Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17)

Probabilistic Non-Negative Matrix Factorization and Its Robust Extensions for Topic Modeling

Minnan Luo,[1,4] **Feiping Nie,**[2*] **Xiaojun Chang,**[3]
Yi Yang,[3] **Alexander Hauptmann,**[4] **Qinghua Zheng**[1]
[1] Shaanxi Province Key Lab. of Satellite and Terrestrial Network Tech. R&D, Department of Computer Science, Xi'an Jiaotong University, P. R. China.
[2] School of Computer Science and Center for Optical Imagery Analysis and Learning, Northwestern Polytechnical University, P. R. China.
[3]Centre for Quantum Computation and Intelligent Systems, University of Technology Sydney.
[4]School of Computer Science, Carnegie Mellon University, PA, USA

Abstract

Traditional topic model with maximum likelihood estimate inevitably suffers from the conditional independence of words given the documents topic distribution. In this paper, we follow the generative procedure of topic model and learn the topic-word distribution and topics distribution via directly approximating the word-document co-occurrence matrix with matrix decomposition technique. These methods include: (1) Approximating the normalized document-word conditional distribution with the documents probability matrix and words probability matrix based on probabilistic non-negative matrix factorization (NMF); (2) Since the standard NMF is well known to be non-robust to noises and outliers, we extended the probabilistic NMF of the topic model to its robust versions using $\ell_{2,1}$-norm and capped $\ell_{2,1}$-norm based loss functions, respectively. The proposed framework inherits the explicit probabilistic meaning of factors in topic models and simultaneously makes the conditional independence assumption on words unnecessary. Straightforward and efficient algorithms are exploited to solve the corresponding non-smooth and non-convex problems. Experimental results over several benchmark datasets illustrate the effectiveness and superiority of the proposed methods.

Introduction

Due to an ever increasing amount of document data, topic modeling plays an important role in the field of document understanding and analyzing (Kuang, Choo, and Park 2015). Typically, topic models are based on the idea that documents are mixtures of topics, where a topic is a probability distribution over words (Steyvers and Griffiths 2007). Th. Hofmann (Hofmann 1999) pioneered to present latent semantic analysis (LSA) (Landauer, Foltz, and Laham 1998) from a statistical perspective and model the word-document co-occurrence information under a framework of aspect model (Hofmann, Puzicha, and Jordan 1999), namely Probabilistic Latent Semantic Analysis (PLSA). PLSA is an important step toward probabilistic modeling of text; however, on one hand, the number of parameters in PLSA model grows linearly with the size of the corpus, and thus it becomes prone to overfitting (Blei, Ng, and Jordan 2003). On the other hand, how to assign a probability to a document outside of the training set is not explicit in the framework of PLSA (Blei, Ng, and Jordan 2003). Considering the exchangeable representations for documents and words, Blei, *et al.* extended the PLSA by introducing a Dirichlet prior on the topic distribution of documents, and proposed a fully Bayesian generalization of latent Dirichlet allocation (LDA) (Blei, Ng, and Jordan 2003). In the past decades, topic model and its variations have been widely used to discover latent semantic structures from collections of text documents (Mcauliffe and Blei 2008), images (Kivinen, Sudderth, and Jordan 2007; Wang, Blei, and Li 2009), audio files (Hoffman, Blei, and Cook 2009) and even biological data (Airoldi et al. 2009).

Topic models indeed attempt to find a low-dimensional representation of the content of a set of documents (Steyvers and Griffiths 2007), which can be modeled from two different views. On one hand, it inherits the explicit probabilistic meaning of factors in aspect model and describes the mixture approximation of the co-occurrence with underlying generative probabilistic semantics. On the other hand, the topic model can be interpreted as a matrix factorization of the conditional distribution of words given the documents. Previous studies (Ding, Li, and Peng 2006; Gaussier and Goutte 2005) have suggested that PLSA and Kullback-Leibler (KL) divergence based Non-negative Matrix Factorization (NMF) indeed optimize the same objective function although they converge to different local minima. It is noteworthy that, as a widely used dimension reduction technique, traditional NMF with Frobenius norm performs well for document clustering and topic modeling (Arora et al. 2012; 2013; Kuang, Choo, and Park 2015), although it lacks explicit probabilistic meaning of factors.

It is based on maximum likelihood estimate that the topic-word distribution and topic distribution are learned in the framework of traditional topic model. Therefore, a fundamental probabilistic assumption is underlying the topic models (including LSA), *i.e.*, words in a document are conditionally independent of each other given the documents topic distribution. In other words, the correlation information among words is completely ignored in the traditional framework of topic models. In fact, the correlation infor-

*Corresponding author. Email: feipingnie@gmail.com

mation might contain significant clues to the content of a document (Steyvers and Griffiths 2007). It is reasonable to believe that some words co-occur more than others, and thus these words usually share the similar frequency. For example, we are given the information that a document contains the word "Trump", which would increase the probability of also observing the word "Clinton".

Instead of using maximum likelihood estimate, in this paper, we learn the topic-word distribution and topics distribution via directly approximating the word-document co-occurrence matrix based on NMF with Frobenius norm, namely probabilistic non-negative matrix factorization for the topic model. This framework inherits the clear probabilistic meaning of factors in topic models and simultaneously makes the independence assumption on words (documents) unnecessary. Considering the outliers with significant loss usually dominate the Frobenius norm based objective function (Nie et al. 2013), the proposed framework provides a flexible way to extend to its robust version by replacing the Frobenius norm with $\ell_{2,1}$-norm or capped $\ell_{2,1}$-norm. Our main contributions are two-fold. On one hand, we propose a probabilistic NMF framework for topic modeling, and intuitively extend to its robust version by using a more robust distance measurement. This framework denies the assumption of conditional independence on words while retaining the explicit probabilistic meaning of factors in the topic model. Thus, it is more appropriate for real-world applications. On the other hand, efficient algorithms are exploited to solve the corresponding probabilistic NMF with different distance measurement for the approximation. Theoretical analysis and experimental results over some benchmark dataset illustrate the effectiveness and superiority of the proposed algorithms.

