Research Meeting Notes

Xi Tan (tan19@purdue.edu)

May 18, 2017

1 Agenda

1. Synthetic Dataset Results.
2. Next step: TBD

2 Generative Process

$$\pi|\alpha \sim nCRP(\alpha) \tag{1}$$

$$\lambda_{uv}(t) = \frac{1}{n_p n_q}\gamma_{pq} + \int_{-\infty}^{t} \beta_{uv}(f(\mathcal{T}_s))e^{-\frac{t-s}{\tau_{uv}}} dN_{vu}(s) \tag{2}$$

$$\lambda_{pq}(t) = \sum_{p=\pi(u),q=\pi(v)} \lambda_{uv}(t) \tag{3}$$

$$M_{new} = \begin{cases} t_{new} \sim HawkesProcess(\lambda_{root}(\cdot)) \\ Z_{u\in S_{new}} \sim Ber\left(\frac{\bar{\lambda}_{u\cdot}(t_{new})}{\sum_u \bar{\lambda}_{u\cdot}(t_{new})}\right) \\ Z_{v\in R_{new}|S_{new}} \sim Ber\left(\frac{\bar{\lambda}_{\cdot v|S}(t_{new})}{\sum_v \bar{\lambda}_{\cdot v|S}(t_{new})}\right) \\ \mathcal{T}_{new} \sim Multinomial(\theta_{S_{new},R_{new}}) \end{cases} \tag{4}$$

where nCRP is the nested Chinese Restaurant Process, and $\beta_{uv}(f(\mathcal{T}_i)) \sim \exp(\mathcal{GP}(0, \kappa_{uv}))$.

$$\theta_{S,R} = \frac{1}{2}\frac{1}{|S|}\sum_{u\in S}\theta_u + \frac{1}{2}\frac{1}{|R|}\sum_{v\in R}\theta_v \tag{5}$$

$$\theta_u, \theta_v \sim Dirichlet(\alpha_k = 1/k) \tag{6}$$

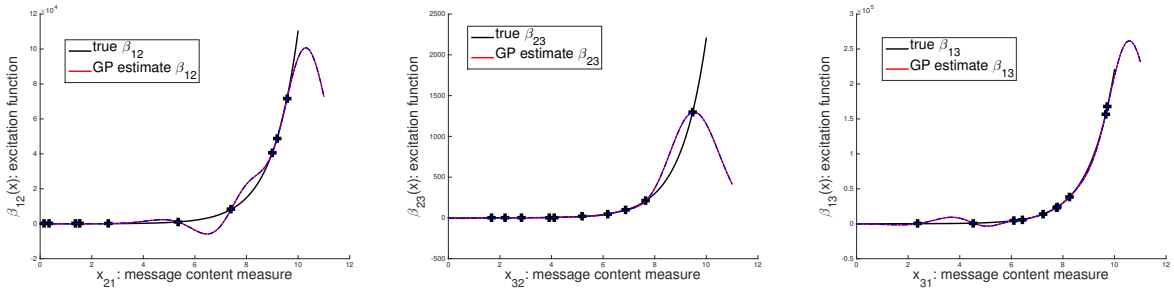
where θ_u, θ_v are individual word distributions of senders and receivers, respectively, and k the number of words in the corpus. $\theta_{S,R}$ is not symmetric.

3 Synthetic Data

Following the generative process described in section 2, we simulated 1000 message communications among 7 individuals (shown in Figure **??**). The clustering tree has two levels, $\{\#1, \#2, \#3\}$ are in cluster 1 (red), $\{\#4, \#5\}$ in cluster 2 (green), and $\{\#6, \#7\}$ in cluster 3 (blue).

The initial rate γ at the root is set to 1, and its fractions are distributed to its offspring proportional to their cluster sizes. The inverse decay rates τ_{uv} are set to 0.1 for all pairs of u, v.

The corpus we used was the top 10,000 words appeared in the NIPS dataset, which consists of 5811 papers published in the Conference on Neural Information Processing Systems (NIPS) during the years 1987



(a) GP plot of β_{12} (b) GP plot of β_{23} (c) GP plot of β_{13}

to 2015. Each message contains 20 words. The seven initial personalized word distributions are randomly generated across the 10,000 words in the corpus through a Dirichlet distribution with identical concentration parameters, i.e., $\alpha_k = 1/10000, \forall k = 1, \cdots, 1000$.

3.1 Quantitative Evaluation

Log-likelihood comparison against other models. We compared our method with existing alternative ones, and from Table 1 we see that our model achieved the best performance in terms of predictive log-likelihood. This is expected given that the data is generated from the model.

	Predictive Log-likelihood
nCRP + HGP	218.64 ($\pm$ 13.86)
CRP + HGP	187.28 ($\pm$ 12.78)
IRM + HP	109.13 ($\pm$ 17.21)
nCRF	129.78 ($\pm$ 9.76)
DHP	198.72 ($\pm$ 16.63)
HP	126.16 ($\pm$ 17.87)

Table 1: Comparing nCRP+HGP with other models. Predictive log-likelihoods with standard deviations (10 runs).

Parameter Estimation. Our next experiment focused on parameter estimation. In particular, we are interested in the inverse decay rate τ. This is because the excitation function is defined at individual level (whereas γ is defined at cluster level), which may lead to meaningful interpretations of individual's behavior and hence its estimation is more interesting. One immediate observation from the posterior of τ is that it has a multi-modal property, which is particularly true when $\tau = 0.5$. This makes the sampling of τ hard, since the multi-modality may lead to identification problems. Fortunately, further experiments show that τ only plays an insignificant role in predicting clusters, keywords and etc., which may due to the fact that when messages are communicated frequently, the effect of the "jump sizes" β dominant the behavior of rate functions.

3.2 Exploration of Model Characteristics

Does GP help?

Does hierarchical clustering structure help? We also compared our model with two manually designed trees: one being the "correct" underlying tree; the other being a "wrong" tree, which put all 7 individuals in one single cluster. Our model which samples trees from nCRP prior recovered the tree structure, and from Table **??** we see that it obtained very similar predictive log-likelihood as that of a correct manual tree, compared to the much worse performance from a wrong manual tree. It is noticed, however, the correct manual tree achieved smaller standard deviation over 10 experiment runs, which is what we expected since the fixed tree reduces randomness of the model.

Posterior Keyword Distributions. Our final experiment on synthetic data concerns the posterior keyword distributions. The leaf nodes in Figure **??** shows the posterior keyword distributions of the 7 individuals. The cluster level keyword distribution is aggregated from its members' distributions (top words of the union of top words), and the root keyword distribution is aggregated from the cluster ones. Thus, the top words in each histogram may not be the same. We also noticed that at the root, the words are almost evenly distributed, which suggests that the most important words across all individuals are almost of the same importance. We may use these top words to identify clusters.