JOINT TOPIC-DOCUMENT MODELING VIA LOW-DIMENSIONAL SPARSE MODELS

Kerui Min[⋆] *Yi Ma*[⋆†]

[⋆]Dept. of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign

[†]Visual Computing Group, Microsoft Research Asia

E-mail: kmin12@illinois.edu, mayi@microsoft.com

ABSTRACT

Topic modeling is a well-known approach for document analysis. In this paper, we propose a new model, and corresponding optimization algorithm for topic modeling. Experimental results on polarity classification demonstrate that the new model provides a more accurate characterization for document corpus, and archived higher classification accuracy compared to *Latent Dirichlet Allocation* (LDA).

Index Terms— topic modeling, non-negative matrix factorization

1. INTRODUCTION

Topic modeling is a well-known computational tool for finding thematic information from a given corpus. In recent years, this approach and its variants have been proved to be very useful in a large number of applications, including text and image classification, clustering, retrieval etc., see, *e.g.* the discussion paper [1] and references therein.

Traditionally, topic modeling was based on the assumption that each document (or more generally, each observed vector) is a combination of a small number of topic vectors [2, 3]. Based on this assumption, quite a few algorithms have been developed, trying to solve the topic vectors given a corpus [4, 3, 5, 6, 7]. As we will see in Section 3, by verifying this assumption on real datasets, we found that the tractional model is highly inaccurate. For example, a document often contains certain keywords (person's name, brand name, etc.) that cannot be explained by common topics. We instead propose a *joint topic-document model* to better model the structure within a document corpus.

2. RELATED WORK

The idea of representing a document by a linear combination of few "topics" can be track back to [4]. The theoretical computer science community studied the studied the behavior of the above algorithm [2]. In machine learning, many probabilistic topic models were developed in recent years, *e.g.* *Latent Dirichlet Allocation* (LDA) [3] and many of its variant like *Correlated Topic Model* (CTM) [5]. Accompanying with the develop of topic models, varying algorithms have been proposed, including Variational Bayesian [3], Gibbs sampling [6] and stochastic gradient descent [8]. Despite its empirical success, it is also shown that the inference of LDA model is NP-Hard [9].

The inference of topic models has an intimate relationship with *non-negative matrix factorization* (NMF), as the non-negativity factors can be associated with a probabilistic interpretation. Very recently, [7] studied how to use NMF for topic modeling, and provided an algorithm with provably approximation under a separability assumption. Many other local search based algorithms also exist, *e.g* [10, 11].

3. OUR APPROACH

3.1. Issues with current approach

Topic modeling is based on the assumption that each document in a corpus is a convex combination of relatively small number of topic vectors. Therefore, the corpus, written as a matrix by stacking the document vectors together, has a low-rank structure. Mathematically, it assumes that the corpus $\boldsymbol{D}$ has the following structure:

$$\boldsymbol{D} \approx \boldsymbol{B}\boldsymbol{S}, \tag{1}$$

where $\boldsymbol{B} \in \mathbb{R}^{m\times k}$ consisting of k topic vectors, acting as the basis of documents, and $\boldsymbol{S} \in \mathbb{R}^{k\times n}$ is the coefficient matrix, $k \ll n$. To obtain a probabilistic interpretation, $\boldsymbol{B}$ and $\boldsymbol{S}$ should satisfying the following condition:

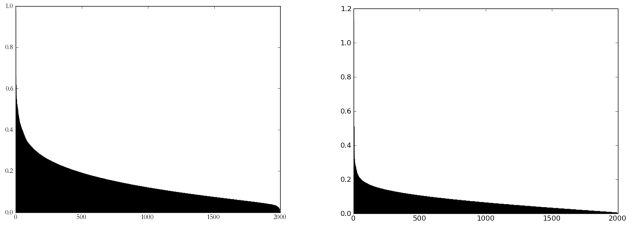
$$\boldsymbol{B} \geq \boldsymbol{0}, \quad \boldsymbol{S} \geq \boldsymbol{0}, \quad \|\boldsymbol{B}_i\|_1 = 1 \text{ for } i \in [k], \tag{2}$$

where $\boldsymbol{B}_i$ is the i-th column of the basis matrix $\boldsymbol{B}$. We note that for applications such as image analysis, the *sum-to-1* condition may not be necessary.

