Locality Preserving Nonnegative Matrix Factorization

Deng Cai[†] **Xiaofei He**[†] **Xuanhui Wang**[‡] **Hujun Bao**[†] **Jiawei Han**[‡]

[†]State Key Lab of CAD&CG, College of Computer Science, Zhejiang University, China
{dengcai, xiaofeihe, bao}@cad.zju.edu.cn

[‡]Department of Computer Science, University of Illinois at Urbana-Champaign
{xwang20, hanj}@cs.uiuc.edu

Abstract

Matrix factorization techniques have been frequently applied in information processing tasks. Among them, Non-negative Matrix Factorization (NMF) have received considerable attentions due to its psychological and physiological interpretation of naturally occurring data whose representation may be parts-based in human brain. On the other hand, from geometric perspective the data is usually sampled from a low dimensional manifold embedded in high dimensional ambient space. One hopes then to find a compact representation which uncovers the hidden topics and simultaneously respects the intrinsic geometric structure. In this paper, we propose a novel algorithm, called *Locality Preserving Non-negative Matrix Factorization* (LPNMF), for this purpose. For two data points, we use KL-divergence to evaluate their similarity on the hidden topics. The optimal maps are obtained such that the feature values on hidden topics are restricted to be non-negative and vary smoothly along the geodesics of the data manifold. Our empirical study shows the encouraging results of the proposed algorithm in comparisons to the state-of-the-art algorithms on two large high-dimensional databases.

1 Introduction

Data representation has been a fundamental problem in many areas of information processing. A good representation can significantly facilitates *learning from example* in terms of learnability and computational complexity [Duda *et al.*, 2000]. On the one hand, each data point may be associated with some hidden topics. For example, a face image can be thought of as a combination of nose, mouth, eyes, etc. On the other hand, from geometrical perspective, the data points may be sampled from a probability distribution supported on a low dimensional submanifold embedded in the high dimensional space. One hopes then to find a compact representation which respects both hidden topics as well as geometric structure.

In order to discover the hidden topics, matrix factorization techniques have been frequently applied [Deerwester *et al.*, 1990; Liu *et al.*, 2008]. For example, the canonical algorithm Latent Semantic Indexing (LSI, [Deerwester *et al.*, 1990]) applies Singular Value Decomposition (SVD) to decompose the original data matrix $\mathbf{X}$ into a product of three matrices, that is, $\mathbf{X} = \mathbf{U}\mathbf{S}\mathbf{V}^{\mathrm{T}}$. $\mathbf{U}$ and $\mathbf{V}$ are orthogonal matrices and $\mathbf{S}$ is a diagonal matrix. The quantities $\mathbf{S}_{ii}$ are called the *singular values* of $\mathbf{X}$, and the column vectors of $\mathbf{U}$ and $\mathbf{V}$ are called left and right *singular vectors*, respectively. By removing those singular vectors corresponding to sufficiently small singular values, we obtain a natural low-rank approximation to the original matrix and each dimension corresponds to a hidden topic. Besides SVD, the other popular matrix factorization techniques include LU-decomposition, QR-decomposition, and Cholesky decomposition.

Recently Non-negative Matrix Factorization (NMF, [Lee and Seung, 1999]) have been proposed and achieved great success due to its theoretical interpretation and practical performance. Previous studies have shown there is psychological and physiological evidence for parts-based representation in human brain [Logothetis and Sheinberg, 1996; Palmer, 1977; Wachsmuth *et al.*, 1994]. The non-negative constraints in NMF lead to a parts-based representation because it allows only additive, not subtractive, combinations. NMF has been shown to be superior to SVD in face recognition [Li *et al.*, 2001] and document clustering [Xu *et al.*, 2003]. The major disadvantage of NMF is that it fails to consider the intrinsic geometric structure in the data.

