Clark Center S295
Stanford University
Stanford, CA, 94305

February 6, 2014

Prof. Sharon Hammes-Schiffer
Deputy Editor of JPC B
The Journal of Physical Chemistry

Editor:

Reviewer 1

> In "Optimal parameter selection in Markov state models for biomolecular conformational dynamics", the authors elaborate on an earlier likelihood-based approach to the selection of the appropriate number of discrete states for crisp partitionings to attempt to obtain a Markov model that balances accuracy with the potential for overfitting that many-state models are capable of. Three different critera are examined: a leave-some-out cross-validation likelihood estimate, as well as two criteria that penalize overfitting by using assumptions of asymptotic normality of the likelihood function (AIC and BIC). Two different "emission probabilities" for the probability of observing a sample x_t from a given discrete state s are considered that would produce a generative model—one that can be used to assign new data to a model: a uniform probability and Gaussian probability. These are compared for consistency with each other, as well as compared with well-established metrics such as the stability of implied timescales of the model with the number of states.
>
> This manuscript will be a useful contribution to the literature on Markov state model construction, but has a number of deficiencies that hinder publication in its current form. Once these are addressed, it could be suitable for publication in JPC B.
>
> Most critically, the manuscript does not clearly explain the described work, fully justify the conclusions, or provide sufficient detail to reproduce the results. A number of points below help address these deficiencies:
>
> The title "Optimal parameter selection..." does not reflect the work contained in the paper, which focuses only on the selection of the number of states. Should the authors which to keep the general title, more work on selecting other kinds of parameters should be included;

otherwise, the title should be amended to something like "Optimal selection of the number of states... using a likelihood approach..."

"As such, the eigenvalues, and correspondingly the relaxation timescales, should remain constant as the lag time increases." The eigenvalues of which matrix should remain constant with what parameter? The eigenvalues of T(t) actually change with t. Perhaps you mean the eigenvalues of some implied rate matrix K where T(t) = exp[Kt]?

"choosing models whose relaxation timescales have converged" Give definition for "relaxation timescales".

"With s(x) as the function mapping conformations into the indicator function basis set" Unclear: do you mean s(x) is the integer index of the basis function that is nonzero (has support) at x? Please be precise. The lack of precision in this manuscript substantially hampers readability and comprehension.

Eq. 1: This neglects a term for the initial conditions x_0. Eq. 1 actually refers to the conditional probability $P[x_t|x_0]$. Is this the quantity you intend to work with, rather than $P[x_t]$? If so, why?

"By convention, we define $P(x|s(x))$ to be zero for all x such that $s(x) \neq s_i$" Be precise! This statement is meaningless. Do you mean "we define $P(x|i) = 0$ whenever $s(x) \neq i$?"

"We propose two possible emission distributions, $P(x|s(x))$" Write this as $P(x|s)$ or $P(x|i)$, to make it clear that your generative probability model MUST be a probability over configurations x given a discrete integer index s or i.

Eq. 3: Rewrite as $P(x|s)$ instead of $P(x_t|s(x_t))$, since these are supposed to be generative distributions for x given state index s.

"A more tractable alternative is a Gaussian emission model..." Why is this even anticipated to be a good model? Won't it force your states to have gaussianlike clusters in them? And isn't a gaussian infinite in support, while you SPECIFICALLY REQUIRE the state membership functions to have finite support because they form a crisp partitioning? Why are you not truncating this model at the state boundaries, which would affect your normalization of the probability density in Eq 5. This inconsistency seems like a major flaw in this work.

"We assume that the overlaps are small..." Is this enforced by some condition on sigma and mu for neighboring states? Has this assump-

tion been checked? This is certainly, absolutely, positively not the case for even the simple model in Fig 3A.

On choice of emission probabilities, an obvious question: Why not choose $P(x|i) \propto exp[-\beta U(x)]\delta_{i,s(x)}$, where $U(x)$ is the potential energy and β the inverse temperature? The normalization constants are ALSO difficult to compute, but have the advantage that they are proportional to the stationary probability elements computed from the transition matrix T.

