Graph Regularized Non-negative Matrix Factorization for Data Representation[*]

Deng Cai[†] Xiaofei He[‡] Jiawei Han[†]

[†] Department of Computer Science, University of Illinois at Urbana-Champaign
[‡] College of Computer Science, Zhejiang University, China

February 2008

Abstract

Recently Non-negative Matrix Factorization (NMF) has received a lot of attentions in information retrieval, computer vision and pattern recognition. NMF aims to find two non-negative matrices whose product can well approximate the original matrix. The sizes of these two matrices are usually smaller than the original matrix. This results in a compressed version of the original data matrix. The solution of NMF yields a natural parts-based representation for the data. When NMF is applied for data representation, a major disadvantage is that it fails to consider the geometric structure in the data. In this paper, we develop a graph based approach for parts-based data representation in order to overcome this limitation. We construct an affinity graph to encode the geometrical information and seek a matrix factorization which respects the graph structure. We demonstrate the success of this novel algorithm by applying it on real world problems.

1 Introduction

The techniques of matrix factorization have become popular in recent years for data representation. In many problems in information retrieval, computer vision and pattern recognition, the input data matrix is of very high dimension. This makes *learning from example* infeasible. One hopes then to find two or more lower

[*]The work was supported in part by the U.S. National Science Foundation NSF IIS-05-13678, NSF BDI-05-15813 and MIAS (a DHS Institute of Discrete Science Center for Multimodal Information Access and Synthesis). Any opinions, findings, and conclusions or recommendations expressed here are those of the authors and do not necessarily reflect the views of the funding agencies.

dimensional matrices whose product provides a good approximation to the original matrix. The canonical matrix factorization techniques include LU-decomposition, QR-decomposition, Cholesky decomposition, and Singular Value Decomposition (SVD).

SVD is one of the most frequently used matrix factorization tool. A singular value decomposition of an $m \times n$ matrix **X** is any factorization of the form

$$\mathbf{X} = \mathbf{U}\mathbf{S}\mathbf{V}^T$$

where **U** is an $m \times m$ orthogonal matrix, **V** is an $n \times n$ orthogonal matrix, and **S** is an $m \times n$ diagonal matrix with $\mathbf{S}_{ij} = 0$ if $i \neq j$ and $\mathbf{S}_{ij} \geq 0$. The quantities $\mathbf{S}_{ii}$ are called the *singular values* of **X**, and the columns of **U** and **V** are called left and right *singular vectors*, respectively. By removing those singular vectors corresponding to sufficiently small singular value, we get a natural low-rank approximation to the original matrix. This approximation is optimal in the sense of reconstruction error and thus optimal for data representation when Euclidean structure is concerned. For this reason, SVD has been applied to various real world applications, such as face recognition (*Eigenface*, [16]) and document representation (*Latent Semantic Indexing*, [6]).

Previous studies have shown there is psychological and physiological evidence for parts-based representation in human brain [13], [17], [11]. The Non-negative Matrix Factorization (NMF) algorithm is proposed to learn the parts of objects like human faces and text documents [8]. NMF aims to find two non-negative matrices whose product provides a good approximation to the original matrix. The non-negative constraints lead to a parts-based representation because they allow only additive, not subtractive, combinations. NMF has been shown to be superior to SVD in face recognition [10] and document clustering [18]. NMF is optimal for learning the parts of objects. However, it fails to consider the geometrical structure of the data space which is essential for data clustering and classification problems.

In this paper, we propose a novel algorithm, called Graph regularized Non-negative Matrix Factorization (GNMF), to overcome the limitation of NMF. We encode the geometrical information of the data space by constructing a nearest neighbor graph. One hopes then to find a new representation space in which two data points are sufficiently close to each other if they are connected in the graph. To achieve this, we design a new matrix factorization objective function and incorporates the graph structure into it. We also develop a optimization scheme to solve the object function based on iterative updates of the two factor matrices. This leads to a new parts-based data representation which respects the geometrical structure of the data space. The convergence proof of our optimization scheme is provided.

The rest of the paper is organized as follows: in Section 2, we give a brief review of NMF. Section 3 introduces our algorithm and give a convergence proof of our optimization scheme. Extensive experimental results on document clustering are presented in Section 4. Finally, we provide some concluding remarks and suggestions for future work in Section 5.

2 A Brief Review of NMF

Non-negative Matrix Factorization (NMF) [8] is a matrix factorization algorithm that focuses on the analysis of data matrices whose elements are nonnegative.

Given a data matrix $\mathbf{X} = [\mathbf{x}_1, \cdots, \mathbf{x}_n] \in \mathbb{R}^{m\times n}$, each column of $\mathbf{X}$ is a sample vector. NMF aims to find two non-negative matrices $\mathbf{U} = [u_{ij}] \in \mathbb{R}^{m\times k}$ and $\mathbf{V} = [v_{ij}] \in \mathbb{R}^{n\times k}$ which minimize the following objective function:

$$O = \|\mathbf{X} - \mathbf{U}\mathbf{V}^T\|_{\mathrm{F}}^2 \tag{1}$$

where $\|\cdot\|_{\mathrm{F}}$ denotes the matrix *Frobenius norm*[1].

