ON GENERATIVE MODELS FOR SEQUENTIAL FORMATION OF CLUSTERS

Petar M. Djurić and Kezi Yu

Department of Electrical & Computer Engineering
Stony Brook University
Stony Brook, NY 11794, USA
e-mail:petar.djuric@stonybrook.edu, kezi.yu@stonybrook.edu

ABSTRACT

In the literature of machine learning, a class of unsupervised approaches is based on Dirichlet process mixture models. These approaches fall into the category of nonparametric Bayesian methods, and they find a wide range of applications including in biology, computer science, engineering, and finance. An important assumption of the Dirichlet process mixture models is that the data are exchangeable. This is a restriction for many types of data whose structures vary over time or space or some other independent variables. In this paper, we address generative models that remove the restriction of exchangeability of the Dirichlet process model, which allows for creation of mixtures with time-varying structures. We also address how these models can be applied to sequential estimation of clusters.

***Index Terms*—** machine learning, Dirichlet processes, time-varying clustering, Chinese restaurant processes with finite capacities

1. INTRODUCTION

A very important task in machine learning is the clustering of observed data [1]. The process of clustering amounts to grouping the data according to their features. A standard approach to clustering is to represent the data according to a mixture distribution where the number of mixing components (mixands) is known beforehand. Each mixand is a representative of a cluster. Typically the parameters of the mixands are unknown, and they have to be estimated during the clustering. A well-known approach to solving this problem is the expectation-maximization (EM) algorithm [2].

When the number of clusters is unknown, one can proceed with hypothesizing that the number of classes is $k = 1, 2, \cdots, K_{\max}$. Upon completing the clustering under each of these hypotheses, one follows up with a model selection and chooses the "best" model of the $K_{\max}$ models. There are also approaches where the maximum number of classes is not predefined and one employs Markov chain Monte Carlo (MCMC) sampling. Well known methods in this category are the reversible jump MCMC sampling [3] and the birth-and-death MCMC sampling [4].

An alternative to these approaches is to use models that are based on Dirichlet process mixtures (DPMs). The DP has been widely used in Bayesian modeling [5], and particularly, as a basis for forming DPMs [6]. DPMs have become quite popular for data clustering in a wide range of disciplines. An early work on their use is presented in [7].

This work was supported by NSF under Award CCF-1320626.

An important assumption of these models is that the data are exchangeable. This entails that there is no change in the structures of the data over time or space or some other independent variables. In reality, many data do not satisfy this assumption. Furthermore, in many settings one wants to process the data sequentially and capture from them how they vary with time. In this paper we address generative models that remove the restriction of exchangeability that the DP model imposes on them. We investigate the first two moments of these processes over time and show how one can infer their parameter(s). We explain them by using the metaphor of a Chinese restaurant process (CRP) with finite capacity.

The paper is organized as follows. In the next section we provide a brief background of DPs. In Section 3, we present generative models of processes that are derived from DPs. In the following section, we investigate the first two moments of these processes as a function of time. In Section 5, we propose methods for estimating the unknown parameter(s) of the process(es). We conclude the paper with final remarks in Section 6.

2. BACKGROUND

A DP is an extension of the Dirichlet distribution and its realizations represent probability distributions. A DP is defined by a concentration parameter α and a base measure H [5].

In the sequel, for easier explanation of the addressed processes, we use the CRP metaphor, and thus, we use it to present the DP. First, suppose that there is a restaurant where customers come and stay for *unlimited* time. Customers are coming to the restaurant one at a time and are seated according to a specific random mechanism. At each table a different dish is served. In our setting, the customers are represented by data samples and the dishes served at the tables are the classes of the data.

The process starts with the arrival of the first customer. This customer is always seated at table 1 (class 1). Then a "dish" θ_1 is drawn for this customer from H. The following customer is seated at table 1 or table 2 with probabilities

$$p_{2,1} = \frac{1}{1+\alpha}, \tag{1}$$

$$p_{2,2} = \frac{\alpha}{1+\alpha}. \tag{2}$$

Once a customer is seated at a new table, a new dish that is served on that table is drawn from the base distribution. We note that each table has its own dish, represented by θ_i. Thus, each table corresponds to a different class, and each table has its own dish (which are actually the parameters of the class).

