Distance Dependent extensions of the Chinese Restaurant Process

Willem Wybo WYBO.WILLEM@GMAIL.COM
Camille Colle CAMILLE.COLLE@UGENT.BE
Pieter-Jan Kindermans PIETERJANKINDERMANS@UGENT.BE
Benjamin Schrauwen BENJAMIN.SCHRAUWEN@UGENT.BE
Ghent University, Electronics and Information Systems department, Sint Pietersnieuwstraat 41, 9000 Ghent, Belgium

Keywords: Chinese Restaurant Process, Distance Dependent Chinese Restaurant Process, Averaged Distance Dependent Chinese Restaurant Process

Abstract

In this paper we consider the clustering of text documents using the Chinese Restaurant Process (CRP) and extensions that take time-correlations into account. To this purpose, we implement and test the Distance Dependent Chinese Restaurant Process (DDCRP) for mixture models on both generated and real-world data. We also propose and implement a novel clustering algorithm, the Averaged Distance Dependent Chinese Restaurant Process (ADDCRP), to model time-correlations, that is faster per iteration and attains similar performance as the fully distance dependent CRP.

1. Introduction

Non-parametric clustering algorithms have been used often in the classification of text documents. These algorithms exist in plenty of variations, that are generally referred to through some metaphor with a restaurant where exotic cuisine is served. The simplest of them is the Chinese Restaurant Process, and it will provide a starting point for the discussions in this paper.

However, these algorithms generally assume that the data is independent and identically distributed (iid). Such an assumption is questionable in many cases. Today's newspaper articles show a high correlation with what was in the news yesterday, and scientific papers tend to be generated in areas were researchers see interest, generally through what was written earlier by their colleagues.

Recently a new clustering algorithm, the Distance Dependent Chinese Restaurant Process, was proposed (Blei & Frazier, 2011), that drops this assumption of iid-draws by including a dependence on the distance between data points. Note that the DDCRP can be used for many applications. In this work we restrict ourselves to document clustering with documents that have a time stamp and are assumed to exhibit time-correlations. We test and compare this algorithm with the normal CRP and also propose a novel variation to it, the Averaged Distance Dependent CRP, where the distance is defined between data points and clusters. We also test this new algorithm and compare it with the previous ones.

This paper is structured as follows: first we provide a short introduction to the algorithms, where the focus is not on completeness but on summarizing the necessary concepts. Next we show what draws from all three processes look like and finally we discuss our testing method and apply it to a real-world dataset.

2. Algorithms

In document clustering, documents are often modelled as a bag of words, in which the order of the words is ignored. Through this assumption, the documents can be modelled as if they are generated by a latent topic, where the topic governs the parameters of the multinomial distribution from which their words are drawn. These parameters are assumed to be drawn from a Dirichlet prior, that is unobserved, but can be learned from the data.

Appearing in *Proceedings of BeneLearn and PMLS 2012*.

More precisely, a list of all unique words $w = (w_1, \ldots, w_N)$ that occur in the documents can be constructed. A document is then represented by a vector $x = (n_1, \ldots, n_N)$, where n_i denotes the number of times the word w_i appears in the document. If a documents belongs to a cluster, it is assumed to have been generated from a multinomial distribution with parameters determined by that cluster. Denoting the parameters of this distribution as $\theta = (\theta_1, \ldots, \theta_N)$, with $\sum_n \theta_n = 1$, the probability to find a certain document x is given by

$$p(x \mid \theta) \sim \theta_1^{n_1} \ldots \theta_N^{n_N} \tag{1}$$

In the algorithms discussed in this paper, the probability of the parameters θ is assumed to follow a Dirichlet distribution. We denote this distribution as G_0 and its parameters as $g_1, \ldots, g_N$. Thus the likelihood of a given vector θ is

$$p(\theta \mid G_0) \sim \theta_1^{g_1 - 1} \ldots \theta_N^{g_N - 1} \tag{2}$$

In mixture models the parameters $g_1, \ldots, g_N$ are determined by the documents already present in the cluster.

2.1. The Chinese Restaurant Process

The Chinese Restaurant process is a process that generates a distribution over partitions (topics) from which sampling is possible (Neal, 2000). The simplest of the sampling methods is based on Gibbs sampling (Bishop, 2007) and through this sampling the CRP becomes a powerful clustering algorithm, based on probabilistic cluster assignments. It has many interesting mathematical properties that are reviewed in (Teh et al., 2006). We repeat only the very basics.

