An introduction to infinite HMMs for single molecule data analysis

I Sgouralis and S Pressé

Abstract

The hidden Markov model (HMM) has been a workhorse of single molecule data analysis and is now commonly used as a standalone tool in time series analysis or in conjunction with other analyses methods such as tracking. Here we provide a conceptual introduction to an important generalization of the HMM which is poised to have a deep impact across Biophysics: the infinite hidden Markov model (iHMM). As a modeling tool, iHMMs can analyze sequential data without a priori setting a specific number of states as required for the traditional (finite) HMM. While the current literature on the iHMM is primarily intended for audiences in Statistics, the idea is powerful and the iHMM’s breadth in applicability outside Machine Learning and Data Science warrants a careful exposition. Here we explain the key ideas underlying the iHMM with a special emphasis on implementation and provide a description of a code we are making freely available. In a companion article, we provide an important extension of the iHMM to accommodate complications such as drift.

1 Introduction

Hidden Markov models (HMMs) [1, 2] provide a method for analyzing sequential time series data and have been a workhorse across fields including Biology [3, 4], Physics [5, 6, 7], and Engineering [8, 9, 10].

The power of HMMs has been heavily exploited in Biophysics [11, 12] in the interpretation of single molecule experiments such as fluorescence resonance energy transfer (FRET) [13]; force spectroscopy [14]; atomic force microscopy [15]; and ion channel patch-clamp [16]. The data from these disparate techniques may be analyzed using HMMs because biomolecules or collections of biomolecules are seen as visiting discrete states and the signal emitted from each state is corrupted by noise; see Fig. 1.

While HMMs have been hugely successful, the method has encountered a fundamental limitation. That is, the number of different states visited over the course of an experiment must be specified in advance [13]. This limitation is too restrictive for complex biomolecules where state numbers may often be difficult to assess a priori. It also presents a problem when the number of states appears to vary from trace to trace as would be expected if some states are only rarely visited. As a work-around, problem- and user-dependent pre- or post-processing steps have therefore been suggested in the biophysical literature to winnow down the number of candidate models or eliminate problematic traces altogether. Popular choices include model-selection tools such as information criteria (e.g. BIC, AIC [11, 17]) or maximum evidence methods (e.g. [12]). However, thanks to recent Mathematics – Bayesian nonparametrics, first introduced in 1973 [18] – an elegant solution is available that largely circumvents these model selection steps.

In this perspectives article, we provide a description of the concepts and implementation of an important new computational tool that exploits Bayesian nonparametrics: the infinite HMM (iHMM) that was heuristically described in Ref. [19] but only fully realized in 2012 [20].

It has recently been suggested that Bayesian nonparametrics are poised to have a deep impact in Biophysics [21] and, in this field, they have just begun to be exploited [22, 23, 24]. However, there is – to our knowledge – no single resource yet available describing the iHMM, its concepts or implementation, as would be required to bring the power of Bayesian nonparametrics to bear on Biophysics and accelerate its inevitable widespread adoption.
Indeed, while the iHMM tackles a conceptually simple problem, it relies on Mathematics whose literature is inaccessible outside a rarified community of Statisticians and Computer Scientists.

It is for this reason that we have organized our perspective article as follows: Section 2.1 describes the structure of the HMM in its finite (traditional) and infinite (nonparametric) realizations; Section 2.2 presents a computational algorithm to perform inference with the iHMM that we make freely available; Section 3 shows results from sample time traces; and Section 4 discusses the potential for further applications to Biophysics.

2 Methods

2.1 Formulation of the iHMM

Here we introduce the HMM and its generalization, the iHMM. To facilitate the presentation, we initially describe the structure of the space within which the biomolecule evolves and subsequently formulate the dynamics. This discussion applies to both HMMs and iHMMs.

In the HMM framework, a system of interest is assumed to alternate successively between different states labeled \( \sigma_k \) where \( k \) takes integer values from 1 to some \( L \). We use \( L \) to denote the total number of states available to the system, no matter if all these states are visited or some remain unvisited during the time course of the measurements. For instance, for an experiment on a single protein, the protein is the “system” and the “states” are conformations such as open/closed conformations of an ion channel [10] or low/high efficiency states in FRET experiments [26]. For an illustration see Fig. 1 (left).

In the standard HMM, we assume that the system’s transitions are governed by Markovian dynamics [1]. This means that the system jumps from a state \( \sigma_k \) to a state \( \sigma_j \), for example from one FRET value to another, in a stochastic manner that depends exclusively on \( \sigma_k \) and not any other state visited in the past. For this reason, all transitions out of state \( \sigma_k \) are fully described by a probability vector \( \tilde{\pi}_{\sigma_k} = (\pi_{\sigma_k \rightarrow \sigma_1}, \pi_{\sigma_k \rightarrow \sigma_2}, \ldots) \), where \( \pi_{\sigma_k \rightarrow \sigma_j} \) is the probability of departing from state \( \sigma_k \) and arriving to \( \sigma_j \).
A note on our notation is appropriate here. Throughout this survey we adopt *tildes* to denote vectors with components over \( S \), the system’s state space \( \{ \sigma_1, \sigma_2, \ldots \} \), such as \( \tilde{\sigma}_n \). Shortly, we will adopt *bars* to denote vectors with components over time.

