ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-7605-5/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICASSP39728.2021.9413656

ON A GUIDED NONNEGATIVE MATRIX FACTORIZATION

Joshua Vendrow[⋆] *Jamie Haddock*[⋆] *Elizaveta Rebrova*[⋆] *Deanna Needell*[⋆]

[⋆]University of California, Los Angeles
Department of Mathematics
520 Portola Plaza, Los Angeles, CA 90095

ABSTRACT

Fully unsupervised topic models have found fantastic success in document clustering and classification. However, these models often suffer from the tendency to learn less-than-meaningful or even redundant topics when the data is biased towards a set of features. For this reason, we propose an approach based upon the nonnegative matrix factorization (NMF) model, deemed *Guided NMF*, that incorporates user-designed seed word supervision. Our experimental results demonstrate the promise of this model and illustrate that it is competitive with other methods of this ilk with only very little supervision information.

Index Terms— supervised topic models, supervised nonnegative matrix factorization, seed words

1. INTRODUCTION

As modern data collection and storage capabilities improve and grow, so do the size and complexity of modern data sets that data practitioners are tasked with turning to actionable knowledge. For this reason, data scientists are increasingly turning to unsupervised dimensionality-reduction and topic modeling techniques to understand the latent trends within their data. These approaches have produced fantastic results in document clustering and classification, see e.g., [1, 2].

However, it has been previously noted that such models can learn topics that are not meaningful or effective in downstream tasks [3]. In particular, these models can be hindered by data in which certain features are so weighted as to bias the models towards topics with these features and away from more balanced and meaningful topics [4].

For this reason, we develop a supervised topic model that incorporates flexible supervision information representing user knowledge of feature importance and associations. Our approach is based upon the popular nonnegative matrix factorization (NMF) [5] and builds upon its supervised variant, semi-supervised NMF (SSNMF) [6]. The key difference in our approach, however, is that our goal is to guide the topic outputs, rather than provide labels for classification. The goal is thus to identify topics within the data that are driven by the seeded features, thereby revealing more meaningful topics for the particular application.

1.1. Nonnegative matrix factorization

NMF is an approach typically applied in unsupervised tasks such as dimensionality-reduction, latent topic modeling, and clustering. Given nonnegative data matrix $\boldsymbol{X} \in \mathbb{R}_{\geq 0}^{m\times n}$ and a user-defined target dimension $k \in \mathbb{N}$, NMF seeks nonnegative factor matrices $\boldsymbol{A} \in \mathbb{R}_{\geq 0}^{m\times k}$, often referred to as the *dictionary* or *topic* matrix, and $\boldsymbol{S} \in \mathbb{R}_{\geq 0}^{k\times n}$, often referred to as the *representation* or *coefficient matrix*, such that $\boldsymbol{X} \approx \boldsymbol{AS}$. There are many formulations of this model (see e.g., [7, 5, 8]) but the most popular utilizes the Frobenius norm,

$$\underset{\boldsymbol{A}\geq 0, \boldsymbol{S}\geq 0}{\arg\min} \|\boldsymbol{X} - \boldsymbol{AS}\|_F^2. \tag{1}$$

Here and throughout, $\boldsymbol{A} \geq 0$ denotes the constraint that $\boldsymbol{A}$ is entry-wise nonnegative. The user-defined parameter k, which represents the target dimension or the number of believed latent topics, governs the quality of reconstruction of the data; generally k is chosen so that $k < \min\{m, n\}$ to ensure non-triviality of the factorization. The columns of $\boldsymbol{A}$ are often referred to as *topics*; the NMF approximations to the data (columns of $\boldsymbol{X}$) are additive nonnegative combinations of these topic vectors. This property of NMF approximations yields interpretability since the strength of relationship between a given data point (column of $\boldsymbol{X}$) and the topics of $\boldsymbol{A}$ is clearly visible in the coefficient vector (corresponding column of $\boldsymbol{S}$). For this reason, NMF has found popularity in applications such as document clustering [1], image and audio processing [9, 10], and financial data mining [11].