The following customers are seated according to a probability mass function defined by the occupancy of the tables by the previous customers. Suppose that when customer n enters the restaurant, tables 1 to m_{n-1} are occupied with $k_{n-1,i}$, $i = 1, 2, \cdots, m_{n-1}$ customers, respectively. Then the probability of seating the nth customer at the ith table is

$$p_{n,i} = \frac{k_{n-1,i}}{n-1+\alpha}, \quad i = 1, 2, \cdots, m_{n-1}, \tag{3}$$

and the probability of seating that customer at a new table, $m_{n-1}+1$, is

$$p_{n,m_{n-1}+1} = \frac{\alpha}{n-1+\alpha}. \tag{4}$$

Thus,

$$m_n = \begin{cases} m_{n-1}+1, & \text{with probability } \frac{\alpha}{n-1+\alpha}, \\ m_{n-1}, & \text{otherwise} \end{cases}. \tag{5}$$

In summary, each table represents a different class, and the customers are observations which are assigned to particular classes.

An important feature of DPs is that their data are *exchangeable*. Also, the number of tables where customers are seated according to them is not limited. In fact, with more customers coming to a restaurant, this number only grows. This process has been used for classification in many disciplines and has been well studied.

In many real-world problems, the number of classes within periods of times varies. We would like to study problems with this kind of dynamics of the data. For example, one might be interested in a sequential processing of signals, where the classification of the signals is time dependent. One approach to this problem is by way of epochs. The processes that we study in this paper belong to this class. An early work on this idea in the context of dynamic topic modeling is [8]. In [9], the authors study evolutionary clustering over epochs with the aim of making the clustering parameters smooth over time. Temporal DP models are studied in [10]. There, the exchangeability of the data is only valid within an epoch. The time-varying DPs investigated in [11] have the property that at each time the random distribution follows a DP.

Another class of approaches is known as dependent DPs [12]. These processes at any time instant with n marginally represent DPs. They have been used for joint estimation of the parameters of the process and the evolution of classes with time [13]. Time sensitive DPs are yet another type of processes that aim at capturing time-dependent clustering [14]. They use temporal weights for the clusters that depend on the cluster history.

3. NEW TYPES OF PROCESSES

Here we describe the new processes by way of modified CRPs. First, consider a restaurant where the customers stay in the restaurant for a limited time. More specifically, let each customer stay in the restaurant for N time units. Once the customer's time in the restaurant expires, the customer leaves the restaurant. Now suppose that the first N customers come to the restaurant and are seated in the usual way according to the probabilities defined by (3) and (4). After the restaurant fills with N customers, the first customer who entered the restaurant, leaves it, and a new customer is seated. Then the second customer who entered the restaurant leaves and a new customer comes in and so on. For the $(N+1)$st customer, the seating probabilities are given by

$$p_{N+1,i} = \begin{cases} \frac{k_{N,i}}{N-1+\alpha}, & i = 1, \cdots, m_N, \\ \frac{\alpha}{N-1+\alpha}, & i = m_N + 1, \end{cases} \tag{6}$$

where $k_{N,i}$ is the number of customers that sit at table i at time N and *after* the first customer left the restaurant. The symbol m_N is the total number of tables that have been occupied since the restaurant was opened. If we denote with z_n the table assigned to the nth customer, we can write

$$k_{N,i} = \sum_{j=2}^{N} \delta_{z_j=i}, \tag{7}$$

where

$$\delta_{z_j=i} = \begin{cases} 1, & z_j = i, \\ 0, & \text{otherwise} \end{cases}. \tag{8}$$

In general, for $n > N$ the seating probabilities are given by

$$p_{n,i} = \begin{cases} \frac{k_{n-1,i}}{N-1+\alpha}, & i = 1, \cdots, m_{n-1}, \\ \frac{\alpha}{N-1+\alpha}, & i = m_{n-1} + 1 \end{cases}, \tag{9}$$

where

$$k_{n-1,i} = \sum_{j=n-N+1}^{n-1} \delta_{z_j=i}. \tag{10}$$

We note that for $n \geq N$, we also have

$$\sum_{i=1}^{m_{n-1}} k_{n-1,i} = N - 1, \tag{11}$$

where m_{n-1} is the total number of tables that have been occupied by time $n-1$ and since the opening of the restaurant. However, we point out that at any time after N, the maximum number of tables that can be occupied is N.