Although the above model is widely accepted, and has been used explicitly or implicitly in [4, 3, 5], this generative model is far from being accurate, empirically. To see this, we analyze the spectral distribution of two datasets: the 20 Newsgroup dataset[1] and the Movie Review dataset[2]. The

[1]Available at http://qwone.com/ jason/20Newsgroups/

[2]Available at http://www.cs.cornell.edu/people/pabo/movie-review-data/



(a) 20 Newsgroup dataset (b) Movie Review dataset

Fig. 1. Singular value distribution of two typical corpora. Both exhibit long tail distribution.

Variance	30%	50%	70%	90%
Rank	100	244	490	1001

Fig. 2. The minimum ranks required to capture 30%, 50%, 70%, and 90% of the variance of the TF·IDF matrix of the 20 Newsgroup dataset.

Movie Review dataset consists of 2000 reviews from the IMDB database, and the 20 Newsgroup is a collection of around 20,000 posts across 20 newsgroups. We randomly select 2000 posts from 20 Newsgroup dataset, and employ the standard pre-process to obtain the TF·IDF matrix $\boldsymbol{D}$ for each dataset. The singular value distributions of the two datasets are shown in Fig 1. The singular value distributions of both datasets qualitatively exhibit very similar heavy-tail distribution, which implies that the data matrices don't have an accurate low-rank approximation. As shown in Fig 2, to capture 90% of the variance of $\boldsymbol{D}$, we should choose k, the number of topics, to be at least 1001. Notice that the rank function only provides an lower-bound of non-negative rank, as we further require $\boldsymbol{B} \geq \boldsymbol{0}$ and $\boldsymbol{S} \geq \boldsymbol{0}$.

Empirical studies such as [3, 8] often choose k to be 50 to 200. As a consequence, the basic model $\boldsymbol{D} \approx \boldsymbol{BS}$ is no longer appropriate as there remains large amount of variation in the data left unexplained by the model. To address this issue, [12] proposed the *joint topic-document model*, $\boldsymbol{D} = \boldsymbol{L} + \boldsymbol{E}$, where $\boldsymbol{L}$ is the low-rank matrix capturing the background, or topic information, and $\boldsymbol{E}$ is a sparse matrix, representing the document-specific keywords or keypharses that cannot be explained by the (low-rank) topic model.

The *joint topic-document model* provides a more accurate model for text corpus. Empirically, however, we found that when we penalize $\boldsymbol{L}$ to be a relatively low-rank matrix, the document-specific counterpart $\boldsymbol{E}$ is not so sparse. In addition, *joint topic-document model* only gives the span of topic distribution. Very often it is necessary to obtain the distribution for each topic and the corresponding mixture-of-topic coefficients for tasks like document classification and clustering.

3.2. Noisy joint topic-document model

Based on the above observations and discussions, we would like to come up with a model that can describe more accurately the corpus data. In particular we would like to model the topics from a corpus as well as those document-specific keywords, and remaining terms that cannot be represented by either the background topic or the document-specific keywords. More specifically, we would like to model our observed data matrix $\boldsymbol{D} \in \mathbb{R}^{m \times n}$ as:

$$\boldsymbol{D} = \boldsymbol{BS} + \boldsymbol{E} + \boldsymbol{N}, \tag{3}$$

where $\boldsymbol{B}$ is topic basis (or topic vector set), $\boldsymbol{E}$ corresponds to the document-specific keywords, and $\boldsymbol{N}$ are the remaining noises. We call this model the *noisy joint topic-document model*. By allowing $\boldsymbol{N}$ to be reasonably large in variance, the document-specific $\boldsymbol{E}$ can be made sufficiently sparse. Motivated by the observation that each document should only contain few topics, we can penalize the mixture-of-topic coefficients $\boldsymbol{S}$ by ℓ_1-norm. We penalize the sparse keywords part $\boldsymbol{E}$ by ℓ_1-norm for the same reason.