Once the system reaches a state \( \sigma_k \), observations are emitted stochastically according to a probability distribution unique to \( \sigma_k \); see Fig. 1(right). We call this the “emission distribution” \( F_{\sigma_k} \). It is often practical to model the emission distributions by a general family and use state-specific parameters \( \phi_{\sigma_k} \) to distinguish its members. For example, to model Gaussian emissions \( x \), as in Fig. 1, we may choose

\[
F_{\sigma_k}(x) = \frac{1}{\sigma_{\sigma_k} \sqrt{2\pi}} \exp\left( -\frac{(x - \mu_{\sigma_k})^2}{2\sigma_{\sigma_k}^2} \right)
\]

where \( \phi_{\sigma_k} = (\mu_{\sigma_k}, \sigma_{\sigma_k}) \) stands for the mean \( \mu_{\sigma_k} \) and standard deviation (width) \( \sigma_{\sigma_k} \) of the observations produced by state \( \sigma_k \). There, for example, we define \( \phi_{\sigma_1} = (\mu_{\sigma_1}, \sigma_{\sigma_1}) \), \( \phi_{\sigma_2} = (\mu_{\sigma_2}, \sigma_{\sigma_2}) \), etc, where \( (\mu_{\sigma_1}, \mu_{\sigma_2}, \ldots) \) are the state FRET efficiencies and \( (\sigma_{\sigma_1}, \sigma_{\sigma_2}, \ldots) \) are the corresponding spreads attributed to noise.

Next, we let \( s_n \) denote the state of the system at the \( n \)^{th} time step of the experiment and \( x_n \) its corresponding observation. Thus, we label with \( n \) the states of the system as it evolves through time forming a sequence \( s_1 \to s_2 \to \cdots \to s_N \), with each \( s_n \) being equal to some state \( \sigma_k \) chosen from \( S \). For completeness, we may also assume that, before the first measurement, the system is at a default state which we denote as \( s_0 \). Again, we mention that \( S \) might contain states that do not appear in this sequence, i.e. states that remain unvisited throughout the experimental time course.

As is it common in the statistical literature – and especially since we will borrow from this notation to introduce the iHMM – we express the HMM compactly using the following scheme

\[
s_n | s_{n-1} \sim \text{Cat}_S(\tilde{\pi}_{s_{n-1}})
\]

\[
x_n | s_n \sim F_{s_n}
\]

and depict it schematically in Fig. 2. The notation “\( \sim \)” we use above indicates that the random variables on the left side are distributed according to, and therefore sampled from, the probability distributions shown on the right side \[27\]. For example, Eq. (2) denotes a sampling from a categorical probability distribution that is supported on \( S \), i.e. \( s_n \) equals a state \( \sigma_k \) that is taken from \( S \) with probability \( \pi_{s_{n-1} \to \sigma_k} \).

As can be seen from Eqs. (2) and (3), the HMM models the experimental output in a doubly stochastic manner: i) the state of the system \( s_n \) evolves stochastically within \( S \), as expressed by Eq. (2); and ii) the observations \( x_n \) from each \( s_n \) are also emitted stochastically, as expressed by Eq. (3). In particular, for single molecule experiments, these two characteristics allow the HMM to elegantly capture: i) the seemingly random biomolecular state switching; and ii) the noise corrupting the measurements \[13, 28\].

From the experimental measurements, we gain access only to the observations \( x_n \), which take the form of a time series \( \bar{x} = (x_1, x_2, \ldots, x_N) \) over some regularly spaced time intervals \( t_n \). In the most general case, the **goals of the HMM** are to estimate: i) the underlying state sequence \( \bar{s} = (s_1, s_2, \ldots, s_N) \) which is unobserved during the measurements; and ii) the model parameters which include \( \phi_{\sigma_k} \) of the emission distributions and the state transition probabilities \( \pi_{\sigma_k \to \sigma_j} \), associated with the states in \( S \). Of course, in practice these goals are feasible only for those states that are visited during the experimental time course.

Normally, this is the extent of the HMM. That is, one fixes \( L \), i.e. the size of the state space \( S \), writes down
Figure 2: **Graphical representation of the HMM.** In the HMM, a biomolecule of interest transitions between unobserved states $s_n$ according to the probability vectors $\tilde{\pi}_{s_n}$ and generates observations $x_n$ according to the probability distributions $F_{s_n}$ that depend on parameters $\phi_{s_n}$. Here, following convention, the $x_n$’s are shaded to denote that these quantities are observed, while the $s_n$’s are hidden. Arrows denote the dependences among the model variables and red lines denote the model parameters.

the likelihood of observing the sequence of observations $\bar{x} = (x_1, x_2, \ldots, x_N)$, and maximizes this likelihood to estimate the quantities of interest. Extensive literature – that has made its way into standard textbooks, for example Ref. [27] – explain each of these steps in detail.

However, in preparation for the iHMM, we take a Bayesian route to accomplish the goals of the HMM [21]. This is a computational overkill for some applications but otherwise essential to its nonparametric generalization.

Following the Bayesian paradigm [21, 29], we assign prior probabilities to model parameters including the emission parameters $\phi_{\sigma_k}$ and transition probability vectors $\tilde{\pi}_{\sigma_k}$. For instance, we may assume a prior $\phi_{\sigma_k} \sim H$ given by a common distribution for all states. However, assigning a prior to $\tilde{\pi}_{\sigma_k}$ is more subtle as any choice must ensure that the predicted transitions stay within the system’s state space $S$. It is specifically the formulation of this prior that fundamentally distinguishes the finite from the infinite variants of the HMM [20], both of which we describe next.

