Distance dependent Chinese restaurant processes

David M. Blei BLEI@CS.PRINCETON.EDU
Department of Computer Science, Princeton University
35 Olden St., Princeton, NJ 08540

Peter Frazier PF98@CORNELL.EDU
Department of Operations Research and Information Engineering , Cornell University
232 Rhodes Hall, Ithaca, NY 14850

Abstract

We develop the distance dependent Chinese restaurant process (CRP), a flexible class of distributions over partitions that allows for non-exchangeability. This class can be used to model dependencies between data in infinite clustering models, including dependencies across time or space. We examine the properties of the distance dependent CRP, discuss its connections to Bayesian nonparametric mixture models, and derive a Gibbs sampler for both observed and mixture settings. We study its performance with time-dependent models and three text corpora. We show that relaxing the assumption of exchangeability with distance dependent CRPs can provide a better fit to sequential data. We also show its alternative formulation of the traditional CRP leads to a faster-mixing Gibbs sampling algorithm than the one based on the original formulation.

1. Introduction

Dirichlet process (DP) mixture models provide a valuable suite of flexible clustering algorithms for high dimensional data analysis. DP mixtures can be described via the Chinese restaurant process (CRP), a distribution over partitions that embodies the assumed prior distribution over cluster structures (Pitman, 2002). The CRP considers a sequence of customers sitting down at tables in a restaurant. Each customer sits at a previously occupied table with probability proportional to the number of customers already sitting there, and at a new table with probability proportional to a concentration parameter. In a CRP mixture, customers are data points, and data sitting at the same table belong to the same cluster. Since the number of occupied tables is random, the resulting posterior distribution of seating assignments provides a distribution of clusterings where the number of clusters is determined by the data.

The customers of a CRP are *exchangeable*—under any permutation of their ordering, the probability of a particular configuration is the same. While exchangeability is a reasonable assumption in some clustering applications, in many it is not. Consider data ordered in time, such as a time-stamped collection of news articles. Each article will tend to cluster with other articles that are nearby in time. Consider spatial data, such as pixels in an image or measurements at geographic locations. Here again, each data will tend to cluster with other data that are nearby in space. The traditional CRP mixture cannot model data like this.

We develop the *distance dependent Chinese restaurant process*, a new CRP in which the random seating assignment of the customers depends on the distances between them. These distances can be based on time, space, or other characteristics. Distance dependent CRPs can recover a number of existing dependent distributions (Ahmed & Xing, 2008; Zhu et al., 2005). They can also be arranged to recover the traditional CRP distribution. The distance dependent CRP expands the palette of infinite clustering models to model inter-data dependence in many ways.

The distance dependent CRP represents the partition with *customer assignments*, rather than table assignments. While the traditional CRP randomly assigns customers to tables, the distance dependent CRP assigns customers to other customers. The random partition of the data, i.e., the table assignments, arises from these customer connections. When used in a Bayesian model, the posterior provides a new tool for flexible clustering of non-exchangeable data.

In Section 2 we develop the distance dependent CRP and discuss its properties. We use it in models of discrete data, both fully-observed and as part of a mixture model. In Section 3 we derive approximate posterior inference algorithms

Appearing in *Proceedings of the 27^{th} International Conference on Machine Learning*, Haifa, Israel, 2010.

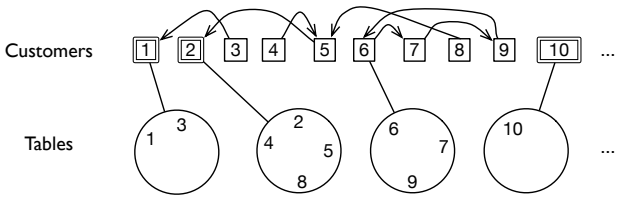

Figure 1. An illustration of the distance dependent CRP. The process operates at the level of customer assignments, where each customer chooses either another customer or no customer according to Eq. (2). Customers that did not choose to connect to another are indicated with a double box; customers that are linked to others are indicated by an arrow. The table assignments, familiar for the CRP, are derived from the customer assignments.

using Gibbs sampling. Finally, in Section 4 we describe an empirical study of three text corpora using the distance dependent CRP with time dependence, and we show that models based on the distance dependent CRP can provide a better fit to sequential data. We also show that its alternative formulation of the traditional CRP leads to a faster-mixing Gibbs sampler than the best one based on the original formulation. The distance dependent CRP provides a new CRP representation and efficient posterior inference algorithms for both non-exchangeable and exchangeable data.

