Guidance, Navigation and Control
Vol. 4, No. 3 (2024) 2450013 (19 pages)

Learning a Subspace and Clustering Simultaneously with Manifold Regularized Nonnegative Matrix Factorization

Feiping Nie* and Huimin Chen†

School of Computer Science
Northwestern Polytechnical University
Xi'an 710072, P. R. China

School of Artificial Intelligence
Optics and Electronics (iOPEN)
Northwestern Polytechnical University
Xi'an 710072, P. R. China

Key Laboratory of Intelligent Interaction
and Applications (Northwestern Polytechnical University)
Ministry of Industry and Information Technology
Xi'an 710072, P. R. China
**feipingnie@gmail.com*
†chenhuimin@mail.nwpu.edu.cn

Heng Huang

Department of Electrical and Computer Engineering
University of Pittsburgh, Pittsburgh, PA 15260, USA

Chris H. Q. Ding

Department of Computer Science and Engineering
The Chinese University of Hong Kong
Shenzhen 518172, P. R. China

Xuelong Li

Institute of Artificial Intelligence (TeleAI)
China Telecom Corporation Limited
31 Jinrong Street, Beijing 100033, P. R. China

Received 21 April 2024
Revised 27 May 2024
Accepted 30 May 2024
Published 26 July 2024

*Corresponding author.

Cite this: Feiping Nie, Huimin Chen, Heng Huang, Chris H. Q. Ding, Xuelong Li, Learning a Subspace and Clustering Simultaneously with Manifold Regularized Nonnegative Matrix Factorization, *Guid. Navig. Control*, Vol. 4, No. 3 (2024) 2450013, doi:10.1142/S2737480724500134

With the incredible growth of high-dimensional data such as microarray gene expression data and web blogs from internet, the researchers are desirable to develop new clustering techniques to address the critical problem created by irrelevant dimensions. Properties of Nonnegative Matrix Factorization (NMF) as a clustering method were studied by relating its formulation to other methods such as K-means clustering. In this paper, by introducing clustering indicator constraints on NMF and incorporating manifold regularization to preserve geometric structures, we propose a novel manifold regularized NMF method that can simultaneously learn subspace and do clustering. As a result, our clustering results can directly assign cluster label to data points. Extensive experimental results show that our method outperforms related other methods.

Keywords: Subspace learning; clustering; nonnegative matrix factorization; manifold learning.

1. Introduction

There is a growing need for cluster analysis to partition data sets into similar groups unsupervisedly.[1,2] For example, digital libraries and the World Wide Web continue to grow exponentially, the ability to find useful information increasingly depends on the indexing infrastructure or search engine. Clustering techniques can be used to discover natural groups in data sets and to identify hidden intrinsic structures without any background knowledge of the data, so they have good generalization capabilities. Up to now, clustering has been used in a variety of areas, including computer vision,[3] data mining,[4] bioinformatics,[5] urban development,[6] information retrieval,[7] etc. To address these applications and many others, a variety of clustering algorithms have been developed. Some of the commonly used clustering methods include K-means,[8] Spectral Clustering (SC),[9] Density-Based Spatial Clustering of Applications with Noise (DBSCAN),[10] etc.

One of the most popular clustering algorithms is the K-means algorithm,[8,11] which has implementation simplicity and low computational complexity,[12] and therefore has broad applicability in many clustering application fields and has been fully studied in the past decades.[13] For example, works such as Refs. 14–16 improved the selection of initial centroids in Lloyd's original K-means to help solve the variance of the number of iterations and final clustering results caused by the random initialization. References 17 and 18 explored the possibility of accelerating Lloyd's original K-means by approximating clustering results. Works such as Refs. 19 and 20 accelerated an exact K-means algorithm to speed up K-means with another strategy. However, such a widely used and researched algorithm still has limitations, that is, it is based on the *a priori* assumption that the data follow a Gaussian distribution, and thus its high accuracy is limited to specific types of data. Correspondingly, another classic clustering technique, SC, has a wider ability to process data of different shapes. Since they consider the potential manifold structure within the data by calculating the affinity between nearest neighbors, they can be used in more complex scenarios. The use of manifold information in SC can greatly improve the clustering results with preserving geometric structures. It is well known that there is a connection between Laplacian–Beltrami operator and graph Laplacian in spectral graph theory.[21,22] Laplacian embedding provides an approximation solution of the ratio cut clustering,[9] and the generalized eigenvectors of the Laplace matrix provide an approximation solution of the normalized cut clustering[23] and min–max clustering.[24] However, such algorithms could produce suboptimal results because they typically follow a two-stage process in optimization.

Despite these advancements, as the dimensionality of real data in many real world applications increases, e.g. gene expression, images, and documents, clusterings containing both K-means and SC suffer from the curse of dimensionality and are susceptible to the impact of redundant features and noise. Specifically, in high-dimensional spaces, the data distribution becomes sparse, rendering traditional distance functions, such as Euclidean distance, less effective for comparing distances between data points. Specifically, in high-dimensional data, many of the dimensions are often irrelevant and only confuse clustering algorithms. As the number of data dimensions increases, distance measures become increasingly meaningless. Moreover, the analysis of high-dimensional data imposes demanding requirements on storage and computation resources. Classic improvements addressed this efficiency issues by embedding the data into a lower dimensional space before applying clustering algorithms. This dimensionality reduction can be achieved through two mainstream methods, one is feature selection,[25–27] and the other is feature extraction.[28,29]

For example, Principle Component Analysis (PCA)[30] and Multidimensional Scaling (MDS)[31] embed data into a linear space with linear transformation; IsoMAP,[32] Local Linear Embedding (LLE),[33] Locality Preserving Projections (LPPs),[34] Laplacian eigenmaps[35] project original data onto a nonlinear lower dimensional manifold and use results from spectral graph theory to approximate the geodesic distances. UDPFS[36] is an unsupervised discriminative projection approach for feature selection, which is a fuzzy K-means clustering model with a regularization term and embedding a projection matrix with a $l_{2,1}$ norm regularization term. FSBC[37] combines the K-means clustering model and a projection matrix constrained by the $l_{2,0}$ norm and incorporates a regularization term on the projection matrix to avoid trivial solutions. BSFS[38] model aims at learning a pseudo-label matrix from a pre-constructed graph via discrete SC, incorporates a balance term to control cluster size in clustering results, and regresses the data back to this pseudo-label matrix using a row-sparse projection matrix. RSOGFS[39] is a feature selection model based on row-sparsity constrained projection and optimized graph, which integrates adaptive graph learning and row-sparse projection learning constrained by an l_{20} norm. WPANFS[40] is a feature selection model with weighed and projected adaptive neighbors that integrating structured graph learning, projection matrix learning, and feature weight vector learning. EGCFS[41] meant to preserve both the local data structure and discriminative information, which is a feature selection model that integrates embedded graph learning, inter-class divergence maximization, and $l_{2,1}$ norm regularized sparse projection learning.