In the CRP, a data point can be assigned to previously formed clusters with a probability that is proportional to the amount of points already in such a cluster. A data point can also be assigned to a new cluster with a certain probability. This can be expressed intuitively trough a metaphor, from which the CRP borrows its name, where customers successively enter a Chinese restaurant and decide to sit at tables (the clusters that will be formed) with a probability proportional to the amount of customers already sitting at the table. They can also decide to sit at a new table with a certain probability proportional to a fixed constant. If we denote the data point that is to be assigned as i, the cluster assignment of i as c_i, the partition of the previous data points defined by the clustering as $\mathbf{z}_{i-1} = (z_1, \ldots, z_n)$ and the constant to which the probability to start a new cluster is proportional as α, the previous considerations can be written as:

$$p(c_i = z_k \mid \mathbf{z}_{i-1}, \alpha) \propto \begin{cases} n_k & \text{if } z_k \in \mathbf{z}_{i-1} \\ \alpha & \text{if } z_k = \text{new table} \end{cases} \tag{3}$$

where n_k denotes the number of data points in cluster k. As described previously, documents are considered to be generated from a multinomial distribution, with parameters determined by the cluster assignment. The probability that a document x_i is assigned to a cluster z_k, containing a set of documents $\{x_j \in z_k\}$, is then also proportional to the predictive probability that this document could have been generated by this cluster. This predictive probability is obtained by integrating out the parameters θ weighted by their Dirichlet likelihood:

$$p(x_i \mid \{x_j \in z_k\}) \propto \int p(x_i \mid \theta) p(\theta \mid \{x_j \in z_k\}, G_0) d\theta \tag{4}$$

The Gibbs sampling algorithm for the CRP then successively removes documents from their cluster and reassigns them to one of the other clusters according to the appropriate probabilities. These probabilities follow from combining equation (3) and (4) and are given by

$$p(c_i = z_k \mid \mathbf{z}_{i-1}, \mathbf{x}, G_0) \propto \begin{cases} n_k \int p(x_i \mid \theta) p(\theta \mid \{x_j \in z_k\}, G_0) d\theta \text{ if } z_k \in \mathbf{z}_{i-1} \\ \alpha \text{ if } z_k = \text{new table} \end{cases} \tag{5}$$

Since the multinomial and Dirichlet are conjugate distributions, a closed form of integral 4 exists, allowing for computationally effective sampling.

2.2. The Distance Dependent Chinese Restaurant Process

The distance dependent Chinese restaurant process can be seen as an extension of the normal CRP. However, now the customers don't sit at tables, but are linked to each other, and the tables arise merely as clusters of connected customers. The customers are assumed to have a sequential variable (e.g.. a time stamp) through which we can calculate a distance d_{ij}. The probability for a customer i to sit with customer j is then

$$p(c_i = j \mid D, \alpha) \propto \begin{cases} f(d_{ij}) & \text{if } j \neq i \\ \alpha & \text{if } j = i \end{cases} \tag{6}$$

where f is some kind of decaying function, D the distance matrix and c_i represents the customer to which i links. To model time correlations one usually takes $d_{ij} = \infty$ if j has a larger time stamp then i, so that no customer can be assigned to a future customer. This decay function can be chosen so as to suit the modellers needs. Good choices for f to model time-correlation include an exponential function or a logistic decay function. If $f(d) = 1$ for $d < \infty$, the distance dependent CRP reduces to the normal CRP.

Since in the DDCRP tables arise as clusters of connected customers, the Gibbs sampler will have a slightly different form than the Gibbs sampler for the normal CRP. Instead of doing probabilistic cluster assignments, the Gibbs sampler will now do probabilistic customer assignments. These assignments probabilities are proportional to (6) and to the appropriate mixture probabilities. If $\mathbf{z}(\mathbf{c}_{-i})$ is a partition with clusters $(z^1, ..., z^n)$, formed by the links c_j $(j \neq i)$, then the probability that the customer i will have a link c_i is given by

$$p(c_i = j \mid \mathbf{c}_{-i}, \mathbf{x}, \alpha, D, G_0) \propto \begin{cases} \alpha \text{ if } c_i = i \\ f(d_{ij}) \text{ if } c_i \text{ does not join two clusters} \\ f(d_{ij}) \frac{p(\mathbf{x}_{z^k(\mathbf{c}_{-i}) \cup z^l(\mathbf{c}_{-i})} \mid G_0)}{p(\mathbf{x}_{z^k(\mathbf{c}_{-i})} \mid G_0) p(\mathbf{x}_{z^l(\mathbf{c}_{-i})} \mid G_0)} \\ \quad \text{if } c_i \text{ joins clusters } k \ \& \ l \end{cases} \tag{7}$$