In the **finite variant** of the HMM [1, 2], before setting the prior on $\tilde{\pi}_{\sigma_k}$, we must fix $L$, the total number of states in $S$ or, in the language of single molecule Biophysics, conformations available to the biomolecule. Once $L$ is fixed, the symmetric *Dirichlet distribution* is a common choice

\[ \tilde{\pi}_{\sigma_k} \sim \text{Dir}_S\left(\frac{\alpha}{L}, \ldots, \frac{\alpha}{L}\right). \]

(4)

Here, $\text{Dir}_S(\alpha/L, \ldots, \alpha/L)$ denotes the Dirichlet distribution supported on $S$ with concentration parameter $\alpha$ (for an explicit definition of the Dirichlet distribution see supporting material). Basically, this prior asserts that lacking any information on the system’s kinetics, the model is likely to place on average equal probability on any possible transition $\sigma_k \rightarrow \sigma_j$. In this prior, the value of $\alpha$ can be used to influence the transitions departing each state. For example, with $\alpha \gg 1$ the prior tends to favor uniform transitions, while with $\alpha \ll 1$ the prior tends to favor sparse transitions. In other words, the resulting sequence $\bar{s}$ tends to contain several different states for $\alpha \gg 1$, while it tends to contain fewer for $\alpha \ll 1$.

The Dirichlet prior in Eq. (4) offers two key advantages: i) it provides a non-informative prior since no specific transitions between the $L$ states are preferentially selected; and ii) it combines well with (i.e. is conjugate to) the categorical distribution of Eq. (2) which greatly simplifies model parameter estimation.

Although such priors help lessen the computational burden, pre-setting $L$: i) ignores the data and thus the arbitrary choice of $L$ may cause under- or over-fitting that, in turn, has far-reaching consequences in estimating state kinetics; and ii) does not allow the model’s complexity to grow in response to newly available data (e.g. a
Figure 3: Graphical representation of the iHMM. The hidden Markov model that formulates the observations to be analyzed (black lines) is shown together with its priors (red lines). For completeness, we also show the concentration parameters $\alpha, \gamma$ and the prior probability distribution on the emission parameters $H$ that fully characterize the iHMM. The key difference from the HMM shown in Fig. 2 is that now the model parameters $\tilde{\pi}_k \sigma$ and $\phi_{k \sigma}$ are treated as random variables similar to the hidden states $s_n$ and observations $x_n$. For details see the main text.

Rarer state visits, visited later or in another time series. Resolving issue ii) in a principled fashion can also help avoid cherry-picking data sets behaving closer to one’s expected or preferred value of $L$. It is to resolve such issues that the iHMM has been developed in the first place.

Now, in the infinite HMM, the key difference is that $S$ is assumed infinite in size [19]. To avoid any confusion, we point out that this assumption is different from forcing the system to visit an infinite number of states, as it might appear at first. Regardless of the size of $S$, the system visits only $K$ different states, where $K$ cannot exceed, for example, the number of steps $N$ in the collected time series. In practice, of course, we expect $K$ to be considerably smaller than $N$.

As we will see, with an infinite state space, the iHMM can recruit as many states as necessary and, in doing so, avoids underfitting by growing the state space as new states are visited (through the effect of the prior) but also avoids overfitting by placing more weight on states already visited (through the effect of the likelihood).

The generalization of the Dirichlet distribution appropriate for the infinite sized $S$ is the Dirichlet process

$$
\tilde{\pi}_\sigma \sim DP_S(\alpha, \tilde{\beta})
$$

where $\tilde{\beta}$ is the “base distribution” of the Dirichlet process (DP) which determines how $\tilde{\pi}_\sigma$ are distributed on average. The notion of a base is only required for the infinite DP and has no direct analog in the finite Dirichlet distribution. By contrast, $\alpha$ is similar to the concentration of the finite HMM since it controls the sparsity of the transition probability vectors $\tilde{\pi}_\sigma$. For instance, large values of $\alpha$ make the model more likely to choose $\tilde{\pi}_\sigma$ that are similar for all states, while low values of $\alpha$ make the model more likely to choose $\tilde{\pi}_\sigma$ that differ considerably.

The DP is at the basis of much of nonparametric Bayesian inference. It was only in 2012 that a hierarchical DP (HDP) was invoked in order to construct the iHMM [20]. The basic idea is to select a random – albeit discrete – base distribution $\tilde{\beta} = (\beta_1, \beta_2, \ldots)$, with members $\beta_k$ that: i) are infinite in number, ii) attain non-negative values, and iii) sum to 1. In the HDP, it is common to sample such a base from a GEM (Griffiths-Engen-McCloskey) distribution realized through a stick-breaking construction [30, 20, 31]

$$
\tilde{\beta} \sim GEM_S(\gamma).
$$

which ensures all these properties. Intuitively, the stick-breaking above is obtained as follows: we start with a stick of length 1. We break the stick at some location $v_1$ and use the broken fraction for the probability of visiting
\[\sigma_1, \text{ thus } \beta_{\sigma_1} = v_1.\] Subsequently, we break the remaining stick at \(v_2\) and use that new fraction for the probability of visiting \(\sigma_2\), thus \(\beta_{\sigma_2} = v_2(1 - \beta_{\sigma_1})\), and so on. In this stick-breaking, the locations \(v_k\) are sampled from a beta distribution \(B(1, \gamma)\), so \(\gamma\) determines the relative location of the break point. Thus \(\gamma\) can be used to set how rapidly \(\beta_{\sigma_k}\) decreases toward zero and ultimately controls the number of states in \(S\) with an appreciable weight of being visited at least once during the experiment.