In this paper, we aim to discover the hidden topics and the intrinsic geometric structure simultaneously. We propose a novel algorithm called *Locality Preserving Non-negative Matrix Factorization* (LPNMF) for this purpose. For two data points, we use KL-divergence to evaluate their similarity on the hidden topics. A nearest neighbor graph is constructed to model the local manifold structure. If two points are sufficiently close on the manifold, then we expect that they have similar representations on the hidden topics. Thus, the optimal maps are obtained such that the feature values on hidden topics are restricted to be non-negative and vary smoothly along the geodesics of the data manifold. We also propose an efficient method to solve the optimization problem. It is important to note that this work is fundamentally based on our previous work GNMF [Cai *et al.*, 2008]. The major difference is that GNMF evaluates the relationship between two matrices using Frobenius norm. While in this work, we use the divergence which has better probabilistic interpretation.

2 A Brief Review of NMF

Non-negative Matrix Factorization (NMF) [Lee and Seung, 1999] is a matrix factorization algorithm that focuses on the analysis of data matrices whose elements are nonnegative.

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

$$O = \sum_{i,j} \left(x_{ij} \log \frac{x_{ij}}{y_{ij}} - x_{ij} + y_{ij} \right) \tag{1}$$

where $\mathbf{Y} = [y_{ij}] = \mathbf{U}\mathbf{V}^T$. The above objective function is lower bounded by zero, and vanishes if and only if $\mathbf{X} = \mathbf{Y}$. It is usually referred as "divergence" of $\mathbf{X}$ from $\mathbf{Y}$ instead of "distance" between $\mathbf{X}$ and $\mathbf{Y}$ because it is not symmetric in $\mathbf{X}$ and $\mathbf{Y}$. It reduces to the Kullback-Leibler divergence, or relative entropy, when $\sum_{ij} x_{ij} = \sum_{ij} y_{ij} = 1$, so that $\mathbf{X}$ and $\mathbf{Y}$ can be regarded as normalized probability distributions. [1]

Although the objective function O in Eq. (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 [Lee and Seung, 2001] presented an iterative update algorithm as follows:

$$\begin{aligned} u_{ik} &\leftarrow u_{ik} \frac{\sum_j \left(x_{ij} v_{jk} / \sum_k u_{ik} v_{jk} \right)}{\sum_j v_{jk}} \\ v_{jk} &\leftarrow v_{jk} \frac{\sum_i \left(x_{ij} u_{ik} / \sum_k u_{ik} v_{jk} \right)}{\sum_i u_{ik}} \end{aligned} \tag{2}$$

It is proved that the above update steps will find a local minimum of the objective function O [Lee and Seung, 2001].

In reality, we have $t \ll m$ and $t \ll n$. Thus, NMF essentially try to find a compressed approximation of the original data matrix, $\mathbf{X} \approx \mathbf{U}\mathbf{V}^T$. We can view this approximation column by column as

$$\mathbf{x}_j \approx \sum_{k=1}^{t} \mathbf{u}_k v_{jk} \tag{3}$$

where $\mathbf{u}_k$ is the k-th column vector of $\mathbf{U}$. Thus, each data vector $\mathbf{x}_j$ is approximated by a linear combination of the columns of $\mathbf{U}$, weighted by the components of $\mathbf{V}$. Therefore $\mathbf{U}$ can be regarded as containing a basis that is optimized for the linear approximation of the data in $\mathbf{X}$. Let $\mathbf{z}_j^T$ denote the j-th row of $\mathbf{V}$, $\mathbf{z}_j = [v_{j1}, \cdots, v_{jk}]^t$. $\mathbf{z}_j$ can be regarded as the new representation of each data point in the new basis $\mathbf{U}$. 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 [Lee and Seung, 2001].

The non-negative constraints on $\mathbf{U}$ and $\mathbf{V}$ 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 [Lee and Seung, 1999]. The advantages of this parts-based representation has been observed in many real world problems such as face analysis [Li *et al.*, 2001], document clustering [Xu *et al.*, 2003] and DNA gene expression analysis [Brunet *et al.*, 2004].

