6.435 Final Report: Temporal Hierarchical Dirichlet Process

Christien Williams, Shen (Sean) Chen

May 2020

1 Introduction

6.435 began by diving into Latent Dirichlet Allocation, a method with which to infer the topics of a document in a corporus. Each topic stood as a distribution over words in a standard vocabulary for the distribution. One limitation of the LDA model is that one must assume the number of topics which are pervasive in the corporus. In the Hierarchical Dirichlet Process (HDP), one can also separate a dataset at hand into groups and then subgroups while retaining links of "statistical strength" between the groups. The said groups can be looked at as the documents in a corporus, and the linked subgroups as topics shared between documents. However, with HDP, a parameter of the number of subgroups doesn't need to be assumed. HDP can then be extended to encompass a temporal analysis, wherein this subdivision of groups can be looked at as potentially evolving over time. Autoregressive/Temporal Hierarchical Dirichlet Mixture models (THDPs) are a model targeted at this extended framework. This project seeks to learn about sequential discrete structures using THDPs. The intermediate milestone is to replicate the experiment of Hierarchical Dirichlet Process, using different MCMC methods to approximate posteriors and compare the performance. Then, this project seeks to emulate the THDPs, ultimately building a model pipeline that can extend non-parametric topic modelling to online learning/streaming data, which could lead to shift of topics or clusters within the data.

The primary goals of this paper are to explain the benefit of time varying Dirichlet Process (DP) models. Along the way, we explain how we built HDP and THDP generative models, and run experiments showing how models can infer distributions assumed to be generated by Dirichlet Processes.

2 Context

2.1 Hierarchical Dirichlet Processes

A Dirichlet Process is a measure on measures. Given a base distribution and a parameter alpha, a DP generates a distribution from the support of the base distribution and with weights defined by the Griffiths-Engen-McCloskey (GEM) distribution. An HDP extends this formulation beyond that of a single level. With an HDP, you can model the distribution of parameters, or latent structure, over a set of groups, where the said parameters are shared/linked between groups. Ultimately:

$$G_0|\gamma, H \sim DP(\gamma, H)$$
$$G_j|\alpha_0, G_0 \sim DP(\alpha_0, G_0)$$

2.2 Non-Parametric Models

Bayesian non-parametric (BNP) models are characterized by their dynamic allocation of parameters. Instead of predefining the number of parameters in the model, and ultimately statically defining the complexity of the model, BNP models purport that the number of parameters can grow in response to the data. Using the example of clustering, a BNP prior permits data to be generated by a continually growing number of clusters. Despite BNP's flexibility, it can fail to capture temporal structure. Consider the context of a stream of data from documents in a corpus over times. If documents are updated over time, the previous belief about the topic distribution in the topic during a previous epoch should be used when updating beliefs. The document in a new epoch shouldn't be treated as a brand new document. This is where Temporal Hierarchical Dirichlet Processes (THDPs) enter the picture.

2.3 Temporal Hierarchical Dirichlet Process

Temporal Hierarchical Dirichlet Processes assume that topic groups/documents/time-dependent clusters are organized across epochs. As a BNP, parameters are unbounded in each epoch, however they may be born and die over time. The actual "parameterization of each component can also evolve over time in a Markovian fashion". The Bayesian dependencies have been shown below [1].

$$G_t|\phi_{1:k}, G_0, \alpha \sim DP\left(\alpha + \sum_k m'_{kt}, \sum_k \frac{m'_{k,t}}{\sum_l m'_{lt} + \alpha}\delta(\phi_k) + \frac{\alpha}{\sum_l m'_{lt} + \alpha}G_0\right)$$

(a) Base function G_t over time t

$$\theta_{tdi}|\theta_{td,1:i-1}, \alpha, \boldsymbol{\psi}_{t-\Delta:t} \sim \sum_{b=1}^{b=B_{td}} \frac{n_{tdb}}{i-1+\alpha}\delta_{\psi_{tdb}} + \frac{\alpha}{i-1+\alpha}\delta_{\psi_{tdb^{\text{new}}}}$$

(b) Params θ over time t, doc d, cluster i

3 Generative Models

3.1 Univariate HDP Generative Model

We started by building a univariate HDP [2] (uHDP) model. Starting here gave us the necessary intuition to then build multivariate and then temporal generative models. In the uHDP, we truncated the number of clusters, k, at 100. For the DP visualization over a finite set, we use a lower tructation number on the clusters for the lower alphas. This is because the majority of the data samples will be concentrated at the first few clusters, hence having a small truncation number k will be sufficient to account for most of the weights in the data.