Nonnegative Matrix Factorization (NMF) was also introduced as a dimension reduction method for pattern analysis.[42,43] Later K-means clustering objective had been written as the maximization of a quadratic function with nonnegativity and orthogonality constraints.[44] Several further developments using NMF for clustering are convex NMF,[45] orthogonal NMF,[46] and equivalence between NMF and probabilistic latent semantic indexing.[47] Furthermore, a graph regularized NMF was proposed to study geometric structure with assuming that the nearby data points are likely to be in the same cluster.[48] In the latest clustering research, NMF has gradually played an important role, e.g. Ref. 49 puts forward a multi-manifold regularized NMF-based MVC framework, which can preserve the local geometric structure. Wang *et al.*[50] proposed an orthogonal NMF-based clustering formulation that equivalently transforms the orthogonality constraint into a set of norm-based nonconvex equality constraints. Wang *et al.*[50] proposed a new ONMF-based clustering formulation that equivalently transforms the orthogonality constraint into a set of norm-based nonconvex equality constraints. Furthermore, Wang *et al.*[51] proposed a multi-view clustering algorithm based on deep SNMF (MCDS), which is computed in the element way to achieve effective clustering for large-scale data sets.

In this paper, we propose a novel NMF method using manifold regularization to simultaneously learn subspace and do clustering. In our approach, instead of using traditional NMF nonnegative constraint on F and G, we constraint G to be a class indicator matrix, which is a special nonnegative matrix. One advantage of our method is that our clustering results can directly assign cluster labels to data points. Previous NMF-based clustering methods require a post-processing step to extract cluster structure from G and the clustering results may not be unique. Considering to capture the manifold structure in data points, we also incorporate a manifold regularization term in our objective to preserve the geometry structure. We derive and introduce an efficient algorithm to optimize our objective with quick convergence. In our optimization results, we not only obtain the clustering indicator matrix, but also learn a lower dimensional subspace that can be used to unsupervisedly reduce data

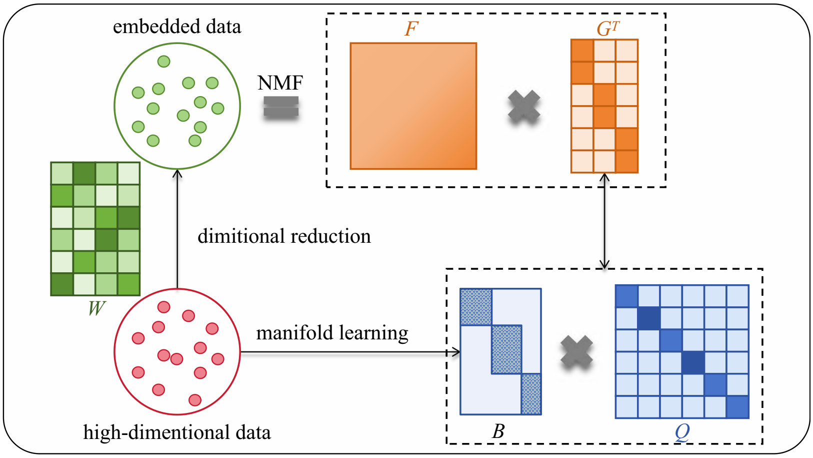


Fig. 1. The flowchart of the proposed model.

dimensionality in classification. The overall clustering flowchart is shown in Fig. 1. The contributions of this paper are summarized below:

- Benefiting from the constraint changes in matrix factorization, an NMF-based clustering that can directly assign cluster labels to data points is proposed, where high-dimensional data are projected into low-dimensional embeddings to perform clustering efficiently and robustly.
- The proposed algorithm incorporates the manifold structure of the samples and is therefore able to adapt to diverse data.
- An algorithm for efficient optimization objectives with fast convergence is also proposed.
- Extensive experiments demonstrate that our proposed method gets better clustering results than previous methods and finds good projection subspace for data dimensionality reduction.

2. Reformulation of the *K*-means Clustering

Suppose we have n data points $x_i \in \mathbb{R}^{d\times 1}(1 \le i \le n)$, the clustering task is to assign each data point to one of the c clusters. Denote the data matrix by $X = [x_1, x_2, \dots, x_n] \in \mathbb{R}^{d\times n}$. We define the class indicator matrix as $G = [G_1^T, G_2^T, \dots, G_n^T]^T \in \mathbb{R}^{n\times c}$, where G_i $(1 \le i \le n) \in \mathbb{R}^{c\times 1}$ is a vector in which one and only one element is 1 and others are zeros. Denote the set of all the class indicator matrices defined above as Ψ.

Traditional *K*-means clustering is an iterative algorithm to update the cluster means and the cluster assignment of data points iteratively. We show that the iterative algorithm is exactly the alternative optimization procedure to solve the following NMF problem:

$$\min_{G\in\Psi, F} \|X - FG^T\|^2. \tag{1}$$

Note that this problem is a little different from the traditional NMF problem. Traditional NMF problem constraints F and G to be nonnegative, while this problem constraints G to be a class indicator matrix, which is a special nonnegative matrix.

As the problem has two variables F and G to be optimized, we can use the alternative optimization method to solve this problem. Specifically, we fix one variable to update the other variable with the optimum one, and iteratively update the two variables until converges.

When we fix G, by setting the derivative with respect to F to zero, we have

$$\begin{aligned} &FG^TG - XG = 0 \\ &\Rightarrow F = XG(G^TG)^{-1}. \end{aligned} \tag{2}$$

Denote the ith column of F by F_i. Based on the definition of the class indicator matrix G, we can see that F_i is the mean of the data points in the current ith cluster.

When we fix F, then $G_i (1 \leq i \leq n)$ are decoupled and we can separately solve the following simpler problem for each $i (1 \leq i \leq n)$:

$$\min_{G \in \Psi} \|x_i - FG_i\|^2. \tag{3}$$

Since G_i $(1 \leq i \leq n) \in \mathbb{R}^{c \times 1}$ is a vector in which one and only one element is 1 and the others are zeros, the solution to the above problem can be easily obtained by

$$G_{ij} = \begin{cases} 1, & j = \arg\min_k \|x_i - F_k\|^2; \\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

The solution indicates that, given current cluster means $F_i (1 \leq i \leq n)$, each data point is assigned to the cluster whose cluster mean is closest to the data point.

From the updated formulas (2) and (4), we can see that the alternative optimization procedure is exactly the same as the traditional K-means iterative procedure. This reformulation of K-means motivates our algorithm proposed in the next section.

