NONPARAMETRIC LEARNING FOR HIDDEN MARKOV MODELS WITH PREFERENTIAL ATTACHMENT DYNAMICS

Asher A. Hensley and Petar M. Djurić

Department of Electrical and Computer Engineering
Stony Brook University, Stony Brook, NY 11794
asher.hensley@stonybrook.edu, petar.djuric@stonybrook.edu

ABSTRACT

We address the learning problem for infinite state Hidden Markov Models (HMMs) with preferential attachment dynamics. Preferential attachment describes a "rich get richer" process causing the HMM self transition probabilities to be proportional to the number of previous self transitions. Furthermore, the length of stay of the process in a particular state follows the Yule-Simon distribution. In describing the generative model of the hidden state processes, we use nonparametric models. We also establish the relationship of the proposed model with the Polya urn scheme and the Chinese restaurant process. The class of HMMs from this paper are applicable to data sets where the time spent in each state follows a power law. Our objective is to estimate the state sequence and the model parameters of the HMM. To that end, we propose a Gibbs sampling procedure. We evaluate the proposed procedure through computer simulations.

Index Terms— Gibbs sampling, power law, Yule-Simon distribution, Chinese Restaurant process, Polya urn

1. INTRODUCTION

In this paper we present a Bayesian nonparametric learning strategy for infinite state Hidden Markov Models (HMMs) with preferential attachment dynamics. Preferential attachment describes a "rich get richer" process commonly associated with the Barabasi-Albert model used for modeling edges in scale-free networks [1]. This class of HMMs offers a convenient framework for modeling time series characterized by long durations of stable statistical behavior, punctuated by abrupt jumps. In particular, these HMMs are relevant for situations where the time spent in each latent state follows a power law. The motivation here is to model unexpected events occurring after long durations of predictable behavior, such as moves in financial markets.

What differentiates this HMM from traditional HMMs is that, despite there being an infinite number of states, only two transitions are possible from each state: (1) a self transition and (2) a new state transition. Because preferential attachment dynamics are in play, the self transition probability is proportional to the number of previous self transitions, causing new state transition probabilities to decrease over time. The rate at which these probabilities change are controllable via a hyper parameter. When a new state is created, the new state parameters are randomly drawn from their prior, and the transition probabilities are reset to their initial conditions.

The number draws before a state transition in these HMMs follow a Yule-Simon (YS) distribution, which is central for deriving the Gibbs sampling update equations. This distribution was first proposed by Yule in 1925 for modeling the size distribution of biological genera [2]. The same result was also derived by Simon in 1955 using an alternative approach to model word frequency distributions in text [3]. Interestingly, the results of Yule and Simon have also been rediscovered many times as described in [4]. However, to our knowledge the YS distribution has never been used in the HMM context considered in this paper.

As discussed later, it will be necessary to infer the parameter (i.e., the power law exponent) of the YS distribution to facilitate the sampling of other latent variables. Standard methods exist for estimating arbitrary power law exponents as described in [5], although maximum likelihood methods specific to the YS distribution have also been proposed (see [6]). For our case, a Bayesian approach would be preferable, which has only recently been been pursued by Leisen et al. in [7] and [8]. In their words, "the Bayesian literature, so far, ignored this distribution" [7].

In [7], two priors for the YS distribution are used: (1) a Jeffreys prior and (2) a loss-based prior. Experimental results are then provided for social network stock indexes, Census surname data, and #1 hits on the Billboard hot 100 chart. In their work, the authors use an MCMC inference strategy. The main finding of the paper is that both priors lead to similar results. However, it is not clear how accurate the results are.

In [8], the authors give a Gibbs sampling scheme for the YS parameter when a Gamma prior is used. In doing so, they show that the full conditional distribution of the YS parameter can be computed explicitly by introducing an auxiliary

This work has been supported by NSF under Awards CCF-1320626 and CCF-1618999.

variable. Results from synthetic data simulations show that their inference procedure can reliably estimate the YS parameter. Experimental results are also given for a word frequency modeling problem. The authors show that the mean of the posterior is identical to the maximum likelihood estimate given by Garcia's method [6].