Related Works

Topic Model

Suppose we have n documents and m words (terms), denoted by sets $\mathcal{D} = \{d_1, d_2, \cdots, d_n\}$ and $\mathcal{W} = \{w_1, w_2, \cdots, w_m\}$, respectively. According to the generative process of topic model in Figure 1, each observation, *i.e.*, the occurrence of a word $w \in \mathcal{W}$ in a document $d \in \mathcal{D}$, is associated with an unobserved variable $z \in \mathcal{Z} = \{z_1, z_2, \cdots, z_K\}$, where K refers to the number of topics. Based on the *conditional independence assumption* that words $w \in \mathcal{W}$ are generated independently of the specific documents $d \in \mathcal{D}$ given the topic $z \in \mathcal{Z}$, *i.e.*, $p(w, d|z) = p(w|z)p(d|z)$ and $p(w|d, z) = p(w|z)$, the generative process can be translated into the corresponding joint probability model $p(w_i, d_j) = \sum_k p(w_i|z_k)p(z_k)p(d_j|z_k)$.

Let $X = [x_{ij}]_{m\times n}$ whose element x_{ij} denotes the term frequency for observation pair (w_i, d_j), *i.e.*, the number of times word w_i occurred in document d_j. Traditional topic models learn the distributions $p(d), p(z|d)$ and $p(w|z)$ by maximizing the following log-likelihood function

$$\mathcal{J}_{PLSA} = \sum_i \sum_j x_{ij} \log \sum_k p(w_i|z_k)p(z_k)p(d_j|z_k). \quad (1)$$

It is noteworthy that the equations $p(\mathcal{W}, d_j) = \prod_i p(w_i, d_j)$ and $p(w_i, \mathcal{D}) = \prod_j p(w_i, |d_j)$ $(\forall j, i)$ are explicitly used in (1). As a result, topic model with maximum likelihood estimate inevitably suffer from the assumption of conditional independence on words, and thus entirely ignores the correlations among words (Steyvers and Griffiths 2007).

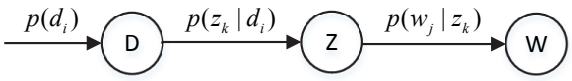


Figure 1: Generative process representation of topic model.

NMF for Topic Modeling

As a widely used dimension reduction technique, standard NMF performs well for document clustering and topic modeling (Arora et al. 2013; Kuang, Choo, and Park 2015; Lee and Seung 2001). Considering the non-robustness of squared ℓ_2-norm to outliers and noises, some robust NMF methods are exploited by using a robust error functions (Ding and Kong 2012; Kong, Ding, and Huang 2011), or via half-quadratic minimization (Du, Li, and Shen 2012). However, these NMF based methods for topic clustering perform less interpretatively since they lack the explicit probabilistic meaning of each factor. Note that studies in (Ding, Li, and Peng 2006; Gaussier and Goutte 2005) pointed out PLSI and KL-divergence based NMF indeed optimize the same objective function with additional constraint. Although KL-divergence based NMF inherits probabilistic meaning of topic model, the corresponding algorithms are typically much slower than those for standard NMF (Xie, Song, and Park 2013). Additionally, it is difficult to extend KL-divergence based NMF to its robust version.

Probabilistic NMF (PNMF)

For a probabilistic interpretation, we normalize the observation matrix $X = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n] \in \mathbb{R}^{m\times n}$ such that each column $\mathbf{x}_j = [x_{1j}, x_{2j}, \cdots, x_{mj}]^\top \in \mathbb{R}^m$ refers to a distribution of words in the j-th document *i.e.*, $\sum_i x_{ij} = 1$ for $j = 1, 2, \cdots, n$. In this sense, the equation $\mathbf{1}^\top X = \mathbf{1}^\top$ holds naturally. Let $U = [u_{ik}]_{m\times K}$ with $u_{ik} = p(w_i|z_k)$ and $V = [v_{jk}]_{n\times K}$ with $v_{jk} = p(z_k|d_j)$. Then the conditional distribution $p(w_i|d_j)$ can be formulated in terms of matrix notation, *i.e.*, $p(w_i|d_j) = \sum_k p(w_i|z_k)p(z_k|d_j) = (UV^\top)_{ij}$. Based on this matrix notation, we employ ℓ_F-norm based error measurement and directly approximate the conditional distribution $p(w|d)$ to actual normalized observation matrix X by solving the following problem

$$\min_{U\geq 0, V\geq 0, \mathbf{1}^\top U = \mathbf{1}, V\mathbf{1} = \mathbf{1}} \mathcal{J}_{\ell_F}^{NMF} = \|X - UV^\top\|_F^2 \quad (2)$$

where the accompanying constraints on decomposed matrices U and V are significantly important since they characterize the explicit probabilistic meaning of factors, and simultaneously facilitate to ensure the equation $\mathbf{1}^\top UV^\top = \mathbf{1}^\top$. For a better understanding, we demonstrate the framework in Figure 2. The optimization problem (2) attempts to find a low-dimensional representation for topic modeling via splitting the normalized observation X into two non-negative

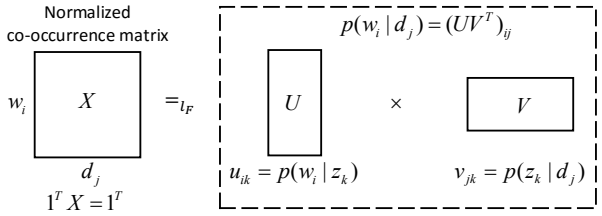


Figure 2: Framework of proposed probabilistic NMF.

matrices with probabilistic constraints. In contrast to the objective $\mathcal{J}_{KL}^{NMF}$, the proposed framework not only inherits the explicit probabilistic meaning of factors but also denies the assumption of conditional independence on words.

