Discrimination-Aware Projected Matrix Factorization

Xuelong Li, *Fellow, IEEE,* Mulin Chen, and Qi Wang, *Senior Member, IEEE*

Abstract—*Non-negative Matrix Factorization* (NMF) has been one of the most popular clustering techniques in machine leaning, and involves various real-world applications. Most existing works perform matrix factorization on high-dimensional data directly. However, the intrinsic data structure is always hidden within the low-dimensional subspace. And the redundant features within the input space may affect the final result adversely. In this paper, a new unsupervised matrix factorization method, *Discrimination-aware Projected Matrix Factorization* (DPMF), is proposed for data clustering. The main contributions are threefold: (1) the linear discriminant analysis is jointly incorporated into the unsupervised matrix factorization framework, so the clustering can be accomplished in the discriminant subspace; (2) the manifold regularization is introduced to perceive the geometric information, and the $\ell_{2,1}$-norm is utilized to improve the robustness; (3) an efficient optimization algorithm is designed to solve the proposed problem with proved convergence. Experimental results on one toy dataset and eight real-world benchmarks show the effectiveness of the proposed method.

Index Terms—Clustering, Linear Discriminant Analysis, Non-negative Matrix Factorization, Subspace Learning

1 Introduction

Non-negative Matrix Factorization (NMF) [1] is a widely-used method for data clustering, and has attracted many researchers in the field of machine learning and data mining. Given the data matrix, NMF approximates it with the product of two non-negative factor matrices. Ding et al. [2] have pointed out that the two factor matrices correspond to the cluster centroid and indicator respectively, thus NMF can obtain the clustering result directly with the cluster indicator, and does not need the post-processing (e.g. k-means). In addition, NMF is able to learn a parts-based representation since its updating rules only allow additive operation [3]. Therefore, it shows good performance in face recognition [4] and document clustering [5], where the objects are parts-based.

Despite its good property on data interpretation, the original NMF has three major disadvantages. First, NMF performs matrix factorization in the input data space directly. However, the high-dimensional data is often lying within a low-dimensional subspace [3], which contains more valuable information. Thus, NMF is limited to capture the discriminant features. Second, because NMF neglects the local relationship of data points, it fails to discover the geometric structure of the data distribution. And this drawback makes NMF unable to handle the data with complicated structures. Third, NMF squares the residue error of each data point with a ℓ_2-norm objective function, so it is easily affected by the outliers.

In the past two decades, many variants of NMF have been put forward to tackle the last two problems. For the second problem, Cai et al. [3] integrated the graph regularization term into the original NMF to exploit the local data structure. Guan et al. [6] proposed an optimal gradient method to speed up the optimization of NMF and regularized-NMF. Zeng et al. [7] introduced the hypergraph to encode the high-order relationship between data points. Zhang et al. [8] and Gao et al. [9] proposed to perform matrix factorization and graph learning simultaneously. To address the third problem, Huang et al. [10] and Zhang et al. [11] factorized the data matrix with a $\ell_{2,1}$-norm objective function and achieved relatively better performance. Zhou et al. [12] developed the divide-and-conquer framework, which shows good performance in handling high-dimensional noisy data. However, the first problem is still not well solved. Zhang et al. [11] proposed to learn the low-dimensional representation of input data, but they neglect the discriminant information. Zafeiriou et al. [13] and Lu et al. [14] combined the Linear Discriminant Analysis (LDA) [15] into matrix factorization for data classification, but these methods are either supervised or semi-supervised, so they cannot be used for data clustering. Moreover, as pointed out by Tao et al. [16], LDA imposes the Gaussian distribution on the input data, which means that it also fails to discover the local data structure. Note that, the first problem may compound the second one, since the data graph constructed with the input data may be unreliable. Therefore, it is important to capture the discriminant features resided within the desired subspace.

In this paper, we propose the Discrimination-aware Projected Matrix Factorization (DPMF) method, which inherits the merits of both LDA and NMF. The major contributions of this research are summarized as follows:

1. The proposed method learns the discriminant subspace with LDA, and performs clustering in the learned subspace. So the redundant features in the input space are avoided.

2. The local data relationship in the desired subspace is captured by the manifold regularization term, and the

• *The authors are with the School of Computer Science and the Center for OPTical IMagery Analysis and Learning (OPTIMAL), Northwestern Polytechnical University, Xi'an 710072, China. Email: xuelong_li@nwpu.edu.cn;chenmulin001@gmail.com;crabwq@gmail.com. Q. Wang is the corresponding author.*

robustness to outliers is improved with the $\ell_{2,1}$-norm.

3. An efficient and effective alternative algorithm is proposed to optimize the proposed problem, and its convergence is proved experimentally.

2 Review of NMF and LDA

2.1 Non-negative Matrix Factorization

Given the data matrix $\mathbf{X} = [x_1, x_2, \cdots, x_n], x_j \in \mathbb{R}^{d\times 1}$ (d and n are the dimensionality and sample number respectively), NMF approximates it with the product of two non-negative matrices:

$$\min_{\mathbf{F}\geq 0, \mathbf{G}\geq 0} ||\mathbf{X} - \mathbf{F}\mathbf{G}^T||_F^2, \tag{1}$$

where $\mathbf{F} \in \mathbb{R}^{d\times c}$ is the cluster centroid, and $\mathbf{G} \in \mathbb{R}^{n\times c}$ is the cluster indicator matrix. However, for any scalar $\delta > 0$, the product of $\delta\mathbf{F}$ and $\mathbf{G}^T/\delta$ will give the same residue error. So the non-negative constraint cannot guarantee the uniqueness of solution.

To circumvent this problem, Huang et al. [10] proposed to impose the orthogonal constraint on $\mathbf{G}$, and then problem (1) is reformulated as

$$\min_{\mathbf{G}\geq 0, \mathbf{G}^T\mathbf{G}=\mathbf{I}_c} ||\mathbf{X} - \mathbf{F}\mathbf{G}^T||_F^2, \tag{2}$$

where $\mathbf{I}_c \in \mathbb{R}^{c\times c}$ is the identity matrix. Then the optimal solution is unique.

