}_{ij}$ are defined in Eq. (10). The mean eigenvalue $\bar{\lambda}$ satisfies the following equation:

$$\det(\bar{\mathbf{A}}) = 0, \tag{12}$$

and the mean eigenvector $\bar{\mathbf{v}}_\lambda$ satisfies the following equation:

$$\bar{\mathbf{A}}|_{\bar{\lambda}}\bar{\mathbf{v}}_\lambda = \mathbf{0}. \tag{13}$$

The first-order Taylor series expansion for the eigenvalue λ as a function of the transition probabilities is

$$\lambda = \bar{\lambda} + \left.\frac{\partial\lambda}{\partial p_{11}}\right|_{\bar{\mathbf{A}}}\Delta p_{11} + \left.\frac{\partial\lambda}{\partial p_{12}}\right|_{\bar{\mathbf{A}}}\Delta p_{12} + \cdots + \left.\frac{\partial\lambda}{\partial p_{KK}}\right|_{\bar{\mathbf{A}}}\Delta p_{KK}, \tag{14}$$

where Δp_{ij} are small perturbations in the parameters. Appendix A shows how to compute the terms $\partial\lambda/\partial p_{ij}|_{\bar{\mathbf{A}}}$ in Eq. (14) efficiently.

Similarly, the first-order Taylor series expansion for the eigenvector $\mathbf{v}_\lambda$ as a function of the transition probabilities is

$$\mathbf{v}_\lambda = \bar{\mathbf{v}}_\lambda + \left.\frac{\partial\mathbf{v}_\lambda}{\partial p_{11}}\right|_{\bar{\mathbf{A}}}\Delta p_{11} + \left.\frac{\partial\mathbf{v}_\lambda}{\partial p_{12}}\right|_{\bar{\mathbf{A}}}\Delta p_{12} + \cdots + \left.\frac{\partial\mathbf{v}_\lambda}{\partial p_{KK}}\right|_{\bar{\mathbf{A}}}\Delta p_{KK}. \tag{15}$$

Appendix B shows how to calculate all the terms $\partial\mathbf{v}_\lambda/\partial p_{ij}|_{\bar{\mathbf{A}}}$ in Eq. (15) efficiently.

2. Multivariate normal approximation

As shown in Sec. II B, the transition probabilities $\mathbf{p_i}$ are distributed according to Dirichlet distributions. If the sample size is sufficiently large, then, by the central limit theorem, the distribution of $\mathbf{p_i}$ converges to a multivariate normal distribution[30] (MVN) with mean $\boldsymbol{\mu_i}$ and covariance matrix $\boldsymbol{\Sigma_i}$ given by

$$\boldsymbol{\mu_i} = \frac{\mathbf{u_i}}{w_i}, \tag{16}$$

$$\boldsymbol{\Sigma_i} = \frac{1}{w_i^2(w_i+1)}[w_i\,\text{diag}(\mathbf{u_i}) - \mathbf{u_i}\mathbf{u_i^T}], \tag{17}$$

where the superscript "**T**" denotes the transpose, diag$(\mathbf{u_i})$ represents a matrix with entries u_{ij} along the diagonal, and the w_i terms are the normalizing weight variables defined in Eq. (10). The covariance matrix in this distribution enforces the constraint that each possible transition probability vector $\mathbf{p_i}$ must sum to unity.

3. Closed-form solutions

Making both the Taylor series and multivariate normal approximations leads to closed-form expressions for the distributions of the eigenvalue λ and its corresponding eigenvector $\mathbf{v}_\lambda$. For notational convenience, we define the deviation vector $\Delta\mathbf{p_i}$, the sensitivity of λ vector $\mathbf{s_i^\lambda}$, and the sensitivity of $\mathbf{v}_\lambda$ matrix $\mathbf{S_i^{v_\lambda}}$,

$$\Delta\mathbf{p_i^T} = [\Delta p_{i1} \ \cdots \ \Delta p_{iK}],$$

$$(\mathbf{s_i^\lambda})^\mathbf{T} = \left[\left.\frac{\partial\lambda}{\partial p_{i1}}\right|_{\bar{\mathbf{A}}} \ \cdots \ \left.\frac{\partial\lambda}{\partial p_{iK}}\right|_{\bar{\mathbf{A}}} \right],$$

$$\mathbf{S_i^{v_\lambda}} = \begin{bmatrix} \left.\frac{\partial v_1^\lambda}{\partial p_{i1}}\right|_{\bar{\mathbf{A}}} & \cdots & \left.\frac{\partial v_1^\lambda}{\partial p_{iK}}\right|_{\bar{\mathbf{A}}} \\ \vdots & \ddots & \\ \left.\frac{\partial v_K^\lambda}{\partial p_{i1}}\right|_{\bar{\mathbf{A}}} & & \left.\frac{\partial v_K^\lambda}{\partial p_{iK}}\right|_{\bar{\mathbf{A}}} \end{bmatrix}. \tag{18}$$

We can then rewrite Eqs. (14) and (15) by grouping K terms at a time as

$$\lambda = \bar{\lambda} + \sum_{i=1}^{K} (\mathbf{s_i^\lambda})^\mathbf{T}\Delta\mathbf{p_i},$$

$$\mathbf{v}_\lambda = \bar{\mathbf{v}}_\lambda + \sum_{i=1}^{K} \mathbf{S_i^{v_\lambda}}\Delta\mathbf{p_i}. \tag{19}$$

The vector $\Delta\mathbf{p_i}$ is equal to $\mathbf{p_i}-\bar{\mathbf{p}}_\mathbf{i}$ and, with the MVN approximation, has mean $\mathbf{0}$ and covariance matrix $\mathbf{\Sigma}_\mathrm{i}$ given by Eq. (17). Linear combinations of MVN random variables are also MVN random variables: If $\mathbf{y}\sim\mathrm{MVN}(\boldsymbol{\mu},\mathbf{\Sigma})$, then $\mathbf{y}'=\mathbf{R}\mathbf{y}+\mathbf{b}$ is $\mathbf{y}'\sim\mathrm{MVN}(\mathbf{R}\boldsymbol{\mu}+\mathbf{b},\mathbf{R}\mathbf{\Sigma}\mathbf{R}^\mathrm{T})$.[30] Therefore, λ has a normal distribution with mean $\bar{\lambda}$ and variance σ^2,[31]

$$\lambda \sim N(\bar{\lambda},\sigma^2), \tag{20}$$

where

$$\sigma^2 = \sum_{i=1}^{K} (\mathbf{s_i^\lambda})^\mathbf{T}\mathbf{\Sigma_i}\mathbf{s_i^\lambda}. \tag{21}$$