"Unfortunately such approaches are intractable for problems of the size encountered in biomolecular simulations because of the need to integrate over all possible Markov models with a given number of parameters." Is this really intractable? That's a very strong claim. Perhaps it is just difficult or complex, since no evidence of intractability is presented?

"N is the number of data points, assumed to be independent and identically distributed" How reasonable is this assumption for molecular simulation data? How do you even quantify the number of "iid normally distributed" data points in a molecular simulation?

"For cross validation, we find it essential to use a prior..." Give complete equation for log likelihood. How do you initialize and sample posterior? Or maximize likelihood?

"We simulated two trajectories with ... Langevin dynamics..." Which langevin integration scheme was used! How were the initial conditions selected?

* "The first trajectory was clustered using the k-centers clustering algorithm..." Cite. What parameters were used for clustering?

"The volumes did not change drastically..." What does this mean, quantitatively?

"...the likelihood remained essentially constant..." What does this mean, quantitatively?

according to the three methods" Which three methods exactly?

Fig. 2: This figure is way too small. What is "score"? Is this a log likelihood? Is lower or higher score representative of lower or higher likelihood? The numbers/units of likelihood are actually meaningful, and should be shown. What is the difference between "volume" and "Gaussian"?

Fig. 2 caption: "form the boundaries of the optimal window" Why is this region the optimal window? It looks like the optimal BIC window is 100-200. And why not just pick THE single optimal number of states? Why do you need a window that contains highly nonoptimal models?

"The AIC and BIC penalize the larger state models much more than the test set log-likelihood..." Why is this discrepancy so large? Doesn't this suggest AIC/BIC are unsuitable if they are not representative of the actual out-of-sample performance on the test set?

"causes the transition matrix to suffer" Clarify?

For Mueller potential with Gaussian likelihood: Would be useful to show log-likelihood as a function of number of states so we can see how much the model prefers 5 states.

Fig. 3: "Since k-means optimizes an objective function that is related to the emission probability" What is this objective function? As I understand it, k-means simply minimizes the sum of intracluster variances, which has nothing to do with your choice of emission probability being uniform or Gaussian or otherwise. Or do you mean you optimize the log-likelihood form of Eq 1 using a voronoi decomposition based on generators that are optimized for each number of states? Also: "The likelihood decreases as new states are added..." Which likelihoods are printed in the figure panels? Test set, uniform, or Gaussian emission probabilities?

¿ * "...AIC is comparable to the test set log-likelihood...The BIC penalizes complexity more strongly..." Why the huge difference between the AIC and BIC?

Fig. 4: Too small. Missing figure title, like "Comparison of likelihood models for Fip35 WW domain trajectory data". "Test set log-likelihoods are omitted in C as the dense matrix algebra is intratable for $k > 10^3$". Intractable? Really? But this is just summing log T_{ij} over N elements, isn't it?

"For tICA, this window is consistent with the convergence of the relaxation timescales" So....looking at convergence of the relaxation timescales is just as good as the new proposed likelihood method?

"Lane et al. proposed...whereas Kellogg et al...Consistent with this spread, we find the optimal number of states to be highly dependent on the space in which the data is clustered." Were these two examples using the same metric of model quality and criteria for selecting the

number of states?

"These results indicate why likelihood-based model selection is superior to approaches relying only on the timescale convergence." How? I don't see clear evidence of this. Which likelihood model is superior specifically? How is this superior to choosing the smallest number of states for which the timescales become flat (assuming that such a thing exists)? The only evidence presented here has judged success by examining concordance among AIC/BIC/leave-some-out cross validation and the timescale metric. In Fig 4, BIC and AIC disagree, while AIC tracks (but does not quantitatively agree with) the test set method. The agreement with the timescale convergence is shown as evidence of correctness. But it seems like SOMETHING is missing to show that this approach (or, more properly, ONE of these likelihood approaches) is superior to simply checking when the timescales go flat. A much more convincing test would be to show that these methods allow one to achieve smaller model error (eg in reproducing some quantitative features of the dynamics—maybe even the timescales?) for some smaller quantity of data when compared with a "gold standard" enormous dataset. The behavior of the optimal number of states as a function of the quantity of data would also be of great interest.

