Non-negative Matrix Factorization for Discrete Data with Hierarchical Side-Information

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 probabilistic framework for efficient non-negative matrix factorization of discrete (count/binary) data with side-information. The side-information is given as a *multi-level* structure, taxonomy, or ontology, with nodes at each level being *categorical-valued* observations. For example, when modeling documents with a two-level side-information (documents being at level-zero), level-one may represent (one or more) authors associated with each document and level-two may represent affiliations of each author. The model easily generalizes to more than two levels (or taxonomy/ontology of arbitrary depth). Our model can learn embeddings of entities present at each level in the data/side-information hierarchy (e.g., documents, authors, affiliations, in the previous example), with appropriate sharing of information across levels. The model also enjoys full local conjugacy, facilitating efficient Gibbs sampling for model inference. Inference cost scales in the number of non-zero entries in the data matrix, which is especially appealing for real-world massive but sparse matrices. We demonstrate the effectiveness of the model on several real-world data sets.

1 INTRODUCTION

Non-negative matrix factorization for discrete data is a fundamental problem in many applications, such as text modeling (Zhou et al., 2012), social network modeling (Yang and Leskovec, 2013), recommender systems (Gopalan et al., 2015), and so on. Often, in addition to the matrix being factorized, there is side-information available along the rows and/or columns, that can be leveraged to handle issues such as data sparsity, the cold-start problem, etc. Several attempts have been made in the recent past (Agarwal and Chen, 2009; Kim et al., 2012; Gopalan et al., 2014; Chaney et al., 2015) to incorporate such side-information when it is given in form *flat-structured* feature vectors/covariates along the rows and/or the columns of the data matrix. In many problems, however, the side-information can more naturally be specified in form of a *hierarchy*, with each node in the hierarchy being a categorical-valued observation. See Fig. 1 for some examples where the side-information is in form of a hierarchy or ontology of categorical-valued observations. Although data exhibiting such structure are prevalent in many applications, existing matrix factorization models cannot properly leverage such forms of side-information arranged in form of multiple layers.

We present a generative Bayesian framework that allows us to leverage such *structural* (e.g., specified hierarchically or via a taxonomy) side-information in the context of non-negative matrix factorization of *discrete* data. Moreover, the proposed framework can handle count as well as binary matrices in a unified manner. In addition to being useful for standard tasks such as matrix completion for count/binary data, our framework can also be used for topic modeling, while leveraging the available side-information. Another appealing aspect of our framework is that, in addition to learning the embeddings for the rows and columns of the data matrix, it can also learn embeddings of the nodes present in the structure that forms the side-information; e.g., for the two examples shown in Fig. 1, our model can learn the embeddings of documents and words, as well as can learn the embeddings for the entities that constitute the side-information - authors and affiliations in Fig 1 (left) and each of the nodes in the label taxonomy in Fig 1 (right). These interpretable embeddings can be useful in other tasks, such as clustering and classification, or for topic modeling at

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.

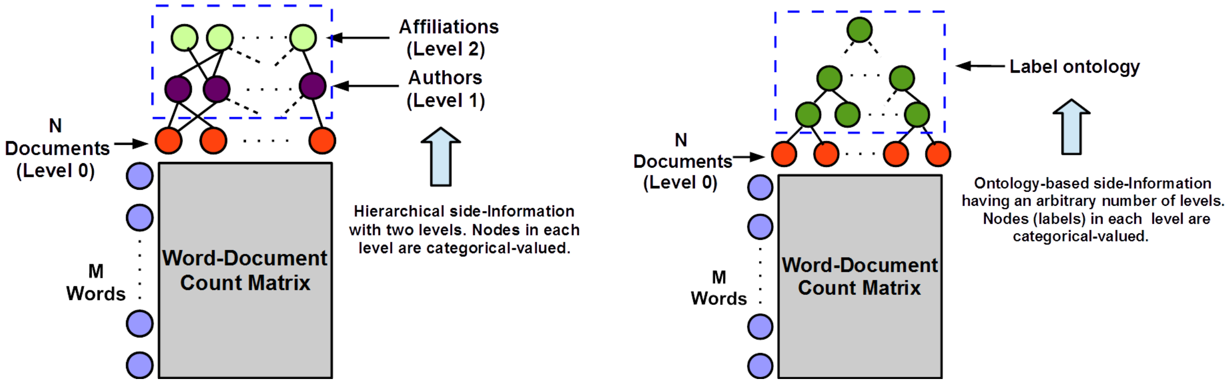
Affiliations (Level 2)
Authors (Level 1)
N Documents (Level 0)
M Words
Word-Document Count Matrix
Hierarchical side-Information with two levels. Nodes in each level are categorical-valued.
Label ontology
N Documents (Level 0)
M Words
Word-Document Count Matrix
Ontology-based side-Information having an arbitrary number of levels. Nodes (labels) in each level are categorical-valued.

Figure 1: Two examples of the type of side-information that our proposed framework can leverage, **Left:** Side-information specified in form of a multi-layer hierarchy with bipartite connections between nodes in adjacent layers. **Right**: Side-information specified in form of an ontology over known labels. In this case, each document is associated with a single categorical label and these labels are the leaf nodes of a label ontology.

multiple resolutions, which significantly emhances the versatility and usefulness of our framework for applications beyond matrix factorization and completion.

Our framework also enjoys full local conjugacy which facilitates closed-form Gibbs sampling for all the model parameters. Moreover, inference in our model (for both count as well as binary matrix case) scales in the number of nonzeros in the data matrix, which makes it scale easily to massive but sparse matrices.