Note that the above process can be viewed as a CRP with a finite capacity (with N tables only, CRPFC(N)), i.e., a restaurant that can serve a finite number of customers (N). As the process evolves, some of the tables that have been occupied in the past become vacant and ready for new customers who will be served with new dishes. From here on, we use this interpretation.

A more complicated process arises when the parameter α is time-varying. Thus, let α now be denoted by α_n. Basically, everything above remains the same except that the seating probabilities become

$$p_{n,i} = \begin{cases} \frac{k_{n-1,i}}{N-1+\alpha_n}, & i = 1, \cdots, m_{n-1}, \\ \frac{\alpha_n}{N-1+\alpha_n}, & i = m_{n-1} + 1 \end{cases}. \tag{12}$$

The dynamics of α_n allow for modeling periods where the number of currently occupied tables becomes small or large. When α_n becomes small, new tables are rarely assigned, and vice versa. We complete this model by describing the time variation of α_n by

$$\alpha_n = g(\alpha_{n-1}, u_n), \tag{13}$$

where $\alpha_n > 0, \forall n$, and $g(\alpha_{n-1}, u_n)$ is a function of the previous value of α and some perturbation u_n with known distribution.

4. MOMENTS OF THE PROCESSES

First we discuss processes with constant α and then processes with time-varying α.

4.1. Processes with constant α

We can make the following claim:

Claim: The mean of occupied tables in a restaurant with a capacity of N tables for $n = 1, 2, \cdots, N$ can be obtained recursively from

$$\mu_n = \mu_{n-1} + \frac{\alpha}{n-1+\alpha}, \quad (14)$$

where $\mu_1 = 1$, and the variance of occupied tables by

$$\sigma_n^2 = \sigma_{n-1}^2 + \frac{(n-1)\alpha}{(n-1+\alpha)^2}, \quad (15)$$

where $\sigma_1^2 = 0$.

The proof of the above claim is rather simple. First we prove the recursive equation for the mean. We start with $\mu_1 = 1$. Define the random Bernoulli variable X_n, $n = 2, 3, \cdots, N$ whose probability mass function is given by

$$p(x_n) = \begin{cases} \frac{\alpha}{n-1+\alpha}, & x = 1 \\ \frac{n-1}{n-1+\alpha}, & x = 0 \end{cases}. \quad (16)$$

We recognize that X_n represents the event that a new table is occupied ($X_n = 1$), or a customer is seated at one of already occupied tables ($X_n = 0$). Now, for the nth time instant we write the recursion

$$\mu_n = \mu_{n-1} P(X_n = 0) + (\mu_{n-1} + 1) P(X_n = 1). \quad (17)$$

Upon the substitution of the probabilities for $X_n = 0$ and $X_n = 1$, we obtain (14).

For the variance, we have

$$\sigma_n^2 = \mathbb{E}(Y_n - \mu_n)^2, \quad (18)$$

where Y_n is the number of occupied tables after n customers entered the restaurant. Since $Y_n = Y_{n-1} + X_n$ and with (14), we write

$$\begin{aligned} \sigma_n^2 &= \mathbb{E}\left(Y_n - \mu_n\right)^2 \\ &= \mathbb{E}\left(Y_{n-1} + X_n - \mu_{n-1} - \frac{\alpha}{n-1+\alpha}\right)^2 \\ &= \mathbb{E}\left(Y_{n-1} - \mu_{n-1}\right)^2 + \mathbb{E}\left(X_n - \frac{\alpha}{n-1+\alpha}\right)^2 \\ &= \sigma_{n-1}^2 + \mathrm{var}(X_n) \quad (19) \\ &= \sigma_{n-1}^2 + \frac{\alpha}{(n-1+\alpha)}\left(1 - \frac{\alpha}{(n-1+\alpha)}\right) \\ &= \sigma_{n-1}^2 + \frac{(n-1)\alpha}{(n-1+\alpha)^2}. \quad (20) \end{aligned}$$