"These results take a step toward fully automating the MSM construction process...by providing a quantitative method for selecting the most suitable complex model." Which method in particular is recommended? The Gaussian method? AIC or BIC? I'm still confused about what the actual, concrete demonstration of superiority is and which method the authors suggest should be used.

Minor issues that the authors should address are below: Abstract: "in regimes in which these techniques encourage over-fitting" It is unclear precisely what "these techniques" refers to.

Two key quantities which define the MSM are thus the state definitions, an indicator function basis over phase space..." There is a substantial quantity of literature on the use of basis functions that are not indicator functions (i.e. "fuzzy" rather than "crisp" partitionings) that should be mentioned. See, for example, the work of Marcus Weber and Susanna Roeblitz (nee Kube).

The model has the training data set as its support rather than phase space, which makes it unable to generalize and assign probability to new data." It may also be worth noting that this is not a generative model, since the next paragraph states that a generative model is desired.

"but is intractble in high dimensions when the states are defined by general polytopes, such as those produced by a data-driven Voronoi tessellation" Be clear that *exact* volume computation is intractable in very high dimension due to the complexity of all known algorithms. Monte Carlo schemes (such as the one used here, or methods such as [http://www.cs.elte.hu/ lovasz/vol4-focs-tr.pdf]) allow the approximation of this volume. However, it's also worth noting that the Gaussian model you propose in Eq. 4 is ALSO INTRACTABLE for EXACTLY THE SAME REASON, since you do not compute the true normalizing constant for essentially identical reasons of computational tractability in high dimension. As such, phrases such as "A more tractable alternative is a Gaussian emission model.." are entirely disingenuous and misleading.

"All MSMs were built using a lag time of 50ns" Why 50 ns?

"identify the simulations folding process" Something isn't quite right with this phrase. What is being referred to?

"the relaxation timescales can continually increase with the number of states used" Even in the limit of an infinite quantity of data, the eigenvalue error bound theorems from Schuette et al. note that this behavior is precisely what we should expect, since the approximation error decreases as the eigenfunctions are better approximated via more states. However, the statistical error grows, though this is not quantified or plotted here.

Reviewer 2:

The authors present an alternative approach for choosing the Markov time for Markov state models that will be of value for automating the construction of these models by removing the need for subjective user input. I have a few concerns, below.

Line 34: Did you mean sacrificing "accuracy" instead of "complexity"?

Can the authors comment on whether k-means or k-centers is preferable based on their Muller potential results?

The authors have access to many other data sets from their own lab and the Shaw group. Would it be possible to show results for other systems (maybe one larger system and one system with more disorder)? I think this would help to establish the generality of their results.

Reviewer 3:

Optimal Parameter Selection in Markov State Models for Biomolecular Conformational Dynamics by McGibbon, Schwantes and Pande. I enjoyed reading this paper. It is well suited for Bill Swope's birthday issue. It addresses an important problem in chemical physics, namely choosing amongst the various possible parameter settings when constructing a Markov State Model. However, the following general points should be addressed in a revision:

I find the title Optimal Parameter Selection in Markov State Models... inappropriate: It suggests that the "Parameter selection in MSMs" is a well-defined problem, and that the solution approach here would be optimal, i.e. not further improvable. Neither is true. I suggest to pick a title that is specific to the content of the paper, e.g. "Likelihood-approach for selecting the number of states / choosing between discretizations in MSMs"

The paper starts with a discussion of approaches that choose the lag time based on implied timescales, and then proposes a way to choose the number of discrete states as an alternative. I can see the idea behind it, but to general reader it will probably be unclear what the connection between the two is. It should be spelled out that there is a change of viewpoint, where instead of fixing the discretization and varying the lagtime, one fixes the lagtime and varies the discretization. And the hope why both approaches are meaningful is that the implied timescales should converge to their true values for either (1) increasing lag time [see Djurdjevac, Sarich, Schtte, MMS 2011 and Prinz, Chodera, Noe, PRX 2014 (http://arxiv.org/abs/1207.0225)], or (2) better approximation of the dominant eigenfunctions via more discrete states [see (19) and Sarich, Noe, Schütte MMS 8, p1154 (2010)].