Optimization Procedure

Note that the objective $\mathcal{J}_{\ell_F}^{NMF}$ is convex in U only or V only, but not convex in both variables together. It is unrealistic to expect an algorithm to solve problem (2) in the sense of finding global minima (Lee and Seung 2001). Thanks to the independence of $p(w_i|z_k)$ and $p(z_k|d_j)$, we address the proposed problem (2) with alternative optimizing algorithm.

Update V with fixed U: For the objective $\mathcal{J}_{\ell_F}^{MFI} = \sum_{j=1}^{n} \|U(\mathbf{v}^j)^\top - \mathbf{x}_j\|_2^2$, we update each row of V each time while fixing the others, *i.e.*,

$$\min_{\mathbf{v}^j \geq \mathbf{0}, \mathbf{v}^j \mathbf{1}=1} \|U(\mathbf{v}^j)^\top - \mathbf{x}_j\|_2^2 \tag{3}$$

where $\mathbf{v}^j$ is the j-th row of matrix V. Removing the unrelated terms to variable $\mathbf{v}^j$, the optimization problem (3) can be reformulated as

$$\min_{\mathbf{v}^j \geq \mathbf{0}, \mathbf{v}^j \mathbf{1}=1} \varphi_j(\mathbf{v}^j) = \mathbf{v}^j Q (\mathbf{v}^j)^\top - 2\mathbf{v}^j \mathbf{c}_j \tag{4}$$

where $Q = U^\top U \in \mathbb{R}^{K \times K}$ and $\mathbf{c}_j = U^\top \mathbf{x}_j \in \mathbb{R}^K$. Subsequently, we propose an Accelerated Projected Gradient (APG) method to solve optimization problem (4) for its simplicity and efficiency (Nesterov 2007). Based on APG method, let the second order Taylor expansion of objective function $\varphi_j(\cdot)$ about an auxiliary variable $\mathbf{y}_t$ in the t-th iteration be $\varphi_j^L(\mathbf{v}^j) = \varphi(\mathbf{y}_t) + \langle \nabla \varphi_j(\mathbf{y}_t), \mathbf{v}^j - \mathbf{y}_t \rangle + \frac{L}{2}\|\mathbf{v}^j - \mathbf{y}_t\|_2^2$. We update $\mathbf{v}^j$ by solving an easier optimization problem $\min_{\mathbf{v}^j \geq \mathbf{0}, \mathbf{v}^j \mathbf{1}=1} \varphi_j^L(\mathbf{v}^j)$ which is equivalent to a Euclidean projection problem on the simplex space, *i.e.*,

$$\mathbf{v}_{t+1}^j = \min_{\mathbf{v}^j \geq \mathbf{0}, \mathbf{v}^j \mathbf{1}=1} \|\mathbf{v}^j - \mathbf{h}_t^j\|_2^2 \tag{5}$$

where $\mathbf{h}_t^j = \mathbf{y}_t - \frac{1}{L}\nabla \varphi_j(\mathbf{y}_t) = \mathbf{y}_t - \frac{2}{L}(\mathbf{y}_t Q - \mathbf{c}_j^\top) \in \mathbb{R}^K$; $L \geq 0$ is a constant. Note that some efficient algorithms for the Euclidean projection problem on the simplex space have been studied thoroughly in the past decades. Interested readers please refer to (Condat 2016; Duchi et al. 2008; Becker et al. 2013; Censor et al. 2012) for more details.

To approximate the auxiliary variable to the solution $\mathbf{v}^j$, we update $\mathbf{y}^t$ according to the following formula

$$\mathbf{y}_{t+1} = \mathbf{v}_t^j + \frac{\sigma_t - 1}{\sigma_{t+1}}(\mathbf{v}_{t+1}^j - \mathbf{v}_t^j) \tag{6}$$

Algorithm 1 APG algorithm for optimization problem (3).

Input: X, Q, $\mathbf{c}$.
Initialize: $\mathbf{v}_0^j = \mathbf{y}_0$, $\sigma_0 = 1$, $t = 0$.
1: **while** not converge **do**
2: Update $\mathbf{v}_{t+1}^j = \min_{\mathbf{v}^j \geq \mathbf{0}, \mathbf{v}^j \mathbf{1}=1} \|\mathbf{v}^j - \mathbf{h}_t^j\|_2^2$ with $\mathbf{h}_t^j = \mathbf{y}_t - \frac{2}{L}(\mathbf{y}_t Q - \mathbf{c}_j^\top)$;
3: Update $\sigma_{t+1} = \frac{1+\sqrt{1+4\sigma_t^2}}{2}$;
4: Update $\mathbf{y}_{t+1} = \mathbf{v}_t^j + \frac{\sigma_t - 1}{\sigma_{t+1}}(\mathbf{v}_{t+1}^j - \mathbf{v}_t^j)$;
5: $t = t + 1$.
6: **end while**

where the acceleration coefficient σ is updated through

$$\sigma_{t+1} = \frac{1}{2}(1 + \sqrt{1 + 4\sigma_t^2}). \tag{7}$$

In summary, we describe the APG algorithm for optimization problem (3) in Algorithm 1. It has been pointed out in (Nesterov 2007) that the APG algorithm converges fast.

