Bayesian nonparametric extensions of Hidden Markov Models

Meenakshi Khosla

December 2018

Abstract

Nonparametric or infinite-dimensional counterparts of Hidden Markov Models have recently shown great promise for segmenting temporal sequences with an unknown number of components. In this technical report, I will focus on the first paper describing infinite-state hidden markov models (iHMMs)[1], while briefly describing the recent developments in modelling over this traditional scheme. This is followed by a discussion on the evolution of inference algorithms for nonparametric HMMs such as the iHMM and its well-known extension-the Hierarchical Dirichlet Process-Hidden Markov Model (HDP-HMM).

1 Introduction

Hidden Markov Models are powerful models for describing sequences, such as speech, genomes, proteins or stock values over time. A major limitation of classical HMMs is that the model is constrained to have a finite number of states beforehand. Domain knowledge is mostly not adequate to specify the dimensionality of the latent space,

and most applications rely on ad-hoc approaches to fix this cardinality. While model selection techniques exist for standard HMMs to determine the number of states, the convergence and divergence among these methods is little understood. Moreover, in certain cases, the cardinality of hidden states and its uncertainty may itself be the variables of interest for inference. This makes nonparametric extensions of HMMs particularly attractive.

Bayesian nonparametric extensions of popular models, like topic models or finite mixture models, have proven promising for diverse statistical inference problems. These methods overcome the limitations of models with finite parametrizations, by allowing the effective cardinality or parameter size to increase as more data is observed. Recently, nonparametric extensions of Hidden Markov Models have garnered significant attention. This modelling scheme is suitable for sequential data encountered in a variety of real-world systems, which may not be well expressed with a finite number of hidden states. Nonparametric HMMs are derived from the theory of dirichlet processes (DPs). Using DPs, the unbounded parameters of these models can be implicitly integrating out to yield a finite number of hyperparameters. Learning and inference procedures defined for classical HMMs are not amenable to nonparametric HMMs, and several papers have lately derived efficient and scalable inference schemes for these models [2, 5, 3]. This has enabled novel applications of nonparametric HMMs in challenging problems, like speaker diarization, motion segmentation, modeling of genetic recombination etc.

The report is organized as follows: Section 2 introduces the Hidden Markov Models and Hierarchical Dirichlet Processes. Section 3 presents their joint treatment in the iHMM and its extensions. Section 4 discusses the inference algorithms for iHMM and HDP-HMM. Section 5 discusses the applications, limitations and future directions. Finally, section 5 presents the summary.

2 Preliminaries

2.1 Hidden Markov Models

The HMM is widely used for segmenting sequential data encountered in various domains such as speech recognition, genomics, stock markets, machine translation etc. It models a Markov process where states $\{s_1, s_2, ...s_T\}$ are not observed; what is observed are entities $\{y_1, y_2, ...y_T\}$ generated by these sequence of discrete states. Conditioned on this state sequence, the observations are assumed to be independent. An HMM is described by three probabilities:

- Transition Probability: Represents the probability of transitioning from state s_j to s_k : $\mathrm{P}(s_k|s_j)$
- Emission Probability: Represents the probability of generating observation y_k from state s_j: $\mathrm{P}(y_k|s_j)$
- Starting Probability: Represents the distribution for the initial state: $\mathrm{P}(s_1)$

A desirable property of classical HMMs is that the probabilities of state sequence given observations can be inferred using an efficient dynamic programming algorithm, known as the forward-backward algorithm. The parameters, i.e. the aforementioned probabilities, can be optimized with the Maximum Likelihood Estimation (MLE) criteria. MLE estimates are obtained within an expectation-maximization (EM) framework, known as the Baum-Welch algorithm, where expectations are computed with respect to the conditional distribution of hidden state sequence given observations $P(\mathbf{s}|\mathbf{y})$.

Existing modelling and inference scheme for HMMs suffer from two key limitations. First drawback is that standard HMMs require us to fix the cardinality of hidden

states in advance. A second limitation of MLE procedure itself is that it is prone to overfitting/underfitting.

