Nonnegative Matrix Factorization Clustering on Multiple Manifolds

Bin Shen, Luo Si
Department of Computer Science, Purdue University
West Lafayette, IN, 47907, USA

Abstract

Nonnegative Matrix Factorization (NMF) is a widely used technique in many applications such as clustering. It approximates the nonnegative data in an original high dimensional space with a linear representation in a low dimensional space by using the product of two nonnegative matrices. In many applications with data such as human faces or digits, data often reside on multiple manifolds, which may overlap or intersect. But the traditional NMF method and other existing variants of NMF do not consider this. This paper proposes a novel clustering algorithm that explicitly models the intrinsic geometrical structure of the data on multiple manifolds with NMF. The idea of the proposed algorithm is that a data point generated by several neighboring points on a specific manifold in the original space should be constructed in a similar way in the low dimensional subspace. A set of experimental results on two real world datasets demonstrate the advantage of the proposed algorithm.

Introduction

Nonnegative Matrix Factorization (NMF) has been applied in many applications such as clustering and classification. It provides a linear representation of nonnegative data in high dimensional space with the product of two nonnegative matrices as a basis matrix and a coefficient matrix. NMF has received substantial attention due to its theoretical interpretation and desired performance.

Several variants of NMF has been proposed recently. Sparseness constraints have been incorporated into NMF to obtain sparse solutions (Hoyer 2004; Kim and Park 2008). Discriminative NMF algorithms have been proposed in (Zafeiriou et al. 2007; Wang et al. 2004) to maximize the between-class distance and minimize the within-class distance when learning the basis and coefficient matrices. Some recent research work suggest data of many applications in a high dimensional Euclidean space are usually embedded in a low dimensional manifold (Roweis and Saul 2000; Niyogi 2003). To explore the local structure on the low dimensional manifold, research work in (Cai et al. 2009; Gu and Zhou 2009) have proposed Locality Preserving NMF and Neighbourhood Preserving NMF, which add constraints between a point and its neighboring one(s).

However, many real world data reside on multiple manifolds (Goldberg et al. 2009). For example, for handwritten digits, each digit forms its own manifold in the feature space. Furthermore, for human faces, the faces of the same person lie on the same manifold and different persons are associated with different manifolds. All existing NMF algorithms do not consider the geometry structure of multiple manifold and cannot model data on a mixture of manifolds, which may overlap or intersect. In particular, the algorithms in (Cai et al. 2009; Gu and Zhou 2009) only consider the situation that all the data samples are drawn from a single manifold. These algorithms create a graph by searching nearest neighbor(s) to preserve the local geometry information: locality or neighborhood. However, the created graph may connect points on different manifolds, which can diffuse information across manifolds and be misleading.

This paper proposes a novel clustering algorithm with NMF that explicitly models the intrinsic geometrical structure of the data on multiple manifolds. The assumption is that data samples are drawn from multiple manifolds, and if one data point can be reconstructed by several neighboring points on the same manifold in the original high dimensional space, it should also be reconstructed in a similar way within the low dimensional subspace by the basis matrix and coefficient matrix. This approach is different from local linear embedding which only studies one manifold in the original space. This paper derives multiplicative updating rules for the proposed algorithm with guaranteed convergence.

The proposed NMF clustering algorithm on multiple manifolds has been evaluated on two real world datasets. It has been shown to generate more accurate and robust results than the K-means and two variants of NMF algorithms. Furthermore, it has been shown that the new algorithm can learn better representation for data in different classes than the traditional NMF method.

The rest part of this paper is organized as follows. It first reviews the NMF algorithm, and then presents the proposed algorithm followed by the optimization method. Furthermore, it describes and analyzes the experimental results. Finally, it concludes and discusses future work.

Review of Nonnegative Matrix Factorization

Given a nonnegative matrix $X \in \mathcal{R}^{m \times n}$, each column of X is a data sample. The NMF algorithm aims to approxi-

mate this matrix by the product of two nonnegative matrices $U \in \mathcal{R}^{m\times k}$ and $V \in \mathcal{R}^{k\times n}$. To achieve this, the following objective function is minimized:

$$\begin{aligned} O &= ||X - UV||_F^2 \\ s.t.\ & U \geq 0, V \geq 0 \end{aligned} \tag{1}$$

where $||.||_F$ denotes the Frobenius norm.

