Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence

The Dynamic Chinese Restaurant Process via Birth and Death Processes

Rui Huang
Department of Statistics
The Chinese University of Hong Kong

Fengyuan Zhu* and **Pheng-Ann Heng**
Department of Computer Science and Engineering
The Chinese University of Hong Kong

Abstract

We develop the Dynamic Chinese Restaurant Process (DCRP) which incorporates time-evolutionary feature in dependent Dirichlet Process mixture models. This model can capture the dynamic change of mixture components, allowing clusters to emerge, vanish and vary over time. All these macroscopic changes are controlled by tracing the birth and death of every single element. We investigate the properties of dependent Dirichlet Process mixture model based on DCRP and develop corresponding Gibbs Sampler for posterior inference. We also conduct simulation and empirical studies to compare this model with traditional CRP and related models. The results show that this model can provide better results for sequential data, especially for data with heterogeneous lifetime distribution.

1 Introduction

Dirichlet Process Mixture (DPM) model has been widely used in statistical data analysis and machine learning, especially in modeling clustered data. Traditional DPM possesses good properties in many aspects, but its static feature makes it not applicable in evolutionary clustering problems. First, the exchangeability assumption embedded in a DPM is usually questionable for data with time dependency. That is, data generated at different time points may not be exchangeable. Instead, the "partial exchangeability" is more realistic (Ahmed and Xing 2008). Second, DPM cannot capture the dynamic change of cluster weights and parameters for an evolutionary clustering model. Third, the "rich-get-richer" phenomenon in DP is problematic in many applications. To solve the evolutionary clustering problem, many authors developed time varying models on dependent data. These models include the time-sensitive DPM model (Zhu, Ghahramani, and Lafferty 2005), the temporal DP mixture model (Ahmed and Xing 2008), the Generalized Pólya-urn process mixture model (Caron, Davy, and Doucet 2012) and the Distance-dependent CRP (Blei and Frazier 2011).

Even though the existing models characterize the dependency of data from different aspects and handle the evolutionary clustering problem, we observe that they share some features in common that could be further improved. First of all, most models operate in discrete time. This implicitly requires a 'discretization' procedure that can lead to loss of information. Furthermore, all existing models apply homogeneous rules to measure the 'influence' of existing data. For example, the recurrent CRP (Ahmed and Xing 2008) deletes all data before the previous time period, which implicitly assumes that data created at the same time will have similar influence on the generation of new data. In this case, the evolution of clusters is expected to share a statistically homogeneous feature.

Unfortunately, the homogeneity assumption may not be valid in many real applications. Intuitively, it is natural to measure the influence of an existing datum (word) on future data by its lifetime. That is, a datum is influential as long as it is "surviving". The death of a datum could be defined by some appropriate criteria based on practical considerations and the death should be observable. In this view, the lifetime distribution is usually heterogeneous. This could be better illustrated by an Internet news modeling example, in which all news is assumed to be generated from certain topics (clusters) over time. In this example, we can claim the death of a word if the news containing it has no more comments from Internet users for two consecutive days. Moreover, a topic is considered to be dead if all words generated from it are dead. From this perspective, we can expect at least two types of heterogeneities. First, the cluster evolution heterogeneity occurs. A piece of news is usually triggered by an underlying event. Therefore, some topics based on certain type of events are expected to be popular for a long time while others are not, and the difference may be substantial. This leads to the dynamic change of cluster weights. More importantly, the heterogeneity in the lifetime of words within a topic occurs. For instance, within the topic "SPORTS", words related to "the World Cup" may be hot for several weeks during the game, while words related to "scandals of athletes" may only last for a few days. This kind of heterogeneity leads to the dynamic change of cluster parameters. However, even though such heterogeneities might be complex, the lifetime can be usually observed under some practical criteria in real applications without substantial cost, and these observations certainly increase our information about the evolutionary clustering problem. In view of this, we de-

*Share equal contribution with Huang.

velop a Dynamic Chinese Restaurant Process (DCRP) which is a novel generalization of the Chinese Restaurant Process (CRP) (Teh 2007). In our model, the emergence and disappearance of data over time are modeled by a birth and death process. For the evolutionary clustering problem, we present a new framework of DCRP mixture model with Gibbs sampling algorithms. The simulation and empirical study show that our DCRP mixture model has good performance both for data with homogeneous and heterogeneous lifetime. For the latter case our model provides substantially improvements compared with other related models.