Although the objective function O in Eqn. (1) is convex in $\mathbf{U}$ only or $\mathbf{V}$ only, it is not convex in both variables together. Therefore it is unrealistic to expect an algorithm to find the global minimum of O. Lee & Seung [9] presented an iterative update algorithm as follows:

$$u_{ij}^{t+1} = u_{ij}^t \frac{\left(\mathbf{X}\mathbf{V}\right)_{ij}}{\left(\mathbf{U}\mathbf{V}^T\mathbf{V}\right)_{ij}} \tag{2}$$

$$v_{ij}^{t+1} = v_{ij}^t \frac{\left(\mathbf{X}^T\mathbf{U}\right)_{ij}}{\left(\mathbf{V}\mathbf{U}^T\mathbf{U}\right)_{ij}} \tag{3}$$

It is proved that the above update steps will find a local mimimum of the objective function O [9].

In reality, we have $k \ll m$ and $k \ll n$. Thus, NMF essentially try to find a compressed approximation

[1] One can use other cost functions to measure how good $\mathbf{U}\mathbf{V}^T$ approximates $\mathbf{X}$[9]. In this paper, we will only focus on the Frobenius norm because of the space limitation.

of the original data matrix, $\mathbf{X} \approx \mathbf{U}\mathbf{V}^T$. We can view this approximation column by column as

$$\mathbf{x}_i \approx \sum_{j=1}^{k} \mathbf{u}_j v_{ij} \tag{4}$$

where $\mathbf{u}_j$ is the j-th column vector of **U**. Thus, each data vector $\mathbf{x}_i$ is approximated by a linear combination of the columns of **U**, weighted by the components of **V**. Therefore **U** can be regarded as containing a basis that is optimized for the linear approximation of the data in **X**. Since relatively few basis vectors are used to represent many data vectors, good approximation can only be achieved if the basis vectors discover structure that is latent in the data [9].

The non-negative constraints on **U** and **V** only allow addictive combinations among different basis. This is the most significant difference between NMF and other other matrix factorization methods, *e.g.*, SVD. Unlike SVD, no subtractions can occur in NMF. For this reason, it is believed that NMF can learn a *parts-based* representation [8]. The advantages of this parts-based representation has been observed in many real world problems such as face analysis [10], document clustering [18] and DNA gene expression analysis [4].

3 Graph Regularized Non-negative Matrix Factorization

By using the non-negative constraints, NMF can learn a parts-based representation. However, NMF performs this learning in the Euclidean space. It fails to to discover the intrinsic geometrical and discriminating structure of the data space, which is essential to the real applications. In this Section, we introduce our *Graph regularized Non-negative Matrix Factorization* (GNMF) algorithm which avoids this limitation by incorporating a geometrically based regularizer.

3.1 The Objective Function

Recall that NMF tries to find a basis that is optimized for the linear approximation of the data which are drawn according to the distribution P_X. One might hope that knowledge of the distribution P_X can be exploited for better discovery of this basis. A natural assumption here could be that if two data points $\mathbf{x}_i, \mathbf{x}_j$ are *close* in the *intrinsic* geometry of the data distribution, then the representations of this two points in the new basis are also close to each other. This assumption is usually referred to as *manifold assumption* [2], which plays an essential rule in developing various kinds of algorithms including dimensionality reduction algorithms [2] and semi-supervised learning algorithms [3, 21, 20].

Let $f_k(\mathbf{x}_i) = v_{ik}$ be function that produce the mapping of the original data point $\mathbf{x}_i$ onto the axis $\mathbf{u}_k$, we use $\|f_k\|_M^2$ to measure the smoothness of f_k along the geodesics in the intrinsic geometry of the data. When we consider the case that the data is a compact submanifold $\mathcal{M} \subset \mathbb{R}^m$, a natural choice for $\|f_k\|_M^2$ is

$$\|f_k\|_M^2 = \int_{\mathbf{x}\in\mathcal{M}} \|\nabla_{\mathcal{M}} f_k\|^2 dP_X(\mathbf{x}) \tag{5}$$

where $\nabla_{\mathcal{M}}$ is the gradient of f_k along the manifold $\mathcal{M}$ and the integral is taken over the distribution P_X.

In reality, the data manifold is usually unknown. Thus, $\|f_k\|_M^2$ in Eqn. (5) can not be computed. Recent studies on spectral graph theory [5] and manifold learning theory [1] have demonstrated that $\|f_k\|_M^2$ can be discretely approximated through a nearest neighbor graph on a scatter of data points.

Consider a graph with n vertices where each vertex corresponds to a data point. Define the edge weight matrix W as follows:

$$\mathbf{W}_{ij} = \begin{cases} 1, & \text{if } \mathbf{x}_i \in N_p(\mathbf{x}_j) \text{ or } \mathbf{x}_j \in N_p(\mathbf{x}_i) \\ 0, & \text{otherwise.} \end{cases} \tag{6}$$

where $N_p(\mathbf{x}_i)$ denotes the set of p nearest neighbors of $\mathbf{x}_i$. Define $\mathbf{L} = \mathbf{D} - \mathbf{W}$, where $\mathbf{D}$ is a diagonal matrix whose entries are column (or row, since $\mathbf{W}$ is symmetric) sums of $\mathbf{W}$, $\mathbf{D}_{ii} = \sum_j \mathbf{W}_{ij}$. $\mathbf{L}$ is called graph Laplacian [5], which is a discrete approximation to the Laplace-Beltrami operator $\triangle_{\mathcal{M}}$ on the manifold [1]. Thus, the discrete approximation of $\|f_k\|_M^2$ can be computed as follows:

$$\begin{aligned} \mathcal{R}_k &= \frac{1}{2}\sum_{i,j=1}^{N} \left(f_k(\mathbf{x}_i) - f_k(\mathbf{x}_j)\right)^2 \mathbf{W}_{ij} \\ &= \sum_{i=1}^{N} f_k(\mathbf{x}_i)^2 \mathbf{D}_{ii} - \sum_{i,j=1}^{N} f_k(\mathbf{x}_i) f_k(\mathbf{x}_j)\mathbf{W}_{ij} \\ &= \sum_{i=1}^{N} v_{ik}^2 \mathbf{D}_{ii} - \sum_{i,j=1}^{N} v_{ik} v_{jk} \mathbf{W}_{ij} \\ &= \mathbf{v}_k^T \mathbf{D} \mathbf{v}_k - \mathbf{v}_k^T \mathbf{W} \mathbf{v}_k \\ &= \mathbf{v}_k^T \mathbf{L} \mathbf{v}_k \end{aligned} \tag{7}$$

$\mathcal{R}_k$ can be used to measure the smoothness of mapping function f_k along the geodesics in the intrinsic geometry of the data set. By minimizing $\mathcal{R}_k$, we get a mapping function f_k which is sufficiently smooth on the data manifold. A intuitive explanation of minimizing $\mathcal{R}_k$ is that if two data points $\mathbf{x}_i$ and $\mathbf{x}_j$ are close (*i.e.* W_{ij} is big), $f_k(\mathbf{x}_i)$ and $f_k(\mathbf{x}_j)$ are similar to each other.

Our GNMF incorporates the $\mathcal{R}_k$ term and minimize the objective function

$$\begin{aligned}\mathcal{O} &= \|\mathbf{X} - \mathbf{U}\mathbf{V}^T\|_{\mathrm{F}}^2 + \lambda \sum_{i=1}^{k} \mathcal{R}_k \\ &= \|\mathbf{X} - \mathbf{U}\mathbf{V}^T\|_{\mathrm{F}}^2 + \lambda \operatorname{Tr}(\mathbf{V}^T\mathbf{L}\mathbf{V})\end{aligned} \tag{8}$$

with the constraint that u_{ij} and v_{ij} are non-negative. $\operatorname{Tr}(\cdot)$ denotes the trace of a matrix. The $\lambda \geq 0$ is the regularization parameter.

3.2 An Algorithm

The objective function $\mathcal{O}$ of GNMF in Eqn. (8) is not convex in both **U** and **V** together. Therefore it is unrealistic to expect an algorithm to find the global minimum of $\mathcal{O}$. In the following, we introduce an iterative algorithm which can achieve a local minimum.

The objective function $\mathcal{O}$ can be rewritten as:

$$\begin{aligned}\mathcal{O} &= \operatorname{Tr}\left((\mathbf{X} - \mathbf{U}\mathbf{V}^T)(\mathbf{X} - \mathbf{U}\mathbf{V}^T)^T\right) + \lambda \operatorname{Tr}(\mathbf{V}^T\mathbf{L}\mathbf{V}) \\ &= \operatorname{Tr}\left(\mathbf{X}\mathbf{X}^T\right) - 2\operatorname{Tr}\left(\mathbf{X}\mathbf{V}\mathbf{U}^T\right) + \operatorname{Tr}\left(\mathbf{U}\mathbf{V}^T\mathbf{V}\mathbf{U}^T\right) \\ &\quad + \lambda \operatorname{Tr}(\mathbf{V}^T\mathbf{L}\mathbf{V})\end{aligned} \tag{9}$$

where the second step of derivation uses the matrix property $\operatorname{Tr}(\mathbf{AB}) = \operatorname{Tr}(\mathbf{BA})$ and $\operatorname{Tr}(\mathbf{A}) = \operatorname{Tr}(\mathbf{A}^T)$. Let ψ_{ij} and ϕ_{ij} be the Lagrange multiplier for constraint $u_{ij} \geq 0$ and $v_{ij} \geq 0$ respectively, and $\Psi = [\psi_{ij}]$, $\Phi = [\phi_{ij}]$, the Lagrange $\mathcal{L}$ is

$$\begin{aligned}\mathcal{L} = &\operatorname{Tr}\left(\mathbf{X}\mathbf{X}^T\right) - 2\operatorname{Tr}\left(\mathbf{X}\mathbf{V}\mathbf{U}^T\right) + \operatorname{Tr}\left(\mathbf{U}\mathbf{V}^T\mathbf{V}\mathbf{U}^T\right) \\ &+ \lambda \operatorname{Tr}(\mathbf{V}^T\mathbf{L}\mathbf{V}) + \operatorname{Tr}(\Psi\mathbf{U}^T) + \operatorname{Tr}(\Phi\mathbf{V}^T)\end{aligned} \tag{10}$$

The partial derivatives of $\mathcal{L}$ with respect to **U** and **V** are:

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial \mathbf{U}} &= -2\mathbf{X}\mathbf{V} + 2\mathbf{U}\mathbf{V}^T\mathbf{V} + \Psi && (11) \\ \frac{\partial \mathcal{L}}{\partial \mathbf{V}} &= -2\mathbf{X}^T\mathbf{U} + 2\mathbf{V}\mathbf{U}^T\mathbf{U} + 2\lambda\mathbf{L}\mathbf{V} + \Phi && (12) \end{aligned}$$