Hence one can learn the above model from the data via solving the following optimization problem

$$\begin{aligned} \min_{\boldsymbol{B},\boldsymbol{S},\boldsymbol{E},\boldsymbol{N}} & \sum_i \|\boldsymbol{S}_i\|_1^2 + \lambda \|\boldsymbol{E}\|_1 + \frac{\tau}{2}\|\boldsymbol{N}\|_F^2 \\ \text{subject to} \quad & \boldsymbol{D} = \boldsymbol{BS} + \boldsymbol{E} + \boldsymbol{N}, \;\; \boldsymbol{B} \geq \boldsymbol{0}, \\ & \boldsymbol{S} \geq \boldsymbol{0}, \;\; \|\boldsymbol{B}_j\|_1 = 1 \text{ for } j \in [k]. \end{aligned} \tag{4}$$

The reason to choose $\|\boldsymbol{S}\|_{1,2}^2 \triangleq \sum_i \|\boldsymbol{S}_i\|_1^2$ as the regularization instead of $\|\boldsymbol{S}\|_1$ is twofold:

1. By choosing $\sum \|\boldsymbol{S}_i\|_1^2$, we enforce *each* document in the given corpus to be a linear combination of a few topics. The norm $\|\boldsymbol{S}\|_1$ enforces the corpus *as a whole* can be sparsely represented by $\boldsymbol{B}$, but does not ensure representations for the documents are evenly sparse.

2. Computationally, the regularizer $\sum_i \|\boldsymbol{S}_i\|_1^2$ allows us to have an efficient algorithm for the optimization problem. We will discuss this issue in detail in the next section.

The above optimization problem also extends the *stable principal component pursuit* [13] by providing a specific generative model for the low-rank topic component.

4. ALGORITHM

In this section, we study how to solve problem (6) efficiently. It is easily seen that the optimization problem is non-trivial due to the non-linear constraint $\boldsymbol{D} = \boldsymbol{BS} + \boldsymbol{E} + \boldsymbol{N}$, where both $\boldsymbol{B}$ and $\boldsymbol{S}$ are unknown. Notice that the bilinear term also occurs in *non-negative matrix factorization*, *dictionary learning*, and *blind source separation*. Our problem formulation

(6) is closely related to dictionary learning, and the solution is inspired by non-negative matrix factorization.

First, we can take care of the non-linear constraint under the *Augmented Lagrangian Method* [14] framework to solve the problem iteratively. Specifically, we solve the following problem

$$\begin{aligned}\min_{\boldsymbol{B},\boldsymbol{S},\boldsymbol{E},\boldsymbol{N}} \; & \|\boldsymbol{S}\|_{1,2}^2 + \lambda \|\boldsymbol{E}\|_1 + \frac{\tau}{2}\|\boldsymbol{N}\|_F^2 \\ & + \frac{\mu}{2}\|\boldsymbol{D} - \boldsymbol{B}\boldsymbol{S} - \boldsymbol{E} - \boldsymbol{N}\|_F^2 \qquad (5) \\ & + \langle \boldsymbol{Y}, \boldsymbol{D} - \boldsymbol{B}\boldsymbol{S} - \boldsymbol{E} - \boldsymbol{N}\rangle \\ \text{subject to} \quad & \boldsymbol{B} \geq \boldsymbol{0},\ \boldsymbol{S} \geq \boldsymbol{0},\ \|\boldsymbol{B}_j\|_1 = 1 \text{ for } j \in [k], \qquad (6)\end{aligned}$$

where $\boldsymbol{Y}$ is the Lagrange multiplier. The remaining constraints are on $(\boldsymbol{B}, \boldsymbol{S})$. Once $(\boldsymbol{B}, \boldsymbol{S})$ is fixed, the problem can be reduce to a typical ℓ_1 minimization problem using a fixed-point algorithm [15]. Alternatively, we can further fix $\boldsymbol{N}$ and optimize $\boldsymbol{E}$, and then fix $\boldsymbol{E}$ and optimize $\boldsymbol{N}$. Therefore, we solve the following three sub-problmes iteratively.

