Scalable Model Selection for Large-Scale Factorial Relational Models

Chunchen Liu * LIU_CHUNCHEN@NEC.CN
Lu Feng * FENG_LU@NEC.CN
NEC Laboratories China

Ryohei Fujimaki RFUJIMAKI@NEC-LABS.COM
Yusuke Muraoka YMURAOKA@NEC-LABS.COM
NEC Knowledge Discovery Research Laboratories

Abstract

With a growing need to understand large-scale networks, factorial relational models, such as binary matrix factorization models (BMFs), have become important in many applications. Although BMFs have a natural capability to uncover overlapping group structures behind network data, existing inference techniques have issues of either high computational cost or lack of model selection capability, and this limits their applicability. For scalable model selection of BMFs, this paper proposes stochastic factorized asymptotic Bayesian (sFAB) inference that combines concepts in two recently-developed techniques: stochastic variational inference (SVI) and FAB inference. sFAB is a highly-efficient algorithm, having both scalability and an inherent model selection capability in a single inference framework. Empirical results show the superiority of sFAB/BMF in both accuracy and scalability over state-of-the-art inference methods for overlapping relational models.

1. Introduction

Relational modeling has been an actively-studied research topic due to its increasing significance in such important applications as social network analyses (Kim et al., 2012), recommendation systems (Hsu, 2005), and bioinformatics (Jaimovich et al., 2006). Relational modeling has two main purposes: (1) to reveal latent group structures underlying the networks (partition entities into several groups on the basis of their connectivities) and (2) to predict unseen links on the basis of known links (i.e. link prediction). With respect to (1), identification of the appropriate number of groups (a.k.a. model selection) is one of the most important challenges in the learning of relational models.

Recently, research interests have been particularly focused on overlapping relational data (a single entity may belong to multiple groups), which is a natural property commonly observed in real-world relational networks. One pioneering work proposes mixed-membership stochastic blockmodels (MMSB) (Airoldi et al., 2008; Gopalan et al., 2012) that express "overlaps" using multinomial distributions. Analogous models like mixed membership relational clustering (Long et al., 2007) and modular structures (Azizi et al., 2014) take node information into consideration. For model selection, state-of-the-art algorithms usually follow non-parametric Bayesian frameworks by employing infinite priors such as hierarchical Dirichlet processes (Kim et al., 2013) and Chinese restaurant processes (Koutsourelakis & Eliassi-Rad, 2008), which automatically select the effective number of groups from among "infinite groups".

An alternative approach for expressing overlapping group structures is factorial modeling. Unlike multinomial modeling, in which individual entities are essentially generated from one group (similar to the situation in which each sample belongs to a single component in standard mixture models), factorial models express entities by combining multiple binary latent features, and, therefore, they can literally express overlapping structures. Binary matrix factorization models (BMFs) (Meeds et al., 2007) are powerful factorial relational models. Although they provide natural overlapping structures, their high computational cost of model selection, arising from non-parametric Bayesian priors, restricts their applicability to large-scale datasets.

To address the problem of Bayesian modeling on massive data, stochastic variational inference (SVI) has been recently developed (Hoffman et al., 2012) and has already been extended to several applications, such as topic models (Wang & Blei, 2012; Ranganath et al., 2013), time series modeling (Johnson & Willsky, 2014), multinomial relational models (Gopalan & Blei, 2013; Kim et al., 2013), and matrix factorization (Hernández-Lobato et al., 2014). SVI relies on external mechanisms for model selection (such as

* Equal contribution.

Proceedings of the 32^{nd} International Conference on Machine Learning, Lille, France, 2015. JMLR: W&CP volume 37.

non-parametric Bayesian priors and an outer loop for cross validation, which substantially increase computational cost in general), and it would be an interesting challenge to develop an SVI-type algorithm with model selection in a single inference framework in which model selection itself would be part of the optimization objective.

This paper proposes a scalable model selection algorithm for BMFs that combines two recently developed techniques: factorized asymptotic Bayesian (FAB) (Eto et al., 2014; Fujimaki & Morinaga, 2012; Fujimaki & Hayashi, 2012; Hayashi & Fujimaki, 2013; Hayashi et al., 2015) inference and SVI. FAB inference is a VB-like inference but has an inherent model selection mechanism in its inference procedure. Starting from a sufficient number of latent features, FAB inference automatically prunes away useless features by maximizing a factorized information criterion (FIC). We first derived FIC/FAB inference for BMFs. Inspired by SVI, we then derived stochastic FAB (sFAB) inference that enables us to perform model selection of BMFs on large-scale networks. sFAB introduces natural regularization that eliminates redundant latent features during stochastic optimization. This "shrinkage" mechanism significantly improves the computational efficiency of sFAB over standard SVI when model selection is required. Empirical results show the superiority of sFAB/BMF in both accuracy and scalability over state-of-the-art inference methods for overlapping relational modeling.

2. Related Work

BMF (Meeds et al., 2007) is the most well-studied factorial relational model because it naturally and expressively reveals overlapping group structures on dyadic data. It employs Beta-Bernoulli priors on latent features to support overlapping modeling, and it also can easily be extended to an infinite model by adopting Indian buffet process (IBP) priors to support automatic model selection for BMF. Latent feature relational models (LFRMs) (Miller et al., 2009) extend BMFs by incorporating covariates (i.e., entity attributes, multiple relations) for link prediction. As a family of LFRMs, max-margin nonparametric latent feature models (MNLFs) (Zhu, 2012) minimize hinge loss, which is a measure of the quality of link prediction, under the principle of maximum entropy discrimination. These BMF families employ non-parametric Bayesian priors for model selection, and Markov Chain Monte Carlo (MCMC) or mean-field variational inference is used for model inference. The biggest challenge is their heavy computation cost, especially on large-scale data.

Recently, more advanced factorial relational models have been proposed. Infinite latent attribute (ILA) models (Palla et al., 2012; 2014) use hierarchical structures to reveal overlapping, where entities are characterized by latent feature vectors, and each feature is partitioned into disjoint subclusters. Infinite multiple relational models (Morup et al., 2013) utilize IBP priors to associate entities with a subset of groups. However, these models allow for neither arbitrary feature interactions nor negative feature correlations, and thus provide less capability to explain data.

SVI has been introduced to relational modeling for Bayesian inference on massive data. SVINET applies SVI to a multinomial-mixture type relational model, though it has no explicit model selection capability (Gopalan & Blei, 2013). A hierarchical Dirichlet processes relational model, which is also multinomial, has been built on SVINET and automatically prunes away useless communities by means of hierarchical Dirichlet processes (Kim et al., 2013). To the best of our knowledge, no existing study has directly extended SVI to factorial models, while there is a study on SVI for probabilistic matrix factorization (Hernández-Lobato et al., 2014) handling continuous latent features.

3. FAB Inference for BMFs

3.1. Binary Matrix Factorization

BMF is a probabilistic model for an $I \times J$ data matrix $\mathbf{X}$ (Meeds et al., 2007), where I and J are the number of rows and columns, respectively. The (i,j)-th entry x_{ij} can be binary, count-valued, or continuous-valued. A row latent feature is denoted by $\mathbf{U} \in \{0,1\}^{I\times K}$, whose i-th row $\mathbf{u}_{i\cdot}$ indicates features associated with the i-th row of $\mathbf{X}$. Similarly, a column latent feature is denoted by $\mathbf{V} \in \{0,1\}^{J\times L}$, whose j-th row $\mathbf{v}_{j\cdot}$ corresponds to the j-th column of $\mathbf{X}$. A weight matrix $\mathbf{W} \in \mathbb{R}^{K\times L}$ is introduced to represent the primary parameters. Then the data matrix $\mathbf{X}$ is drawn from a stochastic process characterized by $\mathbf{UWV}^T$, i.e., a Bernoulli distribution with mean $f_\sigma(\mathbf{UWV}^T)$ for binary data. Here $f_\sigma(x) = 1/(1+\exp(-x))$ is the logistic sigmoid function. Although a previous study used BMFs with Gaussian distribution to model continuous pixel intensity (Meeds et al., 2007), in this paper we extend BMFs to the Bernoulli distribution because many real-world relational matrices are binary.

