Topic-Based Embeddings for Learning from Large Knowledge Graphs

Changwei Hu[1] **Piyush Rai**[12] **Lawrence Carin**[1]
[1] Department of ECE, Duke University [2]CSE Department, IIT Kanpur
{ch237,lcarin}@duke.edu piyush@cse.iitk.ac.in

Abstract

We present a scalable probabilistic framework for learning from multi-relational data, given in form of entity-relation-entity triplets, with a potentially massive number of entities and relations (e.g., in multi-relational networks, knowledge bases, etc.). We define each triplet via a relation-specific bilinear function of the embeddings of entities associated with it (these embeddings correspond to "topics"). To handle massive number of relations and the data sparsity problem (very few observations per relation), we also extend this model to allow sharing of parameters *across* relations, which leads to a substantial reduction in the number of parameters to be learned. In addition to yielding excellent predictive performance (e.g., for knowledge base completion tasks), the interpretability of our topic-based embedding framework enables easy qualitative analyses. Computational cost of our models scales in the number of *positive* triplets, which makes it easy to scale to massive real-world multi-relational data sets, which are usually extremely sparse. We develop simple-to-implement batch as well as online Gibbs sampling algorithms and demonstrate the effectiveness of our models on tasks such as multi-relational link-prediction, and learning from large knowledge bases.

1 INTRODUCTION

Learning from multi-relational data is ubiquitous in problems from a wide variety of areas such as social/biological network analysis (Goldenberg et al., 2010; Jenatton et al., 2012), and modeling of large knowledge bases such as YAGO (Suchanek et al., 2007), NELL (Carlson et al., 2010), Freebase (Bollacker et al., 2008), Google Knowledge Vault project (Dong et al., 2014), etc. Data in these problems usually consist of a sparsely observed set of triplets of the form `entity-relation-entity` and can be represented as a three-way binary sparse tensor $\boldsymbol{\mathcal{Y}}$ of size $N \times N \times R$ where N and R denote the number of entities and relations, respectively. The r-th slice $\mathcal{Y}^r \in \{0,1\}^{N\times N}$ corresponds to the r-th relation type where $\mathcal{Y}^r_{ij} = 1$ (denoting a positive example or a "valid" fact) denotes that existence of the relationship of type r between entities i and j. On the other hand, $\mathcal{Y}^r_{ij} = 0$ means that this relationship is either known to be invalid, or is unknown. Subsequently, at places, we will refer to $\mathcal{Y}^r_{ij}$ as a *fact* (valid/invalid). The number of positives (valid facts) is typically much smaller than the number of negatives (invalid/unknown facts).

Given such data, we may be interested in predicting the existence of the unknown links between entities (e.g., in social/biological networks) or in predicting the validity of *new* facts based on the knowledge of previously known facts (e.g., for knowledge base completion (Nickel et al., 2015)). Other examples of learning from such data include clustering of entities and/or relations, or ranking of entities for a given entity and a relation (e.g., for answering queries from a database).

Commonly used methods for learning from such data include methods based on tensor decomposition (Nickel et al., 2011; Sutskever et al., 2009; Jenatton et al., 2012; Hu et al., 2015), and more generally, methods that learn embeddings of the entities and relations (Socher et al., 2013; Bordes et al., 2013; Wang et al., 2014a; Yang et al., 2014; Dong et al., 2014). These embeddings are usually learned by optimizing some objective that assigns a higher score to an observed (positive) triplet as compared to unobserved (*assumed* negative) triplets, where the score is a function of the embeddings. In Section 4 on Related Work, we discuss these and other methods in more detail.

In this paper, we present a scalable Bayesian framework for the problem of learning from multi-relational data. Our embedding-based framework defines each triplet/fact $\mathcal{Y}^r_{ij}$ to be generated via a bilinear model of the form $p(\mathcal{Y}^r_{ij} = 1|\boldsymbol{u}_i, \boldsymbol{u}_j, \Lambda^r) = f(\boldsymbol{u}_i^\top \Lambda^r \boldsymbol{u}_j)$. Here,

Appearing in Proceedings of the 19^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2016, Cadiz, Spain. JMLR: W&CP volume 41. Copyright 2016 by the authors.

$\boldsymbol{u}_i, \boldsymbol{u}_j$ denote K-dimensional embeddings of entities i and j, respectively, Λ^r is a $K \times K$ matrix which parameterizes the relation type r, and f denotes the link function (defined subsequently in Sec. 2).

In contrast to the existing embedding-based approaches for multi-relational data (Socher et al., 2013; Bordes et al., 2013; Yang et al., 2014; Dong et al., 2014), an attractive aspect of our framework is that the learned embeddings naturally correspond to "topics" (each topic is a *distribution* over entities), which leads to improved interpretability and easy qualitative analyses, e.g., clustering entities/relations based on the topic(s) in which they are most "active".

Another appealing aspect of the proposed framework is its computational scalability. Instead of the commonly employed squared loss or logistic/probit likelihood for modeling binary multi-relational data (Nickel et al., 2011; Sutskever et al., 2009; Jenatton et al., 2012; Socher et al., 2013; Bordes et al., 2013), we leverage the Bernoulli-Poisson likelihood model (Zhou, 2015; Hu et al., 2015) for triplet generation (Sec. 2), which leads to an inference time-complexity that scales in the number of positive triplets in the data. Therefore our framework scales considerably better than the existing approaches for which the computational cost depends on the number of both positive and negative triplets (Nickel et al., 2011; Socher et al., 2013; Sutskever et al., 2009; Bordes et al., 2013; Yang et al., 2014; Dong et al., 2014). Moreover, as we discuss in Sec. 2.3, as compared to logistic/probit models, the Bernoulli-Poisson link function is also a more realistic model for *imbalanced* binary data, which is a characteristic exhibited in most real-world multi-relational data sets for which the number of positive triplets is far fewer than the number of negative triplets.

To handle the potentially massive number of relations commonly encountered in modern multi-relational data sets, we also extend our model to allow sharing of parameters *across* multiple relations, which leads to a substantial reduction in the number of parameters to be learned and also makes the model more robust in cases where the number of observations available for each relation is very small. To accomplish this, we model each of the relation-specific parameter matrices $\{\Lambda^r\}_{r=1}^R$ as a combination of a small set of "basis" relation matrices $\{G^m\}_{m=1}^M$, where $M \ll R$.

Finally, our fully Bayesian framework admits full local conjugacy, which allows deriving closed-form Gibbs sampling updates for all the model parameters. This, combined with the fact that the inference cost only depends on the number of positive triplets in the data, enable a fully Bayesian analysis for large-scale multi-relational data. We also develop an online inference algorithm that can process data in small minibatches and therefore can easily handle data sets that are too massive to deal with using batch algorithms.

2 BAYESIAN NON-NEGATIVE BILINEAR FACTOR MODEL

We first describe the basic setup of our Bayesian framework which is based on a bilinear *non-negative* latent factor model (Fig. 2) for multi-relational data, with the property that leads to scaling in the number of positive triplets. Then, in Section 2.1, we describe in more detail our first model with its properties that lead to efficient, fully Bayesian inference. Subsequently, in Section 2.2, we will generalize the first model to allow further sharing of statistical strength across the parameters of multiple relations. In Section 2.3, we will also provide a justification of why both these models, can more realistically model *imbalanced* binary data (very few positives), such as real-world multi-relational data sets.