Sub-problem 1 Given $(\boldsymbol{B}, \boldsymbol{S}, \boldsymbol{N}, \boldsymbol{Y}, \mu)$, optimize $\boldsymbol{E}$. This is equivalent to solve the following unconstrained optimization program

$$\arg\min_{\boldsymbol{E}} \frac{\lambda}{\mu}\|\boldsymbol{E}\|_1 + \frac{1}{2}\left\|\boldsymbol{X}^{(1)} - \boldsymbol{E}\right\|_F^2, \qquad (7)$$

where $\boldsymbol{X}^{(1)}$ is given by $\boldsymbol{X}^{(1)} \triangleq \boldsymbol{D} - \boldsymbol{B}\boldsymbol{S} - \boldsymbol{N} + \frac{1}{\mu}\boldsymbol{Y}$. It is well-known (see, *e.g.* [15]) that the unique optimal solution is given by the *shrinkage operator* $\mathcal{S}_{\lambda\mu^{-1}}\left(\boldsymbol{X}^{(1)}\right)$, where

$$\mathcal{S}_\nu(\cdot) \triangleq \text{sgn}(\cdot)\max\{|\cdot| - \nu, 0\}, \qquad (8)$$

and $\text{sgn}(\cdot)$ is the sign function. We extend it to matrix domain by applying it element-wise to all entries.

Sub-problem 2 Given $(\boldsymbol{B}, \boldsymbol{S}, \boldsymbol{E}, \boldsymbol{Y}, \mu)$, optimize $\boldsymbol{N}$. Similarly, this is equivalent to solve

$$\arg\min_{\boldsymbol{N}} \frac{\tau}{2}\|\boldsymbol{N}\|_F^2 + \frac{\mu}{2}\left\|\boldsymbol{X}^{(2)} - \boldsymbol{N}\right\|_F^2, \qquad (9)$$

where $\boldsymbol{X}^{(2)} \triangleq \boldsymbol{D} - \boldsymbol{B}\boldsymbol{S} - \boldsymbol{E} + \frac{1}{\mu}\boldsymbol{Y}$.

To obtain a closed-form solution, we take the derivative of the objective function with respect to $\boldsymbol{N}$ and set it to zero.

$$\tau\boldsymbol{N} + \mu\left(\boldsymbol{N} - \boldsymbol{X}^{(2)}\right) = 0. \qquad (10)$$

Hence, we get $\boldsymbol{N} = \mu(\tau + \mu)^{-1}\boldsymbol{X}^{(2)}$.

Sub-problem 3 Given $(\boldsymbol{E}, \boldsymbol{N}, \boldsymbol{Y}, \mu)$, optimize $(\boldsymbol{B}, \boldsymbol{S})$. The original objective function is reduced to the following form

$$\begin{aligned}\arg\min_{\boldsymbol{B},\boldsymbol{S}} \; & \|\boldsymbol{S}\|_{1,2}^2 + \frac{\mu}{2}\left\|\boldsymbol{X}^{(3)} - \boldsymbol{B}\boldsymbol{S}\right\|_F^2 \\ \text{subject to} \; & \boldsymbol{B} \geq \boldsymbol{0},\ \boldsymbol{S} \geq \boldsymbol{0},\ \|\boldsymbol{B}_j\|_1 = 1 \text{ for } j \in [k],\end{aligned}$$

where $\boldsymbol{X}^{(3)} \triangleq \boldsymbol{D} - \boldsymbol{E} - \boldsymbol{N} + \frac{1}{\mu}\boldsymbol{Y}$.