3. Proposed Algorithm

3.1. *Objective*

K-means clustering method has been successfully applied in many data clustering applications. However, when the dimension of data is high and the number of data points is limited, K-means clustering method is usually difficult to provide satisfied result. In practice, the high-dimensional data is first projected onto a low-dimensional subspace via some dimension reduction techniques such as PCA. To achieve better clustering performance, several works have been proposed to perform K-means clustering and dimension reduction iteratively for high-dimensional data.[52–55]

In this work, motivated by the reformulation of the K-means clustering method with Eq. (1), we propose to optimize the subspace W and the class indicator matrix G simultaneously with a unified formulation:

$$\min_{\substack{G \in \Psi, F, \\ W^T W = I}} \|W^T X - FG^T\|^2, \tag{5}$$

where W is a low-dimensional projection matrix, which linearly combines the features in the original high-dimensional space and obtains a set of lower dimensional features through self-expression to project the data in the high-dimensional space into a low-dimensional subspace. The constraint $W^T W = I$ is imposed to avoid trivial solution.

High-dimensional data points usually distributed on a low-dimensional manifold. K-means clustering method did not consider the manifold or geometry structure of the data. The use of manifold information in SC has shown the state-of-the-art clustering performance in many computer vision applications, such as image segmentation.[23,56]

In order to capture the manifold structure in data points, we will incorporate a manifold regularization term in Eq. (5).

First, let us denote $\mathcal{G} = \{\mathcal{X}, A\}$ as an undirected weighted graph with a vertex set $\mathcal{X}$ and an affinity matrix $A \in \mathbb{R}^{n\times n}$, in which each entry A_{ij} of the symmetric matrix A represents the affinity of a pair of vertices. The common choice of A_{ij} is defined by

$$A_{ij} = \begin{cases} \exp\left(-\frac{\|x_i - x_j\|^2}{\sigma^2}\right), & x_i, x_j \text{ are neighbors;} \\ 0 & \text{otherwise,} \end{cases} \tag{6}$$

where σ is the parameter to control the spread of neighbors. The graph Laplacian matrix L is then defined by $L = D - A$, where D is a diagonal matrix with the diagonal elements as $D_{ii} = \sum_j A_{ij}, \forall i$. In clustering literature, the normalized graph Laplacian matrix is often used, which is defined as $\tilde{L} = D^{-\frac{1}{2}} L D^{-\frac{1}{2}} = I - D^{-\frac{1}{2}} A D^{-\frac{1}{2}} = I - \tilde{A}$, where I is an identity matrix and $\tilde{A} = D^{-\frac{1}{2}} A D^{-\frac{1}{2}}$.

To incorporate manifold regularization, we need to minimize the following problem:

$$\min_{G\in\Psi} \operatorname{tr}(G^T \tilde{L} G), \tag{7}$$

where $\operatorname{tr}(\tilde{A})$ denotes the trace operator of matrix $\tilde{A}$. Note that G is constrained to be a class indicator matrix, directly solving this problem is very difficult. We will use the following way to simplify this problem.

First, the problem (7) is equivalent to

$$\max_{G\in\Psi} \operatorname{tr}(G^T \tilde{A} G). \tag{8}$$

The above problem can be further deduced as

$$\min_{G\in\Psi} \|GG^T - \tilde{A}\|^2. \tag{9}$$

Suppose the low-rank approximation of $\tilde{A}$ is $\tilde{A} = BQ(BQ)^T$, where $B \in \mathbb{R}^{n\times c}$ and Q is an arbitrary orthonormal matrix. We can use the eigendecomposition to obtain the low-rank approximation of $\tilde{A}$. Suppose $\Sigma \in \mathbb{R}^{c\times c}$ is the diagonal eigenvalue matrix of $\tilde{A}$, the diagonal elements of Σ contain the c largest eigenvalues of $\tilde{A}$, $P \in \mathbb{R}^{n\times c}$ is the eigenvector matrix corresponding to the eigenvalue matrix Σ. Then B can be calculated by

$$B = P\Sigma^{1/2}. \tag{10}$$

Thus the problem (9) becomes

$$\min_{G\in\Psi, Q^TQ=I} \|GG^T - BQ(BQ)^T\|^2. \tag{11}$$

If matrix G approximates matrix BQ, then the matrix GG^T approximates the matrix $BQ(BQ)^T$. Hence solving the problem (11) can be reasonably transformed to solve the following problem:

$$\min_{G\in\Psi, Q^TQ=I} \|G - BQ\|^2. \tag{12}$$

Combining Eqs. (5) and (12), we propose to solve the following manifold regularized NMF problem:

$$\min_{\substack{G\in\Psi, F,\\ W^TW=I,\\ Q^TQ=I}} \|W^TX - FG^T\|^2 + \mu\|G - BQ\|^2, \tag{13}$$

where μ is a regularization parameter to balance the two terms.

In the next subsection, we introduce an iterative algorithm to solve this problem.

3.2. *Optimization*

When we fix G, the two terms in Eq. (13) are decoupled. Thus we can solve the following two problems, respectively:

$$\min_{W^TW=I,F} \|W^TX - FG^T\|^2, \tag{14}$$

$$\min_{Q^TQ=I} \|G - BQ\|^2. \tag{15}$$

For the first problem (14), we set the derivative of the objective with respect to F to zero, and obtain the solution of F as

$$\begin{aligned} &FG^TG - W^TXG = 0\\ &\Rightarrow F = W^TXG(G^TG)^{-1}. \end{aligned} \tag{16}$$

By substituting Eq. (16) into Eq. (14), the problem becomes

$$\min_{W^TW=I} \|W^TX - FG^T\|^2 = \min_{W^TW=I} tr(W^TS_wW), \tag{17}$$

where

$$S_w = XX^T - XG(G^TG)^{-1}G^TX^T. \tag{18}$$

An interesting observation is that S_w in Eq. (18) is exactly the within-class scatter matrix defined in Linear Discriminant Analysis (LDA).

For the second problem (15), let $H = B^TG$, then the problem is equivalent to

$$\max_{Q^TQ=I} tr(Q^TH). \tag{19}$$

Suppose the Singular Vector Decomposition of H is $H = U\Lambda V^T$, then $tr(Q^TH)$ can be rewritten as

$$\begin{aligned} &tr(Q^TH)\\ &\quad = tr(Q^TU\Lambda V^T)\\ &\quad = tr(\Lambda V^TQ^TU)\\ &\quad = tr(\Lambda Z)\\ &\quad = \sum_i \lambda_{ii}z_{ii}, \end{aligned} \tag{20}$$

where $Z = V^T Q^T U$, λ_{ii} and z_{ii} are the (i,j)th element of matrix λ and Z, respectively.