Update U with fixed V: We rewrite the objective $\mathcal{J}_{\ell_F}^{NMF} = \|X - \sum_k \mathbf{u}_k \mathbf{v}_k^\top\|_F^2$ and update one column of U each time while fixing the other columns, *i.e.*,

$$\min_{\mathbf{u}_k \geq \mathbf{0}, \mathbf{1}^\top \mathbf{u}_k = 1} \phi_k(\mathbf{u}_k) = \|\mathbf{u}_k \mathbf{v}_k^\top - H_k\|_F^2 \tag{8}$$

for $k = 1, 2, \cdots, K$, where $H_k = \sum_{i \neq k} \mathbf{u}_i \mathbf{v}_i^\top - X \in \mathbb{R}^{m \times n}$. This problem can be rewritten as a Euclidean projection problem on the simplex space according to the following Theorem 1.

Theorem 1. *The solution of optimization problem (8) is equivalent to the solution of optimization problem* $\min_{\mathbf{u}_k \geq \mathbf{0}, \mathbf{u}_k^\top \mathbf{1}=1} \|\mathbf{u}_k - \mathbf{h}_k\|_2^2$ *with* $\mathbf{h}_k = \frac{1}{\|\mathbf{v}_k\|_2^2} H_k \mathbf{v}_k \in \mathbb{R}^m$.

Proof. According to some nice properties of trace operator, we reformulate the objective function of problem (8) as

$$\begin{aligned}\phi_k(\mathbf{u}_k) &= Tr\left[\mathbf{u}_k \mathbf{v}_k^\top \mathbf{v}_k \mathbf{u}_k^\top - 2\mathbf{u}_k \mathbf{v}_k^\top H_k^\top + H_k H_k^\top\right] \\ &= \|\mathbf{v}_k\|_2^2 \; Tr[\mathbf{u}_k \mathbf{u}_k^\top - \frac{2}{\|\mathbf{v}_k\|_2^2}(H_k \mathbf{v}_k)^\top \mathbf{u}_k + \frac{1}{\|\mathbf{v}_k\|_2^2} H_k H_k^\top].\end{aligned}$$

Removing the unrelated terms in $\phi_k(\mathbf{u}_k)$ with respect to variable $\mathbf{u}_k$, the problem (8) is equivalent to a Euclidean projection problem on the simplex space, *i.e.*, $\min_{\mathbf{u}_k \geq \mathbf{0}, \mathbf{1}^\top \mathbf{u}_k = 1} \|\mathbf{u}_k - \mathbf{h}_k\|_F^2$, where $\mathbf{h}_k = \frac{1}{\|\mathbf{v}_k\|_2^2} H_k \mathbf{v}_k$. The proof is completed. □

In summary, we describe the alternative algorithm for probabilistic NMF of topic model in Algorithm 2. Note that the main computational cost of Algorithm 2 lies in solving the Euclidean projection problems on the simplex spaces in $\mathbb{R}^m$ and $\mathbb{R}^K$ respectively, where the number of topics K is much smaller than the number of words m in dictionary. In this paper, we solve this problem according to the fast algorithm proposed in (Condat 2016), which performs efficiently for large-scale problems with large dictionaries, with complexity $O(m)$ or $O(m \log m)$. As a result, thanks to the convergence of APG algorithm as well as the efficient optimization of Euclidean projection problem on the simplex space, the proposed Algorithm 2 performs well in practice.

Algorithm 2 Alternative Algorithm for problem (2)

Input: Normalized X satisfying $\mathbf{1}^\top X = \mathbf{1}^\top$.
Initialize: U^0, $t = 0$.
1: **while** not converge **do**
2: $\forall j$, update $\mathbf{v}_{t+1}^j$ with APG Algorithm 1;
3: $\forall k$, update $\mathbf{u}_k^{t+1} = \min_{\mathbf{u}_k \geq \mathbf{0}, \mathbf{1}^\top \mathbf{u}_k = \mathbf{1}} \|\mathbf{u}_k - \mathbf{h}_k^{t+1}\|_2^2$ with $\mathbf{h}_k^{t+1} = \frac{1}{\|\mathbf{v}_k^{t+1}\|_2^2} H_k^{t+1} \mathbf{v}_k^{t+1}$ and $H_k^{t+1} = \sum_{i \neq k} \mathbf{u}_i^t (\mathbf{v}_i^{t+1})^\top - X$;
4: $t = t + 1$.
5: **end while**

For a new document d' with normalized observation $\mathbf{x}' = [x'_1, x'_2, \cdots, x'_m]^\top$ ($\mathbf{1}^\top \mathbf{x}' = 1$) out of the training dataset, we learn its topic distribution $\mathbf{v}' = [p(z_1|d'), p(z_2|d'), \cdots, p(z_K|d')]$ by solving optimization problem $\mathbf{v}' = \min_{\mathbf{v}' \geq 0, \mathbf{v}'\mathbf{1}=1} \|U(\mathbf{v}')^\top - \mathbf{x}'\|_2^2$ with APG Algorithm 1, where the parameter U is estimated over training dataset.

Robust Probabilistic NMF

In this section, we extend the proposed probabilistic NMF of topic model to its robust versions which approximate the observation matrix with $\ell_{2,1}$-norm (PNMF$_{R1}$) and capped $\ell_{2,1}$-norm (PNMF$_{R2}$), respectively.