2 THE MODEL

Here, we will present the model description assuming that the side-information is given with a hierarchy or ontology with two levels; the model can be easily modified to work with arbitrary number of levels.

We assume that we are given a data matrix $\mathbf{X}$ of size $M \times N$, where each column of $\mathbf{X}$ represents an object (e.g., a document). Fig. 1 shows two examples where the observations in $\mathbf{X}$ are count-valued (e.g., word counts for documents). The case when the observations in $\mathbf{X}$ are binary will be discussed subsequently in Sec. 2.4. The side-information for the objects is provided in form of a multi-level structures, such as a hierarchy (Fig. 1-left) or an ontology (Fig. 1-right).

2.1 Background

In the absence of any side-information, the counts matrix $\mathbf{X} \in \mathbb{Z}^{M\times N}$ can be modeled via a Poisson factor analysis (PFA) model as $\mathbf{X} \sim \text{Pois}(\mathbf{U}\mathbf{V}^\top)$ where $\mathbf{U}$ and $\mathbf{V}$ are positive-valued matrices of size $M \times R$ and $N \times R$, respectively, and R denotes the number of latent factors. This construction is also equivalent to assuming that each entry x_{mn} in $\mathbf{X}$ can be written as a sum of R *latent* counts (Dunson and Herring, 2005):

$$x_{mn} = \sum_{r=1}^{R} x_{mnr}, \quad x_{mnr} \sim \text{Pois}(u_{mr}v_{nr}) \quad (1)$$

$$\boldsymbol{u}_{:r} \sim \text{Dir}(\alpha, \ldots, \alpha) \quad (2)$$

$$v_{nr} \sim \text{Ga}\left(g_r, \frac{q_r}{1-q_r}\right) \quad (3)$$

$$g_r \sim \text{Ga}(c_0 g_0, 1/h_0) \quad (4)$$

$$q_r \sim \text{Beta}(c\epsilon, c(1-\epsilon)) \quad (5)$$

Note that each Dirichlet drawn column $\boldsymbol{u}_{:r}$ of $\mathbf{U}$ represents a *distribution* (i.e., a "topic") over the M objects (e.g., words) along the rows of $\mathbf{X}$. Also note that the Poisson-gamma construction (Eq. 1–3) is equivalent to a gamma-negative binomial model (Zhou et al., 2012) for each entry x_{mn} of $\mathbf{X}$.

2.2 Leveraging Multi-Level Side-Information

We would like to leverage the multi-level side-information available for the columns of $\mathbf{X}$ (as shown in Fig. 1). To accomplish this, we augment the PFA generative model using a multi-level conditioning structure imposed on the $N \times R$ factor score matrix $\mathbf{V}$, whose each row $\boldsymbol{v}_n = [v_{n1}, \ldots, v_{nR}]$ denotes the factor scores (or *embedding*) of a level-zero object n.

In particular, to leverage the side-information (i.e., from level-one and above), we first model the r^{th} factor score of object n as a sum of contributions from each of the level-one nodes associated with this object

$$v_{nr} = \sum_{l\in\mathcal{L}_n^{(1)}} v_{nrl} \quad (6)$$

$$v_{nrl} \sim \text{Ga}(g_{lr}, q_r/(1-q_r)) \quad (7)$$

where $\mathcal{L}^{(1)}$ denotes the set of *all* nodes in level-one in the hierarchy and $\mathcal{L}_n^{(1)}$ denotes the subset of these nodes associated with object n from level-zero.

Using gamma-additivity, Eq. 6-7 can be combined as

$$v_{nr} \sim \text{Ga}\left(\sum_{l\in\mathcal{L}_n^{(1)}} g_{lr}, \frac{q_r}{1-q_r}\right) \quad (8)$$

In Eq. 8, g_{lr} denotes the r^{th} factor score of node l at level-one (first level of side-information).

To leverage the level-two side-information, we likewise assume that the factor scores of this level-one node l can, in turn, be written as a sum of contributions from each of the level-two nodes it is associated with:

$$g_{lr} = \sum_{p \in \mathcal{L}_l^{(2)}} g_{lrp} \quad (9)$$

$$g_{lrp} \sim \text{Ga}(h_{pr}, 1/\beta_0) \quad (10)$$

$$h_{pr} \sim \text{Ga}(s, 1/\beta_1) \quad (11)$$

where $\mathcal{L}^{(2)}$ denotes the set of all nodes in level-two of the side-information hierarchy and $\mathcal{L}_l^{(2)}$ denotes the subset of these nodes associated with node l in level-one. Note that Eq. 9-10 can also be combined as

$$g_{lr} \sim \text{Ga}(\sum_{p \in \mathcal{L}_l^{(2)}} h_{pr}, 1/\beta_0) \quad (12)$$

In Eq. 12, h_{pr} denotes the r^{th} factor score of node p at level-two (second level of side-information). Subsequently, we will refer to our model as **PFA-SSI**, as an abbreviation for **P**oisson **F**actor **A**nalysis with **S**tructural **S**ide-**I**nformation.