3 Locality Preserving Non-negative Matrix Factorization

Recall that NMF tries to find a basis that is optimized for the linear approximation of the data. One might hope that knowledge of the geometric structure of the data can be exploited for better discovery of this basis. A natural assumption here could be that if two data points $\mathbf{x}_j, \mathbf{x}_s$ are *close* in the *intrinsic* geometry of the data distribution, then $\mathbf{z}_j$ and $\mathbf{z}_s$, 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* [Belkin and Niyogi, 2001; He and Niyogi, 2003], which plays an essential rule in developing various kinds of algorithms including dimensionality reduction algorithms [Belkin and Niyogi, 2001] and semi-supervised learning algorithms [Belkin *et al.*, 2006].

Recent studies on spectral graph theory [Chung, 1997] and manifold learning theory [Belkin and Niyogi, 2001] have demonstrated that the local geometric structure can be effectively modeled through a nearest neighbor graph on a scatter of data points. Consider a graph with N vertices where each vertex corresponds to a document in the corpus. Define the edge weight matrix W as follows:

$$\mathbf{W}_{js} = \begin{cases} 1, & \text{if } \mathbf{x}_j \in N_p(\mathbf{x}_s) \text{ or } \mathbf{x}_s \in N_p(\mathbf{x}_j) \\ 0, & \text{otherwise.} \end{cases} \tag{4}$$

where $N_p(\mathbf{x}_s)$ denotes the set of p nearest neighbors of $\mathbf{x}_s$.

Again, we can use the divergence between the low dimensional representations in the new basis of two samples to measure the "distance":

$$D(\mathbf{z}_j || \mathbf{z}_s) = \sum_{k=1}^{t} \left(v_{jk} \log \frac{v_{jk}}{v_{sk}} - v_{jk} + v_{sk} \right), \tag{5}$$

since we have $\mathbf{z}_j = [v_{j1}, \cdots, v_{jk}]^t$. Thus, the following term can be used to measure the smoothness of the low dimensional representation varies smoothly along the geodesics in the intrinsic geometry of data.

$$\begin{aligned} \mathcal{R} &= \frac{1}{2} \sum_{j,s=1}^{n} \Big(D(\mathbf{z}_j || \mathbf{z}_s) + D(\mathbf{z}_s || \mathbf{z}_j) \Big) \mathbf{W}_{js} \\ &= \frac{1}{2} \sum_{j,s=1}^{n} \sum_{k=1}^{t} \left(v_{jk} \log \frac{v_{jk}}{v_{sk}} + v_{sk} \log \frac{v_{sk}}{v_{jk}} \right) \mathbf{W}_{js}. \end{aligned} \tag{6}$$

By minimizing $\mathcal{R}$, we get a conditional probability distribution which is sufficiently smooth on the data manifold. A intuitive explanation of minimizing $\mathcal{R}$ is that if two data points $\mathbf{x}_j$ and $\mathbf{x}_s$ are close (*i.e.* $\mathbf{W}_{js}$ is big), $\mathbf{z}_j$ and $\mathbf{z}_s$ are also close to each other. Thus, we can name the new algorithm as *Locality Preserving Non-negative Matrix Factorization* (LPNMF).

[1] One can use other cost functions (*e.g.*, Frobenius norm) to measure how good $\mathbf{U}\mathbf{V}^T$ approximates $\mathbf{X}$. Please refer [Lee and Seung, 2001; Cai *et al.*, 2008] for more details.

Given a data matrix $\mathbf{X} = [x_{ij}] \in \mathbb{R}^{m\times n}$, Our LPNMF aims to find two non-negative matrices $\mathbf{U} = [u_{ik}] \in \mathbb{R}^{m\times t}$ and $\mathbf{V} = [v_{jk}] \in \mathbb{R}^{n\times t}$ which minimize the following objective function:

$$\mathcal{O} = \sum_{i=1}^{m}\sum_{j=1}^{n}\left(x_{ij}\log\frac{x_{ij}}{y_{ij}} - x_{ij} + y_{ij}\right) + \lambda\mathcal{R} \qquad (7)$$

where $\mathbf{Y} = [y_{ij}] = \mathbf{U}\mathbf{V}^T$. The $\lambda \geq 0$ is the regularization parameter.

