DIRICHLET PROCESS MIXTURE MODELS FOR TIME-DEPENDENT CLUSTERING

Kezi Yu and Petar M. Djurić

Department of Electrical and Computer Engineering
Stony Brook University, Stony Brook, NY 11794 (USA)
Email: {kezi.yu,petar.djuric@stonybrook.edu}

ABSTRACT

In many problems of signal processing, an important task is the classification of data. A group of methods that has attracted much interest for this purpose are the nonparametric Bayesian methods, and in particular, those based on the Dirichlet process. A useful metaphor for various generalizations of the Dirichlet process has been the Chinese restaurant process. Often the task of classification must be carried out in a sequential manner, and to that end the concepts from Bayesian non-parametrics cannot be applied straightforwardly. Recently, we introduced the notion of Chinese restaurant process with finite capacity to allow for classification of data on a time-varying basis. In this paper, we introduce the hierarchical Chinese restaurant process with finite capacity to provide further flexibilities to the process of classification. We show a generative model based on the process and then describe how to make online inference using the model. We demonstrate the approach with computer simulations.

1. INTRODUCTION

Classification of data is of significant importance in machine learning [1]. A standard approach is to use mixture models to cluster data. If the number of clusters (mixands) is known beforehand, one could employ algorithms such as expectation-maximization (EM) to estimate the parameters of each mixture and get the clustering result [2]. However, in many cases, it is difficult to decide the number of mixands in advance. One common approach is to fit several models and then select the best of them using model selection techniques [3]. Model selection metrics often include two terms: one that measures how well the data are fit by the model, and the other that quantifies the complexity of the model. The best model is considered the one that yields the best trade-off between performance and complexity.

Another popular approach in dealing with this problem is to use Dirichlet process mixture models (DPMMs) [4]. DPMMs do not specify the number of mixands in advance. Instead, the number of mixands increases as more data are observed. The parameters and the number of mixands are determined by the data via the mechanism of posterior inference. During this process, the tuning of parameters is minimal.

One weakness of the Dirichlet process models is that they cannot be applied in sequential processing for obtaining dynamic clustering of the data. In our recent work [5], we investigated a variation of the Dirichlet process, which we referred to as Chinese restaurant process with finite capacity (CRPFC). According to this model, at any time we process a limited amount of data. Once the number of data samples reaches a limit, new data are processed only after the "oldest" data are removed. This scheme readily enables sequential processing, where the computations do not grow with time.

In this paper, we investigate the CRPFC for modeling time-varying mixtures. In the sequel, a customer and a table represent a data point and a mixand, respectively. We also show how we make inference with CRPFC. We study time-varying Gaussian mixtures and demonstrate with simulations how the number of mixands evolve as new data come and old data are removed. Furthermore, we propose and employ hierarchical CRPFC mixture models to analyze several time series jointly.

The modification of standard DPMMs and extension to a hierarchical structure are applicable in various real-world scenarios. In [6], the authors used standard hierarchical DPMMs to analyze time series signals obtained from different fetuses. The employment of CRPFC mixture models to process the data can provide timely clustering results and real-time assessment of fetal health.

The main contribution of the paper is the extension of the CRPFC to a mixture model. We also propose a hierarchical formulation of the CRPFC and a corresponding mixture model. We apply the proposed models for generation of data and then in reverse, we use the data to make inference about the models, and in particular how the clustering of the data varies with time.

The paper is organized as follows. In the next section, we provide a brief background about DPMMs and hierarchical Dirichlet processes (HDPs). In Section 3, we propose DPMMs and hierarchical DPMMs for time dependent clustering and show how we can make inference using these models. In the following section, we provide simulation results based on these models. Finally, we conclude the paper in Section 5.