We note that unlike the previous two sub-problems where a closed-form optimal solution can be found, this problem is more difficult as the objective function contains the bilinear term $\boldsymbol{B}\boldsymbol{S}$. This is a dictionary learning problem with non-negative constraints, or alternatively, a non-negative matrix factorization problem with sparse regularizer. To make the problem slightly simpler, we can solve the following problem instead.

$$\begin{aligned}\arg\min_{\boldsymbol{B},\boldsymbol{S}} \; & \|\boldsymbol{S}\|_{1,2}^2 + \gamma\|\boldsymbol{B}\|_F^2 + \frac{\mu}{2}\left\|\boldsymbol{X}^{(3)} - \boldsymbol{B}\boldsymbol{S}\right\|_F^2 \\ \text{subject to} \; & \boldsymbol{B} \geq \boldsymbol{0},\ \boldsymbol{S} \geq \boldsymbol{0}, \qquad (11)\end{aligned}$$

where the weight γ is used to strike a balance between the norms of $\boldsymbol{B}$ and $\boldsymbol{S}$. Let $(\boldsymbol{B}', \boldsymbol{S}')$ be a solution of the above optimizations, and define $\boldsymbol{T} \in \mathbb{R}^{k\times k}$ to be

$$\boldsymbol{T} = \text{diag}\left(\|\boldsymbol{B}_1\|_1^{-1}, \cdots, \|\boldsymbol{B}_k\|_1^{-1}\right). \qquad (12)$$

Then, the solution $(\boldsymbol{B}'\boldsymbol{T}, \boldsymbol{T}^{-1}\boldsymbol{S}')$ satisfies the sum-to-1 constraints. Moreover, $\boldsymbol{B}'\boldsymbol{T}$ has the same ℓ_0-norm as $\boldsymbol{B}'$, so the modified solution is sparse as long as the original $\boldsymbol{B}'$ is sparse.

The problem (11) was formulated and studied in [11], where an efficient algorithm, called SNMF, was proposed based on non-negative least squares. Empirically, we find that SNMF provides sufficiently good solutions for sub-problem 3, and the algorithm is insensitive to the choice of γ. Therefore, we fix γ to be a constant throughout our experiments.

Update Lagrange Multiplier The final step in each iteration is to update the Lagrange Multiplier $\boldsymbol{Y}$. Let $\boldsymbol{Y} \leftarrow \boldsymbol{Y} + \mu(\boldsymbol{D} - \boldsymbol{B}\boldsymbol{S} - \boldsymbol{E} - \boldsymbol{N})$, and update μ to be $\rho\mu$ for a constant $\rho > 1$.

5. EXPERIMENTS

5.1. Experimental settings

In this section, we evaluate the proposed topic-document modeling algorithm for polarity classification. We use the Movie Review dataset, which contains 1000 positive reviews and 1000 negative reviews [16]. Therefore, it is a binary classification task. To make a fair comparison, only unigram feature (terms) is used, as it is difficult for traditional topic models like *Latent Dirichlet Allocation* to incorporate other features such as *part-of-speech* tags, or bigram features. Since terms with low document frequency (the number of documents it occurs within a corpus) have little contribution to the topic words, we only keep the top m terms with the highest document frequencies, and construct the TF·IDF matrices accordingly.

Support vector machine (SVM) is used as the classifier for all the subsequent tasks, since it has achieved the highest performance in [16]. For simplicity, we use the linear SVM. The performance accuracy will be evaluated using 5-fold cross-validation using the 2000 samples.