The following iterative multiplicative updating algorithm is proposed in (Lee and Seung 2001) to minimize the objective function:

$$U_{ij} = U_{ij}\frac{(XV^T)_{ij}}{(UVV^T)_{ij}} \tag{2}$$

$$V_{ij} = V_{ij}\frac{(U^TX)_{ij}}{(U^TUV)_{ij}} \tag{3}$$

Proposed Nonnegative Matrix Factorization Clustering on Multiple Manifolds

This section presents the formulation of the proposed Nonnegative Matrix Factorization on Multiple Manifolds(MM-NMF for short). It then describes the optimization algorithm.

Formulation

Given a nonnegative matrix X, each column of which is a data sample. In our target applications with data such as human faces and digits, our assumption is that data samples are drawn from a mixture of manifolds. We use a product of two nonnegative matrices U and V to approximate X while taking the multiple manifold structure information into consideration. The matrix U can be viewed as the basis, while the matrix V can be treated as the coefficients.

Considering that data samples are drawn from different manifolds, the matrix U represent bases for different manifolds, and a data sample $X_{.i}$ only belongs to a manifold (as long as it is not drawn at the intersection). We denote the manifold from which $X_{.i}$ is drawn as M_i. Ideally, there should be a subset of columns of U associated with the manifold M_i. So the corresponding coefficient vector of this sample $V_{.i}$ should have nonzero values only for the entries which correspond to the subset of columns of U associated with the manifold M_i. Thus, the coefficient matrix V should be naturally sparse.

Another important goal is to encode the geometrical information of multiple manifolds. When there are more than one manifolds, we cannot create a graph which directly connects the neighboring samples according to Euclidean distance, since samples close to each other may belong to different manifolds, especially near the intersection of different manifolds.

Based on the above discussion, our MM-NMF algorithm on multiple manifolds should be equipped with two properties: 1) the coefficient matrix V is sparse. In other words, the representation of the samples in the new space is sparse; 2) the local geometrical information on each manifold is preserved. In the following part, we will describe how we formulate MM-NMF with these two desired properties.

Sparse Representation in Output Space

The traditional NMF will learn part-based representation due to its nonnegativity constraint. This means that the data representation in the output space spanned by U is sparse.Our algorithm also inherits the nonnegativity constraint, which will introduce some sparseness.

However, the sparseness introduced by nonnegativity may not be enough and is difficult to control (Hoyer 2004). Some prior research made the matrix sparse by adding an extra sparseness constraint which is usually related with the L1 norm minimization technique (Donoho 2006). We utilize the method in (Kim and Park 2008) , which is denoted as SNMF, to make the coefficient matrix V sparse.

The objective function of SNMF is defined as follows:

$$O_{SNMF} = ||X - UV||_F^2 + \zeta\|U\|_F^2 + \lambda\sum_j \|V_{.j}\|_1^2 \tag{4}$$

where $V_{.j}$ is the jth column of V. The term $\lambda\sum_j \|V_{.j}\|_1^2$ encourages the sparsity, and $\zeta\|U\|_F^2$ is used to control the scale of matrix U. Parameter λ controls the desired sparsity, and ζ is simply set as the maximum value of X.

Since the nonnegativity constraint automatically introduces some degree of sparseness, we will study two cases with and without the extra sparseness regularizer $\lambda\sum_j \|V_{.j}\|_1^2$ in this paper.

Mining Intrinsic Geometry on Individual Manifolds

This section targets on the second property, which is to preserve the local geometrical information on each manifold in the matrix factorization. We try to preserve the neighborhood relation on each manifold. Note that the neighborhood relation is defined on the manifold rather than in the Euclidean space. This local geometry information on each manifold will guide the formulation of sparseness, which is similar with joint sparsity constraint (Turlach, Venablesy, and Wright 2005).

To achieve the goal, we assume there are enough data samples so that any data sample can be well approximated by a linear combination of neighboring data samples on the same manifold, since the manifold is usually smooth (Elhamifar and Vidal 2009).