Using the KKT conditions $\psi_{ij}u_{ij} = 0$ and $\phi_{ij}v_{ij} = 0$, we get the following equations for u_{ij} and v_{ij}:

$$-\big(\mathbf{X}\mathbf{V}\big)_{ij}u_{ij} + \big(\mathbf{U}\mathbf{V}^T\mathbf{V}\big)_{ij}u_{ij} = 0 \tag{13}$$

$$-\big(\mathbf{X}^T\mathbf{U}\big)_{ij}v_{ij} + \big(\mathbf{V}\mathbf{U}^T\mathbf{U}\big)_{ij}v_{ij} + \lambda\big(\mathbf{L}\mathbf{V}\big)_{ij}v_{ij} = 0 \tag{14}$$

These equations lead to the following update rules:

$$u_{ij} \leftarrow u_{ij}\frac{\big(\mathbf{X}\mathbf{V}\big)_{ij}}{\big(\mathbf{U}\mathbf{V}^T\mathbf{V}\big)_{ij}} \tag{15}$$

$$v_{ij} \leftarrow v_{ij}\frac{\big(\mathbf{X}^T\mathbf{U} + \lambda\mathbf{W}\mathbf{V}\big)_{ij}}{\big(\mathbf{V}\mathbf{U}^T\mathbf{U} + \lambda\mathbf{D}\mathbf{V}\big)_{ij}} \tag{16}$$

Regarding these two update rules, we have the following theorem:

Theorem 1 *The objective function $\mathcal{O}$ in Eqn. (8) is nonincreasing under the update rules in Eqn. (15) and (16). The objective function is invariant under these updates if and only if* ***U*** *and* ***V*** *are at a stationary point.*

Theorem 1 grantees that the update rules of **U** and **V** in Eqn. (15) and (16) converge and the final solution will be a local optimum. Please see the Appendix for a detailed proof.

4 Experimental Results

Previous studies show that NMF is very powerful on document clustering [18, 14]. It can achieve similar or better performance than most of the state-of-the-art clustering algorithms, including the popular spectral clustering methods [18]. Assume that a document corpus is comprised of k clusters each of which corresponds to a coherent topic. To accurately cluster the given document corpus, it is ideal to project the documents into a k-dimensional semantic space in which each axis corresponds to a particular topic. In this

Table 1: Statistics of TDT2 and Reuters corpora.

	TDT2	Reuters
No. docs. used	9394	8067
No. clusters used	30	30
Max. cluster size	1844	3713
Min. cluster size	52	18
Med. cluster size	131	45
Avg. cluster size	313	269

semantic space, each document can be represented as a linear combination of the k topics. Because it is more natural to consider each document as an additive rather subtractive mixture of the underlying topics, the combination coefficients should all take non-negative values. These values can be used to decide the cluster membership. This is the main motivation of applying NMF on document clustering. In this section, we also evaluate our GNMF algorithm on document clustering problem.

There are two parameters in our GNMF approach: the number of nearest neighbors p and the regularization parameter λ. Throughout our experiments, we empirically set the number of nearest neighbors p to 5, the value of the regularization parameter λ to 10.

4.1 Data Corpora

We conducted the performance evaluations using the TDT2 [2] and the Reuters[3] document corpora. These two document corpora have been among the ideal test sets for document clustering purposes because documents in the corpora have been manually clustered based on their topics and each document has been assigned one or more labels indicating which topic/topics it belongs to.

The TDT2 corpus consists of data collected during the first half of 1998 and taken from 6 sources, including 2 newswires (APW, NYT), 2 radio programs (VOA, PRI) and 2 television programs (CNN, ABC). It consists of 11201 on-topic documents which are classified into 96 semantic categories. In this experiment, those documents appearing in two or more categories were removed, and only the largest 30 categories were kept, thus leaving us with 9,394 documents in total.

The Reuters corpus contains 21578 documents which are grouped into 135 clusters. Compared with

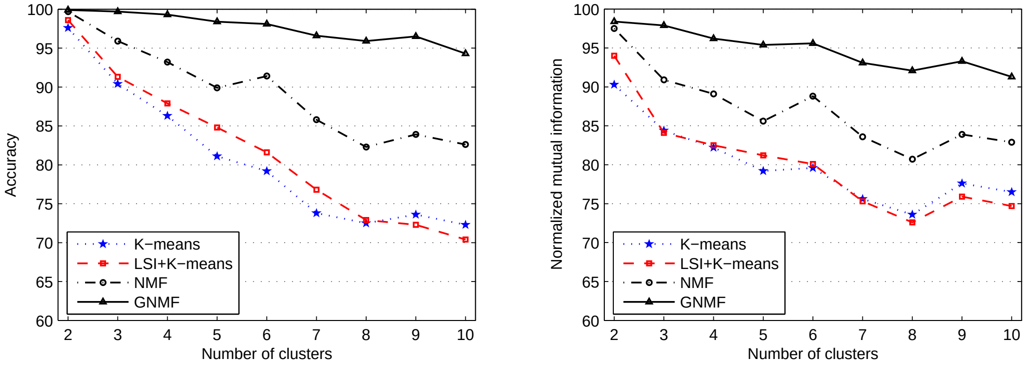
[2] Nist Topic Detection and Tracking corpus at
http://www.nist.gov/speech/tests/tdt/tdt98/index.htm