1.2. Semi-supervised nonnegative matrix factorization

SSNMF is a modified variant of NMF that jointly factorizes a data matrix $\boldsymbol{X} \in \mathbb{R}_{\geq 0}^{m\times n}$ and a supervision information matrix $\boldsymbol{Y} \in \mathbb{R}_{\geq 0}^{c\times n}$ with the goal of learning a dimensionality-reduction model and a model for a supervised learning task (e.g., classification). That is, given data matrix $\boldsymbol{X}$, supervision matrix $\boldsymbol{Y}$, and target dimension $k \in \mathbb{N}$, SSNMF seeks

The authors were partially supported by NSF DMS #2011140 and NSF BIGDATA #1740325.

the dictionary matrix $\boldsymbol{A} \in \mathbb{R}_{\geq 0}^{m \times k}$, representation matrix $\boldsymbol{S} \in \mathbb{R}_{\geq 0}^{k \times n}$, and *supervision matrix* $\boldsymbol{B} \in \mathbb{R}_{\geq 0}^{c \times k}$ such that $\boldsymbol{X} \approx \boldsymbol{AS}$ and $\boldsymbol{Y} \approx \boldsymbol{BS}$. The most popular SSNMF formulation [6] employs a weighted combination of Frobenius norm terms,

$$\underset{\boldsymbol{A},\boldsymbol{S},\boldsymbol{B} \geq 0}{\arg\min} \underbrace{\|\boldsymbol{X} - \boldsymbol{AS}\|_F^2}_{\text{Reconstruction Error}} + \lambda \underbrace{\|\boldsymbol{Y} - \boldsymbol{BS}\|_F^2}_{\text{Classification Error}}; \tag{2}$$

recently, other formulations have been proposed [12].

1.3. Related work

Other supervised variants of NMF (besides equation 2) have been proposed. The works [13, 14, 15] propose models that exploit cannot-link or must-link supervision, while [16] introduces a model with information divergence penalties on the reconstruction and on supervision terms that influence the learned factorization to approximately reconstruct coefficients learned before factorization by a support-vector machine (SVM). Several works [17, 18, 19] propose a supervised NMF model that incorporates Fisher discriminant constraints into NMF for classification. Joint factorization of two data matrices, like that of SSNMF, is described more generally and denoted Simultaneous NMF in [7].

Previous works incorporating feature-level knowledge into topic modeling have predominantly used as their backbone Latent Dirichlet Allocation (LDA) [20]. The authors of [21] guide topics by incorporating *Must-Links* and *Cannot-Links* that increase or decrease, respectively, the probability of two words appearing in the same topic. The authors of [22] guide the formation of topics by adding a constraint on the LDA sampling algorithm to force certain words to only appear in specified topics. Each of the works [23, 24, 25] develop LDA models which incorporate response variables and class labels to improve the learned topic model and its performance on downstream learning tasks.

The work that aligns most closely with our goal is [4], which proposes Seeded LDA. This method accepts sets of seed words and adjusts the LDA model probability distributions to encourage topics to generate words related to those in the seed set. The experimental setup of Seeded LDA is very similar to our own, so within our experiments we provide comparison to Seeded LDA.

Finally, other models utilize other approaches to incorporate feature-level information; [26] utilizes *prototype* supervision information in corpora topic modeling, while that of [27] utilizes n-gram statistics,while [28] is an attempt to extract and utilize *gist* of words in the corpora. This task is also highly related to that of constrained clustering [29].

1.4. Contribution and Organization

Our primary contribution is to propose a simple yet worthwhile approach for topic modeling when the user has a priori knowledge about some of the desired topics. For example, we will showcase a setting from political Twitter data where we wish to learn topics related to specific policies and employ seed words to guide the learned topics towards those desired. Without such guidance, the natural topics identified would largely reflect individual political candidates, thereby obscuring the topics of interest and related documents.

With this as our primary objective, in Section 2 we propose the Guided NMF model and introduce a metric to measure the quality of formed topics. Then, in Section 3, we perform topic modeling experiments on two document analysis data sets and compare our model to Seeded LDA. Finally, in Section 4 we summarize our findings and discuss future work.

2. METHOD

Our proposed method, which we refer to as *Guided NMF*, makes use of seed word (or generally seed-feature) supervision and exploits a model based upon SSNMF. We evaluate this model on corpora topic modeling and classification tasks.