The YS distribution has also seen some use as a prior distribution in various hierarchical models. In [9], it is employed as a prior for the number of random mating units within a population structure. In [10], the YS distribution is a prior for the number of trials in the multinomial distribution that models outliers in Twitter data. However, in these papers little if any emphasis was placed on inferring the YS parameter.

Other nonparametric learning approaches have also been used for infinite state HMMs. Teh et al. introduced the use of Hierarchical Dirichlet Processes in [11] as a prior for the learning problem (HDP-HMM). The algorithm, however, seems to learn more states than necessary to explain the data. Fox et al. addressed this problem by introducing a "sticky" factor in [12] causing the inference algorithm to prefer self transitions. However these models are inappropriate for the power law behavior addressed in this paper.

In this paper, we use an infinite state HMM with preferential attachment dynamics as a means to model time series with long runs of stable behavior followed by unexpected jumps governed by the YS distribution. Our main contribution is a Gibbs sampling algorithm that solves the learning problem. Our algorithm offers an alternative to the HDP-HMM of Teh [11] and Fox [12] with less book keeping for this special class of systems. We demonstrate the performance of the algorithm via computer simulations.

2. MODEL

We consider the inference problem for the following process:

$$\alpha | a, b \sim \text{Ga}(a, b), \tag{1}$$

$$\lambda_j | c, d \sim \text{Ga}(c, d), \tag{2}$$

$$s_t | p_t \sim \text{Ber}(p_t), \quad s_t \in [0, 1] \tag{3}$$

$$z_t = z_{t-1} + s_t, \tag{4}$$

$$x_t | z_t, \boldsymbol{\lambda} \sim \mathcal{N}(0, \lambda_{z_t}^{-1}), \tag{5}$$

where $\mathcal{N}(\mu, \sigma^2)$ denotes the normal distribution with mean μ and variance σ^2, $\text{Ga}(u, v)$ is the gamma distribution with shape parameter u and rate parameter v, and $\text{Ber}(p)$ is the Bernoulli distribution with success probability p with:

$$p_t = \frac{\alpha}{n_j + \alpha}, \tag{6}$$

where n_j is the number of samples the model has remained in the current state j, and p_t is the probability of transition to a new state. If a new state $j + 1$ is created, a new measurement precision λ_{j+1} is drawn from the prior and the count n_{j+1} is initialized to 1. Note that once a state transition occurs, the new state is always created independently of all past states.

It is simple to show that the number of times a state is revisited before transitioning follows the YS distribution,

$$p(n_j = k | \alpha) = \alpha \text{B}(k, \alpha + 1), \tag{7}$$

where $\text{B}(x, y)$ denotes the beta function. The mean and variance of the YS distribution are given by

$$\mathbb{E}(n_j | \alpha) = \frac{\alpha}{\alpha - 1}, \quad \alpha > 1, \tag{8}$$

$$\text{Var}(n_j | \alpha) = \frac{\alpha^2}{(\alpha - 1)^2 (\alpha - 2)}, \quad \alpha > 2. \tag{9}$$

The inference problem we want to solve is to draw samples from the posterior distribution of the latent variables α, $\mathbf{z}$, and $\boldsymbol{\lambda}$ conditioned on a measurement sequence $\mathbf{x}$. It is assumed that the hyper-parameters a, b, c and d are known.

3. RELATIONSHIP TO THE POLYA URN SCHEME AND THE CHINESE RESTAURANT PROCESS

Some readers may recognize the process of states as a modified Polya urn scheme [13]. Suppose we have a Polya urn with two ball colors (say black = self transition and red = new state transition). For each measurement, we draw a ball from the urn. If the ball is black, we stay in the same state and add another black ball to the urn. If the ball is red, we reset the urn to it's initial condition of black and red balls, and then we randomly draw a ball that provides a new state value. It is interesting to point out that a connection between Simon's model and Polya's urn was made by Price in 1975 [14] while studying distributions of bibliometric citations.