2.2 Dirichlet processes

Dirichlet process is central to Bayesian nonparametrics. It defines a probability measure on distribution functions, where each draw from DP is a discrete random distribution itself. DPs are defined using a base distribution (H) and a concentration parameter (α). Formally, consider a distribution H over Θ and a parameter $\alpha \in R^+$. Let $(S_1, S_2, ...S_r)$ be any finite measurable partition of Θ. Then, G is a DP with base distribution H and concentration parameter α, written as $G \sim DP(\alpha, H)$ if,

$$(G(S_1), G(S_2), ..G(S_r)) \sim Dir(\alpha H(S_1), \alpha H(S_2), ...\alpha H(S_r)) \tag{1}$$

2.2.1 Mixture models

DP is commonly used as a prior in mixture models with unknown or countably infinite number of components. First, consider the following Bayesian mixture model with a finite number of K components,

$$\begin{aligned} z_i|\pi &\sim Multinomial(\pi), \qquad \theta_i|H \sim H \\ \pi|\beta &\sim Dir(\beta/k, ...\beta/k), \qquad x_i|z_i, \{\theta_i\}_{i=1}^{\infty} \sim F(\theta_{z_i}) \end{aligned}$$

Here,$\{x_1, x_2, ...x_N\}$ are the N observations which are modelled using discrete indicator variables $\{z_i\}$ taking on values {1,...K}. The mixing proportions π are assigned a Dirichlet prior with pseudocount hyperparameter β. H defines the prior distributiion over the parameters $\{\theta_k\}$ of the component likelihood F.

In the limit $K \to \infty$, the conditional probability of indicator variables after integrating out the DP prior is given as,

$$P(z_i = k | z_{\setminus i}, \beta) = \frac{N_{-d,k}}{N - 1 + \beta} \qquad N_{-d,k} = \sum_{l=1, l \neq i}^{N} \delta(z_l, k)$$

for every represented cluster k. Uninstantiated clusters are created with the remaining probability, $\frac{\beta}{n-1+\beta}$. The infinite parameters can thus be integrated out so that probability of indicator sequences is defined using the finite number of counts (as N is finite). Further, it can be seen that β has a special meaning – it reflects the tendency of the model to generate new components. DP prior is widely used in bayesian nonparametrics because of its infinite dimensionality, i.e., its ability to capture infinite components in a mixture model. It displays a bias towards existing hidden states, in a "rich get richer" fashion. Mixture models help to understand how DPs can be seen as an infinite dimensional generalization of Dirichlet distributions.

2.2.2 Stick breaking construction

Stick-breaking is a popular construction for DPs that helps to see how a DP defines distributions over discrete probability measures. Random draw from a DP, $G_0 \sim DP(\alpha_0, H)$, can be expressed as [8]:

$$G_0 = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k}, \qquad \theta_k | H \sim H, \delta : \text{Indicator function} \tag{2}$$

Here, the proportions π_k can be described using an iterative stick-breaking process dependent on the concentration parameter α_0. Suppose there is a stick of unit length. Let $\beta_k \sim Beta(1, \alpha_0)$ be the fractions we break from the remaining part of the stick at each iteration. π_k are then the fractional lengths from the resulting divisions of

this stick, denoted as $\pi \sim GEM(\alpha_0)$. Mathematically, this is expressed as,

$$\pi_k = \beta_k \prod_{c=1}^{k-1}(1 - \beta_c) \tag{3}$$

2.3 Hierarchical Dirichlet Processes

Consider a scenerio where we have groups of data, where each group can be modelled using distinct components or clusters. It may be desirable to share clusters across these groups, for example, when groups are produced by related or dependent tasks. This situation can be handled by Hierarchical Dirichlet Processes (HDPs) using a shared probability measure across groups. Formally, HDP first defines a base measure G_0 with a DP prior, $G_0 \sim DP(\alpha, H)$. Group distributions $\{G_1, ...G_j, ..G_J\}$ are then sampled from a global DP prior that uses the former base measure as the base distribution, $G_j \sim DP(\alpha_j, G_0)$. The use of a central base measure enables group distributions to share components.

3 Models

Hidden Markov Models can be considered a dynamic variant of the finite mixture models discussed above. This makes DPs a natural choice for extending HMMs to model infinite states. The latent states can be equated with components or clusters in the original DP mixture model formulation. It is important to note that HMMs involve a group of mixture models, each group denoting a distinct value of the current state. Each row of the transition matrix in HMMs, i.e., $P(s_{t+1}|s_t)$ can be modelled as a DP. Modelling each row using independent DPs is inadequate in HMMs for reasons discussed below. As it turns out, the use of hierarchical DPs alleviates this problem, thereby enabling inference in nonparametric HMMs.

3.1 Why do we need a hierarchy?

HDP forms the building block of HMMs. At the fundamental level, a hierarchy allows us to couple the transition dirichlet processes from different states. This is important because in the absence of coupling between different states, the set of states that are transitioned to from state a, for example, would not be the same as the set of states transitioned to from state b. Thus, the sequence will never visit the same state twice with this mechanism.

