Adaptive Markov State Model Generation using Multi-State Transition Interface Sampling

Draft

Amsterdam, July 2014

Jan-Hendrik Prinz
MSKCC, New York
jan.prinz@choderalab.org

The application of small-sized Markov state models (MSMs) has become a standard tool for the analysis of the dynamics in molecular dynamics (MD) simulations and computation of key properties. Here we present a rigorous approach to enhance the generation of a MSM for a system of biomolecules using multi-state transition interface sampling (MSTIS)

Authors

Jan-Hendrik Prinz

Memorial Sloan-Kettering Cancer Center, Zuckerman Research Institute, Zuckerman Research Center, 417 East 68th Street, New York, NY 10065, US

jan.prinz@choderalab.org

Weina Du

University of Amsterdam, The Netherland

weina.du@...

Peter Bolhuis

bolhuis@...

University of Amsterdam, The Netherland

John D. Chodera

Memorial Sloan-Kettering Cancer Center, Zuckerman Research Institute, Zuckerman Research Center, 417 East 68th Street, New York, NY 10065, US

john.chodera@choderalab.org

Keywords / PACS

Time Series Analysis, Timescale Estimation, Markov State Models, Implied Timescales

Introduction

[1]In recent years Markov State Models (MSMs) have matured into a useful tool for the descriptions of dynamics that shows stochastic, but metastable behavior on the timescales of interest [Prinz et al., 2011]. This includes protein dynamics, chemical [cite], physical systems [cite], etc. Its application spans dynamics that can be written using a differential equation that is first order in time and has also been applied for various types of dynamics, e.g. Brownian dynamics, Langevin dynamics, etc... Once a MSM has been succesfully parametrized from data it allows to quickly compute a large variety of target properties like relaxation timescales, mean first passage times, metastable subsets [Röblitz and Weber, 2013, Deuflhard and Weber, 2003] and path related properties like reactive pathways and committor probabilities [Vanden-Eijnden, 2006, Metzner et al., 2009]. The usage of conditional jump probabilities also allows to effectively split the generation of necessary statistics into independent subproblems that require only local equilibrium and can be thus be parallelized as used in adaptive sampling techniques.

Transition Interface Sampling from

[1] MSM Introduction from PMM Theory Draft

General Considerations

TIS uses interfaces

We want to find core sets and regions around it. For a single state it would be best to use the mean first passage time to

Interface definition

We want to define interfaces of a more complex shape than just based on a metric based distance to a cluster center in some space. The Voronoi cell discretization provides a natural way to do this. We just need to define a cutoff for the Voronoi cells to ensure that the interface is bounded. For the innermost interface we do exactly this. Find generators for a Voronoi tesselation with a cut-off ρ and then decide which cell should belong to the inner region of which the boundary defines the interface. After this decision we need to keep all points that are within a 2ρ radius since only these points can contribute to the shape of the interface.

For the next one we do exactly the same thing but allow only generators outside the core region. And then we cluster all points that are outside the previously generated interface using these generators with a new cut-off. And so on for the next interfaces. This will produce a nested set of interfaces.

Voronoi Problems

We need to be aware that we 'cannot'[2] compute the size of each Voronoi cell. At least we do not want to do so and this means that the stationary distribution π contains the unknown factor of the phase space volume of its associated cell. Although the stationary distribution is correct in the sense that it discribes the probability for the process to be in the cell the value can e.g. be large either because the probability density is high or the cell is large and is thus problematic as a measure of importance in terms of a macro state representative or cluster center.

[2] 'cannot' means that this operation is roughly of the order $\mathcal{O}(N^{\dim/2})$ and thus *very* expensive in high dimensions.

Alternatively we will present a way to use average time to reach a state when started from equilibrium or maybe the minimal roundtrip time as a replacement. This is based on the fact that the most stable center to the state with the highest potential energy has the tendency that it is easiest to reach or return to that state compared to other ones [maybe cite Bovier (metastability)]. Eventually we can get a way to estimate the size of the cell from the dynamics.

This problem also means that the left eigenvectors computed suffer from the same problem. Especially if in a dense region we find many cluster centers the probability density might be misleading and so the left eigenvectors

Distance Ideas from Transition matrices

Commute Distance

The average roundtrip time between states i and j

$$\mathrm{cd}\left[i \leftrightarrow j\right] = \left(\mathbf{e}_i - \mathbf{e}_j\right)^{\mathrm{T}} \left(\mathbf{\Pi}\left(\mathrm{Id} - \mathbf{T}\right)\right)^{\text{-}} \left(\mathbf{e}_i - \mathbf{e}_j\right) \tag{1}$$