Note that Z is an orthonormal matrix, i.e. $Z^T Z = I$, so $z_{ii} \leq 1$. On the other hand, $\lambda_{ii} \geq 0$ as λ_{ii} is singular value of H. Therefore, $tr(Q^T H) = \sum_i \lambda_{ii} z_{ii} \leq \sum_i \lambda_{ii}$, and when $z_{ii} = 1 (1 \geq i \geq c)$, the equality holds. That is to say, $tr(Q^T H)$ reaches the maximum when $Z = I$. Recall that $Z = V^T Q^T U$, thus the solution to problem (19) or problem (15) is

$$Q = UZ^T V^T = UV^T. \tag{21}$$

Now we have derived that when we fix G, all the optimums of F, W and Q can be obtained efficiently.

When we fix F, W and Q, then $G_i (1 \leq i \leq n)$ can be decoupled to solve the following simpler problem for each $i (1 \leq i \leq n)$:

$$\min_{G \in \Psi} \|W^T x_i - FG_i\|^2 + \mu \|G_i - Q^T B_i\|^2. \tag{22}$$

Since $G_i (1 \leq i \leq n) \in \mathbb{R}^{c \times 1}$ is a vector in which one and only one element is 1 and the others are zeros, the solution to the above problem can be easily obtained by

$$G_{ij} = \begin{cases} 1, & j = \arg\min_k (\|W^T x_i - F_k\|^2 - 2\mu (BQ)_{ik}), \\ 0 & \text{otherwise.} \end{cases} \tag{23}$$

3.3. *Algorithms' theoretical analysis*

The procedure of the whole algorithm is described in Algorithm 1. During the specific execution of the algorithm, an alternating optimization strategy is adopted to update separately.

Specifically, when G is fixed, the optimums of F, W and Q are efficiently calculated according to problems (16), (17), and (21), respectively, to minimize the

Algorithm 1.

```
Input: Data matrix $X = [x_1, x_2, \ldots, x_n] \in \mathbb{R}^{d \times n}$.
Calculate $A$ by Eq. (6).
Calculate $B$ by Eq. (10).
repeat
    Initialize $G \in \mathbb{R}^{n \times c}$ with arbitrary class indicator matrix.
    Calculate $F$ by Eq. (16).
    Calculate $W$ by Eq. (17).
    Calculate $Q$ by Eq. (21).
    Calculate $G$ by Eq. (23).
until Converges
Output: $W$ and $G$.
```

objective function in the problem (13). Taking F as an example, assuming that the objective function value before the kth update is $\mathrm{obj}(k-1)^F$, then after the iteration of question (16), the objective function value will be updated to $\mathrm{obj}(k)^f$, and there is obviously a relationship of $\mathrm{obj}(k)^F \leq \mathrm{obj}(k-1)^F$. In the same way, after the other two matrices W and Q are updated, we have the relationships $\mathrm{obj}(k)^W \leq \mathrm{obj}(k-1)^W$ and $\mathrm{obj}(k)^Q \leq \mathrm{obj}(k-1)^Q$. Since $\mathrm{obj}(k)^F = \mathrm{obj}(k-1)^W$ and $\mathrm{obj}(k)^W = \mathrm{obj}(k-1)^Q$, we know that the objective function values $\mathrm{obj}(k)^Q \leq \mathrm{obj}(k-1)^F$ after this series of updates. Similarly, when we fix F, W and Q, the optimums of G are also efficiently calculated according to problem (23) to minimize the objective function in the problem (13), and so we have $\mathrm{obj}(k)^G \leq \mathrm{obj}(k-1)^G$. Since $\mathrm{obj}(k)^Q = \mathrm{obj}(k-1)^G$, obj $(k-1)^F = \mathrm{obj}(k-1)$ and $\mathrm{obj}(k)^G = \mathrm{obj}(k)$, we get $\mathrm{obj}(k) \leq \mathrm{obj}(k-1)$. Since the minimum loss values of matrix decomposition and matrix approximation are both 0, the objective function value of problem (13) has a clear lower bound, so that the convergence of the algorithm is guaranteed.

The time complexity of the above calculation process is as follows: When updating F through Eq. (16), matrix multiplication and inverse are required. Since G is a discrete and sparse indicator matrix, the computational complexity of W^TXG can be approximated to $O(nmd)$, and the value of G^TG can be calculated and stored through a $c \times c$ diagonal matrix in advance in $O(nc)$, which participates in subsequent calculations, so the computational time complexity of updating F can be regarded as $O(nmdc)$. When updating W through Eqs. (17) and (18), the calculation of the within class scatter matrix requires $O(ndc + d^2c^2)$, the update of W in problem (17) is by finding the m smallest eigenvalues of S_w, which has a corresponding time complexity of $O(n^2m)$. So the time complexity of updating W is $O(n^2m + d^2c^2)$. The most time-consuming step when updating the indicator Q is the eigenvalue decomposition of H, which takes $O(n^2c)$. Therefore, the time complexity of updating Q is $O(n^2c)$. The time complexity of updating G is $O(ndm + nc^2)$. Based on the above analysis, the time complexity of the proposed optimization algorithm is $O(nmdc + n^2(m + c))$. In practice, the convergence rate of the algorithm is very fast, and it usually converges in five iterations in most cases.

4. Experiments

In this section, we verify the performance of the proposed algorithm (denoted by mNMF) on two-sides: the performance of clustering and the performance of subspace learning.

4.1. *Experimental results on clustering*

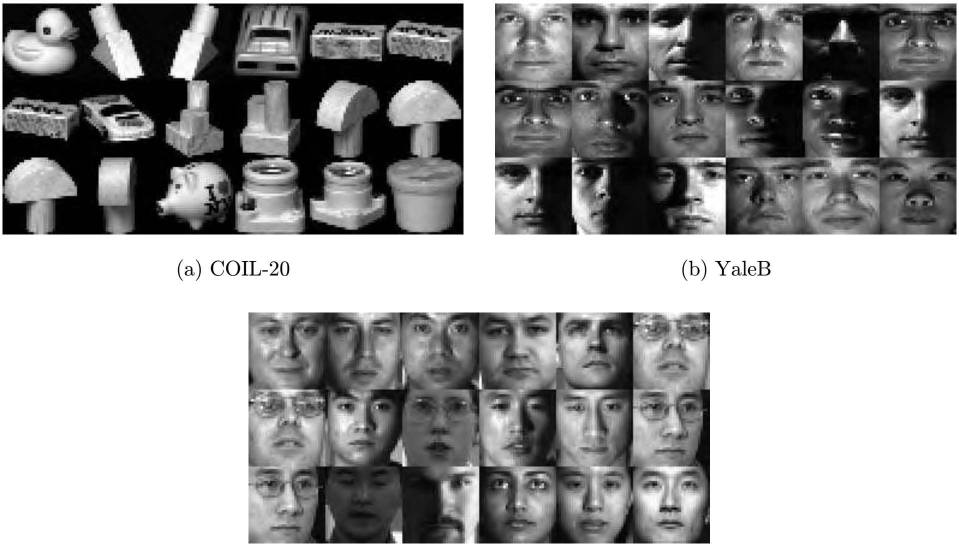
We compare the clustering performance of our algorithm with K-means, PCA+K-means (denoted by PCA+Km), LDA-Km[54] and SC.[57] Ten data sets are used in the experiments, including three UCI data sets, dermatology, ecoli and ionosphere,[a] one object data set, COIL-20, two-digit and character data sets, USPS and Binalpha,