After the arrival of the Nth customer, the $(N+1)$st customer replaces the first customer. We expect that at some time, the mean value and the variance of the number of occupied tables reaches a steady value, and the process of occupied tables is stationary. The condition for stationarity is that the average decrease in number of tables due to the departure of a customer is given by

$$\Delta\mu = \frac{\alpha}{N-1+\alpha}. \quad (21)$$

The reason is that every new customer contributes on average an increase in number of tables given by (21).

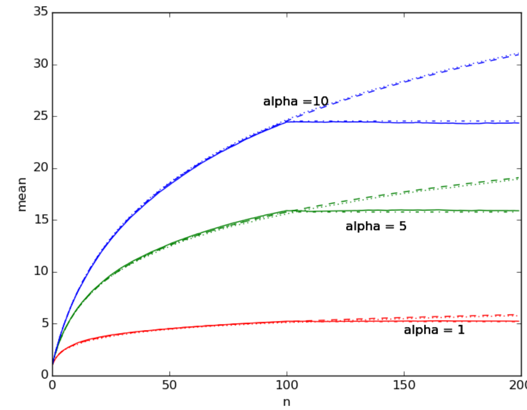

Fig. 1. Evolutions of the process means for different values of α. Solid lines: simulated means of the CRPFC (100), dash-dot lines: theoretical means of the CRPFC (100), dotted lines: theoretical means of the CRP, dashed lines: simulated means of the CRP.

We prove stationarity by induction. When $N = 1$, it is trivial to show that the process is stationary. In fact, then the mean is always equal to one and the variance is zero. Suppose now that a restaurant with $N-1$ tables has achieved stationarity for some n. Then we have $\mu_n = \mu^{(N-1)}$, where $\mu^{(N-1)}$ is the stationary mean of the process with a capacity of $N-1$ tables. Now, at time $n+1$, we decide to increase the capacity of the restaurant by one table so that it now becomes equal to N. At that time we also extend the time duration for the customers to stay in the restaurant, which means that no customers leave the restaurant at $n+1$. Then we can write

$$Y_{n+1} = Y_n + X_{n+1}, \quad (22)$$

where Y_n is the number of customers in the restaurant at time n and X_{n+1} has the same meaning as before. Thus,

$$\begin{aligned} \mu_{n+1} &= \mu^{(N-1)} + \mathbb{E}(X_{n+1}) \\ &= \mu^{(N-1)} + \frac{\alpha}{N-1+\alpha}. \quad (23) \end{aligned}$$

Thus, the new mean is not a function of n, and therefore the process with N tables has also a stationary mean. The argument is similar in showing that the variance is not a function of n once we increase the capacity of the restaurant by one table. Furthermore, this argument suggests that in a process with N tables, the stationarity is achieved at $n = N$. Finally, the stationary values of the mean and variance of this process are given by

$$\mu^{(N)} = \alpha \sum_{k=1}^{N} \frac{1}{k-1+\alpha}, \quad (24)$$

$$\sigma^{2(N)} = \alpha \sum_{k=1}^{N} \frac{k-1}{(k-1+\alpha)^2}. \quad (25)$$

In Fig. 1, we present the evolution of the means of several processes with n (with parameters $\alpha = 1, \alpha = 5$, and $\alpha = 10$, respectively). There we also see the evolution of the means of the CRP with the same respective parameters α. Similarly, in Fig. 2, we display the evolution of the variance with n for processes with parameters $\alpha = 1, \alpha = 5$, and $\alpha = 10$, respectively. The statistics