2.3 Learning Multi-Level Embeddings

Our generative model provides a natural and effective way of learning embeddings of the objects being modeled (e.g., the documents) as well as the embeddings of the nodes that together constitute the multi-level side-information (e.g., the authors and affiliations or the label ontology as shown in Fig. 1). To see this, note that $\boldsymbol{v}_n = [v_{n1}, \dots, v_{nR}]$, $\boldsymbol{g}_l = [g_{l1}, \dots, g_{lR}]$, and $\boldsymbol{h}_p = [h_{p1}, \dots, h_{pR}]$ can be interpreted as *embeddings* of the n^{th} level-zero object, and the l^{th} level-one node and the p^{th} level-two node in the multi-level side-information, respectively. Note that all these embeddings are in the *same* R-dimensional space and hence are "comparable". Since in our model the embeddings correspond to topics, the embeddings allow us to discover the topics associated with each object as well as the topics associated with each constituent node of the side-information. For example, if the side-information is given in form of a label ontology then our model can infer the embedding of each label in the ontology and the topics associated with each label. Such a property makes our framework readily applicable for tasks such as: (1) *supervised* topic modeling (Ramage et al., 2009; Rabinovich and Blei, 2014) with *multi-level supervision*, which most of existing methods are unable to leverage in a proper way (also see Sec. 4 on Related Work); and (2) assigning labels to unlabeled (i.e., test) objects by inferring the embeddings of these objects, using the dictionary $\mathbf{U}$ learned from the labeled training data, applying a standard PFA with dictionary fixed as $\mathbf{U}$, and finding the most similar labels by comparing these inferred embeddings with the embeddings of the set of labels in the training data. Note that, if the side-information is given as a tree/ontology over labels, such an approach would even allow labeling a test object with a non-leaf label, even though the training set objects may only have leaf node labels (somewhat mimicking a zero-shot learning setting).

2.4 Modeling Binary X

If the matrix $\mathbf{X}$ is binary, we can replace the Poisson likelihood for the counts with a Bernoulli-Poisson likelihood for binary data. The Bernoulli-Poisson model Zhou (2015) is based on first drawing a count-valued latent variable from a Poisson and thresholding it at one to generate the binary observation. In our model, this amounts to the following generative model for each binary entry x_{mn} in $\mathbf{X}$

$$x_{mn} = 1(z_{mn} \geq 1), \quad z_{mn} \sim \text{Pois}(\sum_{r=1}^{R} u_{mr} v_{nr}) \quad (13)$$

The rest of the generative model is the same as when $\mathbf{X}$ is count-valued (as described in earlier sections). Marginalizing out z_{mn} leads to the following

$$x_{mn} \sim \text{Ber}\left(1 - \exp(-\sum_{r=1}^{R} u_{mr} v_{nr})\right) \quad (14)$$

In contrast to the logistic/probit likelihood for binary data, the Bernoulli-Poisson construction used here is appealing due for two reasons. The first is that the computations scale in the number of nonzeros in $\mathbf{X}$ rather than the number of observations in $\mathbf{X}$. This is possible because, in the conditional *posterior* of z_{mn}

$$z_{mn} | x_{mn}, \boldsymbol{u}_m, \boldsymbol{v}_n \sim x_{mn} \cdot \text{Pois}_{+}(\sum_{r=1}^{R} u_{mr} v_{nr}) \quad (15)$$

which means $z_{mn} = 0$ with probability one if $x_{mn} = 0$, and therefore need not be sampled if $x_{mn} = 0$. The second reason is that this link function is skewed (towards having very few nonzeros), resembling the complementary loglog function (Piegorsch, 1992; Collett, 2002), unlike the logistic/probit link, and therefore can better model highly sparse binary matrices.

3 Inference *via* Gibbs Sampling

Exact inference in our model is intractable and therefore we resort to approximate inference. Leveraging the Poisson-multinomial equivalence, which allows re-expressing a Poisson random draw as a draw from a multinomial (Dunson and Herring, 2005; Zhou et al., 2012), we obtain a model with full local conjugacy. This allows closed-form Gibbs sampling for all the model parameters. A key aspect of our model is that inference based on Gibbs sampling scales in the number of nonzero entries in $\mathbf{X}$ (for both count as well as binary data which we model as *thresholded counts* as discussed in Sec. 2.4), which makes is especially attractive for massive but sparse matrices. Although, here we only consider *batch* Gibbs sampling, our inference method can be easily extended to perform online Gibbs sampling (Guhaniyogi et al., 2014; Hu et al., 2015), which will allow scaling up to even more massive data sets. We leave this extension to future work.

Sampling the latent counts x_{mnr} and x_{mnrl}: Using the Poisson-multinomial equivalence, if the matrix $\mathbf{X}$ is count-valued, then the latent counts x_{mnr} and x_{mnrl} can be sampled as

$$\{x_{mnr}\} \sim \mathrm{Mult}(x_{mn}; \frac{u_{mr}v_{nr}}{\sum_{r=1}^{R} u_{mr}v_{nr}}) \quad (16)$$

$$\{x_{mnrl}\} \sim \mathrm{Mult}(x_{mnr}; v_{nrl}/v_{nr}) \quad (17)$$

If $\mathbf{X}$ is binary-valued, we need to first sample the latent count z_{mn} for each nonzero x_{mn} from a truncated Poisson. Then x_{mnr} can be sampled as

$$\{x_{mnr}\} \sim \mathrm{Mult}(z_{mn}; \frac{\{u_{mr}v_{nr}\}}{\sum_{r=1}^{R} u_{mr}v_{nr}}) \quad (18)$$

and x_{mnrl} is sampled the same way as in equation 17.