Substituting Eq. (17) for $\mathbf{\Sigma}_\mathrm{i}$, we see that

$$\sigma^2 = \sum_{i=1}^{K} \frac{1}{w_i^2(w_i+1)}(\mathbf{s_i^\lambda})^\mathbf{T}[w_i\,\mathrm{diag}(\mathbf{u_i}) - \mathbf{u_i}\mathbf{u_i^T}]\mathbf{s_i^\lambda}$$

$$= \sum_{i=1}^{K} \frac{1}{w_i^2(w_i+1)}[w_i(\mathbf{s_i^\lambda})^\mathbf{T}\,\mathrm{diag}(\mathbf{u_i})\mathbf{s_i^\lambda} - ((\mathbf{s_i^\lambda})^\mathbf{T}\mathbf{u_i})(\mathbf{u_i^T}\mathbf{s_i^\lambda})]. \tag{22}$$

Similarly, $\mathbf{v}_\lambda$ has a multivariate normal distribution with mean $\bar{\mathbf{v}}_\lambda$ and covariance matrix $\mathbf{\Sigma}^2_{\mathbf{v}_\lambda}$,

$$\mathbf{v}_\lambda \sim \mathrm{MVN}(\bar{\mathbf{v}}_\lambda,\mathbf{\Sigma}^2_{\mathbf{v}_\lambda}), \tag{23}$$

where

$$\mathbf{\Sigma}^2_{\mathbf{v}_\lambda} = \sum_{i=1}^{K} \mathbf{S_i^{v_\lambda}}\mathbf{\Sigma_i}(\mathbf{S_i^{v_\lambda}})^\mathbf{T}$$

$$= \sum_{i=1}^{K} \frac{1}{w_i^2(w_i+1)}[w_i\mathbf{S_i^{v_\lambda}}\,\mathrm{diag}(\mathbf{u_i})(\mathbf{S_i^{v_\lambda}})^\mathbf{T} - (\mathbf{S_i^{v_\lambda}}\mathbf{u_i})(\mathbf{u_i}(\mathbf{S_i^{v_\lambda}})^\mathbf{T})]. \tag{24}$$

4. Computational cost

The closed-form solutions given in Eqs. (20) and (23) require solving for $\bar{\lambda}$ and $\bar{\mathbf{v}}_\lambda$ which take time $O(K^3)$.[29] Appendix A shows that we can find all the partial derivative terms for the eigenvalue in time $O(K^2)$ and Appendix B shows that we can find all the partial derivative terms for the eigenvector in time $O(K^3)$. Since the variance in Eq. (22) for the eigenvalue is the sum of vector dot products (rather than matrix-vector products), we can calculate it in time $O(K^2)$. Similarly, since the covariance matrix in Eq. (24) for the eigenvector is the sum of matrix-vector products (rather than matrix-matrix products), we can calculate it in time $O(K^3)$.

D. Adaptive sampling

As described previously,[19] we can decompose the closed-form normal or multivariate normal distributions to calculate the contribution to the variance from the elements in each row of the transition matrix, corresponding to the transitions from a single state. We can then start new simulations from the states which contribute the most to the variance in order to improve the overall precision.

The variance of the eigenvalue λ decomposes as

$$\sigma^2 = \sum_{i=1}^{K} \frac{\bar{q}_i}{w_i+1},$$

$$\bar{q}_i = (\mathbf{s_i^\lambda})^\mathbf{T}[\mathrm{diag}(\bar{\mathbf{p}}_\mathbf{i}) - \bar{\mathbf{p}}_\mathbf{i}\bar{\mathbf{p}}_\mathbf{i}^\mathbf{T}]\mathbf{s_i^\lambda}, \tag{25}$$

where we have separated out the $\bar{q}_i$ terms which do not depend on the allocation of samples w_i. If we were to add m more samples and assume that the expected transition probabilities remain constant, we can choose the state i which will decrease this variance the most as

$$i = \arg\max\left(\frac{\bar{q}_i}{w_i+1} - \frac{\bar{q}_i}{w_i+m+1}\right). \tag{26}$$

Similar calculations can be performed to obtain the state which contributes the most to any function of the covariance matrix of the eigenvector.

III. RESULTS

To test the closed-form solutions for the distribution for an eigenvalue λ given in Eq. (20) and an eigenvector $\mathbf{v}_\lambda$ given in Eq. (23), we compare the distributions with those obtained from sampling from the posterior transition probability distribution and solving each sample for the eigenvalue or eigenvector of interest. We can test all combinations of the two assumptions given above using different sampling and solving methods. Namely, method 1 will sample from

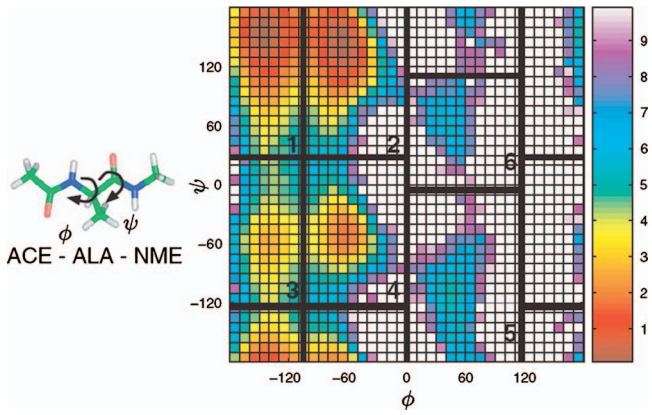

FIG. 1. (Color) Potential of mean force and manual state decomposition for terminally blocked alanine peptide. Left: The terminally blocked alanine peptide with ϕ and ψ backbone torsions labeled. Right: The potential of mean force in the (ϕ,ψ) torsions at 300 K estimated from the parallel tempering simulation. Boundaries defining the six states manually identified by Chodera *et al.* (Ref. 32) are superimposed and the states are labeled. Reproduced with permission from Chodera.

the Dirichlet distributions and solve for the eigenvalues or eigenvectors directly, method 2 will sample from the MVN distributions and solve for the eigenvalues or eigenvectors directly, method 3 will sample from the Dirichlet distributions and substitute into the Taylor series approximations, and method 4 will sample from the MVN distributions and substitute into the Taylor series approximations. In this way, we can determine independently whether the MVN approximations and the Taylor series approximations are valid. The equations derived above [Eqs. (20) and (23)] are simply closed-form solutions to the sampling based method 4.