3.2 Multivariate HDP Generative Model

To build a multivariate HDP generative model, we truncated the number of clusters k at the global topic level to 100. We experimented with different alphas: 1000, 100, 10 and we set the mean to be [0.5, -0.2], and the covariance to be [[1.0, 0.0], [0.0, 1.0]]. We used Griffiths-Engen-McCloskey (GEM) distribution to build our stick breaking process. We know from discussions in class, the larger the second parameter in the beta distribution, the more weight is placed on the possibility of the 2nd value (i.e. the totality of components from i to infinity); thinking of this from the Chinese Restaurant Process (CRP) perspective, there's a higher probability that each successive data point sits at a new table/starts a new cluster; initial atoms thus have smaller weights; on the other hand, with smaller alpha, initial atoms have much larger weight; we see this in action as we vary the betas; with alpha 10, there are fewer atoms with large weight (scaling up to .14), whereas with alpha 1000, there are a lot of atoms close to the maximum weight (scaling to .005)

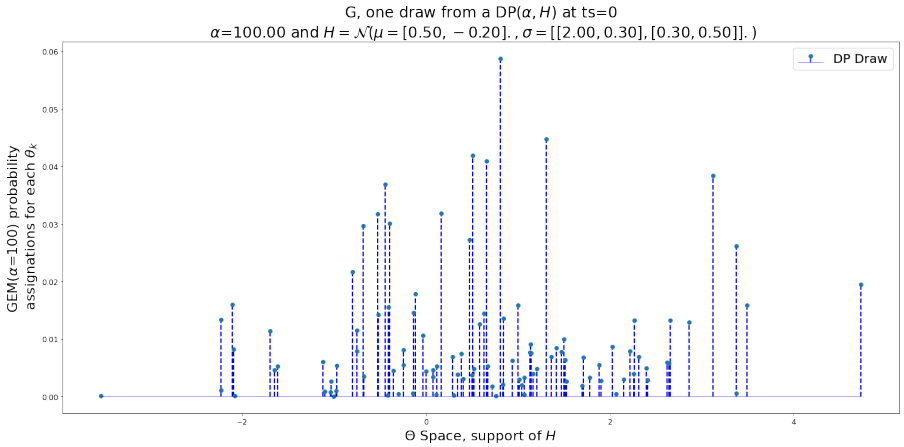
3.3 Multivariate THDP Generative Model

Following the methodology from Ahmed, Amr Xing, Eric. (2012). *'Timeline: A Dynamic Hierarchical Dirichlet Process Model for Recovering Birth/Death and Evolution of Topics in Text Stream.'* [1], we replicated and implemented a Temporal Dirichlet Process Model (TDPM) at the base layer and ultimately an Infinite Temporal Dirichlet Process Model (iTDPM).

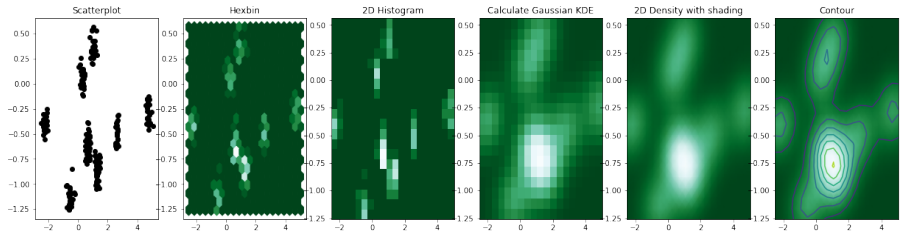
The TDPM begins as follows: The random measure G is time-varying, and the process stipulates that $\Phi_{1:k}$ are the mixture components available in the previous Δ epochs. To put more clearly, $\Phi_{1:k}$ is the collection of unique values of the parameters $\theta_{t:t\Delta}$, where θ_{tn} is the parameter associated with data point x_{tn}. Ahmed, Amr Xing, Eric. (2012) sets m_{kt} to denote the number of "parameters" (atoms) in epoch t associated with component k, then m_{kt}, the prior weight of component k at epoch t is defined as:

This defines a "Δ order process" where the strength of dependencies between epochs is specified with Δ, λ. Ahmed and Xing (2008) showed that these hyper-parameters allow the TDPM to diverge to either a set of independent dirichlet processes at each epoch when is set to Δ=0, and to a global dirichlet process, when the parameters are set to Δ = T and λ = inf. Expanding next to the iTDPM, using the topic modeling analogy, the documents in epoch t are modeled using an epoch specific HDP with high-level base measure denoted as G_{t0}. The epoch-specific base measures G_{t0} are related via the TDPM process of (Ahmed and Xing, 2008). When you integrate the random measures, the result is a recurrent Chinese restaurant franchise process (RCRF).

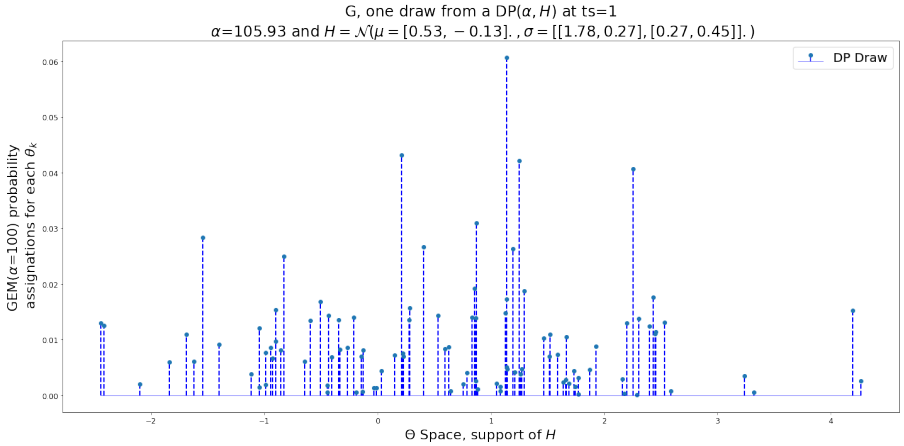
Examples across 3 time steps of the simulated data from THDP generative model with initial assumptions of parameters being the same as Section 3.2 are listed below.