Our	LDA	Our	LDA	Our	LDA
elizabeth (0.12723)	elizabeth (0.00510)	wedding (0.05833)	sandler (0.00791)	batman (0.06913)	effects (0.00797)
queen (0.04298)	foster (0.00449)	singer (0.03481)	wedding (0.00721)	robin (0.02903)	star (0.00713)
england (0.01782)	giles (0.00277)	julia (0.03362)	dvd (0.00645)	mr (0.01424)	special (0.00637)
mary (0.01290)	game (0.00264)	adam (0.01425)	series (0.00410)	joel (0.01008)	batman (0.00401)
rush (0.01159)	anna (0.00263)	romantic (0.01236)	snake (0.00374)	comic (0.00615)	wars (0.00387)
love (0.01112)	fight (0.00256)	comedy (0.00833)	singer (0.00366)	wayne (0.00591)	series (0.00326)
court (0.00971)	video (0.00253)	billy (0.00788)	disc (0.00347)	city (0.00583)	planet (0.00265)
power (0.00900)	eddie (0.00246)	funny (0.00554)	fans (0.00333)	series (0.00557)	trek (0.00245)
country (0.00864)	king (0.00240)	happy (0.00548)	anaconda (0.00323)	villains (0.00529)	earth (0.00236)
political (0.00839)	mortal (0.00237)	steve (0.00546)	x-files (0.00304)	tim (0.00528)	alien (0.00233)

Fig. 3. Comparison of topics learned using our method with LDA for similar topics ($k = 30$). The weight of each term is given in the bracket.

# of topics (k)	LDA	Our method
30	57.2%	**68.2%**
50	61.3%	**70.0%**

Fig. 4. The classification accuracy on the Movie Review dataset, using the distribution of topics information.

# of unigram feature m	$\boldsymbol{D}_i$	$[(\boldsymbol{BS}_i)^T, \boldsymbol{E}_i^T, \boldsymbol{N}_i^T]$
$m = 2000$	85.2%	**87.0%**
$m = 3000$	85.5%	**87.2%**

Fig. 5. Classification accuracy on the Moview Review dataset using the original and new feature space.

5.2. Verify the learned topic

In this experiment, we compare the polarity classification performance of our topic modeling algorithm with LDA using the distribution of topics information, *i.e.* we use $\boldsymbol{S}_i \in \mathbb{R}^k$ to represent the i-th document. For LDA, we use the online LDA algorithm from [8] to learn the topics with default parameters. To ensure the convergence of online LDA, we pass the corpus 200 times. The result is given in Fig 4. Geometrically, the new features are given by projecting the original high-dimensional features to a low-dimensional simplex. We can see that the k-dimensional vector learned by our method significantly outperformed the LDA counterpart, implying that the learned information is semantically more meaningful. In Fig 3, we provide examples of the learned topics compared to LDA.

5.3. Classification using the structure information

The last experiments shows that the decomposed triple $(\boldsymbol{BS}, \boldsymbol{E}, \boldsymbol{N})$ is effective for classification. Recall that our algorithm decompose the corpus $\boldsymbol{D}$ into $\boldsymbol{BS} + \boldsymbol{E} + \boldsymbol{N}$, where $\boldsymbol{BS}$ contains the topic information, $\boldsymbol{E}$ represents those document-specific keywords, and $\boldsymbol{N}$ is the remaining residues. Our intuition is that the above three terms all contain useful information and chould be weighed differently to obtain better performance. To validate this hypothesis, we simply concatenate the decomposed parts together, *i.e.* we use the vector $\left[(\boldsymbol{BS}_i)^T, \boldsymbol{E}_i^T, \boldsymbol{N}_i^T\right] \in \mathbb{R}^{3m}$ as the feature of the i-th document, instead of $\boldsymbol{D}_i \in \mathbb{R}^m$. If the structure of the model does not help extract any additional information, we should expect a performance drop, since the new dimensionality of the new feature space is three times large. Or more specifically, based on the classical statistical learning theory (see, *e.g.* [17]), we know that

$$err_{\text{true}}(h) \leq err_{\text{train}}(h) + \sqrt{\frac{\text{VC}(H)(\ln \frac{2n}{\text{VC}(H)} + 1) + \ln \frac{4}{\delta}}{n}}, \tag{13}$$

whree VC(H) indicates the VC-dimension of the hypothesis space H. For linear classifier, the original feature space has VC(H)$= m + 1$, and the new features space has VC(H')$= 3m + 1$. For sufficiently large n, the second term of the right-hand side of Eq (13) is about 1.73 times larger. The actual performance of the two approaches is given in Fig 5. Clearly, despite the fact that the new feature space is more prone to over-fitting, the structure information does help to improve the accuracy of classification.