The likelihood function of BMFs is described as follows:

$$p(\mathbf{X}|\mathbf{U},\mathbf{V},\mathbf{W}) = \prod_{i=1}^{I}\prod_{j=1}^{J} p(x_{ij}|\mathbf{u}_{i\cdot},\mathbf{v}_{j\cdot},\mathbf{W}) \tag{1}$$
$$= \prod_{i=1}^{I}\prod_{j=1}^{J} f_\sigma(\mathbf{u}_{i\cdot}\mathbf{W}\mathbf{v}_{j\cdot}^T)^{x_{ij}} \left(1 - f_\sigma(\mathbf{u}_{i\cdot}\mathbf{W}\mathbf{v}_{j\cdot}^T)\right)^{1-x_{ij}}.$$

We use fully-factorized Bernoulli priors for $\mathbf{U}$ and $\mathbf{V}$, namely $u_{ik} \sim \mathcal{B}(\alpha_k)$ and $v_{jl} \sim \mathcal{B}(\beta_l)$, that is,

$$p(\mathbf{U}|\boldsymbol{\alpha}) = \prod_{i=1}^{I}\prod_{k=1}^{K} \alpha_k^{u_{ik}} (1-\alpha_k)^{1-u_{ik}}, \tag{2}$$

$$p(\mathbf{V}|\boldsymbol{\beta}) = \prod_{j=1}^{J}\prod_{l=1}^{L} \beta_l^{v_{jl}} (1-\beta_l)^{1-v_{jl}}. \tag{3}$$

The parameter set $\mathcal{P}$ is defined as $\mathcal{P} \equiv \{\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{W}\}$. The joint marginal likelihood is described as $p(\mathbf{X}, \mathbf{U}, \mathbf{V}|\mathcal{P}) = p(\mathbf{U}|\boldsymbol{\alpha})p(\mathbf{V}|\boldsymbol{\beta})p(\mathbf{X}|\mathbf{U}, \mathbf{V}, \mathbf{W})$. We assume a "log-flat" prior on $\mathcal{P}$ (i.e., $(\log p(\mathcal{P}))/N \to 0$, which is asymptotically ignored in the following FIC derivation, as is done in other FAB methods[1] (Fujimaki & Morinaga, 2012; Hayashi & Fujimaki, 2013).

3.2. FIC for BMFs

Consider the following marginal log-likelihood:

$$\log p(\mathbf{X}|\mathcal{M}) = \max_{q} \left\{ \sum_{\mathbf{U},\mathbf{V}} q(\mathbf{U}, \mathbf{V}) \log \frac{p(\mathbf{X}, \mathbf{U}, \mathbf{V}|\mathcal{M})}{q(\mathbf{U}, \mathbf{V})} \right\}, \tag{4}$$

where $p(\mathbf{X}, \mathbf{U}, \mathbf{V}|\mathcal{M}) = \int p(\mathbf{X}, \mathbf{U}, \mathbf{V}|\mathcal{P})\, p(\mathcal{P}|\mathcal{M})\, d\mathcal{P}$. $\mathcal{M}$ and $\mathcal{P}$ are a model and its parameters. The Laplace method (Wong, 2001) is individually applied to $\log p(\mathbf{U}|\boldsymbol{\alpha})$, $\log p(\mathbf{V}|\boldsymbol{\beta})$, and $\log p(\mathbf{X}|\mathbf{U}, \mathbf{V}, \mathbf{W})$. It is worth noting that we follow the asymptotic analysis of the Hessian matrix of log-likelihood in (Hayashi & Fujimaki, 2013) and obtain

$$\log \left|\overline{\mathbf{F}}_W\right| = \sum_{k=1}^{K}\sum_{l=1}^{L} \log \left(\left(\sum_{i=1}^{I}\sum_{j=1}^{J} u_{ik} v_{jl} \right) / (IJ) \right) + O_p(1), \tag{5}$$

where $|\bullet|$ is the determinant of $\bullet$, and $\overline{\mathbf{F}}_W = -\bigtriangledown^2 \log p(\mathbf{X}|\mathbf{U}, \mathbf{V}, \mathbf{W})/(IJ)$. By applying the Laplace method and introducing (5) to (4), we obtain the following FIC of BMF:

$$\begin{aligned} FIC_{BMF}(\mathbf{X}) \equiv \max_{q} \mathbb{E}_q \Bigg[& \log p(\mathbf{X}, \mathbf{U}, \mathbf{V}|\hat{\mathcal{P}}) \\ & - \frac{1}{2}\sum_{k=1}^{K}\sum_{l=1}^{L} \log \left(\sum_{i=1}^{I}\sum_{j=1}^{J} u_{ik} v_{jl} \right) \Bigg] - \frac{D_\alpha}{2}\log I \\ & - \frac{D_\beta}{2}\log J - \frac{D_W - KL}{2}\log(IJ) + H_q(\mathbf{U}, \mathbf{V}), \end{aligned} \tag{6}$$

where q is an arbitrary joint distribution on $\mathbf{U}$ and $\mathbf{V}$, $H_q(\bullet)$ is the entropy of $\bullet$, $D_\star$ is the dimensionality of the subscribed parameter $\star$, and $\hat{\mathcal{P}} = \arg\max_{\mathcal{P}} \log p(\mathbf{X}, \mathbf{U}, \mathbf{V}|\mathcal{P})$. Like other FICs (Fujimaki & Morinaga, 2012; Hayashi & Fujimaki, 2013), it is easy to show the asymptotic consistency of FIC_{BMF} with marginal log-likelihood, i.e., $FIC_{BMF}(\mathbf{X}) = \log p(\mathbf{X}) + \mathcal{O}_p(1)$ under certain regularity conditions[2].

Unlike previously-studied FIC for "single" latent variable models, FIC/BMF has the regularization term $\mathbb{E}_q[-\frac{1}{2}\sum_{k=1}^{K}\sum_{l=1}^{L}\log(\sum_{i=1}^{I}\sum_{j=1}^{J} u_{ik}v_{jl})]$, which comes from the approximation of the Hessian matrix of log-likelihood in (5). It is worth noting that we can see an explicit dependency between two latent features, $\mathbf{U}$ and $\mathbf{V}$, in the regularization term, which yields an effect of pruning away redundant features during the EM-like iterative optimization process. This indicates that FIC/BMF naturally captures a unique characteristic of relational models (i.e., "multiple" latent variables) in its regularization mechanism.

Note that FIC has the regularization effect despite its asymptotic ignorance of priors on $p(\mathcal{P})$. In this sense, its model selection mechanism is essentially different from other prior-based Bayesian model selection such as non-parametric Bayesian methods, automatic relevance determination (Wipf & Nagarajan, 2007), etc.