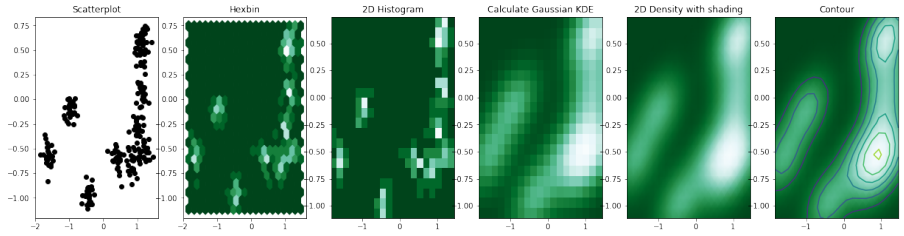
(a) Base function G_0: time step 0



(a) THDP Simulation: time step 0

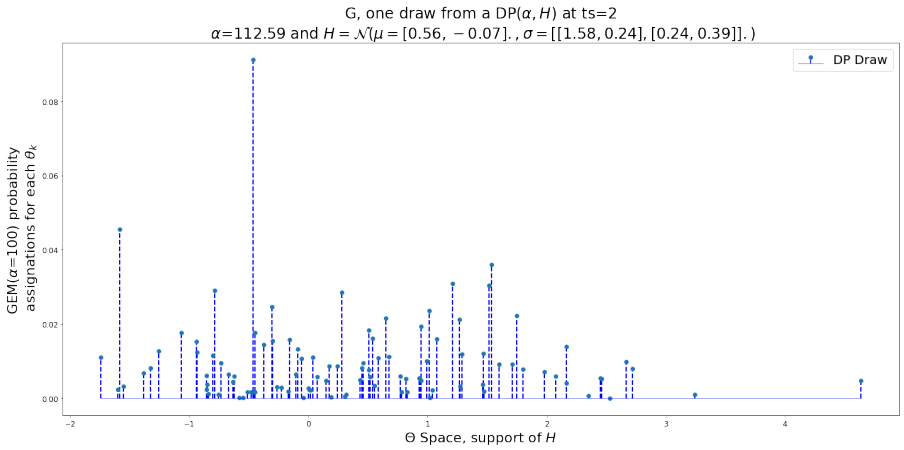


(a) Base function G_0: time step 1

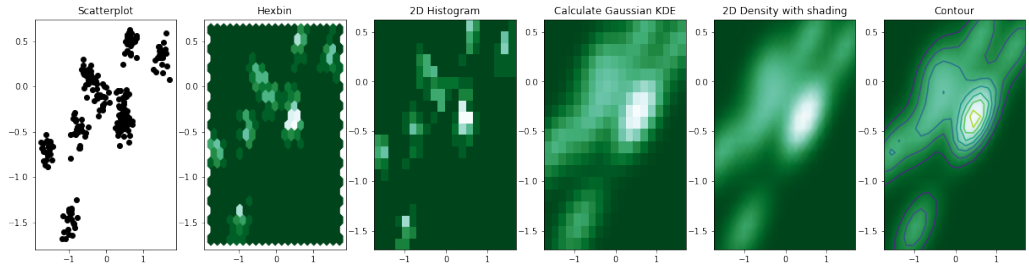


(a) THDP Simulation: time step 1

[section][page=yes]



