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, USA
{kezi.yu, petar.djuric}@stonybrook.edu

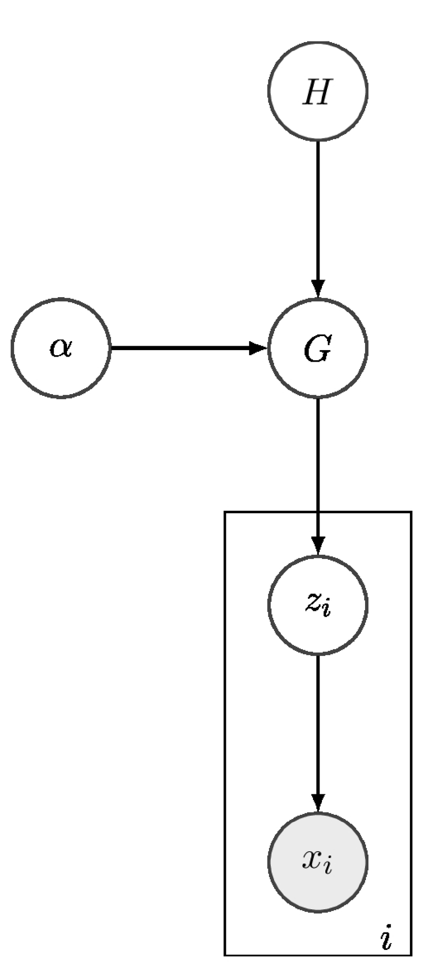
Stony Brook University

Summary

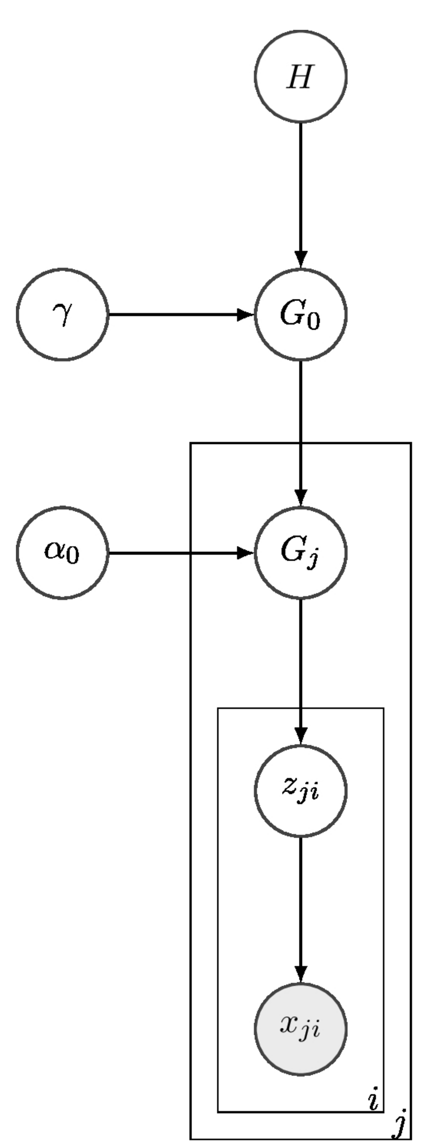
- New models for time-dependent clustering
- Theory based on Dirichlet process mixture models
- Methods capable for processing data sequentially
- Extensions to hierarchical models

Introduction

- In many cases of unsupervised learning tasks, the number of clusters is unknown beforehand.
- By assuming that the data are generated from a **Dirichlet process mixture model (DPMM)** , we can infer the number of clusters as well as their parameters from the data.
- In our work, we proposed a new mixture model based on a variation of the Dirichlet process [1], which is designed for **sequential learning** from data.
- The new model enforces a moving data window of fixed size and processes the data within the window sequentially.
- We extended the model to a **hierarchical model**, allowing for more flexibility.



(a) DPMM



(b) HPDMM

Figure: Graphical model representation of DP-based mixture models.

Background

- A popular approach to model data without prior knowledge of the number of clusters is based on **Dirichlet process (DP)** mixture models.
- In [1], a variation of DP, called **Chinese restaurant process with finite capacity (CRPFC)** was proposed to observe the dynamics of the data across time.
- The key modification is that the capacity of a restaurant is limited to a finite number N. After the capacity is reached, the probability of a customer x_i seated at table k is

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

where n_k^* is the number of customers currently seated at table k.