3.3. FAB algorithm for BMFs

For efficient maximization of (6), we introduce approximations to find a tractable lower bound. First, a mean-field approximation on $\mathbf{U}$ and $\mathbf{V}$ is introduced with the factorized forms of $q(\mathbf{U})$ and $q(\mathbf{V})$ as follows:

$$q(\mathbf{U}) = \prod_{i=1}^{I}\prod_{k=1}^{K} q(u_{ik})^{u_{ik}} \left(1 - q(u_{ik})\right)^{1-u_{ik}}, \tag{7}$$

$$q(\mathbf{V}) = \prod_{j=1}^{J}\prod_{l=1}^{L} q(v_{jl})^{v_{jl}} \left(1 - q(v_{jl})\right)^{1-v_{jl}}. \tag{8}$$

Second, we replace $\hat{\mathcal{P}}$ with $\mathcal{P}$ because, from the definition of $\hat{\mathcal{P}}$, $\log p(\mathbf{X}, \mathbf{U}, \mathbf{V}|\hat{\mathcal{P}}) \geq \log p(\mathbf{X}, \mathbf{U}, \mathbf{V}|\mathcal{P})$ holds for any $\mathcal{P}$. Third, $\log(\sum_i\sum_j u_{ik}v_{jl}) \leq \log \tilde{r}_{kl} + (\sum_i\sum_j u_{ik}v_{jl} - \tilde{r}_{kl})/\tilde{r}_{kl}$ is applied using the concavity of logarithm functions with parameters $\tilde{r}_{kl}$. The above three approximations have already been applied in existing FAB methods, but for BMFs, $\mathbb{E}_q\left[\log p\left(x_{ij}|\mathbf{u}_{i\cdot}, \mathbf{v}_{j\cdot}, \mathbf{W}\right)\right]$ is still intractable. We address this by borrowing the Gaussian lower bound technique of (Jaakkola & Jordan, 1997) as follows:

$$\begin{aligned} & \mathbb{E}_q[\log p(x_{ij}|\mathbf{u}_{i\cdot}, \mathbf{v}_{j\cdot}, \mathbf{W})] \geq x_{ij}\mathbf{u}_{i\cdot}\mathbf{W}\mathbf{v}_{j\cdot}^T + \log f_\sigma(\tilde{\xi}_{ij}) \\ & + \lambda(\tilde{\xi}_{ij})((\mathbf{u}_{i\cdot}\mathbf{W}\mathbf{v}_{j\cdot}^T)^2 - \tilde{\xi}_{ij}^2) - \frac{1}{2}(\mathbf{u}_{i\cdot}\mathbf{W}\mathbf{v}_{j\cdot}^T - \tilde{\xi}_{ij}) \qquad (9) \\ & := g(x_{ij}, \mathbf{u}_{i\cdot}, \mathbf{v}_{j\cdot}, \mathbf{W}, \tilde{\xi}_{ij}) \qquad (10) \end{aligned}$$

where $\lambda(\tilde{\xi}_{ij}) = (0.5 - f_\sigma(\tilde{\xi}_{ij}))/(2\tilde{\xi}_{ij})$, and $\tilde{\xi}_{ij}$ is a newly introduced parameter.

By utilizing above approximations, a tractable lower bound of $FIC_{BMF}(\mathbf{X})$ is derived as follows:

$$\begin{aligned} \mathcal{L}(q, \tilde{\mathcal{R}}, \tilde{\Xi}, \mathcal{P}, \mathbf{X}) = \mathbb{E}_q \Bigg[& \log p(\mathbf{U}|\boldsymbol{\alpha}) + \log p(\mathbf{V}|\boldsymbol{\beta}) \\ & + \sum_{i=1}^{I}\sum_{j=1}^{J} g(x_{ij}, \mathbf{u}_{i\cdot}, \mathbf{v}_{j\cdot}, \mathbf{W}, \tilde{\xi}_{ij}) \\ & - \frac{1}{2}\sum_{k=1}^{K}\sum_{l=1}^{L} \left(\log \tilde{r}_{kl} + \frac{\sum_{i=1}^{I}\sum_{j=1}^{J} u_{ik}v_{jl} - \tilde{r}_{kl}}{\tilde{r}_{kl}} \right) \Bigg] \\ & - \frac{D_\alpha}{2}\log I - \frac{D_\beta}{2}\log J + H_q(\mathbf{U}) + H_q(\mathbf{V}), \end{aligned} \tag{11}$$

[1] This assumption is necessary for asymptotic ignorance of priors in the FIC derivation.

[2] We can skip the proof. We followed the proof in (Hayashi & Fujimaki, 2013).

where $\tilde{\mathcal{R}} = \{\tilde{r}_{kl}\}_{k,l=1}^{K,L}$ and $\tilde{\Xi} = \{\tilde{\xi}_{ij}\}_{i,j=1}^{I,J}$.

An EM-style alternating optimization w.r.t. q, $\mathcal{P}$, $\tilde{\mathcal{R}}$, and $\tilde{\Xi}$ is derived with the stopping condition $(\mathcal{L}^{(t)} - \mathcal{L}^{(t-1)}) < \delta$. Here the superscription (t) denotes the t-th update, and δ is optimization tolerance.

FAB E-step: The FAB E-step optimizes $q(\mathbf{U})$ and $q(\mathbf{V})$ by fixing the parameter set $\{\mathcal{P}, \tilde{\mathcal{R}}, \tilde{\Xi}\}$. Because $q(\mathbf{U})$ and $q(\mathbf{V})$ do not have a closed form solution jointly, alternating updates of $q(\mathbf{U})$ and $q(\mathbf{V})$ are carried out several times in a single E-step. The update equation is found in Appendix A.1.

FAB M-step: The FAB M-step optimizes $\{\mathcal{P}, \tilde{\mathcal{R}}, \tilde{\Xi}\}$ by fixing $q(\mathbf{U})$ and $q(\mathbf{V})$. The update equations are found in Appendix A.2.

FAB Shrinkage step: The FAB E-step induces the sparseness of latent features (for details, please see Appendix A.1). In practice, we apply the following simple thresholding to eliminate redundant features: If $\sum_{i=1}^{I} q u_{ik} \left(\sum_{j=1}^{J} \sum_{l=1}^{L} q v_{jl} \right) < \epsilon$ (or $\sum_{j=1}^{J} q v_{jl} \left(\sum_{i=1}^{I} \sum_{k=1}^{K} q u_{ik} \right) < \epsilon$), then the k-th (l-th) feature will be eliminated with a pre-set threshold value ϵ. The FAB shrinkage step is carried out after the FAB E-step and removes irrelevant features. This reduces the model complexity, which means that the FIC lower bound will increase, and that we can mitigate over-fitting. Moreover, reducing irrelevant features during the EM iteration leads to less computational cost and makes FAB/BMF efficient, as we empirically show in our experiments.

4. Stochastic FAB Inference

Although FAB inference is a computationally efficient inference framework, the batch algorithm is still computationally too expensive to handle huge real-world networks. To address this computational bottleneck, we propose stochastic FAB (sFAB) inference, extending SVI (Hoffman et al., 2012) and its application to relational models (Gopalan & Blei, 2013; Kim et al., 2013). Although sFAB looks similar at first glance to existing SVI methods (e.g., SVINET), the former has an essential and important advantage over the latter, i.e., sFAB has an intrinsic model selection mechanism that eliminates redundant features during the stochastic optimization. This is particularly important in co-clustering scenarios in which model identification is one of the most significant interests (since we have many alternative methods for link prediction scenarios, such as probabilistic matrix factorization). Existing SVI methods achieve model identification with an external model selection mechanism. One simple way would be to use an outer loop for model selection, but this would require additional computational cost. Another way would be the incorporation of non-parametric Bayesian priors (Kim et al., 2013), but this would make the model more complex and also generally require additional computational cost.