2. BACKGROUND

2.1. Dirichlet Process Mixture Models

DPMMs, unlike standard mixture models, allow for the presence of countably infinite number of mixands in the data. This is achieved by using a Dirichlet process as the prior of the mixtures and parameters, i.e.,

$$\begin{aligned} G|\alpha, H &\sim DP(\alpha, H) \\ z_i|G &\sim G \\ x_i|z_i &\sim F(z_i), \end{aligned} \tag{1}$$

where G is a random probability measure, DP stands for Dirichlet process, H is a base measure, α is a concentration parameter, z_i is a random variable, x_i is an observation, and F is the distribution of the mixands.

An insightful description of the Dirichlet process is by the Chinese restaurant process metaphor [7]. Consider a restaurant with infinite number of tables. Let z_i denote the dish that customer i is

This work was supported by NIH under Award 1R21HD080025-01A1 and NSF under Award CCF-1320626.

served. The conditional distribution of z_i given $z_j, j = 1, 2, \cdots, i-1$ is

$$z_i|z_1, \ldots, z_{i-1}, \alpha, H \sim \sum_{k=1}^{K} \frac{m_k}{i-1+\alpha}\delta_{\theta_k} + \frac{\alpha}{i-1+\alpha}H, \quad (2)$$

where m_k is the number of customers seated at table k, θ_k is the dish served at table k, and K is the number of tables already occupied. The dishes are drawn from H. In our setting, we consider that each table corresponds to a cluster. To complete the DPMM, we need to draw the actual serving of the ith customer, x_i. It comes from $F(z_i)$, and in signal processing context, the drawn value represents a data sample.

In theory, the number of mixands generated by DPMMs can be infinite. This number grows logarithmically with the number of data samples (approximately as $O(\alpha log N)$) [9]. The classification of the data given a DPMM is obtained by posterior inference methods, such as Markov chain Monte Carlo (MCMC) sampling [4]. One of these methods, the Gibbs sampling method, and in particular the collapsed Gibbs sampling or blocked Gibbs sampling, is preferable. A collapsed Gibbs sampler integrates out one or more variables whose values are not of importance. This makes the algorithm more efficient since less variables need to be sampled during inference. An alternative approach for approximating the posterior is variational inference [10].

2.2. Hierarchical Dirichlet Processes

DPMMs are useful in tasks of unsupervised classification of data from one single set. Consider now the more general problem of classification by using multiple data sets where the clusters are shared across the whole corpus of data sets. We want to get the clustering information not only within each set, but also jointly over all the sets. Hierarchical Dirichlet process mixture models (HDPMMs) are suitable for dealing with these types of problems.

The HDP can be constructed by recursively drawing the base measure G_j from a Dirichlet process G_0, which itself is also a draw from a Dirichlet process. This will guarantee that all the G_j's share the same support. We use the Chinese restaurant franchise metaphor to further explain the construction process. Consider a restaurant franchise with a shared menu across the restaurants. One dish is ordered at each table in each restaurant, and shared by all the customers seated at that table. Different tables in different restaurants can serve the same dish.

Let x_{ji} be the ith customer in the jth restaurant, and z_{jt} be the dish served at table t in restaurant j. We also introduce K iid random variables $\phi_1, \ldots, \phi_K$ generated from H to represent global dishes. Here we also need a notation for counts. Let n_{jtk} be the number of customers in restaurant j at table t served dish k, and m_{jk} be the number of tables in restaurant j serving dish k. Marginal counts are represented by dots. For example, $n_{jt\cdot}$ represents the number of customers in restaurant j at table t, and $m_{\cdot k}$ represents the number of tables serving dish k.

A new customer in a restaurant can either choose an occupied table according to the probability proportional to the number of customers already seated in the table, or get a new table. In the jth restaurant the choice is based on the probability measure G_j, where

$$G_j|\alpha_0, G_0 \sim DP(\alpha_0, G_0), \quad (3)$$

where

$$G_0|\gamma, H \sim DP(\gamma, H). \quad (4)$$

The conditional distribution of z_{ji} given $z_{j1}, z_{j2}, \ldots, z_{j,i-1}$, α_0, and G_0 is given by