2 The Dynamic Chinese Restaurant Process and DCRP Mixture Model

2.1 The Dynamic Chinese Restaurant Process

A DCRP can be described by the following metaphor. Suppose we have a Chinese restaurant with infinite many tables. Customers are assumed to enter the restaurant in group according to some birth process. The birth time is described by some model M_b. For each customer entering the restaurant, he will choose to sit at a table according to a CRP scheme. Simultaneously, the customers existing in the restaurant will leave according to some death process. That is, each customer will stay in the restaurant for some lifetime (deterministic or stochastic), and then leave. Lifetime distribution is described by M_d. Typically, M_d could be taken as deterministic, exponential or Weibull distribution, or a heterogeneous model. Note that birth and death processes involved in this model are not necessarily continuous-time. For example, if M_b is chosen to be the case that customers enter at equally spaced time points, and M_d is taken to be deterministically unit length, then the model is reduced to the recurrent CRP.

In order to describe the DCRP mathematically, we introduce the following notations:

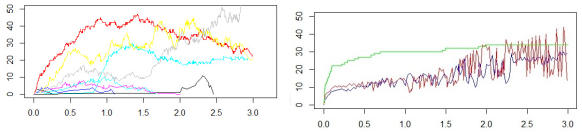
Table 1: Notations for Defining DCRP

SYMBOL	DESCRIPTION
T_i	Birth-time of i-th group
N_i	Number of customers in i-th group
$L_{i,j}$	Lifetime of j-th customer in i-th group
$I_{i,j,t}$	Status indicator of j-th customer in i-th group at time t. $I_{i,j,t} = 1$ if j-th customer in i-th group is in the restaurant at time t and $I_{i,j,t} = 0$ otherwise
$S_{k,t}$	Number of survived customers (arrived before time t) in table k at time t
K_{max}	Largest index of tables have ever been occupied
$z_{i,j}$	Table index of j-th customer in i-th group

With these notations, a DCRP could be formulated as follows.

$$
\begin{aligned}
&T_i \sim M_b \qquad (i = 1, 2, \ldots)\\
&L_{i,j} \sim M_d \qquad (i = 1, 2, \ldots; j = 1, 2, \ldots, N_i)\\
&I_{i,j,t} = \mathbb{1}\{T_i + L_{i,j} > t\}\\
&S_{k,t} = \sum_{m=1}^{i-1} \sum_{n=1}^{N_m} \mathbb{1}\{z_{m,n} = k, I_{m,n,t} = 1\}\\
&z_{i,j} \mid z_{1:i-1,\cdot}, z_{i,1:j-1} \sim \frac{\alpha \mathbb{1}\{z_{i,j} = K_{max} + 1\}}{\sum_{k=1}^{K_{max}} S_{k,T_i} + j - 1 + \alpha} + \\
&\qquad \frac{\sum_{k=1}^{K_{max}} (S_{k,T_i} + \sum_{l=1}^{j-1} \mathbb{1}\{z_{i,l} = k\}) \mathbb{1}\{z_{i,j} = k\}}{\sum_{k=1}^{K_{max}} S_{k,T_i} + j - 1 + \alpha}
\end{aligned}
$$

Several remarks should be mentioned here to provide a clear understanding of DCRP. First, there are two levels of processes in DCRP. One is the birth-death process of customers, which is controlled by the birth time and lifetime of each customer. The other is the latent cluster assignment process, which is a CRP-type scheme depending on the first level birth-death process. Second, DCRP relaxes the exchangeability assumption of traditional CRP, while preserves the "partial exchangeability". That is, the customers within each group are exchangeable, but customers in different groups are not. Finally, it could be imagined that the existence of death process will result in extinction of clusters, and accelerate emergence of new clusters, hence allowing time-varying parameters in the model.



(a) Number of items in different clusters.

(b) Number of clusters.

Figure 1: Simulation results of DCRP

Figure 1 demonstrates the simulation results of DCRP. We set M_b as $Exponential(\lambda)$ with $\lambda = 100$ and M_d as $Exponential(\mu)$ with $\mu_1 = 10$ and $\mu_2 = 50$ representing low and high death rate respectively for comparison. Figure 1a illustrates that the number of items in different clusters can vary over time. The number of items in a cluster can either increase or decrease. Old clusters can fade out while new ones can emerge. Popular clusters can become unpopular and vice versa. Figure 1b shows the evolution of number of clusters over time with pure birth process, birth and death process with low death rate and high rate. With the death process, the number of clusters can either increase or decrease and higher death rate can lead to more significant fluctuation.