(a) Base function G_0: time step 2

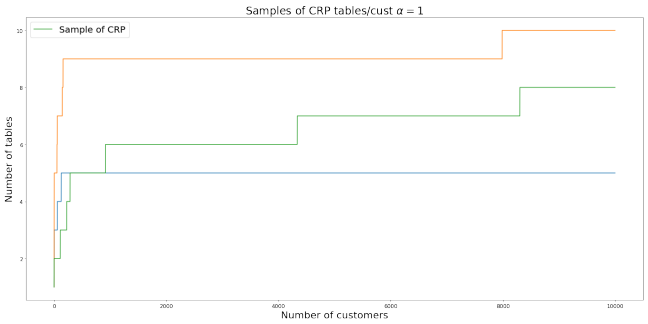


(a) THDP Simulation: time step 2

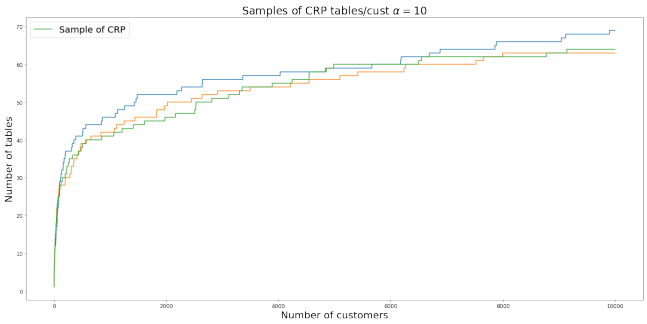
4 Inference Models

4.1 Chinese Restaurant Process

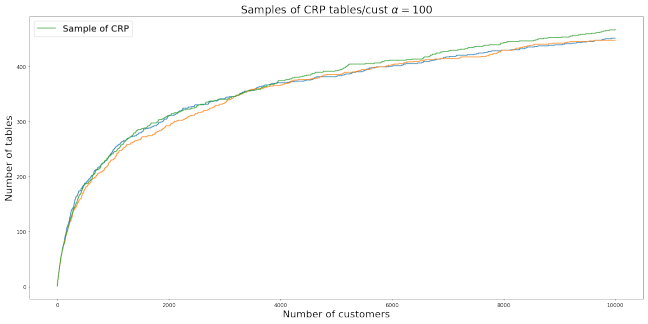
We replicated the Chinese Restaurant Process and visualized the effect of alpha on the number of tables for a given number of customers. Tested CRP on alpha equal to 1, 10, and 100 and visualized the aggregated proportion of data points assigned to each of the clusters/tables.



(a) CRP $\alpha = 1$



(b) CRP $\alpha = 10$



(c) CRP $\alpha = 100$

Figure 8: Chinese Restaurant Process with three differnet αs

4.2 Hierarchical Dirichlet Process (HDP) and the temporal extension (THDP)

To build a Bayesian non-parametric clustering pipeline, we used PyMC3 to replicate the DP and subsequently HDP models, which were tested on clustering the sample data simulated from the generative models mentioned in Section3. We set up MCMC methods for approximations of the model parameters, e.g. NUTS and Metropolis-Hastings. We also applied inference method ADVI to approximate the parameters, but due to the fact that HDP is a model mixed with continuous and discrete variable dependencies, ADVI failed.

We then modified the topic-level base functions G_0 for the HDP model to be varying over time, hence a temporal version of HDP, according to the paper by Amr and Eric [1].

5 Experiment Results

5.1 HDP

To test the replicated HDP Mixture Model mentioned in Section 4.2, we have simulated sample data using the generative model mentioned in Section 3.2, with the mean set to be [0.5, -0.2] and the covariance function being an identity matrix. We also attempted to set up priors over the covariance function, using LKJ Cholesky, but the computational time increased exponentially, hence this would not be considered during our testing mode of the model.

The distribution of a 2-D sample simulation data from HDP generative models could be seen as below. The mode for both dimensions have well reflected the mean assumed in the base function G_0.

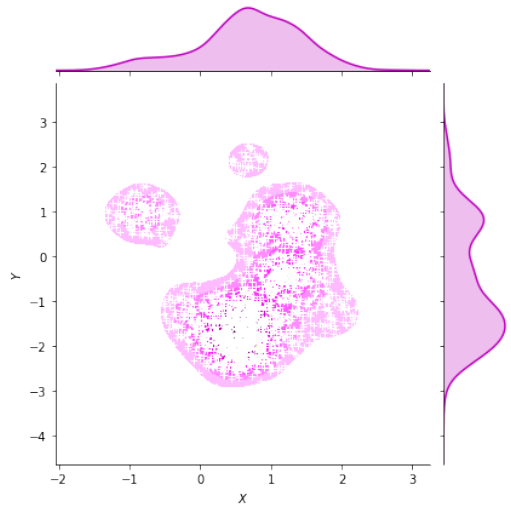
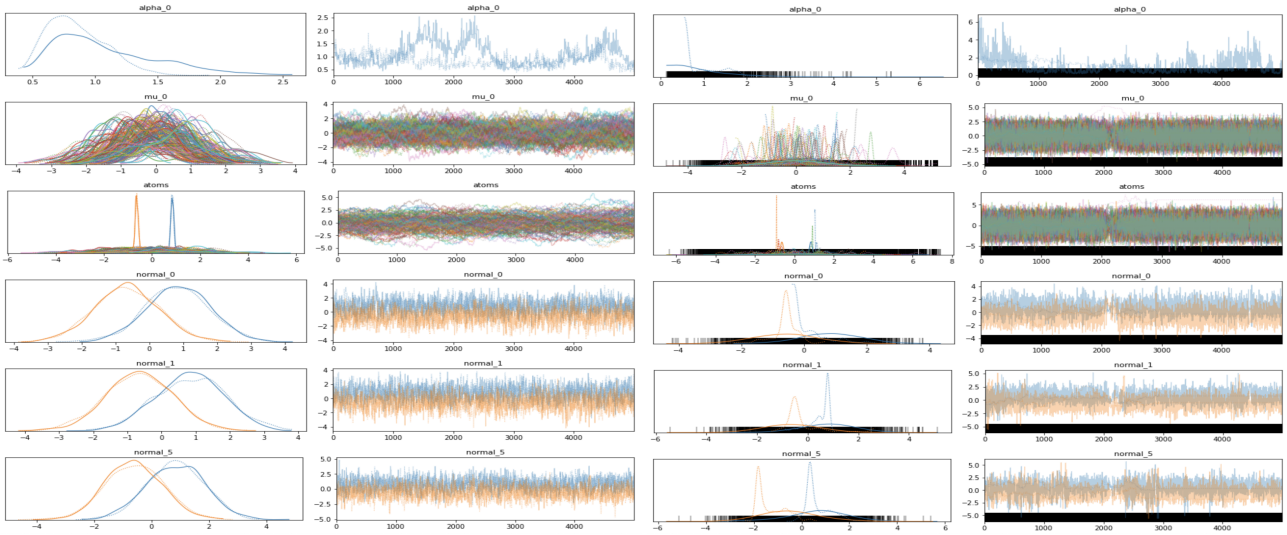