Again, as we are modelling text documents, we have words that are drawn from a multinomial distribution and clusters from a Dirichlet distribution. Hence the posterior probability

$$p(\mathbf{x}_{z^k(\mathbf{c})} \mid G_0) = \int \prod_{i \in z^k(\mathbf{c})} p(x_i \mid \theta) p(\theta \mid G_0) d\theta \tag{8}$$

can again be written in closed form through their respective conjugacy.

2.3. The averaged distance dependent CRP

In this section we introduce a novel clustering algorithm, the averaged distance dependent CRP (ADDCRP), that is a hybrid between the DDCRP and the normal CRP. In this algorithm, the distance is no longer defined between individual data points, but between data points and clusters, through an averaging procedure. If $\mathbf{z}(\mathbf{c}_{-i})$ is a partition $(z^1, \ldots, z^n)$ formed by the cluster-assignments $(c_1, \ldots, c_{i-1})$, and t_i denotes the time stamp of the i'th document, the distance between a data point i and a cluster z^k is defined as

$$d_{ik} = t_i - \frac{1}{|I_{ik}|} \sum_{j \in I_{ik}} t_j \tag{9}$$

where $I_{ik} = \left\{ j | j \in z^k \wedge t_j < t_i \right\}$ is the set of documents in cluster k with time stamps smaller than the time stamp of document i. Other definitions of this distance are also possible, for instance by taking a weighted mean

$$d_{ik} = t_i - \sum_{j \in I_{ik}} \frac{f(t_i - t_j)}{\sum_{l \in I_{ik}} f(t_i - t_l)} t_j \tag{10}$$

and hence attributing more importance to closer data points, or by simply taking the closest point

$$d_{ik} = \min_{j \in I_{ik}} (t_i - t_j) \tag{11}$$

to which we will respectively refer as the weighted ADDCRP and the minimal ADDCRP. Note that if $|I_{ik}| = 0$, we set $d_{ik} = \infty$, so that no data point can be assigned to clusters that contains only later data points.

Cluster assignments are then drawn according to

$$p(c_i = z_k \mid D, \mathbf{z}_{i-1} \alpha) \propto \begin{cases} f(d_{ik}) n_k & \text{for } z_k \in \mathbf{z}_{i-1} \\ \alpha & \text{for } z_k = \text{new table} \end{cases} \tag{12}$$

It can be readily seen that for $f(d) = 1$ if $d < \infty$ this reduces again to the normal CRP. The probabilities of assignment during a Gibbs-sampling run are given by an analogous formula as equation (5) for the normal CRP, but now they are weighted with a factor $f(d)$.

$$p(c_i = z_k \mid \mathbf{z}_{i-1}, \mathbf{x}, G_0) \propto \begin{cases} f(d_{ik}) n_k & \int p(x_i \mid \theta) p(\theta \mid \{x_j \in z_k\}, G_0) d\theta \\ & \text{if } z_k \in z_{i-1} \\ \alpha & \text{if } z_k = \text{new table} \end{cases} \tag{13}$$

3. Drawing from the CRP and its variations

A useful way to gain intuitive understanding in what draws from all three processes look like, is to run them in a generative way. Successive data points are assigned to the previous data points according the formulas (3), (6) and (12) for respectively the normal

CRP, the fully distance dependent CRP and the averaged distance dependent CRP. These assignments are visualized in figure (1) for an exponential decay function $f(d) = e^{-\beta d}$ and for a few different values of the parameters α and β. As one can see, draws from the DDCRP and the ADDCRP look reasonably similar, as the amount of formed clusters and their average length in time are approximately equal, whereas draws from the CRP are very different. This leads us to believe that the DDCRP and the ADDCRP have similar performance.

For testing purposes, we also generated full 'documents' of words drawn from a number of topics. That way we could test and compare models on datasets that are more manageable than real world datasets. We could for instance choose the size of the vocabulary. We now describe the procedure that we used to generate those datasets.