[a] http://www.ics.uci.edu/ mlearn/MLRepository.html.

and four face data sets, Umist, AR, YaleB and PIE. Some data sets are resized, and Table 1 summarizes the details of the data sets used in the experiments, and Fig 2 shows some of the image datasets. We use PCA as a preprocessing to remove the null space of data on all the data sets.

Table 1. Data set description.

Data set	Number	Dimension	Class
Dermatology	366	34	6
Ecoli	336	343	8
Ionosphere	351	34	2
COIL-20	1440	1024	20
USPS	351	34	2
Binalpha	1854	256	10
Umist	575	644	20
AR	840	768	120
YaleB	2414	1024	38
PIE	3329	1024	68



(a) COIL-20

(b) YaleB

(c) PIE

Fig. 2. Part of the image data sets.

SC and mNMF need to determine the neighbors and the parameter σ in (6). In this work, we set the number of neighbors to be 5 in all data sets, and use the self-tune SC[58] method to determine the parameter σ. The dimension of PCA+K-means and mNMF is searched from five candidates ranging from 10 to the dimension of data, and the parameter μ in mNMF is searched from $\{10^{-9}, 10^{-6}, 10^{-3}, 10^{0}, 10^{3}, 10^{6}, 10^{9}\}$. We report the best clustering result from the best parameter.

The results of all clustering algorithms depend on the initialization. To reduce statistical variety, we independently repeat all clustering algorithms for 50 times with random initialization, and then we report the results corresponding to the best objective values.

4.1.1. *Evaluation metrics*

We use the following two popular evaluation metrics to evaluate the performance for all the clustering algorithms.

Clustering Accuracy (ACC) is defined as

$$\mathrm{ACC} = \frac{\sum_{i=1}^{n} \delta(l_i, \mathrm{map}(c_i))}{n},$$

where l_i is the true class label and c_i is the obtained cluster label of x_i, $\delta(x, y)$ is the delta function, and $\mathrm{map}(\cdot)$ is the best mapping function. Note $\delta(x, y) = 1$, if $x = y$; $\delta(x, y) = 0$, otherwise. The mapping function $\mathrm{map}(\cdot)$ matches the true class label and the obtained cluster label and the best mapping is solved by Kuhn–Munkres algorithm. A larger ACC indicates a better performance.

Normalized Mutual Information (NMI) is calculated by

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

where C is a set of clusters obtained from the true labels and C' is a set of clusters obtained from the clustering algorithm. $\mathrm{MI}(C, C')$ is the mutual information metric, and $H(C)$ and $H(C')$ are the entropies of C and C', respectively. See Ref. 59 for more information. NMI is between 0 and 1. Again, a larger NMI value indicates a better performance.

4.1.2. *Experimental results*

The clustering results from various algorithms are reported in Tables 2 and 3. We have the following observations through the tables:

Table 2. Performance comparison of clustering accuracy from K-means, PCA+Km, LDA-Km, SC and mNMF on 10 data sets.

Data set	K-means	PCA+Km	LDA-Km	SC	mNMF
Dermatology	75.96%	75.96%	71.58%	82.79%	**83.06%**
Ecoli	62.91%	**63.99%**	62.80%	42.86%	**63.99%**
Ionosphere	70.66%	70.66%	70.37%	51.28%	**77.78%**
COIL-20	64.10%	67.92%	62.01%	79.31%	**79.79%**
USPS	64.83%	64.89%	66.88%	**68.12%**	68.02%
Binalpha	42.95%	46.30%	46.58%	44.87%	**46.65%**
Umist	45.39%	45.39%	48.52%	**61.04%**	**61.04%**
AR	27.98%	29.17%	24.17%	38.10%	**38.33%**
YaleB	11.06%	12.26%	12.43%	38.61%	**44.86%**
PIE	18.86%	18.56%	22.20%	41.39%	**46.02%**

Table 3. Performance comparison of NMI from K-means, PCA+Km, LDA-Km, SC and mNMF on 10 data sets.

Data set	K-means	PCA+Km	LDA-Km	SC	mNMF
Dermatology	**86.18%**	**86.18%**	85.51%	84.27%	**86.18%**
Ecoli	49.27%	55.53%	**60.50%**	35.59%	50.96%
Ionosphere	12.17%	12.17%	11.90%	1.71%	**21.20%**
COIL-20	77.46%	77.14%	74.85%	**89.28%**	88.91%
Usps	62.70%	62.78%	64.91%	75.89%	**76.45%**
Binalpha	58.52%	59.74%	59.51%	58.35%	**60.86%**
Umist	66.08%	66.08%	65.03%	77.80%	**79.69%**
AR	61.50%	63.18%	58.11%	70.20%	**71.40%**
YaleB	16.09%	17.22%	18.74%	56.38%	**65.67%**
PIE	38.78%	39.02%	39.55%	57.54%	**63.58%**

(1) In general, the performance of PCA+Km is slightly better than K-means in most cases and outperforms LDA-Km in some cases, indicating that the introduction of dimensionality reduction is effective in improving the clustering effect. But at the same time it is also worse than SC and mNMF in other cases, this shows that, compared with clustering methods based solely on division, algorithms that consider the data manifold structure can have better performance.

(2) Specifically, SC is inferior to K-means, PCA+Km, and LDA-Km in some cases, but significantly better than K-means, PCA+Km, and LDA-Km in other cases, especially when the data dimensions are very high. Such observation means that in high-dimensional clustering, manifold-based data analysis methods can more easily mine potential correlations and data structures in the data than other algorithms because these methods can consider local correlations.

(3) Our method mNMF outperforms K-means, PCA+Km, LDA-Km and SC in most cases, which is consistent with the previous theoretical analysis and observations. The reason for such results is that our method simultaneously captures the subspace and manifold structure of the data, can effectively resist the redundancy and noise of dimensions in high-dimensional space, and can also consider the local relationships between data for manifold analysis.