One key aspect of both the proposed models is their departure from the standard logistic/probit link functions for binary-valued triplets, and the use of thresholded counts (Zhou, 2015; Hu et al., 2015) to model the binary-valued triplets. Specifically, each binary-valued triplet $\mathcal{Y}_{ij}^r$ is assumed generated by thresholding a *latent count* $\mathcal{X}_{ij}^r$ at 1, where the latent count $\mathcal{X}_{ij}^r$, in turn, is assumed drawn from a bilinear non-negative latent factor model

$$\mathcal{Y}_{ij}^r = \mathbf{1}(\mathcal{X}_{ij}^r \geq 1), \qquad \mathcal{X}_{ij}^r \sim \text{Poisson}(\boldsymbol{u}_i^\top \Lambda^r \boldsymbol{u}_j) \quad (3)$$

Intuitively, for relation r, the strength of the interaction between entities i and j depends the score $\boldsymbol{u}_i^\top \Lambda^r \boldsymbol{u}_j$, which defines the Poisson rate for $\mathcal{X}_{ij}^r$. Marginalizing out $\mathcal{X}_{ij}^r$ from Eq. 3, we have

$$\mathcal{Y}_{ij}^r \sim \text{Bernoulli}(1 - e^{-\boldsymbol{u}_i^\top \Lambda^r \boldsymbol{u}_j}) \quad (4)$$

Also note that the conditional posterior of the latent count $\mathcal{X}_{ij}^r$ can be written as

$$(\mathcal{X}_{ij}^r | \mathcal{Y}_{ij}^r, \boldsymbol{u}_i, \Lambda^r, \boldsymbol{u}_j) \sim \mathcal{Y}_{ij}^r \cdot \text{Poisson}_+(\boldsymbol{u}_i^\top \Lambda^r \boldsymbol{u}_j) \quad (5)$$

From Eq. 5, if $\mathcal{Y}_{ij}^r = 0$ then $\mathcal{X}_{ij}^r = 0$, almost surely (a.s.), and if $\mathcal{Y}_{ij}^r = 1$ then $\mathcal{X}_{ij}^r \sim \text{Poisson}_+(\boldsymbol{u}_i^\top \Lambda^r \boldsymbol{u}_j)$, a draw from zero-truncated Poisson. Therefore we only need to sample $\mathcal{X}_{ij}^r$ if $\mathcal{Y}_{ij}^r = 1$. We leverage this property in Section 3 to design scalable inference algorithms for our framework. We next describe both of our models which are based on this overall framework.

2.1 Model-1

The complete generative story for the first model, along with the prior distributions over the various

$$\begin{aligned}
\mathcal{Y}^r_{ij} &= \mathbf{1}(\mathcal{X}^r_{ij} \geq 1) \\
\boldsymbol{\mathcal{X}}^r &\sim \text{Poisson}(\mathbf{U}\boldsymbol{\Lambda}^r\mathbf{U}^T) \\
\mathbf{U}_{:,k} &\sim \text{Dirichlet}(a, \ldots, a) \quad \forall k = 1, \ldots, K \\
\boldsymbol{\Lambda}^r_{k_1k_2} &\sim \begin{cases} \text{Gamma}(\epsilon^r d^r_{k_1}, \frac{1}{\beta}), \text{ if } k_1 = k_2 \\ \text{Gamma}(d^r_{k_1} d^r_{k_2}, \frac{1}{\beta}), \text{ if } k_1 \neq k_2 \end{cases} \\
d^r_k &\sim \text{Gamma}(\gamma_0/K, 1/c_0) \\
\epsilon^r &\sim \text{Gamma}(e_0, 1/f_0)
\end{aligned}$$

$$\begin{aligned}
\mathcal{Y}^r_{ij} &= \mathbf{1}(\mathcal{X}^r_{ij} \geq 1) \quad \boldsymbol{\mathcal{X}}^r \sim \text{Poisson}(\mathbf{U}\boldsymbol{\Lambda}^r\mathbf{U}^T) \\
\mathbf{U}_{:,k} &\sim \text{Dirichlet}(a, \ldots, a) \quad \forall k = 1, \ldots, K \\
\Lambda^r_{k_1k_2} &= \sum_{m=1}^{M} \eta_{mr} G^m_{k_1k_2}, \eta_{mr} \sim \text{Gamma}(h_0, \frac{1}{q_0}) \\
G^m_{k_1k_2} &\sim \begin{cases} \text{Gamma}(\epsilon^m d^m_{k_1}, \frac{1}{\beta}), \text{ if } k_1 = k_2 \\ \text{Gamma}(d^m_{k_1} d^m_{k_2}, \frac{1}{\beta}), \text{ if } k_1 \neq k_2 \end{cases} \\
d^m_k &\sim \text{Gamma}(\gamma_0/K, 1/c_0) \quad \epsilon^m \sim \text{Gamma}(e_0, 1/f_0)
\end{aligned}$$

Figure 1: Left: Model-1 with each relation r having its own independent parameter matrix Λ^r. Right: Model-2 with parameter sharing across relations via a set of basis matrices

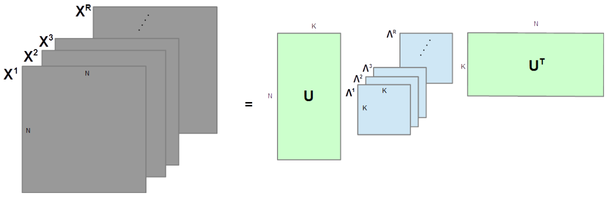

Figure 2: The basic setup of the bilinear latent factor model for multi-relational data

model parameters is shown in Fig. 1 (left). The $N \times K$ matrix $\mathbf{U} = [\boldsymbol{u}_1 \ldots \boldsymbol{u}_N]^\top$ contains the K-dimensional embeddings of each of the N entities. We further assume each N-dimensional *column* of $\mathbf{U}$ is drawn from a Dirichlet, which can therefore be thought of as a *distribution* (or "topic") over the N entities (akin to a topic model). In contrast to the other embedding based models (Nickel et al., 2011; Sutskever et al., 2009; Jenatton et al., 2012; Socher et al., 2013; Bordes et al., 2013), this aspect of our model provides a nice interpretability to the entity embeddings, because each of the K embedding coordinates of an entity can now be thought of as how "active" it is in each of the K topics. This naturally allows to group/cluster the entities based on topics, without having to perform a separate step of running a clustering algorithm over the learned embeddings.

Although the model in Fig. 1 (left) is not originally conjugate, using the Poisson-multinomial equivalence (Dunson and Herring, 2005; Zhou et al., 2012), we are able to develop a Gibbs sampler with closed-form sampling updates for all the model parameters. To see this, note that a Poisson distributed count-valued random variable can be expressed as a sum of Poisson distributed latent counts which, in turn, can be generated by repeatedly sampling from a multinomial. To illustrate this in the context of our model, note that we can express each latent count $\mathcal{X}^r_{ij} \sim \text{Poisson}(\sum_{k_1}^K \sum_{k_2}^K u_{ik_1}\Lambda^r_{k_1k_2}u_{jk_2})$ as a sum of latent counts, i.e., $\mathcal{X}^r_{ij} = \sum_{k_1}^K \sum_{k_2}^K \mathcal{X}_{ik_1k_2j}$ where $\mathcal{X}_{ik_1k_2j} \sim \text{Poisson}(u_{ik_1}\Lambda^r_{k_1k_2}u_{jk_2})$, and then using the Poisson-multinomial equivalence, we have $\forall k_1, k_2$

$$\{\mathcal{X}_{ik_1k_2j}\} \sim \text{Mult}\left(\mathcal{X}^r_{ij}; \frac{\{u_{ik_1}\Lambda^r_{k_1k_2}u_{jk_2}\}}{\sum_{k_1=1}^K \sum_{k_2=1}^K u_{ik_1}\Lambda^r_{k_1k_2}u_{jk_2}}\right)$$

This, coupled with the multinomial-Dirichlet conjugacy, allows us to develop a Gibbs sampler for our model. Section 3 briefly describes the Gibbs sampler (both batch as well as a more efficient online version) and the Supplementary Material provides the additional details.