Unlike SVI which is designed to be a stochastic optimization of evidence lower bounds (ELBO, a.k.a. variational free energy), our objective function is the FIC lower bound described in (11). Let i' and j', respectively, denote the indices of a row and column entities drawn from $\mathbf{X}$ uniformly; $i' \sim \mathcal{U}(1, \cdots, I)$, $j' \sim \mathcal{U}(1, \cdots, J)$, where $\mathcal{U}(\cdot)$ represents the uniform distribution. We can then derive an approximative (stochastic) FIC lower bound whose expectation is equal to that of (11) as follows:

$$\begin{aligned}\mathcal{L}_{i'j'}(q, \tilde{\mathcal{R}}, \tilde{\Xi}, \mathcal{P}, x_{i'j'}) &= I\big(\mathbb{E}_q[\log p(\mathbf{u}_{i'\cdot}|\boldsymbol{\alpha})] \\ &- \mathbb{E}_q[\log q(\mathbf{u}_{i'\cdot})]\big) + J\left(\mathbb{E}_q[\log p(\mathbf{v}_{j'\cdot}|\boldsymbol{\beta})] - \mathbb{E}_q[\log q(\mathbf{v}_{j'\cdot})]\right) \\ &+ IJ\mathbb{E}_q\left[g\left(x_{i'j'}, \mathbf{u}_{i'\cdot}, \mathbf{v}_{j'\cdot}, \mathbf{W}, \tilde{\xi}_{i'j'}\right)\right] \\ &- \mathbb{E}_q\left[\frac{1}{2}\sum_{k=1}^{K}\sum_{l=1}^{L}\left(\log \tilde{r}_{kl} + \frac{IJ u_{i'k} v_{j'l} - \tilde{r}_{kl}}{\tilde{r}_{kl}}\right)\right] + \text{const.}\end{aligned} \tag{12}$$

Next, sFAB is performed in a four-step iteration: (i) randomly subsample an entry $x_{i'j'}$ from the data matrix; (ii) update the variational distributions of associated latent feature vectors $\mathbf{u}_{i'\cdot}$ and $\mathbf{v}_{i'\cdot}$; (iii) prune away the useless features using the thresholding described in Section 3.3; (iv) incrementally update model parameters $\mathcal{P} = \{\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{W}\}$ in accord with the updated $\mathbf{u}_{i'\cdot}$ and $\mathbf{v}_{i'\cdot}$. The only difference from the standard stochastic approximative ELBO is the last term $-\mathbb{E}_q[\frac{1}{2}\sum_{k=1}^{K}\sum_{l=1}^{L}(\log \tilde{r}_{kl} + \frac{IJ u_{i'k} v_{j'l} - \tilde{r}_{kl}}{\tilde{r}_{kl}})]$, but this term yields the essential difference between SVI and sFAB, i.e., model selection capability in stochastic inference.

sFAB E-step The optimization of (12) w.r.t. $q^{(t)}(\mathbf{u}_{i'\cdot})$ results in (19) (in Appendix A.3) and so is $q^{(t)}(\mathbf{v}_{j'\cdot})$. After single sampling of i' and j', we iterate updating $\mathbf{u}_{i'\cdot}$ and updating $\mathbf{v}_{j'\cdot}$ several times. Different sampling strategies, such as node sampling and link sampling (Gopalan & Blei, 2013) are applicable as long as the expectation of an approximative FIC lower bound becomes equivalent to that of the original FIC lower bound (our sampling is similar to "pair sampling" proposed in (Gopalan & Blei, 2013)).

sFAB M-step We first compute the intermediate model parameter $\mathcal{P}^{(t-1/2)}$ by optimizing (12) w.r.t. $\mathcal{P}$. The update equation is found in Appendix A.4. Then we use a weighted average of $\mathcal{P}^{(t-1)}$ and $\mathcal{P}^{(t-1/2)}$ to update the model parameters. We denote the weight as ρ, which will provably lead $\mathcal{P}$ to converge to a local optimum if appropriately chosen. In each iteration, the parameter $\tilde{\Xi}$ also needs to be updated in a timely fashion to make the approximative lower bound tight because $g\left(x_{ij}, \mathbf{u}_{i\cdot}, \mathbf{v}_{j\cdot}, \mathbf{W}, \tilde{\xi}_{ij}\right)$ is close to $\log p\left(x_{ij}|\mathbf{u}_{i\cdot}, \mathbf{v}_{j\cdot}, \mathbf{W}\right)$ when $\mathbf{u}_{i\cdot}\mathbf{W}\mathbf{v}_{j\cdot}^T \in \left[-\tilde{\xi}_{ij}, \tilde{\xi}_{ij}\right]$.

The updated $\mathcal{P}$ might deviate significantly if information on single-edge distributions were used in individual iter-

Algorithm 1 stochastic FAB for BMFs

Initialize $\{q(\mathbf{U}),q(\mathbf{V}),\mathcal{P},\tilde{\mathcal{R}},\tilde{\Xi},\rho\}$.
while convergence criteria or a fixed number of maximum iterations is not met **do**
  Randomly sample a subset S_r of row entities and a subset S_c of column entities.
  for $m = 1, \cdots, M$ **do**
    Optimize $q^{(t,m)}(\mathbf{u}_{i\cdot})$, $\forall i \in S_r$ using (19).
    Optimize $q^{(t,m)}(\mathbf{v}_{j\cdot})$, $\forall j \in S_c$ accordingly.
    Update $\tilde{\mathcal{R}}$ and update $\tilde{\xi}_{ij}$, $\forall i \in S_r, \forall j \in S_c$ using (15) and (18).
  end for
  Prune away the useless features using the thresholding described in Section 3.3.
  Update $\tilde{\mathcal{R}}$ and $\tilde{\Xi}$ using (15) and (18).
  $\boldsymbol{\alpha}^{(t)} = (1-\rho^{(t)})\boldsymbol{\alpha}^{(t-1)} + \rho^{(t)} \frac{\sum_{i\in S_r} q(\mathbf{u}_{i\cdot})}{|S_r|}$
  $\boldsymbol{\beta}^{(t)} = (1-\rho^{(t)})\boldsymbol{\beta}^{(t-1)} + \rho^{(t)} \frac{\sum_{j\in S_c} q(\mathbf{v}_{j\cdot})}{|S_c|}$
  Optimize $\mathbf{W}^{(t-1/2)}$ based on $q^{(t)}(\mathbf{u}_{i\cdot})$ and $q^{(t)}(\mathbf{v}_{j\cdot})$, $\forall i \in S_r, \forall j \in S_c$ using (20).
  $\mathbf{W}^{(t)} = (1-\rho^{(t)})\mathbf{W}^{(t-1)} + \rho^{(t)}\mathbf{W}^{(t-1/2)}$
  Update $\tilde{\xi}_{ij}$, $\forall i \in S_r, \forall j \in S_c$
end while

ations, so a mini-batch $\{x_{ij}|i \in S_r, j \in S_c\}$ is adopted to generate a more informative parameter estimation, as is shown in Algorithm 1. We can easily apply a mini-batch strategy by little changes in the update equations in Appendix A.3 and A.4. Assuming the subsampling ratio to be γ, the speed of sFAB will be scaled up to $1/(\gamma^2)$ times over that with the batch algorithm.

5. Experiments

In the following experiments, we utilized 10-fold cross validation, each time holding out a different 10% of the data (links and non-links).