2. Distance dependent CRPs

The Chinese restaurant process (CRP) induces a probability distribution on partitions. Consider a restaurant with an infinite number of tables, and a sequential process by which customers enter the restaurant and each sit down at a randomly chosen table. After N customers have sat down, their configuration at the tables represents a random partition. Customers sitting at the same table are in the same group of the partition.

In the traditional CRP, the probability of a customer sitting at a table is proportional to the number of other customers already sitting at that table (Pitman, 2002). Let z_i be the index of the ith customer's chosen table, the *table assignment*. Assume that the customers $z_{1:(i-1)}$ occupy K tables, and let n_k be the number of customers sitting at table k. The traditional CRP draws each z_i sequentially,

$$p(z_i = k \mid z_{1:(i-1)}, \alpha) \propto \begin{cases} n_k & \text{for} \quad k \leq K \\ \alpha & \text{for} \quad k = K+1. \end{cases} \tag{1}$$

When all N customers have been seated, their table assignments provide a random partition. Though described sequentially, the CRP is exchangeable. The probability of a particular partition of N customers is invariant to the order in which they sat down.

We now develop the *distance dependent CRP*, a CRP where the customers' table assignments can depend on external distances between them. While the random seating plan of the traditional CRP arises from customers sitting directly at tables, the seating plan of the distance dependent CRP arises from customers sitting with other *customers*.

Let the *customer assignment* c_i be the index of the customer to which the ith customer connects. (For example, in Figure 1 $c_4 = 5$ and $c_5 = 2$, and $c_3 = 1$.) Let d_{ij} be the distance measurement between customers i and j, let D be the set of all distance measurements between customers, and let f be a decay function (described in more detail below). The distance dependent CRP draws the customer assignments independently, conditioned on the distance measurements,

$$p(c_i = j \mid D, \alpha) \propto \begin{cases} f(d_{ij}) & \text{if} \quad i \neq j \\ \alpha & \text{if} \quad i = j. \end{cases} \tag{2}$$

A set of customer assignments maps to a partition, and thus the distance dependent CRP is also a distribution over partitions. The partition is described with *table assignments* $z_{1:N}$, where two customers are sitting at the same table if they are reachable from each other by a sequence of interim customer assignments. The mapping between representations is denoted $R(c_{1:N}) = z_{1:N}$. This relationship is not one-to-one. Two sets of customer assignments can lead to the same table assignment. Figure 1 illustrates the relationship between customer assignments and table assignments.

We highlight two properties of the distance dependent CRP. First, customer assignments do not depend on other customer assignments, only the distances between customers. The generative process of the traditional CRP, while leading to an exchangeable joint distribution, requires that each successive customer is sampled conditional on the previous customers. Second, j ranges over the entire set of customers, and so any customer may sit with any other. Customer assignments can contain cycles (e.g., customer 1 sits with 2 and customer 2 sits with 1) and this still produces a partition.

The distances and decay function let us define a variety of distributions.[1] For example, if each customer is time-stamped, then distance d_{ij} might be the time difference between customers i and j and the decay function can encourage customers to sit with those that are contemporaneous. If each customer is associated with a location in space, then distance d_{ij} might be the Euclidean distance between them and the decay function can encourage customers to sit with those that are in proximity. For many sets of distance measurements, the resulting distribution over partitions is no longer exchangeable. Distance dependent CRPs are appropriate when exchangeability is not a reasonable assumption.

[1] These model properties can be combined—consider $d'_{ij} = f(d_{ij})$—but it makes conceptual sense to separate them. The distances are thought of as innate properties of the customers, while the decay function is a modeling choice that mediates how distances affect the resulting distribution over partitions.

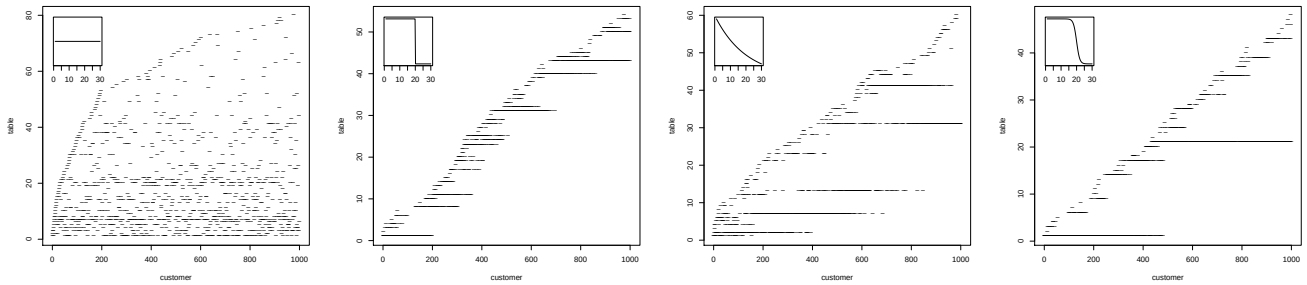

Figure 2. Draws from sequential CRPs. Illustrated are draws for different decay functions, which are inset: (1) The traditional CRP; (2) The window decay function; (3) The exponential decay function; (4) The logistic decay function. The table assignments are illustrated, derived from the customer assignments drawn from the distance dependent CRP. Each customer is represented by a point at the assigned table. The decay functions are functions of the difference in index between the current customer and each previous customer.