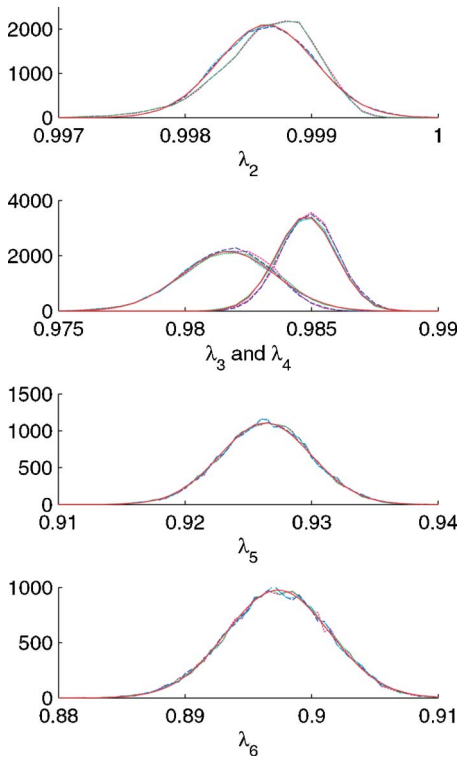

FIG. 2. (Color online) Distributions of the five nonunit eigenvalues of the system shown in Fig. 1. The solid (red) lines indicate the normal distributions calculated using Eq. (22), and the dashed (magneta), dotted (green), long dashed (blue), and dot-dashed (cyan) density plots indicate the distributions generated from the four sampling based methods, Dirichlet sampling and direct solving, MVN sampling and direct solving, Dirichlet sampling and Taylor series substitution, and MVN sampling and Taylor series substitution, respectively, for 20 000 samples.

We apply these methods to calculate the distributions of eigenvalues and eigenvectors in the terminally blocked alanine peptide (Fig. 1) to demonstrate that the multivariate normal and Taylor series approximations are valid. Stable states on the conformational landscape have previously been identified.[32] A set of 30000 shooting trajectories (5000 initiated from equilibrium distributions within each of the six states) at 300 K was obtained from Chodera *et al.*[32] We count transitions between these states at a lag time Δt of 0.1 ps, counting only one transition per trajectory to ensure the independence of the data. The state decomposition is non-Markovian at this lag time; therefore, the eigenvalues and eigenvectors of the transition matrix may not correspond to the true underlying alanine dynamics. However, it is still important to determine the error from sampling in the eigenvalues and eigenvectors, since these values are used in tests for Markovian behavior[5] and clustering of states.[25] Further, the primary focus in this paper is in validating the mathematical modeling of the distributions of eigenvalues and eigenvectors. The counts for this system are

$$\mathbf{Z} = \begin{bmatrix} 4380 & 153 & 15 & 2 & 0 & 0 \\ 211 & 4788 & 1 & 0 & 0 & 0 \\ 169 & 1 & 4604 & 226 & 0 & 0 \\ 3 & 13 & 158 & 4823 & 3 & 0 \\ 0 & 0 & 0 & 4 & 4978 & 18 \\ 7 & 5 & 0 & 0 & 62 & 4926 \end{bmatrix}, \tag{27}$$

and we set the prior $\alpha_{ij}=\frac{1}{6}$, as previously described.[19]

A. Eigenvalue distributions

Figure 2 shows the distributions for the five nonunit eigenvalues as calculated from the normal distribution in Eq. (20) (solid lines) and from the four sampling based methods described above. It is clear that for the fifth and sixth eigenvalues the normal distributions are excellent matches with the sampling based distributions. For the second eigenvalue, there are slight discrepancies between the Dirichlet samples and the MVN samples. For the third and fourth eigenvalues, there appear to be differences between the methods which solve for the eigenvalues directly and those which make the Taylor series approximation.

When there are multiple eigenvalues that are close in magnitude, small perturbations in the transition probabilities may result in the shifting of the rank of the eigenvalues of the corresponding perturbed matrix with respect to the original eigenvalues. Therefore, when the eigenvalues of the perturbed matrix are solved directly, one cannot simply take, for example, the second largest eigenvalue of the perturbed matrix to calculate the distribution of the eigenvalue that was second largest in the original matrix. In this system, the third and fourth eigenvalues overlap in range. In the direct solu-

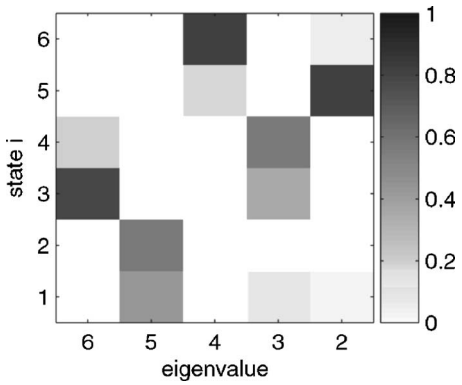

FIG. 3. The percent contribution of each state to the variance for the five nonunit eigenvectors [Eq. (25)].

tions of eigenvalues, the distribution of the third largest eigenvalue is therefore shifted to the right of the distribution of the eigenvalue ranked third in the original matrix. It is possible to match eigenvalues based on their corresponding eigenvectors, but we have not done that here. A benefit of the Taylor series approximation is that it automatically calculates deviations to the particular eigenvalue of interest, and thus is insensitive to these changes in rank.

The Taylor series expansion also immediately decomposes into contributions from the transitions out of each state, as discussed in Sec. II D. Figure 3 shows the contribution of each state to the variance in each eigenvalue (normalized such that the total contribution for each eigenvalue sums to one). These values are $\bar{q}_i$ in Eq. (25). If we wished to determine from which states to start new simulations to reduce the variance in any of the eigenvalues, we would use Eq. (26), since the expected decrease in the variance depends on the current number of samples from a given state. However, since the shooting trajectories have an equal number of samples from each state, we can use Fig. 3 to see that, for example, we should add more samples to state 5 to decrease the variance of the second eigenvalue. It would be very difficult to extract this information if one were to sample possible transition probabilities and solve each sample for the eigenvalues.