Figure 9: Example Simulation from HDP

The MCMC sampling results for the 2-D simulated data, with 10 topics assumed globally, could be seen below. There are four inference results in total. Each plot has two columns, with the left being the distribution of the inferred parameters (from top to bottom: α_0, μ_0 for base function G_0, atoms and three document level μs) and the right being the traceplot for each parameter across the whole markov chain. We have set the sample size for each chain to be 5000 (S=5k), with K and M being the truncated numbers for the topics globally (K) and the number of topics for each document (M). For the case of assuming 50 topics globally and 20 topics at document-level, we could see that the Metropolis-Hastings method has returned a set of much stabler results compared to its Nuts counterpart. The two chains (solid and dotted lines in the distribution plots) for Metropolis-Hastings are also more consistent and aligned compared to Nuts.

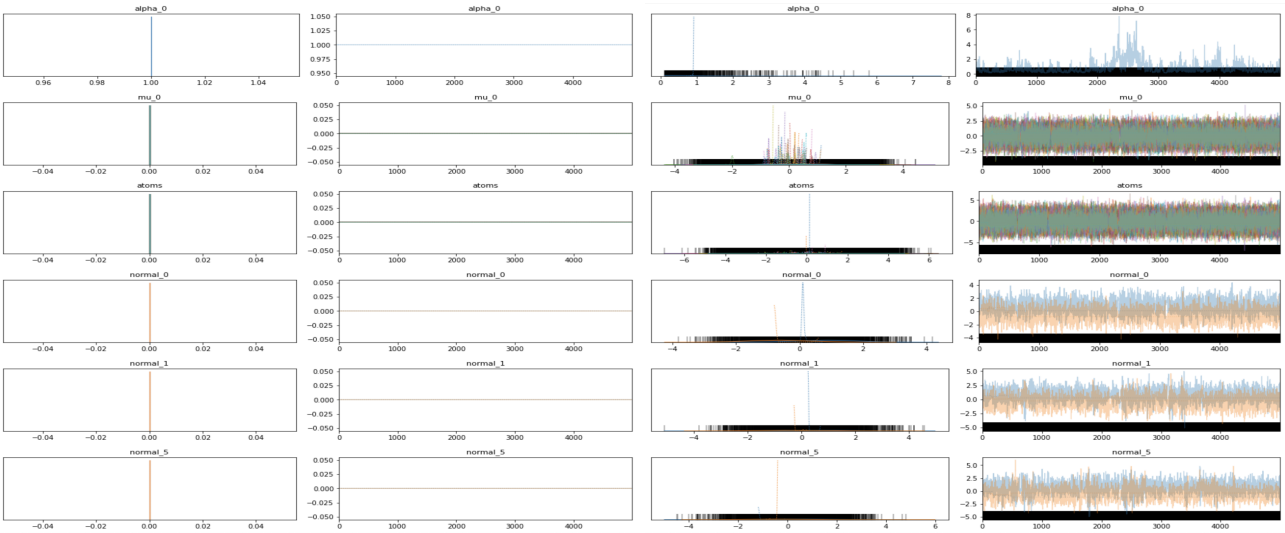
If we truncate the number of topics to K=25 globally and M=15 at document level, the results for Metropolis-Hastings will end up failing. Though Nuts did not return as good a result as in the previous case, it can at least still show that the means for the normal base functions have one dimension being on the negative side and the other being positive. The results show that Nuts is more robust to the truncation of the number of topics compared to Metropolis.