In summary, the iHMM – together with its priors illustrated in Fig. 3 – takes the following form

\[
\begin{align*}
\tilde{\beta} & \sim GEM_S(\gamma) \\
\tilde{\pi}_{\sigma_k} & \sim DP_S(\alpha, \tilde{\beta}) \\
\phi_{\sigma_k} & \sim H \\
s_n|s_{n-1}, \tilde{\pi} & \sim \text{Cat}_S(\tilde{\pi}_{s_{n-1}}) \\
x_n|s_{n}, \tilde{\phi} & \sim F_{s_n}
\end{align*}
\]

where, for notational simplicity, we use \(\tilde{\pi} = \{\tilde{\pi}_{\sigma_1}, \tilde{\pi}_{\sigma_2}, \ldots\}\) and \(\tilde{\phi} = \{\phi_{\sigma_1}, \phi_{\sigma_2}, \ldots\}\) to gather the transition probability vectors and emission parameters associated with all states in \(S\). Here, \(\tilde{\pi}\) can be seen as a bi-infinite matrix whose elements are the probabilities for each possible transition \(\sigma_k \rightarrow \sigma_j\) and it is directly analogous to the transition matrix of the finite HMM.

Unlike the finite HMM where we need to be specific about the size of the state space, the iHMM gives us greater flexibility through the hyperparameters \(\alpha\) and \(\gamma\). For instance, larger concentrations are more appropriate for biomolecules with roughly similar transition rates and several conformations while low concentrations are more appropriate for biomolecules with dissimilar transition rates and few conformations. In general, it is possible to sum over these concentrations if we are truly ignorant of the properties of the biomolecule at hand [20, 32].

As we will see shortly, this powerful formalism allows us to infer state numbers robustly and, even if the number of states is apparent – something which is rarely the case for more complex problems [33] – avoids the critical shortfall of cherry-picking data sets to be analyzed that appear to have similar state numbers.

2.2 Inference on the iHMM

Prior to describing the analysis that iHMMs may offer, we show how to infer quantities such as the hidden state sequence and the parameters from the experimental traces. A working implementation of the algorithm described in this section, equipped with a user interface, can be found in the supporting materials.

To be clear, on account of the infinite size of \(S\), common methods used in finite state HMMs such as Expectation-Maximization or simply EM [1] are inapplicable. Instead, we focus our discussion on a specialized method: the beam sampler [34]. Other methods are also described elsewhere [20, 32].

The beam sampler is a special instance of the Gibbs sampler [35], and although it is beyond the scope of this survey, a brief description about its usage might be beneficial here. Similar to all samplers of this family, we use the beam sampler to generate (pseudo) random sequences of the model variables we are interested in inferring. Specifically, in our case these variables consist of the hidden state sequence \(\tilde{s}\) and the model parameters \(\tilde{\beta}, \tilde{\pi}, \tilde{\phi}\). When generating the sequences, we use Eqs. (7)-(11) and the data \(\tilde{x}\). As a result, the generated sequences have the same statistical properties as if they were consisting of samples from the posterior probability distribution.

\[1\] This code can also be found on the authors’ website as well as on GitHub.
Then, we could generate \( \tilde{P} \) or simply produce histograms that resemble \( \mathbb{P}(\bar{s}, \tilde{\beta}, \tilde{\pi}, \bar{\phi}|\bar{x}) \) as we show in the next section.

Overall, Gibbs and related samplers provide a more general approach than, for example, EM which provides only the maximum of \( \mathbb{P}(\bar{s}, \tilde{\beta}, \tilde{\pi}, \bar{\phi}|\bar{x}) \). Instead, with Gibbs sampling, we fully characterize \( \mathbb{P}(\bar{s}, \tilde{\beta}, \tilde{\pi}, \bar{\phi}|\bar{x}) \) over its whole domain. For an introduction to the general methodology underlying Gibbs sampling we refer the interested reader to Ref. \[36\], while from now on we focus on the iHMM.

Suppose we have a time series of experimental observations \( \bar{x} \). Here we describe the steps involved in generating samples \((\text{superscripted } r)\) of our model variables \( \bar{s}(r), \tilde{\beta}(r), \tilde{\pi}(r), \bar{\phi}(r) \). Given a computed sample \((r-1)\), the naive approach for computing the next one would be by updating each of the involved variables conditioned on the other variables and the data \[35\]. For example, we could generate \( \bar{s}(r) \) by sampling from \( \mathbb{P}(\bar{s} | \bar{\beta}(r-1), \tilde{\pi}(r-1), \bar{\phi}(r-1), \bar{x}) \). Then, we could generate \( \tilde{\beta}(r) \) by sampling from \( \mathbb{P}(\tilde{\beta} | \bar{s}(r), \tilde{\pi}(r-1), \bar{\phi}(r-1), \bar{x}) \), and so on. From the theory of Markov chain Monte Carlo sampling – for example Ref. \[35\] – this approach is guaranteed to produce samples from the full posterior \( \mathbb{P}(\bar{s}, \tilde{\beta}, \tilde{\pi}, \bar{\phi}|\bar{x}) \) as we wish. However, the infinite size of \( S \) makes this approach impractical since in each iteration we need to sample infinite sized \( \tilde{\beta}, \tilde{\pi}, \bar{\phi} \) which is computationally infeasible.