Now we describe how to explore the intrinsic geometry in the data matrix X with size of $M \times N$. Let $X_{.i}$ be the i_{th} sample, which is under consideration now. We want to identify its neighbors on the same manifold rather than the entire Euclidean space. As there are enough samples and the manifold is smooth, the data point can be well approximated by a linear combination of a few nearby samples on the same manifold. Thus, it has a sparse representation over the entire data matrix X. To identify the set of samples that can approximate $X_{.i}$, we use the sparsest linear combination which can approximate $X_{.i}$ by the L1 norm minimization technique (Donoho 2006). We obtain a sparse structure matrix S from the equation of $X = XS$, where the diagonal elements of S are 0. This means any sample can be

expressed by several other samples on the same manifold. We construct the S as follows.

For any $i = 1, ..., N$

$$\begin{aligned} &\min_{S_{.i}} ||S_{.i}||_1 \\ s.t. \quad &X_{.i} = XS_{.i} \\ &S_{ii} = 0 \end{aligned} \tag{5}$$

There may be some noise in real applications and the equality constraint above may not hold, we relax it to the following equation:

$$\begin{aligned} &\min_{S_{.i}} ||S_{.i}||_1 \\ s.t. \quad &\|X_{.i} - XS_{.i}\|_2 < \epsilon \\ &S_{ii} = 0 \end{aligned} \tag{6}$$

Ideally, the nonzero entries in the vector $S_{.i}$ correspond to the samples which lie on the same low dimensional manifold as $X_{.i}$. On the other hand, the nearest samples in Euclidean space from another manifold may not appear in the nonzero entries. ϵ controls the noise energy, and is set to 0.01 here.

MM-NMF Objective Function

We preserve the geometry relation represented by S in the matrix factorization. When the matrix is factorized, we try to preserve geometry constraint from S for V. This can be gained by minimizing

$$\begin{aligned} &\sum_i ||V_{.i} - VS_{.i}||_2 \\ &= ||V - VS||_F \\ &= ||V(I - S)||_F \\ &= tr(V(I - S)(I - S)^T V^T) \\ &= tr(VGV^T) \end{aligned} \tag{7}$$

where $I \in \mathcal{R}^{n \times n}$, and $G = (I - S)(I - S)^T$.

Considering both of the two properties we want to engage: 1, sparseness; 2, local structure preservation on each manifold, the objective function of MM-NMF is defined as:

$$\begin{aligned} O_{MM-NMF} = &||X - UV||_F^2 + \zeta\|U\|_F^2 + \lambda \sum_j \|V_{.j}\|_1^2 \\ &+ \eta tr(VGV^T) \end{aligned} \tag{8}$$

When minimizing the objective function above, we should add constraints that the U and V be nonnegative. The first term is the square fitting error, the second term controls the energy of U, the third term encourages the sparseness, and the last term is to preserve the local geometry structure on each manifold.

Now consider a special case of MM-NMF, where there is no sparseness regularizer($\lambda = 0$). This means we only rely on the nonnegativity constraint to engage the V with first property: sparseness. We name this special case as MM-NMF2. The objective function is:

$$O_{MM-NMF2} = ||X - UV||_F^2 + \zeta\|U\|_F^2 + \eta tr(VGV^T) \tag{9}$$

Optimization

Here we only consider how to optimize the objective function of MM-NMF, since MM-NMF2 is a special of case of MM-NMF.

Since O_{MM-NMF} is not convex with U and V jointly, it is difficult to find the global minimum for O_{MM-NMF}. Instead, we aim to find a local minimum for O_{MM-NMF} by iteratively updating U and V in a similar way with the work (Lee and Seung 2001) for NMF.

Update U Given V, we update U to decrease the value of objective function. In the following equation we follow (Kim and Park 2008) to reformulate the objective function for computational convenience.

$$\begin{aligned} U =& arg\min_{U \geq 0} ||X - UV||_F^2 + \zeta\|U\|_F^2 + \lambda \sum_j \|V_{.j}\|_1^2 \\ &+ \eta tr(VGV^T) \\ =& arg\min_{U \geq 0} ||X - UV||_F^2 + \zeta\|U\|_F^2 \\ =& arg\min_{U \geq 0} ||(X, 0_{m \times k}) - U(V, \sqrt{\zeta} I_k)||_F^2 \\ =& arg\min_{U \geq 0} ||\widetilde{X} - U\widetilde{V}||_F^2 \end{aligned} \tag{10}$$

where $\widetilde{X} = (X, 0_{m \times k})$ and $\widetilde{V} = (V, \sqrt{\zeta} I_k)$.