[3] Reuters-21578 corpus is at
http://www.daviddlewis.com/resources/testcollections/reuters21578/

Table 2: Clustering performance on TDT2

k	Accuracy (%)				Normalized Mutual Information (%)			
	K-means	LSI+K-means	NMF	GNMF	K-means	LSI+K-means	NMF	GNMF
2	97.6±6.7	98.6±5.1	99.7±1.1	99.9±0.3	90.3±24.8	94.0±19.7	97.5±9.7	98.4±2.8
3	90.4±17.3	91.3±15.7	95.9±10.8	99.7±0.4	84.4±25.7	84.1±27.0	90.9±18.7	97.9±2.4
4	86.3±16.4	87.9±17.1	93.2±11.9	99.3±1.6	82.2±20.9	82.5±23.1	89.1±16.6	96.2±7.7
5	81.1±16.9	84.8±15.4	89.9±12.7	98.4±4.7	79.2±17.9	81.2±17.7	85.6±15.8	95.4±7.4
6	79.2±16.1	81.6±15.3	91.4±11.7	98.1±5.0	79.6±15.5	80.1±15.8	88.8±12.5	95.6±6.4
7	73.8±15.3	76.8±15.6	85.8±13.2	96.6±4.7	75.6±16.3	75.3±16.5	83.6±14.0	93.1±6.5
8	72.5±16.3	72.9±14.7	82.3±13.0	95.9±5.3	73.6±16.0	72.6±16.8	80.7±13.4	92.1±6.3
9	73.6±14.6	72.3±14.3	83.9±13.1	96.5±4.9	77.6±12.5	75.9±13.0	83.9±11.4	93.3±6.0
10	72.3±14.5	70.4±12.9	82.6±10.2	94.3±5.7	76.5±13.1	74.7±13.7	82.9±9.9	91.3±6.5
Avg	80.8	81.8	89.4	97.6	79.9	80.0	87.0	94.8

k is the number of clusters



(a)

(b)

Figure 1: (a) Accuracy (b) Normalized mutual information vs. the number of classes on TDT2 corpus

TDT2 corpus, the Reuters corpus is more difficult for clustering. In TDT2, the content of each cluster is narrowly defined, whereas in Reuters, documents in each cluster have a broader variety of content. Moreover, the Reuters corpus is much more unbalanced, with some large clusters more than 200 times larger than some small ones. In our test, we discarded documents with multiple category labels, and only selected the largest 30 categories. This left us with 8067 documents in total. Table 1 provides the statistics of the two document corpora.

In both of the two corpora, the stop words are removed and each document is represented as a *tf-idf* vector.

4.2 Evaluation Metric

The clustering result is evaluated by comparing the obtained label of each document with that provided by the document corpus. Two metrics, the accuracy (AC) and the normalized mutual information metric ($\overline{MI}$) are used to measure the clustering performance [18]. Given a document $\mathbf{x}_i$, let r_i and s_i be the obtained cluster label and the label provided by the corpus, respectively. The AC is defined as follows:

$$AC = \frac{\sum_{i=1}^{n} \delta(s_i, map(r_i))}{n}$$

where n is the total number of documents and $\delta(x, y)$ is the delta function that equals one if $x = y$ and equals zero otherwise, and map(r_i) is the permutation mapping function that maps each cluster label r_i to the equivalent label from the data corpus. The best mapping can be found by using the Kuhn-Munkres algorithm [12].

Let C denote the set of clusters obtained from the ground truth and C' obtained from our algorithm. Their mutual information metric $MI(C, C')$ is defined as follows:

$$MI(C, C') = \sum_{c_i \in C, c'_j \in C'} p(c_i, c'_j) \cdot log_2 \frac{p(c_i, c'_j)}{p(c_i) \cdot p(c'_j)}$$

where $p(c_i)$ and $p(c'_j)$ are the probabilities that a document arbitrarily selected from the corpus belongs to the clusters c_i and c'_j, respectively, and $p(c_i, c'_j)$ is the joint probability that the arbitrarily selected document belongs to the clusters c_i as well as c'_j at the same time. In our experiments, we use the normalized mutual information $\overline{MI}$ as follows:

$$\overline{MI}(C, C') = \frac{MI(C, C')}{\max(H(C), H(C'))}$$

where $H(C)$ and $H(C')$ are the entropies of C and C', respectively. It is easy to check that $\overline{MI}(C, C')$ ranges from 0 to 1. $\overline{MI} = 1$ if the two sets of clusters are identical, and $\overline{MI} = 0$ if the two sets are independent.