2.1. Seed word supervision

We will refer to a keyword identified in the user-provided supervision as a *seed word*, and we will refer to a (possibly weighted) group of seed words as a *seed topic*. We denote a seed topic as a vector $\boldsymbol{v} = (v_1, \ldots, v_m)$ (where m denotes the vocabulary size), where $v_i = 0$ if the ith word in the vocabulary is not in the seed topic and some positive weight otherwise. In general, we expect $\boldsymbol{v}$ to be very sparse because the number of important keywords identified for a topic should be far smaller than the total vocabulary of keywords. We note that by considering each element of $\boldsymbol{v}$ to be a feature rather than a word, we can extend this formulation to any topic modeling task (e.g., in image/video topic modeling tasks). In our experiments, we use $v_i \in \{0, 1\}$ but note that varying weights could improve performance in many applications.

2.2. Guided NMF

Let the data matrix $\boldsymbol{X} \in \mathbb{R}^{m \times n}$ have examples along the columns and features along the rows and suppose we have seed topics $\boldsymbol{v}^{(1)}, \boldsymbol{v}^{(2)}, \ldots, \boldsymbol{v}^{(c)} \in R^m$. Let the *seed matrix* be

$$\boldsymbol{Y} = [\boldsymbol{v}^{(1)}, \boldsymbol{v}^{(2)}, \ldots, \boldsymbol{v}^{(c)}] \in \mathbb{R}_{\geq 0}^{m \times c} \tag{3}$$

Guided NMF is formulated as

$$\min_{\boldsymbol{A} \geq 0, \boldsymbol{S} \geq 0, \boldsymbol{B} \geq 0} \|\boldsymbol{X} - \boldsymbol{AS}\|_F^2 + \lambda \|\boldsymbol{Y} - \boldsymbol{AB}\|_F. \tag{4}$$

We note that this model is symbolically equivalent to standard SSNMF where the data $\boldsymbol{X}$ and seed matrix $\boldsymbol{Y}$ are transposed. Here, the important distinction is the dimension of $\boldsymbol{X}$ to which supervision information is provided. This new perspective yields application when there is available information regarding the latent relationship between individual features and topics, rather than individual data points and classes.

Following application of Guided NMF, we can use the topic supervision matrix $\boldsymbol{B}$ to identify columns of the dictionary matrix $\boldsymbol{A}$ corresponding to the topics that form around our seed words. By examining the corresponding rows of the supervision matrix $\boldsymbol{S}$, we can find the documents that Guided NMF attributes to these topics (interpreting the columns of $\boldsymbol{S}$ as a score for the relationship of each document to a topic, we can classify documents to the topics based on the magnitude of this score). To measure accuracy, we use the widely accepted *receiver operating characteristic* (ROC) metric and corresponding *area under the curve* (AUC). Thus, we use this classification metric as a measure of the quality of topics when they have a one to one correspondence to classes.

3. EXPERIMENTS

In this section, we present results of applying Guided NMF to 20 Newsgroups and a Twitter political dataset. We compare with Seeded LDA in the 20 Newsgroups experiments where we have labels. Code for all experiments can be found in `https://github.com/jvendrow/GuidedNMF` and uses the multiplicative updates method of [12].

3.1. 20 Newsgroups dataset

The 20 Newsgroups dataset is a collection of approximately 20,000 text documents containing the text of messages from 20 different newsgroups on the distributed discussion system Usenet [30]. From this data set, we use a subset of 10 newsgroups with 100 documents each (graphics, hardware, forsale, motorcycles, baseball, medicine, space, guns, mideast, and religion). In a first example, we consider learning 4 topics but guiding those topics via the seed words *pitch*, *medical*, and *space* in hopes of capturing the corresponding topics. In another experiment we use the seed words *motorcycle*, *sale* and *religion* with a similar goal. We choose rank four so as to capture the topics from the leftover document classes separately, allowing the method to more easily guide the remaining topics as we would hope. In Tables 1 and 2, we display the results of running Guided NMF on the newsgroup dataset with these seed words. By including two tables with different seed words, we show how the topics vary based on seed information. We see that for each seed word, a full topic forms around this word that provides clear and salient keywords corresponding to the information within that class.