(a) Metropolis: S = 5k, K = 50, M = 20

(b) Nuts: S = 5k, K = 50, M = 20

Figure 10: MCMC Experiments Results for HDP Model Replication



(a) Metropolis: S = 5k, K = 25, M = 15

(b) Nuts: S = 5k, K = 25, M = 15

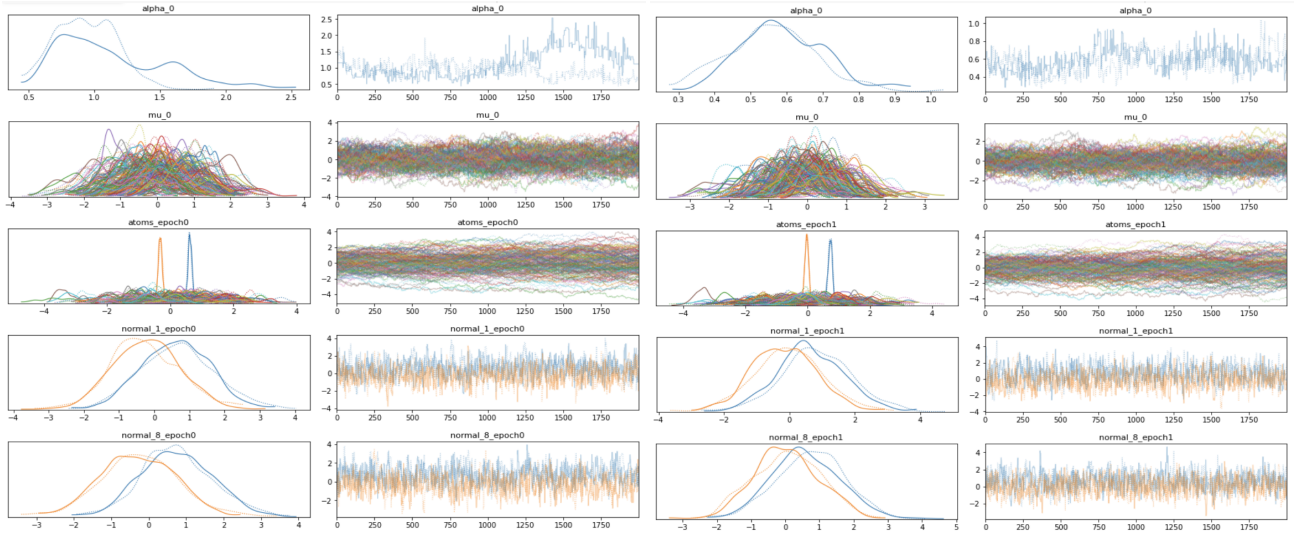
Figure 11: MCMC Experiments Results for HDP Model Replication

5.2 THDP

We tested the temporal HDP model over a simulated time-dependent data from the THDP generative model mentioned in Section 3.3. The following plots have shown the inferred results for two documents across two time steps. The dependencies of base functions G_0 and parameters over time are according to the equations shown in Section 2.3.