2.2 Linear Discriminant Analysis

LDA aims to learn a linear transformation $\mathbf{W} \in \mathbb{R}^{d\times m}$ to project the d-dimensional data into the m-dimensional representation. Given the binary label matrix $\mathbf{Y} \in \mathbb{R}^{n\times c}$, the total-class scatter $\mathbf{S}_t$, between-class scatter $\mathbf{S}_b$ and within-class scatter $\mathbf{S}_w$ are defined as follows [17]:

$$\begin{aligned}
\mathbf{S}_t &= \sum_{i}^{n}(x_i - \mu)(x_i - \mu)^T = \mathbf{X}\mathbf{H}\mathbf{X}^T,\\
\mathbf{S}_b &= \sum_{i=1}^{c} n_i(\mu_i - \mu)(\mu_i - \mu)^T = \mathbf{X}\mathbf{H}\tilde{\mathbf{G}}\tilde{\mathbf{G}}^T\mathbf{H}\mathbf{X}^T,\\
\mathbf{S}_w &= \mathbf{S}_t - \mathbf{S}_b,
\end{aligned} \tag{3}$$

where μ is the mean of all data points, μ_i is the mean of the points in the i-th class, n_i is the number of points in the i-th class. $\tilde{\mathbf{G}} = \mathbf{Y}(\mathbf{Y}^T\mathbf{Y})^{-1/2}$ is the scaled label matrix, and $\mathbf{H} \in \mathbb{R}^{n\times n}$ is $\mathbf{I}_n - \frac{1}{n}\mathbf{1_n}\mathbf{1_n}^T$ ($\mathbf{I}_n$ is the n-dimensional identity matrix and $\mathbf{1_n} \in \mathbb{R}^{n\times 1}$ is a column vector with all its elements as 1).

The objective of LDA is to find the optimal $\mathbf{W}$ to push the points from different classes far away while pulling those from the same class together. Thus, the objective function of LDA can be written as

$$\min_{\mathbf{W}^T\mathbf{S}_t\mathbf{W}=\mathbf{I}_m} \mathrm{Tr}(\mathbf{W}^T\mathbf{S}_w\mathbf{W}), \tag{4}$$

where $\mathbf{I}_m \in \mathbb{R}^{m\times m}$ is the identity matrix, and $\mathrm{Tr}()$ is the trace operator. According to the definitions in Eq. (3), problem (4) can be further transformed into

$$\begin{aligned}
&\min_{\mathbf{W}^T\mathbf{S}_t\mathbf{W}=\mathbf{I}_m} \mathrm{Tr}(\mathbf{W}^T(\mathbf{S}_t - \mathbf{S}_b)\mathbf{W})\\
=&\min_{\mathbf{W}^T\mathbf{S}_t\mathbf{W}=\mathbf{I}_m} \mathrm{Tr}(\mathbf{W}^T\mathbf{X}\mathbf{H}(\mathbf{I}_c - \tilde{\mathbf{G}}\tilde{\mathbf{G}}^T)\mathbf{H}\mathbf{X}^T\mathbf{W})\\
=&\min_{\mathbf{W}^T\mathbf{S}_t\mathbf{W}=\mathbf{I}_m} ||\mathbf{W}^T\mathbf{X}\mathbf{H}(\mathbf{I}_c - \tilde{\mathbf{G}}\tilde{\mathbf{G}}^T)||_F^2.
\end{aligned} \tag{5}$$

3 Discrimination-aware Projected Matrix Factorization

3.1 Methodology

In real-world applications, data with high-dimensionality is often lying within a low-dimensional subspace. To find the discriminant subspace, we propose to integrate LDA and NMF into a unified framework.

In problem (2), when $\mathbf{G}$ is fixed, the optimal $\mathbf{F}$ is computed as $\mathbf{X}\mathbf{G}$. Then the objective value becomes

$$||\mathbf{X} - \mathbf{X}\mathbf{G}\mathbf{G}^T||_F^2 = ||\mathbf{X}(\mathbf{I}_c - \mathbf{G}\mathbf{G}^T)||_F^2, \tag{6}$$

which is equivalent to the objective value of problem (5) if $\mathbf{X}$ and $\mathbf{G}$ are replaced with $\mathbf{W}^T\mathbf{X}\mathbf{H}$ and $\tilde{\mathbf{G}}$ respectively. Furthermore, since $\mathbf{G}$ is considered to be the unique cluster indicator, the optimal $\mathbf{G}$ for problem (2) is equal to the scaled label matrix $\tilde{\mathbf{G}}$ in problem (5). Therefore, LDA can be naturally incorporated in the unsupervised matrix factorization framework as

$$\begin{aligned}
&\min_{\mathbf{F},\mathbf{G},\mathbf{W}} ||\mathbf{W}^T\mathbf{X}\mathbf{H} - \mathbf{F}\mathbf{G}^T||_F^2,\\
&s.t. \mathbf{F} \in \mathbb{R}^{m\times n}, \mathbf{G} \in \mathbb{R}^{n\times c}, \mathbf{W} \in \mathbb{R}^{d\times m},\\
&\mathbf{G} \geq 0, \mathbf{G}^T\mathbf{G} = \mathbf{I}_c, \mathbf{W}^T\mathbf{S}_t\mathbf{W} = \mathbf{I}_m,
\end{aligned} \tag{7}$$

where $\mathbf{G}$ can be also regarded as the pseudo label. And then the discriminant subspace for matrix factorization can be found with the optimal $\mathbf{W}$.