It is fundamentally because of this difficulty that we need the beam sampler which introduces a set of auxiliary variables (slicers) that in each iteration effectively truncate \( S \) to a finite portion \[34, 37\]. In particular, for each time step \( n \) in the dataset consider an auxiliary variable \( u_n \) and let \( \bar{u} = (u_1, u_2, \ldots, u_N) \) gather all of them. Since \( \bar{u} \) consists of auxiliary variables only, we have

\[
\mathbb{P}(\bar{s}, \tilde{\beta}, \tilde{\pi}, \bar{\phi}|\bar{x}) = \sum_{\bar{u}} \mathbb{P}(\bar{u}, \bar{s}, \tilde{\beta}, \tilde{\pi}, \bar{\phi}|\bar{x}) \tag{12}
\]

where the sum on the right hand side is considered over all possible values \( \bar{u} \) may take. Now, we can use the equality in Eq. \[12\] to sample from \( \mathbb{P}(\bar{u}, \bar{s}, \tilde{\beta}, \tilde{\pi}, \bar{\phi}|\bar{x}) \) instead from \( \mathbb{P}(\bar{s}, \tilde{\beta}, \tilde{\pi}, \bar{\phi}|\bar{x}) \). The benefit of doing so is that, when updating \( \tilde{\beta}, \tilde{\pi}, \bar{\phi} \), we have to sample conditioned also on \( \bar{u} \). Thus, by properly choosing the auxiliaries \( u_n \), we make these updates consider only a finite selection of the states \( \sigma_k \). For instance, by having \( u_n \) uniformly distributed over the interval \( (0, \pi_{s_{n-1} \rightarrow s_n}) \), it is sufficient to consider only those \( \sigma_k \) in \( S \) that have a probability of being visited at least once that is greater than \( u_n \). Therefore, we can generate the desired samples while maintaining a finite \( S \) that we expand and compress dynamically according to \( \bar{u} \).

In summary, given \( \bar{s}(r-1), \tilde{\beta}(r-1), \tilde{\pi}(r-1), \bar{\phi}(r-1) \) the beam sampler generates a new set of samples through the following steps:

1. Generate \( \bar{u}(r) \) by sampling from \( \mathbb{P}(\bar{u}|\bar{s}(r-1), \tilde{\pi}(r-1)) \)
2. Expand \( S \) according to \( \bar{u}(r) \)
3. Generate \( \bar{s}(r) \) by sampling from \( \mathbb{P}(\bar{s}|\bar{u}(r), \tilde{\pi}(r-1), \bar{\phi}(r-1), \bar{x}) \)
4. Compress \( S \) according to \( \bar{s}(r) \)
5. Generate \( \tilde{\beta}(r) \) by sampling from \( \mathbb{P}(\tilde{\beta}|\bar{s}(r), \tilde{\pi}(r-1)) \)
6. Generate \( \tilde{\pi}(r) \) by sampling from \( \mathbb{P}(\tilde{\pi}|\bar{s}(r), \tilde{\beta}(r)) \)
7. Generate \( \bar{\phi}(r) \) by sampling from \( \mathbb{P}(\bar{\phi}|\bar{s}(r), \bar{x}) \)
where for simplicity we have dropped additional dependencies in the conditional probability distributions above.

In principle, $\tilde{\beta}$, $\tilde{\pi}$, and $\tilde{\phi}$ are infinite dimensional vectors. However, as mentioned above, the sampler requires the computation of only those components that correspond to the states which $\bar{u}$ allows to be visited. Therefore, in step 4 of the above algorithm we can safely discard those components that are absent from the state sequence $\bar{s}^{(r)}$. By doing so, in each time, we only need to store vectors of finite dimension, $\tilde{\beta}^{(r)}$, $\tilde{\pi}^{(r)}$, and $\tilde{\phi}^{(r)}$. Subsequently, in step 2 of the following iteration we can supplement $\tilde{\beta}^{(r)}$, $\tilde{\pi}^{(r)}$, and $\tilde{\phi}^{(r)}$ with states as necessary. The discarded states, since are unvisited, do not produce any of the observations in $\bar{x}$. Thus, they do not provide any information that could be lost. Similarly, the supplemented states are also not associated with any of the observations in $\bar{x}$. Thus, when we generate them, we simply have to draw samples from their priors. Consequently, the supplemented states neither require nor introduce extra information as it might appear at first. In summary, expanding and compressing $S$ in steps 2 and 4 leaves our inference unaffected.

The supporting materials provide a pseudocode implementation of the algorithm with further details on each of the sampler’s steps for the interested reader.

3 Results

In this section, we demonstrate the use of the iHMM in the analysis of time series data. To benchmark the method, we use synthetic time series where the “ground truth” is known. In a companion Research article – that draws heavily from the formulation described in this paper – we show the performance of the method on actual time traces that exhibit further complications such as drift [38]. For such time traces, we need an additional important generalization to the iHMM; the iHMM must be coupled to a continuous process. For now, our synthetic traces are meant to illustrate realistic data with noise but from which drift is otherwise absent or minimal.

Below we present the results of the analysis on two types of time traces specifically problematic for the HMM:

- **Dataset 1**: time traces where the visited states are in high number and their emissions show significant overlap.
- **Dataset 2**: time traces where some states are rarely visited and the number of states appears to change as more data become available.

Details about the generation of these datasets, as well as more in depth analyses than those we discuss here, can be found in the supporting materials.