We consider several types of decay. The *window decay* $f(d) = 1[d < a]$ only considers customers that are at most distance a from the current customer. The *exponential decay* $f(d) = e^{-d/a}$ decays the probability of each customer exponentially with its distance to the linking customer. The *logistic decay* $f(d) = \exp(-d + a)/(1 + \exp(-d + a))$ is a smooth version of the window decay.

Sequential CRPs. In this paper, we focus our empirical study on the *sequential CRP*, where customers link to themselves or previous customers. Sequential CRPs are constructed with the distances and decay function. The distances satisfy $d_{ij} = \infty$ for $j > i$, and the decay function satisfies $f(\infty) = 0$. This guarantees that no customer links to a later customer, i.e., $p(c_i \leq i \mid D) = 1$. Several previous models can be derived with a sequential CRP. We obtain the model of Ahmed & Xing (2008) by multiplying the window decay function and exponential decay function. We obtain the model of Zhu et al. (2005) with a logistic decay function.

Figure 2 illustrates seating draws from sequential CRPs with each of the decay functions described above. These plots are at the *table* level. Rather than plot the links between customers, we plot the table at which each customer is sitting. Compared to the traditional CRP (also illustrated), customers tend to sit at the same table with other nearby customers. We emphasize that sequential CRPs are only one type of distance dependent CRP. Other distances, combined with Eq. (2), lead to a variety of other non-exchangeable distributions over partitions.

The traditional CRP is a sequential CRP. We can express the traditional CRP as a sequential CRP. We recover the traditional CRP when $f(d) = 1$ for $d \neq \infty$ and $d_{ij} = \infty$ for $i > j$. To see this, consider the marginal distribution of a customer sitting at a particular table, given the previous customers' assignments. The probability of being assigned to each of the other customers at that table is proportional to one. Thus, the probability of sitting at that table is proportional to the number of customers already sitting there. The probability of not being assigned to a previous customer is proportional to the scaling parameter α. This is precisely the traditional CRP distribution of Eq. (1).

Though these models provide the same distribution of partitions, the corresponding Gibbs samplers (for a mixture model based on the CRP) are different. In Section 4 we show that the Gibbs sampler for the dd-CRP construction, which operates at the customer level, is more efficient than the traditional Gibbs sampler, which operates at the table level (Neal, 2000).

2.1. Modeling data with a distance dependent CRP

We described the distance dependent CRP, a prior over partitions. We now describe two applications to Bayesian modeling of discrete data, one in a fully observed model and the other in a mixture model. These examples illustrate how to use the posterior distribution of the partitions, given data and an assumed generating process. We focus on discrete data and we use the terminology of document collections.[2] Observations are collections of words from a fixed vocabulary, organized into documents.

Language modeling. Each document is drawn from a distance dependent CRP for which the tables are embellished with IID draws from a base distribution over terms. (The documents share the same base distribution.) The generative process of a document is as follows. First, for each table draw a term from the base distribution. Second, place the data at tables via random customer assignments. Finally,

[2]CRP-based methods have been extensively applied to text modeling and natural language processing (Teh et al., 2007; Johnson et al., 2007). However, these models apply to any discrete data, such as genetic data and, with modification, to non-discrete data.

assign each data point to the term associated with its table. Subsets of the data exhibit a partition structure by sharing the same table. Note that multiple instances of the same term can be associated with different tables, each having drawn the same term from the base distribution. Instances of different terms are necessarily from different tables.

When using a traditional CRP, this is a simple Dirichlet-smoothed language model, with the Dirichlet random variable marginalized out. Such a model is effective because it captures the "contagious" nature of language—once a word is seen once, it is more likely to be seen again. When we use a sequential CRP, we are assuming that a word is more likely to occur near itself in a document. Words are still considered contagious, but the window of contagion is mediated by the decay function.