In the perspective of manifold learning, the points with close relationship should be grouped into the same cluster. When $\mathbf{W}$ is fixed, we build a data graph $\mathbf{S} \in \mathbb{R}^{n\times n}$ with $\mathbf{W}^T X$, and the geometry structure can be captured by minimizing the following problem

$$\min_{\mathbf{G}\in\mathbb{R}^{n\times c}} \mathrm{Tr}(\mathbf{G}^T\mathbf{L}\mathbf{G}), \tag{8}$$

where $\mathbf{L}$ is the Laplacian matrix of $\mathbf{S}$. According to the feature selection theory [17], the value of $||\mathbf{W}_{i,:}||_2^2$ indicates the significance of the i-th dimension. Then we assume that $||\mathbf{W}_{i,:}||_2^2$ should shrink to zero if the corresponding dimension is incorrelate to the label vector $\mathbf{G}_{i,:}$, leading to the following problem

$$\min_{\mathbf{W}\in\mathbb{R}^{d\times c}} ||\mathbf{W}^T\mathbf{X} - \mathbf{G}^T||_F^2. \tag{9}$$

Finally, by combining the problem (7), (8) and (9) together, and introducing the $\ell_{2,1}$-norm to improve the robustness, we have the objective function of the proposed Discrimination-aware Projected Matrix Factorization (DPMF) method:

$$\begin{aligned}
&\min_{\mathbf{F},\mathbf{G},\mathbf{W}} ||\mathbf{W}^T\mathbf{X}\mathbf{H} - \mathbf{F}\mathbf{G}^T||_{2,1} + \lambda\mathrm{Tr}(\mathbf{G}^T\mathbf{L}\mathbf{G}) +\\
&\qquad \beta||\mathbf{W}^T\mathbf{X} - \mathbf{G}^T||_F^2.\\
&\quad s.t. \mathbf{F} \in \mathbb{R}^{c\times n}, \mathbf{G} \in \mathbb{R}^{n\times c}, \mathbf{W} \in \mathbb{R}^{d\times c},\\
&\qquad \mathbf{G} \geq 0, \mathbf{G}^T\mathbf{G} = \mathbf{I}_c, \mathbf{W}^T\mathbf{S}_t\mathbf{W} = \mathbf{I}_c,
\end{aligned} \tag{10}$$

where λ and β are parameters. When $\mathbf{W}$ is obtained, we update $\mathbf{G}$ and reconstruct the Laplacian graph $\mathbf{L}$ with $\mathbf{W}^T\mathbf{X}$. Then both the manifold learning and matrix factorization can be approached in the desired subspace.

3.2 Optimization Algorithm

Problem (10) is difficult to solve with the constraint $\mathbf{W}^T\mathbf{S}_t\mathbf{W} = \mathbf{I}_c$. So we first disturb the diagonal elements of $\mathbf{S}_t$ with a small enough scalar $\epsilon > 0$. Then $\mathbf{S}_t$ is positive definite, and we can decompose it with the Cholesky decomposition $\mathbf{S}_t = \mathbf{R}^T\mathbf{R}$. Thus, denoting $\mathbf{RW}$ as $\mathbf{P}$, and denoting $(\mathbf{R}^{-1})^T\mathbf{XH}$ and $(\mathbf{R}^{-1})^T\mathbf{X}$ as $\mathbf{A}$ and $\mathbf{B}$ respectively, problem (10) is simplified into

$$\begin{aligned}\min_{\mathbf{F},\mathbf{G},\mathbf{P}} &||\mathbf{P}^T\mathbf{A} - \mathbf{F}\mathbf{G}^T||_{2,1} + \lambda\mathrm{Tr}(\mathbf{G}^T\mathbf{L}\mathbf{G}) + \\ &\beta||\mathbf{P}^T\mathbf{B} - \mathbf{G}^T||_F^2, \\ s.t. &\mathbf{F} \in \mathbb{R}^{c\times n}, \mathbf{G} \in \mathbb{R}^{n\times c}, \mathbf{P} \in \mathbb{R}^{d\times c}, \\ &\mathbf{G} \geq 0, \mathbf{G}^T\mathbf{G} = \mathbf{I}_c, \mathbf{P}^T\mathbf{P} = \mathbf{I}_c.\end{aligned} \tag{11}$$

The above problem is not convex with three variables, so we propose to solve it with the Augmented Lagrangian Multiplier (ALM) [18]. Since both the $\ell_{2,1}$-norm and the manifold regularization depend on $\mathbf{G}$, we introduce two auxiliary variables $\mathbf{E} = \mathbf{P}^T\mathbf{A} - \mathbf{F}\mathbf{G}^T$ and $\mathbf{Z} = \mathbf{G}$. Then problem (11) is equivalent to the following ALM problem

$$\begin{aligned}\min_{\mathbf{E},\mathbf{Z},\mathbf{F},\mathbf{G},\mathbf{P}} &||\mathbf{E}||_{2,1} + \lambda\mathrm{Tr}(\mathbf{Z}^T\mathbf{L}\mathbf{G}) + \beta||\mathbf{P}^T\mathbf{B} - \mathbf{G}^T||_F^2 + \\ &\tfrac{\mu}{2}||\mathbf{P}^T\mathbf{A} - \mathbf{F}\mathbf{G}^T - \mathbf{E} + \tfrac{\mathbf{\Lambda}_1}{\mu}||_F^2 + \\ &\tfrac{\mu}{2}||\mathbf{G} - \mathbf{Z} + \tfrac{\mathbf{\Lambda}_2}{\mu}||_F^2, \\ s.t. &\mathbf{E} \in \mathbb{R}^{c\times n}, \mathbf{Z} \in \mathbb{R}^{n\times c}, \mathbf{F} \in \mathbb{R}^{c\times n}, \mathbf{G} \in \mathbb{R}^{n\times c}, \\ &\mathbf{P} \in \mathbb{R}^{d\times c}, \mathbf{Z} \geq 0, \mathbf{G}^T\mathbf{G} = \mathbf{I}_c, \mathbf{P}^T\mathbf{P} = \mathbf{I}_c,\end{aligned} \tag{12}$$

where $\mu \in \mathbb{R}^{1\times 1}$ is the ALM parameter, and $\mathbf{\Lambda}_1 \in \mathbb{R}^{c\times n}$ and $\mathbf{\Lambda}_2 \in \mathbb{R}^{n\times c}$ are ALM multipliers. Then we optimize each variable iteratively.