Sampling $\boldsymbol{u}_{:r}$: Using the multinomial-Dirichlet conjugacy, $\boldsymbol{u}_{:r}$ is sampled as

$$\boldsymbol{u}_{:r} \sim \mathrm{Dir}(\alpha + x_{1\cdot\cdot}, \ldots, \alpha + x_{M\cdot\cdot}) \quad (19)$$

Sampling v_{nr}: v_{nr} can be updated by $v_{nr} = \sum_{l \in \mathcal{L}_n^{(1)}} v_{nrl}$, where v_{nrl} is sampled as

$$v_{nrl} \sim \mathrm{Ga}(g_{lr} + x_{.nrl}, q_r) \quad (20)$$

Sampling q_r: Using the additive property of Poisson distribution, we have $x_{\cdot nr} \sim \mathrm{Pois}(v_{nr})$. Integrating out v_{nr}, $x_{\cdot nr}$ can be expressed as a draw from the following negative-binomial distribution

$$x_{\cdot nr} \sim \mathrm{NB}(\sum_{l \in \mathcal{L}_n^{(1)}} g_{lr}, q_r) \quad (21)$$

Then q_r can be sampled by using negative-binomial-beta conjugacy as

$$q_r \sim \mathrm{Beta}(c\epsilon + x_{\cdot\cdot r}, c(1-\epsilon) + \sum_{n=1}^{N} \sum_{l \in \mathcal{L}_n^{(1)}} g_{lr}) \quad (22)$$

Sampling g_{lr}: Using the additive property of Poisson distribution, $x_{\cdot nrl} \sim \mathrm{Pois}(u_{\cdot r}v_{nrl})$, which can be further rewritten as $x_{\cdot nrl} \sim \mathrm{Pois}(v_{nrl})$ since $u_{\cdot r} = 1$. Let $\mathcal{D}_l$ be the set of all objects (i.e. documents) associated with node l (i.e. an author) on the first layer, we further obtain the following equation by applying the additive property of Poisson distribution once more

$$\sum_{n \in \mathcal{D}_l} x_{\cdot nrl} \sim \mathrm{Pois}(\sum_{n \in \mathcal{D}_l} v_{nrl}) \quad (23)$$

Since the gamma distribution is infinitely divisible, $\sum_{n \in \mathcal{D}_l} v_{nrl}$ can be expressed as

$$\sum_{n \in \mathcal{D}_l} v_{nrl} \sim \mathrm{Ga}(|\mathcal{D}_l| g_{lr}, q_r/(1-q_r)) \quad (24)$$

Integrating out $\sum_{n \in \mathcal{D}_l} v_{nrl}$ in equation 23 and 24, $\sum_{n \in \mathcal{D}_l} x_{\cdot nrl}$ can be expressed as a negative-binomial distribution by

$$\sum_{n \in \mathcal{D}_l} x_{\cdot nrl} \sim \mathrm{NB}(|\mathcal{D}_l| g_{lr}, q_r) \quad (25)$$

Using scaling property of the gamma distribution,

$$|\mathcal{D}_l| g_{lr} \sim \mathrm{Ga}(\sum_{p \in \mathcal{A}_l} h_{pr}, \frac{|\mathcal{D}_l| g_{lr}}{\beta_0}) \quad (26)$$

Using the data augmentation method for negative-binomial distribution (Zhou and Carin, 2015), g_{lr} can be sampled as

$$f_{lr} \sim \mathrm{CRT}(\sum_{n \in \mathcal{D}_l} x_{\cdot nrl}, |\mathcal{D}_l| g_{lr}) \quad (27)$$

$$g_{lr} \sim \mathrm{Ga}(\sum_{p \in \mathcal{A}_l} h_{pr} + f_{lr}, \frac{1}{\beta_0 - |\mathcal{D}_l| \ln(1-q_r)}) \quad (28)$$

where CRT denotes the Chinese restaurant table (Zhou and Carin, 2015) distribution.

Sampling h_{pr}: According to corollary 2 in (Zhou and Carin, 2015), $f_{lr} \sim \mathrm{Pois}(-|\mathcal{D}_l| g_{lr} \ln(1-q_r))$. As $g_{lr} = \sum_{p \in \mathcal{A}_l} g_{lrp}$, f_{lr} can be augmented as $f_{lr} = \sum_{p \in \mathcal{A}_l} f_{lrp}$, where

$$f_{lrp} \sim \mathrm{Pois}(-|\mathcal{D}_l| g_{lrp} \ln(1-q_r)) \quad (29)$$

Let $\mathcal{M}_p$ be the set containing all level-one nodes associated with a level-two node p. We then have

$$\sum_{f \in \mathcal{M}_p} \ell_{fra} \sim \mathrm{Pois}(-\ln(1-q_r) \sum_{f \in \mathcal{M}_p} (|\mathcal{D}_f| g_{lrp})) \quad (30)$$

$\sum_{l \in \mathcal{M}_p} f_{lrp}$ can be expressed as a negative-binomial distribution by integrating out $\sum_{l \in \mathcal{M}_p} g_{lrp}$,

$$\sum_{l \in \mathcal{M}_p} f_{lrp} \sim \mathrm{NB}(|\mathcal{M}_p| h_{pr}, \frac{-\ln(1-q_r)}{\beta_0 - \ln(1-q_r)}) \quad (31)$$

Using scaling property of the gamma distribution,

$$|\mathcal{M}_p| h_{pr} \sim \mathrm{Ga}(s, |\mathcal{M}_p|/\beta_1) \quad (32)$$

Applying the data augmentation method for negative-binomial distribution once more, h_{pr} is sampled as

$$f'_{pr} \sim \mathrm{CRT}(\sum_{l \in \mathcal{M}_p} f_{lrp}, s) \quad (33)$$

$$h_{pr} \sim \mathrm{Ga}(s + f'_{pr}, \frac{1}{\beta_1 - |\mathcal{M}_p| \ln(1 - Q_r)}) \quad (34)$$

where $Q_r = \frac{-\ln(1-q_r)}{\beta_0 - \ln(1-q_r)}$.