$$z_{ji}|z_{j1}, z_{j2}, \ldots, z_{j,i-1}, \alpha_0, G_0 \sim \sum_{t=1}^{m_{j\cdot}} \frac{n_{jt\cdot}}{i-1+\alpha_0}\delta_{\theta_{jt}} + \frac{\alpha_0}{i-1+\alpha_0}G_0, \quad (5)$$

where θ_{jt} is originally drawn from H. More specifically,

$$\theta_{jt}|\theta_{11}, \theta_{12}, \ldots, \theta_{j,t-1}, \gamma, H \sim \sum_{k=1}^{K} \frac{m_{\cdot k}}{m_{\cdot\cdot}+\gamma}\delta_{\phi_k} + \frac{\gamma}{m_{\cdot\cdot}+\gamma}H, \quad (6)$$

where $\phi_k \sim H$. Again, we see that the probability of choosing a table is proportional to the number of customers sitting at that table.

Finally, the conditional distribution of x_{ji} given z_{ji} is

$$x_{ji}|z_{ji} \sim F(z_{ji}). \quad (7)$$

Thereby, we have completed the construction of the HDPMM. In summary, the mixands of the various data sets are shared across the data sets and are regulated by the same base distribution G_0.

3. MODELS FOR TIME-DEPENDENT CLUSTERING

In this section, we propose two types of Dirichlet process mixture models that allow for time-dependent clustering. The first is based on the CRPFC and the second on a hierarchical version of the CRPFC.

3.1. DPMMs for time-dependent clustering

Here we propose mixture models based on the CRPFC [5]. One can take advantage of the models to observe how the data vary across time. This is particularly useful when the timing of the clusters is critical.

First, we briefly review the concept of CRPFC. In the original setting of the CRP, a customer (a data sample) x_i comes to a restaurant with *infinite* number of tables, and is seated at table k with probability

$$P(z_i = k|z_1, z_2, \cdots, z_{i-1}) \propto \begin{cases} \frac{n_k}{i-1+\alpha}, & \text{if } k \text{ is occupied} \\ \frac{\alpha}{i-1+\alpha}, & \text{if } k \text{ is unoccupied} \end{cases}, \quad (8)$$

where n_k is the number of customers seated at table k. However, in our modified CRP, the capacity of the restaurant is limited to N only. That is, the restaurant has a finite number of tables, equal to N. Furthermore, we assume that the restaurant cannot serve more than N customers at the same time. In particular, we impose the restriction that before a new customer comes in the restaurant, the oldest customer must leave the restaurant. Clearly, for the first N customers, the seating probabilities remain the same as in the original CRP. However, for the customer $i > N$, the probabilities become

$$P(z_i = k|z_{i-N+1}, \cdots, z_{i-1}) \propto \begin{cases} \frac{n_k^*}{N-1+\alpha}, & \text{if } k \text{ is occupied} \\ \frac{\alpha}{N-1+\alpha}, & \text{if } k \text{ is unoccupied} \end{cases}, \quad (9)$$

where n_k^* is the number of customers currently seated at table k. When customer x_i enters the restaurant, only customers x_{i-N+1} to x_{i-1} are still in the restaurant. The earlier customers have already

left the restaurant. With this modification, we can readily show how one can have time-varying structures of clusters [5].

Now we extend this model to a mixture model. As before, each table corresponds to one mixture component, and each customer corresponds to a data sample. Thus, it is straightforward to create a generative model based on this idea. We simply add the generation step

$$x_i|z_i \sim F(z_i). \tag{10}$$

The inference proceeds as follows. Once the restaurant is full (meaning, it serves N customers), and a new customer is about to enter, the customer whose time at the restaurant expired (observation) leaves (the observation is removed from its cluster), and thus the posterior probability of each cluster is changed.