We define the model more formally. Recall $R(c_{1:N})$ maps customer assignments to table assignments. Define $R^*(c_{1:N})$ to contain one customer index from each table. For example, in the sequential CRP, $R^*(c_{1:N})$ can be the first customers to sit at each table, i.e., those customers who link to themselves. (This set is important to incorporate the notion of a base distribution into CRP-based models.) Given a decay function f, distances D, scaling parameter α, and base distribution G_0, N words are drawn:

1. For $i \in [1, N]$, draw $c_i \sim$ dist-CRP(α, f, D).
2. For $i \in [1, N]$, if $c_i \notin R^*(c_{1:N})$ then assign the word $w_i = w_{c_i}$. Otherwise, draw $w_i \sim G_0$.

For each document, we observe a sequence of words $w_{1:N}$ from which we can infer their seating assignments in the distance dependent CRP. (See Section 3 for algorithms for posterior inference.)

Mixture modeling. Bayesian nonparametric methods have been extensively applied to mixture modeling problems, where the inferential goal is to probabilistically divide the data into groups. DP mixtures provide a solution where the number of groups is unbounded. In the Chinese restaurant analogy, the number of tables that the customers (i.e., the data) occupy is unknown in advance. The posterior provides both a grouping and the number of groups.

The second application we study is akin to the CRP mixture or (equivalently) the Dirichlet process (DP) mixture, but differs in that the mixture component for an observation depends on the mixture component for nearby observations. Again, we endow each table with a draw from a base distribution G_0, but here G_0 is a distribution over component parameters and the unit of observation is a document. When analyzing documents with a CRP mixture, the base distribution is typically a Dirichlet (Teh et al., 2007).

Given a decay function f, distances D, scaling parameter α, and an exchangeable Dirichlet distribution with parameter λ, N M-word documents are drawn as follows,

1. For $i \in [1, N]$, draw $c_i \sim$ dist-CRP(α, f, D).
2. For $i \in [1, N]$,
   (a) If $c_i \notin R^*(c_{1:N})$ then set the parameter for the ith customer to $\theta_i = \theta_{c_i}$. Otherwise draw the parameter from the base distribution, $\theta_i \sim$ Dirichlet(λ).
   (b) Draw the ith document, $w_i \sim$ Mult(M, θ_i).

In Section 4, we will study the sequential CRP in this setting, choosing its structure so that contemporaneous documents are more likely to be clustered together. The distances d_{ij} can be the differences between indices in the ordering of the data, or lags between external measurements of distance like date or time.

Relationship to dependent DPs The distance dependent CRP mixture is an alternative to the dependent Dirichlet process (DDP) mixture, which is also an infinite clustering model that accounts for dependencies between the latent component assignments of the data (MacEachern, 1999). DDP mixtures posit collections of dependent random measures that generate the data. The DDP has been extended to sequential, spatial, and other types of dependence (Griffin & Steel, 2006; Duan et al., 2007; Xue et al., 2007).

DDP mixtures and distance dependent CRP mixtures are qualitatively different classes of models. DDP mixtures are Bayesian nonparametric models, based on random measures, while the distance dependent CRP mixtures generally are not.[3] In a DDP, close data are drawn from a similar distribution over clusters; in the distance dependent CRP, close data are likely to to arise from the same cluster.

Notably, DDP mixtures exhibit *marginal invariance*; distance dependent CRPs do not. This means that marginalizing over a particular customer in a DDP mixture gives the same probability distribution as if that customer were not included in the model. The distance dependent CRP does not generally have this property, allowing it to capture the way in which influence might be transmitted from one point to another, such as in a model of disease spread. (We note that many machine learning models, such as conditional random fields or Ising models, do not exhibit marginal invariance.) Thus, in the way they capture dependence, these classes of models make different assumptions. The appropriate choice of model depends on the modeling task at hand.[4]

[3] We have avoided calling the distance dependent CRP a "Bayesian nonparametric" model because it does not necessarily originate from a prior over the infinite space of probability measures, as the CRP originates from the DP. That said, both the CRP and the distance dependent CRP share a characteristic ability to let the data determine their number of clusters.

[4] In a longer paper we show that DDP mixtures and distance dependent CRP mixtures are nearly distinct (Blei & Frazier, 2009). The *only* distance dependent CRP mixture that is equivalent to

DDP mixtures and distance dependent CRP mixtures have a practical difference as well. In posterior inference, statisticians using DDPs must appeal to truncations of the stick-breaking representation of the random measures as the dependency between data precludes more efficient techniques that integrate out the component parameters and proportions. In contrast, distance dependent CRP mixtures are amenable to Gibbs sampling algorithms that integrate out these variables.

3. Posterior inference and prediction

In both applications described above, our goal is to compute the posterior distribution over partitions. As with models based on the traditional CRP, computing this posterior exactly is intractable. We approximate it using Gibbs sampling (Robert & Casella, 2004). We construct a Markov chain for which the stationary distribution is the posterior, run the chain to collect independent samples, and then use those samples to approximate the true posterior.