3.1 Multiplicative Update Rules

The objective function $\mathcal{O}$ of LPNMF in Eq. (7) is not convex in both **U** and **V** together. Therefore it is unrealistic to expect an algorithm to find the global minimum of $\mathcal{O}$. Similar to the NMF, we also have two update rules which can achieve a local minimum.

$$u_{ik} \leftarrow u_{ik}\frac{\sum_j \left(x_{ij}v_{jk}/\sum_k u_{ik}v_{jk}\right)}{\sum_j v_{jk}} \qquad (8)$$

$$\mathbf{v}_k \leftarrow \left(\sum_i u_{ik}\mathbf{I} + \lambda\mathbf{L}\right)^{-1}\begin{bmatrix} v_{1k}\sum_i\left(x_{i1}u_{ik}/\sum_k u_{ik}v_{1k}\right) \\ v_{2k}\sum_i\left(x_{i2}u_{ik}/\sum_k u_{ik}v_{2k}\right) \\ \vdots \\ v_{nk}\sum_i\left(x_{in}u_{ik}/\sum_k u_{ik}v_{nk}\right)\end{bmatrix} \qquad (9)$$

where $\mathbf{v}_k$ is the k-th column of **V** and **I** is a $n \times n$ identity matrix. The matrix **L** is the graph Laplacian [Chung, 1997] of **W**. It is defined as $\mathbf{D}-\mathbf{W}$, where **D** is a diagonal matrix whose entries are column (or row, since **W** is symmetric) sums of **W**, $\mathbf{D}_{jj} = \sum_s \mathbf{W}_{sj}$.

When $\lambda = 0$, it is easy to check that the update rules in Eq. (8) and (9) reduce to the update rules of original NMF. When $\lambda > 0$, we have the following theorem:

Theorem 1. *The objective function $\mathcal{O}$ in Eq. (7) is nonincreasing under the update rules in Eq. (8) and (9). 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 Eq. (8) and (9) 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 [Xu *et al.*, 2003]. It can achieve similar or better performance than most of the state-of-the-art clustering algorithms, including the popular spectral clustering methods [Shi and Malik, 2000]. 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 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 LPNMF algorithm on document clustering problem.

The largest 30 categories in TDT2 [2] and Reuters[3] are used in our experiment. 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). The Reuters data set was originally collected and labeled by Carnegie Group, Inc. and Reuters, Ltd. In this experiment, those documents appearing in two or more categories were removed. Finally we have 9,394 documents in TDT2 corpus and 8,076 documents in Reuters corpus. In both of the two data sets, the stop words are removed and each document is represented as a *tf-idf* vector.

4.1 Evaluation Metric

The clustering result is evaluated by comparing the obtained label of each document with that provided by the document corpus. The normalized mutual information metric (NMI) is used to measure the performance [Xu *et al.*, 2003; Cai *et al.*, 2005].

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

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

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 NMI as follows:

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

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

4.2 Performance Evaluations and Comparisons

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

- Canonical kmeans clustering method (kmeans in short).
- Two representative spectral clustering methods: Average Association (AA in short) [Zha *et al.*, 2001], and

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

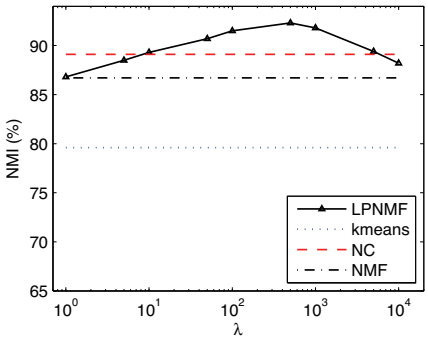
[3] Reuters-21578 corpus is at http://www.daviddlewis.com/resources/testcollections/reuters21578/