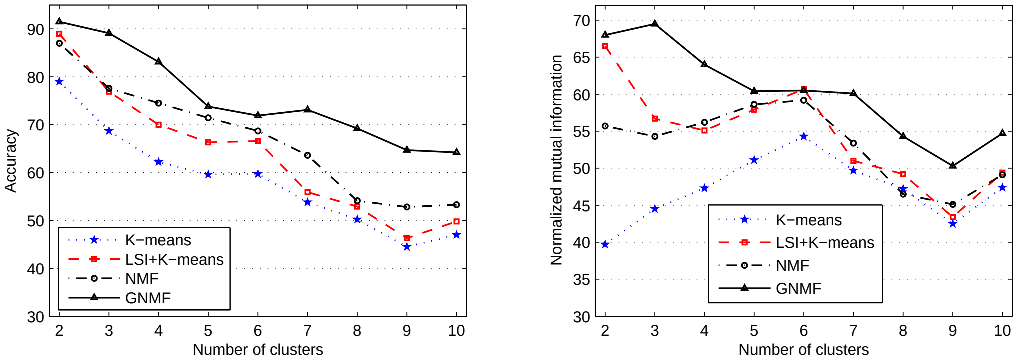
4.3 Performance Evaluations and Comparisons

To demonstrate how the document clustering performance can be improved by our method, we compared GNMF with other three popular document clustering algorithms as follows:

Table 3: Clustering performance on Reuters

k	Accuracy (%)				Normalized Mutual Information (%)			
	K-means	LSI+K-means	NMF	GNMF	K-means	LSI+K-means	NMF	GNMF
2	79.0±17.4	89.0±15.9	87.0±15.3	91.5±13.3	39.7±37.3	66.5±38.0	55.7±39.6	68.0±34.6
3	68.7±14.7	76.9±19.7	77.6±15.9	89.1±11.6	44.5±26.6	56.7±33.8	54.3±31.5	69.5±24.8
4	62.2±13.7	70.0±20.2	74.5±17.3	83.1±13.9	47.3±21.8	55.1±28.1	56.2±25.0	64.0±19.9
5	59.6±16.5	66.3±21.4	71.4±15.0	73.8±12.9	51.1±22.3	57.9±26.4	58.6±22.7	60.4±19.9
6	59.7±16.6	66.6±20.0	68.7±16.1	71.9±10.9	54.3±23.0	60.7±26.2	59.2±22.4	60.5±18.8
7	53.8±16.5	55.9±19.5	63.6±14.0	73.1±10.9	49.7±20.3	51.0±22.9	53.4±18.8	60.1±14.3
8	50.2±17.3	52.9±22.0	54.1±16.4	69.2±10.1	47.2±21.3	49.2±25.6	46.5±20.4	54.3±16.7
9	44.5±16.7	46.3±19.2	52.8±15.2	64.7±10.7	42.5±20.9	43.4±23.4	45.1±19.3	50.3±15.0
10	47.0±17.6	49.8±20.1	53.3±13.9	64.2±11.4	47.4±19.9	49.4±22.7	49.1±18.3	54.7±14.3
Avg	58.3	63.7	67.0	75.6	47.1	54.4	53.1	60.2

k is the number of clusters



(a)

(b)

Figure 2: (a) Accuracy (b) Normalized mutual information vs. the number of clusters on Reuters corpus

- Canonical K-means clustering method (K-means in short).

- K-means clustering in the Latent Semantic Indexing subspace (LSI+K-means in short). LSI [6] is the most well known dimensionality reduction algorithm in document analysis. It is essentially based on SVD and try to project the document into a *latent semantic subspace*. The document cluster structure is expected to be more explicit in this semantic subspace. Interestingly, Zha *et al*. [19] has shown that K-means clustering in the LSI subspace has close connection with Average Association [15], which is a popular spectral clustering algorithm. They showed that if the inner product is used to measure the document similarity and construct the graph, K-means after LSI is equivalent to average association.

- Nonnegative Matrix Factorization based clustering (NMF in short). We implemented a normalized cut weighted version of NMF as suggested in [18].

Table 2 and 3 show the evaluation results using the TDT2 and the Reuters corpus, respectively. The evaluations were conducted with the cluster numbers ranging from two to ten. For each given cluster number K, 50 test runs were conducted on different randomly chosen clusters. Both the average and variance of the performance are reported in the tables.

These experiments reveal a number of interesting points:

- The non-negative matrix factorization based methods, both NMF and GNMF, outperform the other two methods, which suggests the superiority of NMF in discovering the hidden topic structure than other matrix factorization methods, *e.g.*, SVD.

- Our GNMF approach gets significantly better performance than the ordinary NMF. This shows that by considering the intrinsic geometrical structure of the data, GNMF can learn a better compact representation in the sense of semantic structure.

- The improvement of GNMF over other methods is more significant on the TDT2 corpus than the Reuters corpus. One possible reason is that the document clusters in TDT2 are generally more compact and focused than the clusters in Reuters. Thus, the nearest neighbor graph constructed over TDT2 can better capture the geometrical structure of the document space.

5 Conclusions and Future Work

We have presented a novel method for matrix factorization, called Graph regularized Non-negative Matrix Factorization (GNMF). GNMF models the data space as a submanifold embedded in the ambient space and performs the non-negative matrix factorization on this manifold in question. As a result, GNMF can have more discriminating power than the ordinary NMF approach which only considers the Euclidean structure of the data. Experimental results on document clustering show that GNMF provides better representation in the sense of semantic structure.

Several questions remain to be investigated in our future work:

1. There is a parameter λ which controls the smoothness of our GNMF model. GNMF boils down to

original NMF when $\lambda = 0$. Thus, a suitable value of λ is critical to our algorithm. It remains unclear how to do model selection theoretically and efficiently.

2. It would be very interesting to explore different ways of constructing the document graph to model the semantic structure in the data. There is no reason to believe that the nearest neighbor graph is the only or the most natural choice. For example, for web page data it may be more natural to use the hyperlink information to construct the graph.