Models

1. Mixture models based on CRPFC:
   - We choose the emission distribution to be multi-variate Gaussian.
   - The sampling probability of tables after capacity is reached is

$$P(z_i = k | z_{-i}, x) = \begin{cases} b\frac{n^*_{-i,k}}{N-1+\alpha} \int P(x_i|\theta) \Big[\prod_{\substack{j \neq i, \\ j \in J}} P(x_j|\theta) \Big] H(\theta) \mathrm{d}\theta, & \text{if occupied} \\ b\frac{\alpha}{N-1+\alpha} \int P(x_i|\theta) H(\theta) \mathrm{d}\theta, & \text{if unoccupied} \end{cases} \qquad (2)$$

2. Hierarchical mixture models:
   - We use metaphors similar to the Chinese restaurant franchise (CRF) process.
   - The seating probability in restaurant j after the capacity is reached is

$$P(z_{ji} = t | z_{j,i-N+1:i-1}) \propto \begin{cases} \frac{n^*_{jt}}{N-1+\alpha} & \text{if } k \text{ is occupied} \\ \frac{\alpha}{N-1+\alpha} & \text{if } k \text{ is unoccupied} \end{cases} \qquad (3)$$

   - The dish probability remains the same as the standard HDP mixture models [2]

$$P(\theta_{jt} = \phi_k | \theta_{11}, \cdots, \theta_{j,t-1}) \propto \begin{cases} \frac{m_k}{M+\gamma}, & \text{if } \phi_k \text{ is drawn already} \\ \frac{\gamma}{M+\gamma}, & \text{if } \phi_k \text{ is new} \end{cases} \qquad (4)$$

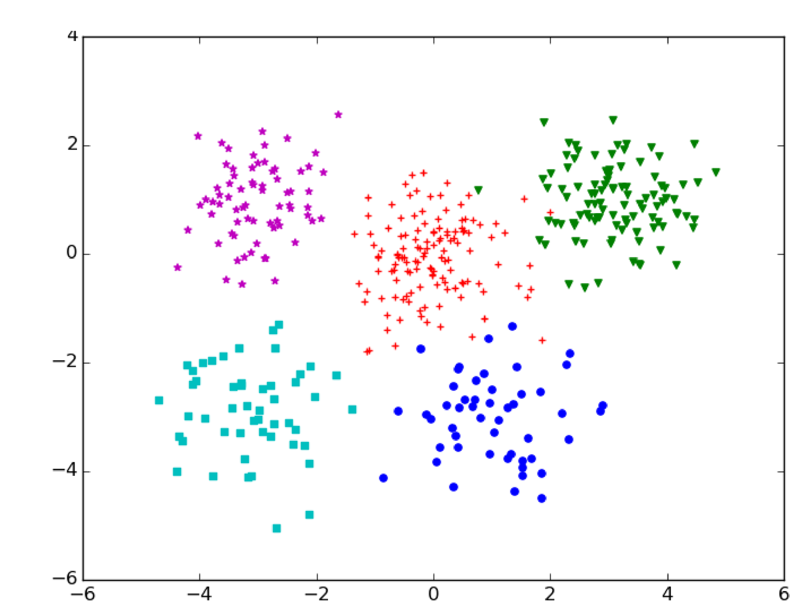
References

[1] P. M. Djurić and K. Yu.
On generative models for sequential formation of clusters.
In *Signal Processing Conference (EUSIPCO), 2015 23rd European*, pages 2786–2790. IEEE, 2015.