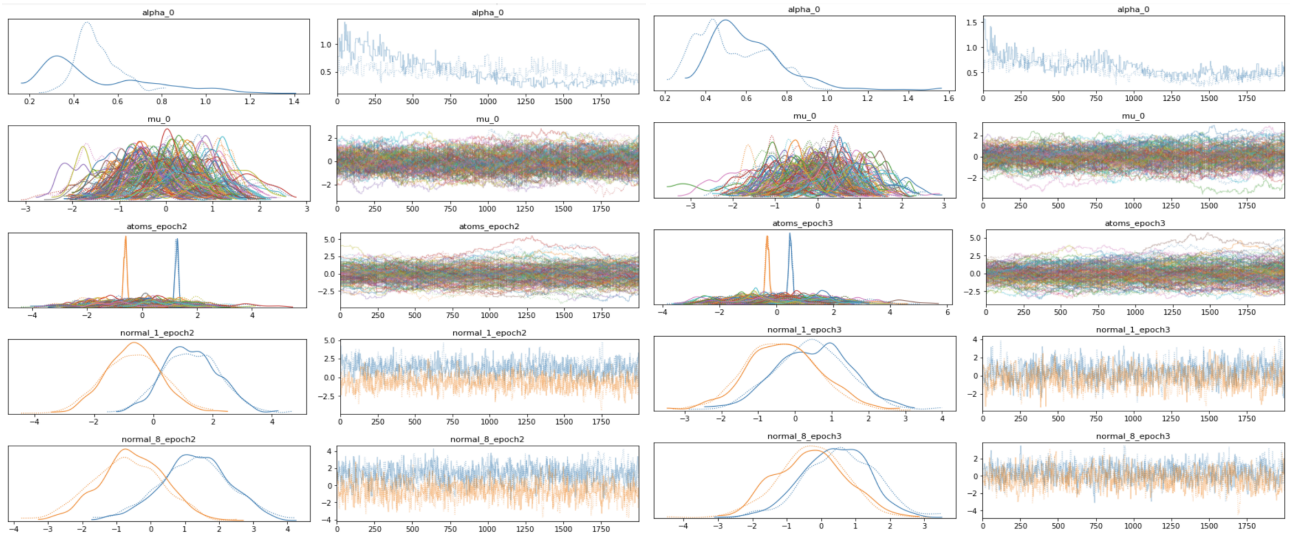
Due to the limitation of computational power, we did not manage to train the THDP model on the whole simulated dataset from THDP generative model, which contains 10 documents and 10 time steps. However, we can still observe from the results for atoms parameter (The third row for each of the distribution plots). The shift of mus for both of the 2-dimensional means to the negative side is obvious, which is what was defined for the parameters in the generative model. we have not tested but are optimistic about the performance of our replicated THDP model in terms of capturing the time-varying features of the clusters after it is being run on a much larger dataset.

[section][page=yes]



(a) document 0 at time 0 (b) document 1 at time 0

Figure 12: Inference results for two documents in a corpus at time 0



(a) document 0 at time 1 (b) document 1 at time 1

Figure 13: Inference results for two documents in a corpus at time 1

6 Conclusions and Future Steps

We have succeeded in replicating the HDP and Temporal HDP model with MCMC inferences and have built a set of generative models for both HDP and THDP for simulation purposes. The models have shown promising results for correctly inferring the means assumed to generate the simulated data.

The next steps include the following:

Firstly, we need to include priors for the variance functions and allow the covariance to be shifting over time just like the means. This is because to capture the time-varying pattern of the clusters, it is important to capture their variability over time.

Secondly, we need to test the THDP model on text data extensively. We intended to test the models utilizing Word2Vec Embeddings to transform text data into numerical vectors with predefined dimensions. We have set up a pipeline that fits the HDP on the Word2Vec embedded text data. Due to time limit, we did not end up testing our THDP model using Word2Vec embeddings.

References

[1] Eric P. Xing Amr Ahmed. Timeline: A dynamic hierarchical dirichlet process model for recovering birth/death and evolution of topics in text stream. *UAI*, 2010.

[2] Matthew J. Beal Yee Whye Teh, Michael I. Jordan and David M. Blei. Hierarchical dirichlet process.