2.2 DCRP Mixture Model

We can establish a dependent DP mixture model by using DCRP as prior. Consider a DCRP mixture model as follows.

$$\begin{aligned} \theta_k &\sim H \qquad k = 1, 2, \dots \\ (z_{i,\cdot}, T_i, L_{i,\cdot}, N_i) \mid \mathcal{F}_{T_i^-} &\sim DCRP(\alpha; M_b, M_d) \\ & \qquad i = 1, 2, \dots \\ x_{i,j} \mid z_{i,j} &\sim F(\theta_{z_{i,j}}) \end{aligned}$$

where $\mathcal{F}_{T_i^-}$ denotes all the information up to time T_i. In this model, data are observed in groups. Observable variables include T_i, N_i, $L_{i,j}$, and $x_{i,j}$. It should be noticed that the $L'_{i,j}s$ can be both uncensored or right-censored (Lawless 2011), which means we can fully observe the lifetime of a dead datum, while we can only observe the living time for a survived datum. Latent variables include the $z_{i,j}$ and θ_k. Our problem of interest is that upon observing all data in a time interval $[0, t]$ (where t is chosen based on practical consideration), we want to make inference on the latent variables, hence estimate the whole model. After that, predictions could be made based on the estimated model. Note that M_b and M_d here are important for prediction, and they could be estimated by a variety of statistical approaches based on the observations.

2.3 Inference

In this section we show how to make inference on the DCRP mixture model. Suppose that in a time interval $[0, t]$, we have made the following observations:

- Birth time of i-th group: $T_i, i = 1, 2, \dots, n$
- Number of data in i-th group: $N_i, i = 1, 2, \dots, n$
- Lifetime of j-th customer in i-th group: $L_{i,j}, i = 1, 2, \dots, n-1; j = 1, 2, \dots, N_i$ (uncensored or right-censored)
- Value of j-th datum in i-th group: $x_{i,j}$

Our objective is to estimate the birth model M_b, death model M_d, and find the posterior distribution of latent cluster assignment indicators $z_{i,j}$ and cluster-specific parameters θ_k.

Estimation of M_b and M_d Estimating M_b and M_d based on the observations $T_1, \dots, T_n$ and $L_{1,\cdot}, \dots, L_{n-1,\cdot}$ is a general statistics problem, and a variety of parametric or nonparametric methods could be applied here based on the nature of observed data. It should be noted that M_d is important to capture the scheme of dependency (homogeneous or heterogeneous), and both M_b and M_d are needed for predicting new data. Although the estimation of M_b and M_d might be complicated, we want to emphasize that it is not a trouble if our main objective is modeling the observed data, rather than predicting the future. In fact, as we will see later, the posterior inference of latent cluster assignments and cluster parameters does not depend on M_b and M_d given the observed birth time and lifetime.

As an illustrative example, for simplicity we assume $T_1, T_2 - T_1, \dots, T_n - T_{n-1} \overset{i.i.d.}{\sim} Exp(\lambda)$ and $L_{i,j} \overset{i.i.d.}{\sim} Exp(\mu)$. Since the birth time $T_1, T_2, \dots, T_n$ can be fully observed, it is easy to get the maximum likelihood estimator (MLE) of λ

$$\hat{\lambda} = \frac{n}{T_n}$$

For estimation of μ, note that the observations can be right-censored, with censoring time $c_{i,j} = t - T_i$, we can derive the likelihood function of the observed lifetimes following (Lawless 2011)

$$\begin{aligned} L(\mu) &= \prod_{i=1}^{n-1} \prod_{j=1}^{N_i} (\mu e^{-\mu t_{i,j}})^{\delta_{i,j}} (e^{-\mu t_{i,j}})^{1-\delta_{i,j}} \\ &= \mu^r \exp\left(-\mu \sum_{i=1}^{n-1} \sum_{j=1}^{N_i} t_{i,j} \right) \end{aligned}$$

where $t_{i,j} = \min\{L_{i,j}, c_{i,j}\}$, $\delta_{i,j} = \mathbb{1}\{L_{i,j} < c_{i,j}\}$ and $r = \sum_{i=1}^{n-1} \sum_{j=1}^{N_i} \delta_{i,j}$. Therefore, the MLE for μ is

$$\hat{\mu} = \frac{r}{\sum_{i=1}^{n-1} \sum_{j=1}^{N_i} t_{i,j}}$$

Posterior Distribution of $z_{i,j}$ and θ_k To simplify the notations, let $\mathbf{X} = (x_{1,1}, \dots, x_{n,N_n})$, $\mathbf{N} = (N_1, \dots, N_n)$, $\mathbf{T} = (T_1, \dots, T_n)$, $\mathbf{L} = (L_{1,1}, \dots, L_{n,N_n})$, $\mathbf{Z} = (z_{1,1}, \dots, z_{n,N_n})$, $\mathbf{\Theta} = (\theta_1, \dots, \theta_m)$, where $m = \sum_{i=1}^{n} N_i$ is the maximum possible number of clusters given the observations. In order to draw samples from the joint posterior distribution of $(\mathbf{Z}, \mathbf{\Theta})$ via a Gibbs Sampler, we want to derive the conditional distributions $P(\mathbf{Z} \mid \mathbf{\Theta}, \mathbf{X}, \mathbf{N}, \mathbf{T}, \mathbf{L})$ and $P(\mathbf{\Theta} \mid \mathbf{Z}, \mathbf{X}, \mathbf{N}, \mathbf{T}, \mathbf{L})$.