The updating rule for U to (Lee and Seung 2001)reduce the objective function can be either of the following ones, which can be proven in a similar way in (Lee and Seung 2001; Gu and Zhou 2009),

$$U_{ij} = U_{ij} \frac{(\widetilde{X}\widetilde{V}^T)_{ij}}{(U\widetilde{V}\widetilde{V}^T)_{ij}} \tag{11}$$

$$U_{ij} = U_{ij} \sqrt{\frac{(\widetilde{X}\widetilde{V}^T)_{ij}}{(U\widetilde{V}\widetilde{V}^T)_{ij}}} \tag{12}$$

Update V Now we decrease the objective function with respect to V given U.

$$\begin{aligned} V &= arg\min_V O_{MM-NMF} \\ &= arg\min_V ||X - UV||_F^2 + \lambda \sum_j \|V_{.j}\|_1^2 + \eta tr(VGV^T) \\ &= arg\min_V ||\begin{pmatrix} X \\ 0_{1 \times n} \end{pmatrix} - \begin{pmatrix} U \\ \sqrt{\lambda} e_{1 \times k} \end{pmatrix} V||_F^2 + \eta tr(VGV^T) \\ &= arg\min_V ||\widetilde{X} - \widetilde{U}V||_F^2 + \eta tr(VGV^T) \end{aligned} \tag{13}$$

where $\widetilde{X} = \begin{pmatrix} X \\ 0_{1\times n} \end{pmatrix}$ and $\widetilde{U} = \begin{pmatrix} U \\ \sqrt{\lambda} e_{1\times k} \end{pmatrix}$.

This optimization step can be done efficiently using the following updating rule, which can be proven in a similar way in (Gu and Zhou 2009),

$$V_{ij} = V_{ij}\sqrt{\frac{(\widetilde{U}^T\widetilde{X} + \eta V G^-)_{ij}}{(\widetilde{U}^T\widetilde{U}V + \eta V G^+)_{ij}}} \tag{14}$$

where

$$\begin{aligned} G &= G^+ - G^- \\ G^+_{ij} &= \frac{|G_{ij}| + G_{ij}}{2} \\ G^-_{ij} &= \frac{|G_{ij}| - G_{ij}}{2} \end{aligned} \tag{15}$$

Convergence Analysis

Since both updating methods for U and V are non-increasing, and the objective function clearly has a lower bound, for example, 0, thus the algorithm will converge.

Experimental Results

Data Set

We conduct clustering experiments on real data sets: the ORL face database, and the USPS handwritten digits.

The ORL face database contains ten images for each of the forty human subjects, which are taken at different times, under different lighting condition, with different facial expression and with/without glasses. All faces are resized to 32×32 for efficiency. Figure 1 shows some sample images from ORL data set.

Figure 1: Examples of ORL face database

The USPS digit dataset contains 8-bit gray scale images of "0" through "9". Each image is of size 16×16, and there are 1100 images for each class. Figure 2 shows some images from this dataset.

Evaluation Metric

To evaluate the performance of clustering, we use two metrics (Xu, Liu, and Gong 2003; Cai, He, and Han 2008): 1, accuracy; 2, normalized mutual information(NMI).

The clustering algorithm is tested on N samples. For a sample x_i, the cluster label is denoted as r_i, and ground true label is t_i. The accuracy is defined as follows:

$$accuracy = \frac{\sum_{i=1}^{N} \delta(t_i, map(r_i))}{N} \tag{16}$$

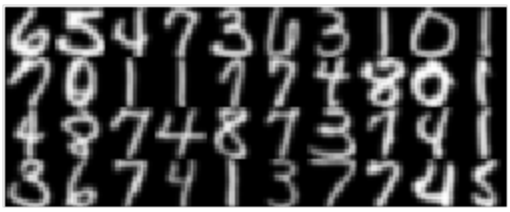

Figure 2: Examples of USPS Handwritten Digits

where $\delta(x, y)$ is equal to 1 if x is equal to y, and 0 otherwise. Function $map(x)$ is the best permutation mapping function gained by Kuhn-Munkres algorithm (Lovasz and Plummer 1986), which maps cluster to the corresponding predicted label. So, we can easily see that the more labels of samples are predicted correctly, the greater the accuracy is.

