Efficient Inference in the Infinite Multiple Membership Relational Model

Morten Mørup and Mikkel N. Schmidt
Section for Cognitive Systems, Technical University of Denmark
{mm,mns}@imm.dtu.dk

Summary: The Indian Buffet Process (IBP) is a stochastic process on binary features that has been applied to modeling communities in complex networks [4, 5, 6]. Inference in the IBP is challenging as the potential number of possible configurations grows as 2^{KN} where K is the number of latent features and N the number of nodes in the network. We presently consider the performance of three MCMC sampling approaches for the IBP; standard Gibbs sampling, joint Gibbs sampling and non-conjugate split/merge sampling. Our results indicate that including joint sampling significantly improves on parameter inference over standard Gibbs sampling while split-merge sampling appears useful for improving the inference as measured by burn-in-time of the sampler.

Introduction: Recently the Indian Buffet Process (IBP) [1] has been applied to modeling overlapping communities in networks[4, 5, 6]. We currently focus on the model proposed in [6] that is given by the following generative process

$$
\begin{aligned}
\boldsymbol{Z} &\sim IBP(\alpha), \\
\sigma &\sim Beta(\beta_c^+, \beta_c^-), \\
\eta_{lk} &\sim Beta(\beta_c^+, \beta_c^-) \\
\eta_{lk} &\sim Beta(\beta_o^+, \beta_o^-) \quad (\text{for } l \neq k) \\
Y_{ij} &\sim Bernoulli(1 - (1-\sigma)\prod_{lk}(1-\eta_{lk})^{z_{il}z_{jk}}).
\end{aligned}
$$

Thus, within-community link probabilities are drawn according to $Beta(\beta_c^+, \beta_c^-)$ whereas between community elements according to $Beta(\beta_o^+, \beta_o^-)$. A property of the above generative model is that features act as independent causes of links, and thus if a vertex gets an additional feature it will result in an increased probability of linking to other vertices formed by the noisy-or process $\pi_{ij} = 1 - (1-\sigma)\prod_{\ell k}(1-\eta_{\ell k})^{z_{i\ell}z_{jk}}$, In contrast to the model proposed by [5], where negative weights leads to features that inhibit links, the above model is more restricted. However, contrary to [5] the model scales in the number of links rather than the size of the network [6] and directly generalizes the infinite relational model (IRM) of [3, 7], thus the name infinite multiple-membership relational model (IMRM)[6].

We will investigate two approaches to enhance inference in the model over standard Gibbs sampling: joint Gibbs sampling over multiple latent features and split/merge sampling. We note that we can trivially absorb the noise parameter σ into $\boldsymbol{\eta}$ by forming a community $\eta_{00} = \sigma$ with no interaction to other communities, i.e. $\eta_{l0} = 0 \forall l$, $\eta_{0k} = 0 \forall k$ such that $z_{0i} = 1 \forall i$, i.e., by representing the noise term as an additional community that all nodes are members of. Letting the elements of the matrix $\boldsymbol{P}$ be given by $p_{\ell k} = \log(1 - \eta_{\ell k})$ we have for the joint distribution