B. Eigenvector distributions

In addition to the distributions of the eigenvalues, we are also interested in the distribution of the eigenvectors. Figure 4 shows the mean and variance calculated from the closed-form distribution [Eq. (23)] and the four sampling based methods for the eigenvector components corresponding to the third (top panel) and fifth (bottom panel) eigenvalues. The inset in the top panel shows the full distributions for the second eigenvector component. Since the eigenvalues may shift in rank with the perturbations to the transition probabilities, the eigenvector component distribution generated from solving for the eigenvectors directly is clearly bimodal. Because the third and fourth eigenvalues overlap in range (as shown in Fig. 2), the eigenvectors calculated by ranking the eigenvalues in order actually correspond to different processes. As in the case with the eigenvalues, the Taylor series

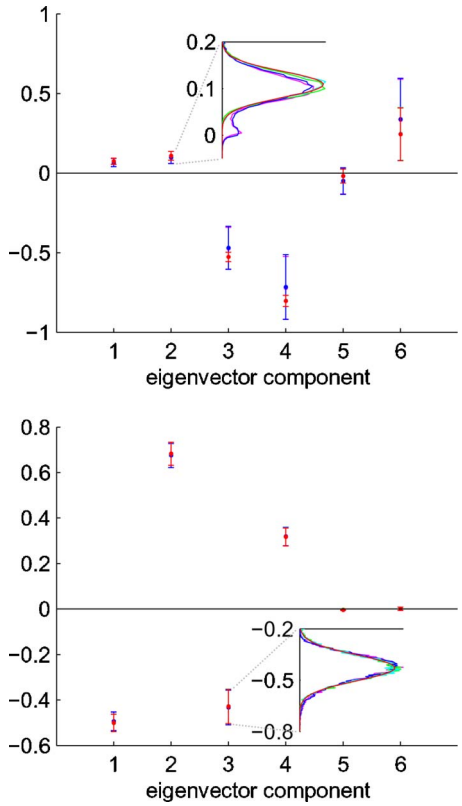

FIG. 4. (Color online) Distributions of the eigenvector components corresponding to the third (top panel) and fifth (bottom panel) eigenvalues (as calculated either from the MVN distribution or from the samples obtained by methods 1–4 described above). The insets show the actual distribution for the second eigenvector component (top inset) and third eigenvector component (bottom inset).

methods are insensitive to rank ordering changes, and calculate the true, unimodal distribution of the eigenvector components.

The fifth eigenvalue, however, is well separated from the other eigenvalues, and the full distribution of the third eigenvector component (shown in the right inset) is a good approximation of the actual distribution. The distributions of eigenvector components are not independent—they also have some covariance between them, which is not shown here. While we have only shown the mean and variances for two of the eigenvectors and the full distributions for two of the components, the results are similar across eigenvectors and components (data not shown).

The variance of each eigenvector component can again be decomposed into contributions from transitions leaving each state. Figure 5 shows this decomposition for the eigenvectors corresponding to the third (top panel) and fifth (bottom panel) eigenvalues. The values shown are the percent contribution to the sum of the variances of all the components. We can see that for the third eigenvector, the sixth component has the most variance and can be improved by adding more samples from states five and six. For the fifth eigenvector, the fifth and sixth components are quite precise, and the remaining four components depend to different degrees on the transitions from the first four states. Again, this information would be very difficult to extract by sampling from the transition matrix and solving the eigenvectors for each sample.

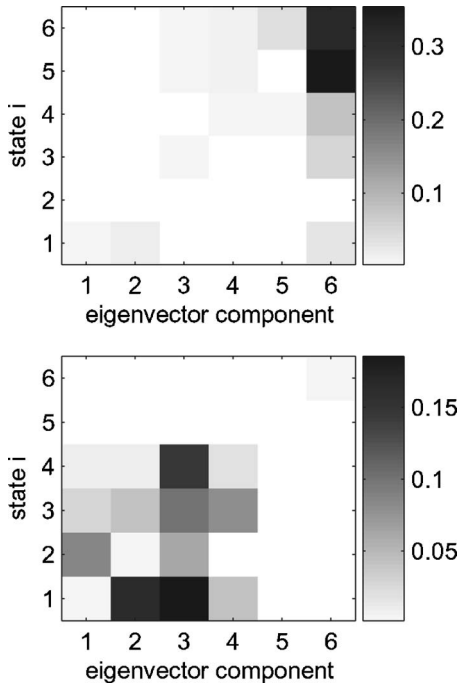

FIG. 5. The contributions to the variance of the eigenvector components as decomposed by transitions from each state. The top panel corresponds to the third eigenvalue and the bottom panel corresponds to the fifth eigenvalue.

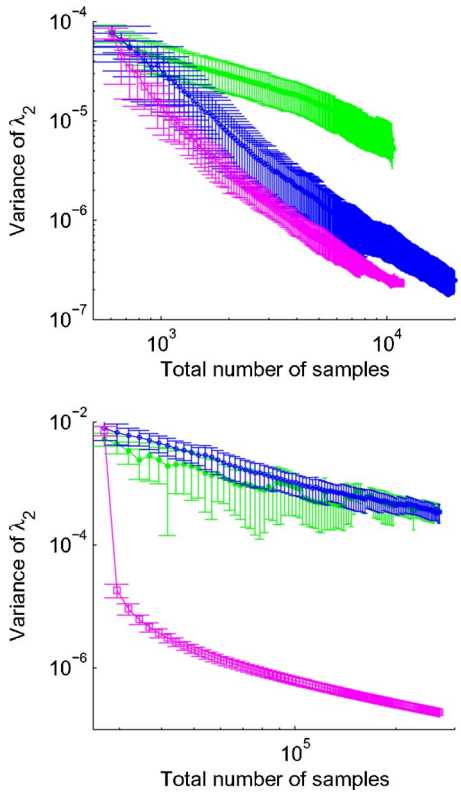

FIG. 6. (Color online) The mean and standard deviation of the variance of the largest nonunit eigenvalue of the six-state model of terminally blocked alanine peptide (top panel) and the 2454-state model of the villin headpiece (bottom panel) as a function of the total number of samples for 20 independent trials of the equilibrium (dots), even (open circles), and adaptive (squares) sampling algorithms.

C. Adaptive sampling

In addition to efficiently calculating the uncertainties in the eigenvalues and eigenvectors, we also wish to use these estimates to improve the sampling as described in Sec. II D. We compare the adaptive sampling algorithm to equilibrium sampling, where the number of trajectories initiated from each state is proportional to the equilibrium probability of the state, and even sampling, where an equal number of trajectories are initiated from each state.

Assume that we can take m transition samples in each round, we can decide where to allocate the samples before each round, and that we have a limit on the total number of samples. In the equilibrium sampling algorithm, for each new transition sample, we will choose the state from which to initiate the sample with a probability equal to the equilibrium probability of the state. The equilibrium probability of a state is the eigenvector, properly normalized, corresponding to the unit eigenvalue of the transition probability matrix, calculated from all the simulation data. The even sampling algorithm will always take the same number of samples from each state in each round, m/K. In the simplest implementation of the adaptive sampling algorithm, we calculate the contribution from each row to the variance of the quantity of interest and add all m samples to the row that will decrease the variance the most.

