Supplementary Material: Jointly Learning Non-negative Projection and Dictionary with Discriminative Graph Constraints for Classification

Weiyang Liu[1], Zhiding Yu[2], Yandong Wen[2], Rongmei Lin[3], Meng Yang[4]
[1]Peking University, [2]Carnegie Mellon University
[3]Emory University, [4]Shenzhen University

Summary

The JNPDL model is well motivated by the current drawbacks of dictionary learning approaches, while each constraints are also well designed (the novel discriminative graph constraints are proposed, and all constrains are designed to be easily optimized). Aiming to bridge the gap between features and dictionary, we do think the proposed idea of learning a projection for features jointly with the dictionary is worth noticing. Consider that the given training data is usually not naturally discriminative, yhus the discrimination of the learned dictionary will be limited by the training data. Jointly learning the discriminative feature projection and discriminative dictionary together helps to improve each other. The key of JNPDL is to learn a discriminative projection (non-negativity is one way to discrimination) that can better work with the dictionary.

In the supplementary material, we give the detailed optimization framework for the JNPDL model in **Appendix A**, and then provide a detailed proof in **Appendix B** to show the convergence of the updating rules for $\boldsymbol{P}$ and $\boldsymbol{M}$ is theoretically guaranteed. In other words, the proposed multiplicative updating algorithm for the non-negative projection $\boldsymbol{P}$ in the paper is proved to be convergent.

Appendix A: Optimization Framework

We adopt a standard iterative learning framework to jointly learn the sparse representation $\boldsymbol{X}$, the non-negative projection matrix $\boldsymbol{P}$, the intermediate non-negative basis matrix $\boldsymbol{M}$ and the dictionary $\boldsymbol{D}$. The proposed algorithm is shown in Algorithm 1. The non-negative projection learning converges as we prove in **Appendix B**.

Non-negative projection learning

To learn the non-negative projection, we optimize $\boldsymbol{P}, \boldsymbol{M}$ with $\boldsymbol{D}, \boldsymbol{X}$ fixed. Thus the JNPDL model is rewritten as

$$\begin{aligned}\min_{\boldsymbol{P}\geq\boldsymbol{0},\boldsymbol{M}\geq\boldsymbol{0}} \ \Big\{ & \|\boldsymbol{PY}-\boldsymbol{DX}\|_F^2+\sum_{i=1}^{K}\|\boldsymbol{PY}_i-\boldsymbol{D}_i\boldsymbol{X}_i^i\|_F^2 \\ & +\alpha_1\|\boldsymbol{Y}-\boldsymbol{MPY}\|_F^2+\alpha_1\beta\cdot\text{Tr}(\hat{\boldsymbol{P}}\boldsymbol{Y}\boldsymbol{L}_p\boldsymbol{Y}^T\hat{\boldsymbol{P}}^T) \\ & +\alpha_1\beta\cdot\text{Tr}(\tilde{\boldsymbol{P}}\boldsymbol{Y}\boldsymbol{L}_p^p\boldsymbol{Y}^T\tilde{\boldsymbol{P}}^T)+\alpha_1\|\boldsymbol{M}-\boldsymbol{P}^T\|_F^2\Big\}\end{aligned} \tag{1}$$

which is essentially a projective non-negative matrix factorization problem [6, 2]. We use the multiplicative iterative solution [6, 2, 4] to solve Eq. (1). Specifically, we transform it into tractable sub-problems and optimize $\boldsymbol{M}$ and $\boldsymbol{P}$ by a multiplicative non-negative iterative procedure.

Because $\boldsymbol{M}$ is the basis matrix, it is necessary to require each column $\boldsymbol{m}_i$ to have unit l_2 norm, i.e., $\|\boldsymbol{m}_i\|=1$. This extra constraint makes the optimization more complicated, so we compensate the norms of the basis matrix into the coefficient matrix as in [4] and replace $\alpha_1\beta\text{Tr}(\hat{\boldsymbol{P}}\boldsymbol{Y}\boldsymbol{L}_p\boldsymbol{Y}^T\hat{\boldsymbol{P}}^T)+\alpha_1\beta\text{Tr}(\tilde{\boldsymbol{P}}\boldsymbol{Y}\boldsymbol{L}_p^p\boldsymbol{Y}^T\tilde{\boldsymbol{P}}^T)$ with

$$\alpha_1\beta\cdot\Big(\text{Tr}(\hat{\boldsymbol{Q}}\hat{\boldsymbol{P}}\boldsymbol{Y}\boldsymbol{L}_p\boldsymbol{Y}^T\hat{\boldsymbol{P}}^T\hat{\boldsymbol{Q}}^T)+\text{Tr}(\tilde{\boldsymbol{Q}}\tilde{\boldsymbol{P}}\boldsymbol{Y}\boldsymbol{L}_p^p\boldsymbol{Y}^T\tilde{\boldsymbol{P}}^T\tilde{\boldsymbol{Q}}^T)\Big) \tag{2}$$

where $\hat{\boldsymbol{Q}}$ equals $\text{diag}\{\|\boldsymbol{m}_1\|,\cdots,\|\boldsymbol{m}_q\|\}$ and $\tilde{\boldsymbol{Q}}$ equals $\text{diag}\{\|\boldsymbol{m}_{q+1}\|,\cdots,\|\boldsymbol{m}_s\|\}$.