2.2 Model-2: Sharing Parameters Across Relations

Model-1 parametrizes each relation type r by Λ^r, a $K \times K$ matrix. In real-world multi-relational data sets with potentially thousands of relation types, often many relation types may be similar to each other and therefore, instead of modeling each Λ^r independently, it may be more appropriate to jointly model these in order to share statistical strength across relations. This significantly reduces the number of parameters that need to be learned and can also be helpful to handle the data sparsity problem, i.e., when the number of triplets observed per relation is very small. Our second model, with generative story shown in Fig. 1 (right), allows such a sharing by modeling the parameters Λ^r of each relation type as a linear combination of M "basis" relation parameter matrices $\{G^m\}_{m=1}^M$ shared by all the relations, i.e.,

$$\Lambda^r = \sum_{m=1}^{M} \eta_{mr} G^m$$

Fig. 3 illustrates this idea pictorially. If two relations r and r' are similar, their combination weight vectors η_r and $\eta_{r'}$ are also expected to be similar. Note that, in this model, we can also view the combination weights $\eta_r = [\eta_{1r}, \ldots, \eta_{Mr}] \in \mathbb{R}_+^M$ as an embedding of relation type r. Therefore, unlike other bilinear models (Nickel et al., 2011; Sutskever et al., 2009), this model also provides a vector embedding for relations as well. Interestingly, this model structurally resembles a non-negative variant of Tucker tensor factorization model (Kolda and Bader, 2009) with $\mathbf{U}$ as the factor matrix of the entity dimension, $\boldsymbol{\eta} = [\eta_1, \dots, \eta_R]$ as the factor matrix of the relation dimension, and $\{G^m\}_{m=1}^M$ being the *core tensor*. Just like model-1, leveraging the Poisson-multinomial equivalent allows us to to develop closed-form Gibbs sampling updates for all the model parameters (Section 3.2 provides further details).

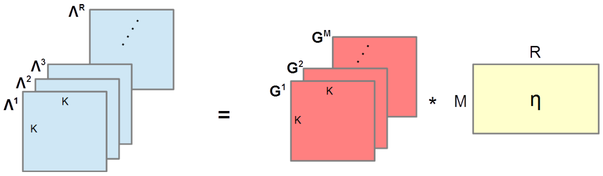

Figure 3: Parameter sharing across relations for Model-2

2.3 Connection with the complementary log-log link function

It is interesting to note that the form of the likelihood function $\mathcal{Y}_{ij}^r \sim \text{Bernoulli}(1 - e^{-\boldsymbol{u}_i^\top \Lambda^r \boldsymbol{u}_j})$, arising in both our models, resembles the complementary log-log (cloglog) function (Piegorsch, 1992; Collett, 2002) often used to model imbalanced binary data. In particular, the rate of growth of the function $p(\mathcal{Y}_{ij}^r = 1) = 1 - e^{-\boldsymbol{u}_i^\top \Lambda^r \boldsymbol{u}_j}$ (where $\boldsymbol{u}_i^\top \Lambda^r \boldsymbol{u}_j$ is non-negative) along the Y axis from 0.5 to 1 tends to be much slower than the rate it drops from 0.5 to 0. Therefore, our generative model more realistically captures the data imbalance in real-world multi-relational data sets that have very few positive triplets.

3 INFERENCE

As discussed in Section 2.1, using data-augmentation and Poisson-multinomial equivalence (Dunson and Herring, 2005; Zhou et al., 2012), we are able to derive closed-form Gibbs sampling updates for all the model parameters. Note again that since we only need to sample the latent counts $\mathcal{X}_{ij}^r$ for the positive triplets, our sampling algorithms scale in the number of positive triplets, thereby leading to a very efficient inference. In Section 3.1 and 3.2, we first describe the batch Gibbs sampling algorithms for model-1 and model-2, respectively, and then, in Section 3.3, we briefly describe an online Gibbs sampler for both models (further details given in the Supplementary Material). In the rest of the exposition, we refer to the model-1 as **BPBFM-1** and model-2 as **BPBFM-2**, where BPLFM denotes **B**ernoulli-**P**oisson **B**ilinear **F**actor **M**odel to reflect our generative model.

3.1 Gibbs Sampling for BPBFM-1

Note that, for BPBFM-1 (generative model shown in Fig. 1-left), the latent count for each triplet $\mathcal{Y}_{ij}^r$ is defined as $\mathcal{X}_{ij}^r = \sum_{k_1=1}^K \sum_{k_2=1}^K \mathcal{X}_{ik_1k_2j}$. In what follows, we will also make use of the following quantities, defined in terms of the $\mathcal{X}_{ik_1k_2j}$'s:

$$\mathcal{X}_{ik\cdot\cdot}^r = \sum_{k_2=1}^K \sum_{j=1}^N \mathcal{X}_{ikk_2j}^r \tag{6}$$

$$\mathcal{X}_{ik\cdot\cdot} = \sum_{r=1}^R \mathcal{X}_{ik\cdot\cdot}^r \tag{7}$$

$$\mathcal{X}_{\cdot k_1k_2\cdot}^r = \sum_{i=1}^N \sum_{j=1}^N \mathcal{X}_{ik_1k_2j}^r \tag{8}$$

Sampling $\mathcal{X}_{ij}^r$: For each triplet $\mathcal{Y}_{ij}^r = 1$, the latent count $\mathcal{X}_{ij}^r$ can be sampled as

$$\mathcal{X}_{ij}^r \sim \mathcal{Y}_{ij}^r \cdot \text{Poisson}_+\left(\sum_{k_1=1}^K \sum_{k_2=1}^K u_{ik_1} \Lambda_{k_1k_2}^r u_{jk_2}\right) \tag{9}$$

Note that this only needs to be done for the observed triplets (i.e., if $\mathcal{Y}_{ij}^r = 1$).

Sampling $\mathcal{X}_{ik_1k_2j}^r$: Due to the Poisson-multinomial equivalence, $\mathcal{X}_{ik_1k_2j}^r$ can be sampled as

$$\{\mathcal{X}_{ik_1k_2j}^r\} \sim \text{Mult}\left(\mathcal{X}_{ij}^r; \frac{\{u_{ik_1}\Lambda_{k_1k_2}^r u_{jk_2}\}}{\sum_{k_1=1}^K \sum_{k_2=1}^K u_{ik_1}\Lambda_{k_1k_2}^r u_{jk_2}}\right) \tag{10}$$

Sampling $\mathbf{U}_{:,k}$: Using Dirichlet-multinomial conjugacy, each column of $\mathbf{U}$ can be sampled as

$$\mathbf{U}_{:,k} \sim \text{Dirichlet}(a + \mathcal{X}_{1k\cdot\cdot}, a + \mathcal{X}_{2k\cdot\cdot}, \dots, a + \mathcal{X}_{Nk\cdot\cdot}) \tag{11}$$

Sampling d_k^r:. Using the additive property of the Poisson draws, we have

$$\mathcal{X}_{\cdot k_1k_2\cdot}^r \sim \text{Poisson}\left(\sum_{i=1}^N \sum_{j=1}^N u_{ik_1}\Lambda_{k_1k_2}^r u_{jk_2}\right) \tag{12}$$

Marginalizing out $\Lambda_{k_1k_2}^r$ from Eq.(12), we have

$$\mathcal{X}_{\cdot k_1k_2\cdot}^r \sim \text{NegBin}((\epsilon^r)^{\delta_{k_1k_2}} d_{k_1}^r (d_{k_2}^r)^{1-\delta_{k_1k_2}}, p_{k_1k_2}) \tag{13}$$

where $\delta_{k_1k_2} = 1$ if $k_1 = k_2$, and $\delta_{k_1k_2} = 0$ otherwise, and NegBin denotes the Negative Binomial distribution. In the above, $p_{k_1k_2}$ is define as $p_{k_1k_2} = \frac{\theta_{k_1k_2}}{\theta_{k_1k_2}+\beta}$ where $\theta_{k_1k_2} = \sum_{i=1}^N \sum_{j=1}^N u_{ik_1}u_{jk_2}$. Using Eq.(13) and the data augmentation scheme proposed in ((Zhou et al., 2012)), d_k^r can be sampled by first sampling $\ell_{kk_2}^r \sim \sum_{t=1}^{X_{\cdot k_1k_2\cdot}^r} \text{Bernoulli}\left(\frac{(\epsilon^r)^{\delta_{k_1k_2}} d_{k_1}^r (d_{k_2}^r)^{1-\delta_{k_1k_2}}}{(\epsilon^r)^{\delta_{k_1k_2}} d_{k_1}^r (d_{k_2}^r)^{1-\delta_{k_1k_2}}+t-1}\right)$ and then sampling d_k^r as $d_k^r \sim \text{Gamma}\left(\frac{\gamma_0}{K} + \sum_{k_2=1}^K \ell_{kk_2}^r, \frac{1}{c_0 - \sum_{k_2}^K (\epsilon^r)^{\delta_{kk_2}} (d_{k_2}^r)^{1-\delta_{kk_2}} \ln(1-p_{kk_2})}\right)$.