1. Terminally blocked alanine peptide

For the alanine peptide, we choose to adaptively sample in order to reduce the variance of the largest nonunit eigenvalue, which corresponds to transitions between states 1, 2, 3, and 4 and states 5 and 6. In general, one would probably determine some function of the variances of multiple eigenvalues to minimize. Given our dataset, we can simulate the sampling algorithms by randomly selecting trajectories without replacement from the set of shooting trajectories initiated from each state. In this way, we can estimate either how many shooting trajectories would be necessary to achieve a given precision in the eigenvalue or the possible precision with a given number of total trajectories. The transition matrix used to calculate the equilibrium probabilities of each state was the count matrix given in Eq. (27), normalized such that each row summed to one.

We began by setting the prior $\alpha_{ij}=\frac{1}{6}$ and selecting 100 trajectories at random from each of the six states. In the equilibrium sampling algorithm, for each round, the initial state for each of 60 additional trajectories were selected with a probability equal to the equilibrium probability of each state. A random trajectory from each of these states was then selected. In the even sampling algorithm, in each round ten random trajectories were added from each state. In the adaptive sampling algorithm, in each round we determined which state would decrease the variance of λ_2 the most [Eq. (26)] and added 60 random trajectories from that state. For each of the sampling algorithms, this procedure was repeated until all the trajectories from any state were selected.

Figure 6 (top panel) shows the variance of λ_2 as a function of the number of samples for the equilibrium, even, and adaptive sampling algorithms. Plotted are the mean and standard deviation of the variance of λ_2 over 20 independent

trials of each of the three algorithms. It is clear that the adaptive sampling algorithm outperforms the even sampling algorithm on average by a factor of around 5. The equilibrium sampling strategy performs very poorly for this example, with a variance of one to two orders of magnitude more than the even or adaptive sampling strategies for the same number of total simulations.

2. Villin headpiece

The adaptive sampling algorithm was also performed on a 2454-state model built from simulation data of the 36-residue alpha-helical villin headpiece. A complete description of the simulation details and Markovian state model construction is given by Jayachandran *et al.*[10] This model was constructed at a lag time of 10 ns and has been shown to reproduce the simulation data at this resolution. Because the median number of transition samples from a state was less than 500, we chose to run the sampling algorithms on the dataset *with replacement*. As in the alanine example, we compare the variance of the largest nonunit eigenvalue for the three sampling algorithms.

The prior is initialized to $\alpha_{ij}=1/2454$ and for each algorithm we began by selecting ten transition samples at random from each state (with replacement). In the equilibrium sampling algorithm, the equilibrium distribution of each state was calculated from the transition probability matrix, obtained from Jayachandran *et al.*[10] In each round, the initial state for each of 2454 additional samples were selected with a probability equal to the equilibrium probability of the state. A random transition sample from each of these states was then selected. In the even sampling algorithm, in each round, one random sample was added from each state. In the adaptive sampling algorithm, in each round we determined which state would decrease the variance of λ_2 the most [Eq. (26)] and added 2454 random transition samples from that state. For each of the sampling algorithms, this procedure was repeated for a total of 100 rounds.

Figure 6 (bottom panel) shows the mean and variance of the variance of λ_2 as a function of the total number of transition samples for 20 independent trials each of the equilibrium, even, and adaptive sampling algorithms. After a few rounds, the variance of λ_2 from the adaptive sampling algorithm is over three orders of magnitude less than the variance from the equilibrium and even sampling algorithms.

IV. DISCUSSION AND CONCLUSIONS

Given that we can generate a large number of molecular dynamics trajectories using distributed computing methods, such as Folding@Home, it is important to develop efficient techniques for analyzing the data. One compact way to model the data is to build a graph of the important states of the molecule and model kinetics as a Markov chain on this graph. In a previous work, we discussed methods for calculating the uncertainties due to finite sampling in kinetic properties such as the mean first passage time and other solutions to linear equations that can be calculated from the model.

However, many applications of the model use the eigenvalues and eigenvectors of the transition matrix, since these correspond with the aggregate rates and the participants in those rates of the system. For example, the eigenvalues directly correspond with the rates between sets of states, and thus can be compared with experiments, the eigenvalues and their implied timescales are used in tests for Markovian behavior,[5] and the eigenvectors guide a clustering of states based on kinetic similarity.[23] In all these applications, it is useful to know the uncertainty of the eigenvalues and eigenvectors, since the uncertainties may influence any decisions made with these values.

A main contribution of this work is the efficient closed-form solution of the uncertainty in the eigenvalues and eigenvectors of the transition matrix caused by finite sampling. By making two simple approximations, that the distribution of transition probabilities is well approximated by a multivariate normal distribution and that a first-order Taylor series expansion is adequate to describe the eigenvalues and eigenvectors, we have shown how to calculate the distribution of eigenvectors and eigenvalues. The closed-form solution can be calculated in roughly the same amount of time as simply solving for the eigenvalues and eigenvectors of the expected transition matrix, and therefore is much more efficient than any sampling based scheme. The method is thus scalable to large systems with many states. In addition, we may expect the transition counts to be sparse, and, as we previously discussed,[19] it is possible to leverage sparse matrix techniques to solve for a limited number of eigenvalues and eigenvectors of the system[33,34] and to update the transition counts using bordered systems.[35]

We have shown on a simple alanine peptide system that the distributions of eigenvectors and eigenvalues are in good agreement with those obtained from sampling possible transition probability matrices and solving for the eigenvalues and eigenvectors of each sample. An additional benefit of the closed-form solution is that it automatically accounts for changes in the rank of the eigenvalues due to perturbations in the transition matrix. There is no need to determine the correspondence of eigenvalues between different samples of the transition matrix. While we only presented results on this six-state system for ease of visualization, we have tested these methods on larger systems with similar results.

One downside of these methods is that they assume that the counts from state i to state j are all independently observed. However, this assumption is only used in the derivation of the posterior transition probability distribution. If we were to relax this assumption, we could still use the Taylor series approximations with samples from whatever distribution we believe the transition probabilities arise from. For example, if we have data at shorter intervals than the lag time Δt, we could use overlapping segments to calculate the transition counts. These counts would no longer be independent, but as long as a multivariate normal approximation to their distribution could be calculated, the closed-form distributions derived here could easily be modified. Some other properties, such as enforcing detailed balance, may not easily be approximated by multivariate normal distributions. In these cases, if one can generate samples from the distribution, one can substitute these into Eqs. (14) and (15) to ap-

proximate the eigenvalues and eigenvectors of interest. This is still much more efficient than solving for the eigenvalues and eigenvectors directly for each sample.