Update E: when fixing all the variables expect $\mathbf{E}$, we have

$$\min_{\mathbf{E}\in\mathbb{R}^{c\times n}} ||\mathbf{E}||_{2,1} + \frac{\mu}{2}||\mathbf{E} - \mathbf{M}||_F^2, \tag{13}$$

where $\mathbf{M} = \mathbf{P}^T\mathbf{A} - \mathbf{F}\mathbf{G}^T + \frac{\mathbf{\Lambda}_1}{\mu}$. According to Huang et al. [10], the optimal $\mathbf{E}$ is computed as

$$\mathbf{E}_{:,i} = \begin{cases} (1 - \frac{1}{\mu||\mathbf{M}_{:,i}||_2})\mathbf{M}_{:,i}, \text{if } ||\mathbf{M}_{:,i}||_2 \geq \frac{1}{\mu} \\ 0, \qquad\qquad \text{else} \end{cases}. \tag{14}$$

Update Z: when updating $\mathbf{Z}$, problem (12) becomes

$$\min_{\mathbf{Z}\geq 0, \mathbf{Z}\in\mathbb{R}^{n\times c}} \lambda\mathrm{Tr}(\mathbf{Z}^T\mathbf{L}\mathbf{G}) + \tfrac{\mu}{2}||\mathbf{G} - \mathbf{Z} + \tfrac{\mathbf{\Lambda}_2}{\mu}||_F^2, \tag{15}$$

which can be further reduced into a closed-form problem

$$\min_{\mathbf{Z}\geq 0} ||\mathbf{Z} - \mathbf{T}||_F^2, \tag{16}$$

where $\mathbf{T} = \mathbf{G} + \frac{\mathbf{\Lambda}_2}{2} - \frac{\lambda}{\mu}\mathbf{L}\mathbf{G}$. Therefore, the optimal $\mathbf{Z}$ is

$$\mathbf{Z}_{ij} = max(\mathbf{T}_{ij}, 0). \tag{17}$$

Update F: optimizing problem (12) w.r.t. $\mathbf{F}$ yields the following sub-problem:

$$\min_{\mathbf{F}\in\mathbb{R}^{c\times n}} ||\mathbf{P}^T\mathbf{A} - \mathbf{F}\mathbf{G}^T - E + \tfrac{\mathbf{\Lambda}_1}{\mu}||_F^2. \tag{18}$$

Because $\mathbf{G}^T\mathbf{G} = \mathbf{I}_c$, the above problem is reformulated as

$$\min_{\mathbf{F}\in\mathbb{R}^{c\times n}} ||\mathbf{F} - (\mathbf{P}^T\mathbf{A} - \mathbf{E} + \tfrac{\mathbf{\Lambda}_1}{\mu})\mathbf{G}||_F^2, \tag{19}$$

so the optimal $\mathbf{F}$ is $(\mathbf{P}^T\mathbf{A} - \mathbf{E} + \frac{\mathbf{\Lambda}_1}{\mu})\mathbf{G}$.

Update G: to update $\mathbf{G}$, problem (12) is reduced to

$$\begin{aligned}\min_{\mathbf{G}} &\lambda\mathrm{Tr}(\mathbf{Z}^T\mathbf{L}\mathbf{G}) + \beta||\mathbf{P}^T\mathbf{B} - \mathbf{G}^T||_F^2 + \\ &\tfrac{\mu}{2}||\mathbf{P}^T\mathbf{A} - \mathbf{F}\mathbf{G}^T - \mathbf{E} + \tfrac{\Lambda_1}{\mu}||_F^2 + \\ &\tfrac{\mu}{2}||\mathbf{G} - \mathbf{Z} + \tfrac{\mathbf{\Lambda}_2}{\mu}||_F^2, \\ s.t. &\mathbf{G}^T\mathbf{G} = \mathbf{I}_c, \mathbf{G} \in \mathbb{R}^{n\times c}.\end{aligned} \tag{20}$$

By expanding the objective function and removing the irrelevant terms, we get

$$\begin{aligned}\min_{\mathbf{G}\in\mathbb{R}^{n\times c}} &||\mathbf{G} - \mathbf{K}||_F^2, \\ s.t. &\mathbf{G}^T\mathbf{G} = \mathbf{I}_c,\end{aligned} \tag{21}$$

where $\mathbf{K} = 2\frac{\beta}{\mu}\mathbf{B}^T\mathbf{P} - \frac{\lambda}{\mu}\mathbf{L}\mathbf{Z} + (\mathbf{P}^T\mathbf{A} - \mathbf{E} + \frac{\mathbf{\Lambda}_1}{\mu})^T\mathbf{F} + \mathbf{Z} - \frac{\mathbf{\Lambda}_2}{\mu}$. And Huang et al. [10] proved that the optimal solution of the above problem is

$$\mathbf{G} = \mathbf{U}\mathbf{V}^T, \tag{22}$$

where $\mathbf{U} \in \mathbb{R}^{n\times c}$ and $\mathbf{V} \in \mathbb{R}^{c\times c}$ are the left and right singular vectors of the compact singular value decomposition of $\mathbf{K}$.