5.1. Baseline Methods

We compared the following four state-of-the-art methods with sFAB/BMF and FAB/BMF as baselines: 1) BMF with variational Bayesian inference (VB/BMF) (Miller, 2011), 2) BMF with Gibbs sampling (MCMC/BMF) (Meeds et al., 2007), 3) ILA with Gibbs sampling (Palla et al., 2014), and 4) SVINET (mixed membership stochastic block model with SVI) (Gopalan et al., 2012). The first two methods were selected to verify the superiority of the sFAB inference for relational modeling against the other inference methods. The third was selected as the latest and the most advanced factorial relational model and for a comparison with prior-based model selection (ILA uses IBP for model selection). The fourth was selected as state-of-the-art SVI-based relational modeling.

We used the latest implementations provided by the authors for SVINET (in C++)[3] and ILA (in MATLAB)[4], and implemented VB/BMF and MCMC/BMF in MATLAB on our own. sFAB/BMF and FAB/BMF were implemented in C++ to compare their computational time with that of SVINET. For model selection, ILA and MCMC/BMF select the model using IBP or Beta-Bernoulli priors, while SVINET and VB/BMF employ an outer loop of cross validation. sFAB/BMF and FAB/BMF select the feature numbers automatically using the shrinkage effect, as we have explained in Sections 3.2 and 3.3.

Prediction accuracy, clustering accuracy, and computational efficiency were evaluated using test log-likelihood, normalized mutual information (NMI), and elapsed time, respectively. We defined log-likelihood on test data as the test log-likelihood. The learning rate for sFAB/BMF was set to 0.5 for small datasets ($N < 1000$) and 0.2 for large datasets ($N \geq 1000$), in consideration of the balance between accuracy and efficiency. SVINET set the mini-batch to the entire set of links and used a learning rate of 1. That means its learning rate depended on the sparsity of a given network. We followed the experimental setting of the original paper of ILA (Palla et al., 2012) and ran 500 MCMC iterations for ILA and 1000 iterations for MCMC/BMF. sFAB/BMF, FAB/BMF, and SVINET all employed $\delta = 1 \times 10^{-5}$ as the optimization tolerance.

5.2. Description of Datasets

We adopted the "benchmark" tool (Lancichinetti & Fortunato, 2009) that was utilized in the SVINET paper (Gopalan & Blei, 2013) to generate synthetic "overlapping" networks with different scales, different group numbers, and different densities (dense and sparse). Further, eight real network datasets (namely, Zachary's Karate Club (Zachary, 1977), summer school survey network[5], U.S. Political Books[6], NIPS coauthorship network (Globerson et al., 2007), Facebook, autonomous systems, the collaboration network of Arxiv Astro Physics (AstroPh for short), and Arxiv High Energy Physics (HepPh for short)[7]) with different scales (node numbers ranged from 34 to 10K) were used to evaluate sFAB/BMF and FAB/BMF. The number of nodes and edges are summarized in Table 1. With the exception of the summer school survey network, these real networks were all undirected. For the NIPS coauthorship network, we extracted a subset of the 234 most connected authors from the overall network, in accord with previous studies (Miller et al., 2009; Palla et al., 2012). We also extracted a subset with 10K nodes from AstroPh and HepPh since SVINET

[3]https://github.com/premgopalan/svinet.
[4]http://mlg.eng.cam.ac.uk/konstantina/ILA/
[5]http://clique.ucd.ie/data
[6]http://www.orgnet.com
[7]http://snap.stanford.edu/data

Table 1. Comparisons for Test Log-likelihood on Real Networks. Standard deviations are showed in parentheses. The best and second best results for each dataset are highlighted in **bold** and *italic* respectively.

data	nodes	edges	sFAB/BMF	FAB/BMF	VB/BMF	MCMC/BMF	SVINET	ILA
Z.K.Club	34	78	$-0.259(.0229)$	$\mathbf{-0.230(.0419)}$	$-0.264(.0568)$	$-0.269(.0712)$	$-0.400(.0556)$	$-0.323(.0778)$
SumSchool	73	1138	$-0.265(.0279)$	$\mathbf{-0.248(.0201)}$	$-0.253(.0128)$	$-0.270(.0333)$	–	$-0.344(.0243)$
PolBooks	105	441	$-0.206(.0208)$	$\mathbf{-0.186(.0052)}$	$-0.256(.0317)$	$-0.268(.0427)$	$-0.260(.0144)$	$-0.494(.0465)$
NIPS	234	832	$-0.066(.0056)$	$\mathbf{-0.060(.0054)}$	$-0.100(.0195)$	$-0.106(.0271)$	$-0.084(.0099)$	$-0.108(.0095)$
Facebook	4.0K	88.2K	$-0.026(.0003)$	$\mathbf{-0.024(.0003)}$	–	–	$-0.027(.0001)$	–
as-733	5.2K	10.5K	$-0.007(.0001)$	$\mathbf{-0.005(.0002)}$	–	–	$-0.008(.0000)$	–
AstroPh	10K	144K	$-0.0143(.0001)$	$\mathbf{-0.0135(.0002)}$	–	–	$-0.0138(.0001)$	–
HepPh	10K	116K	$-0.008(.0001)$	$\mathbf{-0.007(.0003)}$	–	–	$-0.010(.0001)$	–

would have required an unrealistically long time for execution.

5.3. Synthetic Data

In all the simulation experiments below, we set the initial K as 40 for sFAB/BMF and FAB/BMF, and performed inference with $K = 2, \cdots, 40$ for VB/BMF and SVINET. ILA was excluded due to its huge computational cost. In terms of network densities, the probability of generating a link within a community (p_{in}) was about 0.7 for dense networks, and about 0.35 for sparse networks. For both dense and sparse networks, the probability of having a link between two nodes belonging to different communities (p_{out}) was roughly 0.01.

We first investigated the performance of sFAB/BMF, FAB/BMF, VB/BMF, MCMC/BMF, and SVINET on the synthetic networks with 500 nodes. The left-hand column in Figure 1 shows prediction accuracy (a) and clustering accuracy (b) in four different settings (K=10,30 $\times$ dense/sparse). In terms of difference in (batch) inference methods, FAB/BMF outperformed the other methods (VB and MCMC) and the same results were obtained in model selection (see Table 2 (top)). The two stochastic inference methods, sFAB/BMF and SVINET, performed competitively, but we should also note that the model in the benchmark tool (Lancichinetti & Fortunato, 2009) is a multinomial relational model and that this setting is rather advantageous to SVINET. Compared to FAB/BMF, sFAB/BMF suffered a little accuracy loss but still performed reasonably well. We have omitted comparisons in terms of computational speed on the small synthetic networks due to the inconsistency among implementation language platforms used for the various algorithms.