On optimizing M with P, D, X fixed. We can further rewrite Eq. (18) as Tr$(\boldsymbol{M}\boldsymbol{G}_m\boldsymbol{M}^T)$ where

$$\begin{aligned}\boldsymbol{G}_m &= \boldsymbol{G}_{m+} - \boldsymbol{G}_{m-}\\ &= \begin{bmatrix}\hat{\boldsymbol{P}}\boldsymbol{Y}(\alpha_1\beta\boldsymbol{B}_p)\boldsymbol{Y}^T\hat{\boldsymbol{P}}^T & 0\\ 0 & \tilde{\boldsymbol{P}}\boldsymbol{Y}(\alpha_1\beta\boldsymbol{B}_p^p)\boldsymbol{Y}^T\tilde{\boldsymbol{L}}^T\end{bmatrix}\odot\boldsymbol{I}\\ &\quad-\begin{bmatrix}\hat{\boldsymbol{P}}\boldsymbol{Y}(\alpha_1\beta\boldsymbol{W}_p)\boldsymbol{Y}^T\hat{\boldsymbol{P}}^T & 0\\ 0 & \tilde{\boldsymbol{P}}\boldsymbol{Y}(\alpha_1\beta\boldsymbol{W}_p^p)\boldsymbol{Y}^T\tilde{\boldsymbol{L}}^T\end{bmatrix}\odot\boldsymbol{I}\end{aligned} \tag{3}$$

where $\odot$ denotes the element-wise matrix multiplication, and $\boldsymbol{I}$ is an identity matrix. Then we put the non-negative constraints into the objective function with respect to $\boldsymbol{M}$, and define $\boldsymbol{\psi}_{ij}$ as the Lagrange multiplier for $\boldsymbol{M} \geq \boldsymbol{0}$. With $\boldsymbol{\Psi} = [\psi_{ij}]$, the Lagrange $\mathcal{L}(\boldsymbol{W})$ is defined as

$$\begin{aligned}\mathcal{L}(\boldsymbol{M}) =&\|\boldsymbol{P}\boldsymbol{Y}-\boldsymbol{D}\boldsymbol{X}\|_F^2+\sum\|\boldsymbol{P}\boldsymbol{Y}_i-\boldsymbol{D}_i\boldsymbol{X}_i^i\|_F^2+\\ &\alpha_1\|\boldsymbol{Y}-\boldsymbol{M}\boldsymbol{P}\boldsymbol{Y}\|_F^2+\text{Tr}(\boldsymbol{M}\boldsymbol{G}_m\boldsymbol{M}^T)\\ &+\alpha_1\|\boldsymbol{M}-\boldsymbol{P}^T\|_F^2+\text{Tr}(\boldsymbol{\Psi}\boldsymbol{M}^T)\end{aligned}. \tag{4}$$

Thus the partial derivative of $\mathcal{L}$ with respect to $\boldsymbol{M}$ is

$$\begin{aligned}\frac{\partial\mathcal{L}}{\partial\boldsymbol{M}} =&-2\alpha_1\boldsymbol{Y}\boldsymbol{Y}^T\boldsymbol{P}^T+2\alpha_1\boldsymbol{M}\boldsymbol{P}\boldsymbol{Y}\boldsymbol{Y}^T\boldsymbol{P}^T\\ &+2\boldsymbol{M}\boldsymbol{G}_m+2\alpha_1\boldsymbol{M}-2\alpha_1\boldsymbol{P}^T+\boldsymbol{\Psi}\end{aligned}. \tag{5}$$

According to the Karush-Kuhn-Tucker (KKT) condition ($\psi_{ij}\boldsymbol{M}_{ij}=0$) and $\frac{\partial\mathcal{L}}{\partial\boldsymbol{M}}=0$, we can obtain the update rule:

$$\boldsymbol{M}_{ij}^{(t+1)} = \boldsymbol{M}_{ij}^{(t)}\frac{(\alpha_1\boldsymbol{Y}\boldsymbol{Y}^T\boldsymbol{P}^T+\boldsymbol{M}\boldsymbol{G}_{m-}+\alpha_1\boldsymbol{P}^T)_{ij}}{(\alpha_1\boldsymbol{M}\boldsymbol{P}\boldsymbol{Y}\boldsymbol{Y}^T\boldsymbol{P}^T+\boldsymbol{M}\boldsymbol{G}_{m+}+\alpha_1\boldsymbol{M})_{ij}}. \tag{6}$$

On optimizing P with M, D, X fixed. After updating $\boldsymbol{M}$, we normalize the column vectors of $\boldsymbol{M}$ and multiply the norm to the projective matrix $\boldsymbol{P}$, namely $\boldsymbol{P}_i \leftarrow \boldsymbol{P}_i \times \|\boldsymbol{m}_i\|_2,\ \boldsymbol{m}_i \leftarrow \boldsymbol{m}_i/\|\boldsymbol{m}_i\|_2, \forall i$. By using the normalized $\boldsymbol{M}$, we simplify Eq. (18) as $\alpha_1\beta\text{Tr}(\hat{\boldsymbol{P}}\boldsymbol{Y}\boldsymbol{L}_p\boldsymbol{Y}^T\hat{\boldsymbol{P}}^T)+\alpha_1\beta\text{Tr}(\tilde{\boldsymbol{P}}\boldsymbol{Y}\boldsymbol{L}_p^p\boldsymbol{Y}^T\tilde{\boldsymbol{P}}^T)$. Thus the Lagrange $\mathcal{L}(\boldsymbol{P})$ is

$$\begin{aligned}\mathcal{L}(\boldsymbol{P}) =& \|\boldsymbol{P}\boldsymbol{Y}-\boldsymbol{D}\boldsymbol{X}\|_F^2+\sum\|\boldsymbol{P}\boldsymbol{Y}_i-\boldsymbol{D}_i\boldsymbol{X}_i^i\|_F^2+\\ &\alpha_1\beta\text{Tr}(\hat{\boldsymbol{P}}\boldsymbol{Y}\boldsymbol{L}_p\boldsymbol{Y}^T\hat{\boldsymbol{P}}^T)+\alpha_1\beta\text{Tr}(\tilde{\boldsymbol{P}}\boldsymbol{Y}\boldsymbol{L}_p^p\boldsymbol{Y}^T\tilde{\boldsymbol{P}}^T)+\\ &\alpha_1\|\boldsymbol{Y}-\boldsymbol{M}\boldsymbol{P}\boldsymbol{Y}\|_F^2+\alpha_1\|\boldsymbol{M}-\boldsymbol{P}^T\|_F^2+\text{Tr}(\boldsymbol{\Phi}\boldsymbol{P}^T)\end{aligned} \tag{7}$$