For the second metric, let C denote the cluster centers of the ground truth, and C' denote the cluster centers by clustering algorithm. The mutual information is defined as follows:

$$MI(C, C') = \sum_{c\in C; c'\in C'} p(c, c') log \frac{p(c, c')}{p(c)p(c')} \tag{17}$$

where $p(c)$ and $p(c')$ are the probabilities that a document belongs to cluster c and c' respectively, and $p(c, c')$ is the probability that a document jointly belongs to cluster c and cluster c'. The normalized mutual information(NMI) is defined as follows:

$$NMI(C, C') = \frac{MI(C, C')}{max((H(C)), (H(C')))} \tag{18}$$

where $H(C)$ and $H(C')$ are the entropies of C and C'. We can easily see that NMI measures the dependency of two distributions, and higher value of NMI means greater similarity between two distributions.

Clustering on Human Faces

We conduct ten independent experiments on the ORL face dataset. In each experiment, we randomly select ten subjects, and get 100 images since there are ten images for each subject. Then the clustering algorithms are run on the 100 images. Thus X is a matrix with the size of 1024×100. U is a matrix with size of 1024×10 and V is of size 10×100. We don't list the performance SNMF, because according to our experimental results SNMF cannot outperform NMF, which means the best choice of the parameter λ is 0. When λ is set to 0, SNMF is equivalent to NMF.

As mentioned above, ζ is simply fixed as the maximum value of the input matrix. The parameter λ is searched within $[0, 0.5]$. However, this parameter is not critical as we will see MM-NMF2, which simply sets λ to 0, still has competitive performance. The parameter η reflects how reliable the structure information encoded by S is. Generally speaking, if there are more samples, the S will be more reliable since the samples on each manifold will be denser, and the η should have larger value. In this paper, η is set by searching only on the grid of $\{0.1, 1, 10, 100, 1000\}$. In the

Method	K-means	NMF	MM-NMF2	MM-NMF
1	71.0	67.0	66.0	**82.0**
2	60.0	54.0	**80.0**	79.0
3	58.0	59.0	81.0	**84.0**
4	58.0	61.0	**79.0**	78.0
5	**70.0**	52.0	68.0	64.0
6	63.0	62.0	73.0	**75.0**
7	66.0	65.0	72.0	**80.0**
8	48.0	61.0	**67.0**	64.0
9	59.0	65.0	54.0	**72.0**
10	55.0	**59.0**	**59.0**	56.0
Average	60.8	60.5	69.9	**73.4**
Std.	7.0	4.8	8.9	9.2

Table 1: Clustering Accuracy(%) on ORL face database

Method	K-means	NMF	MM-NMF2	MM-NMF
1	80.4	69.3	77.8	**88.8**
2	73.4	62.1	82.1	**83.6**
3	72.4	68.8	82.2	**82.6**
4	69.1	66.2	81.3	**83.5**
5	**80.5**	58.6	75.2	70.3
6	68.8	63.2	75.5	**80.6**
7	76.2	77.8	79.7	**85.2**
8	59.5	63.1	67.8	**70.1**
9	67.4	64.3	63.6	**72.8**
10	**65.4**	62.9	64.1	64.9
Average	71.3	65.6	74.9	**78.2**
Std.	6.6	5.3	7.2	8.0

Table 2: Clustering NMI(%) on ORL face database

experiments on face clustering, our λ for MM-NMF is 0.01, and η for MM-NMF and MM-NMF2 is 1.

The accuracy and NMI of different algorithms are shown in table 1 and table 2, respectively. The average performance and standard deviation of the algorithms are listed in the bottom lines of the tables.