First, to derive $P(\mathbf{Z} \mid \mathbf{\Theta}, \mathbf{X}, \mathbf{N}, \mathbf{T}, \mathbf{L})$, we have

$$\begin{aligned} P(\mathbf{Z} \mid \mathbf{\Theta}, \mathbf{X}, \mathbf{N}, \mathbf{T}, \mathbf{L}) \propto P(\mathbf{Z} \mid \mathbf{N}, \mathbf{T}, \mathbf{L}) \times \\ \prod_{i=1}^{n} \prod_{j=1}^{N_i} f(x_{i,j} \mid \theta_{z_{i,j}}) \end{aligned} \tag{1}$$

where $P(\mathbf{Z} \mid \mathbf{N}, \mathbf{T}, \mathbf{L})$ is the prior distribution of $\mathbf{Z}$, which is given by the DCRP and can be further decomposed as

$$\begin{aligned} P(\mathbf{Z} \mid \mathbf{N}, \mathbf{T}, \mathbf{L}) = & P(z_{n,\cdot} \mid z_{1,\cdot}, \dots, z_{n-1,\cdot}) \times \\ & P(z_{n-1,\cdot} \mid z_{1,\cdot}, \dots, z_{n-2,\cdot}) \dots \\ & P(z_{2,\cdot} \mid z_{1,\cdot}) \end{aligned} \tag{2}$$

Here $z_{i,\cdot} = (z_{i,1}, \dots, z_{i,N_i})$ denotes all the cluster assignment indicators in group i.

Then, to derive $P(\mathbf{\Theta} \mid \mathbf{Z}, \mathbf{X}, \mathbf{N}, \mathbf{T}, \mathbf{L})$, we assume all $\theta'_k s$ are conditionally independent given $(\mathbf{Z}, \mathbf{X}, \mathbf{N}, \mathbf{T}, \mathbf{L})$. Therefore,

$$\begin{aligned} P(\mathbf{\Theta} \mid \mathbf{Z}, \mathbf{X}, \mathbf{N}, \mathbf{T}, \mathbf{L}) &\propto \prod_{k=1}^{m} P(\theta_k \mid \mathbf{Z}, \mathbf{X}, \mathbf{N}, \mathbf{T}, \mathbf{L}) \\ &\propto \prod_{k=1}^{m} \left(h(\theta_k) \prod_{z_{i,j} \mathbb{1}\{T_i + L_{i,j} > t\} = k} f(x_{i,j} \mid \theta_k) \right) \end{aligned} \tag{3}$$

where $h(\theta_k)$ is the prior distribution of θ_k and the last production is taken over all the survived data points in cluster k at time t.

2.4 Gibbs Sampling Algorithms

Based on the construction of DCRP mixture model and posterior inference, we can now develop a Gibbs sampler to simulate samples from the joint posterior $P(\mathbf{Z}, \mathbf{\Theta} \mid \mathbf{X}, \mathbf{N}, \mathbf{T}, \mathbf{L})$. The algorithm consists of two steps: First, given $\mathbf{\Theta}^{(k)}, \mathbf{X}, \mathbf{N}, \mathbf{T}, \mathbf{L}$, we sample $\mathbf{Z}^{(k+1)}$ from $P(\mathbf{Z} \mid \mathbf{\Theta}^{(k)}, \mathbf{X}, \mathbf{N}, \mathbf{T}, \mathbf{L})$. Second, given $\mathbf{Z}^{(k+1)}, \mathbf{X}, \mathbf{N}, \mathbf{T}, \mathbf{L}$, we sample $\mathbf{\Theta}^{(k+1)}$ from $P(\mathbf{\Theta} \mid \mathbf{Z}^{(k+1)}, \mathbf{X}, \mathbf{N}, \mathbf{T}, \mathbf{L})$.

We start with the first step. To sample $\mathbf{Z}$ from $P(\mathbf{Z} \mid \mathbf{\Theta}, \mathbf{X}, \mathbf{N}, \mathbf{T}, \mathbf{L})$ we sequentially sample $z_{i,\cdot}, i = 1, 2, \ldots, n$ from $P(z_{i,\cdot} \mid z_{-i,\cdot}, \mathbf{\Theta}, \mathbf{X}, \mathbf{N}, \mathbf{T}, \mathbf{L})$, where $z_{-i,\cdot}$ denote all the cluster indicators excluding i-th group.