Our state process is also related to other nonparametric models such as the Chinese Restaurant Process (CRP) [15], Chinese Restaurant Franchise [11], and Dirichlet Process [16]. Imagine we have an infinite number of Chinese restaurants (i.e., a franchise) each with just *one* table. In the standard CRP, if there are customers at the 1st table only, there is some probability that a new customer will choose to sit at a second table. However, because there is only one table in each restaurant, instead of starting a new table the new customer will be directed to a new restaurant and sit at its first and only table. This process will then repeat forever.

4. INFERENCE

Here we outline a Gibbs sampling procedure designed to draw samples from the latent variable posterior conditioned on the observed data. The general inference procedure is given in **Algorithm 1**. Over the next several subsections we describe the steps of Algorithm 1 in more detail.

```
□ Algorithm 1: Main Gibbs Sampler
  Initialize: $\alpha, \boldsymbol{\lambda}, \mathbf{z}, a, b, c, d$
  Repeat:
    for $t = 1, ..., T$
      Draw $z_t \sim p(z_t | \mathbf{z}_{-t}, \boldsymbol{\lambda}, x_t, \alpha)$
    end
    Draw $\alpha \sim p(\alpha|\mathbf{z}, a, b)$
    Draw $\lambda_j|$ all $x_t$ assigned to state $j$
```

4.1. Initialization

The algorithm can either be initialized from scratch or from the output of a previous Gibbs sampling run. In either case, we will assume $a, b, c,$ and d are known. When beginning from scratch, we set α to an arbitrary starting value. The state assignments $\mathbf{z}$ and parameters $\boldsymbol{\lambda}$ are then initialized using **Algorithm 2**:

```
□ Algorithm 2: $\mathbf{z}$, $\boldsymbol{\lambda}$ Initialization
  Initialize: $\alpha, a, b, c, d$
  Assign $z_1 = n_1 = j = 1$
  Draw $\lambda_1|x_1$
  for $t = 2, ..., T$
    Draw $z_t \sim p(z_t|n_j, \lambda_j, x_t, \alpha)$
    if $z_t = j$
      Assign $n_j = n_j + 1$
      Draw $\lambda_k|$ all $x_t$ assigned to state $j$
    elseif $z_t = j + 1$
      Assign $j = j + 1$ and $n_j = 1$
      Draw $\lambda_j|x_t$
    end
  end
  Return $\mathbf{z}$, $\boldsymbol{\lambda}$
```

where

$$p(z_t = j|n_j, \lambda_j, x_t, \alpha) \propto \frac{n_j}{n_j + \alpha}\mathcal{N}(x_t|0, \lambda_j^{-1}), \tag{10}$$

and

$$p(z_t = j+1|n_j, \lambda_j, x_t, \alpha) \propto \frac{\alpha}{n_j + \alpha}\text{St}(x_t|0, c/d, 2c). \tag{11}$$

The notation $\text{St}(x_t|\eta, \tau, \nu)$ denotes a Student's-*t* distribution with mean η, precision τ, and degrees of freedom ν.

4.2. Sampling z_t

Sampling the state indicator variables z_t requires examining a few different cases, with examples given in Fig. 2. Thus, when running Algorithm 1, the sampling case must be determined first, before drawing an outcome and updating z_t. Here we describe the outcome probabilities for each case while assuming the last sample of $z_t = j$.

Right Boundary Case:

$$p(z_t = j|\mathbf{z}_{-t}, \boldsymbol{\lambda}, x_t, \alpha) \propto \frac{n_j - 1}{n_j + \alpha}\mathcal{N}(x_t|0, \lambda_j^{-1}),$$

$$p(z_t = j+1|\mathbf{z}_{-t}, \boldsymbol{\lambda}, x_t, \alpha) \propto \frac{n_{j+1}}{n_{j+1} + \alpha + 1}\mathcal{N}(x_t|0, \lambda_{j+1}^{-1}),$$

$$p(z_t = k|\mathbf{z}_{-t}, \boldsymbol{\lambda}, x_t, \alpha) \propto \frac{\alpha}{1 + \alpha}\text{St}(x_t|0, c/d, 2c).$$