Beside drawing assignments c_i for data points x_i by using formulas (3), (6) and (12), words should, in a fully generative context, be drawn from the predictive probabilities

$$p(x_i \mid \{x_j \in z_k\}) = \int p(x_i \mid \theta) p(\theta \mid \{x_j \in z_k\}, G_0) d\theta \tag{14}$$

where z_k is the mixture component to which c_i points. For practical reasons we followed a simpler approach, where our vocabulary, containing W words, was split up in a number of C different classes. For each mixture component, we selected one or two of those classes as main 'topics', and generated documents largely from those main topics. On average we selected about two thirds of the words in a document from those main topics in a random fashion. The other third of the words was selected from random other topics in the vocabulary. For all our tests on generated data we used $W \simeq 800 - 900$, $C = 16$ and on average 600 words per document.

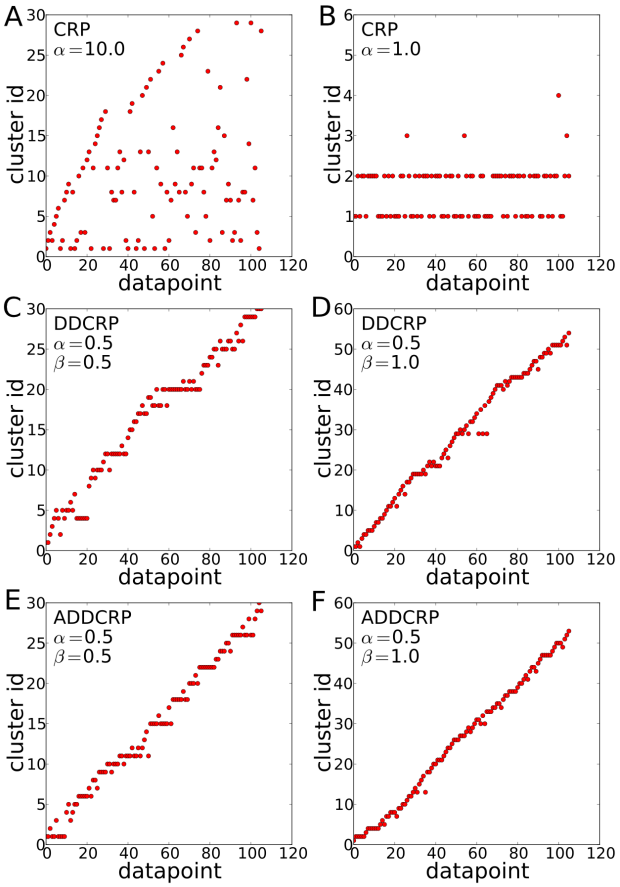

Figure 1. Draws from (a), (b) the normal CRP (c), (d) the fully distance dependent CRP and (e), (f) the averaged distance dependent CRP

4. Runtime

In this paragraph we consider the scaling properties of the runtime of all three algorithms as a function of the number of documents N. In the CRP and the ADDCRP a document is assigned to one of the cluster according to equations (5) (CRP) and (12) (DDCRP). If there are K clusters, one has to calculate this likelihood K times. Since the samplers successively remove and reassign all elements once every iteration, an iteration scales as $O(KN)$. In the DDCRP, one calculates the link assignment probabilities for every document (7). Nevertheless it is only necessary to evaluate the computationally heavy mixture integrals (8) for clusters that merge. This can be done in advance and thus this step also scales as K. However, the bookkeeping in the DDCRP is more involved, as one has to track the cluster through a list of links. This causes the algorithm to scale as $O(KN^2)$ rather than $O(KN)$. We show this effect in figure 2 for data generated by the procedure describe in paragraph 3. Note that these considerations only apply to the runtime per iteration. As Blei and Frazier note, the number of iterations until convergence can be much smaller in the DDCRP, as that sampler can move whole chunks of clusters at once, whereas the samplers for the CRP and ADDCRP only move one element at a time. This depends however heavily on the data at hand, and we did not notice this effect in our experiments.