For $i = 1$, we have

$$P(z_{1,\cdot} \mid z_{-1,\cdot}, \mathbf{\Theta}, \mathbf{X}, \mathbf{N}, \mathbf{T}, \mathbf{L}) \propto P(z_{1,\cdot} \mid \mathbf{\Theta}, \mathbf{N}, \mathbf{T}, \mathbf{L}) \times P(z_{2:n,\cdot} \mid z_{1,\cdot}, \mathbf{\Theta}, \mathbf{N}, \mathbf{T}, \mathbf{L}) \prod_{j=1}^{N_1} f(x_{1,j} \mid \theta_{z_{1,j}}) \tag{4}$$

For $i = 2, 3, \ldots, n-1$, we have

$$P(z_{i,\cdot} \mid z_{-i,\cdot}, \mathbf{\Theta}, \mathbf{X}, \mathbf{N}, \mathbf{T}, \mathbf{L}) \propto P(z_{i,\cdot} \mid z_{1:i-1,\cdot}, \mathbf{\Theta}, \mathbf{N}, \mathbf{T}, \mathbf{L}) \times P(z_{i+1:n,\cdot} \mid z_{1:i,\cdot}, \mathbf{\Theta}, \mathbf{N}, \mathbf{T}, \mathbf{L}) \prod_{j=1}^{N_i} f(x_{i,j} \mid \theta_{z_{i,j}}) \tag{5}$$

For $i = n$, we have

$$P(z_{n,\cdot} \mid z_{-n,\cdot}, \mathbf{\Theta}, \mathbf{X}, \mathbf{N}, \mathbf{T}, \mathbf{L}) \propto P(z_{n,\cdot} \mid z_{1:n-1,\cdot}, \mathbf{\Theta}, \mathbf{N}, \mathbf{T}, \mathbf{L}) \prod_{j=1}^{N_n} f(x_{n,j} \mid \theta_{z_{n,j}}) \tag{6}$$

Then, we complete the whole algorithm by finishing the second step, which is to update $\mathbf{\Theta}$ from $P(\mathbf{\Theta} \mid \mathbf{Z}, \mathbf{X}, \mathbf{N}, \mathbf{T}, \mathbf{L})$. Note that this is exactly given by (3). Therefore, we have

$$P(\mathbf{\Theta} \mid \mathbf{Z}, \mathbf{X}, \mathbf{N}, \mathbf{T}, \mathbf{L}) \propto \prod_{k=1}^{m} \left(h(\theta_k) \prod_{z_{i,j} \mathbb{1}\{T_i + L_{i,j} > t\} = k} f(x_{i,j} \mid \theta_k) \right) \tag{7}$$

2.5 Prediction

It is often desirable to predict the distribution of new data based on the estimated model. Since the model is time-varying, this predictive distribution will depend on the birth time of a new datum. Suppose that we have observations $\mathbf{X}, \mathbf{N}, \mathbf{T}, \mathbf{L}$ in $[0, t]$, and we want to predict the distribution of a new data point at time $t' > t$, then the predictive distribution is given by

$$P(x \mid t', \mathbf{\Theta}, \mathbf{Z}, \mathbf{X}, \mathbf{N}, \mathbf{T}, \mathbf{L}) = \sum_{z=1}^{K_n+1} P(z \mid t', \mathbf{Z}, \mathbf{X}, \mathbf{N}, \mathbf{T}, \mathbf{L}) f(x \mid \theta_z) \tag{8}$$

where $K_n + 1$ is the existing number of clusters.

Obviously, analytical solution of posterior predictive distribution could hardly be obtained. Even though we can integrate out the latent variables in equation (8) with the posterior distributions, the calculation will be very tedious, even impossible. However, this problem can be easily solved with Monte Carlo simulations based on the samples $(\mathbf{Z}^{(j)}, \mathbf{\Theta}^{(j)})$ drawn from the joint posterior distribution $P(\mathbf{Z}, \mathbf{\Theta} \mid \mathbf{X}, \mathbf{N}, \mathbf{T}, \mathbf{L})$.

3 Comparison with Related Works

There exist a variety of generalized CRP to model the evolution of mixture models over time. One such model is the time-sensitive DP Mixture model (Zhu, Ghahramani, and Lafferty 2005) where the contribution of each item in a cluster decays exponentially. The temporal DP mixture model (Ahmed and Xing 2008) is another approach which deletes all previous items over discrete time. In this way, the weight of each component is updated recurrently. The Distance-dependent CRP (Blei and Frazier 2011) is a more general model where the influence of each item varies by a general decay function. The Generalized Pólya-urn Process mixture model (Caron, Davy, and Doucet 2012) presents the variation of parameters by deleting particles with a fixed distribution. The common assumption of these generalized CRP models is that the influence rule of all data on the dynamic mixture should be homogeneous, which is relaxed in our approach.