The variance of the distributions for the eigenvalues and eigenvectors easily decompose into contributions relating to the transition probabilities from each state. We showed how to use adaptive sampling techniques to leverage this information and intelligently select simulations from our data set to reduce the variance of the largest nonunit eigenvalue for the six-state alanine system and for a 2454-state model of the villin headpiece. The gain in precision or the reduction in the total number of samples for the alanine system was modest because there were only six states in the model. However, for the villin system, the gain in precision from the adaptive sampling algorithm was over three orders of magnitude. While most of the states in the villin system were moderately populated at equilibrium, the largest nonunit eigenvalue was only sensitive to the transitions from a handful of states. We fully expect to see similar benefits for other molecular systems. The ability to calculate errors in a closed-form manner allows one to easily perform adaptive sampling techniques to reduce the uncertainty. Because the required simulation time for sampling one transition is on the order of one CPU day for many protein systems, an iterative, adaptive sampling algorithm will be easy to integrate into a distributed computing framework, such as Folding@Home.

In conclusion, we have developed error analysis methods to calculate the distributions of eigenvalues and eigenvectors in a Markovian state model caused by finite sampling. We have shown that the approximate solutions are in good agreement with the actual distributions, and are computationally far more efficient. We have also shown how to perform adaptive sampling to reduce the computational cost needed to build a model with a given precision.

ACKNOWLEDGMENTS

This paper has greatly benefited from many useful discussions with Kishore Singhal. The authors would also like to thank John Chodera for providing the alanine simulation data and for helpful comments on the manuscript and Guha Jayachandran for providing the model of the villin headpiece. The authors acknowledge the support from NSF (Grant No. 03-17072) and NIH (Grant No. GM062868).

APPENDIX A: EIGENVALUE SENSITIVITY ANALYSIS

In this appendix we show how the terms in the Taylor series expansion in Eq. (14) can be computed efficiently. For details, see Vlach and Singhal.[36]

The objective is to find $\partial\lambda/\partial p_{ij}$, where λ is an eigenvalue and is defined such that

$$\det\left(\underbrace{\mathbf{P}-\lambda\mathbf{I}}_{\mathbf{A}}\right)=0. \tag{A1}$$

Consider the factors of $\mathbf{A}$,

$$\mathbf{A}=\mathbf{L}\mathbf{U}, \tag{A2}$$

where $\mathbf{L}$ is a lower triangular matrix and $\mathbf{U}$ is an upper triangular matrix with unit entries along the diagonal. The determinant of a product is the product of the determinants,

$$\det(\mathbf{A})=\det(\mathbf{L})\det(\mathbf{U}). \tag{A3}$$

The determinants of triangular matrices are simply the product of the diagonal elements. Since the matrix $\mathbf{U}$ has unit values along its diagonal, its determinant is equal to one. Thus, for the determinant of $\mathbf{A}$ to equal zero, the matrix $\mathbf{L}$ must have a zero element along its diagonal. Assume that this zero element is in the last row: $l_{KK}=0$ (partial or full pivoting may be needed to ensure this[36]).

We can relate the partial derivative of λ to the derivative of l_{KK} using the chain rule,

$$\frac{dl_{KK}}{dp_{ij}}=\frac{\partial l_{KK}}{\partial\lambda}\frac{\partial\lambda}{\partial p_{ij}}+\frac{\partial l_{KK}}{\partial p_{ij}}=0, \tag{A4}$$

where the derivative must equal zero since the value of l_{KK} is fixed at zero for λ to be an eigenvalue. To find the terms $\partial l_{KK}/\partial\lambda$ and $\partial l_{KK}/\partial p_{ij}$, we differentiate Eq. (A2) above by a general parameter h,

$$\frac{\partial\mathbf{A}}{\partial h}=\frac{\partial\mathbf{L}}{\partial h}\mathbf{U}+\mathbf{L}\frac{\partial\mathbf{U}}{\partial h}. \tag{A5}$$

We define vectors $\mathbf{x}$ and $\mathbf{x^a}$ as follows:

$$\mathbf{U}\mathbf{x}=\mathbf{e_K},$$

$$\mathbf{L^T}\mathbf{x^a}=\mathbf{0}, \tag{A6}$$

where $\mathbf{e_K}$ is the column vector corresponding to the Kth column of the identity matrix. We force a nontrivial solution for $\mathbf{x^a}$ by setting

$$x_K^a=1. \tag{A7}$$

We pre- and postmultiply Eq. (A5) by these vectors:

$$(\mathbf{x^a})^\mathbf{T}\frac{\partial\mathbf{A}}{\partial h}\mathbf{x}=(\mathbf{x^a})^\mathbf{T}\frac{\partial\mathbf{L}}{\partial h}\mathbf{U}\mathbf{x}+(\mathbf{x^a})^\mathbf{T}\mathbf{L}\frac{\partial\mathbf{U}}{\partial h}\mathbf{x}. \tag{A8}$$

Substituting the definitions in Eq. (A6) into Eq. (A8) gives

$$(\mathbf{x^a})^\mathbf{T}\frac{\partial\mathbf{A}}{\partial h}\mathbf{x}=(\mathbf{x^a})^\mathbf{T}\frac{\partial\mathbf{L}}{\partial h}\mathbf{e_K}+\mathbf{0}^\mathbf{T}\frac{\partial\mathbf{U}}{\partial h}\mathbf{x}. \tag{A9}$$

The first term on the right hand side can be reduced since $\mathbf{L}$ is a lower triangular matrix. Postmultiplying its derivative by $\mathbf{e_K}$ gives a vector in which all entries are zero, except the last one, which is $\partial l_{KK}/\partial h$. We premultiply by $(\mathbf{x^a})^\mathbf{T}$, which gives $x_K^a\partial l_{KK}/\partial h=\partial l_{KK}/\partial h$, since we defined $x_K^a=1$ in Eq. (A7). The second term is equal to zero. Therefore, Eq. (A9) can be rewritten as

$$(\mathbf{x^a})^\mathbf{T}\frac{\partial\mathbf{A}}{\partial h}\mathbf{x}=\frac{\partial l_{KK}}{\partial h}. \tag{A10}$$

We can now calculate the remaining terms in Eq. (A4) by setting h in Eq. (A10) equal to either λ or p_{ij},

$$\left((\mathbf{x^a})^{\mathbf{T}}\frac{\partial \mathbf{A}}{\partial \lambda}\mathbf{x}\right)\frac{\partial \lambda}{\partial p_{ij}}+\left((\mathbf{x^a})^{\mathbf{T}}\frac{\partial \mathbf{A}}{\partial p_{ij}}\mathbf{x}\right)=0. \tag{A11}$$