In both applications, the state of the chain is a set of customer assignments $c_{1:N}$. These assign each observation x_i to a customer j, where j is either the index of another customer or is equal to i. (We generically call observations x_i, corresponding to words in the language modeling application and documents in the mixture modeling application.)

The Gibbs sampler iteratively draws from the conditional distribution of each assignment given the other assignments and the observations. Let c_{-i} be the seating assignment with the outgoing link of customer i removed (i.e., the other assignments for sampling the ith assignment). Other customers still can link to i, and note that i is in its own table if no other customer links to it. For example in Figure 1, if $i = 5$ then c_{-i} defines 5 tables, where the table $\{2, 4, 5, 8\}$ under $c_{1:N}$ is split into the tables $\{2\}$ and $\{4, 5, 8\}$. The conditional distribution of c_i given the other assignments c_{-i}, the observed data $x_{1:N}$, and the base measure G_0 is,

$$p(c_i \mid c_{-i}, x_{1:N}, G_0) \propto p(c_i)p(x_{1:N} \mid c_i, c_{-i}, G_0). \quad (3)$$

The first term is the distance dependent CRP prior of Eq. (2). The second term is the likelihood of the data under the partition induced by $\{c_{-i}, c_i\}$.

3.1. Language modeling

We first consider the language modeling application, i.e., when we directly observe observations from G_0. Observations are a sequence of N words $w_{1:N}$ and recall that $R^*(c_{1:N})$ contains one customer index from each table (e.g., in Figure 1, we could take $R^*(c_{1:N}) = \{1, 2, 6, 10\}$).

a DDP mixture is the traditional CRP embellished with a base measure, i.e., the Dirichlet process mixture.

Eq. (3) becomes

$$p(c_i \mid c_{-i}, w_{1:N}, G_0) \propto p(c_i)1[w_i = w_{c_i}] \prod_{j \in R^*(c_{1:N})} p(w_j \mid G_0). \quad (4)$$

The product term is $p(w_{1:N} \mid c_i, c_{-i})$, which is the probability of independently drawing each table's unique word from G_0. This term can be further decomposed into two terms, one that depends on the conditioned variable c_i and one that does not,

$$\prod_{j \in R^*(c_{1:N})} p(w_j \mid G_0) = h(c_i, G_0) \prod_{j \in R^*(c_{-i})} p(w_j \mid G_0).$$

The quantity $h(c_i, G_0)$ is defined below. Since the product over $R^*(c_{-i})$ does not depend on c_i, Bayes rule provides $p(c_i \mid c_{-i}, w_{1:N}, G_0) \propto p(c_i)1[w_i = w_{c_i}]h(c_i, G_0)$.

To compute $h(c_i, G_0)$ consider the change from $R(c_{-i})$ to $R(c_{1:N})$. That is, how does the link c_i change the seating assignment? There are two cases: (1) The tables remain the same, i.e., $R(c_{-i}) = R(c_{1:N})$; this occurs if customer i links to a customer that is already seated at his table. (2) Two tables are joined; this occurs if customer i links to a customer at another table. In the first case, $h(c_i, G_0) = 1$. In the second $h(c_i, G_0) = 1/p(w_i \mid G_0)$, since w_i starts two tables under under c_{-i} but only one under $c_{1:N}$.

Finally, to sample from Eq. (4), we consider the two cases above with the two cases for $p(c_i)$ from Eq. (2), i.e., $c_i = i$ and $c_i \neq i$. (The four resulting cases reduce to three because when $c_i = i$, $R(c_{-i}) = R(c_{1:N})$.)

Note that we derived the Gibbs sampler for the general distance based CRP, as opposed to the sequential CRP. This sampler could be used, for example, to cluster image codewords in a spatial setting, or to cluster members of a social network. Moreover, this sampler does not hinge on discrete data; it is easily adapted to continuous data or count data.

Sampling the base distribution. We further consider a hierarchical model where G_0 is drawn from a prior, $G_0 \sim H(\lambda)$. The term $p(w_i \mid G_0)$ is replaced with the integral

$$p(w_i \mid c_i, c_{-i}, w_{-i}, \lambda) = \int p(w_i \mid G_0)p(G_0 \mid \{w_k : k \in R^*(c_{1:N})\}, \lambda)dG_0.$$

In the equation, the conditional distribution of G_0 only depends on those data in $R^*(c_{1:N})$ because each table contains exactly one draw from G_0. When H and G_0 form a conjugate pair, this integral is a straightforward Bayesian computation (Bernardo & Smith, 1994).