References

[1] M. Belkin. *Problems of Learning on Manifolds*. PhD thesis, University of Chicago, 2003.

[2] M. Belkin and P. Niyogi. Laplacian eigenmaps and spectral techniques for embedding and clustering. In *Advances in Neural Information Processing Systems 14*, pages 585–591. MIT Press, Cambridge, MA, 2001.

[3] M. Belkin, P. Niyogi, and V. Sindhwani. Manifold regularization: A geometric framework for learning from examples. *Journal of Machine Learning Research*, 2006.

[4] J.-P. Brunet, P. Tamayo, T. R. Golub, and J. P. Mesirov. Metagenes and molecular pattern discovery using matrix factorization. *Proceedings of the National Academy of Sciences*, 101(12):4164–4169, 2004.

[5] F. R. K. Chung. *Spectral Graph Theory*, volume 92 of *Regional Conference Series in Mathematics*. AMS, 1997.

[6] S. C. Deerwester, S. T. Dumais, T. K. Landauer, G. W. Furnas, and R. A. harshman. Indexing by latent semantic analysis. *Journal of the American Society of Information Science*, 41(6):391–407, 1990.

[7] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977.

[8] D. D. Lee and H. S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401:788–791, 1999.

[9] D. D. Lee and H. S. Seung. Algorithms for non-negative matrix factorization. In *Advances in Neural Information Processing Systems 13*. 2001.

[10] S. Z. Li, X. Hou, H. Zhang, and Q. Cheng. Learning spatially localized, parts-based representation. In *2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'01)*, pages 207–212, 2001.

[11] N. K. Logothetis and D. L. Sheinberg. Visual object recognition. *Annual Review of Neuroscience*, 19:577–621, 1996.

[12] L. Lovasz and M. Plummer. *Matching Theory*. Akad*é*miai Kiad*ó*, North Holland, Budapest, 1986.

[13] S. E. Palmer. Hierarchical structure in perceptual representation. *Cognitive Psychology*, 9:441–474, 1977.

[14] F. Shahnaza, M. W. Berrya, V. Paucab, and R. J. Plemmonsb. Document clustering using nonnegative matrix factorization. *Information Processing & Management*, 42(2):373–386, 2006.

[15] J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905, 2000.

[16] M. Turk and A. Pentland. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3(1):71–86, 1991.

[17] E. Wachsmuth, M. W. Oram, and D. I. Perrett. Recognition of objects and their component parts: Responses of single units in the temporal cortex of the macaque. *Cerebral Cortex*, 4:509–522, 1994.

[18] W. Xu, X. Liu, and Y. Gong. Document clustering based on non-negative matrix factorization. In *Proc. 2003 Int. Conf. on Research and Development in Information Retrieval (SIGIR'03)*, pages 267–273, Toronto, Canada, Aug. 2003.

[19] H. Zha, C. Ding, M. Gu, X. He, , and H. Simon. Spectral relaxation for k-means clustering. In *Advances in Neural Information Processing Systems 14*, pages 1057–1064. MIT Press, Cambridge, MA, 2001.

[20] D. Zhou, O. Bousquet, T. Lal, J. Weston, and B. Sch*ö*lkopf. Learning with local and global consistency. In *Advances in Neural Information Processing Systems 16*, 2003.

[21] X. Zhu and J. Lafferty. Harmonic mixtures: combining mixture models and graph-based methods for inductive and scalable semi-supervised learning. In *ICML '05: Proceedings of the 22nd international conference on Machine learning*, pages 1052–1059, 2005.

Appendix (Proofs of Theorem 1):

The objective function $\mathcal{O}$ of GNMF in Eqn. (8) is certainly bounded from below by zero. To prove Theorem 1, we need to show that $\mathcal{O}$ is nonincreasing under the update steps in Eqn. (15) and (16). Since the second term of $\mathcal{O}$ is only related to **V**, we have exactly the same update formula for **U** in GNMF as the original NMF. Thus, we can use the convergence proof of NMF to show that $\mathcal{O}$ is nonincreasing under the update step in Eqn. (15). Please see [9] for details.

Now we only need to prove that $\mathcal{O}$ is nonincreasing under the update step in Eqn. (16). we will follow the similar procedure described in [9]. Our proof will make use of an auxiliary function similar to that used in the Expectation-Maximization algorithm [7]. We begin with the definition of the *auxiliary function*.

Definition $G(v, v')$ is an *auxiliary function* for $F(v)$ if the conditions

$$G(v, v') \geq F(v), \quad G(v, v) = F(v)$$

are satisfied.

The auxiliary function is very useful because of the following lemma.

Lemma 2 *If G is an auxiliary function of F, then F is nonincreasing under the update*

$$v^{(t+1)} = \arg\min_{v} G(v, v^{(t)}) \tag{17}$$

Proof

$$F(v^{(t+1)}) \leq G(v^{(t+1)}, v^{(t)}) \leq G(v^{(t)}, v^{(t)}) = F(v^{(t)})$$

∎

Now we will show that the update step for **V** in Eqn. (16) is exactly the update in Eqn. (17) with a proper auxiliary function.