Update P: when solving $\mathbf{P}$, problem (12) is transformed into

$$\begin{aligned}\min_{\mathbf{P}\in\mathbb{R}^{d\times c}} &\beta||\mathbf{P}^T\mathbf{B} - \mathbf{G}^T||_F^2 + \\ &\tfrac{\mu}{2}||\mathbf{P}^T\mathbf{A} - \mathbf{F}\mathbf{G}^T - \mathbf{E} + \tfrac{\mathbf{\Lambda}_1}{\mu}||_F^2, \\ s.t. &\mathbf{P}^T\mathbf{P} = \mathbf{I}_c,\end{aligned} \tag{23}$$

which can be further reduced to

$$\begin{aligned}\min_{\mathbf{P}\in\mathbb{R}^{d\times c}} &\mathrm{Tr}(\mathbf{P}^T\mathbf{J}\mathbf{P}) - \mathrm{Tr}(\mathbf{P}^T\mathbf{Q}), \\ s.t. &\mathbf{P}^T\mathbf{P} = \mathbf{I}_c,\end{aligned} \tag{24}$$

where

$$\begin{aligned}\mathbf{J} &= \beta\mathbf{B}\mathbf{B}^T + \tfrac{\mu}{2}\mathbf{A}\mathbf{A}^T, \\ \mathbf{Q} &= 2\beta\mathbf{B}\mathbf{G} + \mu\mathbf{A}(\mathbf{F}\mathbf{G}^T + \mathbf{E} - \tfrac{\mathbf{\Lambda}_1}{\mu})^T.\end{aligned} \tag{25}$$

Note that problem (24) is non-convex because $\mathbf{J}$ is not a positive semi-definite matrix. Denoting the largest eigenvalue of matrix $\mathbf{J}$ as λ_{max}, problem (24) can be converted into the following form

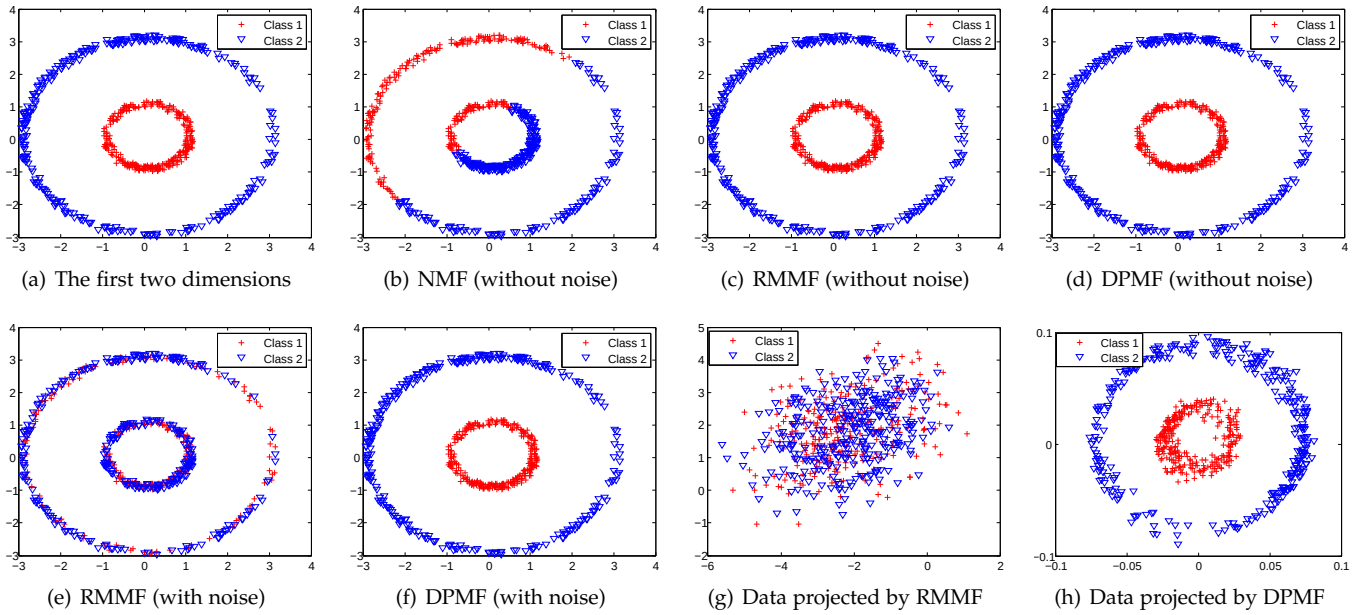
$$\begin{aligned}\max_{\mathbf{P}\in\mathbb{R}^{d\times c}} &\mathrm{Tr}[\mathbf{P}^T(\lambda_{max}\mathbf{I}_c - \mathbf{J})\mathbf{P}] + \mathrm{Tr}(\mathbf{P}^T\mathbf{Q}), \\ s.t. &\mathbf{P}^T\mathbf{P} = \mathbf{I}_c.\end{aligned} \tag{26}$$

Since $(\lambda_{max}\mathbf{I}_c - \mathbf{J})$ is positive semi-definite, problem (26) is a standard convex problem on the Stiefel manifold, and can be solved by the Generalized Power Iteration (GPI) method [19].

Update μ, $\mathbf{\Lambda}_1$ and $\mathbf{\Lambda}_2$: the ALM parameters are updated as follows:

$$\begin{aligned}\mathbf{\Lambda}_1 &= \mathbf{\Lambda}_1 + \mu(\mathbf{P}^T\mathbf{A} - \mathbf{F}\mathbf{G}^T - \mathbf{E}), \\ \mathbf{\Lambda}_2 &= \mathbf{\Lambda}_2 + \mu(\mathbf{G} - \mathbf{Z}), \\ \mu &= \rho\mu,\end{aligned} \tag{27}$$

where the parameter ρ controls the convergence speed.



(a) The first two dimensions (b) NMF (without noise) (c) RMMF (without noise) (d) DPMF (without noise)

(e) RMMF (with noise) (f) DPMF (with noise) (g) Data projected by RMMF (h) Data projected by DPMF

Fig. 1. Performance of NMF, RMMF and DPMF on the toy dataset.

4 Experiments

In this section, the effectiveness of the proposed DPMF is demonstrated on one toy dataset and eight real-world datasets. And the convergence behavior and the parameter sensitivity of DPMF are also discussed.

4.1 Performance on Toy Dataset

To evaluate whether DPMF is able to exploit the geometry structure and project the input data to the discriminant subspace, a toy dataset is constructed.

Dataset. As shown in Figure 1, the dataset consists of the data points from two classes. And the data dimensionality is six, where the data is distributed in concentric circles in the first two dimensions while the other four dimensions are noises randomly generated from 0 to 3.