We proceed by way of example. Consider a multivariate Gaussian mixture model where the number of mixands of the mixture and/or its parameters vary with time. It is well known that the conjugate prior of multivariate Gaussian distributions with unknown mean and covariance is the normal-inverse-Wishart distribution. Thus, we choose the base distribution H of the DP to be the normal-inverse-Wishart distribution. When the number of customers is less than the restaurant capacity, the sampling scheme follows Algorithm 3 in [4] according to the following probabilities:

$$P(z_i = k|z_{-i}, x) = \begin{cases} b\frac{n_{-i,k}}{n-1+\alpha}\int P(x_i|\theta)[\prod_{j\neq i} P(x_j|\theta)]H(\theta)\mathrm{d}\theta, & \text{if occupied} \\ b\frac{\alpha}{n-1+\alpha}\int P(x_i|\theta)H(\theta)\mathrm{d}\theta, & \text{if unoccupied} \end{cases}, \tag{11}$$

where b is the normalizing factor, n is the number of customers in the restaurant, $n_{-i,k}$ is the number of customers at table k excluding customer i, z_{-i} are all the indices of the tables of all the customers in the restaurant except that of the ith customer, and similarly x are the first i observations. Here, we integrated out the parameters θ to make the algorithm more efficient. The integrals result in multivariate t-distributions that can readily be evaluated for a given x_i.

After the capacity is reached, that is when $i > N$, the sampling probabilities of the tables become

$$P(z_i = k|z_{-i}, x) = \begin{cases} b\frac{n^*_{-i,k}}{N-1+\alpha}\int P(x_i|\theta)[\prod_{\substack{j\neq i,\\ j\in J}} P(x_j|\theta)]H(\theta)\mathrm{d}\theta, & \text{if occupied} \\ b\frac{\alpha}{N-1+\alpha}\int P(x_i|\theta)H(\theta)\mathrm{d}\theta, & \text{if unoccupied} \end{cases}, \tag{12}$$

where $n^*_{-i,k}$ is the number of customers sitting at table k, excluding customer i, and J is the set of indices of the customers currently in the restaurant. Since the conjugate prior is used, the integral can be solved analytically. The marginal distributions $p(x_i|z_{-i}, x)$ are again multivariate t-distributions. The parameters of these distributions when we remove the "oldest" sample can readily be obtained with minimal computation.

3.2. Hierarchical DPMMs for time-dependent clustering

One can readily extend the CRPFC mixture model to a hierarchical structure where the capacity of a restaurant in a franchise is finite but the number of dishes remains unlimited. The mixture model based on hierarchical CRPFC is similar to the standard HDP mixture model. The idea of drawing samples from an HDP mixture model amounts to sampling the table assignment according to the distribution of customers in the local restaurant and choosing dishes from the hierarchical probability measure G_0 [11]. For hierarchical CRPFC mixture models, let x_{ji} still be the ith customer in restaurant j, and n_{jt} denote the number of customers in restaurant j seated at table t. The seating probabilities in restaurant j after the capacity is reached is given by

$$P(z_{ji} = t|z_{j,i-N+1:i-1}) \propto \begin{cases} \frac{n^*_{jt}}{N-1+\alpha} & \text{if } k \text{ is occupied} \\ \frac{\alpha}{N-1+\alpha} & \text{if } k \text{ is unoccupied} \end{cases} \tag{13}$$

Now we proceed with assigning dishes to each table. Since the distribution of dishes is generated from a standard Dirichlet process, the probability is not affected by the finite capacity limitation. Let m_k denote the number of tables serving dish ϕ_k in the entire franchise, and M denote the total number of different dishes served in all franchise. Then the dish probability is given by

$$P(\theta_{jt} = \phi_k|\theta_{11}, \cdots, \theta_{j,t-1}) \propto \begin{cases} \frac{m_k}{M+\gamma}, & \text{if } \phi_k \text{ is drawn already} \\ \frac{\gamma}{M+\gamma}, & \text{if } \phi_k \text{ is new} \end{cases} \tag{14}$$

where if ϕ_k is new, it is generated from H. Once the θ_{jt}s are drawn, we need to obtain x_{ji} from $F(\theta_{ji})$, which completes the generating process of the hierarchical CRPFC mixture model. The inference process can be deduced based on the generation process as was done in Section 3.1.