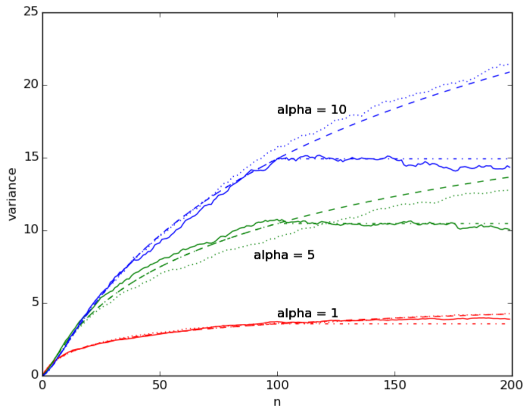

Fig. 2. Evolutions of process variances for different values of α. Solid lines: simulated variances of the CRPFC (100), dash-dot lines: theoretical variances of the CRPFC (100), dotted lines: theoretical variances of the CRP, dashed lines: simulated variances of the CRP.

were obtained from 1000 realizations of each process. Figures 1 and 2 also show the theoretical means and variances. From the figures we can clearly see how the mean and variance of the CRPFC settle at a steady value as soon as there are N customers in the restaurant.

4.2. Processes with time-varying α

We can derive similar expressions for the means and variances of the processes when α varies with time. It is not difficult to show that now for $n = 2, 3, \cdots, N$, we have

$$\mu_n = \mu_{n-1} + \frac{\alpha_n}{n - 1 + \alpha_n}, \tag{26}$$

where $\mu_1 = 1$, and the variance of occupied tables is given by

$$\sigma_n^2 = \sigma_{n-1}^2 + \frac{(n-1)\alpha_n}{(n - 1 + \alpha_n)^2}, \tag{27}$$

where $\sigma_1^2 = 0$. For $n > N$, we have

$$Y_n = Y_{n-1} - Q_n + X_n, \tag{28}$$

where Q_n is a Bernoulli random variable with $P(Q_n = 1) = p_n$ and where Q_n takes the value of one if the customer who leaves the restaurant was sitting alone before leaving. Then we can write

$$\mu_n = \mu_{n-1} - p_n + \frac{\alpha_n}{N - 1 + \alpha_n}. \tag{29}$$

For the variance, we have

$$\begin{aligned} \sigma_n^2 &= \mathbb{E}(Y_n - \mu_n)^2 \\ &= \mathbb{E}\left(Y_{n-1} - Q_n + X_n - \mu_{n-1} - p_n + \frac{\alpha_n}{N-1+\alpha}\right)^2 \\ &= \sigma_{n-1}^2 - 2\,\mathrm{cov}(Y_{n-1}, Q_n) + p_n(1-p_n) \\ &+ \frac{(N-1)\alpha_n}{(N-1+\alpha_n)^2}, \end{aligned} \tag{30}$$

where $\mathrm{cov}(Y_{n-1}, Q_n)$ is the covariance between the random variables Y_{n-1} and Q_n, i.e.,

$$\mathrm{cov}(Y_{n-1}, Q_n) = \mathbb{E}\left((Y_{n-1} - \mu_{n-1})(Q_n - p_n)\right). \tag{31}$$

In Fig. 3, we show the evolution of the means of two process with n and where $N = 100$. The parameter α of the first process (solid line) abruptly changes its value at two time instants, at $n = 200$ and $n = 500$. At $n = 200$ it increases its value from $\alpha = 2$ to $\alpha = 8$. The second change is from $\alpha = 8$ back to $\alpha = 2$. The other process (dashed line) changes its value of α in a similar way at the same time instants, but the values are now $\alpha = 5$ and $\alpha = 15$. In Fig. 4, we plotted the evolution of the variance of the same processes (with corresponding solid and dashed lines). All the results were obtained from 500 realizations.