Robust Approximation with $\ell_{2,1}$-norm

We assume the fitting error follows Laplacian distribution. The idea of robust probabilistic NMF with $\ell_{2,1}$-norm can be formulated as

$$\min_{U \geq 0, V \geq 0, \mathbf{1}^\top U = \mathbf{1}, V\mathbf{1} = \mathbf{1}} \Psi(U, V) = \|X - UV^\top\|_{2,1} \quad (9)$$

where the $\ell_{2,1}$-norm of matrix W is defined as $\|W\|_{2,1} = \sum_j \|\mathbf{w}_j\|_2$, which satisfies the three conditions of norm (Kong, Ding, and Huang 2011). The $\ell_{2,1}$-norm based error measurement isolates the outliers by imposing sparsity on the corresponding column of error matrix $X - UV^\top$ (Liu, Liu, and Sun 2015; Chang et al. 2014; Nie et al. 2010). Let $f(x) = \sqrt{x}$ and $g_j(U, \mathbf{v}_j) = \|\mathbf{x}_j - U(\mathbf{v}^j)^\top\|_2^2$, then the objective function $\Psi(U, V) = \sum_{j=1}^n \|\mathbf{x}_j - U(\mathbf{v}^j)^\top\|_2 = \sum_{j=1}^n f(g_j(U, \mathbf{v}_j))$, where f is a concave function. As a result, we follow the **concave duality** theorem (Zhang 2009) and exploit an efficient re-weighted Algorithm 3 to address optimization problem (9), where the key step lies in solving the following optimization problem

$$\min_{U \geq 0, V \geq 0, \mathbf{1}^\top U = \mathbf{1}, V\mathbf{1} = \mathbf{1}} \sum_{j=1}^n f'(g_j(U, \mathbf{v}_j)) g_j(U, \mathbf{v}_j). \quad (10)$$

where $\delta_j = f'(g_j(U, \mathbf{v}_j))$ denotes the super-gradient of concave function f at point $g_j(U, \mathbf{v}_j)$, *i.e.*,

$$\delta_j = \frac{1}{2\|\mathbf{x}_j - U(\mathbf{v}^j)^\top\|_2}. \quad (11)$$

Algorithm 3 Re-weighted Algorithm for problem (9)

Input: X satisfying $\mathbf{1}^\top X = \mathbf{1}^\top$.
Initialize: $\delta_j^0 = I$ $(\forall j)$, $t = 0$
1: **while** not converge **do**
2: Update U^{t+1} and V^{t+1} by solving problem (10);
3: Update weight δ_j^{t+1} $(\forall j)$ according to Eq. (11);
4: $t = t + 1$.
5: **end while**
Output: U, V.

Algorithm 4 Alternative Algorithm for problem (10)

Input: Normalized X satisfying $\mathbf{1}^\top X = \mathbf{1}^\top$.
1: **while** not converge **do**
2: $\forall j$, update $\mathbf{v}_{t+1}^j$ with APG Algorithm 1;
3: $\forall k$, update $\mathbf{u}_k^{t+1} = \min_{\mathbf{u}_k \geq \mathbf{0}, \mathbf{1}^\top \mathbf{u}_k = \mathbf{1}} \|\mathbf{u}_k - \mathbf{h}_k^{t+1}\|_2^2$ with $\mathbf{h}_k^{t+1} = \frac{1}{(\mathbf{v}_k^{t+1})^\top \Delta \mathbf{v}_k^{t+1}} M_k^{t+1} \Delta \mathbf{v}_k^{t+1}$ and $M_k^{t+1} = X - \sum_{i \neq k} \mathbf{u}_i^t (\mathbf{v}_i^{t+1})^\top$;
4: $t = t + 1$.
5: **end while**

We specify the objective function above as $\Psi_\delta(U, V) = \sum_{j=1}^n \delta_j \|\mathbf{x}_j - U(\mathbf{v}^j)^\top\|_2^2 = \|(X - UV^\top)\Delta^{\frac{1}{2}}\|_F^2$, where $\Delta = diag(\delta_1, \delta_2, \cdots, \delta_n)$ is a $n \times n$ diagonal matrix. In this case, the problem (10) is very similar to the problem (2) expect for a multiplier Δ. In this paper, we solve the problem (10) by iteratively updating U and V with the other one fixed until convergence. When variable U is fixed, the optimization problem (10) with respect to each row of matrix V turns to optimization problem (4), and thus can be solved efficiently. With fixed V, since $\forall k$, $\Psi_\delta(U, V) = \|\mathbf{u}_k \mathbf{v}_k^\top \Delta^{\frac{1}{2}} - M_k \Delta^{\frac{1}{2}}\|_F^2$ holds with $M_k = X - \sum_{i \neq k} \mathbf{u}_i \mathbf{v}_i^\top \in \mathbb{R}^{m \times n}$, we update the k-th column of U while keeping the other columns fixed, by solving the following problem

$$\min_{\mathbf{u}_k \geq \mathbf{0}, \mathbf{1}^\top \mathbf{u}_k = 1} \|\mathbf{u}_k \mathbf{v}_k^\top \Delta^{\frac{1}{2}} - M_k \Delta^{\frac{1}{2}}\|_F^2. \quad (12)$$