4.2. *Experimental results on subspace learning*

Our method optimizes the subspace and clustering simultaneously, and thus is not only a clustering method but also an unsupervised subspace learning method. We compare the subspace learning performance of our algorithm with two classical unsupervised subspace learning methods, PCA and LPP,[60] on face recognition problem. Four face data sets Umist, AR, YaleB and PIE are used in this experiment. The nearest neighbor classifier is used for classification after dimension reduction. Five images per class are randomly chosen as the training data set and remaining images are used as the test data set. We also report the mean recognition accuracy and standard deviation over 20 random splits in Table 4 and Fig. 3. Figure 4 plots

Table 4. Recognition performance (Mean Recognition Accuracy ± Standard Deviation %) of PCA, LPP and mNMF over 20 random splits on four face data sets.

Data set	PCA	LPP	mNMF
Umist	**82.84 ± 2.20%**	75.66 ± 2.20%	**82.84 ± 2.20%**
AR	49.71 ± 2.25%	51.77 ± 2.25%	**67.00 ± 2.14%**
YaleB	46.89 ± 1.46%	54.90 ± 1.46%	**69.78 ± 2.23%**
PIE	58.49 ± 1.15%	78.31 ± 1.15%	**88.25 ± 1.07%**

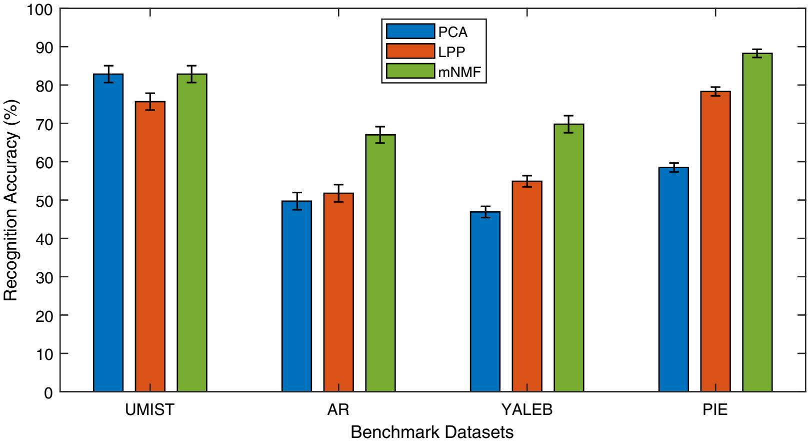
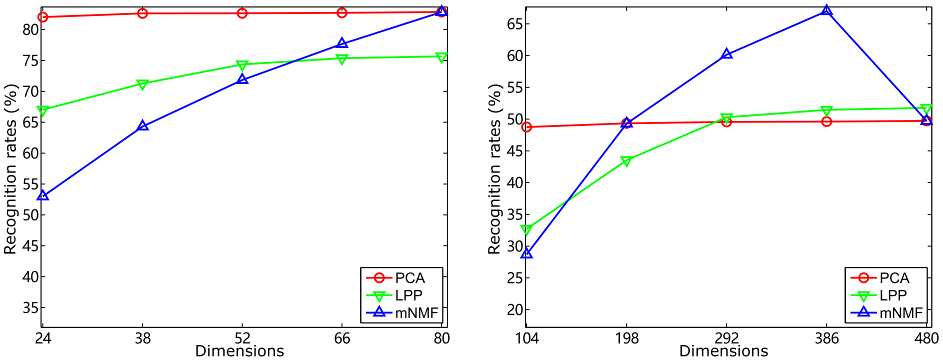


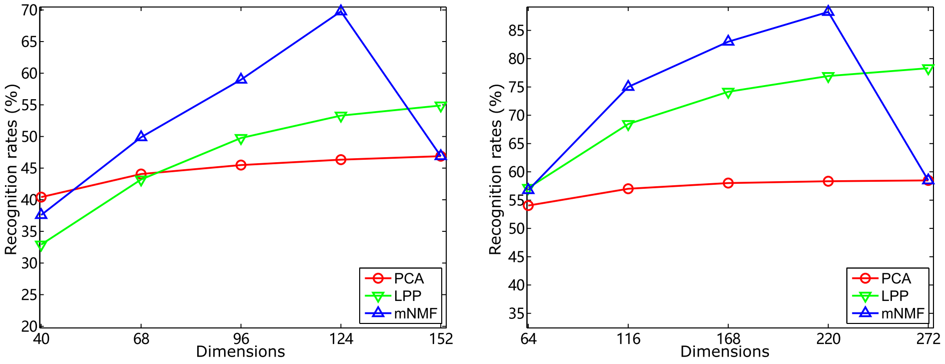
Fig. 3. Visualization of recognition performance of PCA, LPP and mNMF.



(a) Umist

(b) AR

Fig. 4. Recognition rates (%) with different feature dimensions on the Umist, AR, YaleB and PIE data sets.



(c) YaleB

(d) PIE

Fig. 4. (*Continued*)

the recognition accuracy with respect to the number of features. We have the following observations:

(1) For PCA and LPP, there is no consistent winner across all four face data sets. PCA outperforms LPP on the Umist data set, but is not as good as LPP on the AR, YaleB and PIE data sets. This shows that it is difficult for us to define which one is better between LPP and PCA, because they are both algorithms that perform dimensionality reduction based on the similarity between data.

(2) Our method mNMF significantly outperforms PCA and LPP on AR, YaleB and PIE data sets, which verifies that mNMF is also an effective unsupervised subspace learning method. At the same time, the method proposed in this paper is not only a subspace learning method, but also has pseudo-labels as information providers to participate in guiding the dimensionality reduction process, so more accurate recognition accuracy can be obtained after dimensionality reduction.

5. Conclusions

In this paper, we proposed a novel NMF method using manifold regularization to simultaneously learn subspace and do clustering. Different from traditional NMF, we constraint G to be a class indicator matrix, which is a special nonnegative matrix. One advantage of our method is that our clustering results can directly assign cluster labels to data points. Previous NMF-based clustering methods require a post-processing step to extract cluster structure from G and the clustering results may not be unique. We also incorporate a manifold regularization term in our objective to preserve the geometry structure. An efficient algorithm is derived to optimize our objective with quick convergence. Extensive experiments demonstrate that our proposed method gets better clustering results than previous clustering methods and finds good projection subspace for data dimensionality reduction.