Table 1: Clustering performance on TDT2 (% NMI)

k	kmeans	AA	NC	NMF	LPNMF
2	82.7	74.7	97.7	95.7	**98.3**
3	83.4	82.2	**95.1**	89.0	**95.1**
4	80.7	76.7	89.6	86.2	**94.4**
5	78.1	75.9	92.1	83.6	**93.9**
6	81.6	79.5	**92.3**	91.9	92.1
7	79.2	79.7	87.7	84.8	**88.6**
8	78.1	74.2	83.1	81.7	**86.6**
9	79.8	75.6	87.7	86.6	**90.9**
10	73.1	69.2	76.7	81.2	**83.4**
Avg	79.6	76.4	89.1	86.7	**91.5**

Table 2: Clustering performance on Reuters (% NMI)

k	kmeans	AA	NC	NMF	LPNMF
2	44.9	39.2	56.3	60.2	**64.0**
3	37.5	39.2	51.2	50.7	**55.4**
4	50.4	43.9	57.8	61.1	**62.9**
5	46.9	42.1	50.6	51.3	**53.1**
6	55.1	45.1	59.1	59.3	**59.9**
7	54.0	44.9	54.0	56.3	**57.7**
8	41.8	32.4	39.7	40.4	**44.7**
9	43.9	36.2	45.1	45.9	**48.1**
10	54.8	44.7	51.5	55.4	**55.5**
Avg	47.7	40.9	51.7	53.4	**55.7**



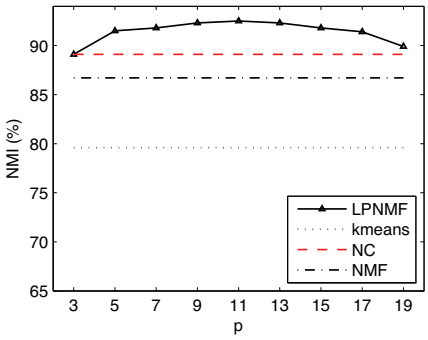


Figure 1: The performance of LPNMF vs. parameters λ and p. The LPNMF is very stable with respect to both the parameter λ and p. It achieves consistent good performance with the λ varying from 100 to 1000 and p varying from 5 to 17.

Normalized Cut (NC in short) [Shi and Malik, 2000]. Interestingly, Zha *et al*. [Zha *et al.*, 2001] has shown that the AA is equivalent to that of the SVD followed by the kmeans clustering method if the inner product is used to measure the document similarity.

- Nonnegative Matrix Factorization based clustering (NMF in short).

There are two parameters in our LPNMF 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 100.

Table 1 and 2 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, 20 test runs were conducted on different randomly chosen clusters and the average performance is reported in the tables. These experiments reveal a number of interesting points:

- The non-negative matrix factorization based methods, both NMF and LPNMF, outperform the AA method (SVD+kmeans), which suggests the superiority of NMF in discovering the hidden topic structure than other matrix factorization methods, *e.g.*, SVD.
- Our LPNMF approach gets significantly better performance than the ordinary NMF. This shows that by considering the intrinsic geometrical structure of the data, LPNMF can learn a better compact representation in the sense of semantic structure.
- The improvement of LPNMF 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.

4.3 Parameters Selection

Our LPNMF model has two essential parameters: the number of nearest neighbors p and the regularization parameter λ. Figure 1 show how the performance of LPNMF varies with the parameters λ and p (We only show the result on TDT2 corpus due to the space limitation. The curves are similar on Reuters corpus). As we can see, the LPNMF is very stable with respect to both the parameter λ and p. It achieves consistent good performance with the λ varying from 100 to 1000 and p varying from 5 to 17.

5 Conclusion

We have presented a novel method for matrix factorization, called Locality Preserving Non-negative Matrix Factorization (LPNMF). LPNMF 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, LPNMF 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 LPNMF provides better representation in the sense of semantic structure.