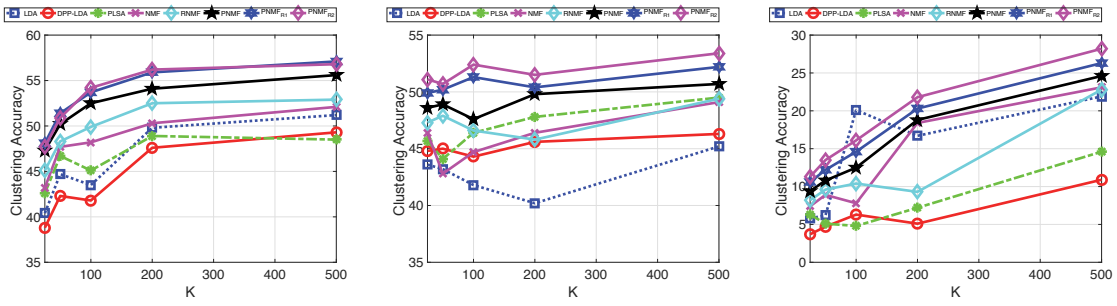
According to Theorem 1, this problem is equivalent to a Euclidean projection problem on the simplex space, *i.e.*, $\min_{\mathbf{u}_k \geq \mathbf{0}, \mathbf{1}^\top \mathbf{u}_k = 1} \|\mathbf{u}_k - \mathbf{h}_k\|_2^2$, where $\mathbf{h}_k = \frac{1}{\mathbf{v}_k^\top \Delta \mathbf{v}_k} M_k \Delta \mathbf{v}_k \in \mathbb{R}^m$. We summarize the alternative algorithm for optimization problems (10) in Algorithm 4.

Robust Approximation with Capped $\ell_{2,1}$-norm

Consider the better robustness of capped $\ell_{2,1}$-norm, (Zhang 2013) and (Gong, Ye, and Zhang 2013), we go further to use capped $\ell_{2,1}$-norm to measure the error of observation matrix approximation. This idea can be formulated as the following optimization problem

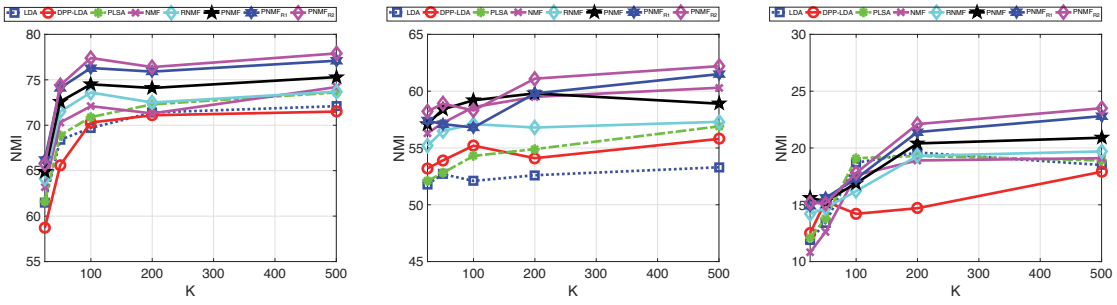
$$\min_{U \geq 0, V \geq 0, \mathbf{1}^\top U = \mathbf{1}, V\mathbf{1} = \mathbf{1}} \sum_{j=1}^n \min(\|\mathbf{x}_j - U(\mathbf{v}^j)^\top\|_2, \theta) \quad (13)$$

where $\theta > 0$ is a thresholding parameter. Note that this optimization problem turns to the $\ell_{2,1}$-norm based probelm (9) if thresholding θ is set as positive infinite.



(a) Reuters (b) TDT (c) 20 Newsgroups

Figure 3: Clustering performance in terms of ACC on three datasets. Performance is shown in percentages.



(a) Reuters (b) TDT (c) 20 Newsgroups

Figure 4: Clustering performance in terms of NMI on three datasets. Performance is shown in percentages.

Let functions $f_\theta(x) = \min(\sqrt{x}, \theta)$ and $g_j(U, \mathbf{v}_j) = \|\mathbf{x}_j - U(\mathbf{v}^j)^\top\|_2^2$. Then the objective of optimization problem (13) can be rewritten as $\Phi(U, V) = \sum_j f_\theta(g_j(U, \mathbf{v}_j))$, where $f_\theta(x)$ is a concave function. As a result, we use the re-weighted Algorithm 3 to solve the non-convex and non-smooth challenging problem (13), and the key step lies in solving optimization problem

$$\min_{U\geq 0, V\geq 0, \mathbf{1}^\top U=\mathbf{1}, V\mathbf{1}=\mathbf{1}} \sum_j \delta_j \|\mathbf{x}_j - U(\mathbf{v}^j)^\top\|_2^2 \qquad (14)$$

according to Algorithm 4, where $\delta_j = f'_\theta(g_j(U, \mathbf{v}_j))$ is calculated by

$$\delta_j = \begin{cases} \frac{1}{2\|\mathbf{x}_j - U(\mathbf{v}^j)^\top\|_2}, & \|\mathbf{x}_j - U(\mathbf{v}^j)^\top\|_2^2 \leq \theta; \\ 0, & \|\mathbf{x}_j - U(\mathbf{v}^j)^\top\|_2^2 \geq \theta. \end{cases} \qquad (15)$$

Note that the determination of topic distribution for a new document d' does not depend on the re-weighting matrix Δ for both of the proposed robust probabilistic NMF of topic model. Therefore, given the estimated U from training dataset, the procedure of testing for a new document is identical to the probabilistic NMF based on ℓ_F-norm.

Experiment

In this section, we present experimental results on three benchmark datasets in terms of three evaluation metrics, which demonstrate the effectiveness and superiority of the proposed approaches.

Datasets Description and Baseline Methods

Three datasets are utilized in the experiments. The first dataset is the Reuters dataset (Cai and He 2012). We remove the categories with less than 100 documents, resulting in 9 categories and 7,195 documents. 70% documents are used for training, and the rest are used for testing. The second dataset is a subset of the NIST Topic Detection and Tracking (TDT) corpus (Cai, He, and Han 2005) which consists of 9,394 documents from the largest 30 categories. In a similar fashion, we use 70% documents for training and the rest for testing. The third dataset is the 20 Newsgroups (20-News), which contains 18,846 documents from 20 categories. 60% documents are used for training, and the rest are used for testing. To focus on the important words, we remove the stop words and use a vocabulary of 5,000 words with the largest document frequency.