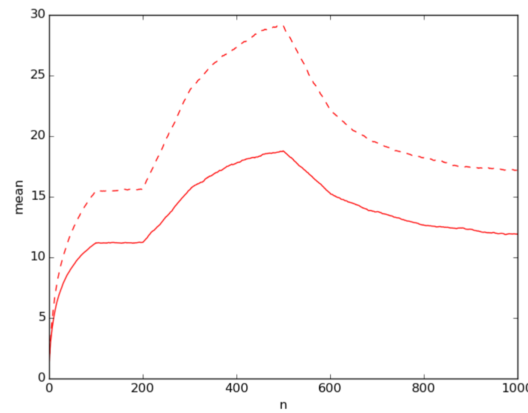

Fig. 3. Evolutions of means of two processes with time-varying α parameters. See text for explanation.

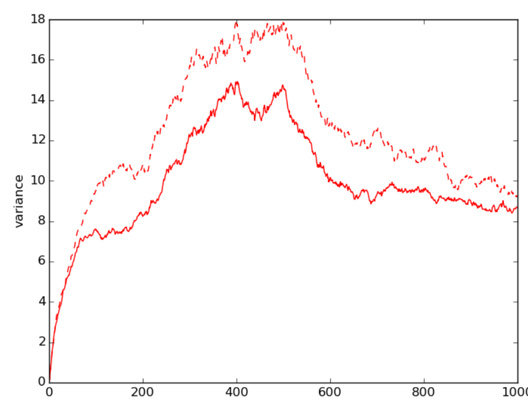

Fig. 4. Evolutions of variances of two processes with time-varying α parameters. See text for explanation.

5. INFERENCE

5.1. Processes with constant α

Here we address the estimation of α from a sequence of labels z_n (table assignments). We assume that the estimation is from labels after the restaurant has already at least N seated customers. Suppose that we have available J observations. For the likelihood we write

$$p(z_{1:J}|\alpha) = \prod_{j=1}^{J} \frac{\alpha^{\delta_{z_j=\text{new}}} k_{i,j}^{\delta_{z_j=i}}}{N-1+\alpha}, \quad (32)$$

where $\delta_{z_j=\text{new}} = 1$ if the jth customer is assigned a new table and is zero otherwise, and $k_{i,j}$ is the number of customers sitting on table i when the jth observation is made. We readily show that the maximum likelihood (ML) estimate is given by

$$\widehat{\alpha} = \frac{(N-1)K}{J-K}, \quad (33)$$

where

$$K = \sum_{j=1}^{J} \delta_{z_j=\text{new}}. \quad (34)$$

We see that the sufficient statistic for estimating α is K, the total number of newly assigned tables in J assignments. The range of K is $K \in \{0, 1, \cdots, J\}$. For $K = 0$, we obtain $\widehat{\alpha} = 0$, and for $K = J$, $\widehat{\alpha} = \infty$.

The ML estimate of α is biased. In fact, one can readily show that its expected value is infinite. Since K is a binomial random variable with parameters J and $\frac{\alpha}{N-1+\alpha}$, we have

$$\mathbb{E}(\widehat{\alpha}) = \sum_{k=0}^{J} \frac{(N-1)k}{J-k} \binom{J}{k} \left(\frac{\alpha}{N-1+\alpha}\right)^{k} \left(\frac{N-1}{N-1+\alpha}\right)^{J-k}, \quad (35)$$

and the claim immediately follows because the event $k = J$ has a finite probability and the factor $1/(J-k)$ in the summation will force the sum to be infinite.

5.2. Processes with time-varying α

For processes with time-varying α, the estimation of α_n is more challenging. This is a nonlinear problem that can be addressed by particle filtering [15]. With particle filtering, we aim at tracking the posterior distribution of α_n, given the observations $z_{1:J} \equiv \{z_1, z_2, \cdots, z_J\}$, $p(\alpha_n|z_{1:J})$. To that end we use a state space model, where we represent the state equation by

$$\alpha_n \sim p(\alpha_n|\alpha_{n-1}), \quad (36)$$

and the observation equation by

$$z_n \sim p(z_n|\alpha_n) = Cat(z_{n-N+1:n-1}, \alpha_n), \quad (37)$$

where $Cat(\cdot, \cdot)$ stands for a Categorical distribution with probabilities defined by the observations $z_{n-N+1:n-1}$ and α_n as in (12). With particle filtering we represent the posterior $p(\alpha_{n-1}|z_{1:n-1})$ by a random measure composed of particle $\alpha_{n-1}^{(m)}$ and weights $w_{n-1}^{(m)}$. At time n, if we draw particles from

