Transfer Learning for Hierarchically Supervised Topic Models

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

Abstract

We present a new framework for topic modeling that allows leveraging *hierarchical* side-information associated with the documents, specified as a multi-level taxonomy or ontology structure. For example, a scholarly document with a two-level side-information will have its authors' identities in the first level and, and the author affiliations in the second level. Our framework is based on non-negative matrix factorization of count data (e.g., word counts) and learns embeddings of the entities present at each level in the data/side-information hierarchy (e.g., documents, authors, affiliations, in the previous example), with appropriate transfer of information across levels. Although here we consider document modeling, the framework can be readily applied to the more general problem of non-negative matrix factorization of count-valued matrices with hierarchical side-information. The framework also enjoys full local conjugacy, facilitating efficient Gibbs sampling for model inference. Inference scales in the number of non-zero entries in data matrix, which is especially appealing for massive but sparse matrices. We demonstrate the effectiveness of this framework on several real-world datasets.

1 Introduction

Topic models [1, 9] provide a useful way for uncovering topics or thematic structures in documents. In some cases, the documents may be associated with meta-data or labels (supervision) that can be leveraged for improved topic modeling[7, 5]. Most of the exising methods of this type, however, do not assume/exploit *structural* forms of supervision, e.g., a taxonomy of labels/categories for the documents, or other natural types of multi-level supervision, such as document authors and their corresponding affiliations. See Fig. 1 for some examples where the the data is naturally represented as a matrix of word counts for each document, with associated side-information. Although data exhibiting such structure are prevalent in many applications, existing methods cannot properly leverage such forms of side-information arranged in form of multiple layers. This problem setting naturally calls for effective and efficient transfer learning across the multiple layers of supervision.

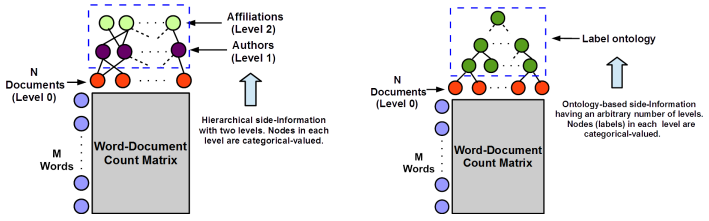


Figure 1: Two examples of the type of side-information that our proposed framework can leverage, **Left:** Side-information specified in form of a multi-layer hierarchy with bipartite connections between nodes in adjacent layers. **Right**: Side-information specified in form of an ontology over known labels.

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 factor-

ization of count data, by transferring information across the multiple levels of hierarchy of taxonomy associated with the data. In addition to being useful for standard tasks such as matrix completion for count 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 that constitute 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 *multiple resolutions* (e.g., in Fig 1-left, topics can be naturally associated to authors and affiliations). This significantly enhances the versatility and usefulness of our framework to applications beyond matrix factorization and completion. Our framework also enjoys full local conjugacy which facilitates closed-form Gibbs sampling for all model parameters. Moreover, inference in our model 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 as a hierarchy or ontology with two levels; the model can be easily implemented using Gibbs sampling and modified to work with arbitrary number of levels. Here we assume that we are given a data matrix $\mathbf{X}$ of size $M \times N$, where each column of $\mathbf{X}$ represents a document. The side-information for documents is provided in form of a multi-level structures, such as a hierarchy (Fig. 1-left) or an ontology (Fig. 1-right). In the absence of any side-information, the counts matrix $\mathbf{X} \in \mathbb{Z}^{M \times N}$ can be modeled using a Poisson Factor Analysis (PFA) model as $\mathbf{X} \sim \text{Pois}(\mathbf{U}\mathbf{V}^\top)$ where $\mathbf{U}$ and $\mathbf{V}$ are non-negative matrices of size $M \times R$ and $N \times R$, respectively. This is equivalent to assuming that each count-valued observation x_{mn} is a sum of R latent counts [2, 9, 3]. Therefore, we have

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

$$v_{nr} \sim \text{Ga}(g_r, q_r/(1-q_r)), \qquad q_r \sim \text{Beta}(c\epsilon, c(1-\epsilon)) \tag{2}$$

$$\boldsymbol{u}_{:r} \sim \text{Dir}(\alpha, \ldots, \alpha), \qquad g_r \sim \text{Ga}(c_0 g_0, 1/h_0) \tag{3}$$

Note that each Dirichlet drawn column $\boldsymbol{u}_{:r}$ of $\mathbf{U}$ represents a "topic". Also note the Poisson-gamma construction (Eq. 1–2) is equivalent to a gamma-negative binomial model [9].

2.1 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

$$v_{nr} = \sum_{l \in \mathcal{L}_n^{(1)}} v_{nrl}, \qquad v_{nrl} \sim \text{Ga}(g_{lr}, q_r/(1-q_r)) \tag{4}$$

where $\mathcal{L}^{(1)}$ denotes the set of *all* nodes in level-one and $\mathcal{L}_n^{(1)}$ denotes the subset of these nodes associated with object n from level-zero. Using gamma-additivity, Eq. 4 can be combined as

$$v_{nr} \sim \text{Ga}(\sum_{l \in \mathcal{L}_n^{(1)}} g_{lr}, q_r/(1-q_r)) \tag{5}$$

In Eq. 5, 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 is written as a sum of contributions from each of the level-two nodes it is associated with:

$$g_{lr} = \sum_{p \in \mathcal{L}_l^{(2)}} g_{lrp}, \qquad g_{lrp} \sim \text{Ga}(h_{pr}, 1/\beta_0), \qquad h_{pr} \sim \text{Ga}(s, 1/\beta_1) \tag{6}$$

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. 6 can also be combined as $g_{lr} \sim \text{Ga}(\sum_{p \in \mathcal{L}_l^{(2)}} h_{pr}, 1/\beta_0)$, where 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.