Acknowledgments

This work was supported in part by the National Key Basic Research Foundation of China under Grant 2009CB320801, the Program for Changjiang Scholars and Innovative Research Team in University (IRT0652,PCSIRT), the National Science Foundation of China under Grant 60875044, the U.S. National Science Foundation grants IIS-08-42769 and BDI-05-15813, and the Air Force Office of Scientific Research MURI award FA9550-08-1-0265. 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.

Appendix (Proofs of Theorem 1):

The objective function $\mathcal{O}$ of LPNMF in Eq. (7) 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 Eq. (8) and (9). Since the second term of $\mathcal{O}$ is only related to $\mathbf{V}$, we have exactly the same update formula for $\mathbf{U}$ in LPNMF 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 Eq. (8). Please see [Lee and Seung, 2001] for details.

Now we only need to prove that $\mathcal{O}$ is nonincreasing under the update step in Eq. (9). we will follow the similar procedure described in [Lee and Seung, 2001]. Our proof will make use of an auxiliary function similar to that used in the Expectation-Maximization algorithm [Dempster *et al.*, 1977]. We begin with the definition of the *auxiliary function*.

Definition $G(\mathbf{V}, \mathbf{V}')$ is an *auxiliary function* for $F(\mathbf{V})$ if the conditions

$$G(\mathbf{V}, \mathbf{V}') \geq F(\mathbf{V}), \quad G(\mathbf{V}, \mathbf{V}) = F(\mathbf{V})$$

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*

$$\boldsymbol{V}^{(q+1)} = \arg\min_{\boldsymbol{V}} G(\boldsymbol{V}, \boldsymbol{V}^{(q)}) \tag{10}$$

Proof.

$$F(\mathbf{V}^{(q+1)}) \leq G(\mathbf{V}^{(q+1)}, \mathbf{V}^{(q)}) \leq G(\mathbf{V}^{(q)}, \mathbf{V}^{(q)}) = F(\mathbf{V}^{(q)})$$

□

Now we will show that the update step for $\mathbf{V}$ in Eq. (9) is exactly the update in Eq. (10) with a proper auxiliary function.

We rewrote the objective function $\mathcal{O}$ of LPNMF in Eq. (7) as follows

$$\begin{aligned}\mathcal{O} = &\sum_{i,j}\left(x_{ij}\log\frac{x_{ij}}{\sum_k u_{ik}v_{jk}} - x_{ij} + \sum_k u_{ik}v_{jk}\right)\\ &+\frac{\lambda}{2}\sum_{j,s,k}\left(v_{jk}\log\frac{v_{jk}}{v_{sk}} + v_{sk}\log\frac{v_{sk}}{v_{jk}}\right)\mathbf{W}_{js}\end{aligned} \tag{11}$$

Lemma 3. *Function*

$$\begin{aligned}&G(\boldsymbol{V}, \boldsymbol{V}^{(q)})\\ =&\sum_{i,j}\left(x_{ij}\log x_{ij} - x_{ij} + \sum_k u_{ik}v_{jk}\right)\\ &-\sum_{i,j,k}\left(x_{ij}\frac{u_{ik}v_{jk}^{(q)}}{\sum_k u_{ik}v_{jk}^{(q)}}\left(\log u_{ik}v_{jk} - \log\frac{u_{ik}v_{jk}^{(q)}}{\sum_k u_{ik}v_{jk}^{(q)}}\right)\right)\\ &+\frac{\lambda}{2}\sum_{j,s,k}\left(v_{jk}\log\frac{v_{jk}}{v_{sk}} + v_{sk}\log\frac{v_{sk}}{v_{jk}}\right)\boldsymbol{W}_{js}\end{aligned}$$

is an auxiliary function for

$$\begin{aligned}F(\boldsymbol{V}) = &\sum_{i,j}\left(x_{ij}\log\frac{x_{ij}}{\sum_k u_{ik}v_{jk}} - x_{ij} + \sum_k u_{ik}v_{jk}\right)\\ &+\frac{\lambda}{2}\sum_{j,s,k}\left(v_{jk}\log\frac{v_{jk}}{v_{sk}} + v_{sk}\log\frac{v_{sk}}{v_{jk}}\right)\boldsymbol{W}_{js}\end{aligned}$$