The datasets we analyze are shown in Fig. 4. Specifically, we used these time series to estimate the posterior distributions over: i) the number, $K$, of different conformations visited by the biomolecule, and ii) the parameters describing the emission distributions, such as the mean value for each conformation $\mu_{\sigma_k}$, for those states. To accomplish this, we generate samples from $P(\bar{u}, \bar{s}, \tilde{\beta}, \tilde{\pi}, \tilde{\phi} | \bar{x})$ using the beam sampler described earlier.

Figure 5 shows how the number of different states visited in each sampled state sequence $\bar{s}^{(r)}$ for dataset 1, which we denote with $K^{(r)}$, changes through the sampler’s iterations. Again, we mention that $K$ counts only those states that are visited within the given trace. Initially, $K^{(1)}$ equals the number of states in the state space as specified in the initialization of our method. After approximately 50 iterations, $K^{(r)}$ drops to 5 which is the correct number of states attained by our hypothetical biomolecule.

Subsequently, $K^{(r)}$ stays at 5 and occasionally adds extra states that are rapidly eliminated. This behavior is characteristic of our algorithm [35]. In the initial iterations, known as “burn-in”, the sampler randomly explores
Figure 4: Synthetic datasets resembling a hypothetical biomolecule undergoing transitions between discrete states that we analyzed with the iHMM. Left: time series $\tilde{x} = (x_1, \ldots, x_N)$ of noisy observations. During the measuring period the biomolecule attains 5 conformations $\sigma_1, \ldots, \sigma_5$. The number of conformations are a priori unknown and the iHMM seeks to determine the probability over the number of states as well as their properties given the data available. In dataset 1, the biomolecule transitions often through every state. By contrast, in dataset 2 transitions to some states are rare. As a result, all states in dataset 1 are almost equally visited throughout the experiment time course, while in dataset 2 higher states are visited, by chance, only toward the end of the trace. Right: corresponding emission distributions $F_{\sigma_k}$ as obtained by simply binning the observations (blue) and by plotting the exact ones used for the simulations (red). For both datasets, the emission distributions show significant overlap. In all panels, dotted lines indicate the exact mean values $\mu_{\sigma_k}$ of the emission distributions.
Figure 5: After some iterations, the sampler used in the iHMM to analyze dataset 1 of Fig. 4 eventually converges to the correct number of states. The number of visited states $K^{(r)}$ (top) and the means of the emission distributions $\mu_{\sigma_k}^{(r)}$ (bottom) change throughout the sampler’s iterations. Unlike the HMM that uses a finite and fixed state space, the iHMM learns the number of available states and grows/shrinks the state-space as required by the data.

possible choices for the model variables that eventually evolve closer to the ground truth. Once they approach the ground truth, the sampler begins generating them based on the corresponding posterior distribution $P(K|\bar{x})$. So, after burn-in, values with high $P(K|\bar{x})$ are sampled more often than those with low $P(K|\bar{x})$.

The same is true of the other variables we try to estimate. For example, in Fig. 5, we also show how the sampled means of the emission distributions $\mu_{\sigma_k}^{(r)}$ change through the iterations. As with $K^{(r)}$, during burn-in, the sampler explores different choices which eventually converge to the ground truth. Afterwards, the sampler occasionally explores new emissions, for example around iterations 150–250, that quickly disappear.

To obtain more quantitative results, we may drop the burn-in samples from the generated sequences and use the remainder of the samples to produce histograms. For example, in Fig. 5, we show histograms of $K^{(r)}$ and $\mu_{\sigma_k}^{(r)}$. These histograms approximate the corresponding posterior probability distributions and thus provide a full characterization of the estimated variables. In these cases, the breadth of the histograms arises because the data we analyze are limited and noisy, so the estimates are associated with some uncertainty that is reflected in the broadening of the posterior distributions.

As can be seen, dataset 1 – while containing a large number of states which are not well separated – can be successfully analyzed by means of the iHMM. However, dataset 2 poses another difficulty. Namely, the biomolecule visits some states only rarely. In experiments, this would give rise to traces with an unequal number of states.
We may use samples from the iHMM posterior probability to infer the size of the state space and the location of each state. In particular, we illustrate histograms for $P(K | \bar{x})$ (top) and $P(\mu_\alpha | \bar{x})$ (bottom) using the dataset 1 of Fig. 4. On both panels dashed lines indicate the exact (ground truth) values used to produce the data in Fig. 4.
Figure 7: We may use the iHMM to estimate portions of the complete state space such as those contained in different segments of dataset 2 provided in Fig. 4. Upper: shows estimated noiseless traces for two cases: i) using a limited segment of the full trace; and ii) using the full trace. Although only the later case allows an estimate of all 5 states, both cases provide similar estimates over those states which they mutually visit. Lower: Corresponding estimates of the number of states contained in each trace.

One may naively suggest that longer traces should be collected but, due to experimental limitations (e.g. early photobleaching in FRET experiments), such long traces might not always be available. Instead, a large number of shorter traces, each containing an unknown portion of the complete state space, may need to be analyzed separately and the results combined [39]. For this reason, it is crucial to treat all traces on an equal footing without a priori assuming a different number of states individually for each one as the HMM would require.

To illustrate how the iHMM handles such cases, we use the iHMM to analyze the data of Fig. 4 (bottom trace) twice. First, using only the initial 300 time steps which visits only 2 out of the 5 states, and subsequently using the complete trace which contains all 5 states.