My main concern is that the paper suggests that the best result is obtained by maximizing the present likelihood. But stationary and dynamical expectation values such as the relaxation timescales are well defined by the underlying dynamics, and the ultimate objective of any estimation procedure must be to correctly approximate them. Using the present likelihood could systematically give wrong estimates for various reasons. An inappropriate statistical model for the local output function (which is really hard to make) would lead to biased estimates. Then there is uncertainty in choosing the Bayesian criterion - AIC, BIC or something else? If the lag time chosen is too short, it will not lead to correct estimates for feasible numbers of states. The differences in Fig 4 (see Details below) indeed suggests that the present criterion fails to provide the correct estimate in at

least some cases. If a new estimator is derived, it needs to be shown that it converges to the correct result in some limit. I suggest to build a fine discrete model, e.g. by direct gridding of the Mller potential, and constructing a simple MCMC dynamics on neighboring gridpoints, and then considering different clusterings on this data set. There, the exact timescales are known, and an embedding in a geometric space is given, such that the method can be applied such that the present method can be applied as is. It then needs to be shown that the present likelihood approach selects MSMs with timescales that coincide with the true values.In addition, it would be useful to compare the results here with implied timescale plots. Do the results agree in the cases where ITS do converge?

Please check the timescales in Fig 2(A,B) and 4(A,B) - they look identical, but shouldnt. Details / specific points:

"ever-more states". Ever-more states or ever-larger lag time? (when did number of states come in?)

"Often the choice is made to use the model whose timescales better match an experimental observable:" I hope thats not true and I wouldnt generally accuse the community of practising this. When timescales dont converge, this is a numerical or statistical problem with the data or the estimators used. Ignoring this fact and deliberately choosing a model that coincidentially delivers the numbers one hopes to get would be simply cheating.

Gaussians could be truncated and normalized - it seems that renormalization would require to solve an even harder problem than for uniform output probabilities, because youd need to integrate the Gaussian density over a high-dimensional polytope.

"We assume that the overlaps are small" But are they small in your application? Ensuring this puts requirements on the sizes of states and sigma. How is it checked that this is then a good model?

Are you sure that the timescales shown in Fig 2A(right) and 2B(right) are from different data? They look identical to me.

Im surprised that there are three slow relaxation processes for the Mller potential before the second gap. The potential looks like there should be only two (due to three metastable states). Were MSMs build with reversible transition matrix estimation, and are all eigenvalues thus real-valued?

Muller potential: lag time was chosen to be 30 steps. How is this

choice justified? Since the timescales in Fig. 2 are not independent of N, this must mean that at least for some state decompositions the ITS are not converged in the lag time. In these cases, its not even appropriate to describe the state-to-state dynamics by a Markov chain at all. Is the statistical model still meaningful in this case

Mller potential, Gaussian likelihood: How were the Gaussian sigma-parameters chosen? The k-means states in Fig 3 have very different sizes and their centers are sometimes very close, so the assumption that output distributions shall not overlap seems to be hard to realize.

WW domain, lag time was chosen to be 50 ns - same comment as for the Mller potential, above

WW domain, Gaussian likelihood - same comment as for the Mller potential, above

Are you sure that the timescales shown in Fig. 4A(right) and 4B(right) are from different data? They look identical to me.

Doesnt Fig 4c suggest that theres a problem here? For the likelihood-based choice made, the slowest timescale is estimated between 1500 and 3000 ns with RMSD clustering. In contrast, the TICA-based estimates suggest a timescale of around 5000 ns. So if the TICA results is indeed correct, the method should be going for larger state number in the RMSD case.

In some timescale vs number of state plots the timescales descreases with increasing number of states (e.g. Fig 2A). Thats surprising. Can the authors explain that? Has a prior been used that may be responsible for this behavior?

Sincerely,

Robert T. McGibbon, Christian R. Schwantes, Vijay S. Pande