From the results, we can see that our algorithms MM-NMF2 and MM-NMF outperform traditional NMF by about ten percent, and MM-NMF is slightly better than MM-NMF2, which means the sparseness regularizer is helpful. In other words, the sparseness introduced by nonnegativity constraint is not enough for this task, and we need extra sparseness regularization.

According to our experiments, the performance is not sensitive to the value of η. For example, when η is set on 0.1, the average accuracy of MM-NMF is 68.6%, which is still much better than NMF.

After the matrix factorization, we gain a nonnegative basis matrix U with size of 1024×10. Ideally, each column should represent a human subject. We visualize these basis learned by NMF and MM-NMF in figure 3 and figure 4, respectively. From these basis, we can easily see that MM-NMF learns better representation for each class. This is mainly because they are manifold specific.

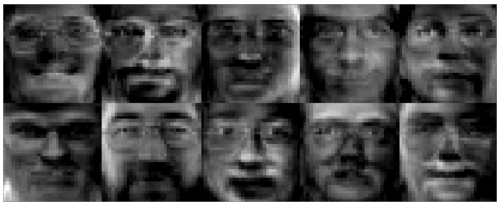

Figure 3: Face Basis Learned by NMF

Figure 4: Face Basis Learned by MM-NMF

Clustering on Handwritten Digits

For experiment on USPS digits, we randomly select 1000 samples from the dataset. Thus X is a matrix with the size of 256×1000, U is of size 256×10 and V is of size 10×1000. The scheme to set parameters here is the same as face clustering. The λ here is 0.4, and η is set to 100.

The accuracy and NMI of different algorithms are shown in table 3 and table 4, respectively.

From the results, we can also see that our algorithms MM-NMF2 and MM-NMF outperform traditional NMF by about ten percent in both accuracy and NMI, and MM-NMF is slightly better than MM-NMF2, which again shows the sparseness regularizer is helpful. Here K-means also has comparable good performance. It outperforms traditional NMF, but it still cannot outperform our MM-NMF.

In the experiments on digits, we have more examples than experiments on faces, thus the η has larger value as mentioned above. Experiments are also conducted when η is set as 10 and 1, and MM-NMF still generates robust results

Method	K-means	NMF	MM-NMF2	MM-NMF
1	64.5	54.4	**69.0**	64.3
2	56.8	48.1	**65.9**	63.6
3	56.9	55.4	**62.0**	61.2
4	50.9	52.3	55.9	**64.5**
5	58.4	53.2	61.0	**61.5**
6	65.8	52.2	**68.4**	65.9
7	63.8	59.1	53.8	**65.5**
8	**65.8**	50.5	54.6	**65.8**
9	**65.0**	47.5	56.8	60.1
10	62.6	55.7	66.3	**68.1**
Average	61.1	52.8	61.3	**64.1**
Std.	5.0	3.6	5.8	2.5

Table 3: Clustering Accuracy(%) on USPS Digits Dataset

Method	K-means	NMF	MM-NMF2	MM-NMF
1	59.7	51.3	**60.9**	58.8
2	61.3	46.4	**65.0**	59.6
3	58.5	49.9	58.9	**60.3**
4	54.0	46.7	53.5	**60.8**
5	**59.0**	45.8	56.0	57.4
6	63.2	50.0	**66.2**	61.6
7	59.4	55.6	53.1	**61.2**
8	63.2	50.9	59.2	**64.4**
9	**63.9**	47.3	58.5	58.5
10	59.6	53.4	62.6	**64.0**
Average	60.2	49.7	59.4	**60.9**
Std.	2.9	3.2	4.4	2.3

Table 4: Clustering NMI(%) on USPS Digits Dataset

with the average accuracy as 61.2% and 60.7%, which are still much better than NMF.

In experiments, each column of U should represent a digit. We visualize these basis U learned by NMF and MM-NMF in figure 5 and figure 6, respectively. From these two figure, we can easily see that basis learned by MM-NMF have much clearer interpretation than basis learned by NMF.