Next we evaluated sFAB/BMF, FAB/BMF, and SVINET on the synthetic networks with 5000 nodes. The right-hand column in Figure 1 shows prediction accuracy (c), clustering accuracy (d), and elapsed time (e) for four different settings (K=10,30 $\times$ dense/sparse). In terms of prediction accuracy, sFAB/BMF, FAB/BMF, and SVINET appeared comparable, and SVINET sometimes performed slightly better (again, this setting is rather advantageous to SVINET). On the other hand, we can also see a clear

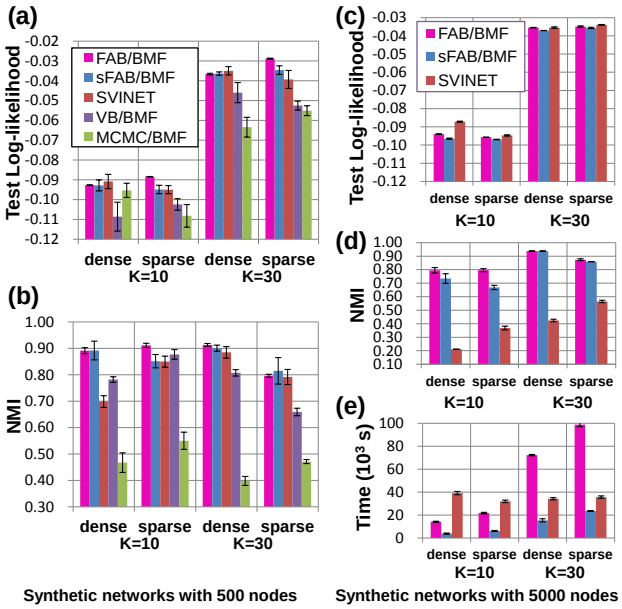

Figure 1. Comparisons on synthetic networks.

advantage for sFAB/BMF and FAB/BMF over SVINET in terms of clustering performance (NMI). One explanation is that SVINET assigns one node to a group when one link of the node belongs to that group, and that it overestimates the overlaps existing in the network. On the other hand, (s)FAB/BMF provide clear grouping assignments by estimating feature matrices. Comparison of sFAB/BMF with FAB/BMF shows that sFAB/BMF significantly reduced the computational cost without losing much accuracy. Further, sFAB/BMF is much more computationally efficient than SVINET. This is because 1) the shrinkage mechanism in sFAB/BMF automatically eliminates redundant features and eventually accelerates the algorithm over EM iterations, and 2) SVINET does not have an intrinsic model selection mechanism and needs an outer loop for model selection. Results here (particularly for clustering accuracy and efficiency) indicate that sFAB/BMF is more powerful than SVINET in detecting overlapping group structures (clustering).

The top half in Table 2 shows comparisons w.r.t. model selection. MCMC/BMF performed unstably (over- or under-estimation, depending on data). sFAB/BMF, FAB/BMF, and SVINET performed competitively and slightly better

Table 2. Comparisons for Model Selection Capability on synthetic data (top) and real-world data (bottom). Estimated feature numbers are shown with the standard deviations in parentheses.

N	K	Density	BMF				SVINET
			sFAB	FAB	VB	MCMC	
500	10	dense	11(1)	11(1)	10(1)	14(3)	10(0)
500	10	sparse	11(1)	10(1)	9(2)	9(1)	10(0)
500	30	dense	31(1)	31(1)	27(2)	21(5)	31(2)
500	30	sparse	33(1)	32(3)	27(3)	10(1)	32(2)
5000	10	dense	14(1)	12(1)	–	–	10(0)
5000	10	sparse	16(1)	13(1)	–	–	14(1)
5000	30	dense	30(0)	30(0)	–	–	33(2)
5000	30	sparse	34(1)	31(2)	–	–	31(1)

data	BMF				ILA	SVINET
	sFAB	FAB	VB	MCMC		
Z.K.Club	3(0)	5(1)	7(2)	5(1)	31(7)	3(1)
SumSchool	15(1)	15(1)	6(2)	6(1)	35(9)	–
PolBooks	8(1)	9(1)	11(2)	9(3)	73(8)	5(1)
NIPS	14(1)	15(1)	9(0)	14(3)	27(7)	18(2)
Facebook	62(2)	64(2)	–	–	–	75(11)
as-733	31(1)	28(4)	–	–	–	2(0)
AstroPh	96(2)	87(5)	–	–	–	100(0)
HepPh	83(4)	88(5)	–	–	–	93(4)

than VB/BMF. SVINET performed comparably with FAB/BMF and slightly better than sFAB/BMF. Overall, FAB/BMF and sFAB/BMF outperformed the other methods on the synthetic data despite the fact that the "true" model assumes multinomial relations.

5.4. Real-world Data

We evaluated the performance of sFAB/BMF, FAB/BMF, VB/BMF, MCMC/BMF, SVINET, and ILA on the small real networks ($N < 1000$) and compared the performance for sFAB/BMF and FAB/BMF with that for SVINET on the large real networks ($N \geq 1000$). On the small real networks, we set the initial K as 20 for sFAB/BMF and FAB/BMF, and performed inference with $K = 2, \cdots, 20$ for VB/BMF and SVINET. On the large networks, the upper bound of K was extended to 100. SVINET software supported undirected networks, and we omitted it on the summer school survey network, which is directed.

We have summarized in Table 1 the test log-likelihood for each algorithm as an evaluation metric for prediction accuracy . As can be seen, FAB/BMF outperformed all the other methods on all of the datasets. Combining results in Table 3, we find that sFAB/BMF sacrificed some precision for time, though the loss of precision was not large for those large-scale datasets. sFAB/BMF performed the second best after FAB/BMF on datasets other than the summer school survey network and AstroPh. VB/BMF and MCMC/BMF achieved comparable performance on the small networks, while SVINET was more suitable for the large networks. Results shown in Table 2 (bottom) suggest that ILA may have suffered from over-fitting, as its feature dimensionality is much larger than those of the others. Our experience indicates that non-parametric Bayesian priors prefer more complicated models than do true ones[8], and IBP in ILA might be the relevant factor. As

[8]There are theoretical studies for the case of mixture models (Miller & Harrison, 2013; 2014).

Table 3. Comparisons for Elapsed Time on Real Networks. Standard deviations are shown in parentheses. s, h, d, and m denote second, hour, day, and month respectively.

data	sFAB/BMF	FAB/BMF	SVINET
Z.K.Club	$1.07s(0.43s)$	$1.00s(0.46s)$	$228s(159s)$
SumSchool	$6.24s(2.17s)$	$7.68s(5.11s)$	–
PolBooks	$5.89s(2.85s)$	$8.76s(1.61s)$	$129s(50.5s)$
NIPS	$27.5s(10.8s)$	$113s(12.1s)$	$118s(14.0s)$
Facebook	$12.0h(1.29h)$	$87.7h(10.0h)$	$156h(2.13h)$
as-733	$13.0h(0.61h)$	$65.8h(2.37h)$	$75.7h(0.51h)$
AstroPh	$7.93d(0.89d)$	$22.1d(5.01d)$	$> 1m$
HepPh	$6.35d(0.17d)$	$21.7d(0.77d)$	$> 1m$

Table 2 shows, in most cases, sFAB/BMF and FAB/BMF chose the smallest number of features and provided the most compact and precise models. Further, sFAB/BMF and FAB/BMF tended to choose similar model complexities. This indicates that FAB model selection capability works well to mitigate over-fitting and can be integrated perfectly with stochastic optimization. Finally, we compared the elapsed time of sFAB/BMF , FAB/BMF, and SVINET, as is shown in Table 3. It is rather surprising that FAB/BMF performed even faster than SVINET. We observed two reasons for this. First, FAB/BMF performs model selection in a single EM iteration loop and therefore is significantly advantageous over SVINET which requires an outer loop for model selection. Note that, although the outer loop of SVINET can be parallelized, the computational cost is still lower bounded by the maximum K value. Second, SVINET took far more iterations for convergence. Intuitively speaking, (s)FAB/BMF can eliminate "poorly-fitted" latent features during EM iteration, and this accelerates convergence. Finally, sFAB/BMF significantly improved computational efficiency, particularly on large-scale networks.