where ϕ is the Lagrange multiplier for constraint $\boldsymbol{P}_{ij}\geq 0$ and $\boldsymbol{\Phi}=[\phi_{ij}]$. After setting $\frac{\partial\mathcal{L}}{\partial\boldsymbol{P}}=\boldsymbol{0}$ and applying KKT condition ($\phi_{ij}\boldsymbol{P}_{ij}=0$), we obtain the update rule for $\boldsymbol{P}$:

$$\boldsymbol{P}_{ij}^{(t+1)} = \boldsymbol{P}_{ij}^{(t)}\frac{\begin{pmatrix}\boldsymbol{D}\boldsymbol{X}\boldsymbol{Y}^T+\sum\boldsymbol{D}_i\boldsymbol{X}_i^i\boldsymbol{Y}_i^T+\alpha_1\boldsymbol{M}^T\boldsymbol{Y}\boldsymbol{Y}^T\\ +\alpha_1\boldsymbol{M}^T+\alpha_1\beta\begin{bmatrix}\hat{\boldsymbol{P}}^t\boldsymbol{Y}\boldsymbol{W}_p\boldsymbol{Y}^T\\ \tilde{\boldsymbol{P}}^t\boldsymbol{Y}\boldsymbol{W}_p^p\boldsymbol{Y}^T\end{bmatrix}\end{pmatrix}_{ij}}{\begin{pmatrix}\boldsymbol{P}^t\boldsymbol{Y}\boldsymbol{Y}^T+\sum\boldsymbol{P}^t\boldsymbol{Y}_i\boldsymbol{Y}_i^T+\alpha_1\boldsymbol{P}^t+\\ \alpha_1\boldsymbol{M}^T\boldsymbol{M}\boldsymbol{P}^t\boldsymbol{Y}\boldsymbol{Y}^T+\alpha_1\beta\begin{bmatrix}\hat{\boldsymbol{P}}^t\boldsymbol{Y}\boldsymbol{B}_p\boldsymbol{Y}^T\\ \tilde{\boldsymbol{P}}^t\boldsymbol{Y}\boldsymbol{B}_p^p\boldsymbol{Y}^T\end{bmatrix}\end{pmatrix}_{ij}}. \tag{8}$$

Now both Eq. (6) and Eq. (41) are non-negative update. We prove that the convergence of the updating rule for $\boldsymbol{P}$ and $\boldsymbol{M}$ can be guaranteed. Detailed proof can refer to the supplementary material (Appendix A).

Discriminative dictionary learning

The discriminative dictionary learning is optimized using a standard iterative optimization procedure that is widely adopted in sparse coding.

On optimizing X with D, P, M fixed. With $\boldsymbol{D}, \boldsymbol{P}$ fixed, the optimization of the JNPDL model becomes

$$\begin{aligned}\min_{\boldsymbol{X}}\{&\|\boldsymbol{P}\boldsymbol{Y}-\boldsymbol{D}\boldsymbol{X}\|_F^2+\sum_{i=1}^{K}\|\boldsymbol{P}\boldsymbol{Y}_i-\boldsymbol{D}_i\boldsymbol{X}_i^i\|_F^2+\alpha_3\|\boldsymbol{X}\|_1\\ &+\sum_{i=1}^{K}\sum_{j=1,j\neq i}^{K}\|\boldsymbol{D}_j\boldsymbol{X}_i^j\|_F^2+\alpha_2\text{Tr}(\boldsymbol{X}^T(\boldsymbol{L}')\boldsymbol{X})\}\end{aligned} \tag{9}$$

Algorithm 1 Training Procedure of JNPDL

Input: Training samples $\boldsymbol{Y} = \|\boldsymbol{Y}_1, \cdots, \boldsymbol{Y}_N\|$, intrinsic graph $\boldsymbol{W}_c, \boldsymbol{W}_p$, penalty graph $\boldsymbol{W}_c^p, \boldsymbol{W}_p^p$, parameters $\alpha_1, \alpha_2, \alpha_3, \beta$, iteration number T.

Output: Non-negative projection matrix $\boldsymbol{P}$, dictionary $\boldsymbol{D}$, coding coefficient matrix $\boldsymbol{X}$.

Step1: Initialization

$t = 1$.

Randomly initialize columns in $\boldsymbol{D}^0, \boldsymbol{M}^0$ with unit l_2 norm.

Initialize $\boldsymbol{x}_{i,1\leq i\leq N}$ with $((\boldsymbol{D}^0)^T(\boldsymbol{D}^0) + \lambda_2 \boldsymbol{I})^{-1}(\boldsymbol{D}^0)^T\boldsymbol{y}_i$ where $\boldsymbol{y}_i$ is the ith training sample (regardless of label).

Step2: Search local optima

while not convergence or $t < T$ **do**

  Solve $\boldsymbol{P}^t, \boldsymbol{M}^t$ iteratively with fixed $\boldsymbol{D}^{t-1}$ and $\boldsymbol{X}^{t-1}$ via Eq. (1).

  Solve $\boldsymbol{X}^t$ with fixed $\boldsymbol{M}^t$, $\boldsymbol{D}^{t-1}$ and $\boldsymbol{P}^t$ via Eq. (10).

  Solve $\boldsymbol{D}^t$ with fixed $\boldsymbol{M}^t$, $\boldsymbol{P}^t$ and $\boldsymbol{X}^t$ via Eq. (12).