Mean First Passage Time

Definition of $\mathbf{A}$ and $\mathbf{W}$

$$\mathbf{A} \equiv \mathrm{Id} - \mathbf{T} \tag{2}$$

$$\begin{aligned} \mathbf{W} &\equiv \mathrm{Id} - \mathbf{A}\mathbf{A}^{\#} \\ &= \mathbf{J}\mathbf{\Pi} \end{aligned}$$

Generlized inverse (Use the Jordan Decomposition and use the pseudo inverse only on the diagonal part

$$\begin{aligned}\mathbf{A}^{\#} &\equiv \mathrm{JE}[\mathbf{A}]\mathrm{JD}\left[\mathbf{A}\right]^{\dagger}\mathrm{JE}[\mathbf{A}]^{-1} \\ &= (\mathbf{A}+\mathbf{W})^{-1}-\mathbf{W} \\ &= (\mathrm{Id}-\mathbf{T}+\mathbf{J}\boldsymbol{\Pi})^{-1}-\mathbf{J}\boldsymbol{\Pi}\end{aligned}$$

The matrix $\mathbf{A}^{\#}$ can be interpreted as a list of vectors that contain the expected difference in counts between timeseries started from equilibrium and timeseries started in each state.

$$\mathbf{A}^{\#}_{j} = \sum_{k=0}^{\infty}\left(\mathbf{e}_j^{\mathrm{T}}\mathbf{T}^k - \boldsymbol{\pi}^{\mathrm{T}}\right) \tag{3}$$

The indexing with dg, like $\mathbf{A}_{dg}$, means to take matrix $\mathbf{A}$ and set all off-diagonal entries to zero. Mean First Passage Time

$$\begin{aligned}\mathbf{M} = \mathrm{mfpt}\left[i\to j\right] &= \mathbf{A}_{\mathbf{x}|x_0=i}\left[\min_{n\geq 1} x_n = j\right] \\ &= \left(\mathrm{Id}-\mathbf{A}^{\#}+\mathbf{J}\mathbf{A}^{\#}_{dg}\right)\boldsymbol{\Pi}^{-1} \\ &= \left(\mathrm{Id}-(\mathbf{A}+\mathbf{W})^{-1}+\mathbf{W}+\mathbf{J}\mathbf{A}^{\#}_{dg}\right)\boldsymbol{\Pi}^{-1} \\ &= \left(\mathrm{Id}-(\mathbf{A}+\mathbf{J}\boldsymbol{\Pi})^{-1}+\mathbf{J}\boldsymbol{\Pi}+\mathbf{J}\mathbf{A}^{\#}_{dg}\right)\boldsymbol{\Pi}^{-1}\end{aligned}$$

This can be used to compute the expected time to reach a state when started in equilibrium (first discovery time)

$$\begin{aligned}\mathrm{fdt}[k] &= \mathbf{A}_{\mathbf{x}|x_0\sim\pi}\left[\min_{n} x_n = k\right] \\ &= \left(\boldsymbol{\pi}^{\mathrm{T}}\mathbf{M}\right)_k \\ &= 1+\frac{\mathbf{A}^{\#}_{kk}}{\pi_k}\end{aligned}$$

We can also compute the variance in the mean first passage time by

$$\begin{aligned}\mathbf{V}_{ij} &= \mathrm{vmfpt}[i\to j] = \mathbf{A}_{\mathbf{x}|x_0=i}\left[\left(\min_{n\geq 1} x_n = j-\mathbf{M}_{ij}\right)^2\right] \\ \mathbf{V}_{ij} &= \mathbf{B}_{ij}-\left(\mathbf{M}_{ij}\right)^2\end{aligned}$$

with the squared fluctuations given by

$$\mathbf{B} = \mathbf{M}\left(2\mathbf{A}^{\#}_{dg}\mathbf{D}+\mathrm{Id}\right)+2\left(\mathbf{A}^{\#}\mathbf{M}-\mathbf{J}(\mathbf{A}^{\#}\mathbf{M})_{dg}\right) \tag{4}$$

which can be simplified for the diagonal elements

$$\mathbf{B}_{dg} = 2\mathbf{D}\mathbf{A}^{\#}_{dg}\mathbf{D}+\mathbf{D} \tag{5}$$

I assume we could use this to check how fast a certain region might be explored up to a certain percentage.

Multi-state Committor

Probability to enter a state i before visiting any other state $j \neq i$.

Using OOM we can also change the expression for the committor to trajectories of a maximal length which is very useful. See Multi-State MFPT for the extension to the distribution for a trajectory length k.

This directly gives also the probability to not hit any core!

Assume that we have

$$\mathbf{T}_V = \mathrm{diag}\left(\{\begin{cases}1 & \text{if } i \in \mathbf{V} \\ 0 & \text{else}\end{cases} \mid i \in \Omega\}\right) \mathbf{T} \tag{6}$$

which is the the transition matrix with all columns set to zero that belong to states that are not in $\mathbf{V}$ the set of states not in a core, also referred to as the *void*.