Proof. It is straightforward to verify that $G(\mathbf{V}, \mathbf{V}) = F(\mathbf{V})$. To show that $G(\mathbf{V}, \mathbf{V}^{(q)}) \geq F(\mathbf{V})$, we use convexity of the log function to derive the inequality

$$-\log\left(\sum_k u_{ik}v_{jk}\right) \leq -\sum_k\left(\alpha_k\log\frac{u_{ik}v_{jk}}{\alpha_k}\right)$$

which holds for all nonnegative α_k that sum to unity. Setting

$$\alpha_k = \frac{u_{ik}v_{jk}^{(q)}}{\sum_k u_{ik}v_{jk}^{(q)}},$$

we obtain

$$\begin{aligned}&-\log\left(\sum_k u_{ik}v_{jk}\right) \leq\\ &-\sum_k\left(\frac{u_{ik}v_{jk}^{(q)}}{\sum_k u_{ik}v_{jk}^{(q)}}\left(\log u_{ik}v_{jk} - \log\frac{u_{ik}v_{jk}^{(q)}}{\sum_k u_{ik}v_{jk}^{(q)}}\right)\right).\end{aligned}$$

From this inequality it follows that $G(\mathbf{V}, \mathbf{V}^{(q)}) \geq F(\mathbf{V})$. □

Theorem 1 then follows from the application of Lemma 2:

Proof of Theorem 1. The minimum of $G(\mathbf{V}, \mathbf{V}^{(q)})$ with respect to $\mathbf{V}$ is determined by setting the gradient to zero:

$$\begin{aligned}\sum_i u_{ik} - &\sum_i x_{ij}\frac{u_{ik}v_{jk}^{(q)}}{\sum_k u_{ik}v_{jk}^{(q)}}\frac{1}{v_{jk}}\\ &+\frac{\lambda}{2}\sum_s\left(\log\frac{v_{jk}}{v_{sk}} + 1 - \frac{v_{sk}}{v_{jk}}\right)\mathbf{W}_{js} = 0,\\ &1 \leq j \leq n, \quad 1 \leq k \leq t\end{aligned} \tag{12}$$

Because of the log term, it is really hard to solve the above equations system. Let us recall the motivation of the regularization term. We hope that if two data points $\mathbf{x}_j$ and $\mathbf{x}_s$ are close (*i.e.* $\mathbf{W}_{js}$ is big), $\mathbf{z}_j$) will be close to $\mathbf{z}_s$ and v_{jk}/v_{sk} will approximately be 1. Thus, we can use the following approximation:

$$\log(x) \approx 1 - \frac{1}{x}, \quad x \to 1.$$

The above approximation is based on the first order expansion of Tylor series of log function. With this approximation, the equations in Eq. (12) can be written as

$$\begin{aligned} \sum_i u_{ik} - \sum_i x_{ij} \frac{u_{ik} v_{jk}^{(q)}}{\sum_k u_{ik} v_{jk}^{(q)}} \frac{1}{v_{jk}} \\ + \frac{\lambda}{v_{jk}} \sum_s (v_{jk} - v_{sk}) \mathbf{W}_{js} = 0, \\ 1 \le j \le n, \quad 1 \le k \le t \end{aligned} \tag{13}$$

Let $\mathbf{D}$ denote a diagonal matrix whose entries are column (or row, since $\mathbf{W}$ is symmetric) sums of $\mathbf{W}$, $\mathbf{D}_{jj} = \sum_s \mathbf{W}_{js}$. Define $\mathbf{L} = \mathbf{D} - \mathbf{W}$. Let $\mathbf{v}_k$ denote the k-th column of $\mathbf{V}$, $\mathbf{v}_k = [v_{1k}, \cdots, v_{nk}]^T$. It is easy to verify that $\sum_s (v_{jk} - v_{sk}) \mathbf{W}_{js}$ equals to the j-th element of vector $\mathbf{L}\mathbf{v}_k$.