4 RELATED WORK

Our work has interesting parallels with some existing methods that attempt to leveraging side-information when modeling discrete data; for example, methods that can incorporate supervision in Latent Dirichlet Allocation (LDA) based topic models (Rosen-Zvi et al., 2004; Mcauliffe and Blei, 2008; Lacoste-Julien et al., 2009; Wang and Blei, 2011; Zhu et al., 2012), and recent work on utilizing side-information in matrix factorization models for count/binary data (Gopalan et al., 2014; Acharya et al., 2015). These class of methods are, however, limited in the type of side-information that can be leveraged, as they usually do not assume any structure within the associated side-information (which is usually given in form of a flat feature vector or a single binary/multi-class label associated with each object).

Among other related work, our framework is somewhat similar in spirit to hierarchically supervised LDA model (Perotte et al., 2011) and the topic model for taxonomies (Bakalov et al., 2012). However, these methods are designed strictly for leveraging very specific, taxonomy-based side-information, whereas our framework can handle more general forms of structural side-information, and a taxonomy is just one of the examples of such side-information.

Structural side-information can also be utilized for specific cases using specialized hierarchical generative models such as the hierarchical Dirichlet Process (Teh et al., 2006). However, in order to properly utilize the type of multi-level side-information our framework can easily utilize, such models would require significant modeling sophistication. Moreover, inference can be considerably more challenging in such models.

In addition to being richer in terms of the types of structural side-information that can be leveraged, our fully Bayesian framework is also conceptually simpler in construction as compared to the aforementioned class of methods. At the same time, our model is easily amenable to efficient inference, and has several interesting propeties that the existing methods lack (e.g., learning embeddings the objects being modeled as well as embeddings of the nodes at all levels in the side-information), and is applicable in a wide variety of applications, such as topic modeling, recommender systems, network modeling, while leverging *structural* side-information in a principled way.

5 EXPERIMENTS

We evaluate our model, both quantitatively (in its ability to predict *missing* data in the matrix **X**) and quatitatively (interpretability of the topics and embeddings learned by the model), by performing experiments on six real-world data sets. For four of the data sets, the matrix **X** has count-valued observations, whereas for the remaining two (Cora and CiteSeer [1]), the observations in **X** are binary. The description of each data set and the associated side-information is given below:

- **20 Newsgroup:** This data [2] consists of 18,774 documents (vocabulary size 5638) organized into 20 groups where each of the groups can be further classified into a super-group (there are a total of seven super-groups). Thus the side-information can be thought of as a two-level taxonomy. For this data, **X** is 5638 × 18774 word-count matrix.
- **State of the Union:** This dataset includes 225 state of the union messages (vocabulary size 7518) delivered annually by 41 presidents of the US from 1790 to 2014 (Wang and McCallum, 2006). Party affiliation for each president is also available (Independent, Federalist, Democratic-Republican, Democrat, Whig, Republican). Thus the side-information is a two-level taxonomy. For this data, **X** is 7518 × 225 word-count matrix.
- **Scholars:** This dataset includes abstracts of 20,149 papers (vocabulary size 8663 words) written by 2,425 researchers associated with 200 affiliations at a US university‘(Hu et al., 2015). The side-information is a two-level hierarchy. For this data, **X** is 8663 × 20149 word-count matrix.
- **NIPS:** 2484 articles (vocabulary size 14036) of the NIPS conferences from 1988 to 2003. The corpus consists of 2865 authors. For this data [3], the side-information only consists of a single level (author identities). For this data, **X** is 14036 × 2484 word-count matrix.
- **Cora:** The data contains 2708 research papers from 7 sub-areas of machine learning: case-based reasoning, genetic algorithms, neural networks, probabilistic methods, reinforcement learning, rule learning, and theory. There are overall 5429 citations (links) between the papers.
- **CiteSeer:** The Citeseer data set contains 3312 papers which can be classified into 6 categories: Agents, AI, DB, IR, ML, and HCI. There are overall 4591 citations (links) between the papers.

[1] http://preview.tinyurl.com/jq4sag6

[2] http://qwone.com/~jason/20Newsgroups/

[3] http://ai.stanford.edu/~gal/Data/NIPS/

Table 1: Loglikelihood comparison between PFA and PFA-SSI for State of the Union (STOU), 20 newsgroup (20 News), Scholar, and NIPS data sets. 10% data was held out as testing data, and 90% used as training. Results are averaged over 10 random splits of training and test.

Methods	STOU	20 Newsgroup	Scholar	NIPS
PFA Zhou et al. (2012)	-23232	-522876	-506905	-345853
PFA-SSI	**-22168**	**-397969**	**-389060**	**-293404**

Table 2: Loglikelihood, AUC and AUC-PR comparison between PFA and PFA-SSI for Cora and CiteSeer datasets. 10% data was held out as testing data, and 90% used as training. Results are averaged over 10 random splits of training and test.