4. SIMULATIONS

In this section, we provide results of simulations of the proposed mixture models. First we show results of the CRPFC mixture model, and then the hierarchical CRPFC mixture model.

4.1. CRPFC Mixture Model

A time series of 400 data points, each generated from a two-dimensional Gaussian distribution, was classified by DPMM for temporal clustering. We set $N = 100$. The data sets were generated from a mixture of five multivariate Gaussian distributions with different means and an identical covariance matrix. The whole simulation process was carried out as follows:

1. Classification of the first 100 data samples using a standard DPMM.
2. Removal of the first sample in the current data window from the clusters inferred from the previous step, and recomputation of the parameters (mean and covariance) of the clusters.
3. Addition of the next data sample, and assignment of the cluster according to the mechanism described by the CRPFC mixture model.
4. Repetition of step (2) until all the data are processed.

In Fig. 1, we plotted all the samples in time series 2, with different symbols representing the clusters that the samples belong to. In Fig. 2, we plotted the classification of the data where different colors represent different clusters. The means of the clusters are represented with stars. Clearly, the method identified correctly both the number of clusters and the class of the specific observations. In Fig. 3, we display the time variation of the number of samples in a particular cluster. This cluster was only present among the first 300 samples. According to the simulation results, the CRPFC mixture model has satisfactory performance in classifying data and capturing the dynamics of each cluster.

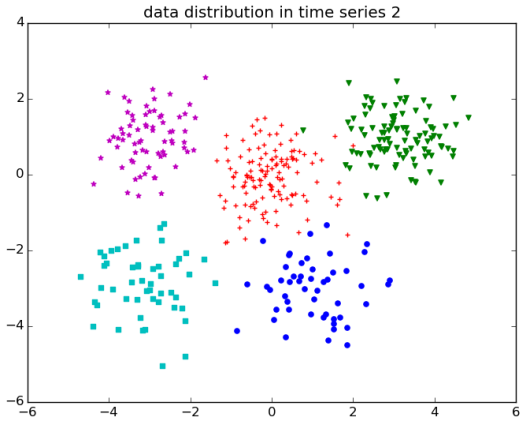


Fig. 1: Sample distribution in time series 2. Different symbols represent samples from differet clusters.

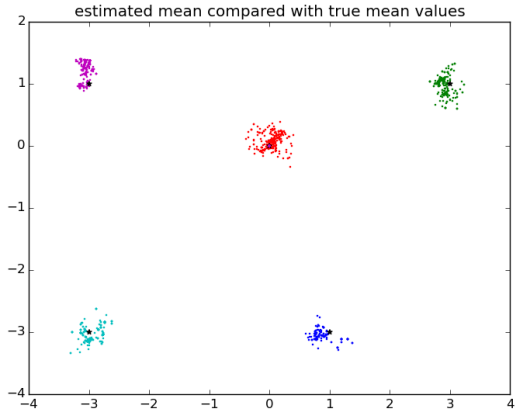


Fig. 2: Classified samples. Different colors represent different clusters, and the stars are located at the true true mean values.

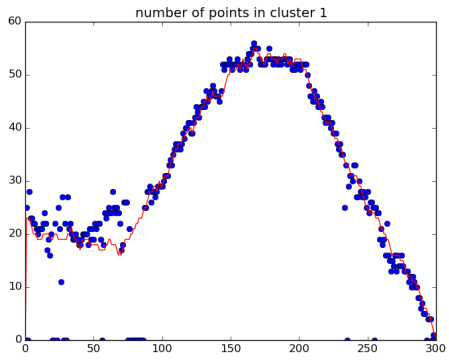
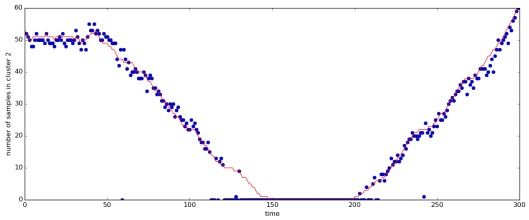


Fig. 3: Variation of the number of samples in a cluster with time. The horizontal axis represents time and the vertical axis is the number of data samples in the cluster. The red line is the true number of samples in the cluster.