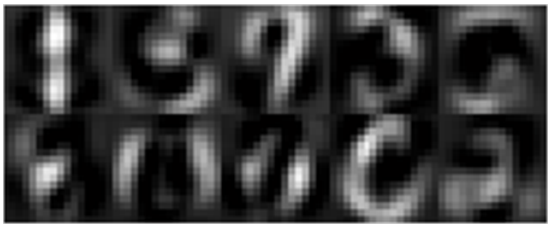

Figure 5: Digit Basis Learned by NMF

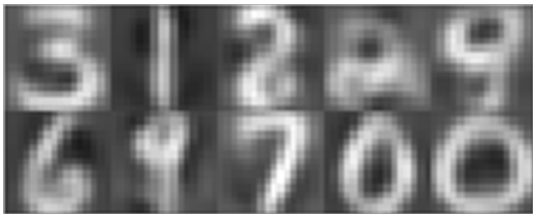

Figure 6: Digit Basis Learned by MM-NMF

Conclusion

We observe that when the data are supported on more than one manifold, a good representation should be sparse since any data sample is drawn from a single manifold. Also, the local geometry structure should be preserved on each manifold rather than the entire set of manifolds. Based on the observation, we propose a nonnegative matrix factorization algorithm for data samples drawn from a mixture of manifolds, which is never considered before. The sparseness is encouraged and the local geometric structure on each of the manifolds is preserved in our our algorithm. The corresponding objective function can be efficiently minimized by two iterative multiplicative updating rules. Experimental results on real datasets show that our algorithm gains better clustering results and learns better interpretable representation for each class. For future, clustering for data with multiple labels on multiple manifolds is a promising direction.

References

Cai, D.; He, X.; Wang, X.; Bao, H.; and Han, J. 2009. Locality preserving nonnegative matrix factorization. In *IJCAI'09: Proceedings of the 21st international joint conference on Artifical intelligence*, 1010–1015. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Cai, D.; He, X.; and Han, J. 2008. Graph regularized non-negative matrix factorization for data representation. In *UIUC Computer Science Research and Tech Reports*.

Donoho, D. L. 2006. For most large underdetermined systems of equations, the minimal l1-norm near-solution approximates the sparsest near-solution. *Communications on pure and applied mathematics* 59(7):907–934.

Elhamifar, E., and Vidal, R. 2009. Sparse subspace clustering. In *IEEE Conference on Computer Vision and Pattern Recognition.*, 2790–2797.

Goldberg, A.; Zhu, X.; Singh, A.; Xu, Z.; and Nowak, R. 2009. Multi-manifold semi-supervised learning. In *Twelfth International Conference on Artificial Intelligence and Statistics*.

Gu, Q., and Zhou, J. 2009. Neighborhood preserving nonnegative matrix factorization. In *The 20th British Machine Vision Conference*.

Hoyer, P. O. 2004. Non-negative matrix factorization with sparseness constraints. *J. Mach. Learn. Res.* 5:1457–1469.

Kim, J., and Park, H. 2008. Sparse nonnegative matrix factorization for clustering. In *CSE Technical Reports*. Georgia Institute of Technology.

Lee, D. D., and Seung, H. S. 2001. Algorithms for non-negative matrix factorization. In *NIPS*, 556–562. MIT Press.

Lovasz, L., and Plummer, M. D. 1986. *Matching Theory*. Akademiai Kiado.

Niyogi, P. 2003. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation* 15:1373 – 1396.

Roweis, S. T., and Saul, L. K. 2000. Nonlinear dimensionality reduction by locally linear embedding. *Science* 290(5500):2323 – 2326.

Turlach, B. A.; Venablesy, W. N.; and Wright, S. J. 2005. Simultaneous variable selection. *Technometrics* 27:349 – 363.

Wang, Y.; Jia, Y.; Hu, C.; and Turk, M. 2004. Fisher non-negative matrix factorization for learning local features. In *Asian Conference on Computer Vision*.

Xu, W.; Liu, X.; and Gong, Y. 2003. Document clustering based on non-negative matrix factorization. In *SIGIR '03: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval*, 267–273. New York, NY, USA: ACM.

Zafeiriou, S.; Tefas, A.; Buciu, I.; and Pitas, I. 2007. Exploiting discriminant information in non-negative matrix factorization with application to frontal face verification. *IEEE Transactions on Neural Networks* 17(3):683 – 695.