The equations system in Eq. (13) can be rewritten as

$$\sum_i u_{ik} \mathbf{I} \mathbf{v}_k + \lambda \mathbf{L} \mathbf{v}_k = \begin{bmatrix} v_{1k}^{(q)} \sum_i \left(x_{i1} u_{ik} / \sum_k u_{ik} v_{1k}^{(q)} \right) \\ \vdots \\ v_{nk}^{(q)} \sum_i \left(x_{in} u_{ik} / \sum_k u_{ik} v_{nk}^{(q)} \right) \end{bmatrix}$$
$$1 \le k \le t.$$

Thus, the update rule of Eq. (10) takes the form

$$\mathbf{v}_k^{(q+1)} = \left(\sum_i u_{ik} \mathbf{I} + \lambda \mathbf{L} \right)^{-1} \begin{bmatrix} v_{1k}^{(q)} \sum_i \left(x_{i1} u_{ik} / \sum_k u_{ik} v_{1k}^{(q)} \right) \\ \vdots \\ v_{nk}^{(q)} \sum_i \left(x_{in} u_{ik} / \sum_k u_{ik} v_{nk}^{(q)} \right) \end{bmatrix}$$
$$1 \le k \le t.$$

Since G is an auxiliary function, F is nonincreasing under this update. □

References

[Belkin and Niyogi, 2001] M. Belkin and P. Niyogi. Laplacian eigenmaps and spectral techniques for embedding and clustering. In *NIPS 14*. 2001.

[Belkin *et al.*, 2006] M. Belkin, P. Niyogi, and V. Sindhwani. Manifold regularization: A geometric framework for learning from examples. *Journal of Machine Learning Research*, 7:2399–2434, 2006.

[Brunet *et al.*, 2004] Jean-Philippe Brunet, Pablo Tamayo, Todd R. Golub, and Jill P. Mesirov. Metagenes and molecular pattern discovery using matrix factorization. *Proceedings of the National Academy of Sciences*, 101(12):4164–4169, 2004.

[Cai *et al.*, 2005] Deng Cai, Xiaofei He, and Jiawei Han. Document clustering using locality preserving indexing. *IEEE Transactions on Knowledge and Data Engineering*, 17(12):1624–1637, December 2005.

[Cai *et al.*, 2008] Deng Cai, Xiaofei He, Xiaoyun Wu, and Jiawei Han. Non-negative matrix factorization on manifold. In *Proc. Int. Conf. on Data Mining (ICDM'08)*, 2008.

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

[Deerwester *et al.*, 1990] 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.

[Dempster *et al.*, 1977] 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.

[Duda *et al.*, 2000] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*. Wiley-Interscience, Hoboken, NJ, 2nd edition, 2000.

[He and Niyogi, 2003] Xiaofei He and Partha Niyogi. Locality preserving projections. In *NIPS 16*. 2003.

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

[Lee and Seung, 2001] Daniel D. Lee and H. Sebastian Seung. Algorithms for non-negative matrix factorization. In *NIPS 13*. 2001.

[Li *et al.*, 2001] Stan Z. Li, XinWen Hou, HongJiang Zhang, and QianSheng Cheng. Learning spatially localized, parts-based representation. In *CVPR'01*, 2001.

[Liu *et al.*, 2008] Wei Liu, Dacheng Tao, and Jianzhuang Li. Transductive component analysis. In *Proc. Int. Conf. on Data Mining (ICDM'08)*, 2008.

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

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

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

[Wachsmuth *et al.*, 1994] 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.

[Xu *et al.*, 2003] Wei Xu, Xin Liu, and Yihong Gong. Document clustering based on non-negative matrix factorization. In *SIGIR'03*, 2003.

[Zha *et al.*, 2001] H. Zha, C. Ding, M. Gu, X. He, and H. Simon. Spectral relaxation for k-means clustering. In *NIPS 14*. 2001.