3.2 Infinite HMM

Beal et al.[1] proposed a two-level hierarchical model for the transition dynamics in HMMs, known as the infinite HMM (iHMM). Conditional on the current state (s_t), the next state (s_{t+1}) is modelled using a DP with concentration parameter β.

$$P(s_{t+1} = j | s_t = i, n, \beta) = \frac{n_{ij}}{\sum_{l=1}^{K} n_{il} + \beta}, \qquad n_{ij} = \sum_{t'=1}^{t-1} \delta(s_{t'}, i)\delta(s_{t'+1}, j) \qquad j \in \{1, ..., K\} \tag{4}$$

This top level DP thus favors typical trajectories, while allowing novel transitions with a finite probability $\frac{\beta}{\sum_{l=1}^{K} n_{il}+\beta}$. With this probability, the transitions are controlled by a second DP with a different concentration parameter γ.

$$P(s_{t+1} = j | s_t = i, n^0, \gamma) = \frac{n_j^0}{\sum_{l=1}^{K} n_l^0 + \gamma} \tag{5}$$

This DP and its counts n^0 are known as the oracle. Here, n_l^0 represents the number of times the state l was transitioned to from an oracle DP. The oracle counts

are updated separately from the transition counts, n, of the first-level DP. Under the second-level DP, a completely new state is transitioned to with a finite probability of $\frac{\gamma}{\sum_{l=1}^{K} n_l^0 + \gamma}$.

The model also introduces an interesting self-transitioning bias by initializing the self-transition counts to a finite value denoted by α. The transition dynamics in this model are thus completely controlled by three parameters: (a) α controls the probability for self transitions, (b) β controls the sparsity of transition matrix by influencing the tendency to explore new transitions, and (c) γ controls the expected number of states by influencing the probability with which a new states is sampled. These three parameters represent the priors for the transition dynamics, and are capable of creating a wide variety of state trajectories.

The emission mechanism is modelled analogously, except that there is no notion of self transitions. Thus, there are only two parameters for describing emissions $\{\alpha_e, \beta_e\}$.

3.3 HDP-HMM

The former approach is not strictly a hierarchical DP in the bayesian sense because the DP parameters for all groups (i.e., each row of the transition matrix) are not derived from a common base distribution. Rather, the hierarchy is imposed through a coupling between transition DPs using oracles. Teh et al.[2] introduced a formal Hidden Dirichlet Process - Hidden Markov Model (HDP-HMM), that has now been widely applied across various domains, such as speaker diarization[4], visual scene recognition[6], gene expression [7] etc. Their generative approach for different variables is formulated mathematically using the stick-breaking construction,