Left Boundary Case:

$$p(z_t = j|\mathbf{z}_{-t}, \boldsymbol{\lambda}, x_t, \alpha) \propto \frac{n_j - 1}{n_j + \alpha}\mathcal{N}(x_t|0, \lambda_j^{-1}),$$

$$p(z_t = j-1|\mathbf{z}_{-t}, \boldsymbol{\lambda}, x_t, \alpha) \propto \frac{n_{j-1}}{n_{j-1} + \alpha + 1}\mathcal{N}(x_t|0, \lambda_{j-1}^{-1}),$$

$$p(z_t = k|\mathbf{z}_{-t}, \boldsymbol{\lambda}, x_t, \alpha) \propto \frac{\alpha}{1 + \alpha}\text{St}(x_t|0, c/d, 2c).$$

Double Boundary Case:

$$p(z_t = j-1|\mathbf{z}_{-t}, \boldsymbol{\lambda}, x_t, \alpha) \propto \frac{n_{j-1}}{n_{j-1} + \alpha + 1}\mathcal{N}(x_t|0, \lambda_{j-1}^{-1}),$$

$$p(z_t = j+1|\mathbf{z}_{-t}, \boldsymbol{\lambda}, x_t, \alpha) \propto \frac{n_{j+1}}{n_{j+1} + \alpha + 1}\mathcal{N}(x_t|0, \lambda_{j+1}^{-1}),$$

$$p(z_t = k|\mathbf{z}_{-t}, \boldsymbol{\lambda}, x_t, \alpha) \propto \frac{\alpha}{1 + \alpha}\text{St}(x_t|0, c/d, 2c).$$

After drawing each new sample of z_t, some book keeping will be required. For example, when adding/deleting state indexes, existing state indexes will need to be modified for continuity. Additionally, when a new state is created (k is drawn), a new value of λ must also be drawn. The necessary steps will depend on the implementation of the algorithm. Note that if z_t does not match one of the above cases, it is not updated, with the exception of the first and last measurements being special cases of one of the above cases.

4.3. Sampling α

For sampling α, we use the Gibbs sampler based on the method of Leisen et al. which exploits the YS distribution (see (7)) as shown in **Algorithm 3** (see [8]):

```
□ Algorithm 3: Alpha Gibbs Sampler
  Initialize: $\alpha, n_1, n_2, ..., n_L, a, b$
  Repeat:
    Draw $w_j = \text{Beta}(\alpha + 1, n_j) \quad j = 1, ..., L$
    Draw $\alpha \sim \text{Ga}(a^*, b^*)$
```

where Beta(u, v) is the beta distribution and,

$$a^* = a + L \quad \text{and} \quad b^* = b - \sum_{j=1}^{L} \ln w_j \tag{12}$$

4.4. Sampling λ

To sample λ_j conditioned on a set of N measurements $\{x_t, x_{t+1}, ..., x_{t+N-1}\}$, we have

$$\lambda_j|x_t, x_{t+1}, ..., x_{t+N-1} \sim \text{Ga}(c^*, d^*), \tag{13}$$

where

$$c^* = c + \frac{N}{2} \quad \text{and} \quad d^* = d + \frac{1}{2}\sum_{i=t}^{t+N-1} x_i^2. \tag{14}$$