There also exist a number of dependent DP based on other constructions. The pioneering work of MacEachern introduced the "single-p DDP model" (MacEachern 2000) which considers a DDP as a collection of stochastic process. However, it does not consider the varying of the collection size over time. Grifin and Steel introduced the "order-based" DDP (Griffin and Steel 2006) based on the stick-breaking construction which reorders the stick-breaking ratio over time. The "Local" DP (Chung and Dunson 2011) is a generalized version of "order-based" DP, which regroups stick-breaking ratios locally. Teh introduced the Hierarchical Dirichle process (Teh et al. 2006) where the base distribution of a child DP is its parent DP. This model has been extended to the dynamic HDP (Ren, Dunson, and Carin 2008) by combining the weighted mixture formulation with HDP. There are also a number of models which construct DDP from the perspective of the relation between DP, Gamma Process and Poisson Process (Rao and Teh 2009) (Lin, Grimson, and Fisher III 2010). Considering from different perspectives, the inference and sampling of these approaches are also different from our DCRP mixture model.

4 Experiment

4.1 Simulation Study

In this section, we perform a simulation study to demonstrate the application of the DCRP mixture model for modeling evolutionary clustered data. Also, we compare the experiment results with the following benchmark models: the traditional CRP (CRP), the recurrent CRP (rCRP) and the Distance-dependent CRP (dd-CRP) with exponential decay function which has best performance in the original paper.

Experiment Setting We generate two datasets from a time-varying mixture model. In dataset 1, lifetimes of generated data are homogeneous while they are heterogeneous in dataset 2. We utilize the observed lifetimes of data to estimate the decay function in dd-CRP. We use the same base distribution H for each model and the concentration parameter α of each model is set to be 2.

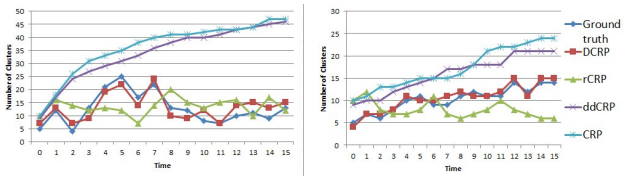


Figure 2: Number of clusters for dataset 1 (left) and dataset 2 (right).

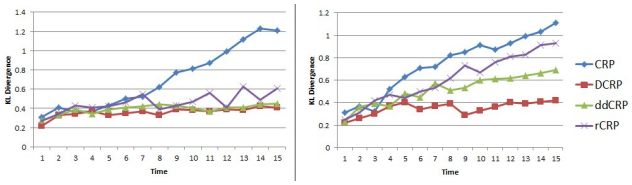


Figure 3: Kullback-Leibler Distance between estimated model and true model dataset 1 (left) and dataset 2 (right).

Results In our Gibbs sampler we ran three chains with different initial values and a sample of size 2000 is collected after 1200 iterations (burn-in time). We use the potential scale reduction factor (Brooks and Gelman 1998) as the test criterion for convergence. The reslut shows that all parameters converge before the burn-in time.

Figure 2 demonstrates the number of clusters in estimated models over time. It could be observed that our model can better capture the dynamic change of the number of clusters over time for both datasets. In figure 3, we compare the performance of each model with the criterion of the Kullback-Leibler divergence between the true model and the estimated model over time. In dataset 1, all generalized CRP models have significant improvement over the traditional CRP while the difference of the results of these models is marginal. This is because the lifetime distribution of each data is homogeneous. On the contrary, when the lifetime distribution of each data is heterogeneous, our model has a clear improvement over both traditional CRP and other generalized CRP models.

4.2 Empirical Study

This section evaluates the proposed DCRP mixture model on two datasets and compares the results with the traditional CRP, the rCRP, the dd-CRP as well as two state-of-the-art dynamic topic models: the topic over time (TOT) (Wang and McCallum 2006) and the continuous time dynamic topic model (CTDTM) (Wang, Blei, and Heckerman 2012).

Dataset Description We apply our model on two real datasets:

Twitter We randomly select 10,000 Twitter users with their full set of tweets between January 1st, 2013 and May 31st, 2013, resulting 74,593 distinct tweets. Each tweet contains its content, the time it was posed and the time of its replies. We consider the post time of a tweet as its "birth" time and the time of its last reply plus 2 hours as its "death" time.