Methods	Cora			CiteSeer		
	Loglike	AUC	AUC-PR	Loglike	AUC	AUC-PR
PFA Zhou et al. (2012)	-9057	0.699	0.740	-9973	0.545	0.670
PFA-SSI	**-3682**	**0.808**	**0.841**	**-4042**	**0.788**	**0.814**

All of our experiments were performed on a standard desktop with 12 GB RAM. For each data set, we set the number of topics (R) to be 200, which serves as an upper bound on the number of topics and the model can prune away the unnecessary topics due to the beta-negative binomial construction (Zhou et al., 2012) of our model. In all our experiments, we fix the hyper-parameters β_0 and β_1 to 1, $\epsilon = 1/R$, and the Dirichlet hyperparameter α was fixed at 0.1. These hyperparameter settings worked well for all the data sets.

5.1 Predicting Held-out Data

We evaluate our model on predicting missing data in the matrix $\mathbf{X}$ by holding out 10% of the observations and predicting them via our non-negative matrix factorization approach, using the remaining 90% data as training data. Each experiment was repeated 10 times and the average accuracies are reported.

Baseline: We compare our model with Poisson Factor Analysis (PFA) Zhou et al. (2012), which is a state-of-the-art non-negative matrix factorization method and also subsumes many other discrete matrix factorization methods (including gamma-Poisson count matrix factorization, LDA, etc.) as special cases. Also note that the PFA model of Zhou et al. (2012) can only handle count data. Therefore, to apply this baseline for the two binary data sets, we modified the PFA implementation ourselves by replacing the Poisson likelihood model by the Bernoulli-Poisson model. Also, we are unable to provide here comparison with other baselines because, to the best of our knowledge, none of the existing methods can incorporate the type of multi-level side-information available for the count/binary matrices we use in our experiments.

Table 1 shows the results for the cases when $\mathbf{X}$ is count-valued and Table 2 shows the results for the cases when $\mathbf{X}$ is binary-valued. For the count-valued data sets, we report the heldout log-likelihood. For the binary-valued data sets, we report the heldout log-likelihood as well as area under the ROC curve (AUC) and area under the precision-recall curve (AUC-PR). As shown in Table 1 and Table 2, our model significantly outperforms PFA on all the data sets, which shows our model's ability in leveraging structural side-information in an effective way.

5.2 Qualitative Analyses

We perform qualitative analyses of our results on various data sets using the topics and the embeddings learned by our model.

20 Newsgroup Data: For this data, Table 3 shows the most prominent topic associated with each of the 20 groups of the level-one side-information. Note that our model learns embeddings of each of these groups and the non-negative embeddings of each group can be used to identify the most active topic associated with that group. Likewise, Table 4 shows the most prominent topic associated with each of the 7 super-groups of the level-two side-information. As Table 3 and Table 4 show, the topics inferred are closely related to the corresponding groups/super-groups. Using the inferred group/super-group embeddings, we also compute cosine similarities between groups and between groups and supergroups. Fig. 2 shows the plots of the estimated similarities. As the plots show, similarities between groups that belong to the same super-group are high, as reflected by the block-diagonal pattern in Fig. 2 (left). Likewise, each group has a higher inferred similarity with its own super-group as compared to other super-groups, as shown in in Fig. 2 (right). These results show that the embeddings learned by our model are meaningful and are consistent with the ground-truth.

State of the Union Data: For the State of the Union data, we use the inferred embeddings of presidents and parties to compute president-president similarity and president-party similarity. The resulting plots are shown in Fig. 3. It is interesting to note that the president-president inferred similarity plot shows a

Table 3: Most prominent topic for each group in 20 newsgroup data

atheism	graphics	win.misc	pc.hardware	mac.hardware	win.x	forsale	autos	motorcycles	baseball
religion	image	windows	dos	windows	file	sale	car	car	game
real	graphics	file	windows	drive	window	offer	bike	bike	year
god	bit	pc	system	mac	program	st	cars	work	baseball
book	data	mail	drive	card	output	shipping	oil	dod	team
true	computer	ac	scsi	mb	server	condition	dod	phone	hit
question	software	os	card	system	entry	price	ca	problem	players
liar	processing	dos	mb	scsi	mit	email	engine	engine	cs
hockey	crypt	electronics	med	space	christian	guns	mideas	politics.misc	religion.misc
game	db	data	doctor	space	mary	fire	israel	government	god
ca	key	circuit	patients	nasa	entry	fbi	jews	cramer	bible
hockey	encryption	signal	msg	billion	church	indiana	israeli	optilink	jesus
espn	chip	input	disease	cost	win	compound	arab	clayton	christian
team	government	output	day	extra	rules	uiuc	jewish	state	christians
nhl	clipper	pin	chronic	based	sin	tanks	land	clinton	christ
year	keys	loop	medical	station	scripture	news	read	white	word

Table 4: Most prominent topic for each supergroup in 20 newsgroup data

religion	comp	auto	sport	sci	politics	forsale
god	image	bike	game	space	gun	sale
bible	graphics	dod	hockey	launch	government	offer
jesus	bit	ca	period	satellite	crime	st
christian	data	back	april	satellites	control	shipping
christians	computer	car	espn	technology	firearms	condition
christ	software	bmw	play	commercial	news	price
word	processing	front	team	system	criminal	email