We compared the proposed three alternatives, *i.e.* PNMF, PNMF$_{R1}$ and PNMF$_{R2}$, with the following baseline methods: Latent Dirichlet Allocation (LDA) (Blei, Ng, and Jordan 2003), LDA regularized with Determinantal Point Process prior (DPP-LDA) (Zou and Adams 2012), PLSA (Hofmann, Puzicha, and Jordan 1999), NMF for Latent Semantic Analysis (LSA) (Stevens et al. 2012) and robust NMF with $\ell_{2,1}$-norm (RNMF)(Kong, Ding, and Huang 2011). The parameters in the compared approaches are tuned using 5-fold cross validation. The value of θ used in PNMF$_{R2}$ is set according to the ratio of the outliers (between 0.03 and

Table 1: Perplexity on three datasets. Smaller number indicates better performance.

		LDA	DPP-LDA	PLSA	NMF	RNMF	PNMF	PNMF$_{R1}$	PNMF$_{R2}$
Reuters	$K = 25$	1357	1348	1336	1275	1232	1211	1202	**1179**
	$K = 50$	1339	1331	1324	1267	1205	1182	1173	**1152**
	$K = 100$	1328	1322	1313	1258	1174	1158	1127	**1103**
	$K = 200$	935	926	921	846	817	786	757	**732**
	$K = 500$	911	903	895	782	764	755	709	**684**
TDT	$K = 25$	2712	2678	2654	2432	2413	2057	1847	**1802**
	$K = 50$	2713	2681	2657	2416	2416	2032	1832	**1815**
	$K = 100$	2716	2684	2645	2418	2419	2016	1816	**1793**
	$K = 200$	2722	2686	2659	2404	2395	2011	1785	**1743**
	$K = 500$	2435	2412	2384	2219	2174	1843	1638	**1596**
20 Newsgroups	$K = 25$	874	863	867	832	826	**789**	801	794
	$K = 50$	876	859	866	827	815	803	795	**782**
	$K = 100$	876	847	853	811	802	785	778	**762**
	$K = 200$	857	842	849	789	774	792	762	**730**
	$K = 500$	773	751	765	712	701	684	659	**624**

0.05). The number of topics for each dataset varies among $\{25, 50, 100, 200, 500\}$.

Clustering Performance

In this section, we verify the effectiveness of the learned representations for testing data on k-means clustering. In the experiment, we fix the number of clusters as the ground truth number of categories. For a fair comparison, we repeat the k-means 50 times with different random initializations and return the solution with the lowest loss value. Following related studies (Cai, He, and Han 2005; Blei, Ng, and Jordan 2003) on topic models for clustering, we leverage two popular evaluation metrics, namely clustering accuracy (ACC) and normalized mutual information (NMI) (Cai, Zhang, and He 2010), to measure the performance of clustering on the testing data after various unsupervised dimension reduction approaches. The experimental results of values of ACC and NMI over three datasets are reported in Figure 3 and Figure 4, respectively. We observe from the experimental results that: (1) The values of ACC and NMI regarding different methods have an overall tendency to grow as the increase topics. However, the growth rate slows down when more topics involved. (2) Since the proposed probabilistic NMF approaches do not rely on the conditional independence of words, they achieve comparable even better performance in contrast to LDA, DPP-LDA, and PLSA. (3) Due to the probabilistic constraint on the components of U and V, the proposed probabilistic NMF methods perform better than LSA with NMF and RNMF. It is noteworthy that both RNMF and the proposed PNMF$_{R1}$ employ the $\ell_{2,1}$-norm for the loss measurement; however, PNMF$_{R1}$ achieves better performances thanks to the probabilistic constraints. (4) The proposed PNMF$_{R2}$ algorithm outperform other approaches in almost all the cases on both ACC and NMI, followed by PNMF$_{R1}$ among the compared methods.

Perplexity on Testing Data

Following the setting in (Salakhutdinov and Hinton 2009), we further compute the perplexity on the held-out test set to assess the document modeling power of the compared algorithms. A lower perplexity score indicates better generalization performance. The experimental results on the three benchmark datasets are reported in Table 1. For each model, we observe that the value of perplexity decrease with the increase of topics over all datasets. Thanks to the robustness of $\ell_{2,1}$ and capped $\ell_{2,1}$-norm based error measurement, the proposed robust extensions of probabilistic NMF models achieve the best performance among the compared methods over nearly all of the datasets. Additionally, the proposed probabilistic NMF with ℓ_F-norm also show better performance than other conventional methods which rely on the conditional independence of words given the documents topic distribution.

Conclusion

In this paper, we proposed three new and straightforward NMF for the topic model by directly approximating the word-document co-occurrence matrix with matrix decomposition technique with probabilistic constraints. The proposed framework inherits the explicit probabilistic meaning of factors in topic models and simultaneously makes the conditional independence assumption on words unnecessary. Intuitively, it is capable of being extended to its robust versions. Some efficient algorithms are exploited to solve the proposed optimization problems. Experimental results on some benchmark datasets illustrate the effectiveness and superiority of the proposed methods.

Acknowledgments

This research was funded in part by National Science Foundation of China (Nos 61502377, 61532015 and 61532004), the National Key Research and Development Program of

China (No. 2016YFB1000903), China Postdoctoral Science Foundation (No. 2015M582662), the National Science Foundation (No. IIS-1638429) and the U. S. Army Research Office (W911NF-13-1-0277).