Sampling ϵ^r: ϵ^r can be sampled as $\epsilon^r \sim \text{Gamma}(e_0 + \sum_{k=1}^{K} \ell_{kk}^r, \frac{1}{f_0 - \sum_k^K d_k^r \ln(1-p_{kk})})$

Sampling $\Lambda_{k_1k_2}^r$: Using Gamma-Poisson conjugacy, $\Lambda_{k_1k_2}^r$ can be sampled as $\text{Gamma}((\epsilon^r)^{\delta_{k_1k_2}} d_{k_1}^r (d_{k_2}^r)^{1-\delta_{k_1k_2}} + \mathcal{X}_{\cdot k_1k_2\cdot}^r, \frac{1}{\beta+\theta_{k_1k_2}})$.

3.2 Gibbs Sampling for BPBFM-2

Proceeding in a manner similar to as we did for BPBFM-1, we can express each latent count $\mathcal{X}_{ij}^r$ in BPBFM-2 (which models each Λ^r as $\Lambda^r = \sum_{m=1}^{M} \eta_{mr} G^m$) as a sum of the following form: $\mathcal{X}_{ij}^r = \sum_{k_1=1}^{K} \sum_{k_2=1}^{K} \sum_{m=1}^{M} X_{ik_1k_2mj}^r$ where

$$\mathcal{X}_{ik_1k_2mj}^r \sim \text{Poisson}(u_{ik_1} \eta_{mr} G_{k_1k_2}^m u_{jk_2})$$

We further define

$$\mathcal{X}_{\cdot k_1k_2m\cdot}^{\cdot} = \sum_{i=1}^{N} \sum_{j=1}^{N} \sum_{r=1}^{R} \mathcal{X}_{ik_1k_2mj}^r$$

$$\mathcal{X}_{\cdots m\cdot}^r = \sum_{i=1}^{N} \sum_{j=1}^{N} \sum_{k_1=1}^{K} \sum_{k_2=1}^{K} \mathcal{X}_{ik_1k_2mj}^r$$

Using additive property of Poisson distribution

$$\mathcal{X}_{\cdot k_1k_2m\cdot}^{\cdot} \sim \text{Poisson}(\theta_{k_1k_2} G_{k_1k_2}^m \sum_{r=1}^{R} \eta_{mr}) \tag{14}$$

$$\mathcal{X}_{\cdots m\cdot}^r \sim \text{Poisson}(\eta_{mr} \sum_{k_1=1}^{K} \sum_{k_2=1}^{K} \theta_{k_1k_2} G_{k_1k_2}^m) \tag{15}$$

The Gibbs sampler updates for BPBFM-2 depend on these quantities. For brevity, we provide the detailed update equations in the Supplementary Material.

3.3 Online Gibbs Sampling

Although the Gibbs sampler we presented in the previous sections is efficient for moderate-sized data sets, when the number of observed triplets (and the number of entities and/or relations), batch Gibbs sampling can be prohibitive to run and/or may have slow mixing. We therefore also develop online Gibbs sampling algorithms for both of our models. The proposed online Gibbs sampling algorithms for our models are based on the idea of the recently developed Bayesian Conditional Density Filtering (BCDF) framework (Guhaniyogi et al., 2014). They key idea in BCDF is to process data in small minibatches, and maintain and update sufficient statistics of the model parameters with each new minibatch of the data. In our models, these sufficient statistics are the latent counts. We briefly outline the online Gibbs sampler for BPBFM-1 below:

Denoting I_t as indices of valid triplets in minibatch selected at iteration t, and I as the indices of all the valid triplets in training data. Define $\mathcal{X}_{ik\cdot\cdot}^{r,t} = \frac{|I|}{|I_t|} \sum_{k_2=1}^{K} \sum_{j=1,ij\in I_t}^{N} \mathcal{X}_{ikk_2j}^r$, $\mathcal{X}_{ik\cdot\cdot}^t = \frac{|I|}{|I_t|} \sum_{r=1}^{R} \mathcal{X}_{ik\cdot\cdot}^{r,t}$, and $\mathcal{X}_{\cdot k_1k_2\cdot}^{r,t} = \frac{|I|}{|I_t|} \sum_{i,j\in I_t}^{N} \mathcal{X}_{ik_1k_2j}^{r,t}$, where $|I|$ and I_t are cardinalities of the two sets. Then similar to batch Gibbs Sampling, define following quantities for $t \leq 2$: $\mathcal{X}_{ik\cdot\cdot}^{r,t} = (1-\rho)\mathcal{X}_{ik\cdot\cdot}^{r,t-1} + \rho\frac{|I|}{|I_t|} \sum_{k_2=1}^{K} \sum_{j=1,ij\in I_t}^{N} \mathcal{X}_{ikk_2j}^r$, $\mathcal{X}_{ik\cdot\cdot}^t = (1-\rho)\mathcal{X}_{ik\cdot\cdot}^{t-1} + \rho\frac{|I|}{|I_t|} \sum_{r=1}^{R} \mathcal{X}_{ik\cdot\cdot}^{rt}$, and $\mathcal{X}_{\cdot k_1k_2\cdot}^{r,t} = (1-\rho)\mathcal{X}_{\cdot k_1k_2\cdot}^{r,t-1} + \rho\frac{|I|}{|I_t|} \sum_{ij,ij\in I_t}^{N} \mathcal{X}_{ik_1k_2j}^{r,t}$. Here $\rho = (t + t_0)^{-w}$ is a decaying learning rate, as used in other online inference algorithms, such as stochastic variational inference (Hoffman et al., 2013). Here, $t_0 > 0$ and $w \in (0.5, 1]$ are required to guarantee convergence.

We omit the full details of the online Gibbs sampler here due to the lack of space. The Supplementary Material provides more details of the update equations for online inference for both BPBFM-1 and BPBFM-2.

4 RELATED WORK

There has been a significant amount of recent interest in the problem of learning from multi-relational data, both in the social/biological network analysis problems, as well as in modeling of large knowledge bases (such as YAGO, NELL, Freebase, etc.) that consist of a massive number of triplet-based facts of the form entity-relation-entity, involving very large number of entities and relations. Because a natural representation for multi-relational data is in form of a three-way tensor (or a collection of matrices), a number of methods, closely related to each other, such as those based on tensor decomposition, collective matrix factorization, and generalizations of stochastic blockmodels for multi-relational data, have been proposed for learning from such data (Nickel et al., 2011; Sutskever et al., 2009; Jenatton et al., 2012; Rai et al., 2014; Zhu, 2012).

Another class of methods, especially in the context of learning from large knowledge bases, involve learning the embeddings of entities and relations, and using these embeddings to predict the unknown facts (or "links") involving pairs of entities, given a specific relation. This class of methods typically includes (i) models such as those based on minimizing an energy function (Bordes et al., 2011; Socher et al., 2013; Bordes et al., 2014) involving valid entity-relation-entity triplets; and (ii) *translated embedding* based approaches (Bordes et al., 2013; Wang et al., 2014a; Yang et al., 2014) that embed the entities and the relations in a vector space based on the criteria that, for each valid triplet (h, r, t), the distance between the embeddings of the head and the tail, after the head entity embedding translated by the embedding of the relation, is smaller than the corresponding distance computed for an invalid triplet generated by perturbing either the head or the tail entity. We use some of these methods as baselines in our experiments.

There has been relatively little work on Bayesian methods for learning from large multi-relational data and knowledge bases (Sutskever et al., 2009; Rai et al., 2014; Zhu, 2012; Hu et al., 2015). Although some of the existing Bayesian methods, such as the Bayesian Clustered Tensor Factorization (Sutskever et al., 2009), do provide nice modeling flexibility (e.g., discovering clusters of entities and relations in addition to being applicable for tasks such as link prediction), these methods are not able to scale to large modern-day knowledge bases. In contrast, our proposed framework offers the various benefits of a generative, fully Bayesian model, in addition to being easily scalable for large multi-relational data sets, due to its dependence only on the positive triplets and the computational efficiency of the accompanying batch and online Gibbs sampling algorithms.