  $t \leftarrow t + 1$.

end while

Step3: Output

Output $\boldsymbol{P} = \boldsymbol{P}^t$, $\boldsymbol{D} = \boldsymbol{D}^t$ and $\boldsymbol{X} = \boldsymbol{X}^t$.

where $\boldsymbol{L}' = \boldsymbol{L}_c - \boldsymbol{L}_c^p$. Eq. (9) can be solved using feature sign search algorithm [1] after certain formulation based on [7, 3]. We optimize $\boldsymbol{X}$ class by class. Following [1], we update $\boldsymbol{X}_i$ one by one in the ith class. We define $\boldsymbol{x}_{i,j}$ as the coding coefficients of the jth sample in the ith class and reformulated the problem as

$$\begin{aligned}\min_{\boldsymbol{X}}\{&\|\boldsymbol{P}\boldsymbol{Y}_i - \boldsymbol{D}\boldsymbol{x}_{i,j}\|_F^2 + \|\boldsymbol{P}\boldsymbol{Y}_i - \boldsymbol{D}_i\boldsymbol{x}_{i,j}^i\|_F^2 \\ &+ \sum_{n=1,n\neq i}^{K} \|\boldsymbol{D}_n\boldsymbol{x}_{i,j}^n\|_F^2 + \alpha_2 Q(\boldsymbol{x}_{i,j}) + \alpha_3\|\boldsymbol{x}_{i,j}\|_1\end{aligned} \tag{10}$$

where $Q(\boldsymbol{x}_{i,j}) = \alpha_2(\boldsymbol{x}_{i,j}^T\boldsymbol{X}_i\boldsymbol{L}_j' + (\boldsymbol{X}_i\boldsymbol{L}_j')^T\boldsymbol{x}_{i,j} - \boldsymbol{x}_{i,j}^T\boldsymbol{L}_{jj}')$ in which $\boldsymbol{L}_i'$ is the ith column of $\boldsymbol{L}$, and $\boldsymbol{L}_{ii}'$ is the entry in the ith row and ith column of $\boldsymbol{L}$. $\boldsymbol{x}_{i,j}$ can be solved via feature sign search algorithm as in [7, 3].

On optimizing $\boldsymbol{D}$ with $\boldsymbol{P}, \boldsymbol{X}, \boldsymbol{M}$ fixed. By fixing $\boldsymbol{P}$, $\boldsymbol{M}$ and $\boldsymbol{X}$, the JNPDL model is rewritten as

$$\begin{aligned}\min_{\boldsymbol{D}}\{&\|\boldsymbol{P}\boldsymbol{Y} - \boldsymbol{D}\boldsymbol{X}\|_F^2 + \sum_{i=1}^{K}\|\boldsymbol{P}\boldsymbol{Y}_i - \boldsymbol{D}_i\boldsymbol{X}_i^i\|_F^2 \\ &+ \sum_{i=1}^{K}\sum_{j=1,j\neq i}^{K}\|\boldsymbol{D}_j\boldsymbol{X}_i^j\|_F^2\}\end{aligned} \tag{11}$$

for which we update $\boldsymbol{D}$ class by class sequentially. When we update $\boldsymbol{D}_v$, the sub-dictionaries $\boldsymbol{D}_i, i \neq v$ associated to the other classes will be fixed. Thus Eq. (11) can be further rewritten to

$$\begin{aligned}\min_{\boldsymbol{D}_i, i\in\{1,2,\cdots,K\}}\{&\|\boldsymbol{P}\boldsymbol{Y}_i - \boldsymbol{D}_i\boldsymbol{X}_i\|_F^2 + \|\boldsymbol{P}\boldsymbol{Y}_i - \boldsymbol{D}_i\boldsymbol{X}_i^i\|_F^2 \\ &+ \sum_{j=1,j\neq i}^{K}\|\boldsymbol{D}_j\boldsymbol{X}_i^j\|_F^2\}\end{aligned} \tag{12}$$

which is essentially a quadratic programming problem and can be directly solved by the algorithm presented in [5] (update $\boldsymbol{D}_i$ atom by atom). Note that each atom in the dictionary should have unit l_2 norm.

Appendix B: Proof of the convergence of updating rules for P and M

Proof. Before proving the convergence of the updating rules, we first introduce some necessary preliminaries.

Definition 1. *Function $\mathcal{G}(\boldsymbol{A}, \boldsymbol{A}')$ is an auxiliary function for function $\mathcal{F}(\boldsymbol{A})$ if the conditions*

$$\mathcal{G}(\boldsymbol{A}, \boldsymbol{A}') \geq \mathcal{F}(\boldsymbol{A}), \;\; \mathcal{G}(\boldsymbol{A}, \boldsymbol{A}) = \mathcal{F}(\boldsymbol{A}) \tag{13}$$

are satisfied.

Lemma 1. *If $\mathcal{G}(\boldsymbol{A}, \boldsymbol{A}')$ is an auxiliary function for $\mathcal{F}(\boldsymbol{A})$, then $\mathcal{F}(\boldsymbol{A})$ is non-increasing to $\boldsymbol{A}$ under the update*

$$\boldsymbol{A}^{t+1} = \arg\min_{\boldsymbol{A}} \mathcal{G}(\boldsymbol{A}, \boldsymbol{A}^t) \tag{14}$$

where t denotes the tth iteration.

Proof. From Eq.(14), we construct the following relation:

$$\mathcal{F}(\boldsymbol{A}^t) = \mathcal{G}(\boldsymbol{A}^t, \boldsymbol{A}^t) \geq \mathcal{G}(\boldsymbol{A}^{t+1}, \boldsymbol{A}^t). \tag{15}$$

Because $\mathcal{G}(\boldsymbol{A}, \boldsymbol{A}')$ is an auxiliary function for $\mathcal{F}(\boldsymbol{A})$, we can obtain the following inequality from Eq. (13):

$$\mathcal{G}(\boldsymbol{A}^{t+1}, \boldsymbol{A}^t) \geq \mathcal{F}(\boldsymbol{A}^{t+1}) \tag{16}$$

which leads to

$$\mathcal{F}(\boldsymbol{A}^{t+1}) \leq \mathcal{F}(\boldsymbol{A}^t). \tag{17}$$

Thus $\mathcal{F}(\boldsymbol{A})$ is non-increasing with respect to $\boldsymbol{A}$ under the updating rule in Eq. (14). The lemma is proved □

We first consider the scenario when $\boldsymbol{P}$ is fixed. With $\boldsymbol{P}$ fixed, we rewrite the optimization objective (Eq. (10) in the paper) as

$$\begin{aligned}\mathcal{F}(\boldsymbol{M}) =& \alpha_1\|\boldsymbol{Y} - \boldsymbol{MPY}\|_F^2 + Tr(\boldsymbol{MG}_m\boldsymbol{M}^T) \\ &+ \alpha_1\|\boldsymbol{M} - \boldsymbol{P}^T\|_F^2\end{aligned}. \tag{18}$$