NIPS We use dataset of all NIPS papers from 1987 to 2003 (17 years). We record the date of publication and the date of its last citation plus two years, and consider they are the "birth" and "death" time of this document.

Language Modeling We evaluate the proposed DCRP mixture model on the two datasets with a comparison of benchmark models. For all topic models, we set $\alpha = 2$ for both datasets while other hyper-parameters are set as suggested in the original papers. The decay functions in dd-CRP model for both datasets are trained by the observed lifetime of each data. The twitter data are discretized on daily basis for discrete-time models. For all models requiring Markov Chain Monte Carlo, we run three chains with different initial value. The burn-in time is set to be 1200 iterations and the sample size is 2000. Convergence diagnostics is conducted with methods mentioned above, which shows that all parameters converge within the burn-in period. We measure the performance of each model by Bayes factor (BF) (Berger and Pericchi 1996) over the traditional CRP.

The results of each model on the two datasets are illustrated in Table 2. The proposed model achieves a significant improvement over baseline models for both datasets, especially for the twitter data which has a more significant intrinsic heterogeneity on topics over time.

Figure 4 illustrates an example of the evolution of words distribution within a topic over time on Twitter dataset. From 01/10/2013 to 02/04/2013, the topic "SPORTS" was more about the NFL games because of the NFL Playoffs and Super Bowl. After that, the topic becomes relatively general. Since 04/21/2013, the topic is more about basketball games because of the start of NBA Playoffs. Figure 5 provides the evolution of several topics discovered in NIPS data by DCRP model. It clearly demonstrates the variation of popularity of each topics over time.

Figure 6 describes two selected topics discovered by DCRP, rCRP and ddCRP. In the twitter dataset, the topic of "SPORTS" is identified by all models, and the shown results are top 10 words with highest probabilities within this topic recorded at May 31st, 2013. We observe that rCRP and dd-CRP classified the special issue of "Boston Marathon Explosion" into the topic "SPORTS", while our DCRP model identified it into a new cluster which is more about terrorism. The interpretation is like this: intuitively, "Boston Marathon" undoubtedly belongs to topic "SPORTS", but "Boston Marathon Explosion" is more likely to be generated from the topic "TERRORISM" rather than "SPORTS". Before the explosion, there is no topic about terrorism in our dataset. After the explosion, our model detected the emergence of a new topic about "TERRORISM", which seems to be a more reasonable result. For NIPS data, DCRP clearly identified the topic of "PCA". It also successfully identified

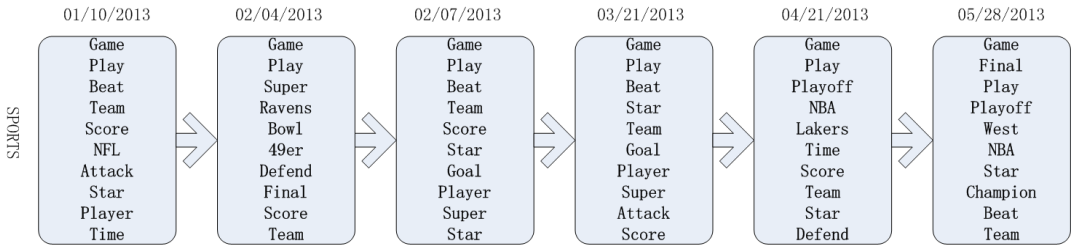


Figure 4: Dynamics of topic "SPORTS" for Twitter data over time.

the emergence of the "exponential PCA" technique developed around the year of 2003. While rCRP and dd-CRP model failed to identify that nonlinear dimensional reduction and metric learning are two techniques different from PCA (though very similiar), and they mixed these three topics together. However, DCRP has another two different topics that covers nonlinear dimensional reduction and metric learning respectively.

Table 2: Bayes Factors for Twitter and NIPS Dataset

Model	Twitter	NIPS
DCRP	104.3	53.5
dd-CRP	35.2	22.7
r-CRP	27.5	23.4
TOT	33.8	24.2
CTDTM	34.3	27.2

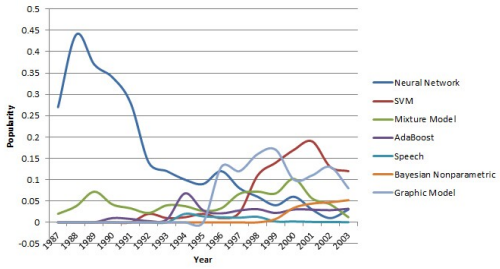


Figure 5: Dynamics of different topics for NIPS paper over time.

Predictive Topic Coherence We examine the predictive power of DCRP mixture model on the two datasets comparing with benchmark models. For each dataset, we remove the last 20 percent of data for cross validation. We evaluate the quality of each model by estimating the topic coherence(Mimno et al. 2011) of hold-out data with the predictive distribution. The parameter setting of each model is the same as last section.