$$
\begin{aligned}
p(\boldsymbol{Y}\boldsymbol{Z}, \boldsymbol{\eta}|\alpha, \beta_c^+, \beta_c^-, \beta_o^+, \beta_o^-) &\propto \left(\Big[\prod_{(i,j)\in\mathcal{Y}_1}(1 - e^{\boldsymbol{z}_i^\top \boldsymbol{P}\boldsymbol{z}_j})\Big]\exp\Big[\sum_{(i,j)\in\mathcal{Y}_0}\boldsymbol{z}_i^\top \boldsymbol{P}\boldsymbol{z}_j\Big]\right) \\
&\quad \left(\prod_k \frac{\eta_{kk}^{\beta_c^+-1}(1-\eta_{kk})^{\beta_c^--1}}{Beta(\beta_c^+,\beta_c^-)}\right)\left(\prod_{l\neq k}\frac{\eta_{\ell k}^{\beta_o^+-1}(1-\eta_{\ell k})^{\beta_o^--1}}{Beta(\beta_o^+,\beta_o^-)}\right) \\
&\quad \left(\frac{\alpha^K \prod_{k=1}^K \frac{(N-m_k)!(m_k-1)!}{N!}}{\prod_{\boldsymbol{h}\in[0,1]^N} K_{\boldsymbol{h}}!}\right),
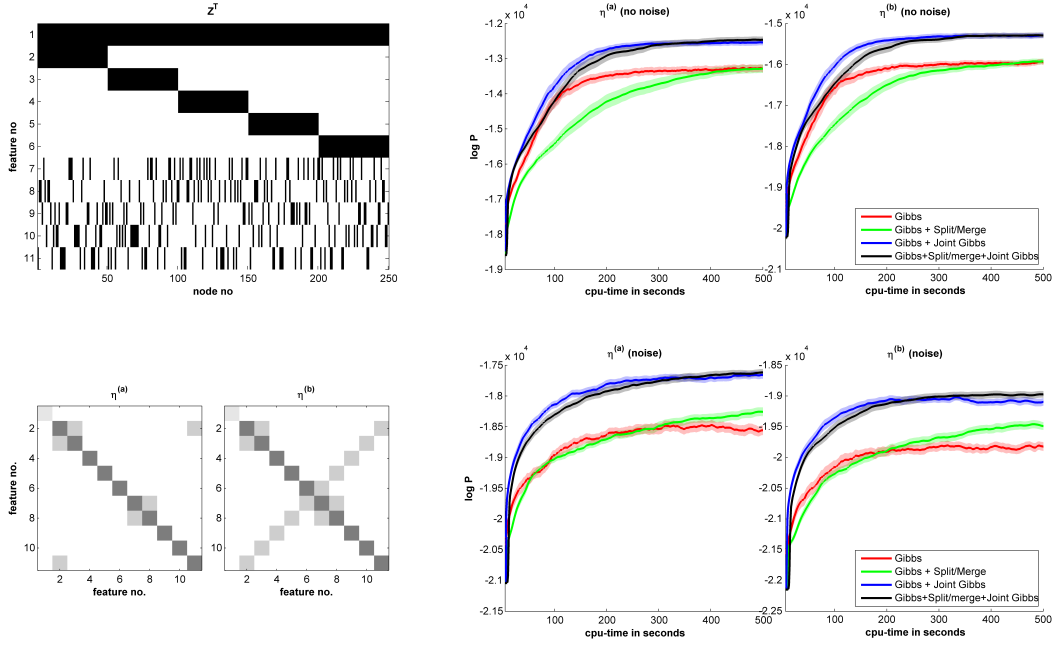\end{aligned}
$$

1

Figure 1: **Left panel:** The parameters used to generate the data. The generated networks have 250 nodes, a total of 10 clusters each containing 50 nodes such that each node belonged to two communities. Four undirected networks are considered generated by the two matrices $\boldsymbol{\eta}^{(a)}$ and $\boldsymbol{\eta}^{(b)}$ with and without noise given at the bottom left for the noise case (i.e. $\sigma = 0.1$). **Right panels:** Inference performance for IMRM for $\boldsymbol{\eta}$ with 3 and 7 non-zero entries in the off-diagonal respectively. Given in each plot is the joint posterior likelihood $\log P$ versus cpu-time for different combinations of inference procedures. Solid lines indicate the mean and shaded regions two times the standard deviation of the mean for 100 random initializations. In the analysis in the top row $\sigma = 0$ (i.e. no-noise) and in the bottom row $\sigma = 0.1$ (i.e., noise included). (In the analysis $\beta_c^+ = \beta_c^- = 1$, $\beta_o^+ = 0.1$, $\beta_o^- = 1$ favoring low densities in the off-diagonal elements of $\boldsymbol{\eta}$ and $\alpha = \log(250)$,).

where m_k is the number of vertices belonging to class k and $K_{\boldsymbol{h}}$ is the number of columns of $\boldsymbol{Z}$ equal to $\boldsymbol{h}[1]$.

Inference procedures: Let there be a total of K active components. The number of states of these K components are for the IBP 2^{KN}, thus the IBP invokes an exponential explosion in the number of possible states. This makes inference in the IBP challenging compared to the Chinese restaurant process with K^N possible states. We investigate three MCMC strategies for inference in the IBP given below. These update will be used interlaced with inference of $\boldsymbol{\eta}$ that is sampled element-wise by a random walk Metropolis-Hastings procedure.