Competitors. The proposed DPMF is compared with the original NMF [1], and Robust Manifold Matrix Factorization (RMMF) [11]. On the other hand, RMMF projects the data into a low-dimensional representation, and learns the local data manifold with graph regularization. For DPMF, the similarity graph is initialized with an efficient method [20] and the transformation matrix is initialized with the method in [21]. The graph of RMMF is built with the self-tune Gaussian method [22].

Performance. The first row in Figure 1 visualizes the clustering performance of different methods. It can be seen that the original NMF can not cluster the data into the correct class even without noise dimensions. This is because NMF neglects the local geometry structure. RMMF and the proposed DPMF work well on the noise data, which benefits from the exploitation of local data relationship.

The second row in Figure 1 (e) and (f) show the clustering results of RMMF and DPMF on the noise data. RMMF while DPMF still performs well. From Figure 1 (g) and (h), we can see that RMMF fails to project the data into the correct subspace, while the subspace learned by DPMF preserves the intrinsic structure of original data successfully, which indicates that DPMF is more capable of capturing the discriminant features. DPMF inherits the merit of LDA, so it has the capability to learn the valuable features and find the discriminant subspace. The noises dimensions do not exist in the subspace, so DPMF clusters the data correctly. This result indicates that the incorporation of LDA does improve the clustering performance of DPMF.

4.2 Performance on Real-World Datasets

In this part, eight real-world datasets are employed to verify the effectiveness of DPMF on data clustering. The widely-used clustering ACCuracy (ACC) and Normalized Mutual Information (NMI) [8] are taken as evaluation metrics.

Dataset. The real-world datasets used in the experiments include one object image dataset, i.e., COIL20 [23], three face image datasets, i.e., JAFFE [24], UMIST [25] and YALE [26], one biology dataset, i.e., SRBCT [27], and three datasets from the UCI Machine Learning Repository [28], i.e., Dermatology, Movement and Semeion.

Competitors. Six state-of-the-art clustering methods are taken for comparison, including k-means, Normalized Cut (NCut) [29], NMF [1], Graph-regularized NMF (GNMF) [3], Robust Manifold NMF (RMNMF) [10] and Robust Manifold Matrix Factorization (RMMF) [11].

Initialization and parameter setting. The similarity graph of each method is constructed by the approach suggested by the authors. Particularly, for NCut, the data graph is built with the self-tune Gaussian method [22]. For GNMF and RMNMF, the bipartite graphs are constructed by finding the five nearest neighbors. The initialization strategy of RMMF and DPMF are the same as in Section 4.1. Since the initial values of μ, Λ_1 and Λ_2 in DPMF have very little influence on the final results, we set them empirically.

In addition, the regularization parameters of GNMF and RMNMF are searched in the range of $\{10^{-3}, 10^{-2}, ..., 10^{0}\}$,

TABLE 1
ACC of different methods on real-world datasets. The best results are in bold face. And the second best results are underlined.

Datasets	k-means	NCut	NMF	GNMF	RMNMF	RMMF	DPMF
COIL20	0.6176	0.5328	0.4570	0.8049	0.5874	0.8693	**0.8853**
JAFFE	0.7178	0.7300	0.7075	0.8310	0.7563	0.8814	**0.9108**
UMIST	0.4123	0.4450	0.3376	0.5583	0.4197	0.5813	**0.6104**
YALE	0.4812	0.5158	0.3467	0.3818	0.3570	0.5327	**0.5576**
SRBCT	0.4494	0.3735	0.4301	0.4217	0.4193	0.5213	**0.5542**
Dermatology	0.6973	0.8876	0.7022	0.7978	0.7221	0.8901	**0.9536**
Movement	0.4403	0.4569	0.3566	0.4694	0.4063	0.5372	**0.5561**
Semeion	0.5976	0.5207	0.3772	0.5888	0.5498	0.5821	**0.6033**

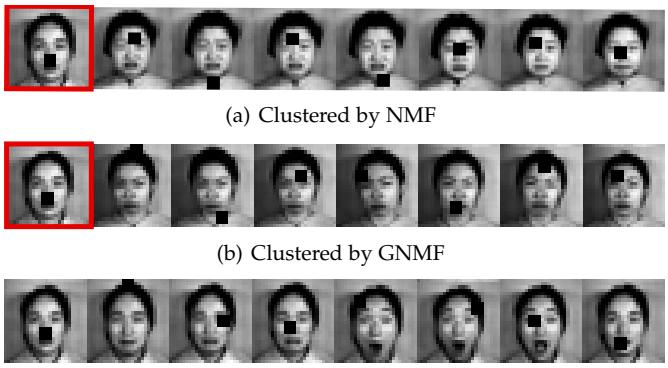
TABLE 2
NMI of different methods on real-world datasets. The best results are in bold face. And the second best results are underlined.

Datasets	k-means	NCut	NMF	GNMF	RMNMF	RMMF	DPMF
COIL20	0.7505	0.6799	0.5894	0.8787	0.7113	0.9247	**0.9351**
JAFFE	0.8160	0.8216	0.7343	0.9030	0.7926	0.8917	**0.9197**
UMIST	0.6253	0.6343	0.4866	0.7726	0.5877	0.7714	**0.7923**
YALE	0.5730	0.5704	0.4107	0.4549	0.4220	0.5841	**0.6032**
SRBCT	0.1571	0.1061	0.1822	0.1776	0.1559	0.2636	**0.3340**
Dermatology	0.7872	0.8488	0.6780	0.8365	0.7619	0.8624	**0.9118**
Movement	0.5751	0.5951	0.4200	0.6150	0.4967	0.6560	**0.6834**
Semeion	0.5310	0.4851	0.3199	0.5746	0.4688	0.5932	**0.6201**

and the parameters of RMMF and DPMF are searched from $\{10^{-8}, 10^{-6}, ..., 10^{2}\}$. For all the NMF methods except GNMF, the clustering result is obtained with the learned indicator matrix directly. And for GNMF, k-means is used for post-processing because its $\mathbf{G}$ does not have a clear cluster structure. And for those that are sensitive to the initialization, we run them with the best parameters for twenty times and report the average ACC and NMI.