Matrix $\mathbf{A}$ is defined in Eq. (2), and it is easy to see that $\partial\mathbf{A}/\partial\lambda=-\mathbf{I}$ and $\partial\mathbf{A}/\partial p_{ij}=\mathbf{e_i}\mathbf{e_j}^{\mathbf{T}}$. Therefore,

$$\frac{\partial \lambda}{\partial p_{ij}}=\frac{(\mathbf{x^a})^{\mathbf{T}}\mathbf{e_i}\mathbf{e_j}^{\mathbf{T}}\mathbf{x}}{(\mathbf{x^a})^{\mathbf{T}}\mathbf{x}}=\frac{x_i^a x_j}{(\mathbf{x^a})^{\mathbf{T}}\mathbf{x}}. \tag{A12}$$

We wish to evaluate $\partial\lambda/\partial p_{ij}$ terms at the matrix $\bar{\mathbf{A}}$, which corresponds to the expected values of the parameters. The determination of the vectors $\bar{\mathbf{x}}=\mathbf{x}|_{\bar{\mathbf{A}}}$ and $\bar{\mathbf{x}}^{\mathbf{a}}=\mathbf{x}^{\mathbf{a}}|_{\bar{\mathbf{A}}}$ involves decomposing the matrix $\bar{\mathbf{A}}$ into factors $\bar{\mathbf{L}}$ and $\bar{\mathbf{U}}$ and then solving the following sets of linear equations,

$$\bar{\mathbf{U}}\bar{\mathbf{x}}=\mathbf{e_K}, \tag{A13}$$

$$\bar{\mathbf{L}}^{\mathbf{T}}\bar{\mathbf{x}}^{\mathbf{a}}=\mathbf{0}. \tag{A14}$$

We can then find all the $\partial\lambda/\partial p_{ij}|_{\bar{\mathbf{A}}}$ terms by simply normalizing the outer product of $\bar{\mathbf{x}}^{\mathbf{a}}$ and $\bar{\mathbf{x}}:\bar{\mathbf{x}}^{\mathbf{a}}\bar{\mathbf{x}}^{\mathbf{T}}/(\bar{\mathbf{x}}^{\mathbf{a}})^{\mathbf{T}}\bar{\mathbf{x}}$.

Factoring $\bar{\mathbf{A}}$ into its **LU** factors takes time $O(\frac{1}{3}K^3)$. The solutions of Eqs. (A13) and (A14) take time $O(K^2)$ for forward or backward substitution. All the terms $\partial\lambda/\partial p_{ij}|_{\bar{\mathbf{A}}}$ can be computed from these two solutions in $O(K^2)$ operations. This must be done independently for each eigenvalue for which the uncertainty is desired.

APPENDIX B: EIGENVECTOR SENSITIVITY ANALYSIS

In this appendix, we show how the partial derivative terms in Eq. (15) can be computed efficiently. We start with the following eigenvector equation [Eq. (3)]:

$$\mathbf{A}\mathbf{v}_\lambda=\mathbf{0}.$$

To differentiate this equation with respect to a parameter p_{ij}, we must use the chain rule, since $\mathbf{A}$ is a function of p_{ij} and λ, and λ is a function of p_{ij}.

$$\left(\frac{\partial \mathbf{A}}{\partial p_{ij}}+\frac{\partial \mathbf{A}}{\partial \lambda}\frac{\partial \lambda}{\partial p_{ij}}\right)\mathbf{v}_\lambda+\mathbf{A}\frac{\partial \mathbf{v}_\lambda}{\partial p_{ij}}=\mathbf{0}. \tag{B1}$$

From Eq. (2), $\partial\mathbf{A}/\partial\lambda=-\mathbf{I}$, $\partial\mathbf{A}/\partial p_{ij}=\mathbf{e_i}\mathbf{e_j}^{\mathbf{T}}$, and $\partial\lambda/\partial p_{ij}=x_i^a x_j/(\mathbf{x^a})^{\mathrm{T}}\mathbf{x}$, as derived in Appendix A, along with the definitions for $\mathbf{x}$ and $\mathbf{x^a}$. We therefore have the system of linear equations,

$$\mathbf{A}\frac{\partial \mathbf{v}_\lambda}{\partial p_{ij}}=-\left(\frac{\partial \mathbf{A}}{\partial p_{ij}}-\frac{\partial \lambda}{\partial p_{ij}}\mathbf{I}\right)\mathbf{v}_\lambda, \tag{B2}$$

where all the terms on the right hand side are known. We wish to evaluate the partial derivative terms at the matrix $\bar{\mathbf{A}}$, which corresponds to the expected values of the parameters,

$$\bar{\mathbf{A}}\frac{\partial \mathbf{v}_\lambda}{\partial p_{ij}}\bigg|_{\bar{\mathbf{A}}}=-\left(\frac{\partial \mathbf{A}}{\partial p_{ij}}-\frac{\partial \lambda}{\partial p_{ij}}\mathbf{I}\right)\bar{\mathbf{v}}_\lambda. \tag{B3}$$

However, the matrix $\bar{\mathbf{A}}$ is singular, so we must enforce one additional constraint. As eigenvectors are only determined to within a constant factor, they are often normalized such that their magnitude is equal to one.

$$(\mathbf{v}_\lambda)^{\mathbf{T}}\mathbf{v}_\lambda=1. \tag{B4}$$

Differentiating this constraint gives

$$2(\mathbf{v}_\lambda)^{\mathbf{T}}\frac{\partial \mathbf{v}_\lambda}{\partial p_{ij}}=0. \tag{B5}$$

Combining this constraint with Eq. (B3) gives.

$$\begin{bmatrix}\bar{\mathbf{A}}\\(\bar{\mathbf{v}}_\lambda)^{\mathbf{T}}\end{bmatrix}\cdot\frac{\partial \mathbf{v}_\lambda}{\partial p_{ij}}=\begin{bmatrix}-\left(\left(\partial\mathbf{A}/\partial p_{ij}\right)-\left(\partial\lambda/\partial p_{ij}\right)\mathbf{I}\right)\bar{\mathbf{v}}_\lambda\\0\end{bmatrix}. \tag{B6}$$

Equation (B6) can be separated into two parts based on the terms on the right hand side. Substituting in the values for the partial derivatives of $\mathbf{A}$ and λ, we get

$$\begin{bmatrix}\bar{\mathbf{A}}\\(\bar{\mathbf{v}}_\lambda)^{\mathbf{T}}\end{bmatrix}\cdot\mathbf{b_{ij}^1}=\begin{bmatrix}-\mathbf{e_i}(\mathbf{e_j})^{\mathbf{T}}\bar{\mathbf{v}}_\lambda\\0\end{bmatrix},$$

$$\begin{bmatrix}\bar{\mathbf{A}}\\(\bar{\mathbf{v}}_\lambda)^{\mathbf{T}}\end{bmatrix}\cdot\mathbf{b_{ij}^2}=\begin{bmatrix}(x_i^a x_j/(\mathbf{x^a})^{\mathbf{T}}\mathbf{x})\bar{\mathbf{v}}_\lambda\\0\end{bmatrix}, \tag{B7}$$

$$\frac{\partial \mathbf{v}_\lambda}{\partial p_{ij}}=\mathbf{b_{ij}^1}+\mathbf{b_{ij}^2}.$$

