**ORIE 6741 Midterm Report Outline**

Due: 11/21/2017

This should be a 4-5 page report, which serves as a check-point. It should have roughly the same sections as your final report (introduction, related work, methodology, experiments, discussion), where some sections or parts of sections may be in progress. The introduction and related work sections should be well developed.

The grading scheme for the midterm report is: 60% for the method, 35% for the design of experiments and current progress, 5% for the planned activities.

README: This is meant to be a rough approximation of the final draft. The main points and structure should be outlined. Relevant equations and their references should be included. Any figures should also be noted/included.

**Abstract**

Concise summary of the paper: main points to cover; (1) motivation, (2) methodology (this should include some contribution/insight), (3) current experiments and results, (4) future planned experiments and expected results.

**Introduction**

1. Motivation: Time-series data is difficult to model. They are non-stationary and difficult to predict. However, there are many benefits to being able to predict them (finance, earthquakes, weather, etc). There are benefits to capturing *characteristic segments* from the time-series data that may correlate with *change points*. By modeling the model with bayesian nonparametrics, we allow the complexity of the model to grow with the data available which is intuitively appropriate given the non-stationary nature of the time-series data. A predictive model that can identify latent patterns within the data can be a powerful tool for understanding the data (for example, in weather patterns) and also potentially for online prediction (for example, stock prices). The model should be easily interpretable. The model should be updated online as more data is collected as we would like nonparametric models to do.
2. Contribution: We propose an IBP clustering process to cluster training time-series data. This allows for flexibility of learning multiple latent features rather than traditional clustering which rigidly assigns data points to single clusters. This allows for more descriptive clusters. We then propose gaussian process (GP) with spectral mixture (SM) kernel to learn covariance of the clusters. SM allows for more expressive correlations to be discovered by modeling the fourier transform of the kernel as a gaussian mixture. We can also gain additional insight by looking at the spectral density of the kernels. We then propose a method for online clustering (by using chi squared goodness of fit testing) and learning which allows our model to grow with more added data. Short term online inference is also possible with our model.

**Related Work**

1. Modeling time-series with combination of nonparametrics and dynamics models:

* These models can be restrictive and do not have the inherent flexibility of purely nonparametric models. (see proposal)

2. Purely Non-parametric models:

* These models have more flexibility and can grow in complexity as more observations are made. Generally cluster multiple time-series trajectories but not find characteristic segments (section clusters) which may give insight on the non-stationarity of the time-series. (see proposal)

3. IBP for clustering in general:

* IBP has been able to discover latent features in applications such as image processing. It has also been used for clustering problems. We believe we can leverage the flexibility of latent features to clustering time-series trajectories and learn latent features of clusters.

4. IBP for discovering Kernel Structure:

* They predefine kernels to learn unique compositions of kernels. Only allows for addition of kernels, and previous kernels that are defined as latent features. This does not really allow for super expressive correlations. (see proposal)

5. SM kernels (GP):

* Previous success of using SM kernels for extrapolation. This does not require hand crafted kernels. Moving more towards automatic kernel structure discovery and can theoretically model any stationary kernel. (AG Wilson)

6. Chi-Squared tests:

* Chi-squared tests to evaluate filter health for tracking applications. (Bar-Shalom)

**Methodology**

1. Clustering

* Constructing IBP prior & Interpretation

Latent feature modelling concern with finding the hidden structure of data.In short, given data set (X), we can decompose it to a binary matrix (Z) and a weight matrix(A).Each row of the matrix Z represent object while each column represent the latent feature.One way to interpret the Z matrix is that whether each object possess the latent feature or not. Indian buffet process is a nonparametric method which give a prior over binary matrix.[Ref] give a close form of the IBP process. In traditional clustering method, one data point can only belong to one cluster which shown in the pic below.A drawback is that the number of cluster (K) need to be predefined.However, in nonparametric model, the number of cluster grow as the size of data increase.