Performance. Tabel 1 and 2 show the quantitative comparison of different methods. The proposed DPMF achieves the highest ACC and NMI on all datasets, which indicates the good performance. Particularly, SRBCT is a biology dataset with very high-dimensionality (2308). DPMF projects the data into a very low-dimensional (4) subspace, and shows promising performance. This phenomenon demonstrates that the intrinsic data structure is exactly lying within the low-dimensional subspace. NCut, GNMF and RMNMF show better performance than k-means and NMF, because they respect the data structure by exploiting the local data correlation. RMMF achieves the second best performance in most of the cases because it learns the low-dimensional representation of high-dimensional data. But it is still inferior to the proposed DPMF for two reasons: (1) RMMF neglects the discriminant information, while DPMF learns the valuable features by discriminant analysis; (2) RMMF builds the data graph with the input data directly, while DPMF updates the graph in each iteration and finally learns the data structure in the desired subspace.

In order to evaluate the robustness of DPMF, we further conduct experiments on the JAFFE dataset. To simulate the outliers, each face image in the dataset is randomly occluded with a 6×6 black area. Figure 2 shows some representative clustering results of NMF, GNMF and DPMF, and the red box indicates that the image is partitioned into the wrong cluster. We can see that NMF and GNMF cluster the first face image into the wrong group, while DPMF partitions the faces captured from the same person correctly. NMF and GNMF square the residue error of each sample, so they are prone to outliers. While DPMF is robust to outliers with the $\ell_{2,1}$-norm objective function.

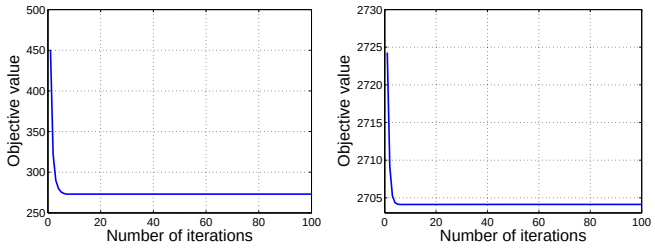


(a) Clustered by NMF

(b) Clustered by GNMF

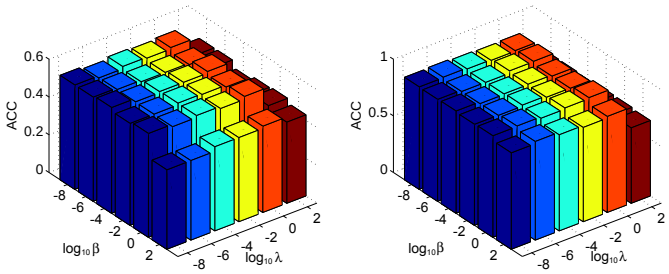
(c) Clustered by DPMF

Fig. 2. Clustering results of NMF, GNMF and DPMF on occluded face images. Red square boxes indicate the incorrect results. NMF and GNMF cluster the first image into the wrong class, while DPMF performs well.



(a) YALE (b) Dermatology

Fig. 3. Convergence curves of the proposed optimization method on (a) YALE and (b) Dermatology.



(a) YALE (b) Dermatology

Fig. 4. Classification accuracy of DPMF on (a) YALE and (b) Dermatology with varying β and λ. It can be seen that the performance is not sensitive to the parameters within a wide range.

4.3 Convergence Study and Parameter Sensitivity

The convergence behavior of DPMF is first discussed. The parameter μ increases exponentially during the iteration, so the problem (12) converges to the original problem (10) finally. And for problem (12), the optimization algorithm decomposes it into five sub-problems, among which the first four provide the closed form solutions, and the convergence of fifth one has already been proven by Nie et al. [19]. So the objective value decreases in the optimization of each sub-problem, and finally converges to a local optimal value. As shown in Figure 3, the optimization always converges within five iterations. Therefore, the proposed optimization algorithm is effective and efficient.

In addition, the parameter sensitivity of DPMF is also investigated. There are two important parameters (i.e., λ and β) in DPMF. λ controls the weight of the manifold regularization term, and β balances the transformation learning term. As the values of λ and β vary within the range of $\{10^{-8}, 10^{-6}, ..., 10^{2}\}$, the clustering accuracy of DPMF is visualized in Figure 4. It can be seen that DPMF shows stable performance across a wide range of λ and β. And the performance decreases when λ and β are 10^2, because the fitting error of the matrix factorization term becomes very large in this situation.

5 Conclusions

In this research, an unsupervised Discriminant-aware Projected Matrix Factorization (DPMF) method is presented for data clustering. In order to discover the intrinsic geometry structure, DPMF projects the data into the low-dimensional subspace, which contains more discriminant information. In addition, the manifold regularization is performed in the learned subspace to capture the local relationship between samples, and $\ell_{2,1}$-norm is used to improve the robustness to outliers. And the proposed objective function can be solved efficiently with the suggested optimization algorithm. Extensive experiments conducted on toy and real-world datasets show the satisfying performance of DPMF, and validate its superiority over the state-of-the-art competitors.

Acknowledgment

This work was supported by the National Natural Science Foundation of China under Grant 61871470, U1864204 and 61773316, and Project of Special Zone for National Defense Science and Technology Innovation.