We denote $\mathcal{F}_{ij}$ as the part of $\mathcal{F}(\boldsymbol{M})$ relevant to $\boldsymbol{M}_{ij}$, and then compute the first-order and the second-order derivative as follows:

$$\begin{aligned}\mathcal{F}'_{ij}(\boldsymbol{M}) =& \alpha_1(-2\boldsymbol{YY^T P}^T + 2\boldsymbol{MPYY}^T\boldsymbol{P}^T)_{ij} \\ &+ (2\boldsymbol{MG}_m)_{ij} + \alpha_1(2\boldsymbol{M} - 2\boldsymbol{P}^T)_{ij}\end{aligned}, \tag{19}$$

$$\mathcal{F}''_{ij}(\boldsymbol{M}) = 2\alpha_1(\boldsymbol{PYY}^T\boldsymbol{P}^T + \boldsymbol{G}_m + \boldsymbol{I})_{jj} \tag{20}$$

where $\boldsymbol{I}$ denotes an identity matrix with matched size. We construct the function $\mathcal{G}(\boldsymbol{M}_{ij}, \boldsymbol{M}_{ij}^t)$ as

$$\begin{aligned}&\mathcal{G}(\boldsymbol{M}_{ij}, \boldsymbol{M}_{ij}^t) = \mathcal{F}_{ij}(\boldsymbol{M}_{ij}^t) + \mathcal{F}'_{ij}(\boldsymbol{M}_{ij}^t)(\boldsymbol{M}_{ij} - \boldsymbol{M}_{ij}^t) \\ &+ \frac{\alpha_1(\boldsymbol{M}^t\boldsymbol{PYY}^T\boldsymbol{P}^T + \boldsymbol{M}^t\boldsymbol{G}_{m+} + \boldsymbol{M}^t)_{ij}}{\boldsymbol{M}_{ij}^t}(\boldsymbol{M}_{ij} - \boldsymbol{M}_{ij}^t)^2\end{aligned}. \tag{21}$$

Lemma 2. *$\mathcal{G}(\boldsymbol{M}_{ij}, \boldsymbol{M}_{ij}^t)$ in Eq. (21) is an auxiliary function for the function $\mathcal{F}_{ij}(\boldsymbol{M})$.*

Proof. Because it is easily obtained that $\mathcal{G}(\boldsymbol{M}_{ij}, \boldsymbol{M}_{ij}) = \mathcal{F}_{ij}(\boldsymbol{M}_{ij})$, we only need to prove that $\mathcal{G}(\boldsymbol{M}_{ij}, \boldsymbol{M}_{ij}^t) \geq \mathcal{F}_{ij}(\boldsymbol{M}_{ij})$. We first compute the Taylor series expansion of $\mathcal{F}_{ij}(\boldsymbol{M})$ as

$$\begin{aligned}\mathcal{F}_{ij}(\boldsymbol{M}_{ij}) =& \mathcal{F}_{ij}(\boldsymbol{M}_{ij}^t) + \mathcal{F}'_{ij}(\boldsymbol{M}_{ij}^t)(\boldsymbol{M}_{ij} - \boldsymbol{M}_{ij}^t) \\ &+ \frac{1}{2}\mathcal{F}''_{ij}(\boldsymbol{M}_{ij}^t)(\boldsymbol{M}_{ij} - \boldsymbol{M}_{ij}^t)^2\end{aligned}. \tag{22}$$

Because the following inequalities are satisfied:

$$\begin{aligned}(\boldsymbol{M}^t\boldsymbol{PYY}^T\boldsymbol{P}^T)_{ij} &= \sum_v \left(\boldsymbol{M}_{iv}^t(\boldsymbol{PYY}^T\boldsymbol{P}^T)_{vj}\right) \\ &\geq \boldsymbol{M}_{ij}^t(\boldsymbol{PYY}^T\boldsymbol{P}^T)_{jj},\end{aligned} \tag{23}$$

$$\begin{aligned}(\boldsymbol{M}^t\boldsymbol{G}_{m+})_{ij} &= \sum_v \left(\boldsymbol{M}_{iv}^t(\boldsymbol{G}_{m+})_{vj}\right) \\ &\geq \boldsymbol{M}_{ij}^t(\boldsymbol{G}_m)_{jj},\end{aligned} \tag{24}$$

$$\boldsymbol{M}_{ij}^t \geq \boldsymbol{M}_{ij}^t\boldsymbol{I}_{jj}, \tag{25}$$

we can let the following relation hold:

$$\begin{aligned}&\frac{\alpha_1(\boldsymbol{M}^t\boldsymbol{PYY}^T\boldsymbol{P}^T + \boldsymbol{M}^t\boldsymbol{G}_{m+} + \boldsymbol{M}^t)_{ij}}{\boldsymbol{M}_{ij}^t} \\ &\geq (\boldsymbol{PYY}^T\boldsymbol{P}^T + \boldsymbol{G}_m)_{jj}.\end{aligned} \tag{26}$$

Therefore, we can prove that $\mathcal{G}(\boldsymbol{M}_{ij}, \boldsymbol{M}_{ij}^t) \geq \mathcal{F}_{ij}(\boldsymbol{M}_{ij})$ holds. The lemma is proved. □

Theorem 1. *The updating rule for $\boldsymbol{M}$ can be obtained by minimizing the auxiliary function $\mathcal{G}(\boldsymbol{M}_{ij}, \boldsymbol{M}_{ij}^t)$.*