We then split the rank deficient matrix into eigenvalues and left and right eigenvectors like

$$\mathbf{T}_V = \mathbf{R}_V \,\mathrm{diag}(\lambda_V)\mathbf{L}_V \tag{7}$$

then the max length committor is given by

$$C_k = \mathbf{R}_V \,\mathrm{diag}(\frac{\lambda_V^{len} - 1}{\lambda_V - 1})\mathbf{L}_V \mathbf{T}_{:,k} \tag{8}$$

with *len* being the maximal length of the trajectory and k the core state selected. The probabilities to not reach any core is given by

$$p_{not} = 1 - \sum_k C_k \tag{9}$$

We can also get the probability to reach a state in exactly n states by

$$p_k(n) = \mathbf{R}_V \,\mathrm{diag}(\lambda_V^n)\mathbf{L}_V \mathbf{T}_{:,k} \tag{10}$$

with this object we can write the committor as

$$C_k = \sum_{n=0}^{len} p_k(n) \tag{11}$$

Multi-state MFPT

Average time to visit a state i before visiting any other state $j \neq i$.

$$\mathbf{M} = \mathrm{cmfpt}\,[i \to j \mid \neg V] = \mathbf{A}_{\mathbf{x}|x_0=i, x_i \notin V}\left[\min_{n \geq 1} x_n = j\right] \tag{12}$$

Using OOM the expressions can be written down easily. I have not yet found a way to reduce the complexity by one order, but it works.

One can also get the distribution of times since OOMs allow to specify which type of trajectories should be used, e.g. we can ask what is the average time to visit state i among all trajectories that do not end up prior in another core state j and have a maximal length of t_{max}.

From the complete distribution we can also compute higher moments like variances of the *mfpt* or *confidence intervals*. We compute

$$\begin{aligned} m_k(n) &= C_k^{-1} \sum_{n=0}^{len} (1+k) p_k(n) \\ &= C_k^{-1} \mathbf{R}_V \operatorname{diag}\left(\frac{1-(n+2)\lambda_V^{n+1}+(n+1)\lambda_V^{n+2}}{(1-\lambda_V)^2}\right) \mathbf{L}_V \mathbf{T}_{:,k} \end{aligned}$$

which simplifies to

$$m_k(\infty) = C_k^{-1} \mathbf{R}_V \operatorname{diag}((1-\lambda_V)^{-2}) \mathbf{L}_V \mathbf{T}_{:,k}$$

and in the case, where we have only one core this should reduce to the known mfpt as

$$m_k(\infty) = \mathbf{R}_V \operatorname{diag}((1-\lambda_V)^{-2}) \mathbf{L}_V \mathbf{T}_{:,k}$$

which provides an alternative to the previously given expressions to compute the mfpt. Similarly we can derive also variances in the mfpt

$$\begin{aligned} v_k(n) &= C_k^{-1} \sum_{n=0}^{len} (1+k)^2 p_k(n) \\ &= C_k^{-1} \mathbf{R}_V \operatorname{diag}\left(\frac{-\left(2n^2+6n+3\right)\lambda_V^{n+2}+(n+2)^2\lambda_V^{n+1}+(n+1)^2\lambda_V^{n+3}-\lambda_V-1}{(\lambda_V-1)^3}\right) \mathbf{L}_V \mathbf{T}_{:,k} \end{aligned}$$

The most complex operation is the eigendecomposition which has to be done once for a fixed set of core sets. The rest is only matrix multiplications.

Example

[fill]

Conclusion

[fill]

Acknowledgements

Thank you's

References

P. Deuflhard and M. Weber. Robust Perron cluster analysis in conformation dynamics. *ZIB Report*, 09, 2003.

P. Metzner, C. Schütte, and E. Vanden-Eijnden. Transition path theory for Markov jump processes. *Multiscale Model. Sim.*, 7(3): 1192–1219, 2009. DOI: 10.1137/070699500.

J.-H. Prinz, H. Wu, M. Sarich, B. G. Keller, M. Senne, M. Held, J. D. Chodera, C. Schütte, and F. Noé. Markov models of molecular kinetics: generation and validation. *J. Chem. Phys.*, 134(17):174105, May 2011. DOI: 10.1063/1.3565032.

S. Röblitz and M. Weber. Fuzzy spectral clustering by PCCA+: application to Markov state models and data classification. *Adv Data Anal Classif*, 7(2):147–179, May 2013. DOI: 10.1007/s11634-013-0134-6.

E. Vanden-Eijnden. *Transition Path Theory*, volume 703. Springer Verlag, 2006. DOI: 10.1007/3-540-35273-2.