References

[1] D. Lee and H. Seung, "Learning the parts of objects with nonnegative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, 1999.
[2] C. Ding, T. Li, and M. Jordan, "Convex and semi-nonnegative matrix factorizations," *PAMI*, vol. 32, no. 1, pp. 45–55, 2010.
[3] D. Cai, X. He, J. Han, and T. Huang, "Graph regularized nonnegative matrix factorization for data representation," *PAMI*, vol. 33, no. 8, pp. 1548–1560, 2011.
[4] Q. Wang, J. Wan, F. Nie, B. Liu, C. Yan, and X. Li, "Hierarchical feature selection for random projection," *TNNLS*, vol. 30, no. 5, pp. 1581–1586, 2019.
[5] Q. Wang, Z. Qin, F. Nie, and X. Li, "Spectral embedded adaptive neighbors clustering," *TNNLS*, vol. 30, no. 4, pp. 1265–1271, 2019.
[6] N. Guan, D. Tao, Z. Luo, and B. Yuan, "Nenmf: An optimal gradient method for nonnegative matrix factorization," *IEEE Transactions on Signal Processing*, vol. 60, no. 6, pp. 2882–2898, 2012.
[7] K. Zeng, J. Yu, C. Li, J. You, and T. Jin, "Image clustering by hypergraph regularized non-negative matrix factorization," *Neurocomputing*, vol. 138, pp. 209–217, 2014.
[8] L. Zhang, Q. Zhang, B. Du, J. You, and D. Tao, "Adaptive manifold regularized matrix factorization for data clustering," in *IJCAI*, 2017, pp. 3399–3405.
[9] H. Gao, F. Nie, and H. Huang, "Local centroids structured nonnegative matrix factorization," in *AAAI*, 2017, pp. 1905–1911.
[10] J. Huang, F. Nie, H. Huang, and C. Ding, "Robust manifold nonnegative matrix factorization," *KDD*, vol. 8, no. 3, pp. 11:1–11:21, 2013.
[11] L. Zhang, Q. Zhang, B. Du, D. Tao, and J. You, "Robust manifold matrix factorization for joint clustering and feature extraction," in *AAAI*, 2017, pp. 1662–1668.
[12] T. Zhou, W. Bian, and D. Tao, "Divide-and-conquer anchoring for near-separable nonnegative matrix factorization and completion in high dimensions," in *International Conference on Data Mining, Dallas*, 2013, pp. 917–926.
[13] S. Zafeiriou, A. Tefas, I. Buciu, and I. Pitas, "Exploiting discriminant information in nonnegative matrix factorization with application to frontal face verification," *TNN*, vol. 17, no. 3, pp. 683–695, 2006.
[14] Y. Lu, Z. Lai, Y. Xu, X. Li, D. Zhang, and C. Yuan, "Nonnegative discriminant matrix factorization," *TCSVT*, vol. 27, no. 7, pp. 1392–1405, 2017.
[15] K. Fukunaga, *Introduction to statistical pattern recognition*. Academic Press, 1972.
[16] D. Tao, X. Li, X. Wu, and S. Maybank, "Geometric mean for subspace selection," *IEEE Transactions on Pattern Analysis and Machine Intellegence*, vol. 31, no. 2, pp. 260–274, 2009.
[17] Y. Yang, H. Shen, Z. Ma, Z. Huang, and X. Zhou, "$l_{2,1}$-norm regularized discriminative feature selection for unsupervised learning," in *IJCAI*, 2011, pp. 1589–1594.
[18] F. Nie, H. Wang, H. Huang, and C. Ding, "Joint schatten p -norm and ℓ_p -norm robust matrix completion for missing value recovery," *KIS*, vol. 42, no. 3, pp. 525–544, 2015.
[19] F. Nie, R. Zhang, and X. Li, "A generalized power iteration method for solving quadratic problem on the stiefel manifold," *SCIS*, vol. 60, no. 11, p. 112101, 2017.
[20] F. Nie, X. Wang, M. Jordan, and H. Huang, "The constrained laplacian rank algorithm for graph-based clustering," in *AAAI*, 2016, pp. 1969–1976.
[21] F. Nie, W. Zhu, and X. Li, "Unsupervised feature selection with structured graph optimization," in *AAAI*.
[22] L. Manor and P. Perona, "Self-tuning spectral clustering," in *NIPS*, 2004, pp. 1601–1608.
[23] D. Cai, X. He, J. Han, and T. Huang, "Graph regularized nonnegative matrix factorization for data representation," *PAMI*, vol. 33, no. 8, p. 1548, 2011.
[24] M. Lyons, J. Budynek, and S. Akamatsu, "Automatic classification of single facial images," *PAMI*, vol. 21, no. 12, pp. 1357–1362, 1999.
[25] H. Wechsler, a. F. P. Phillips, and V. Bruce, and T. S. Huang, "Face recognition: From theory to applications," *Springer-Verlage*, 1998.
[26] X. He, S. Yan, Y. Hu, P. Niyogi, and H. Zhang, "Face recognition using laplacianfaces," *PAMI*, vol. 27, no. 3, pp. 328–340, 2005.

[27] J. Khan, J. Wei, M. Ringnr, L. Saal, M. Ladanyi, F. Westermann, F. Berthold, M. Schwab, C. Antonescu, and C. Peterson, "Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks," *Nature Medicine*, vol. 7, no. 6, pp. 673–679, 2001.
[28] M. Lichman, "UCI machine learning repository," 2013.
[29] J. Shi and J. Malik, "Normalized cuts and image segmentation," *PAMI*, vol. 22, no. 8, pp. 888–905, 2000.

Xuelong Li (M'02-SM'07-F'12) is currently a Full Professor with the School of Computer Science and the Center for OPTical Imagery Analysis and Learning (OPTIMAL), Northwestern Polytechnical University, Xi'an 710072, China.

Mulin Chen received the B.E. degree in software engineering and the M.E. degree in computer application technology from Northwestern Polytechnical University, Xi'an, China, in 2014 and 2016 respectively. He is currently pursuing the Ph.D. degree with the School of Computer Science and the Center for Optical Imagery Analysis and Learning (OPTIMAL), Northwestern Polytechnical University, Xi'an, China. His current research intersts include computer vision and machine learning.

Qi Wang (M'15-SM'15) received the B.E. degree in automation and the Ph.D. degree in pattern recognition and intelligent systems from the University of Science and Technology of China, Hefei, China, in 2005 and 2010, respectively. He is currently a Professor with the School of Computer Science and the Center for OPTical IMagery Analysis and Learning (OPTIMAL), Northwestern Polytechnical University, Xi'an, China. His research interests include computer vision and pattern recognition.