2.2 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 $\boldsymbol{v}_n = [v_{n1}, \ldots, v_{nR}]$, $\boldsymbol{g}_l = [g_{l1}, \ldots, g_{lR}]$, , and $\boldsymbol{h}_p = [h_{p1}, \ldots, h_{pR}]$ 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 [6, 4] with *multi-level supervision*, which most of existing methods are unable to leverage in a proper way; 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.

3 EXPERIMENTS

We evaluate our model, both quantitatively (in its ability to predict *missing* data in the matrix $\mathbf{X}$) and quatitatively (interpretability of the topics and embeddings learned by the model), by performing experiments on three real-world data sets described below:

20 Newsgroup: This data 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.

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 [8]. Party affiliation information for each president is also available (Independent, Federalist, Democratic-Republican, Democrat, Whig, and Republican). Thus the side-information can be thought of as a two-level taxonomy.

NIPS: 2484 articles (vocabulary size 14036) of the NIPS conferences from 1988 to 2003. The corpus consists of 2865 authors. For this data, the side-information only consists of one level (author).

Table 1: Loglikelihood comparison between PFA and PFA-SSI.

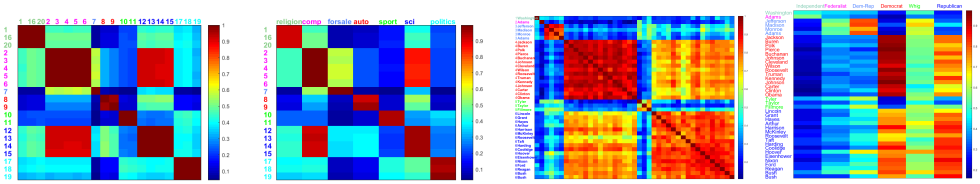
Methods	STOU	20 Newsgroup	NIPS
PFA [9]	-23232	-522876	-345853
PFA-SSI	**-22168**	**-397969**	**-293404**

3.1 Predicting Held-out Data

We evaluate our model on predicting missing data in $\mathbf{X}$ by holding out 10% of the observations and predicting them via our approach, using the remaining 90% data as training data. We compare our model with Poisson Factor Analysis (PFA) [9], which is a state-of-the-art non-negative matrix factorization method and also subsumes many other discrete factorization methods (including gamma-Poisson count matrix factorization, LDA, etc.) as special cases. Table 1 shows the log-likelihood for the held-out data. As is shown, our model significantly outperforms PFA on all datasets, which shows our model's ability in leveraging structural side-information in an effective way.

Table 2: Most prominent topic for five groups (left) and five super-groups (right) in 20 newsgroup

atheism	graphics	mac.hardware	forsale	autos	religion	auto	sport	sci	politics
religion	image	windows	sale	car	god	bike	game	space	gun
real	graphics	drive	offer	bike	bible	dod	hockey	launch	government
god	bit	mac	st	cars	jesus	ca	period	satellite	crime
book	data	card	shipping	oil	christian	back	april	satellites	control
true	computer	mb	condition	dod	christians	car	espn	technology	firearms
question	software	system	price	ca	christ	bmw	play	commercial	news



(a) (b) (c) (d)

Figure 2: **20 newsgroups and STOU data.** Inferred group-group (a) and president-president (b) similarities at level-one. Inferred group-supergroup (b) and president-party (d) similarities between the level-one and the level-two nodes. The numbers in (a) (b) 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.The numbers before each president in (c) 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.

Table 3: Two most prominent topics (for time-period of 1988-2003) for five 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

3.2 Qualitative Analyses

We also perform qualitative analyses of our results using the topics and the embeddings learned by our model.

20 Newsgroup Data: For this data, Table 2 shows the most prominent topic associated with five groups of the level-one (groups) and level-two (supergroups) 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. As shown in the table, 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(a)(b) 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 (a). 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 (b). These results show that the embeddings learned by our model are meaningful and 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. 2(c)(d). It is interesting to note that the president-president inferred similarity plot shows a 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 3 shows two most active topics for each of five of the authors in NIPS data. As Table 3 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.

4 CONCLUSION

We have presented a probabilistic framework for incorporating structural side-information in non-negative matrix factorization for count-valued data (e.g., a matrix of word counts for a collection of text documents, such as scholarly publications). 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.

References

[1] D. Blei, A. Ng, and M. Jordan. Latent dirichlet allocation. *JMLR*, 2003.

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

[3] C. Hu, P. Rai, C. Chen, M. Harding, and L. Carin. Scalable bayesian non-negative tensor factorization for massive count data. In *ECML*, 2015.

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

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

[6] 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.

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

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

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