References

Airoldi, E. M.; Blei, D. M.; Fienberg, S. E.; and Xing, E. P. 2009. Mixed membership stochastic blockmodels. In *NIPS*.

Arora, S.; Ge, R.; Kannan, R.; and Moitra, A. 2012. Computing a nonnegative matrix factorization–provably. In *STOC*.

Arora, S.; Ge, R.; Halpern, Y.; Mimno, D. M.; Moitra, A.; Sontag, D.; Wu, Y.; and Zhu, M. 2013. A practical algorithm for topic modeling with provable guarantees. In *ICML*.

Becker, S.; Cevher, V.; Koch, C.; and Kyrillidis, A. 2013. Sparse projections onto the simplex. *ICML*.

Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.* 3:993–1022.

Cai, D., and He, X. 2012. Manifold adaptive experimental design for text categorization. *IEEE Trans. Knowl. Data Eng.* 24(4):707–719.

Cai, D.; He, X.; and Han, J. 2005. Document clustering using locality preserving indexing. *IEEE Trans. Knowl. Data Eng.* 17(12):1624–1637.

Cai, D.; Zhang, C.; and He, X. 2010. Unsupervised feature selection for multi-cluster data. In *ACM SIGKDD*.

Censor, Y.; Chen, W.; Combettes, P. L.; Davidi, R.; and Herman, G. T. 2012. On the effectiveness of projection methods for convex feasibility problems with linear inequality constraints. *Computational Optimization and Applications* 51(3):1065–1088.

Chang, X.; Nie, F.; Yang, Y.; and Huang, H. 2014. A convex formulation for semi-supervised multi-label feature selection. In *AAAI*.

Condat, L. 2016. Fast projection onto the simplex and the l1 ball. *Mathematical Programming, Series A* 158(1):575–585.

Ding, C., and Kong, D. 2012. Nonnegative matrix factorization using a robust error function. In *ICASSP*.

Ding, C. H. Q.; Li, T.; and Peng, W. 2006. Nonnegative matrix factorization and probabilistic latent semantic indexing: Equivalence chi-square statistic, and a hybrid method. In *AAAI*.

Du, L.; Li, X.; and Shen, Y.-D. 2012. Robust nonnegative matrix factorization via half-quadratic minimization. In *ICDM*.

Duchi, J.; Shalev-Shwartz, S.; Singer, Y.; and Chandra, T. 2008. Efficient projections onto the l 1-ball for learning in high dimensions. In *ICML*, 272–279.

Gaussier, E., and Goutte, C. 2005. Relation between plsa and nmf and implications. In *ACM SIGIR*.

Gong, P.; Ye, J.; and Zhang, C. 2013. Multi-stage multi-task feature learning. *J. Mach. Learn. Res.* 14(1):2979–3010.

Hoffman, M.; Blei, D.; and Cook, P. R. 2009. Finding latent sources in recorded music with a shift-invariant hdp. In *DAFx*.

Hofmann, T.; Puzicha, J.; and Jordan, M. I. 1999. Learning from dyadic data. In *NIPS*.

Hofmann, T. 1999. Probabilistic latent semantic indexing. In *ACM SIGIR*. ACM.

Kivinen, J. J.; Sudderth, E. B.; and Jordan, M. I. 2007. Learning multiscale representations of natural scenes using dirichlet processes. In *ICCV*.

Kong, D.; Ding, C.; and Huang, H. 2011. Robust nonnegative matrix factorization using l21-norm. In *CIKM*.

Kuang, D.; Choo, J.; and Park, H. 2015. Nonnegative matrix factorization for interactive topic modeling and document clustering. In *Partitional Clustering Algorithms*. 215–243.

Landauer, T. K.; Foltz, P. W.; and Laham, D. 1998. An introduction to latent semantic analysis. *Discourse processes* 25(2-3):259–284.

Lee, D. D., and Seung, H. S. 2001. Algorithms for nonnegative matrix factorization. In *NIPS*.

Liu, H.; Liu, Y.; and Sun, F. 2015. Robust exemplar extraction using structured sparse coding. *IEEE Trans. Neural Netw. Learn. Syst.* 26(8):1816–1821.

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

Nesterov, Y. 2007. Gradient methods for minimizing composite objective function. *CORE*.

Nie, F.; Huang, H.; Cai, X.; and Ding, C. H. 2010. Efficient and robust feature selection via joint $\ell_{2,1}$-norms minimization. In *NIPS*.

Nie, F.; Wang, H.; Huang, H.; and Ding, C. H. 2013. Adaptive loss minimization for semi-supervised elastic embedding. In *IJCAI*.

Salakhutdinov, R., and Hinton, G. E. 2009. Replicated softmax: an undirected topic model. In *NIPS*.

Stevens, K.; Kegelmeyer, P.; Andrzejewski, D.; and Buttler, D. 2012. Exploring topic coherence over many models and many topics. In *EMNLP-CoNLL*.

Steyvers, M., and Griffiths, T. 2007. Probabilistic topic models. *Handbook of Latent Semantic Analysis* 427(7):424–440.

Wang, C.; Blei, D.; and Li, F.-F. 2009. Simultaneous image classification and annotation. In *CVPR*.

Xie, B.; Song, L.; and Park, H. 2013. Topic modeling via nonnegative matrix factorization on probability simplex. In *NIPS Workshop on Topic Models: Computation, Application, and Evaluation*.

Zhang, T. 2009. Multi-stage convex relaxation for learning with sparse regularization. In *NIPS*.

Zhang, T. 2013. Multi-stage convex relaxation for feature selection. *Bernoulli* 19(5B):2277–2293.

Zou, J. Y., and Adams, R. P. 2012. Priors for diversity in generative latent variable models. In *NIPS*.