Proof. We let the derivative of $\mathcal{G}(\boldsymbol{M}_{ij}, \boldsymbol{M}_{ij}^t)$ with respect to $\boldsymbol{M}_{ij}$ equal to zero, namely

$$\begin{aligned}&\frac{\partial \mathcal{G}(\boldsymbol{M}_{ij}, \boldsymbol{M}_{ij}^t)}{\partial \boldsymbol{M}_{ij}}\\ &= \frac{2\alpha_1(\boldsymbol{M}^t\boldsymbol{P}\boldsymbol{Y}\boldsymbol{Y}^T\boldsymbol{P}^T + \boldsymbol{M}^t\boldsymbol{G}_{m+} + \boldsymbol{M}^t)_{ij}}{\boldsymbol{M}_{ij}^t}(\boldsymbol{M}_{ij} - \boldsymbol{M}_{ij}^t)\\ &\quad + \mathcal{F}'_{ij}(\boldsymbol{M}_{ij}^t).\\ &= 0\end{aligned} \tag{27}$$

from which we can derive

$$\boldsymbol{M}_{ij}^{t+1} = \boldsymbol{M}_{ij}^t \frac{(\alpha_1\boldsymbol{Y}\boldsymbol{Y}^T\boldsymbol{P}^T + \boldsymbol{M}^t\boldsymbol{G}_{m-} + \alpha_1\boldsymbol{P}^T)_{ij}}{(\alpha_1\boldsymbol{M}^t\boldsymbol{P}\boldsymbol{Y}\boldsymbol{Y}^T\boldsymbol{P}^T + \boldsymbol{M}^t\boldsymbol{G}_{m+} + \alpha_1\boldsymbol{M}^t)_{ij}}. \tag{28}$$

which is identical to the updating rule that we use in the paper. Thus the lemma is proved □

Then we consider the other scenario when $\boldsymbol{M}$ is fixed. After updating the matrix $\boldsymbol{M}$ via Eq. (28), we normalize the column vectors $\boldsymbol{m}_i$ of $\boldsymbol{M}$ and consequently convey the norm to the projective matrix $\boldsymbol{P}$, namely

$$\begin{aligned}\boldsymbol{P}_i &\leftarrow \boldsymbol{P}_i \times \|\boldsymbol{m}_i\|\\ \boldsymbol{m}_i &\leftarrow \boldsymbol{m}_i/\|\boldsymbol{m}_i\|\end{aligned} \tag{29}$$

where $\boldsymbol{P}_i$ is the ith column vector of the projection matrix $\boldsymbol{P}$. Considering Eq. (29) and the fixed $\boldsymbol{M}$, we can rewrite the optimization objective (Eq. (10) in the paper) as

$$\begin{aligned}\mathcal{F}(\boldsymbol{P}) = &\|\boldsymbol{P}\boldsymbol{Y} - \boldsymbol{D}\boldsymbol{X}\|_F^2 + \sum\|\boldsymbol{P}\boldsymbol{Y}_i - \boldsymbol{D}_i\boldsymbol{X}_i^i\|_F^2 +\\ &\alpha_1\beta\text{Tr}(\hat{\boldsymbol{P}}\boldsymbol{Y}\boldsymbol{L}_p\boldsymbol{Y}^T\hat{\boldsymbol{P}}^T) + \alpha_1\beta\text{Tr}(\tilde{\boldsymbol{P}}\boldsymbol{Y}\boldsymbol{L}_p^p\boldsymbol{Y}^T\tilde{\boldsymbol{P}}^T)\cdot\\ &+ \alpha_1\|\boldsymbol{Y} - \boldsymbol{M}\boldsymbol{P}\boldsymbol{Y}\|_F^2 + \alpha_1\|\boldsymbol{M} - \boldsymbol{P}^T\|_F^2)\end{aligned} \tag{30}$$

By denoting $\mathcal{F}_{ij}$ as the part of $\mathcal{F}(\boldsymbol{P})$ relevant to $\boldsymbol{P}_{ij}$, we have the following derivatives:

$$\begin{aligned}&\mathcal{F}'_{ij}(\boldsymbol{P}) = 2(\boldsymbol{P}\boldsymbol{Y}\boldsymbol{Y}^T)_{ij} - 2(\boldsymbol{D}\boldsymbol{X}\boldsymbol{Y}^T))_{ij} + 2(\sum\boldsymbol{P}\boldsymbol{Y}_i\boldsymbol{Y}_i^T)_{ij}\\ &- 2\sum(\boldsymbol{D}_i\boldsymbol{X}_i^i\boldsymbol{Y}_i^T)_{ij} - 2\alpha_1(\boldsymbol{M}^T\boldsymbol{Y}\boldsymbol{Y}^T)_{ij} + 2\alpha_1(\boldsymbol{M}^T\boldsymbol{M}\boldsymbol{P}\boldsymbol{Y}\boldsymbol{Y}^T)_{ij},\\ &+ 2\alpha_1\beta\begin{bmatrix}\hat{\boldsymbol{P}}\boldsymbol{Y}\boldsymbol{L}_p\boldsymbol{Y}^T\\ \hat{\boldsymbol{P}}\boldsymbol{Y}\boldsymbol{L}_p^p\boldsymbol{Y}^T\end{bmatrix}_{ij} + \alpha_1(2\boldsymbol{P} - 2\boldsymbol{M}^T)_{ij}\end{aligned} \tag{31}$$

$$\begin{aligned}&\mathcal{F}''_{ij}(\boldsymbol{P}) = 2(\boldsymbol{Y}\boldsymbol{Y}^T)_{jj} + 2(\sum\boldsymbol{Y}_i\boldsymbol{Y}_i^T)_{jj} + 2\alpha_1(\boldsymbol{M}^T\boldsymbol{M})_{ii}(\boldsymbol{Y}\boldsymbol{Y}^T)_{jj}\\ &+ 2\alpha_1\beta\begin{bmatrix}\boldsymbol{Y}\boldsymbol{L}_p\boldsymbol{Y}^T\\ \boldsymbol{Y}\boldsymbol{L}_p^p\boldsymbol{Y}^T\end{bmatrix}_{jj} + 2\alpha_1\boldsymbol{I}_{jj}\end{aligned}. \tag{32}$$