ORCID

Feiping Nie https://orcid.org/0000-0002-0871-6519
Huimin Chen https://orcid.org/0000-0002-7968-4948
Heng Huang https://orcid.org/0000-0002-3483-8333
Chris H. Q. Ding https://orcid.org/0009-0009-3374-1941

References

1. Laith Mohammad Qasim Abualigah *et al.*, *Feature Selection and Enhanced Krill Herd Algorithm for Text Document Clustering* (Springer, 2019).
2. Guoqing Chao, Shiliang Sun and Jinbo Bi, A survey on multiview clustering, *IEEE Trans. Artif. Intell.* **2**, 146–168 (2021).
3. Wang Zeng, Sheng Jin, Wentao Liu, Chen Qian, Ping Luo, Wanli Ouyang and Xiaogang Wang, Not All Tokens Are Equal: Human-centric Visual Analysis Via Token Clustering Transformer, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2022).
4. Zineb Dafir, Yasmine Lamari and Said Chah Slaoui, A survey on parallel clustering algorithms for big data, *Artif. Intell. Rev.* **54**, 2411–2443 (2021).
5. Absalom E. Ezugwu, Amit K. Shukla, Moyinoluwa B. Agbaje, Olaide N. Oyelade, Adán José-García and Jeffery O. Agushaka, Automatic clustering algorithms: A systematic review and bibliometric analysis of relevant literature, *Neural Comput. Appl.* **33**, 6247–6306 (2021).
6. Asma Belhadi, Youcef Djenouri, Kjetil Nørvåg, Heri Ramampiaro, Florent Masseglia and Jerry Chun-Wei Lin, Space-time series clustering: Algorithms, taxonomy, and case study on urban smart cities, *Eng. Appl. Artif. Intell.* **95**, 103857 (2020).
7. Shuai Zhao, Linchao Zhu, Xiaohan Wang and Yi Yang, Centerclip: Token Clustering for Efficient Text-video Retrieval, in *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval* (2022).
8. Stuart P. Lloyd, Least squares quantization in PCM, *Bell Telephone Laboratories Paper* (Marray Hill, 1957).
9. Pak K. Chan, Martine Schlag and Jason Y. Zien, Spectral K-way ratio-cut partitioning and clustering, *IEEE Trans. Comput.-Aid Des. Integr. Circuits Syst.* **13**, 1088–1096 (1994).
10. Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu *et al.*, A density-based algorithm for discovering clusters in large spatial databases with noise, in *KDD'96: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining* (Portland, Oregon, USA, 1996), pp. 226–231.
11. J. MacQueen, Some methods for classification and analysis of multivariate observations, in *Proc. 5th Berkeley Symposium* (Statistics and Probability, University of California Press, Berkeley, CA, 1967), pp. 281–297.
12. Abiodun M. Ikotun, Absalom E. Ezugwu, Laith Abualigah, Belal Abuhaija and Jia Heming, K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data, *Inf. Sci.* **622**, 178–210 (2023).
13. Shuyin Xia, Daowan Peng, Deyu Meng, Changqing Zhang, Guoyin Wang, Elisabeth Giem, Wei Wei and Zizhong Chen, Ball k-means: Fast adaptive clustering with no bounds, *IEEE Trans. Pattern Anal. Mach. Intell.* **44**, 87–99 (2022).
14. Olivier Bachem, Mario Lucic, S. Hamed Hassani and Andreas Krause, Approximate K-means++ in sublinear time, *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 30 (Arizona, USA, 2016), pp.1457–1467.

15. Olivier Bachem, Mario Lucic, Hassani Hassani and Andreas Krause, Fast and Provably Good Seedings for k-means, in *Advances in Neural Information Processing Systems*, eds. D. Lee, M. Sugiyama, U. Luxburg, I. Guyon and R. Garnett (Curran Associates, 2016), pp. 55–63.
16. James Newling and François Fleuret, K-medoids for K-means Seeding, in *Advances in Neural Information Processing Systems*, eds. I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan and R. Garnett (Curran Associates, 2017), pp. 5195–5203.
17. Cheng-Hao Deng and Wan-Lei Zhao, Fast k-means Based on k-NN graph, in *2018 IEEE 34th International Conference on Data Engineering (ICDE)* (2018), pp. 1220–1223.
18. Qinghao Hu, Jiaxiang Wu, Lu Bai, Yifan Zhang and Jian Cheng, Fast K-means for Large Scale Clustering, in *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, CIKM '17* (Association for Computing Machinery, New York, NY, USA, 2017), pp. 2099–2102.
19. Yufei Ding, Yue Zhao, Xipeng Shen, Madanlal Musuvathi and Todd Mytkowicz, Yinyang K-means: A Drop-in Replacement of the Classic K-means with Consistent Speedup, in *Proceedings of the 32nd International Conference on Machine Learning*, eds. F. Bach and D. Blei (Lille, France, 2015), pp. 579–587.
20. 20. Petr Ryšavý and Greg Hamerly, Geometric Methods to Accelerate k-means Algorithms, in *Proceedings of the 2016 SIAM International Conference on Data Mining (SDM)* (Miami, Florida, USA, 2016), pp. 324–332.
21. Fan R. Chung, Spectral graph theory, *CBMS Regional Conference Series in Mathematics*, Vol. 92 (American Mathematical Society, 1997), p. 1–212.
22. F. R. K. Chung, A. Grigor'yan and S.-T. Yau, Higher eigenvalues and isoperimetric inequalities on Riemannian manifolds and graphs, *Comm. Anal. Geom.* **8**, 969 (2000).
23. Jianbo Shi and Jitendra Malik, *IEEE. Trans. Pattern Anal. Mach. Intell.* **22**, 888–905 (2000).
24. Chris H. Q. Ding, Xiaofeng He, Hongyuan Zha, Ming Gu and Horst D. Simon, A min-max cut algorithm for graph partitioning and data clustering, *Proceedings of the IEEE International Conference on Data Mining (ICDM)* (San Jose, California, USA, 2001), pp. 107–114.
25. Feiping Nie, Danyang Wu, Rong Wang and Xuelong Li, Truncated robust principle component analysis with a general optimization framework, *IEEE Trans. Pattern Anal. Mach. Intell.* **44**, 1081–1097 (2020).
26. Zheng Wang, Feiping Nie, Canyu Zhang, Rong Wang and Xuelong Li, Worst-case discriminative feature learning via max-min ratio analysis, *IEEE Trans. Pattern Anal. Mach. Intell.* **46**, 641–658 (2024).
27. Feiping Nie, Xia Dong, Zhanxuan Hu, Rong Wang and Xuelong Li, Discriminative projected clustering via unsupervised LDA, *IEEE Trans. Neural Networks Learn. Syst.* **34**, 9466–9480 (2023).
28. Xiaoping Li, Yadi Wang and Rubén Ruiz, A survey on sparse learning models for feature selection, *IEEE Trans. Cybern.* **52**, 1642–1660 (2022).
29. Heng Tao Shen, Yonghua Zhu, Wei Zheng and Xiaofeng Zhu, Half-quadratic minimization for unsupervised feature selection on incomplete data, *IEEE Trans. Neural Netw. Learn. Syst.* **32**, 3122–3135 (2021).
30. I. T. Jolliffe, *Principal Component Analysis*, 2nd edn., Springer Series in Statistics (Springer-Verlag, New York, 2002).
31. Trevor F. Cox and Michael A. A. Cox, Multidimensional scaling (Chapman and Hall, 2001).