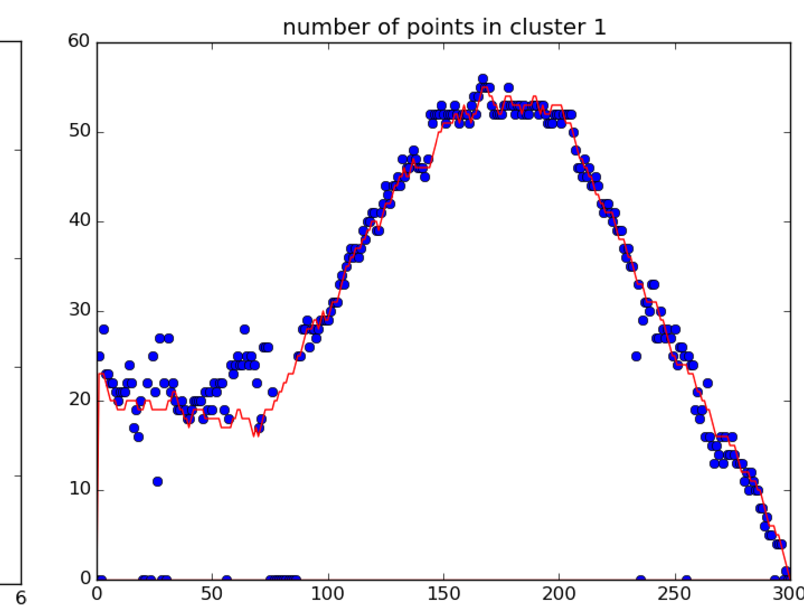
[2] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei.
Hierarchical dirichlet processes.
Journal of the american statistical association, 2012.

Results

1. Simulation results of CRPFC mixture models:

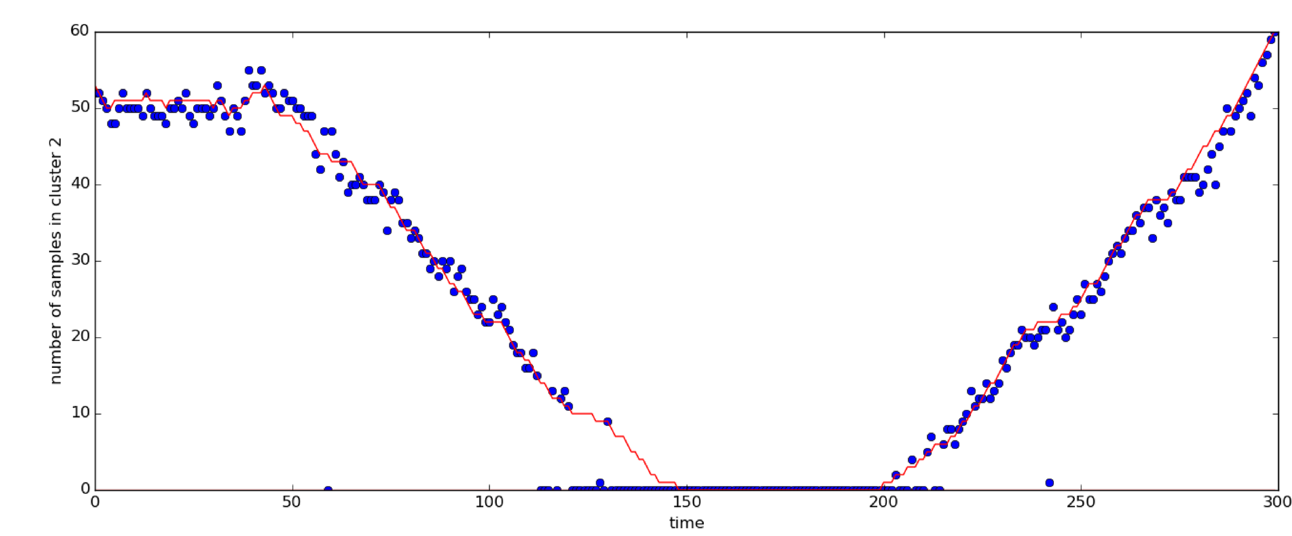


(a) Simulated data distribution. Different symbols represent samples from different clusters.