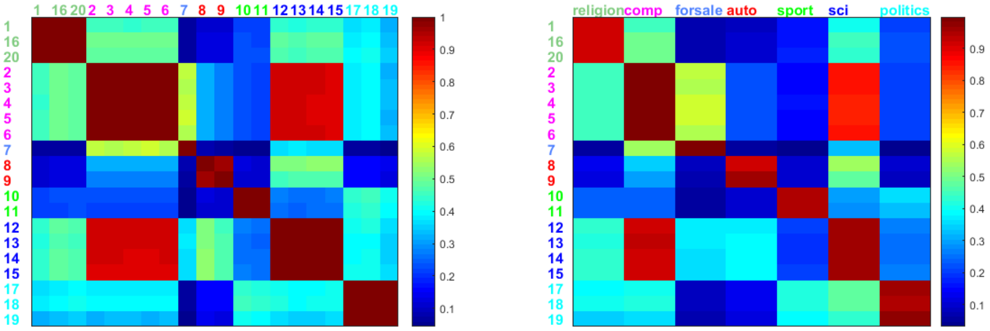

Figure 2: **20 newsgroups data. Left Figure:** Inferred similarities between the level-one nodes (i.e., between the 20 groups) in the side-information. **Right Figure:** Inferred similarities between the level-one and the level-two nodes (i.e., 20 groups and 7 super-groups) in the side-information. The numbers are indices for groups, and numbers with same color indicate that the corresponding groups are associated with the same supergroup. The indices for groups are as follows. 1: alt.atheism; 2: comp.graphics; 3: comp.os.ms-windows.misc; 4: comp.sys.ibm.pc.hardware; 5: comp.sys.mac.hardware; 6: comp.windows.x; 7: misc.forsale; 8: rec.autos; 9: rec.motorcycles; 10: rec.sport.baseball; 11: rec.sport.hockey; 12: sci.crypt; 13: sci.electronics; 14: sci.med; 15: sci.space; 16: soc.religion.christian; 17: talk.politics.guns; 18: talk.politics.mideast; 19: talk.politics.misc; 20: talk.religion.misc.

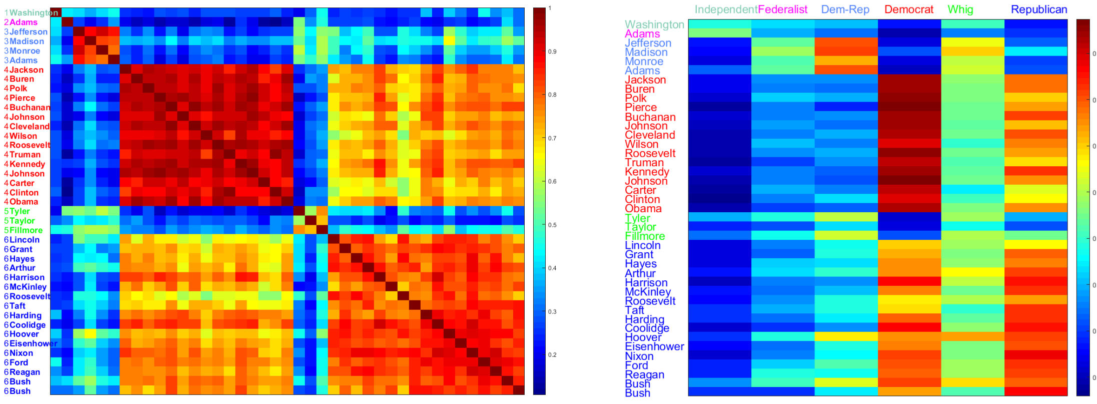

Figure 3: **State of the union data. Left Figure:** Inferred similarities between the presidents in the embedding space. The numbers before each president are labels for parties. 1: Independent; 2: Federalist; 3: Democratic-Republican; 4: Democrat; 5: Whig; 6: Republican. In the legend, the names of all presidents from the same party are shown in the same color. **Right Figure:** Inferred similarities between presidents (level-one nodes) and parties (level-two nodes).

Table 5: Two of the most prominent topics (for considered time-period of 1988-2003) for five of the authors in NIPS data

Alex Smola		Zoubin Ghahramani		Geoff Hinton		Michael Jordan		Peter Bartlett	
functions	data	variables	gaussian	units	objects	probability	space	theorem	tree
linear	basis	em	mixture	hidden	experts	parameters	local	bound	data
kernel	set	models	components	weights	view	likelihood	dimensional	case	training
support	functions	field	data	hinton	hierarchical	bayesian	cluster	proof	decision
set	radial	monte carlo	independent	information	recognition	prior	structure	dimension	test
vector	gaussian	networks	density	inputs	parts	distribution	nearest	upper	trees
space	training	inference	covariance	net	gating	estimation	points	class	machine
regression	test	belief	matrix	set	level	density	dimensionality	set	sets
svm	supervised	markov	variance	internal	multiple	maximum	high	theory	boosting
quadratic	techniques	sampling	ica	bias	information	posterior	manifold	lemma	margin

Table 6: Five most similar authors (for considered time-period of 1988-2003) for five of the authors in NIPS data

Alex Smola	Zoubin Ghahramani	Geoff Hinton	Michael Jordan	Peter Bartlett
Chris Burges	D Titterington	DE Rumelhart	Nir Friedman	Robert Williamson
Pavel Laskov	Lawrence Saul	J McClelland	Sathiya Keerthi	D Helmbold
Ayhan Demiriz	Brendan Frey	J Elman	Miguel Carreira-Perpinan	John Shawe-Taylor
Vladimir Vapnik	David Barber	Antony Bloesch	Tony Jebara	E Sontag
Charles Micchelli	Tommy Jaakkola	Ryotaro Kamimura	David MacKay	V Maiorov

block-diagonal structure (for better visualization, the president indices are ordered based on the party indices), with presidents from the same party inferred to be highly similar with each-other. This suggests that the side-information from level-two nodes (parties) is effectively transferred to level-one nodes (presidents).