Finally, we would like to note that the Bernoulli-Poisson link for binary data has also been used recently in (Zhou, 2015). However, there are several key differences from our proposed framework: (1) the model in (Zhou, 2015) can only deal with a single relation type, whereas our framework allows learning from multi-relational networks and knowledge bases that consist of multiple types of relations; (2) in addition to handling multiple types of relations (BPBFM-1), our second model (BPBFM-2) allows further sharing of statistical strengths *across* multiple relations; and (3) while the model in (Zhou, 2015) relies on batch Gibbs sampling, we also develop *online* Gibbs sampling algorithms for both of our models, which allows us to apply these models to large-scale multi-relational data sets and knowledge bases.

5 EXPERIMENTS

We evaluate both of our proposed models **BPBFM-1** and BPBFM-2 on several benchmark data sets consisting of both moderate-sized multi-relational data sets, as well as large-scale benchmark knowledge bases (Section 5.1 provides more details of these data sets). In our experiments, we compare both our models (BPBFM-1 and BPBFM-2) with several state-of-the-art methods that include: (1) two bilinear latent factor models: **RESCAL** (Nickel et al., 2011) and **LFM** (Jenatton et al., 2012); (2) Bayesian logistic tensor factorization (Rai et al., 2014) (**BLTF**) for link-prediction from multi-relational data; (3) Zero-Truncated Poisson CP factorization (**ZTP-CP**) (Hu et al., 2015) (4) Embedding based methods - **TransE** (Bordes et al., 2013) and **TransH** (Wang et al., 2014a), both of which are recently proposed state-of-the-art methods for modeling large-scale knowledge bases. We report results on both quantitative comparisons (in terms of link-prediction/knowledge base completion accuracies), as well as qualitative analyses (e.g., analyzing clusters of entities/relations inferred by our models).

All our experiments were performed on a standard desktop with 24 GB RAM. In all our experiments, the hyperparameters β_0, β_1 and β were fixed to 1, , which worked well in practice.

5.1 Datasets

We use two groups of data sets in our experiments. The first group consists of three moderate-sized multi-relational data sets

- **Kinship:** This is a 104 × 104 × 26 binary tensor (Nickel et al., 2011) containing 26 types of relations among a set of 104 individuals.
- **Nations:** This is a $14 \times 14 \times 56$ dataset describes the relations among 14 countries with respect to 56 types of interactions.
- **UMLS:** This is a $135 \times 135 \times 49$ dataset describes the causal influence among 135 biomedical concepts with respect to 49 types of interactions.

The other group consists of three large knowledge bases, and includes Freebase-15K, Wordnet-100K, and NELL-50K. Table- 1 shows statistics of these data sets.

Table 1: Statistics of FB-15K, NELL-50K, and WN-100K

Datasets	FB-15K	NELL-50K	WN-100K
Entity #	14,951	29,904	38,696
Relation #	1345	233	11
Valid Triplet # (Train)	483,142	57365	112,581
Valid Triplet # (Test)	118,142	21,412	42,176

5.2 Experiments on Multi-Relational Data

In our first set of experiments, we evaluate both our models (using both batch as well as online inference) on the three multi-relational data sets (Kinship, UMLS, Nations) for the task of link-prediction, as well as for doing qualitative analyses (clustering entities and relations based on the topic-based embeddings inferred by our models).

5.2.1 Link Prediction and Computational Efficiency

We compare both our model with LFM, BLTF, and ZTPCP. For all methods, we set $K = 30$, and use 90%

Table 2: Most Prominent Entities in Topics Inferred from UMLS

Topic 1 (Group)	Topic 2 (Function)	Topic 3 (Chemical)	Topic 4 (Method/Procedure)
professional/occupational group	cell function	nucleic acid	molecular bio research tech
population group	organism function	steroid	human-caused phenomenon/process
age group	physiologic function	amino acid	laboratory procedure
group	molecular function	carbohydrate	diagnostic procedure
family group	organ/tissue function	lipid	laboratory/test result

Table 3: Most Prominent Relations in Topics Inferred From UMLS

Topic 1 (isa/part)	Topic 2 (diagnose/treat)	Topic 3 (Experiments)	Topic 4 (Adjacent/Surround)
part of	treats	manifestation of	adjacent to
isa	prevents	measurement of	connected to
issue in	diagnoses	evaluation of	surrounds
conceptual part of	complicates	indicates	traverses

Table 4: AUC Comparison for Multi-relational Data

Datasets	Kinship	UMLS	Nation
RESCAL (Nickel et al., 2011)	0.968	0.973	0.872
LFM (Jenatton et al., 2012)	**0.999**	0.991	0.836
BLTF (Rai et al., 2014)	0.983	0.988	0.856
ZTP-CP (Hu et al., 2015)	0.932	0.989	0.889
BPBFM-1 (Batch Gibbs)	0.971	0.988	0.882
BPBFM-1 (Online Gibbs)	0.975	0.988	0.891
BPBFM-2 (Batch Gibbs)	0.946	0.981	0.871
BPBFM-2 (Online Gibbs)	0.976	**0.994**	**0.896**

Table 5: Computational time comparison

Datasets	Kinship	UMLS	Nation
LFM (Jenatton et al., 2012)	1.8704	6.4563	0.0726
BPBFM-1 (Batch Gibbs)	0.1993	0.1433	0.0709
BPBFM-2 (Batch Gibbs)	0.7142	0.4005	0.1240

of valid triplets as training dataset and the remaining as testing. For LFM, we use default settings for other parameters as the code shared online [1]. For model 2, we set $M = 50$. All other parameters in our two models are randomly initialized; however smarter initializations of the embeddings can also be used. Receiver Operating Characteristic AUC (AUC) and Precision Recall AUC (AUC-PR) are used to evaluate the performance for link prediction. As shown in Table 4, except for Kinship, our models achieve comparable or better AUC as compared to the other methods.

We also compare per-iteration computation times of our model with LFM (both methods implemented in Matlab) on the three small datasets, as shown in Table 5. We do not report timings of other baselines because the implementations are not directly comparable (Matlab vs Python vs C). Since the computational cost of our model scales only in the number of nonzeros in the data, we gain maximum speed-up for UMLS and minimum speed-up for Nation dataset. This is consistent with the fraction of nonzero entries for three data sets (Kinship: 0.0384, UMLS: 0.0076, Nation: 0.1844).

5.2.2 Qualitative Analysis on UMLS Data

Since each column of the matrix $\mathbf{U}$ inferred by our models corresponds to a topic, we use the columns of $\mathbf{U}$ to rank most prominent entities in each topic (based on the magnitude of entries in that column), as shown in Table 2. Likewise, using the $M \times R$ non-negative matrix η inferred by BPBFM-2, we can group similar relation types by treating each row of η as a "topic" and sorting the entries in that row to rank the relations (Table 3 shows the top 4 relations for each topic).

5.3 Experiments on Large Knowledge Bases

In our second set of experiments, we evaluate our models on large knowledge bases (Freebase15K, Wordnet-100K, and NELL-50K) on two tasks: knowledge base completion (predicting the validility of held-out triplets) and qualitative analyses (grouping entities and relations using the topic based embeddings learned by our models, as we did previously for UMLS data).

5.3.1 Knowledge Base Completion

For this task, in Table 8, we compare BPBFM-1 with two state-of-the-art knowledge base embedding methods - TransE (Bordes et al., 2013) and TransH (Wang et al., 2014a). We also provide, in Table 9, a separate comparison between BPBFM-1 and BPBFM-2 to discuss the benefits of using BPBFM-2 which is able to share information across the different relations.

For the comparison between BPBFM-1 and BPBFM-2 (Table 9), for all three data sets, we set $K = 10$ for both models. For BPBFM-2, we set $M = 5$ for WN-100K since it has only 11 relations, and $M = 80$ for Freebase-15K and NELL-50 as the number of relations is much larger (Freebase-15K has 1345 relations and NELL-50K has 233 relations). As shown in Table 9, BPBFM-2 outperforms BPBFM-1, even if K is as small as 10.