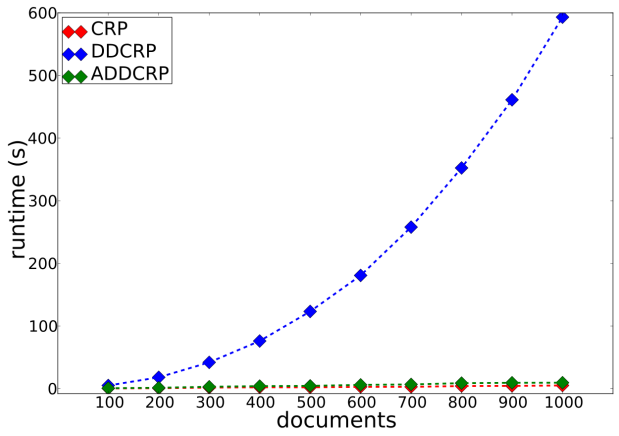

Figure 2. Average runtime per iteration as a function of the number of documents

5. Prediction

The goal of analysing data with time-correlations is often to estimate the likelihood of future data points. Suppose that N documents, represented by the vector $\mathbf{x}$, are clustered by an algorithm in a partition $(z^1, \ldots, z^n)$ by the cluster assignments $\mathbf{c} = (c_1, \ldots, c_N)$ (in the case of the distance dependent CRP, $\mathbf{c}$ denotes the element assignments). The total predictive likelihood for a given later document x_{new} is then given by

$$p(x_{new} \mid \mathbf{x}, D, G_0, \alpha) = \sum_{c_{new}} \sum_{\mathbf{c}} p(c_{new} \mid \mathbf{c}, D, \alpha) \; p(x_{new} \mid c_{new}, \mathbf{c}, \mathbf{x}, G_0) p(\mathbf{c} \mid \mathbf{x}, D, \alpha, G_0) \quad (15)$$

where D is the distance matrix. The last factor on the right hand side denotes the probability of a given partition, given the model and the data. Since only sequential data is considered, the assignment of a new data point does not change the probability of the previous assignments. Hence the sum over $\mathbf{c}$ can be estimated by averaging over different clustering runs. The second factor is the probability of the new data point under the assignment c_{new}. This can be computed with the standard inference methods. The first factor is the probability of that assignment. As can be seen from (3), (6) and (12), this probability is in the most general case only dependent on the previous assignments and the distance matrix. Under the distance dependent CRP this probability becomes independent of the previous assignments and (15) reduces to

$$p(x_{new} \mid \mathbf{x}, D, G_0, \alpha) = \sum_{c_{new}} p(c_{new} \mid D, \alpha) \; \sum_{\mathbf{c}} p(x_{new} \mid c_{new}, \mathbf{c}, \mathbf{x}, G_0) p(\mathbf{c} \mid \mathbf{x}, D, \alpha, G_0) \quad (16)$$

Under the normal CRP it follows that these assignment probabilities are all equal, and therefore equation (16) simplifies further to

$$p(x_{new} \mid \mathbf{x}, D, G_0, \alpha) = \frac{1}{N} \sum_{c_{new}} \sum_{\mathbf{c}} p(x_{new} \mid c_{new}, \mathbf{c}, \mathbf{x}, G_0) p(\mathbf{c} \mid \mathbf{x}, D, \alpha, G_0) \quad (17)$$

This can be seen from the interpretation of a normal CRP as a distance dependent CRP with $f(d) = 1$ if $d < \infty$.

6. Held-out likelihood as a comparative test

To test the performance of our model we follow the approach of Blei and Frazier (Blei & Frazier, 2011) and compute the so-called held-out likelihood. Suppose a dataset contains M documents and the clustering is performed on a smaller number N of earlier documents. The held-out likelihood of a later document is then defined as the predictive likelihood for this document, given the N earlier documents and the clustering. This is nothing else than equation (15). The held-out likelihood is a measure of how well the held-out data can be predicted by the mixture components created by the clustering algorithm. Hence algorithms that achieve a higher likelihood can be seen as better performing in this context.

Blei and Frazier computed this held-out likelihood on real-world datasets of newspaper articles and scientific papers. They concluded that in most cases the DDCRP is a better model than the normal CRP. We replicate those tests on another dataset and also perform them for the ADDCRP. We used abstracts from award winning papers from the National Science Foundation[1], with submission dates ranging from 1989 to 1995, where we removed a standard list of stop-words and words that appeared only once in an abstract.

If we look at the held-out likelihood in function of the training set size, figure (3a), we see that the likelihood increases as the training set size is increased, the expected behaviour.

[1] `http://kdd.ics.uci.edu/databases/nsfabs/`

In figure (3b) we compare the different clustering models. Both the DDCRP and the ADDCRP perform better than the CRP. Their relative performance is however dependent on the choice of the decay parameter, which has to be adjusted to the data.