3.2. Mixture modeling

In mixture modeling, observations are a sequence of documents $\vec{w}_{1:N}$. Each document is assigned to a customer, and

the Gibbs sampler iteratively draws from the conditional distribution of each customer assignment in Eq. (3).

To compute the likelihood term, we introduce more notation. If $r \in R^*(c_{1:N})$, let S_r be the set of customer indices that are at the same table as r. For example, S_2 in Figure 1 is $\{2,4,5,8\}$. With this notation, the likelihood term is

$$p(\vec{w}_{1:N} \mid c_{-i}, c_i = j, \lambda) = \prod_{r \in R^*(c_{1:N})} p(\vec{w}_{S_r} \mid \lambda). \quad (5)$$

When G_0 is conjugate to the data likelihood, e.g., Dirichlet(λ) in the case of multinomial data, each term is a ratio of normalizing constants (Bernardo & Smith, 1994).

Finally, note that for sequential D and specific decay functions, this provides an alternative approximate inference method for the models of Ahmed & Xing (2008) and Zhu et al. (2005). Moreover, since the traditional CRP can be expressed as a sequential CRP, this is a new Gibbs sampler for CRP mixture models. We compare this sampler to the traditional collapsed Gibbs sampler in Section 4.

Prediction. In Section 4, we estimate the log likelihood of held-out documents given a set of observed documents. This is a measure of the predictive power of the model. We approximate this quantity from Gibbs samples.

Assume that D is sequential. Let $\vec{w}$ be a held out document, $\vec{w}_{1:N}$ be the previously observed documents, and $c^{(b)}_{1:N}$ be the bth sample from a Gibbs run of B samples. The predictive likelihood is

$$p(\vec{w} \mid \vec{w}_{1:N}, D, \alpha) \approx (1/B) \sum_{b=1}^{B} p(\vec{w} \mid c^{(b)}_{1:N}, \vec{w}_{1:N}),$$

where each term is

$$p(\vec{w} \mid c^{(b)}_{1:N}, \vec{w}_{1:N}) = \sum_{c=1}^{N+1} p(c \mid D, \alpha) p(\vec{w} \mid c, c^{(b)}_{1:N}, \vec{w}_{1:N}).$$

The first term is the prior probability of the new customer assignment given in Eq. (2). (The new customer sits alone when $c = N + 1$.) The conditional probability of the new document in the second term is a ratio of probabilities: The numerator is the marginal probability of the new document and the previous documents assigned to the partition of c; the denominator is the marginal probability of the previous documents assigned to the partition of c.

When D is not sequential, the predictive distribution is more complicated—it must be estimated as a ratio of likelihoods. This is possible, though potentially computationally intensive, by running two Gibbs samplers: one that includes the missing observation as a hidden variable, and one that does not. Other methods of estimating ratios of likelihoods might also be available (Meng & Gelman, 1998).

4. Empirical study

We studied the distance dependent CRP in the language modeling and mixture settings on three text data sets. We focus on time dependence, where the sequential ordering of the data is respected via the decay function and distance measurements (see Section 2). We explored a number of different parameterizations of the decay function. As we show below, the distance dependent CRP gives better fits to text data in both modeling settings.

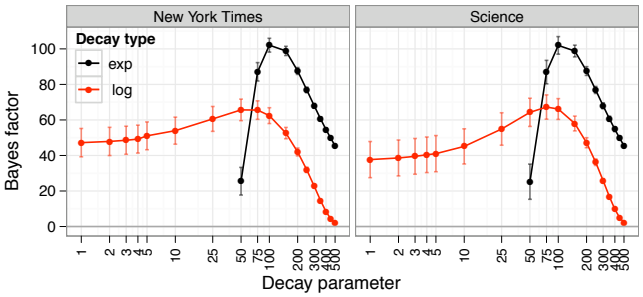

Figure 3. Bayes factors of the decayed CRP versus the traditional CRP on documents from *Science* and the *New York Times*. The black line at zero indicates an equal fit between the traditional CRP and decayed CRP. Also illustrated are standard errors across documents. In the logistic decay function, the slope parameter is fixed at 50. Values for the exponential decay at parameters less than 50 have large negative Bayes factors and are not illustrated. The shapes of the curves are similar for the two corpora; both are written in the same language.

Further, we compared the traditional Gibbs sampler for DP mixtures to the Gibbs sampler for the distance dependent CRP formulation of DP mixtures. The sampler presented here mixes faster than the traditional sampler.

Language modeling The fully-observed decayed CRP models were evaluated on two data sets: a collection of 100 OCR'ed documents from the journal *Science* and a collection of 100 world news articles from the *New York Times*. We assess sampler convergence visually, examining the autocorrelation plots of the log likelihood of the state of the chain (Robert & Casella, 2004).