3.2. Twitter political dataset

The Twitter political data set [31] is a data set of tweets sent by political candidates during the 2016 election season. In Table 3, we display the results of running a regular NMF on the data set. We see that most topics focus on a specific candidate or campaign slogan rather than a political issue.

Table 1. Topic keywords learned for a rank 4 Guided NMF on the 20 Newsgroups dataset with the seed words *pitch*, *medical*, and *space*. We see that a clear topic forms from each keyword matching one desired newsgroup class.

Topic 1	Topic 2	Topic 3	Topic 4
pitch	*medical*	*space*	people
expected	tests	nasa	know
curveball	disease	shuttle	think
stiffness	diseases	launch	time
loosen	prejudices	sci	use
shoulder	services	lunar	new
shea	graduates	orbit	see
rotation	health	earth	say
game	patients	station	us
giants	available	mission	god

Table 2. Topic keywords learned for a rank 4 Guided NMF on the 20 Newsgroups dataset with the seed words *motorcycle*, *sale*, and *religion*. We see that a clear topic forms from each keyword matching one desired newsgroup class.

Topic 1	Topic 2	Topic 3	Topic 4
motorcycle	*sale*	*religion*	people
bike	offer	christian	know
dod	condition	judaism	think
wheelie	shipping	freedom	time
shaft	asking	christians	use
bikes	includes	islam	new
rider	mb	compulsion	space
riding	excellent	avi	see
scene	price	life	say
ski	best	gunpoint	us

To uncover "hidden" topics concerning political issues, we run Guided NMF on this data set with two seed words, *economy* and *obamacare*, two issues discussed during the 2016 election, and display the results in Table 4. We see that a topic forms that around each seed word, and the topic keywords provide additional context for the seeded issue; we see that the main economic concerns are jobs and taxes, and the discussion relating to Obamacare focuses on repeal, for which some Republican candidates advocated.

3.3. Ablation and Comparison

In all the text-based experiments above, we used only a single seed word per class and achieved salient and interpretable results. Here, we explore the impact of adding additional seed words and also varying the rank of the factorization. We also provide comparisons to Seeded LDA [4]. To compute AUC for Seeded LDA, we use the metric described in Section 2.2, but rather than using the $\boldsymbol{S}$ matrix as in the case of NMF, we instead use the document-topic distribution variables. For the

Table 3. Topic keywords learned by a rank 8 NMF on the Twitter political dataset. We see that most topics center around one of the political candidates.

Topic 1	**Topic 2**	**Topic 3**	**Topic 4**
thank	govpencein	gopdebate	tedcruz
trump2016	indiana	imwithhuck	cruz
maga[1]	indiana_edc	jeb	cruzcrew
great	state	tonight	ted
america	jobs	president	choosecruz
Topic 5	**Topic 6**	**Topic 7**	**Topic 8**
kasich	hillary	randpaul	fitn
john	trump	iowa	new
johnkasich	people	iacaucus	hampshire
ohio	donald	caucus	johnkasich
gov	president	tonight	nh

[1]Here "maga" abbreviates "makeamericagreatagain."

Table 4. Topic keywords learned by a rank 8 Guided NMF on the Twitter political dataset with the seed words *economy* and *obamacare*. The first two topics form around these seeds, with meaningful related keywords appearing below them.

Topic 1	**Topic 2**	**Topic 3**	**Topic 4**
economy	*obamacare*	govpencein	gopdebate
jobs	fullrepeal	indiana	kasich
tax	repeal	indiana_edc	randpaul
plan	replace	state	john
create	fight	jobs	tonight
Topic 5	**Topic 6**	**Topic 7**	**Topic 8**
tedcruz	hillary	johnkasich	people
thank	trump	new	need
cruz	donald	fitn	must
cruzcrew	clinton	kasich	berniesanders
ted	president	hampshire	country

space topic we use the seed words *space, lunar, nasa, launch, rocket, moon, shuttle* and *orbit* and for the baseball topic we use the seed words *pitch, baseball, team, ball, game, season, base* and *field*. We choose these seed words from keywords commonly appearing in space or baseball NMF topics.