5.4 Qualitative Analyses on Freebase15K, NELL-50K, and WN-100K

The topic-based non-negative embeddings learned by our model can be useful for qualitative analyses. To illustrate this, we show results of our qualitative analyses on Freebase15K, NELL-50K, and WN-100K. Six of the factors (each factor represents a topic) inferred by BPBFM-2 are presented in table 6, and for each topic, we show top-8 entities. As the larger datasets contain more and richer entities than the smaller dataset, we

[1] `http://tinyurl.com/q6a66ro`

Table 6: Most Prominent Entities in Topics Inferred for FB-15K, WN-100K and NELL-50K

FB-15K		WN-100K		NELL-50K	
Topic 1 (Biology)	Topic 2 (Country)	Topic 3 (Film)	Topic 4 (Position)	Topic 5 (Football)	Topic 6 (Band)
animal kingdom	britain	valentine's day	midfielder	washington redskins	zeppelin
worm genus	america	harry potter	forward(football)	eagles	beatles
edible nut	france	new york stories	defender	colts	dream theater
family	emerald isle	love actually	goalkeeper	seahawks	poison
anacardiaceae	japan	grindhouse	winger	dallas cowboys	ramones
bird footed dinosaur	canada	terror in the aisles	head coach	packers	iron maiden
accipitridae	italia	who framed roger rabbit	forward(hockey)	oakland raiders	blondie
aschelminthes	deutschland	Om Shanti Om	infielder	bills	rush

Table 7: Most Prominent Relations in Topics Inferred for FB-15K

Topic 1 (Education)	Topic 2 (Film/Award)	Topic 3 (Family Relation)	Topic 4 (Individual Info)	Topic 5 (Sports)
institution	film	sibling	nationality	/sports/team
degree	actor	split to	location	/sports/position
major field of study	nominated for	parents	gender	/football/positions
student	award nominee	child	place of birth	/football/position
specialization	honored for	children	cause of death	/football/team

can see a diverse set of factors, such as biology, countries, films, sports and musical bands, and all entities in each factor seem to be closely related to each other.

Table 8: AUC-PR for Knowledge Bases

Datasets	FB-500K	NELL-50K	WN-100K
TransE	0.645	0.623	0.674
TransH	0.744	0.681	0.613
BPBFM-1	**0.780**	**0.774**	**0.681**

Table 9: AUC and AUC-PR (A-PR below) comparison between BPBFM-1 and BPBFM-2

Datasets	FB-500K		NELL-50K		WN-100K	
	AUC	A-PR	AUC	A-PR	AUC	A-PR
BPBFM-1	0.724	0.648	0.742	0.793	0.648	0.622
BPBFM-2	**0.727**	**0.665**	**0.783**	**0.806**	**0.725**	**0.717**

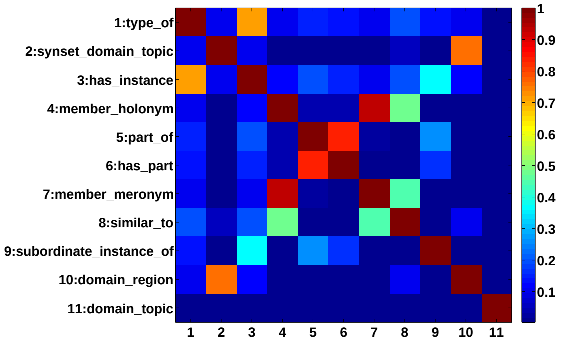

Figure 4: Similarities between Relations in Wordnet

In addition, to indicate BPBFM-2's ability in learning the prominent relations in different topics, we present five topics of relations, and for each topic show top-5 relations. The results are shown in Table 7. In factor 1, we find that these relations are all about education, and the other relations in the other four topics make sense as well.

We also show the inferred similarities between different relations by treating each column of matrix η as the feature vector for the corresponding relation. As WN-100K has only 11 relations, it is an ideal dataset to intuitively show relation similarities via a correlation plot. The consine similarity among relations are shown in figure 4. In the plot three pairs of relations, {type_of, has_instance}, {part_of, has_part}, and {member_holonym, member_meronym}, are found similar to each other, which makes a lot sense because if we switch the head and tail in a triplet{head, relation, tail}, the two relations in each pair are basically the same relation.

Figure 4 the pairwise similarities between the 11 relations present in the Wordnet data as computed using the inferred η matrix by BPBFM-2. As shown in the plot, the model is able to correctly infer the similarities between the relations.

6 CONCLUSION

We have presented a scalable and fully Bayesian bilinear non-negative latent factor model to analyze large multi-relational data. A rich generative modeling framework enables to our models to not just learn embeddings of entities and relations and perform tasks such as link-prediction and knowledge-based completion, but also gives interpretable results for further qualitative analyses. In particular, the topic-based embeddings learned by our models can be useful in itself, e.g., for grouping entities and/or relations in terms of the topics they represent. Computational cost that scales w.r.t. the number of positive triplets makes our framework an ideal choice for learning from real-world multi-relations data that are massive (in terms of number of entities and relations) yet have very few positive triplets. Our framework can be extended in several directions; for example, allowing new entities and/or relations to be added; adding a temporal dimension (e.g., a fact may be a true over a period of time but not forever (Dong et al., 2014)); or incorporating other sources of information, e.g., a text corpus in addition to the knowledge base (Wang et al., 2014b).

ACKNOWLEDGMENTS

This research was supported in part by ARO, DARPA, DOE, NGA and ONR.

References

K. Bollacker, C. Evans, P. Paritosh, T. Sturge, and J. Taylor. Freebase: a collaboratively created graph database for structuring human knowledge. In *SIGMOD*, 2008.

A. Bordes, J. Weston, R. Collobert, and Y. Bengio. Learning structured embeddings of knowledge bases. In *AAAI*, 2011.

A. Bordes, N. Usunier, A. Garcia-Duran, J. Weston, and O. Yakhnenko. Translating embeddings for modeling multi-relational data. In *NIPS*, 2013.

A. Bordes, X. Glorot, J. Weston, and Y. Bengio. A semantic matching energy function for learning with multi-relational data. *Machine Learning*, 2014.

A. Carlson, J. Betteridge, B. Kisiel, B. Settles, E. R. Hruschka Jr, and T. M. Mitchell. Toward an architecture for never-ending language learning. In *AAAI*, 2010.

D. Collett. *Modelling binary data*. CRC press, 2002.

X. Dong, E. Gabrilovich, G. Heitz, W. Horn, N. Lao, K. Murphy, T. Strohmann, S. Sun, and W. Zhang. Knowledge vault: A web-scale approach to probabilistic knowledge fusion. In *KDD*, 2014.

D. B. Dunson and A. H. Herring. Bayesian latent variable models for mixed discrete outcomes. *Biostatistics*, 2005.

A. Goldenberg, A. X. Zheng, S. E. Fienberg, and E. M. Airoldi. A survey of statistical network models. *Foundations and Trends® in Machine Learning*, 2 (2), 2010.

R. Guhaniyogi, S. Qamar, and D. B. Dunson. Bayesian conditional density filtering. *arXiv preprint arXiv:1401.3632*, 2014.

M. D. Hoffman, D. M. Blei, C. Wang, and J. Paisley. Stochastic variational inference. *JMLR*, 2013.

C. Hu, P. Rai, and L. Carin. Zero-truncated poisson tensor factorization for massive binary tensors. In *UAI*, 2015.

R. Jenatton, N. L. Roux, A. Bordes, and G. R. Obozinski. A latent factor model for highly multi-relational data. In *NIPS*, 2012.

T. G. Kolda and B. W. Bader. Tensor decompositions and applications. *SIAM review*, 51(3):455–500, 2009.

M. Nickel, V. Tresp, and H.-P. Kriegel. A three-way model for collective learning on multi-relational data. In *ICML*, 2011.

M. Nickel, K. Murphy, V. Tresp, and E. Gabrilovich. A review of relational machine learning for knowledge graphs: From multi-relational link prediction to automated knowledge graph construction. *arXiv preprint arXiv:1503.00759*, 2015.

W. W. Piegorsch. Complementary log regression for generalized linear models. *The American Statistician*, 1992.

P. Rai, Y. Wang, S. Guo, G. Chen, D. Dunson, and L. Carin. Scalable bayesian low-rank decomposition of incomplete multiway tensors. In *ICML*, 2014.

R. Socher, D. Chen, C. D. Manning, and A. Ng. Reasoning with neural tensor networks for knowledge base completion. In *NIPS*, 2013.