32. J. B. Tenenbaum, V. de Silva and J. C. Langford, A global geometric framework for nonlinear dimensionality reduction, *Science* **290**, 2319–2323 (2000).
33. Sam T. Roweis and Lawrence K. Saul, Nonlinear dimensionality reduction by locally linear embedding, *Science* **290**, 2323–2326 (2000).
34. Xiaofei He and Partha Niyogi, Locality preserving projections, *NIPS* (Vancouver, British Columbia, Canada, 2004), pp. 153–160.
35. Mikhail Belkin and Partha Niyogi, Laplacian eigenmaps and spectral techniques for embedding and clustering, *NIPS* (Vancouver, Canada, 2001), pp. 585–591.
36. Rong Wang, Jintang Bian, Feiping Nie and Xuelong Li, Unsupervised discriminative projection for feature selection, *IEEE Trans. Knowl. Data Eng.* **34**, 942–953 (2022).
37. Peng Zhou, Jiangyong Chen, Mingyu Fan, Liang Du, Yi-Dong Shen and Xuejun Li, Unsupervised feature selection for balanced clustering, *Knowl.-Based Syst.* **193**, 105417 (2020).
38. Peng Zhou, Jiangyong Chen, Liang Du and Xuejun Li, Balanced spectral feature selection, *IEEE Trans. Cybern.* **53**, 4232–4244 (2023).
39. Feiping Nie, Xia Dong, Lai Tian, Rong Wang and Xuelong Li, Unsupervised feature selection with constrained $l_{2,0}$-norm and optimized graph, *IEEE Trans. Neural Netw. Learn. Syst.* **33**, 1702–1713 (2022).
40. Zhengxin Li, Feiping Nie, Danyang Wu, Zhanxuan Hu and Xuelong Li, Unsupervised feature selection with weighted and projected adaptive neighbors, *IEEE Trans. Cybern.* **53**, 1260–1271 (2023).
41. Rui Zhang, Yunxing Zhang and Xuelong Li, Unsupervised feature selection via adaptive graph learning and constraint, *IEEE Trans. Neural Netw. Learn. Syst.* **33**, 1355–1362 (2022).
42. Daniel D. Lee and H. Sebastian Seung, Learning the parts of objects by non-negative matrix factorization, *Nature* **401**, 788–791 (1999).
43. Pentti Paatero and Unto Tapper, Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values, *Environmetrics* **5**, 111 (1994).
44. Chris Ding, Xiaofeng He and Horst D. Simon, On the equivalence of nonnegative matrix factorization and spectral clustering, *Proceedings of the 2005 SIAM International Conference on Data Mining (SDM)* (2005), pp. 606–610.
45. Chris Ding, Tao Li and Michael I. Jordan, Convex and semi-nonnegative matrix factorizations, *IEEE Trans. Pattern Anal. Mach. Intell.* **32**, 45–55 (2009).
46. Chris Ding, Tao Li, Wei Peng and Haesun Park, Orthogonal nonnegative matrix tri-factorizations for clustering, *Proceedings of the International Conference on Knowledge Discovery and Data Mining (KDD 2006)* (2006), pp. 126–135.
47. Chris Ding, Tao Li and Wei Peng, Nonnegative matrix factorization and probabilistic latent semantic indexing: Equivalence, chi-square statistic, and a hybrid method, *Proceedings of the National Conference on Artificial Intelligence* (2006), pp. 137–143.
48. Deng Cai, Xiaofei He, Xiaoyun Wu and Jiawei Han, Non-negative matrix factorization on manifold, in *Proceedings of the 25th International Conference on Machine Learning (ICML)* (Helsinki, Finland, 2008), pp. 63–72.
49. Linlin Zong, Xianchao Zhang, Long Zhao, Hong Yu and Qianli Zhao, Multi-view clustering via multi-manifold regularized non-negative matrix factorization, *Neural Netw.* **88**, 74–89 (2017).
50. Shuai Wang, Tsung-Hui Chang, Ying Cui and Jong-Shi Pang, Clustering by orthogonal NMF model and non-convex penalty optimization, *IEEE Trans. Signal Process.* **69**, 5273–5288 (2021).

51. Dexian Wang, Tianrui Li, Wei Huang, Zhipeng Luo, Ping Deng, Pengfei Zhang and Minbo Ma, A multi-view clustering algorithm based on deep semi-NMF, *Inform. Fusion* **99**, 101884 (2023).
52. Tao Li, Sheng Ma and Mitsunori Ogihara, Document clustering via adaptive subspace iteration, in *Proceedings of the 45th Annual International ACM SIGIR Conference (SIGIR)* (Sheffield, UK, 2004), pp. 218–225.
53. Fernando De la Torre and Takeo Kanade, Discriminative cluster analysis, in *Proceedings of the 23th International Conference on Machine Learning (ICML)* (Pennsylvania, USA, 2006), pp. 241–248.
54. Chris H. Q. Ding and Tao Li, Adaptive dimension reduction using discriminant analysis and K-means clustering, in *Proceedings of the 24th International Conference on Machine Learning (ICML)* (Oregon, USA, 2007), pp. 218–225.
55. Jieping Ye, Zheng Zhao and Huan Liu, Adaptive distance metric learning for clustering, in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)* (Minnesota, USA, 2007), pp. 521–528.
56. Stella X. Yu and Jianbo Shi, Multiclass spectral clustering, in *Proceedings of the 9th International Conference on Computer Vision (ICCV)* (Nice, France, 2003), pp. 313–319.
57. Andrew Y. Ng, Michael I. Jordan and Yair Weiss, On spectral clustering: Analysis and an algorithm, in *Advances in Neural Information Processing Systems (NIPS)* (Vancouver, Canada, 2001), pp. 849–856.
58. Lihi Zelnik-Manor and Pietro Perona, Self-tuning spectral clustering, in *Advances in Neural Information Processing Systems (NIPS)* (Vancouver, Canada, 2004), pp. 1601–1608.
59. Deng Cai, Xiaofei He and Jiawei Han, Document clustering using locality preserving indexing, *IEEE Trans. Knowl. Data Eng.* **17**, 1624–1637 (2005).
60. Xiaofei He, Shuicheng Yan, Yuxiao Hu, Partha Niyogi and Hong-Jiang Zhang, Face recognition using laplacianfaces, *IEEE Trans. Pattern Anal. Mach. Intell.* **27**, 328–340 (2005).