In Tables 5 and 6, we display AUC scores for Guided NMF and Seeded LDA over a variety of settings for rank and number of seed words. We see that Guided NMF consistently has an AUC above 0.8 for all rank and number of seed word choices. In the case of particular interest in our setting, namely when few seed words are supplied and/or only a small number of topics are desired, Guided NMF significantly outperforms Seeded LDA; we note that with a higher rank the desired topics are more likely to form naturally, making the task easier. With many seed words and a high rank, Seeded LDA only slightly outperforms our method. This can likely be attributed to differences between NMF and LDA.

Table 5. AUC scores for the 20 Newsgroups dataset on documents for the space class.

Rank	Method	# Seed words			
		1	2	4	8
4	Guided NMF	**0.83**	**0.88**	**0.88**	**0.87**
	Seeded LDA	0.31	0.42	0.74	0.86
6	Guided NMF	**0.86**	**0.87**	0.88	0.87
	Seeded LDA	0.37	0.5	**0.91**	**0.89**
10	Guided NMF	**0.88**	0.89	0.89	0.89
	Seeded LDA	0.45	**0.95**	**0.95**	**0.95**

Table 6. AUC scores for the 20 Newsgroups dataset on documents for the baseball class.

Rank	Method	# Seed words			
		1	2	4	8
4	Guided NMF	**0.89**	**0.9**	**0.9**	**0.9**
	Seeded LDA	0.31	0.42	0.74	0.86
6	Guided NMF	**0.9**	**0.9**	0.9	**0.9**
	Seeded LDA	0.37	0.5	**0.91**	0.89
10	Guided NMF	**0.87**	0.9	0.9	0.9
	Seeded LDA	0.45	**0.95**	**0.95**	**0.95**

4. CONCLUSION

We propose an NMF-based model, that we call Guided NMF, which incorporates seed topic supervision to guide learned topics towards meaningful and coherent sets of features. Our initial experiments illustrate the promise of this model in text-based topic modeling applications. This model could be extended to image/video applications, where the supervision provided encourages object localization and segmentation.

5. REFERENCES

[1] W. Xu, X. Liu, and Y. Gong, "Document clustering based on non-negative matrix factorization," in *Proc. ACM SIGIR Conf. on Research and Development in Inform. Retrieval*, 2003, pp. 267–273.

[2] F. Shahnaz, M. Berry, V. Pauca, and R. Plemmons, "Document clustering using nonnegative matrix factorization," *Inform. Process. Manag.*, vol. 42, no. 2, pp. 373–386, 2006.

[3] J. Chang, S. Gerrish, C. Wang, J. L. Boyd-Graber, and D. M. Blei, "Reading tea leaves: How humans interpret topic models," in *Adv. Neur. In.*, 2009, pp. 288–296.

[4] J. Jagarlamudi, H. Daumé III, and R. Udupa, "Incorporating lexical priors into topic models," in *Proc. Conf. Euro. Chapter Assoc. Comp. Ling.*, 2012, pp. 204–213.

[5] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788, 1999.

[6] H. Lee, J. Yoo, and S. Choi, "Semi-supervised nonnegative matrix factorization," *IEEE Signal Proc. Let.*, vol. 17, no. 1, pp. 4–7, 2009.

[7] A. Cichocki, R. Zdunek, A. H. Phan, and S. Amari, *Nonnegative matrix and tensor factorizations: applications to exploratory multi-way data analysis and blind source separation*, John Wiley & Sons, 2009.

[8] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Proc. Adv. Neur. In.*, 2001, pp. 556–562.

[9] D. Guillamet and J. Vitria, "Non-negative matrix factorization for face recognition," in *Proc. Catalonian Conf. on Artif. Intel.* Springer, 2002, pp. 336–344.

[10] A. Cichocki, R. Zdunek, and S. Amari, "New algorithms for non-negative matrix factorization in applications to blind source separation," in *Proc. Int. Conf. Acoust. Spe. Sig. Process.* IEEE, 2006, vol. 5, pp. V–V.