F. M. Suchanek, G. Kasneci, and G. Weikum. YAGO: a core of semantic knowledge. In *WWW*, 2007.

I. Sutskever, J. B. Tenenbaum, and R. R. Salakhutdinov. Modelling relational data using bayesian clustered tensor factorization. In *NIPS*, 2009.

Z. Wang, J. Zhang, J. Feng, and Z. Chen. Knowledge graph embedding by translating on hyperplanes. In *AAAI*, 2014a.

Z. Wang, J. Zhang, J. Feng, and Z. Chen. Knowledge graph and text jointly embedding. In *EMNLP*, 2014b.

B. Yang, W.-t. Yih, X. He, J. Gao, and L. Deng. Embedding entities and relations for learning and inference in knowledge bases. *arXiv preprint arXiv:1412.6575*, 2014.

M. Zhou. Infinite edge partition models for overlapping community detection and link prediction. In *AISTATS*, 2015.

M. Zhou, L. A. Hannah, D. Dunson, and L. Carin. Beta-negative binomial process and poisson factor analysis. In *AISTATS*, 2012.

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

Supplementary: Scalable Bilinear Non-negative Latent Factor Models for Multi-Relational Data

1 Gibbs Sampling for BPBFM-2

Just as we did for BPBFM-1, we can express each latent count $\mathcal{X}_{ij}^r$ in BPBFM-2 (which models each Λ^r as $\Lambda^r = \sum_{m=1}^M \eta_{mr} G^m$) as a sum of the following form: $\mathcal{X}_{ij}^r = \sum_{k_1}^K \sum_{k_2}^K \sum_m^M X_{ik_1k_2mj}^r$ where $\mathcal{X}_{ik_1k_2mj}^r \sim \text{Poisson}(u_{ik_1} \eta_{mr} G_{k_1k_2}^m u_{jk_2})$. We further define $\mathcal{X}_{\cdot k_1k_2m\cdot}^{\cdot} = \sum_i^N \sum_j^N \sum_{r=1}^R \mathcal{X}_{ik_1k_2mj}^r$ and $\mathcal{X}_{\cdots m\cdot}^r = \sum_i^N \sum_j^N \sum_{k_1=1}^K \sum_{k_2=1}^K \mathcal{X}_{ik_1k_2mj}^r$, using additive property of Poisson distribution

$$\mathcal{X}_{\cdot k_1k_2m\cdot}^{\cdot} \sim \text{Poisson}(\theta_{k_1k_2} G_{k_1k_2}^m \sum_{r=1}^R \eta_{mr}) \quad (1)$$

$$\mathcal{X}_{\cdots m\cdot}^r \sim \text{Poisson}(\eta_{mr} \sum_{k_1=1}^K \sum_{k_2=1}^K \theta_{k_1k_2} G_{k_1k_2}^m) \quad (2)$$

With these defined, we proceed to give the update equations for the Gibbs sampler for BPBFM-2.

Sampling $\mathcal{X}_{ij}^r$: X_{ij}^r is sampled just as in model-1.

Sampling $\mathcal{X}_{ik_1k_2mj}^r$: $\mathcal{X}_{ik_1k_2mj}^r$ can be sampled as

$$\mathcal{X}_{ik_1k_2j}^r \sim \text{Mult}(\mathcal{X}_{ij}^r; \frac{u_{ik_1} \eta_{mr} G_{k_1k_2}^m u_{jk_2}}{\sum_{k_1=1}^K \sum_{k_2=1}^K u_{ik_1} \Lambda_{k_1k_2}^r u_{jk_2})} \quad (3)$$

Sampling $\mathbf{U}_{:,k}$: Using Dirichlet-multinomial conjugacy, each column of $\mathbf{U}$ can be sampled as

$$\mathbf{U}_{:,k} \sim \text{Dir}(a + \mathcal{X}_{1k\cdots}^{\cdot}, a + \mathcal{X}_{2k\cdots}^{\cdot}, \ldots, a + \mathcal{X}_{Nk\cdots}^{\cdot}) \quad (4)$$

where $\mathcal{X}_{ik\cdots}^{\cdot} = \sum_{k_2=1}^K \sum_{j=1}^N \sum_{m=1}^M \sum_{r=1}^R \mathcal{X}_{ikk_2mj}^r$.

Sampling d_m^m:. Marginalizing out $G_{k_1k_2}^m$ from Eq.(1), we have

$$\mathcal{X}_{\cdot k_1k_2m\cdot}^{\cdot} \sim \text{NegBin}((\epsilon^m)^{\delta_{k_1k_2}} d_{k_1}^m (d_{k_2}^m)^{1-\delta_{k_1k_2}}, p_{k_1k_2}) \quad (5)$$

where $p_{k_1k_2} = \frac{\theta_{k_1k_2} \sum_{r=1}^R \eta_{mr}}{\theta_{k_1k_2} \sum_{r=1}^R \eta_{mr} + \beta}$. Using the data augmentation scheme proposed we used for BPBFM-1, d_k^m can be sampled by first sampling

$$\ell_{kk_2}^m \sim \sum_{t=1}^{X_{\cdot k_1k_2m\cdot}^{\cdot}} \text{Bern}(\frac{(\epsilon^m)^{\delta_{k_1k_2}} d_{k_1}^m (d_{k_2}^m)^{1-\delta_{k_1k_2}}}{(\epsilon^m)^{\delta_{k_1k_2}} d_{k_1}^m (d_{k_2}^m)^{1-\delta_{k_1k_2}} + t - 1}) \quad (6)$$

and then sampling

$$d^m \sim \quad (7)$$
$$\text{Ga}(\frac{\gamma_0}{K} + \sum_{k_2}^K \ell_{kk_2}^m, \frac{1}{c_0 - \sum_{k_2}^K (\epsilon^m)^{\delta_{kk_2}} (d_{k_2}^m)^{1-\delta_{kk_2}} \ln(1 - p_{kk_2})})$$

Sampling ϵ^m: ϵ^m can be sampled as

$$\epsilon^r \sim \text{Ga}(e_0 + \sum_k^K \ell_{kk}^r, \frac{1}{f_0 - \sum_k^K d_k^m \ln(1 - p_{kk})}) \quad (8)$$

Sampling $G_{k_1k_2}^m$: Using Gamma-Poisson conjugacy, $G_{k_1k_2}^m$ can be sampled by

$$G_{k_1k_2}^m \sim \quad (9)$$
$$\text{Ga}((\epsilon^m)^{\delta_{k_1k_2}} d_{k_1}^r (d_{k_2}^r)^{1-\delta_{k_1k_2}} + X_{\cdot k_1k_2m\cdot}^{\cdot}, \frac{1}{\beta + \theta_{k_1k_2} \sum_{r=1}^R \eta_{mr}})$$

Sampling η_{mr}: Using equation (2) and Gamma-Poisson conjugacy, η_{mr} can be sampled by

$$\eta_{mr} \sim \text{Ga}(h_0 + \mathcal{X}_{\cdots m\cdot}^r, \frac{1}{q_0 + \sum_{k_1=1}^K \sum_{k_2=1}^K \theta_{k_1k_2} G_{k_1k_2}^m}) \quad (10)$$

2 Online Gibbs Sampling

In this section, we provide the details of the online Gibbs sampling algorithms for both of our models. Our online Gibbs sampling algorithms are based on the idea of the recently developed Bayesian Conditional Density Filtering (BCDF) framework (Guhaniyogi et al., 2014). They key idea in BCDF is to process data in small minibatches, and maintain and update sufficient statistics of the model parameters with each new minibatch of the data. In our models, these sufficient statistics are the latent counts.