We compare models by estimating the Bayes factor, the ratio of the document's probability under the decayed CRP to its probability under the traditional CRP (Kass & Raftery, 1995). For a decay function f, we estimate

$$BF_{f,\alpha} = p(w_{1:N} \mid \text{dist-CRP}_{f,\alpha}) / p(w_{1:N} \mid \text{CRP}_{\alpha}) \quad (6)$$

A value greater than one indicates an improvement of the distance dependent CRP over the traditional CRP. We use the method of Geyer & Thompson (1992) to estimate this ratio from posterior samples.

Figure 3 illustrates the average Bayes factors across documents for various decay functions, where the distance between words is the number of words between them. The logistic decay function always provides a better model than

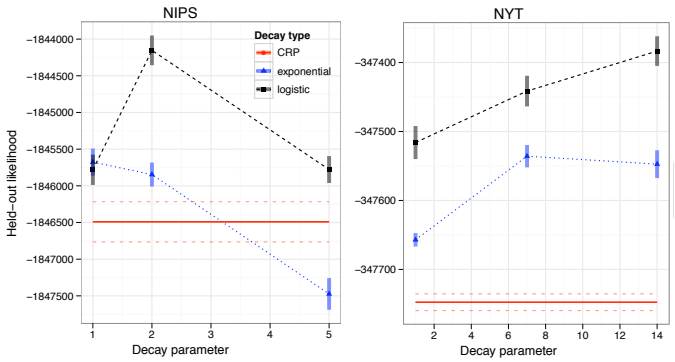

Figure 4. Predictive held-out log likelihood for the last year of NIPS and last three days of the *New York Times* corpus. Error bars denote standard errors across MCMC samples. On the NIPS data, the distance dependent CRP outperforms the traditional CRP for the logistic decay with a 2 day parameter. On the *New York Times* data, the distance dependent CRP outperforms the traditional CRP in almost all settings tested.

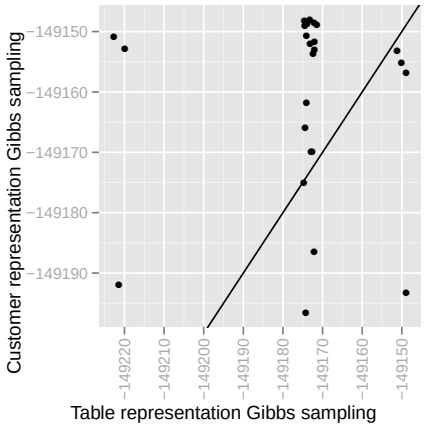

Figure 5. The log probability of MAP estimates from the table based Gibbs sampler (Neal, 2000) compared to the customer based Gibbs sampler, i.e., the sampler from Section 3 with the CRP. The data are 500 articles from the New York Times. Each point is a run of the sampler, started from the same place (each customer at a unique table), but with different permutations of the data. The customer-based sampler finds better MAP estimates in 19/26 runs, The sampler of Section 3 for CRP mixtures mixes faster than the traditional sampler.

the traditional CRP; the exponential decay function provides a better model at certain settings of its parameter. The hierarchical setting is pictured, with a Dirichlet prior on the unobserved G_0; the shapes of the curves are similar in the non-hierarchical settings.

Mixture modeling We examined the distance dependent CRP mixture on two text corpora. We analyzed one month of the *New York Times* (NYT) time-stamped by day, containing 2,777 articles, 3,842 unique terms and 530K observed words. We also analyzed 12 years of NIPS papers time-stamped by year, containing 1,740 papers, 5,146 unique terms, and 1.6M observed words. Distances D were differences between time-stamps.

In both corpora we use the last 250 articles as held out data. In the NYT data, this is three days of news; in the NIPS data, these are papers from the 11th and 12th year. (We retain the time stamps of the held-out data.) We evaluate the models by estimating the predictive likelihood of the held out data. The results are in Figure 4. On the NYT corpus, the distance dependent CRPs definitively outperform the traditional CRP. A logistic decay with a window of 14 days performs best. On the NIPS corpus, the logistic decay function with a decay parameter of 2 years outperforms the traditional CRP. In general, these results show that non-exchangeable models given by the distance dependent CRP mixture provide a better fit than the exchangeable CRP mixture.

Comparison to the traditional Gibbs sampler The distance dependent CRP can express a number of flexible models. However, as we describe in Section 2, it can also re-express the traditional CRP. In the mixture model setting, the Gibbs sampler of Section 3.2 thus provides an alternative algorithm for approximate posterior inference in DP mixtures. We compare this Gibbs sampler to the widely used collapsed Gibbs sampler for DP mixtures, i.e., Algorithm 3 from Neal (2000), which is applicable when the base measure G_0 is conjugate to the data likelihood.