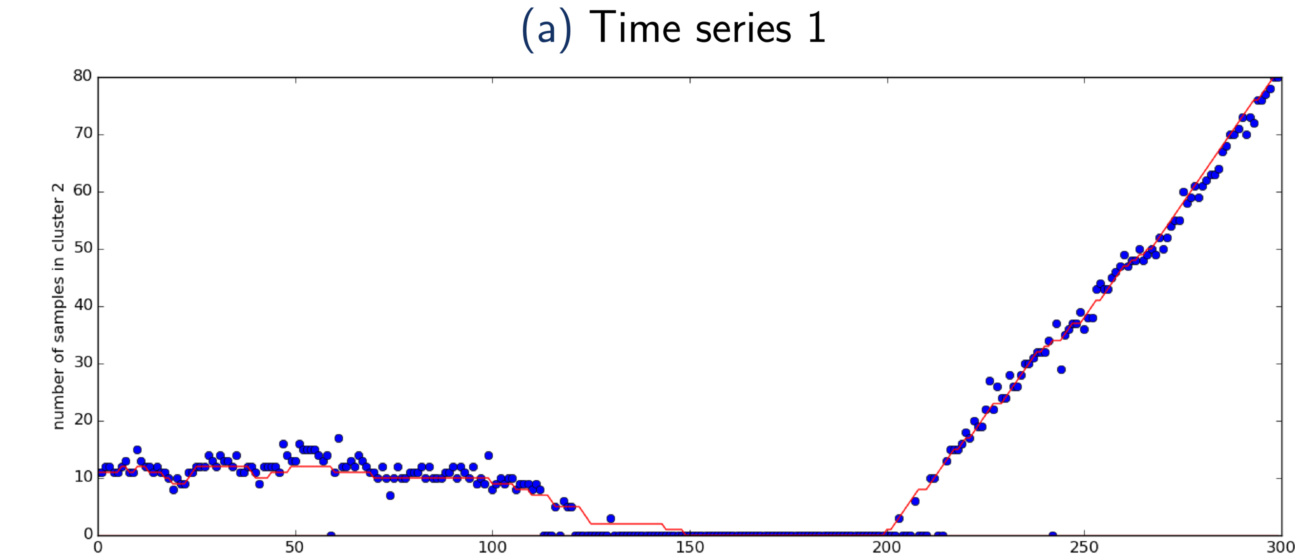


(b) Variation of the number of samples in a cluster with time.

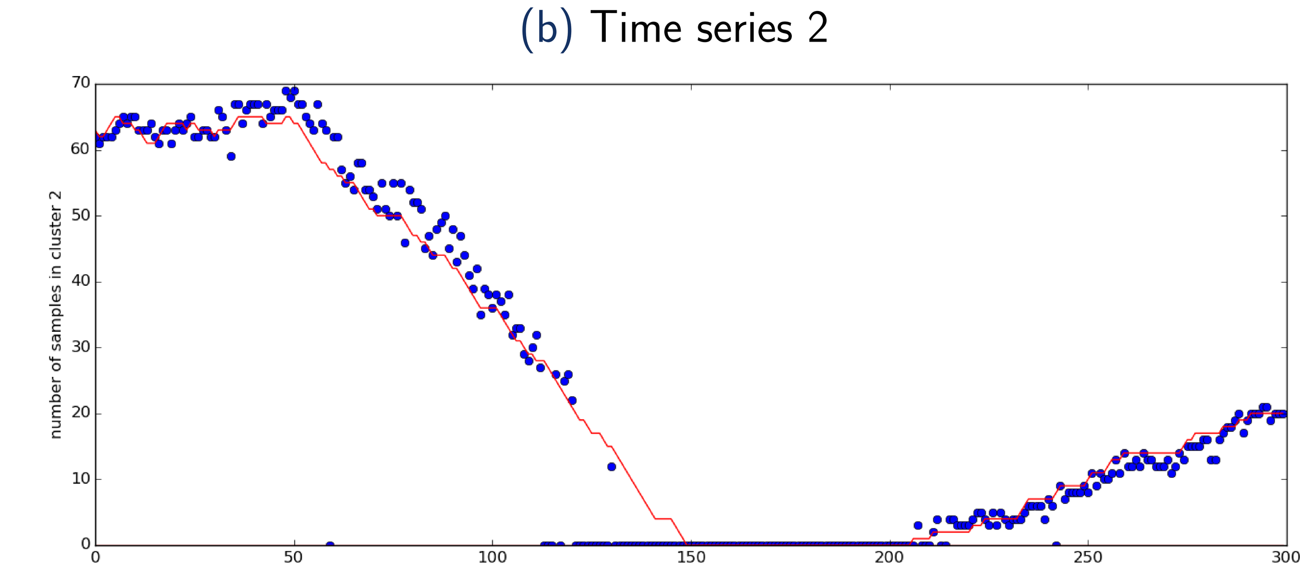
2. Simulation results of hierarchical CRPFC mixture models:



(a) Time series 1



(b) Time series 2



(c) Time series 3

Figure: 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.