Standard Gibbs Sampling: Following [1], a Gibbs sampler for the latent binary features $\boldsymbol{Z}$ can be derived. Consider sampling the kth feature of vertex v_i: If one or more other vertices also possess the feature, i.e., $m_{-ik} = \sum_{j \neq i} z_{jk} > 0$, the posterior marginal is given by $p(z_{ik} = 1 | \boldsymbol{Z}_{-(ik)}, \boldsymbol{P}, \boldsymbol{Y}) \propto p(\boldsymbol{Y}|\boldsymbol{Z}, \boldsymbol{P})\frac{m_{-ik}}{N}$. When evaluating the likelihood term, only the terms that depend on z_{ik} need be computed and the Gibbs sampler can be implemented efficiently by reusing computation and by up and down dating variables. In addition to sampling existing features, $K^{(i)} = \text{Poisson}\left(\frac{\alpha}{N}\right)$ new features should also be associated with v_i. Following [4] we sample the new features by Metropolis-Hastings using the prior as proposal density.

Joint Gibbs Sampling: Let $\mathcal{C}$ denote the set of components to jointly sample and let further $|\mathcal{C}|$ denote the number of components considered. We will Gibbs sample from the distribution of all $2^{|\mathcal{C}|}$ possible combinations of assignments given by $p(z_{i\mathcal{C}} = [0,1]^{|\mathcal{C}|}|\boldsymbol{Z}_{-(i\mathcal{C})}, \boldsymbol{P}, \boldsymbol{Y}) \propto p(\boldsymbol{Y}|\boldsymbol{Z}, \boldsymbol{P}) \prod_{k \in \mathcal{C}} \frac{m_{-ik}^{z_{ik}}(N-m_{-ik})^{1-z_{ik}}}{N}$. Again we exploit that when evaluating the likelihood terms, only the terms that depend on $z_{i\mathcal{C}}$ need be computed and the Gibbs sampler can be implemented

2

efficiently by reusing computation and by up and down dating variables. In our implementation we sample random subsets of features for $|\mathcal{C}| = \{4, 5, 6\}$.

Non-conjugate Split-Merge Sampling: Split-merge sampling in the IBP has previously been discussed briefly by [4] and [5]. We use the framework proposed in [2] for the Dirichlet process mixture model (DPMM) that exploits an intermediate launch state. This launch state we define by three restricted Gibbs sampling sweeps for $\boldsymbol{Z}$ and $\boldsymbol{\eta}$. Two types of proposals were used evenly; a split/merge move that randomly selects two features to split or merge, and a birth/death move that chooses either to introduce an additional feature from the noise cluster (birth) or to merge an existing cluster into the noise cluster given by σ (death).

Results and Discussion: Figure 1 shows the results of different combination of the inference procedures. As can be seen the best performing inference is not standard Gibbs sampling. A major improvement for all considered problems was found when combining Gibbs sampling with joint sampling. Furthermore, split/merge moves also appeared to slightly speed up the burn-in period of the sampler for this non-conjugate problem. To conclude, our results indicate that joint sampling is a very efficient sampling strategy for the IBP relative to standard Gibbs sampling while split/merge moves also appears to be useful.

References

[1] Thomas L. Griffiths and Zoubin Ghahramani. Infinite latent feature models and the Indian buffet process. In *Neural Information Processing Systems, Advances in (NIPS)*, pages 475–482, 2006.

[2] Sonia Jain and Radford M. Neal. A split-merge markov chain monte carlo procedure for the dirichlet process mixture model. *Journal of Computational and Graphical Statistics*, 13(1):158–182, 2004.

[3] Charles Kemp, Joshua B. Tenenbaum, Thomas L. Griffiths, Takeshi Yamada, and Naonori Ueda. Learning systems of concepts with an infinite relational model. In *Artificial Intelligence, Proceedings of the National AAAI Conference on*, 2006.

[4] Edward Meeds, Zoubin Ghahramani, Radford M. Neal, and Sam T. Roweis. Modeling dyadic data with binary latent factors. In *Advances in Neural Information Processing Systems (NIPS)*, volume 19, pages 977–984, 2007.

[5] Kurt T. Miller, Thomas L. Griffiths, and Michael I. Jordan. Nonparametric latent feature models for link prediction. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1276–1284, 2009.

[6] M. Mørup, M. N. Schmidt, and L. K. Hansen. Infinite multiple membership relational modeling for complex networks. *Presented first time at NIPS Workshop on Networks Across Disciplines in Theory and Application 2010, published IEEE MLSP 2011*.

[7] Zhao Xu, Volker Tresp, Kai Yu, and Hans-Peter Kriegel. Learning infinite hidden relational models. *Uncertainity in Artificial Intelligence (UAI2006)*, 2006.