Figure 7 shows the results of the analyses. Specifically, in the upper panel we show the estimated noiseless traces and in the lower panel the estimated numbers of states. As can be seen, the iHMM successfully estimates the correct number in each trace. Further, as can be seen from the denoised traces, the portion of the state space, i.e. states $\sigma_1$ and $\sigma_2$, that is common is estimated accurately from both traces.
4 Perspectives

In this survey we presented the infinite hidden Markov model [19, 20], a relatively recent development in Statistics which has already found diverse applicability, among other fields, in Genomics [40], Genetics [41], Finance [42, 43], Tracking [44, 45], Machine learning [46, 47], and also Biophysics [22, 23, 24].

As a modeling tool, the iHMM inherits the characteristics of its predecessor – the finite hidden Markov model – but offers greater flexibility in the modeling and analysis of the experimental data. An important advantage is that it does not require the underlying system to have a limited state space as would its predecessor. This feature is of particular importance to Biophysics where little information is often available to fix the number of biomolecular states a priori.

Despite the difficulty in prespecifying the complexity of the state space, the finite HMM is a routine choice for the analysis of single molecule time series [11, 12, 13]. Nevertheless, as the experimental techniques improve and data from complex biomolecules are becoming available, the shortcomings of the finite HMM have become more apparent. To this end, methods to circumvent the limitations of HMM have been developed independently of the progress made on the same problems that have exploited Bayesian nonparametrics.

While it has previously been suggested that single molecule analysis may benefit from Bayesian nonparametrics [22], to date only limited applications of these methods have yet been explored. This is, in part, because models like the iHMM and its implementation are scattered over various sources but also because of the important language barrier between fields where Bayesian nonparametrics have changed the course of data analysis and the Physical Sciences. It is in part for this reason that we believe methods continue to be developed to address consequences of the finiteness of the state space of the HMM in the Physical Sciences.

The iHMM, described here, presents great advantages in the analysis of single molecule observations. For one, it elegantly addresses those challenges that are presented by the HMM, namely the problem of having to select a number of states when the emission distributions are unclear. But also, to address the problem – currently often addressed in a user-dependent fashion – of how to proceed when different traces present a seemingly different number of states. The key idea is that, just as the HMM performs de-noising while parametrizing transition kinetics, the iHMM takes this logic one step further by additionally learning the number of states in a self-consistent manner.

While greatly advantageous, the iHMM itself presents an important challenge for experimental data with drift, a common feature of Biophysical time traces [26, 13]. Indeed without its explicit consideration, drift would be interpreted as the population of artifact states by a method willing to recruit extra states to accommodate the data. This is a key challenge that has so far not been considered anywhere. Without careful consideration, drift would render the power of the iHMM into a key disadvantage. While our main emphasis here has been to feature the power of the iHMM to Biophysics, our companion Research article [38] now tackles the problem of extending iHMMs to deal with drift.

References

[1] L Rabiner and B Juang. An introduction to hidden Markov models. IEEE ASSP Magazine, 3(1):4–16, 1986.

[2] SR Eddy. What is a hidden Markov model? Nature Biotechnology, 22(10):1315–1316, 2004.

[3] BJ Yoon. Hidden Markov models and their applications in biological sequence analysis. Current Genomics, 10(6):402–415, 2009.
[4] A Krogh, M Brown, IS Mian, K Sjölander, and D Haussler. Hidden Markov models in computational biology: Applications to protein modeling. *Journal of Molecular Biology*, 235(5):1501–1531, 1994.

[5] RL Streit and RF Barrett. Frequency line tracking using hidden Markov models. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 38(4):586–598, 1990.

[6] A Pikrakis, S Theodoridis, and D Kamarotos. Classification of musical patterns using variable duration hidden Markov models. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(5):1795–1807, 2006.

[7] N Dasgupta, P Runkle, L Couchman, and L Carin. Dual hidden Markov model for characterizing wavelet coefficients from multi-aspect scattering data. *Signal Processing*, 81(6):1303–1316, 2001.

[8] S Fine, Y Singer, and N Tishby. The hierarchical hidden Markov model: Analysis and applications. *Machine Learning*, 32(1):41–62, 1998.

[9] BH Juang and LR Rabiner. Hidden Markov models for speech recognition. *Technometrics*, 33(3):251–272, 1991.

[10] JL Marroquin, EA Santana, and S Botello. Hidden Markov measure field models for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(11):1380–1387, 2003.

[11] SA McKinney, C Joo, and T Ha. Analysis of single-molecule FRET trajectories using hidden Markov modeling. *Biophysical Journal*, 91(5):1941–1951, 2006.

[12] JE Bronson, J Fei, JM Hofman, RL Gonzalez, and CH Wiggins. Learning rates and states from biophysical time series: a Bayesian approach to model selection and single-molecule fret data. *Biophysical Journal*, 97(12):3196–3205, 2009.

[13] M Blanco and NG Walter. Analysis of complex single-molecule FRET time trajectories. *Methods in Enzymology*, 472:153–178, 2010.

[14] X Long, JW Parks, CR Bagshaw, and MD Stone. Mechanical unfolding of human telomere G-quadruplex DNA probed by integrated fluorescence and magnetic tweezers spectroscopy. *Nucleic Acids Research*, 41(4):2746–2755, 2013.

[15] M Kruithof and J van Noort. Hidden Markov analysis of nucleosome unwrapping under force. *Biophysical Journal*, 96(9):3708–3715, 2009.

[16] L Venkataramanan and FJ Sigworth. Applying hidden Markov models to the analysis of single ion channel activity. *Biophysical Journal*, 82(4):1930–1942, 2002.

[17] JB Munro, RB Altman, N O’Connor, and SC Blanchard. Identification of two distinct hybrid state intermediates on the ribosome. *Molecular Cell*, 25(4):505–517, 2007.