The experiment results for each datasets have been illustrated in table 3. The proposed DCRP model performs better comparing with baseline models.

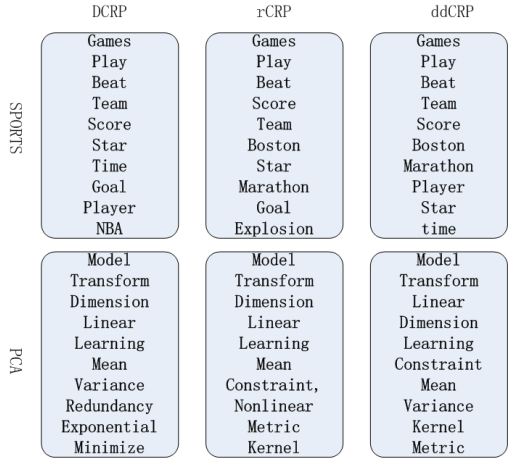


Figure 6: Comparison of Different Models

Table 3: Predictive Topic Coherence for Twitter and NIPS Data

Model	Twitter	NIPS
DCRP	-15830	-20340
dd-CRP	-17500	-21150
r-CRP	-17070	-21930
CRP	-18330	-23430
TOT	-16930	-22730
CTDTM	-17120	-21940

5 Discussion

In this paper we developed the dynamic Chinese Restaurant Process and corresponding DCRP mixture model for modeling and prediction of sequential data with complex time dependency. This model can successfully handle the case when lifetime of data is observable, and provide substantially better results for heterogeneous lifetime. The simulation and empirical study show that our DCRP mixture model consistently achieves superior performance over other related models. In the future, we are interested in investigating the case that "lifetime" of a data is not observable.

References

Ahmed, A., and Xing, E. P. 2008. Dynamic non-parametric mixture models and the recurrent chinese restaurant process:

with applications to evolutionary clustering. In *SDM*, 219–230. SIAM.

Berger, J. O., and Pericchi, L. R. 1996. The intrinsic bayes factor for model selection and prediction. *Journal of the American Statistical Association* 91(433):109–122.

Blei, D. M., and Frazier, P. I. 2011. Distance dependent chinese restaurant processes. *The Journal of Machine Learning Research* 12:2461–2488.

Brooks, S. P., and Gelman, A. 1998. General methods for monitoring convergence of iterative simulations. *Journal of computational and graphical statistics* 7(4):434–455.

Caron, F.; Davy, M.; and Doucet, A. 2012. Generalized polya urn for time-varying dirichlet process mixtures. *arXiv preprint arXiv:1206.5254*.

Chung, Y., and Dunson, D. B. 2011. The local dirichlet process. *Annals of the Institute of Statistical Mathematics* 63(1):59–80.

Griffin, J. E., and Steel, M. J. 2006. Order-based dependent dirichlet processes. *Journal of the American statistical Association* 101(473):179–194.

Lawless, J. F. 2011. *Statistical models and methods for lifetime data*, volume 362. John Wiley & Sons.

Lin, D.; Grimson, E.; and Fisher III, J. W. 2010. Construction of dependent dirichlet processes based on poisson processes.

MacEachern, S. N. 2000. Dependent dirichlet processes. *Unpublished manuscript, Department of Statistics, The Ohio State University*.

Mimno, D.; Wallach, H. M.; Talley, E.; Leenders, M.; and McCallum, A. 2011. Optimizing semantic coherence in topic models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 262–272. Association for Computational Linguistics.

Rao, V., and Teh, Y. W. 2009. Spatial normalized gamma processes. In *Advances in neural information processing systems*, 1554–1562.

Ren, L.; Dunson, D. B.; and Carin, L. 2008. The dynamic hierarchical dirichlet process. In *Proceedings of the 25th international conference on Machine learning*, 824–831. ACM.

Teh, Y. W.; Jordan, M. I.; Beal, M. J.; and Blei, D. M. 2006. Hierarchical dirichlet processes. *Journal of the american statistical association* 101(476).

Teh, Y. W. 2007. Dirichlet processes: Tutorial and practical course. *Machine Learning Summer School*.

Wang, X., and McCallum, A. 2006. Topics over time: a non-markov continuous-time model of topical trends. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, 424–433. ACM.

Wang, C.; Blei, D.; and Heckerman, D. 2012. Continuous time dynamic topic models. *arXiv preprint arXiv:1206.3298*.

Zhu, X.; Ghahramani, Z.; and Lafferty, J. 2005. Time-sensitive dirichlet process mixture models. Technical report, DTIC Document.