$$\alpha^{(m)} \sim p(\alpha_n|\alpha_{n-1}^{(m)}), \quad (38)$$

the weights of $\alpha_n^{(m)}$ are computed by

$$w_n^{(m)} \propto w_{n-1}^{(m)}\, Cat(z_{n-N+1:n-1}, \alpha_n^{(m)}), \quad (39)$$

where $\propto$ signifies proportionality. The minimum mean squared error (MMSE) estimate of α_n is obtained from

$$\widehat{w}_n = \sum_{m=1}^{M} w_n^{(m)} \alpha_n^{(m)}. \quad (40)$$

Before the next time instant $n+1$, one may implement resampling so that the estimate of the time-varying α does not degrade with time.

6. CONCLUSION

In this paper we investigated generative models for sequential formation of clusters. One can can readily extend these models to create mixture models with time-varying structures. The variability of the structures is defined by a concentration parameter and the “capacity” of the process. We considered models where the concentration parameters are constant and time-varying. For processes with constant concentration parameters, we obtained their maximum likelihood estimates, and for processes with time-varying concentration parameters, we proposed a particle filtering method for their tracking.

7. REFERENCES

[1] C. M. Bishop, *Pattern Recognition and Machine Learning*, Springer, New York, 2006.

[2] G. McLachlan and T. Krishnan, *The EM Algorithm and Extensions*, John Wiley & Sons, 2007.

[3] S. Richardson and P. J. Green, “On Bayesian analysis of mixtures with an unknown number of components (with discussion),” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 59, no. 4, pp. 731–792, 1997.

[4] M. Stephens, “Bayesian analysis of mixture models with an unknown number of components-an alternative to reversible jump methods,” *Annals of Statistics*, vol. 28, no. 1, pp. 40–74, 2000.

[5] T. S. Ferguson, “A Bayesian analysis of some nonparametric models,” *Annals of Statistics*, vol. 1, no. 2, pp. 209–230, 1973.

[6] N. L. Hjort, C. Holmes, P. Müller, and S. G. Walker, *Bayesian Nonparametrics*, Cambridge University Press, 2010.

[7] M. D. Escobar and M. West, “Bayesian density estimation and inference using mixtures,” *Journal of the American Statistical Association*, vol. 90, no. 430, pp. 577–588, 1995.

[8] D. M. Blei and J. D. Lafferty, “Dynamic topic models,” in *Proceedings of the 23rd International Conference on Machine Learning*, 2006, pp. 113–120.

[9] T. Xu, Z. Zhang, P. S. Yu, and B. Long, “Dirichlet process based evolutionary clustering,” in *Proceedings of the IEEE International Conference on Data Mining*, 2008.

[10] A. Ahmed and E. P. Xing, “Dynamic non-parametric mixture models and the recurrent Chinese restaurant process with applications to evolutionary clustering,” in *Proceedings of SDM*. SIAM, 2008, pp. 219–230.

[11] F. Caron, M. Davy, and A. Doucet, “Generalized Polya urn for time-varying Dirichlet process mixtures,” in *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, 2007.

[12] S. N. MacEachern, “Decision theoretic aspects of dependent nonparametric processes,” in *Proceedings of the World of the International Society for Bayesian Analysis*, 2000.

[13] N. Srebro and S. Roweis, “Time-varying topic models using dependent Dirichlet processes,” *Univeristy of Toronto, Technical report*, 2005.

[14] X. Zhu, Z. Ghahramani, and J. Lafferty, “Time-sensitive Dirichlet process mixture models,” *Carnegie Mellon University, Technical report*, 2005.

[15] A. Doucet, N. de Freitas, and N. Gordon, Eds., *Sequential Monte Carlo Methods in Practice*, Springer, New York, 2001.