The auxiliary function of $\mathcal{F}_{ij}(\boldsymbol{P})$ is designed as

$$\begin{aligned}&\mathcal{G}(\boldsymbol{P}_{ij}, \boldsymbol{P}_{ij}^t) = \mathcal{F}_{ij}(\boldsymbol{P}_{ij}^t) + \mathcal{F}'_{ij}(\boldsymbol{P}_{ij}^t)(\boldsymbol{P}_{ij} - \boldsymbol{P}_{ij}^t)\\ &+ \frac{\begin{pmatrix}\boldsymbol{P}^t\boldsymbol{Y}\boldsymbol{Y}^T + \sum\boldsymbol{P}^t\boldsymbol{Y}_i\boldsymbol{Y}_i^T + \alpha_1(\boldsymbol{P}^t) +\\ \alpha_1\boldsymbol{M}^T\boldsymbol{M}\boldsymbol{P}\boldsymbol{Y}\boldsymbol{Y}^T + \alpha_1\beta\begin{bmatrix}\hat{\boldsymbol{P}}^t\boldsymbol{Y}\boldsymbol{B}_p\boldsymbol{Y}^T\\ \tilde{\boldsymbol{P}}^t\boldsymbol{Y}\boldsymbol{B}_p^p\boldsymbol{Y}^T\end{bmatrix}\end{pmatrix}_{ij}}{\boldsymbol{P}_{ij}^t}(\boldsymbol{P}_{ij} - \boldsymbol{P}_{ij}^t)^2\end{aligned}. \tag{33}$$

Lemma 3. *$\mathcal{G}(\boldsymbol{P}_{ij}, \boldsymbol{P}_{ij}^t)$ in Eq. (33) is an auxiliary function for the function $\mathcal{F}_{ij}(\boldsymbol{P})$.*

Proof. Because obviously $\mathcal{G}(\boldsymbol{P}_{ij}, \boldsymbol{P}_{ij}) = \mathcal{F}_{ij}(\boldsymbol{P}_{ij})$, we only need to prove that $\mathcal{G}(\boldsymbol{P}_{ij}, \boldsymbol{P}_{ij}^t) = \mathcal{F}_{ij}(\boldsymbol{P}_{ij})$. We first obtain the Taylor series expansion of $\mathcal{F}_{ij}(\boldsymbol{P})$ as

$$\begin{aligned}\mathcal{F}_{ij}(\boldsymbol{P}_{ij}) = &\mathcal{F}_{ij}(\boldsymbol{P}_{ij}^t) + \mathcal{F}'_{ij}(\boldsymbol{P}_{ij}^t)(\boldsymbol{P}_{ij} - \boldsymbol{P}_{ij}^t)\\ &+ \frac{1}{2}\mathcal{F}''_{ij}(\boldsymbol{P}_{ij}^t)(\boldsymbol{P}_{ij} - \boldsymbol{P}_{ij}^t)^2\end{aligned}. \tag{34}$$

Since the following relations hold:

$$\begin{aligned}(\boldsymbol{P}^t\boldsymbol{Y}\boldsymbol{Y}^T)_{ij} &= \sum_v \left(\boldsymbol{P}_{iv}^t(\boldsymbol{Y}\boldsymbol{Y}^T)_{vj}\right)\\ &\geq \boldsymbol{P}_{ij}^t(\boldsymbol{Y}\boldsymbol{Y}^T)_{jj}\end{aligned}, \tag{35}$$

$$
\begin{aligned}
(\sum_i \boldsymbol{P}^t\boldsymbol{Y}_i\boldsymbol{Y}_i^T)_{ij} &= \sum_v \left(\boldsymbol{P}^t_{iv}(\sum \boldsymbol{Y}_i\boldsymbol{Y}_i^T)_{vj}\right) \\
&\geq \boldsymbol{P}^t_{ij}(\sum \boldsymbol{Y}_i\boldsymbol{Y}_i^T)_{jj}
\end{aligned}, \tag{36}
$$

$$
\begin{aligned}
(\boldsymbol{M}^T\boldsymbol{M}\boldsymbol{P}^t\boldsymbol{Y}\boldsymbol{Y}^T)_{ij} &= \sum_v \left((\boldsymbol{M}^T\boldsymbol{M}\boldsymbol{P}^t)_{iv}(\boldsymbol{Y}\boldsymbol{Y}^T)_{vj}\right) \\
&\geq (\boldsymbol{M}^T\boldsymbol{M}\boldsymbol{P}^t)_{ij}(\boldsymbol{Y}\boldsymbol{Y}^T)_{jj} \\
&= \sum_v \left((\boldsymbol{M}^T\boldsymbol{M})_{iv}\boldsymbol{P}^t_{vj}\right)(\boldsymbol{Y}\boldsymbol{Y}^T)_{jj} \\
&\geq \boldsymbol{P}^t_{ij}(\boldsymbol{M}^T\boldsymbol{M})_{ii}(\boldsymbol{Y}\boldsymbol{Y}^T)_{jj}
\end{aligned}, \tag{37}
$$

$$
\begin{aligned}
\begin{bmatrix}\hat{\boldsymbol{P}}^t\boldsymbol{Y}\boldsymbol{B}_p\boldsymbol{Y}^T \\ \tilde{\boldsymbol{P}}^t\boldsymbol{Y}\boldsymbol{B}_p^p\boldsymbol{Y}^T\end{bmatrix}_{ij} &= \begin{cases} \sum_v \left(\hat{\boldsymbol{P}}^t_{iv}(\boldsymbol{Y}\boldsymbol{B}_p\boldsymbol{Y}^T)_{vj}\right), & \text{if} j \leq q \\ \sum_v \left(\tilde{\boldsymbol{P}}^t_{iv}(\boldsymbol{Y}\boldsymbol{B}_p^p\boldsymbol{Y}^T)_{vj}\right), & \text{otherwise} \end{cases} \\
&\geq \begin{cases} \hat{\boldsymbol{P}}^t_{ij}(\boldsymbol{Y}\boldsymbol{B}_p\boldsymbol{Y}^T)_{jj}, & \text{if} j \leq q \\ \tilde{\boldsymbol{P}}^t_{ij}(\boldsymbol{Y}\boldsymbol{B}_p^p\boldsymbol{Y}^T)_{jj}, & \text{otherwise} \end{cases} \\
&\geq \begin{cases} \hat{\boldsymbol{P}}^t_{ij}(\boldsymbol{Y}\boldsymbol{L}_p\boldsymbol{Y}^T)_{jj}, & \text{if} j \leq q \\ \tilde{\boldsymbol{P}}^t_{ij}(\boldsymbol{Y}\boldsymbol{L}_p^p\boldsymbol{Y}^T)_{jj}, & \text{otherwise} \end{cases} \\
&= \boldsymbol{P}^t_{ij}\begin{bmatrix}\boldsymbol{Y}\boldsymbol{L}_p\boldsymbol{Y}^T \\ \boldsymbol{Y}\boldsymbol{L}_p^p\boldsymbol{Y}^T\end{bmatrix}_{jj}
\end{aligned}, \tag{38}
$$