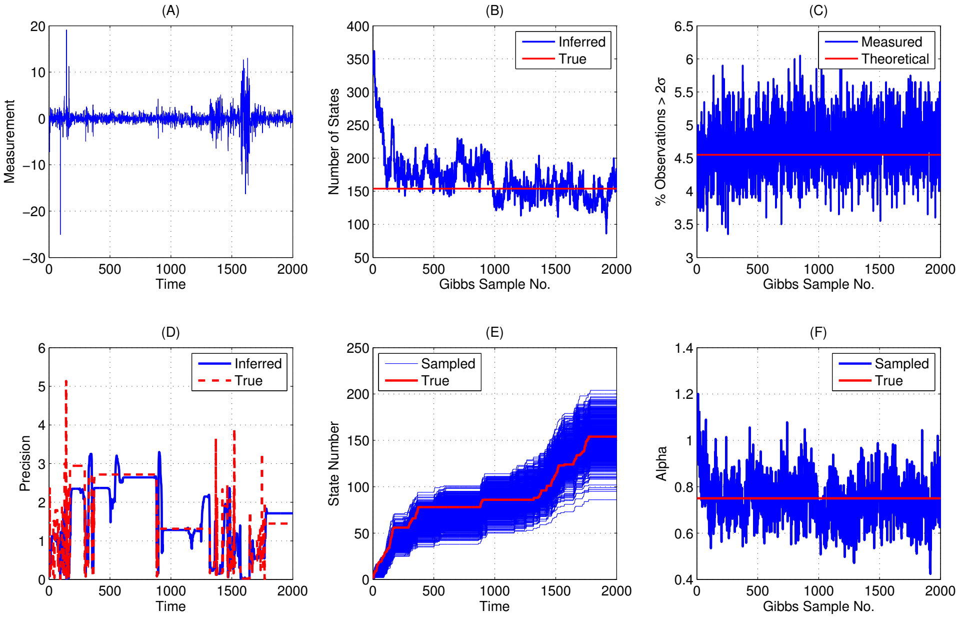

Fig. 1. (A) The measurement sequence. (B) The number of sampled states as a function of Gibbs sample number. (C) The percentage of measurements greater than 2 standard deviations based on the sampled precisions for each state. (D) The average sampled precisions (after burn in) vs. the true precisions. (E) The sampled latent state sequences (after burn in) vs the true latent state sequence. (F) The sampled alpha values as a function of Gibbs sample number.

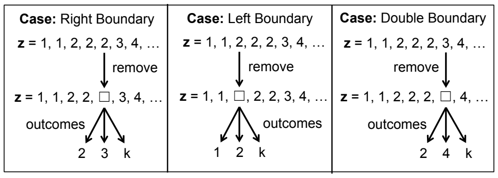
Case: Right Boundary
z = 1, 1, 2, 2, 2, 3, 4, …
remove
z = 1, 1, 2, 2, □, 3, 4, …
outcomes
2 3 k
Case: Left Boundary
z = 1, 1, 2, 2, 2, 3, 4, …
remove
z = 1, 1, □, 2, 2, 3, 4, …
outcomes
1 2 k
Case: Double Boundary
z = 1, 1, 2, 2, 2, 3, 4, …
remove
z = 1, 1, 2, 2, 2, □, 4, …
outcomes
2 4 k

Fig. 2. Gibbs sampling state indicator examples

5. SIMULATIONS

To test our inference algorithm(s) we generated a set of 2000 synthetic measurements using equations (1)-(5), with $a, b, c, d = 1$ and $\alpha = 0.75$. This set of measurements is shown in Fig. 1-A. We then ran Algorithm 2 to initialize $\mathbf{z}$ and $\boldsymbol{\lambda}$, where we initialized $\alpha = \mathbb{E}(\alpha|a, b) = 1$.

Next, we ran Algorithm 1 for 2,000 iterations. For each iteration, Algorithm 3 was also run for 999 iterations (burn-in), and the $1,000^{th}$ sample was returned to Algorithm 1 before proceeding. Each iteration of Algorithm 1 yielded a slightly different partition of the measurement sequence. This number of partitions as a function of Gibbs sample is shown Fig. 1-B, where the true number of partitions is shown as a horizontal red line. Fig. 1-E shows where each state change (i.e., partition) occurred as a function of time, each step being a state transition. Note that the state indexes increase linearly, meaning that the final state index ($t = 2,000$) is also the number of partitions as shown in Fig. 1-B.