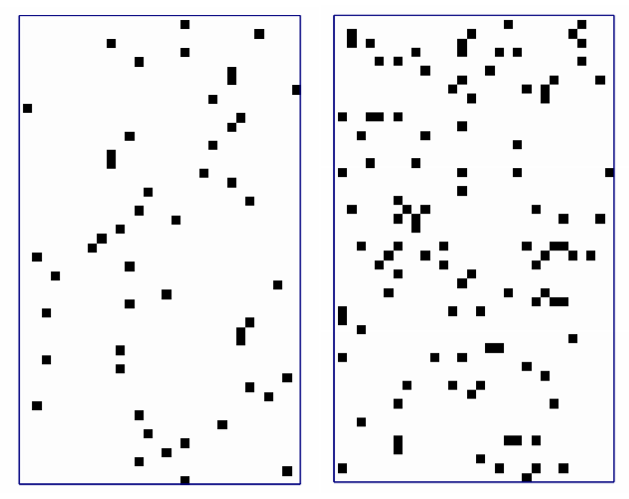
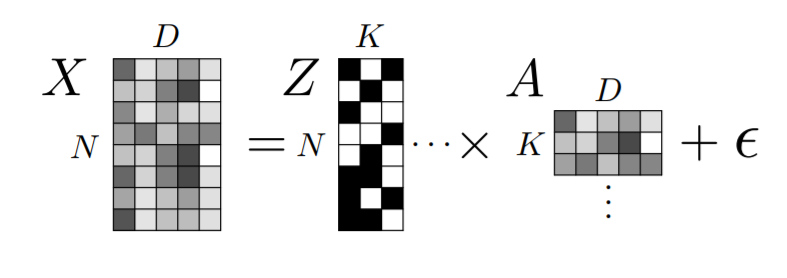


Figure: (Left) traditional clustering method assigned each object to one cluster. (IBP) one object can belong to multiple cluster (From The Indian Buffet Process and Extensions Zoubin Ghahramani 2009) -- To do = Recreate from Motorcycle data (Kmean vs IBP)

* Inference IBP

In this project, we will consider Gaussian latent feature modelling as shown in the figure below.



The goal is to find the posterior of Z given data (X) with N rows and D dimensions .A weight matrix (A)is assumed to be from Gaussian distribution. In the original paper of IBP [Ref], inference is done via Gibbs sampling since we can calculate the likelihood and conditional probability of Z .Metropolis Hasting is then used to find the noise parameters of X and A. From our experiment, this method is computationally expensive and a good initialization is crucial.

* Fast Inference for IBP using mean variational method

[Ref] provide a fast implementation of Indian buffet process by using mean variational method. Also, restart parameter is introduced to avoid bad initialization. From our experiment this implementation is significantly faster than the first version. (**To add more about mean variational method)**

1. Stealing Code and online inference/clustering

* We learn a different SM kernel with GP for each cluster. This allows for more expressive pattern discovery. Due to the non-stationary nature of the data it makes sense to learn a different correlation structure for each cluster. Given new online data we can perform chi-squared goodness of fit tests given the predictive distributions over the outputs defined by the hyperparameters of the kernels fit to each cluster. We can use this to determine in a probabilistic principled manner whether the new data belongs in our clusters or not. Another approach could potentially be to compare the spectral density of the new data with the spectral density of the learned kernels to determine if they should be long to the same pattern.

**Experiments**

1. Clustering non-stationary data

* We used motorcycle data set which is commonly used for non-stationary data [Ref]. By using IBP, the result in the figure **(to add figure)** show the latent features for each data point which help us finding cluster of the data. We also repeat the same experiment where cluster is separated. Traditional method like k-mean clustering will find these data points as different cluster **(to add figure)** However, from the experiment IBP process can find that these data point govern the same latent features and belong to the same cluster. As a result, this potentially help us fitting the model better.

2. SM kernel learning and chi-squared:

* Implement SM kernel over patterns and see the fit/extrapolation
* Use test data. Perform chi-squared test over the test data and compare with current clusters. See how they fit. Also potentially compare spectral density of new data with old clusters and see fits.

3. Metric of evaluation of final results

* Some quantification of the experimental results in the final project.

**Discussion**

1. Talk about novel method, contribution, current experiments.
2. How do we expect the pieces to fit together, what the final project will look like.
3. Future experiments that we will not have conducted by midterm report due date.

**References**