6. Summary and Concluding Remarks

This paper has proposed a scalable inference method with model selection for large-scale BMFs, created by combining two recently developed technologies: FAB and SVI. sFAB is a highly-efficient algorithm, having both scalability and an inherent model selection capability in a single inference framework. Experimental results on both simulation and real datasets show that sFAB inference outperforms other inference technologies, such as VB and MCMC, and is more computationally efficient than SVINET when model selection is needed. A number of interesting areas remain for future work. For web-scale relational modeling, distributed stochastic optimization might be required. Also, SVI has been extended for streaming scenarios (Tamara et al., 2013). Extending sFAB's model selection capability to such streaming scenarios would be another interesting possibility to explore.

A. Appendix: Update Equations

A.1. FAB E-step

Let us consider the update of $q(\mathbf{U})$ as an example. The notation $\mathbf{U}_{-ik}$ refers to the set of entries in the matrix $\mathbf{U}$ other than u_{ik}.

For convenience, let qu_{ik} and qv_{jl} denote $q(u_{ik})$ and $q(v_{jl})$, respectively. Then, taking the gradient of $\mathcal{L}$ w.r.t. qu_{ik} and setting it to zero, we can obtain its closed-form solution:

$$
\begin{aligned}
&qu_{ik}^{(t,m)} = \\
&f_\sigma\Bigg(\log\frac{\alpha_k^{(t-1)}}{1-\alpha_k^{(t-1)}} + \sum_{l=1}^{L}\sum_{j=1}^{J}\left(x_{ij}-\frac{1}{2}\right)qv_{jl}^{(t,m-1)}w_{kl}^{(t-1)} \\
&+\sum_{j=1}^{J}\lambda(\tilde{\xi}_{ij}^{(t-1)})\Bigg(\sum_{l=1}^{L}qv_{jl}^{(t,m-1)}\left(w_{kl}^{(t-1)}\right)^2 \\
&+2\sum_{l=1}^{L}\sum_{q<l}qv_{jl}^{(t,m-1)}qv_{jq}^{(t,m-1)}w_{kl}^{(t-1)}w_{kq}^{(t-1)} \\
&+2\sum_{l=1}^{L}\sum_{p\neq k}qv_{jl}^{(t,m-1)}qu_{ip}^{(t,m-1)}w_{kl}^{(t-1)}w_{pl}^{(t-1)} \\
&+2\sum_{l=1}^{L}\sum_{q\neq l}\sum_{p\neq k}qv_{jl}^{(t,m-1)}qv_{jq}^{(t,m-1)}qu_{ip}^{(t,m-1)}w_{kl}^{(t-1)}w_{pq}^{(t-1)}\Bigg) \\
&-\sum_{l=1}^{L}\sum_{j=1}^{J}\frac{qv_{jl}^{(t,m-1)}}{2\tilde{r}_{kl}^{(t-1)}}\Bigg).
\end{aligned}
\tag{13}
$$

where the superscription (t,m) denotes the m-th update of the t-th E-step. The update of qv_{jl} is obtained in a similar manner. The last term of the R.H.S. of (13) originates from the regularization term in (6), i.e., $\mathbb{E}_q[-\frac{1}{2}\sum_{k=1}^{K}\sum_{l=1}^{L}\log(\sum_{i=1}^{I}\sum_{j=1}^{J}u_{ik}v_{jl})]$. We obtain $\tilde{r}_{kl} = IJ\alpha_k\beta_l$ according to (14) and (15). Roughly speaking, the smaller α_k is, the smaller qu_{ik} becomes (and vice versa), and this induces sparseness of latent features (many α_ks go zero during the EM iteration).

A.2. FAB M-step

The update equations for α, β and $\tilde{R}$ are obtained as follows:

$$
\alpha_k^{(t)} = \frac{\sum_{i=1}^{I}qu_{ik}^{(t)}}{I}, \quad \beta_l^{(t)} = \frac{\sum_{j=1}^{J}qv_{jl}^{(t)}}{J} \tag{14}
$$

$$
\tilde{r}_{kl}^{(t)} = \sum_{i=1}^{I}\sum_{j=1}^{J}qu_{ik}^{(t)}qv_{jl}^{(t)} \tag{15}
$$

Also, the update equation of weight matrix $\mathbf{W}$ is obtained as follows:

$$
w_{kl}^{(t)} = -\frac{\mathcal{F}\left(\mathbf{W}_{-kl}^{(t-1)}, q^{(t)}, \tilde{\Xi}^{(t-1)}, \mathbf{X}\right)}{2\sum_{i=1}^{I}\sum_{j=1}^{J}qu_{ik}^{(t)}qv_{jl}^{(t)}\lambda(\tilde{\xi}_{ij}^{(t-1)})}, \tag{16}
$$

$$
\begin{aligned}
&\mathcal{F}\left(\mathbf{W}_{-kl}^{(t-1)}, q^{(t)}, \tilde{\Xi}^{(t-1)}, \mathbf{X}\right) \\
=&\sum_{i=1}^{I}\sum_{j=1}^{J}qu_{ik}^{(t)}qv_{jl}^{(t)}\left(x_{ij}-\frac{1}{2}\right) \\
&+2\sum_{i=1}^{I}\sum_{j=1}^{J}qu_{ik}^{(t)}qv_{jl}^{(t)}\lambda(\tilde{\xi}_{ij}^{(t-1)})\Bigg(\sum_{q\neq l}qv_{jq}^{(t)}w_{kq}^{(t-1)} \\
&+\sum_{p\neq k}qu_{ip}^{(t)}w_{pl}^{(t-1)}+\sum_{q\neq l}\sum_{p\neq k}qu_{ik}^{(t)}qv_{jl}^{(t)}w_{pq}^{(t-1)}\Bigg),
\end{aligned}
\tag{17}
$$

where $\mathbf{W}_{-kl}$ denotes the entries in the matrix $\mathbf{W}$ other than w_{kl}. Given $\{q(\mathbf{U}), q(\mathbf{V}), \mathcal{P}, \tilde{\mathcal{R}}\}$, $\tilde{\Xi}$ is optimized by setting

$$
\tilde{\xi}_{ij}^{(t)} = \left(\mathbb{E}_q\left[\left(\sum_{k=1}^{K}\sum_{l=1}^{L}u_{ik}v_{jl}w_{kl}\right)^2\right]\right)^{\frac{1}{2}} \tag{18}
$$

A.3. sFAB E-step

Let us consider the update of $q(\mathbf{u}_{i'\cdot})$ as an example. Taking the gradient of $\mathcal{L}_{i'j'}$ in (12) w.r.t. $qu_{i'k}$ and setting it to zero, we can obtain the closed-form solution:

$$
\begin{aligned}
&qu_{i'k}^{(t,m)} = \\
&f_\sigma\Bigg(\log\frac{\alpha_k^{(t-1)}}{1-\alpha_k^{(t-1)}} + J\sum_{l=1}^{L}\left(x_{i'j'}-\frac{1}{2}\right)qv_{j'l}^{(t,m-1)}w_{kl}^{(t-1)} \\
&+J\lambda(\tilde{\xi}_{i'j'}^{(t-1)})\Bigg(\sum_{l=1}^{L}qv_{j'l}^{(t,m-1)}\left(w_{kl}^{(t-1)}\right)^2 \\
&+2\sum_{l=1}^{L}\sum_{q<l}qv_{j'l}^{(t,m-1)}qv_{j'q}^{(t,m-1)}w_{kl}^{(t-1)}w_{kq}^{(t-1)} \\
&+2\sum_{l=1}^{L}\sum_{p\neq k}qv_{j'l}^{(t,m-1)}qu_{i'p}^{(t,m-1)}w_{kl}^{(t-1)}w_{pl}^{(t-1)} \\
&+2\sum_{l=1}^{L}\sum_{q\neq l}\sum_{p\neq k}qv_{j'l}^{(t,m-1)}qv_{j'q}^{(t,m-1)}qu_{i'p}^{(t,m-1)}w_{kl}^{(t-1)}w_{pq}^{(t-1)}\Bigg) \\
&-J\sum_{l=1}^{L}\frac{qv_{j'l}^{(t,m-1)}}{2\tilde{r}_{kl}^{(t-1)}}\Bigg).
\end{aligned}
\tag{19}
$$

where the superscription (t,m) denotes the m-th update of the t-th E-step. The update of $qv_{j'l}$ is obtained in a similar manner.