[11] R. de Fréin, K. Drakakis, S. Rickard, and A. Cichocki, "Analysis of financial data using non-negative matrix factorization," in *Proc. Int. Mathematical Forum*. Hikari, 2008, vol. 3(38), pp. 1853–1870.

[12] J. Haddock, L. Kassab, S. Li, A. Kryshchenko, R. Grotheer, E. Sizikova, C. Wang, T. Merkh, R. W. M. A. Madushani, M. Ahn, D. Needell, and K. Leonard, "Semi-supervised nonnegative matrix factorization models for topic modeling in learning tasks," 2020, Submitted.

[13] Y. Chen, M. Rege, M. Dong, and J. Hua, "Non-negative matrix factorization for semi-supervised data clustering," *Knowl. Inf. Syst.*, vol. 17, no. 3, pp. 355–379, 2008.

[14] W. Fei, L. Tao, and Z. Changshui, "Semi-supervised clustering via matrix factorization," in *Proc. SIAM Int. Conf. on Data Mining*, 2008.

[15] Y. Jia, S. Kwong, J. Hou, and W. Wu, "Semi-supervised non-negative matrix factorization with dissimilarity and similarity regularization," *IEEE T. Neur. Net. Lear.*, 2019.

[16] Y. Cho and L. K. Saul, "Nonnegative matrix factorization for semi-supervised dimensionality reduction," *arXiv preprint arXiv:1112.3714*, 2011.

[17] Y. Jia, Y. Wang, C. Turk, and M. Hu, "Fisher non-negative matrix factorization for learning local features," in *Proc. Asian Conf. Comp. Vis.* Citeseer, 2004, pp. 27–30.

[18] Y. Xue, C. S. Tong, W. Chen, W. Zhang, and Z. He, "A modified non-negative matrix factorization algorithm for face recognition," in *Proc. Int. Conf. on Pattern Recognition*. IEEE, 2006, vol. 3, pp. 495–498.

[19] S. Zafeiriou, A. Tefas, I. Buciu, and I. Pitas, "Exploiting discriminant information in nonnegative matrix factorization with application to frontal face verification," *IEEE T. Neural Networ.*, vol. 17, no. 3, pp. 683–695, 2006.

[20] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, no. Jan, pp. 993–1022, 2003.

[21] D. Andrzejewski, X. Zhu, and M. Craven, "Incorporating domain knowledge into topic modeling via Dirichlet forest priors," in *Proc. Int. Conf. Mach. Learn.*, 2009, pp. 25–32.

[22] D. Andrzejewski and X. Zhu, "Latent dirichlet allocation with topic-in-set knowledge," in *Proc. NAACL HLT Workshop Semi-Supervised Learn. Nat. Lang. Proc.*, 2009, pp. 43–48.

[23] J. D. Mcauliffe and D. M. Blei, "Supervised topic models," in *Adv. Neur. In.*, 2008, pp. 121–128.

[24] S. Lacoste-Julien, F. Sha, and M. I. Jordan, "DiscLDA: Discriminative learning for dimensionality reduction and classification," in *Adv. Neur. In.*, 2009, pp. 897–904.

[25] D. Ramage, D. Hall, R. Nallapati, and C. D. Manning, "Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora," in *Proc. Conf. Empirical Methods in Nat. Lang. Proc.*, 2009, pp. 248–256.

[26] A. Haghighi and D. Klein, "Prototype-driven learning for sequence models," in *Proc. Human Lang. Tech. Conf. NAACL*, 2006, pp. 320–327.

[27] H. M. Wallach, "Topic modeling: beyond bag-of-words," in *Proc. Int. Conf. Mach. Learn.*, 2006, pp. 977–984.

[28] T. L. Griffiths, M. Steyvers, and J. B. Tenenbaum, "Topics in semantic representation.," *Psychol. Rev.*, vol. 114, no. 2, pp. 211, 2007.

[29] S. Basu, I. Davidson, and K. Wagstaff, *Constrained clustering: Advances in algorithms, theory, and applications*, CRC Press, 2008.

[30] K. Lang, "20 newsgroups," Jan 2008.

[31] J. Littman, L. Wrubel, and D. Kerchner, "2016 United States Presidential Election Tweet Ids," 2016.