Simple algebra shows that one can introduce the K vectors $\mathbf{c_i}$ and the vector $\mathbf{d}$ and solve the following sets of equations:

$$\begin{bmatrix}\bar{\mathbf{A}}\\(\bar{\mathbf{v}}_\lambda)^{\mathbf{T}}\end{bmatrix}\cdot\mathbf{c_i}=\begin{bmatrix}-\mathbf{e_i}\\0\end{bmatrix}\quad\forall\, i,$$

$$\begin{bmatrix}\bar{\mathbf{A}}\\(\bar{\mathbf{v}}_\lambda)^{\mathbf{T}}\end{bmatrix}\cdot\mathbf{d}=\begin{bmatrix}\bar{\mathbf{v}}_\lambda\\0\end{bmatrix}, \tag{B8}$$

and then compute

$$\frac{\partial \mathbf{v}_\lambda}{\partial p_{ij}}=\frac{\mathbf{c_i}}{(v_\lambda)_j}+\left(\frac{(\mathbf{x^a})^{\mathbf{T}}\mathbf{x}}{x_i^a x_j}\right)\mathbf{d}. \tag{B9}$$

We can solve each of the systems of equations in Eq. (B8) by augmenting the **LU** factors of $\bar{\mathbf{A}}$, calculated to determine the sensitivity of the eigenvalue λ, in time $O(K^2)$. Since there are $K+1$ equations, the total time to calculate all the sensitivity terms is $O(K^3)$.

[1] M. Shirts and V. Pande, Science **290**, 1903 (2000).
[2] M. E. Karpen, D. J. Tobias, and C. L. Brooks, Biochemistry **32**, 412 (1993).
[3] H. Grubmuller and P. Tavan, J. Chem. Phys. **101**, 5047 (1994).
[4] N. Singhal, C. Snow, and V. S. Pande, J. Chem. Phys. **121**, 415 (2004).
[5] W. C. Swope, J. W. Pitera, and F. Suits, J. Phys. Chem. B **108**, 6571 (2004).
[6] W. C. Swope, J. W. Pitera, F. Suits *et al.*, J. Phys. Chem. B **108**, 6582 (2004).
[7] S. Sriraman, I. G. Kevrekidis, and G. Hummer, J. Phys. Chem. B **109**, 6479 (2005).
[8] M. Andrec, A. K. Felts, E. Gallicchio, and R. M. Levy, Proc. Natl. Acad. Sci. U.S.A. **102**, 6801 (2005).
[9] V. Schultheis, T. Hirschberger, H. Carstens, and P. Tavan, J. Chem. Theory Comput. **1**, 515 (2005).
[10] G. Jayachandran, V. Vishal, and V. S. Pande, J. Chem. Phys. **124**, 164902 (2006).

[11] S. P. Elmer and V. S. Pande, J. Chem. Phys. **121**, 12760 (2004).
[12] S. P. Elmer, S. Park, and V. S. Pande, J. Chem. Phys. **123**, 114902 (2005).
[13] P. Kasson, N. W. Kelley, N. Singhal, M. Vrljic, A. T. Brunger, and V. S. Pande, Proc. Natl. Acad. Sci. U.S.A. **103**, 11916 (2006).
[14] D. Chandler, J. Chem. Phys. **68**, 2959 (1978).
[15] J. E. Adams and J. D. Doll, Surf. Sci. **111**, 492 (1981).
[16] A. F. Voter and J. D. Doll, J. Chem. Phys. **82**, 80 (1985).
[17] M. Weber and T. Galliat, Konrad Zuse Zentrum Berlin Report No. 02-12, 2002 (unpublished).
[18] M. Weber, Ph.D. thesis, Konrad Zuse Zentrum Berlin, Berlin, Germany, 2006.
[19] N. Singhal and V. S. Pande, J. Chem. Phys. **123**, 204909 (2005).
[20] C. Schutte, Ph.D. thesis, Konrad Zuse Zentrum Berlin, Berlin, Germany, 1999.
[21] C. Schutte, A. Fischer, W. Huisinga, and P. Deuflhard, J. Comput. Phys. **151**, 146 (1999).
[22] W. Huisinga, Ph.D. thesis, Free University of Berlin, Berlin, Germany, 2001.
[23] C. Schutte and W. Huisinga, in *Handbook of Numerical Analysis: Special Volume on Computational Chemistry*, Vol. X, edited by P. G. Ciaret and J.-L. Lions (Elsevier, New York, 2002).
[24] P. Deuflhard and M. Weber, Numer. Linear Algebra Appl. **398**, 161 (2005).
[25] J. D. Chodera, N. Singhal, V. S. Pande, K. A. Dill, and W. C. Swope, J. Chem. Phys. **126**, 155101 (2007).
[26] F. Noe, I. Horenko, C. Schutte, and J. C. Smith, J. Chem. Phys. **126**, 155102 (2007).
[27] N. L. Johnson, S. Kotz, and N. Balakrishnan, *Discrete Multivariate Distributions* (Wiley, New York, 1997).
[28] S. Kotz, N. Balakrishnan, and N. L. Johnson, *Continuous Multivariate Distributions* (Wiley, New York, 2000).
[29] G. Golub and C. van Loan, *Matrix Computations*, 3rd ed. (The Johns Hopkins University Press, London, 1996).
[30] C. R. Rao, *Linear Statistical Inference and its Applications*, 2nd ed. (Wiley, New York, 1973).
[31] If the distribution of multiple eigenvalues is desired, it is possible to group terms in Eq. (14) similarly to how we group terms for the eigenvectors in Eq. (15) to find the covariance matrix between the eigenvalues.
[32] J. D. Chodera, W. C. Swope, J. W. Pitera, and K. A. Dill, Multiscale Model. Simul. **5**, 1214 (2006).
[33] A. Ruhe, SIAM J. Sci. Comput. (USA) **19**, 1535 (1998).
[34] R. B. Lehoucq, D. C. Sorensen, and C. Yang, *ARPACK Users' Guide: Solution of Large-Scale Eigenvalue Problems with Implicitly Restarted Arnoldi Methods* (SIAM Publications, Philadelphia, 1998).
[35] J. R. Bunch and D. J. Rose, J. Math. Anal. Appl. **48**, 574 (1974).
[36] J. Vlach and K. Singhal, *Computer Methods for Circuit Analysis and Design* (Van Nostrand Reinhold, New York, 1983).