2.1 Online Gibbs Sampling for BPBFM-1

Denoting I_t as indices of valid triplets in minibatch selected at iteration t, and I as the indices of all the valid triplets in training data. Define $\mathcal{X}_{ik\cdot\cdot}^{r,t} = \frac{|I|}{|I_t|} \sum_{k_2=1}^K \sum_{j=1,ij \in I_t}^N \mathcal{X}_{ikk_2j}^r$, $\mathcal{X}_{ik\cdot\cdot}^t = \frac{|I|}{|I_t|} \sum_{r=1}^R \mathcal{X}_{ik\cdot\cdot}^{r,t}$, and $\mathcal{X}_{\cdot k_1k_2\cdot}^{r,t} = \frac{|I|}{|I_t|} \sum_{i,j \in I_t}^N \mathcal{X}_{ik_1k_2j}^{r,t}$, where $|I|$ and I_t are cardinalities of the two sets. Then similar to batch Gibbs Sampling, define following quantities for $t \leq 2$: $\mathcal{X}_{ik\cdot\cdot}^{r,t} = (1-\rho)\mathcal{X}_{ik\cdot\cdot}^{r,t-1} + \rho \frac{|I|}{|I_t|} \sum_{k_2=1}^K \sum_{j=1,ij \in I_t}^N \mathcal{X}_{ikk_2j}^r$, $\mathcal{X}_{ik\cdot\cdot}^t = (1-\rho)\mathcal{X}_{ik\cdot\cdot}^{t-1} + \rho \frac{|I|}{|I_t|} \sum_{r=1}^R \mathcal{X}_{ik\cdot\cdot}^{rt}$, and $\mathcal{X}_{\cdot k_1k_2\cdot}^{r,t} = (1-\rho)\mathcal{X}_{\cdot k_1k_2\cdot}^{r,t-1} + \rho \frac{|I|}{|I_t|} \sum_{ij,ij \in I_t}^N \mathcal{X}_{ik_1k_2j}^{r,t}$. Here $\rho = (t + t_0)^{-w}$ is a decaying learning rate, as used in other online inference algorithms, such as stochastic variational inference (Hoffman et al., 2013). Here, $t_0 > 0$ and $w \in (0.5, 1]$ are required to guarantee convergence. With these defined, online Gibbs sampling at iteration t proceeds as:

Sampling $\mathbf{U}_{:,k}$: Each column of $\mathbf{U}$ can be sampled as

$$\mathbf{U}_{:,k} \sim \text{Dir}(a + \mathcal{X}^t_{1k\cdot\cdot}, a + \mathcal{X}^t_{2k\cdot\cdot}, \ldots, a + \mathcal{X}^t_{Nk\cdot\cdot}) \quad (11)$$

Sampling d_k^r: d_k^r can be sampled by first sampling

$$\ell^r_{kk_2} \sim \sum_{t=1}^{X^{r,t}_{\cdot k_1 k_2 \cdot}} \text{Bern}(\frac{(\epsilon^r)^{\delta_{k_1k_2}} d^r_{k_1} (d^r_{k_2})^{1-\delta_{k_1k_2}}}{(\epsilon^r)^{\delta_{k_1k_2}} d^r_{k_1} (d^r_{k_2})^{1-\delta_{k_1k_2}} + t - 1}) \quad (12)$$

and then sampling

$$d_k^r \sim \quad (13)$$
$$\text{Ga}(\frac{\gamma_0}{K} + \sum_{k_2}^{K} \ell^r_{kk_2}, \frac{1}{c_0 - \sum_{k_2}^K (\epsilon^r)^{\delta_{kk_2}} (d^r_{k_2})^{1-\delta_{kk_2}} \ln(1 - p_{kk_2})})$$

Sampling ϵ^r: ϵ^r can be sampled as

$$\epsilon^r \sim \text{Ga}(e_0 + \sum_{k}^{K} \ell^r_{kk}, \frac{1}{f_0 - \sum_k^K d_k^r \ln(1 - p_{kk})}) \quad (14)$$

Sampling $\Lambda^r_{k_1k_2}$: $\Lambda^r_{k_1k_2}$ can be sampled by

$$\Lambda^r_{k_1k_2} \sim \quad (15)$$
$$\text{Ga}((\epsilon^r)^{\delta_{k_1k_2}} d^r_{k_1} (d^r_{k_2})^{1-\delta_{k_1k_2}} + \mathcal{X}^{r,t}_{\cdot k_1k_2\cdot}, \frac{1}{\beta + \theta_{k_1k_2}})$$

$\mathcal{X}^r_{ij}$, ϵ^r and $\mathcal{X}^r_{ik_1k_2j}$ are sampled the same way as the batch Gibbs sampling.

2.2 Online Gibbs Sampling for BPBFM-2

Similar to online BPBFM-1, we define $\mathcal{X}^{\cdot,t}_{\cdot k_1k_2m\cdot} = (1-\rho)\mathcal{X}^{\cdot,t-1}_{\cdot k_1k_2m\cdot} + \rho\frac{|I|}{|I_t|}\sum_{ij,ij\in I_t}\sum_{r=1}^R \mathcal{X}^r_{ik_1k_2mj}$, $\mathcal{X}^{r,t}_{\cdots m\cdot} = (1-\rho)\mathcal{X}^{r,t-1}_{\cdots m\cdot} + \rho\frac{|I|}{|I_t|}\sum_{ij,ij\in I_t}\sum_{k_1=1}^K\sum_{k_2=1}^K \mathcal{X}^r_{ik_1k_2mj}$, and $\mathcal{X}^{\cdot,t}_{ik\cdots} = (1-\rho)\mathcal{X}^{\cdot,t-1}ik\cdots + \rho\frac{|I|}{|I_t|}\sum_{k_2=1}^K\sum_{j=1,ij\in I_t}^N\sum_{m=1}^M\sum_{r=1}^R \mathcal{X}^r_{ikk_2mj}$. With these defined, we proceed to give the update equations for the online Gibbs sampler for model-2:

Sampling $\mathbf{U}_{:,k}$: Each column of $\mathbf{U}$ can be sampled as

$$\mathbf{U}_{:,k} \sim \text{Dir}(a + \mathcal{X}^{\cdot,t}_{1k\cdots}, a + \mathcal{X}^{\cdot,t}_{2k\cdots}, \ldots, a + \mathcal{X}^{\cdot,t}_{Nk\cdots}) \quad (16)$$

Sampling d_m^m: d_k^m can be sampled by first sampling

$$\ell^m_{kk_2} \sim \sum_{t=1}^{X^{\cdot,t}_{\cdot k_1k_2m\cdot}} \text{Bern}(\frac{(\epsilon^m)^{\delta_{k_1k_2}} d^m_{k_1} (d^m_{k_2})^{1-\delta_{k_1k_2}}}{(\epsilon^m)^{\delta_{k_1k_2}} d^m_{k_1} (d^m_{k_2})^{1-\delta_{k_1k_2}} + t - 1}) \quad (17)$$

and then sampling

$$d_k^m \sim \quad (18)$$
$$\text{Ga}(\frac{\gamma_0}{K} + \sum_{k_2}^{K} \ell^m_{kk_2}, \frac{1}{c_0 - \sum_{k_2}^K (\epsilon^m)^{\delta_{kk_2}} (d^m_{k_2})^{1-\delta_{kk_2}} \ln(1 - p_{kk_2})})$$

Sampling $G^m_{k_1k_2}$: Using Gamma-Poisson conjugacy, $G^m_{k_1k_2}$ can be sampled by

$$G^m_{k_1k_2} \sim \quad (19)$$
$$\text{Ga}((\epsilon^m)^{\delta_{k_1k_2}} d^r_{k_1} (d^r_{k_2})^{1-\delta_{k_1k_2}} + X^{\cdot,t}_{\cdot k_1k_2m\cdot}, \frac{1}{\beta + \theta_{k_1k_2}\sum_{r=1}^R \eta_{mr}})$$

Sampling η_{mr}: η_{mr} can be sampled by

$$\eta_{mr} \sim \text{Ga}(h_0 + \mathcal{X}^{r,t}_{\cdots m\cdot}, \frac{1}{q_0 + \sum_{k_1=1}^K \sum_{k_2=1}^K \theta_{k_1k_2} G^m_{k_1k_2}}) \quad (20)$$

$\mathcal{X}^r_{ij}$, $\mathcal{X}^r_{ik_1k_2mj}$, ϵ^m ϵ^m can be sampled the same way as the batch Gibbs sampling.

References

R. Guhaniyogi, S. Qamar, and D. B. Dunson. Bayesian conditional density filtering. *arXiv preprint arXiv:1401.3632*, 2014.

M. D. Hoffman, D. M. Blei, C. Wang, and J. Paisley. Stochastic variational inference. *JMLR*, 2013.