Interestingly, although we have not done any special engineering towards this task, we found that the performance is on par with the best performance reported in [18], where a much more sophisticated algorithm based on sentence-level analysis and a subjectivity dataset was designed specifically for the sentimental analysis.

6. CONCLUSION

In this paper, to address issues of previous topic models, we proposed the *noisy joint topic-document model* and provide an efficient iterative algorithm for topic modeling based on it. The model is generic, we believe it can be used for document classification, clustering, and beyond.

7. REFERENCES

[1] D. Blei, "Probabilistic topic models," in *Communications of the ACM*, 2012, vol. 55, pp. 77–84.

[2] C. H. Papadimitriou, H. Tamaki, P. Raghavan, and S. Vempala, "Latent semantic indexing: a probabilistic analysis," in *Proceedings of the seventeenth ACM symposium on Principles of database systems*, 1998, pp. 159–168.

[3] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," in *J. Mach. Learn. Res*, 2003, number 3, pp. 993–1022.

[4] S. T. Dumais, G. W. Furnas, T. K. Landauer, S. Deerwester, and R. Harshman, "Using latent semantic analysis to improve access to textual information," in *Proceedings of CHI*, 1988, pp. 281–286.

[5] D. M. Blei and J. D. Lafferty, "Correlated topic models," in *NIPS*, 2005.

[6] I. Porteous, D. Newman, A. Ihler, A. Asuncion, P. Smyth, and M. Welling, "Fast collapsed gibbs sampling for latent dirichlet allocation," in *Proc. of the 14th ACM SIGKDD*, 2009, pp. 569–577.

[7] S. Arora, R. Ge, and A. Moitra, "Learning topic models - going beyond svd," in *CoRR abs/1204.1956*, 2012.

[8] M. D. Hoffman, D. M. Blei, and F. R. Bach, "Online learning for latent dirichlet allocation," in *NIPS*, 2010, pp. 856–864.

[9] D. Sontag and D. Roy, "Complexity of inference in latent dirichlet allocation," in *NIPS*, 2011.

[10] C.-J. Lin, "Projected gradient methods for non-negative matrix factorization," in *Neural Computation*, 2007, vol. 19, pp. 2756–2779.

[11] H. Kim and H. Park:, "Nonnegative matrix factorization based on alternating nonnegativity constrained least squares and active set method," in *SIAM J. Matrix Analysis Applications*, 2008, vol. 30, pp. 713–730.

[12] K. Min, Z. Zhang, J. Wright, and Y. Ma, "Decomposing background topics from keywords by principal component pursuit," in *CIKM*, 2010, pp. 269–278.

[13] Z. Zhou, X. Li, J. Wright, E. J. Candes, and Y. Ma, "Stable principal component pursuit," in *ISIT*, 2010, pp. 1518–1522.

[14] D. Bertsekas, "Constrained optimization and lagrange multiplier method," in *Academic Press*, 1982.

[15] E. T. Hale, W. Yin, and Y. Zhang, "Fixed-point continuation for ℓ_1-minimization: Methodology and convergence," in *SIAM Journal on Optimization*, 2008, vol. 19, pp. 1107–1130.

[16] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan, "Thumbs up? sentiment classification using machine learning techniques," in *Proceedings of EMNLP*, 2002, pp. 79–86.

[17] V. N. Vapnik, *The nature of statistical learning theory*, Springer-Verlag New York, Inc., New York, NY, USA, 1995.

[18] Bo Pang and Lillian Lee, "A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts," in *Proceedings of ACL*, 2004, pp. 271–278.