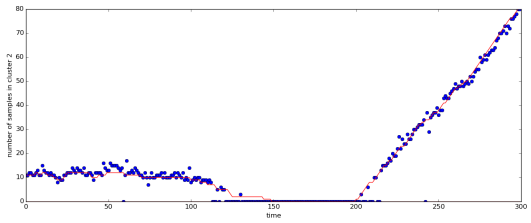
4.2. Hierarchical CPRFC Mixture Model

The same data generating process was used as in Section 3.2. We worked with three independent time series. These series shared clusters as described before. We used a hierarchical mixture model with the capacity of 100 of each restaurant to classify the data.

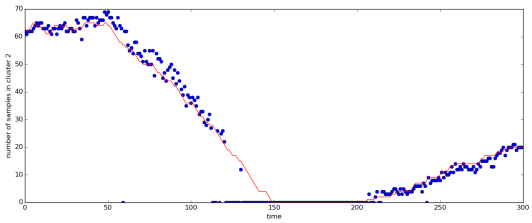
We present some results in Fig. 4. There we see the number of points of the same cluster in the different time series as they change with time. The red line is the true value, and the blue dots are the values inferred from the data. For each of the time series we only show results of one cluster due to lack of space. The results show good agreement between the true and estimated values of the cluster sizes.



(a) Times series 1



(b) Time series 2



(c) Time series 3

Fig. 4: Number of points of the same cluster in different time series as functions of time. The red line represents the true values, and the blue dots are the values inferred from data.

5. CONCLUSION

In this paper, we proposed Dirichlet process mixture models for time-dependent clustering. These models have a finite horizon of operation at any time which enables them to capture time varying structures. We explained how data are generated using these models, and how one can make inference from them. Our simulation results showed that the mixture models can accurately capture the evolution of clusters, namely their appearance, disappearance and reappearance. Furthermore, for the same purpose we also studied hierarchical mixture models for joint processing of data in a corpus of data.

6. REFERENCES

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

[2] T. K. Moon, "The expectation-maximization algorithm," *IEEE Signal Processing Magazine*, vol. 13(6), pp. 47–60, 1996.

[3] G. Claeskens and N. L. Hjort, *Model Selection and Model Averaging*, Cambridge: Cambridge University Press, 2008.

[4] R. Neal, "Markov chain sampling methods for Dirichlet process mixture models," *Journal of Computational and Graphical Statistics*, vol. 9(2), pp. 249-265, 2000.

[5] P. M. Djurić and K. Yu, "On generative models for sequential formation of clusters," *Proceedings of the European Signal Processing Conference*, Nice, France, 2015.

[6] K. Yu, J. G. Quirk, and P. M. Djurić "Fetal heart rate analysis by hierarchical Dirichlet process mixture models," *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, Shanghai, China, 2016.

[7] S. J. Gershman and D. M. Blei, "A tutorial on Bayesian nonparametric models," *Journal of Mathematical Psychology*, vol. 56(1), pp. 1-12, 2012.

[8] J. Sethuraman, "A constructive definition of Dirichlet priors," No. FSU-TR-M-843. Florida State University, 1991.

[9] C. E. Antoniak, "Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems," *The Annals of Statistics*, pp. 1152-1174, 1974.

[10] D. M. Blei and M. I. Jordan, "Variational inference for Dirichlet process mixtures," *Bayesian Analysis*, vol. 1(1), pp. 121-143, (2006).

[11] Y. W. Teh and M. I. Jordan, "Hierarchical nonparametric models with applications," *Bayesian Nonparametrics*, vol.1, 2010.

[12] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei, "Hierarchical Dirichlet processes," *Journal of the American Statistical Association*, vol. 101(476), 2006.