Next we plotted the average precisions (i.e., λ_j) vs. the true precisions as a function of time in Figure 1-D. This was done by discarding the first 1,000 Gibbs samples and then averaging the remaining Gibbs samples. To gauge the accuracy of the sampled precision values as a function of Gibbs sample, we plotted the percent of measurements greater than two standard deviations away from the mean in Fig. 1-C based on their respective precisions. Note that the theoretical value (shown in red) is based on an infinite number of points. Finally, the sampled alpha values as a function of Gibbs sample number are shown in Fig. 1-F.

6. CONCLUSIONS

In this paper we have presented a nonparametric learning strategy for infinite state HMMs with preferential attachment dynamics. These models are particularly relevant for time series exhibiting long runs of stable behavior with erratic jumps, such as those seen in finance. The presented Gibbs sampling algorithms provide a straightforward way to conduct batch inference on these types of systems. Future work will be necessary to make predictions about new measurements and move towards an online recursive learning approach. Generalizations of this include accounting for time varying α, heteroscedastic observations, and unknown values for the hyper parameters a, b, c, and d. Additionally, state parameters could be modified to evolve according to a dynamical system.

7. REFERENCES

[1] A. L. Barabási and R. Albert, "Emergence of scaling in random networks," *Science*, vol. 286, no. 5439, pp. 509–512, 1999.

[2] G.U. Yule, "A mathematical theory of evolution, based on the conclusions of Dr. J. C. Willis, F. R. S.," *Philosophical Transactions of the Royal Society of London. Series B*, vol. 213, pp. 21–87, 1925.

[3] H. A. Simon, "On a class of skew distribution functions," *Biometrika*, vol. 42, no. 3/4, pp. 425–440, 1955.

[4] M. V. Simkin and V. P. Roychowdhury, "Re-inventing Willis," *Physics Reports*, vol. 502, no. 1, pp. 1–35, 2011.

[5] A. Clauset, C. R. Shalizi, and M. E. J. Newman, "Power-law distributions in empirical data," *SIAM Review*, vol. 51, no. 4, pp. 661–703, 2009.

[6] J. M. G. Garcia, "A fixed-point algorithm to estimate the Yule–Simon distribution parameter," *Applied Mathematics and Computation*, vol. 217, no. 21, pp. 8560–8566, 2011.

[7] F. Leisen, L. Rossini, and C. Villa, "Objective Bayesian analysis of the Yule-simon distribution with applications," *arXiv preprint arXiv:1604.05661*, 2016.

[8] F. Leisen, L. Rossini, and C. Villa, "A note on the posterior inference for the Yule-Simon distribution," *arXiv preprint arXiv:1604.07304*, 2016.

[9] J. Corander, M. Gyllenberg, and T. Koski, "Random partition models and exchangeability for Bayesian identification of population structure," *Bulletin of Mathematical Biology*, vol. 69, no. 3, pp. 797–815, 2007.

[10] V. K. Pillutla, Z. Fang, P. Devineni, D. Koutra, C. Faloutsos, and J. Tang, "On skewed multi-dimensional distributions: the FusionRP model, algorithms, and discoveries," *unpublished*.

[11] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei, "Hierarchical Dirichlet processes," *Journal of the American Statistical Association*, 2012.

[12] E. Fox, E. B. Sudderth, M. I. Jordan, and A. S. Willsky, "Nonparametric Bayesian learning of switching linear dynamical systems," in *Advances in Neural Information Processing Systems*, 2009, pp. 457–464.

[13] A. A. Markov, "Extension of the law of large numbers to dependent quantities," *Izv. Fiz.-Matem. Obsch. Kazan Univ.(2nd Ser)*, vol. 15, pp. 135–156, 1906.

[14] D. S. Price, "A general theory of bibliometric and other cumulative advantage processes," *Journal of the American Society for Information Science*, vol. 27, no. 5, pp. 292–306, 1976.

[15] D. J. Aldous, "Exchangeability and related topics," in *École d'Été de Probabilités de Saint-Flour XIII1983*, pp. 1–198. Springer, 1985.

[16] T. S. Ferguson, "A Bayesian analysis of some nonparametric problems," *The Annals of Statistics*, pp. 209–230, 1973.