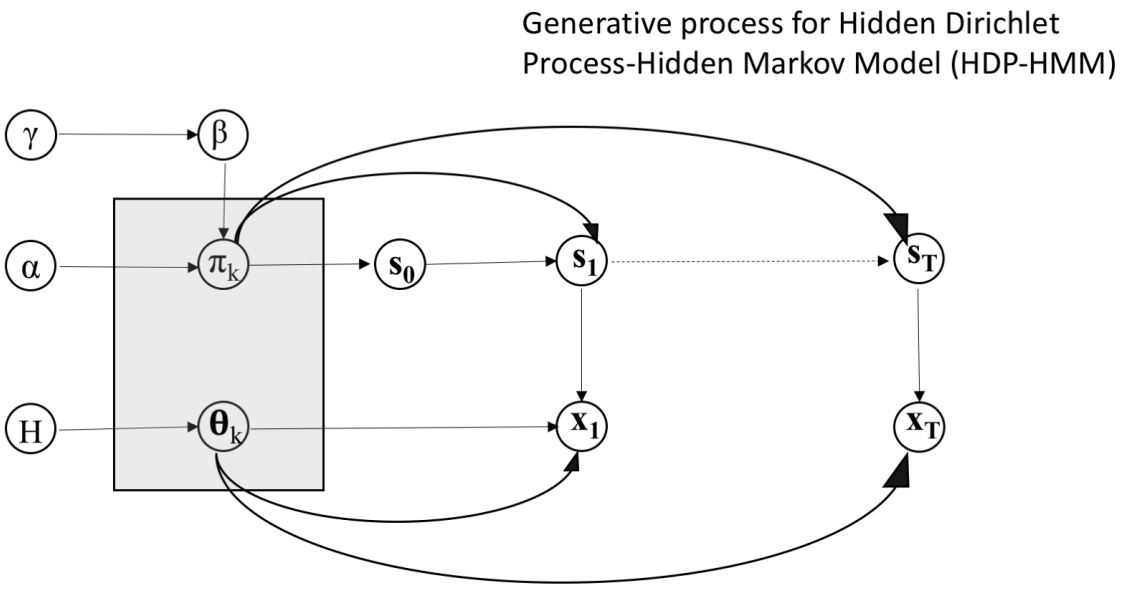

Figure 1: A schematic of HDP-HMM

$$\text{Stick-breaking prior:} \qquad \beta|\gamma \sim GEM(\gamma)$$

$$\text{State-specific transition distribution:} \qquad \pi_k|\alpha_0, \beta \sim DP(\alpha_0, \beta), \qquad k = 1, 2....$$

$$\text{State-specific emission parameters:} \qquad \theta_k|H \sim H, \qquad k = 1, 2....$$

$$\text{Hidden states:} \qquad s_t|\{\pi_k\}_{k=1}^{\infty}, s_{t-1} \sim \pi_{s_{t-1}}, \qquad t = 1, ..., T$$

$$\text{Observations:} \qquad x_t|s_t, \{\theta_k\}_{k=1}^{\infty} \sim F(\theta_{s_t}) \qquad t = 1, ..., T$$

The generative process for HDP-HMM is depicted in Figure 1.

3.4 Sticky HDP-HMM

Fox et al. [3] propose an easy extension of the HDP-HMM to incorporate a self transition bias, similar to Beal et al. [1]. They show that in the absence of an extra

self transition bias, the model creates redundant states with rapid transitions among them. To combat this, they propose the following modification to the transition distribution,

$$\begin{aligned} \beta|\gamma &\sim GEM(\gamma) \\ \pi_k|\alpha_0, \beta, \kappa &\sim DP(\alpha_0 + \kappa, \frac{\alpha_0\beta + \kappa\delta_k}{\alpha_0 + \kappa}), \qquad k = 1, 2... \end{aligned}$$

Intuitively, this increases the probability of self transitions by an amount proportional to κ. Their model, known as the sticky HDP-HMM, is especially useful for data with state persistence such as audio recordings. This approach showed impressive performance in speaker diarization.

4 Inference and learning

In the terminology of HMMs, inference refers to deriving the hidden state sequence given model parameters, whereas learning corresponds to finding the parameters of the probability distributions (transition or emission distributions). As described above, inference in classical HMMs is based on an efficient dynamic programming algorithm, whereas the learning procedure relies on an expectation maximization algorithm. Inference in nonparametric HMM is not as straightforward. Most current learning and inference schemes for nonparametric HMMs are based on Gibbs sampling.

4.1 Approximate Gibbs sampling

The original implementation of iHMM relies on an approximate Gibbs sampling procedure for both learning and inference. Instead of computing full conditionals, Beal

et al. [1] proposed an approximate inference strategy for iHMMs where the sample updates for state s_t are based only upon the neighbors (s_{t-1}, s_{t+1}, y_t) and transition/emission counts after excluding the counts contributed by s_t. This reduces the computational operations to be linear in the number of time points. The sampled hidden state sequence are used to obtain the posterior probabilities, $\mathrm{P}(v|\{s_t\}_{t=1}^T)$ of the hyperparameters v, where $v = \{\alpha, \beta, \gamma, \beta_e, \gamma_e\}$. The values of these hyperparameters are updated based on their maximum a posteriori (MAP) estimate.

4.2 Gibbs sampling with auxiliary variables

In contrast to iHMMs, the HDP-HMM defines transitions using well-defined underlying probability measures or priors, thereby enabling a full posterior bayesian inference. In the HDP-HMM formalism, Teh et al. [2] propose an efficient Gibbs sampling technique with the use of auxiliary variables. Their sampling scheme is general for all HDPs. The variables of interest in HDP-HMM are $v = \{\{s_t\}_{t=1}^T, \{\pi_k, \theta_k\}_{k=1}^L, \beta\}$. The authors first propose to use a finite representation of the posteriors. Only the instantiated components K out of the possibly infinite L components are explicitly modelled, whereas all the the unrepresented components are pooled together as a single variable u. Thus, assuming the stick-breaking weights are arranged in order, we can set $\beta_u = \sum_{k=K+1}^{L} \beta_k$ and take $\beta = (\beta_1, ..., \beta_K, \beta_u)$. In this scenario, we need to only record the transition counts n_{jk} for 1 $\leq$k$\leq$K. Since π $\sim$DP, it can be integrated out from the conditional probabilities. The analytical expressions for the remaining conditional are described below. For notational convenience in subsequent formulae, the superscripted elements in front of the minus sign represent indices that are excluded from the corresponding sequence.