The Gibbs sampler for the distance dependent CRP iteratively samples the customer assignment of a data point; the collapsed Gibbs sampler iteratively samples the cluster assignment of each data point. The practical difference between the two algorithms is that the distance dependent CRP based sampler can change several customers' cluster assignment via a single customer assignment. This allows for larger moves in the state space of the posterior and, we will see below, faster mixing of the sampler.

Moreover, the computational complexity of the two samplers is the same. Both require computing the change in likelihood of adding or removing either a set of points (in the distance dependent CRP case) or a single point (in the traditional CRP case) to each cluster. Whether adding or removing one or a set of points, this amounts to computing a ratio of normalizing constants for each cluster. This is where the bulk of the computation of each sampler lies.

To compare the samplers, we analyzed documents from the *New York Times* collection under a CRP mixture with scaling parameter equal to one and uniform Dirichlet base measure. Figure 5 illustrates the log probability of the MAP estimate of the partition structure under the CRP for each sampler.

Each point represents a permutation of the data and 1000 Gibbs iterations. (Even though the CRP is exchangeable, both Gibbs samplers are sensitive to the initial permutation of the data.) The data were started from the state where each customer is at a unique table. Out of 25 random permutations, the distance dependent CRP algorithm (representing the traditional CRP with a customer assignment representation) finds a more likely partition under the CRP than the table assignment sampler. This indicates that Gibbs sampler under the customer assignment representation is mixing faster than the traditional alternative.

5. Discussion

We developed the distance dependent Chinese restaurant process, a distribution over partitions that accommodates a flexible and non-exchangeable seating assignment distribution. The distance dependent CRP hinges on the customer assignment representation. We derived a general-purpose Gibbs sampler based on this representation, and examined sequential models of text.

The distance dependent CRP opens the door to a number of further developments in infinite clustering models. We plan to explore spatial dependence in models of natural images, and multi-level models akin to the hierarchical Dirichlet process (Teh et al., 2007). Moreover, the simplicity and fixed dimensionality of the corresponding Gibbs sampler suggests that a variational method is worth exploring as an alternative deterministic form of approximate inference.

6. Acknowledgments

David M. Blei is supported by ONR 175-6343 and NSF CAREER 0745520. The authors thank the anonymous reviewers for constructive comments.

References

Ahmed, A. and Xing, E. Dynamic non-parametric mixture models and the recurrent Chinese restaurant process with applications to evolutionary clustering. In *International Conference on Data Mining*, 2008.

Bernardo, J. and Smith, A. *Bayesian theory*. John Wiley & Sons Ltd., Chichester, 1994.

Blei, D. and Frazier, P. Distance dependent Chinese restaurant processes. *arXiv*, (0910.1022), 2009.

Duan, J., Guindani, M., and Gelfand, A. Generalized spatial Dirichlet process models. *Biometrika*, 94:809–825, 2007.

Geyer, C. and Thompson, E. Constrained Monte Carlo maximum likelihood for dependent data. *Journal of the American Statistical Association*, 54(657–699), 1992.

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

Johnson, M., Griffiths, T., and S., Goldwater. Adaptor grammars: A framework for specifying compositional nonparametric Bayesian models. In Schölkopf, B., Platt, J., and Hoffman, T. (eds.), *Advances in Neural Information Processing Systems 19*, pp. 641–648, Cambridge, MA, 2007. MIT Press.

Kass, R. and Raftery, A. Bayes factors. *Journal of the American Statistical Association*, 90(430):773–795, 1995.

MacEachern, S. Dependent nonparametric processes. In *ASA Proceedings of the Section on Bayesian Statistical Science*, 1999.

Meng, X. and Gelman, A. Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. *Statistical Science*, 13(2):163–185, 1998.

Neal, R. Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 9(2):249–265, 2000.

Pitman, J. *Combinatorial Stochastic Processes*. Lecture Notes for St. Flour Summer School. Springer-Verlag, New York, NY, 2002.

Robert, C. and Casella, G. *Monte Carlo Statistical Methods*. Springer Texts in Statistics. Springer-Verlag, New York, NY, 2004.

Teh, Y., Jordan, M., Beal, M., and Blei, D. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2007.

Xue, Y., Dunson, D., and Carin, L. The matrix stick-breaking process for flexible multi-task learning. In *International Conference on Machine Learning*, 2007.

Zhu, X., Ghahramani, Z., and Lafferty, J. Time-sensitive Dirichlet process mixture models. Technical Report CMU-CALD-05-104, Carnegie Mellon University, 2005.