A.4. sFAB M-step

We compute the intermediate model parameter $\mathcal{P}^{(t-1/2)}$ by taking the noisy gradient of $\mathcal{L}_{i'j'}$ in (12) w.r.t. $\mathcal{P}$ and setting it to zero. As a result, the update equation of intermediate weight matrix $\mathbf{W}^{(t-1/2)}$ is as follows:

$$
w_{kl}^{(t-1/2)} = -\frac{\mathcal{F}_{i'j'}\left(\mathbf{W}_{-kl}^{(t-1)}, qu_{i'k}^{(t)}, qv_{j'l}^{(t)}, \tilde{\xi}_{i'j'}^{(t-1)}, x_{i'j'}\right)}{2qu_{i'k}^{(t)}qv_{j'l}^{(t)}\lambda(\tilde{\xi}_{i'j'}^{(t-1)})}, \tag{20}
$$

$$
\begin{aligned}
&\mathcal{F}_{i'j'}\left(\mathbf{W}_{-kl}^{(t-1)}, qu_{i'k}^{(t)}, qv_{j'l}^{(t)}, \tilde{\xi}_{i'j'}^{(t-1)}, x_{i'j'}\right) \\
=&qu_{i'k}^{(t)}qv_{j'l}^{(t)}\left(x_{i'j'}-\frac{1}{2}\right)+2qu_{i'k}^{(t)}qv_{j'l}^{(t)}\lambda(\tilde{\xi}_{i'j'}^{(t-1)}) \\
&\Bigg(\sum_{q\neq l}qv_{j'q}^{(t)}w_{kq}^{(t-1)}+\sum_{p\neq k}qu_{i'p}^{(t)}w_{pl}^{(t-1)} \\
&+\sum_{q\neq l}\sum_{p\neq k}qu_{i'k}^{(t)}qv_{j'l}^{(t)}w_{pq}^{(t-1)}\Bigg).
\end{aligned}
\tag{21}
$$

References

Airoldi, E., Blei, D., Fienberg, S., and Xing, E. Mixed membership stochastic blockmodels. *JMLR*, 2008.

Azizi, E., Airoldi, E., and Galagan, J. Learning modular structures from network data and node variables. In *ICML*, 2014.

Eto, R., Fujimaki, R., Morinaga, S., and Tamano, H. Fully-automatic bayesian piecewise sparse linear models. In *AISTATS*, 2014.

Fujimaki, R. and Hayashi, K. Factorized asymptotic bayesian hidden markov models. In *ICML*, 2012.

Fujimaki, R. and Morinaga, S. Factorized asymptotic bayesian inference for mixture modeling. In *AISTATS*, 2012.

Globerson, A., Chechik, G., Pereira, F., and Tishby, N. Euclidean embeddding of co-occurrence data. *JMLR*, 2007.

Gopalan, P. and Blei, D. Efficient discovery of overlapping communities in massive networks. *PNAS*, 2013.

Gopalan, P., Mimno, D., Gerrish, S., Freedman, M., and Blei, D. Scalable inference of overlapping communities. In *NIPS*, 2012.

Hayashi, K. and Fujimaki, R. Factorized asymproric bayesian inference for latent feature models. In *NIPS*, 2013.

Hayashi, K., Maeda, S., and Fujimaki, R. Rebuilding factorized information criterion: Asymptotically accurate marginal likelihood. In *ICML*, 2015.

Hernández-Lobato, J., Houlsby, N., and Ghahramni, Z. Stochastic inference for scalable probabilistic modeling of binary matrices. In *ICML*, 2014.

Hoffman, M., Blei, D., Wang, C., and Paisley, J. Stochastic variational inference. *JMLR*, 2012.

Hsu, W. Relational graphical models for collaborative filtering and recommendation of computational workflow components. In *Workshop of IJCAI*, 2005.

Jaakkola, T. and Jordan, M. A variational approach to bayesian logistic regression models and their extensions. In *AISTATS*, 1997.

Jaimovich, A., Eledan, G., Margalit, H., and Friedman, N. Towards an integrated protein-protein interaction network: a relational markov network approach. *Journal of Computational Biology*, 2006.

Johnson, M. and Willsky, A. Stochastic variational inference for bayesian time series models. In *ICML*, 2014.

Kim, D., Hughes, M., and Sudderth, E. The nonparametric metadata dependent relational model. In *ICML*, 2012.

Kim, D., Gopalan, P., Blei, D., and Sudderth, E. Efficient online inference for bayesian nonparametric relational models. In *NIPS*, 2013.

Koutsourelakis, P. and Eliassi-Rad, T. Finding mixed-memberships in social networks. In *AAAI*, 2008.

Lancichinetti, A. and Fortunato, S. Benchmarks for testing community detection algorithms on directed and weighted graphs with overlapping communities. *Phys Rev E Stat Nonlin Soft Matter Phys*, 2009.

Long, B., Zhang, Z., and Yu, P. A probabilistic framework for relational clustering. In *KDD*, 2007.

Meeds, E., Ghahramani, Z., Neal, R., and Roweis, S. Modeling dyadic data with binary latent factors. In *NIPS*, 2007.

Miller, J. and Harrison, M. A simple example of dirichlet process mixture inconsistency for the number of components. In *NIPS*, 2013.

Miller, J. and Harrison, M. Inconsistency of pitman-yor process mixtures for the number of components. *JMLR*, 2014.

Miller, K. *Bayesian nonparametric latent feature models*. PhD thesis, University of California, Berkeley, 2011.

Miller, K., Griffiths, T., and Jordan, M. Nonparametric latent feature models for link prediction. In *NIPS*, 2009.

Morup, M., Schmidt, M., and Hansen, L. Infinite multiple membership relational modeling for complex metworks. In *Workshop of NIPS*, 2013.

Palla, K., Knowles, D., and Ghahramani, Z. An infinite latent attribute model for network data. In *ICML*, 2012.

Palla, K., Knowles, D., and Ghahramani, Z. Relational learning and network modelling using infinite latent attribute models. *IEEE Trans. PAMI*, 2014.

Ranganath, R., Wang, C., Blei, D., and Xing, E. An adaptive learning rate for stochastic variational inference. In *ICML*, 2013.

Tamara, B., Nicholas, B., Andre, W., and Ashia, W. Streaming variational bayes. In *NIPS*, 2013.

Wang, C. and Blei, D. Truncation-free online variational inference for bayesian nonparametric models. In *NIPS*, 2012.

Wipf, D. and Nagarajan, S. A new view of automatic relevance determination. In *NIPS*, 2007.

Wong, R. *Asymptotic Approximation of Integrals (Classics in Applied Mathematics)*. SIAM, 2001.

Zachary, W. An information flow model for conflict and fission in small groups. *Journal of Anthropological Research*, 1977.

Zhu, J. Max-margin nonparametric latent feature models for link prediction. In *ICML*, 2012.