- Sampling θ_k: The posterior of θ_k given $\{ s_t\}_{t=1}^T$, $\{x_t\}_{t=1}^T$ and the prior H is given

as,

$$P(\theta_k|\{s_t\}_{t=1}^T, \beta, \{x_t\}_{t=1}^T, \theta^{-k}) \propto H(\theta_k) \prod_{t:s_t=k} F(x_t|\theta_k)$$

This posterior is directly used for sampling $\{\theta_k\}_{k=1}^K$

- Sampling s: Conditional probability of s_t for posterior sampling is given as,

$$P(s_t = k|s^{-t}, \beta, \{x_t\}_{t=1}^T, \{\theta_k\}_{k=1}^u) \propto (\alpha_0\beta_k + n_k^{-t})F(x_t|\theta_k), \qquad k = 1, ..., K, u$$

- Sampling β: An auxiliary variable method is used for sampling β. Formally, an auxiliary variable m is introduced that describes a backward message pass from state s_t to s_{t-1},

$$m_{t,t-1}(s_{t-1}) \propto \sum_{s_t} P(s_t|\pi_{s_{t-1}})F(x_t|\theta_{s_t})m_{t+1,t}(s_t) \qquad t \leq T$$

This variable allows sampling of β as $\beta \sim Dir(\gamma/L + m_{.1}, ..., \gamma/L + m_{.L})$

5 Summary

Nonparametric HMMs are particularly useful because of their flexibility and the ability to capture rich unconstrained structures in the sequential data. This is also demonstrated by their ability to accurately predict ground truth temporal segmentations in various application domains. Inference in nonparametric HMMs is challenging and most current methods rely on Gibbs sampling for both parameter learning and state inference. For time-series data, Gibbs sampling procedures generally show poor mixing with very slow convergence rates due to strong dependencies between consec-

utive time steps. A few inference methods have been explored as an alternative to Gibbs sampling in the infinite-state HMM setting. Stochastic and memoized variational inference methods recently demonstrated the scalability of these models to handle large sequences [5].

6 Notations

Dir(β) Dirichlet distribution with parameter β

DP(α, H) Dirichlet Process with concentration α and base distribution H

δ_{θ_k} Indicator function: $\delta_{\theta_k}(x) = 1$ when $x = \theta_k$, 0 otherwise

GEM(γ) Stick-breaking construction with parameter γ

References

[1] Matthew J. Beal, Zoubin Ghahramani and Carl Edward Rasmussen. *The infinite Hidden Markov Model*. Advances in Neural Information Processing Systems, 2001.

[2] Yee Whye Teh, Michael I. Jordan, Matthew J. Beal and David M. Blei. *Hierarchical Dirichlet Processes*. Journal of the American Statistical Association, 101:476, 1566-1581, DOI: 10.1198/016214506000000302.

[3] Emily B. Fox, Eric B. Sudderth, Michael I. Jordan and Alan S. Wilsky. *An HDP-HMM for systems with state persistence*. Proceedings of the 25th International Conference on Machine learning, 2008.

[4] Emily B. Fox, Eric B. Sudderth, Michael I. Jordan and Alan S. Wilsky. *A sticky HDP-HMM with application to speaker diarization*. Ann. Appl. Stat. 5 (2011), no. 2A, 1020–1056. doi:10.1214/10-AOAS395.

[5] Michael C. Hughes, William Stephenson and Eric B. Sudderth. *Scalable adaptation of state complexity for nonparametric hidden Markov models*. Advances in Neural Information Processing Systems, 2015.

[6] Jyri J. Kivinen, Erik B. Sudderth and Michael I. Jordan. *Learning Multiscale Representations of Natural Scenes Using Dirichlet Processes*. IEEE International Conference on Computer Vision, 2007.

[7] Matthew J. Beal and P. Krishnamurthy. *Gene expression time course clustering with countably infinite hidden Markov models*. In Proc. Conference on Uncertainty in Artificial Intelligence, 2006.

[8] J. Sethuraman. *A constructive definition of Dirichlet priors*. Statistica Sinica, 4: 639–650, 1994