NIPS Data: We next look at the topics inferred from the NIPS data. Using the inferred embeddings for each author, we rank the most prominent topics for each author (based on the embedding scores). Table 5 shows two most active topics for each of five of the authors in NIPS data. As Table 5 shows, the inferred most prominent topics for each of these authors are consistent with what these authors were best known for the time-period (1988-2003) covered by this data collection. We also perform an experiment to find the most similar authors for a given author. For this, we use the author embeddings to compute author-author similarity and, in Table 6, show the five most similar authors for each authors from a set of five authors. The results in Table 5 and Table 6 show that the inferred embeddings can provide a good explanation of the data.

5.3 Classification via inferred embeddings

The embeddings learned by our model can also be useful for classication tasks. To demonstrate this, we perform an experiment on multiclass classification. For this experiment, we use the 20 newsgroup data, which is divided into a training set consisting of 11269 documents and a test set consisting of 7507 documents. We use the training set to train our PFA-SSI model and use the word-topic matrix $\mathbf{U}$ and the label embedding matrix $\mathbf{G}$ learned from the training data to predict the labels for test set documents as follows: we infer the document embedding $\mathbf{V}_{test}$ of each test document by sampling from the posterior conditioning on $\mathbf{U}$ and then find the most similar label for each test document by comparing the inferred test document's embedding with embedding of each label. As a baseline, we fit an LDA model on training and test documents, train a multiclass SVM on the topic proportions and labels of the training data, and use the learned classifer to predict the labels for test documents. Our model gave a classification accuracy of 63.7% whereas the LDA+SVM baseline gave a classification accuracy of 61.5%. This experiment shows that our model, although not originally designed for classification tasks, can nevertheless achieve reasonable classification accuracies because the supervision enhances the discriminative power of the embeddings learned by our model.

6 CONCLUSION

We have presented a probabilistic framework for incorporating structural side-information in non-negative matrix factorization for count and binary data. Our fully Bayesian framework is conceptually simple, computationally scalable, and leads to improved performance on predicting held-out data. The topics and the embeddings learned by our model can be useful for various other downstream tasks (e.g., classification) or for qualitative analyses.

The flexibility of our generative model, which can model both count as well as binary data under a unified framework, and the ease of inference, makes our framework particularly attractive for applications involving discrete data with structural side-information. Our framework can also be extended to handle binary/count tensor data (Hu et al., 2015; Schein et al., 2015) with structural side-information given along one more more of the tensor modes. For our model, it is also easy to perform online variational inference or online Gibbs sampling, which will allow analyzing even more massive data sets using our model. We leave such developments to future work.

ACKNOWLEDGMENTS

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

References

A. Acharya, D. Teffer, J. Henderson, M. Tyler, M. Zhou, and J. Ghosh. Gamma process poisson factorization for joint modeling of network and documents. In *ECML-PKDD*. 2015.

D. Agarwal and B.-C. Chen. Regression-based latent factor models. In *KDD*, 2009.

A. Bakalov, A. McCallum, H. Wallach, and D. Mimno. Topic models for taxonomies. In *JCDL*, 2012.

A. J. Chaney, D. M. Blei, and T. Eliassi-Rad. A probabilistic model for using social networks in personalized item recommendation. In *RecSys*, 2015.

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

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

P. Gopalan, L. Charlin, and D. Blei. Content-based recommendations with poisson factorization. In *NIPS*, 2014.

P. Gopalan, J. Hofman, and D. Blei. Scalable recommendation with hierarchical poisson factorization. In *UAI*, 2015.

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

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

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

S. Lacoste-Julien, F. Sha, and M. I. Jordan. Disclda: Discriminative learning for dimensionality reduction and classification. In *NIPS*, 2009.

J. D. Mcauliffe and D. M. Blei. Supervised topic models. In *NIPS*, 2008.

A. J. Perotte, F. Wood, N. Elhadad, and N. Bartlett. Hierarchically supervised latent dirichlet allocation. In *NIPS*, 2011.

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

M. Rabinovich and D. Blei. The inverse regression topic model. In *ICML*, 2014.

D. Ramage, D. Hall, R. Nallapati, and C. D. Manning. Labeled lda: A supervised topic model for credit attribution in multi-labeled corpora. In *EMNLP*, 2009.

M. Rosen-Zvi, T. Griffiths, M. Steyvers, and P. Smyth. The author-topic model for authors and documents. In *UAI*, 2004.

A. Schein, J. Paisley, D. M. Blei, and H. Wallach. Bayesian poisson tensor factorization for inferring multilateral relations from sparse dyadic event counts. In *KDD*, 2015.

Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical dirichlet processes. *Journal of the american statistical association*, 101(476), 2006.

C. Wang and D. M. Blei. Collaborative topic modeling for recommending scientific articles. In *KDD*, 2011.

X. Wang and A. McCallum. Topics over time: a non-markov continuous-time model of topical trends. In *KDD*, 2006.

J. Yang and J. Leskovec. Overlapping community detection at scale: a nonnegative matrix factorization approach. In *WSDM*, 2013.

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

M. Zhou and L. Carin. Negative binomial process count and mixture modeling. *PAMI*, 37(2):307–320, 2015.

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

J. Zhu, A. Ahmed, and E. P. Xing. Medlda: maximum margin supervised topic models. *JMLR*, 13(1):2237–2278, 2012.