[18] TS Ferguson. A bayesian analysis of some nonparametric problems. *The annals of statistics*, pages 209–230, 1973.

[19] MJ Beal, Z Ghahramani, and CE Rasmussen. The infinite hidden Markov model. In *Advances in Neural Information Processing Systems*, pages 577–584, 2001.

[20] YW Teh, MI Jordan, MJ Beal, and DM Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 2012.

[21] KE Hines. A primer on Bayesian inference for biophysical systems. *Biophysical Journal*, 108(9):2103–2113, 2015.

[22] KE Hines, JR Bankston, and RW Aldrich. Analyzing single-molecule time series via nonparametric Bayesian inference. *Biophysical Journal*, 108(3):540–556, 2015.
[23] K Palla, DA Knowles, and Z Ghahramani. A reversible infinite hmm using normalised random measures. In *International Conference on Machine Learning*, 2014.

[24] CP Calderon and K Bloom. Inferring latent states and refining force estimates via hierarchical dirichlet process modeling in single particle tracking experiments. *PloS one*, 10(9):e0137633, 2015.

[25] AE Conicella, GH Zerze, J Mittal, and NL Fawzi. ALS mutations disrupt phase separation mediated by $\alpha$-helical structure in the TDP-43 low-complexity c-terminal domain. *Structure*, 24(9):1537–1549, 2016.

[26] R Roy, S Hohng, and T Ha. A practical guide to single-molecule FRET. *Nature Methods*, 5(6):507–516, 2008.

[27] C Bishop. Pattern recognition and machine learning (information science and statistics), 1st edn. 2006. corr. 2nd printing edn, 2007.

[28] E Nir, X Michalet, KM Hamadani, TA Laurence, D Neuhauser, Y Kovchegov, and S Weiss. Shot-noise limited single-molecule FRET histograms: comparison between theory and experiments. *Journal of Physical Chemistry B*, 110(44):22103–22124, 2006.

[29] JW Yoon, A Bruckbauer, WJ Fitzgerald, and D Klenerman. Bayesian inference for improved single molecule fluorescence tracking. *Biophysical Journal*, 94(12):4932–4947, 2008.

[30] J Sethuraman. A constructive definition of Dirichlet priors. *Statistica Sinica*, pages 639–650, 1994.

[31] Jim Pitman. Poisson–dirichlet and gem invariant distributions for split-and-merge transformations of an interval partition. *Combinatorics, Probability & Computing*, 11(05):501–514, 2002.

[32] EB Fox, EB Sudderth, MI Jordan, and AS Willsky. A sticky HDP-HMM with application to speaker diarization. *The Annals of Applied Statistics*, pages 1020–1056, 2011.

[33] M Pirchi, G Ziv, I Riven, SS C, N Zohar, Y Barak, and G Haran. Single-molecule fluorescence spectroscopy maps the folding landscape of a large protein. *Nature Communications*, 2:493, 2011.

[34] J Van Gael, Y Saatci, YW Teh, and Z Ghahramani. Beam sampling for the infinite hidden Markov model. In *Proceedings of the 25th International Conference on Machine Learning*, pages 1088–1095. ACM, 2008.

[35] Ch Robert and G Casella. *Monte Carlo statistical methods*. Springer Science & Business Media, 2013.

[36] C Robert and G Casella. *Introducing Monte Carlo Methods with R*. Springer Science & Business Media, 2009.

[37] SG Walker. Sampling the Dirichlet mixture model with slices. *Communications in Statistics–Simulation and Computation*, 36(1):45–54, 2007.

[38] I Sgouralis and S Pressé. ICON: an adaptation of infinite HMMs for time traces with drift. *Biophysical Journal* (Submitted), 2016.

[39] MR Blanco, JS Martin, ML Kahlscheuer, R Krishnan, J Abelson, A Laederach, and NG Walter. Single molecule cluster analysis dissects splicing pathway conformational dynamics. *Nature Methods*, 2015.

[40] C Yau, O Papaspiiliopoulos, GO Roberts, and C Holmes. Bayesian non-parametric hidden Markov models with applications in genomics. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(1):37–57, 2011.

[41] EP Xing, KA Sohn, et al. Hidden Markov Dirichlet process: Modeling genetic inference in open ancestral space. *Bayesian Analysis*, 2(3):501–527, 2007.

[42] A Dufays. Infinite-state Markov-switching for dynamic volatility. *Journal of Financial Econometrics*, page nbv017, 2015.
[43] S Shi and Y Song. Identifying speculative bubbles using an infinite hidden Markov model. *Journal of Financial Econometrics*, page nbu025, 2014.

[44] D Kuettel, MD Breitenstein, L Van Gool, and V Ferrari. What’s going on? discovering spatio-temporal dependencies in dynamic scenes. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 1951–1958. IEEE, 2010.

[45] EB Fox, EB Sudderth, and AS Willsky. Hierarchical Dirichlet processes for tracking maneuvering targets. In *Information Fusion, 2007 10th International Conference on*, pages 1–8. IEEE, 2007.

[46] YW Teh and MI Jordan. Hierarchical Bayesian nonparametric models with applications. *Bayesian nonparametrics*, 1, 2010.

[47] W Hu, G Tian, X Li, and S Maybank. An improved hierarchical Dirichlet process-hidden Markov model and its application to trajectory modeling and retrieval. *International Journal of Computer Vision*, 105(3):246–268, 2013.