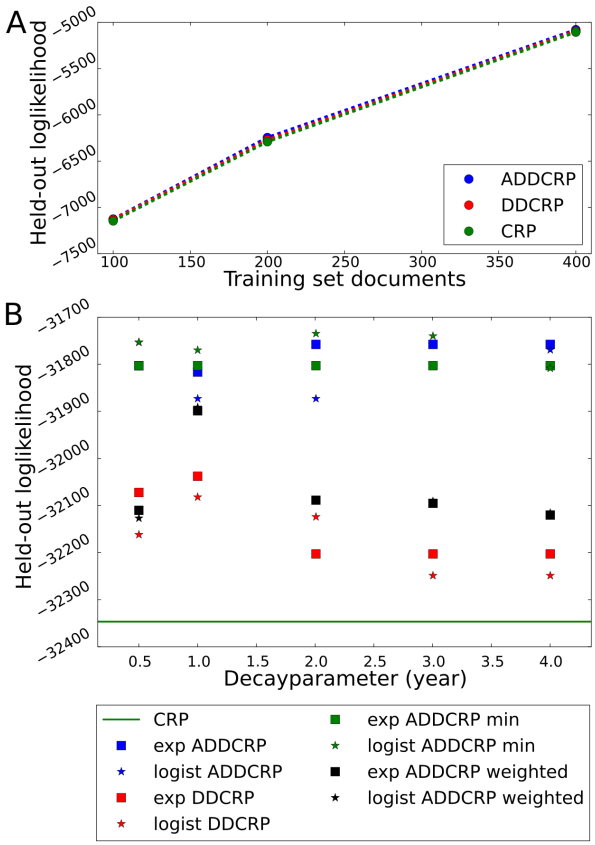

Figure 3. (a) The held-out likelihood in function of the number of training documents, for the CRP, DDCRP and ADDCRP. Here the held-out likelihood is averaged over 10 test documents. (b) Held-out likelihood in function of the decay parameters for the various models for two different decay functions, exponential: $f(d;a) = \exp(-d/a)$, and the logistic decay function: $f(d;a,b) = 1/\left(1+\exp([-d+a]/b)\right)$. Where b is fixed at 0.5 year. Note that as the CRP has no dependency on the decay parameter, it is just a constant. Here the held-out likelihood is averaged over 50 test-documents

7. Conclusion

We have implemented and tested two methods, a new one (ADDCRP) and an already existing one (DDCRP), for clustering time-correlated data in the context of mixture models applied to document modelling. We introduced the ADDCRP as a method to model time-correlations that scales in a similar way as the CRP if the document number is increased. Implementation-wise, the ADDCRP has the same complexity as the CRP, whereas the DDCRP is more complex because the clusters have to be tracked through a list of links. We found that the ADDCRP also achieved better performance than the DDCRP on our data, according to the held-out likelihood measure. The performance is however dependent on the choice of the decay parameter. We have shown that for their main goal, the prediction of future data, all the distance dependent variants reach better performance than the original CRP. Hence they form an interesting subject of future research. For instance, a question that arises naturally is whether an algorithm can be found that can learn the decay parameter β from the data. If some probabilistic framework could be found for this, algorithms as discussed here could become very powerful. Another question that relates to our newly proposed algorithm, is whether it can also be applied successfully to other domains where non-parametric clustering algorithms are used. Applications where the DDCRP can outperform the CRP are numerous (image compression, language modelling,...). Hence an interesting topic for future research would be to investigate whether the ADDCRP can match the performance of the DDCRP in those applications. If performance would turn out to be equivalent, the modellers choice will ultimately be determined by the trade-off between the convergence time and the runtime per iteration, both can vary heavily dependent on the application. Modelling time- and other correlations through a distance dependence is not yet explored very thoroughly in the context of non-parametric clustering algorithms, and as such these questions are very interesting to address in future research.

References

Bishop, C. M. (2007). *Pattern recognition and machine learning (information science and statistics)*. Springer. 1st ed. 2006. corr. 2nd printing edition.

Blei, D. M., & Frazier, P. (2011). Distance dependant chinese restaurant process. *Journal of Machine Learning Research*, *12*, 2461–2488.

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

Teh, Y. W., Jordan, M. I., Beal, M. J., & Blei, D. M. (2006). Hierarchical dirichlet processes. *Journal of the American Statistical Association*, *101(476)*, 1566–1581.