We rewrote the objective function $\mathcal{O}$ of GNMF in Eqn. (8) as follows

$$\begin{aligned}\mathcal{O} &= \|\mathbf{X} - \mathbf{U}\mathbf{V}^T\|_{\mathrm{F}}^2 + \lambda\,\mathrm{Tr}(\mathbf{V}^T\mathbf{L}\mathbf{V}) \\ &= \sum_{i=1}^{m}\sum_{j=1}^{n}(x_{ij} - \sum_{l=1}^{k} u_{il}v_{jl})^2 + \lambda \sum_{l=1}^{k}\sum_{i=1}^{n}\sum_{j=1}^{n} v_{jl}L_{ji}v_{il}\end{aligned} \tag{18}$$

Considering any element v_{ab} in $\mathbf{V}$, we use F_{ab} to denote the part of $\mathcal{O}$ which is only relevant to v_{ab}. It is easy to check that

$$F'_{ab} = \left(\frac{\partial \mathcal{O}}{\partial \mathbf{V}}\right)_{ab} = \big(-2\mathbf{X}^T\mathbf{U} + 2\mathbf{V}\mathbf{U}^T\mathbf{U} + 2\lambda\mathbf{L}\mathbf{V}\big)_{ab} \tag{19}$$

$$F''_{ab} = 2\big(\mathbf{U}^T\mathbf{U}\big)_{bb} + 2\lambda\mathbf{L}_{aa} \tag{20}$$

Since our update is essentially element-wise, it is sufficient to show that each F_{ab} is nonincreasing under the update step of Eqn. (16).

Lemma 3 *Function*

$$\begin{aligned}G(v, v_{ab}^{(t)}) =& F_{ab}(v_{ab}^{(t)}) + F'_{ab}(v_{ab}^{(t)})(v - v_{ab}^{(t)}) \\ &+ \frac{\big(\boldsymbol{V}\boldsymbol{U}^T\boldsymbol{U}\big)_{ab} + \lambda\big(\boldsymbol{D}\boldsymbol{V}\big)_{ab}}{v_{ab}^{(t)}}(v - v_{ab}^{(t)})^2\end{aligned} \tag{21}$$

is an auxiliary function for F_{ab}, the part of $\mathcal{O}$ which is only relevant to v_{ab}.

Proof Since $G(v, v) = F_{ab}(v)$ is obvious, we need only show that $G(v, v_{ab}^{(t)}) \geq F_{ab}(v)$. To do this, we compare the Taylor series expansion of $F_{ab}(v)$

$$\begin{aligned}F_{ab}(v) =& F_{ab}(v_{ab}^{(t)}) + F'_{ab}(v_{ab}^{(t)})(v - v_{ab}^{(t)}) \\ &+ \big[\big(\mathbf{U}^T\mathbf{U}\big)_{bb} + \lambda\mathbf{L}_{aa}\big](v - v_{ab}^{(t)})^2\end{aligned} \tag{22}$$

with Eqn. (21) to find that $G(v, v_{ab}^{(t)}) \geq F_{ab}(v)$ is equivalent to

$$\frac{\big(\mathbf{V}\mathbf{U}^T\mathbf{U}\big)_{ab} + \lambda\big(\mathbf{D}\mathbf{V}\big)_{ab}}{v_{ab}^{(t)}} \geq \big(\mathbf{U}^T\mathbf{U}\big)_{bb} + \lambda\mathbf{L}_{aa}. \tag{23}$$

We have

$$\big(\mathbf{V}\mathbf{U}^T\mathbf{U}\big)_{ab} = \sum_{l=1}^{k} v_{al}^{(t)}\big(\mathbf{U}^T\mathbf{U}\big)_{lb} \geq v_{ab}^{(t)}\big(\mathbf{U}^T\mathbf{U}\big)_{bb} \tag{24}$$

and

$$\begin{aligned}\lambda\big(\mathbf{DV}\big)_{ab} &= \lambda\sum_{j=1}^{m}\mathbf{D}_{aj}v_{jb}^{(t)} \geq \lambda\mathbf{D}_{aa}v_{ab}^{(t)} \\ &\geq \lambda\big(\mathbf{D}-\mathbf{W}\big)_{aa}v_{ab}^{(t)} = \lambda\mathbf{L}_{aa}v_{ab}^{(t)} \end{aligned} \tag{25}$$

.

Thus, Eqn. (23) holds and $G(v, v_{ab}^{(t)}) \geq F_{ab}(v)$. ▮

We can now demonstrate the convergence of Theorem 1:

Proof of Theorem 1 Replacing $G(v, v_{ab}^{(t)})$ in Eqn. (17) by Eqn. (21) results in the update rule:

$$\begin{aligned} v_{ab}^{(t+1)} &= v_{ab}^{(t)} - v_{ab}^{(t)}\frac{F'_{ab}(v_{ab}^{(t)})}{2\big(\mathbf{V}\mathbf{U}^T\mathbf{U}\big)_{ab} + 2\lambda\big(\mathbf{DV}\big)_{ab}} \\ &= v_{ab}^{(t)}\frac{\big(\mathbf{X}^T\mathbf{U} + \lambda\mathbf{WV}\big)_{ab}}{\big(\mathbf{V}\mathbf{U}^T\mathbf{U} + \lambda\mathbf{DV}\big)_{ab}} \end{aligned} \tag{26}$$

Since Eqn. (21) is an auxiliary function, F_{ab} is nonincreasing under this update rule. ▮