$$
\boldsymbol{P}^t_{ij} \geq \boldsymbol{P}^t_{ij}\boldsymbol{I}_{jj}, \tag{39}
$$

we can have $\mathcal{G}(\boldsymbol{P}_{ij}, \boldsymbol{P}^t_{ij}) \geq \mathcal{F}(\boldsymbol{P}_{ij})$. Therefore the lemma is proved. □

Theorem 2. *The updating rule for* $\boldsymbol{P}$ *can be obtained by minimizing the auxiliary function* $\mathcal{G}(\boldsymbol{P}_{ij}, \boldsymbol{P}^t_{ij})$.

Proof. Let $(\partial\mathcal{G}(\boldsymbol{P}_{ij}, \boldsymbol{P}^t_{ij}))/(\partial\boldsymbol{P}_{ij}) = 0$, and we have

$$
\begin{aligned}
&\frac{2\begin{pmatrix}\boldsymbol{P}^t\boldsymbol{Y}\boldsymbol{Y}^T + \sum \boldsymbol{P}^t\boldsymbol{Y}_i\boldsymbol{Y}_i^T + \alpha_1(\boldsymbol{P}^t) + \\ \alpha_1\boldsymbol{M}^T\boldsymbol{M}\boldsymbol{P}\boldsymbol{Y}\boldsymbol{Y}^T + \alpha_1\beta\begin{bmatrix}\hat{\boldsymbol{P}}^t\boldsymbol{Y}\boldsymbol{B}_p\boldsymbol{Y}^T \\ \tilde{\boldsymbol{P}}^t\boldsymbol{Y}\boldsymbol{B}_p^p\boldsymbol{Y}^T\end{bmatrix}\end{pmatrix}_{ij}}{\boldsymbol{P}^t_{ij}}(\boldsymbol{P}_{ij} - \boldsymbol{P}^t_{ij}) \\
&+ \mathcal{F}'_{ij}(\boldsymbol{P}^t_{ij}) = 0
\end{aligned} \tag{40}
$$

from which we can derive the updating rule for $\boldsymbol{P}$

$$
\boldsymbol{P}^{(t+1)}_{ij} = \boldsymbol{P}^{(t)}_{ij}\frac{\begin{pmatrix}\boldsymbol{D}\boldsymbol{X}\boldsymbol{Y}^T + \sum \boldsymbol{D}_i\boldsymbol{X}_i^i\boldsymbol{Y}_i^T + \alpha_1\boldsymbol{M}^T\boldsymbol{Y}\boldsymbol{Y}^T \\ + \alpha_1\boldsymbol{M}^T + \alpha_1\beta\begin{bmatrix}\hat{\boldsymbol{P}}^t\boldsymbol{Y}\boldsymbol{W}_p\boldsymbol{Y}^T \\ \tilde{\boldsymbol{P}}^t\boldsymbol{Y}\boldsymbol{W}_p^p\boldsymbol{Y}^T\end{bmatrix}\end{pmatrix}_{ij}}{\begin{pmatrix}\boldsymbol{P}^t\boldsymbol{Y}\boldsymbol{Y}^T + \sum \boldsymbol{P}^t\boldsymbol{Y}_i\boldsymbol{Y}_i^T + \alpha_1\boldsymbol{P}^t + \\ \alpha_1\boldsymbol{M}^T\boldsymbol{M}\boldsymbol{P}^t\boldsymbol{Y}\boldsymbol{Y}^T + \alpha_1\beta\begin{bmatrix}\hat{\boldsymbol{P}}^t\boldsymbol{Y}\boldsymbol{B}_p\boldsymbol{Y}^T \\ \tilde{\boldsymbol{P}}^t\boldsymbol{Y}\boldsymbol{B}_p^p\boldsymbol{Y}^T\end{bmatrix}\end{pmatrix}_{ij}}. \tag{41}
$$

Thus the theorem is proved. □

According to **Lemma 1**, **Theorem 1** and **Theorem 2**, we have proved that the convergence of the updating rules for $\boldsymbol{P}$ and $\boldsymbol{M}$ can be theoretically guaranteed. □

References

[1] Honglak Lee, Alexis Battle, Rajat Raina, and Andrew Y Ng. Efficient sparse coding algorithms. In *NIPS*, 2006.

[2] Xiaobai Liu, Shuicheng Yan, and Hai Jin. Projective nonnegative graph embedding. *IEEE T. IP*, 19(5):1126–1137, 2010.

[3] Jiwen Lu, Gang Wang, Weihong Deng, and Pierre Moulin. Simultaneous feature and dictionary learning for image set based face recognition. In *ECCV*. 2014.

[4] Changhu Wang, Zheng Song, Shuicheng Yan, Lei Zhang, and Hong-Jiang Zhang. Multiplicative nonnegative greph embedding. In *CVPR*, 2009.

[5] Meng Yang, Lei Zhang, Jian Yang, and David Zhang. Metaface learning for sparse representation based face recognition. In *ICIP*, 2010.

[6] Zhirong Yang and Erkki Oja. Linear and nonlinear projective nonnegative matrix factorization. *IEEE T. NN*, 21(5):734–749, 2010.

[7] Miao Zheng, Jiajun Bu, Chun Chen, Can Wang, Lijun Zhang, Guang Qiu, and Deng Cai. Graph regularized sparse coding for image representation. *IEEE T. IP*, 20(5):1327–1336, 2011.