Low-Rank Correlation Analysis for Discriminative Subspace Learning

Jiacan Zheng[1], Zhihui Lai[1,2(✉)], Jianglin Lu[1], and Jie Zhou[1]

[1] Computer Vision Institute, College of Computer Science and Software Engineering, Shenzhen University, Shenzhen 518060, China
lai_zhi_hui@163.com

[2] Shenzhen Institute of Artificial Intelligence and Robotics for Society, Shenzhen, China

Abstract. Linear dimensionality reduction is a commonly used technique to solve the curse of dimensionality problem in pattern recognition. However, learning a discriminative subspace without label is still a challenging problem, especially when the high-dimensional data is grossly corrupted. To address this problem, we propose an unsupervised dimensionality reduction method called Low-Rank Correlation Analysis (LRCA). The proposed model integrates the low-rank representation and the linear embedding together with a seamless formulation. As such, the robustness and discriminative ability of the learned subspace can be effectively promoted together. An iterative algorithm equipped with alternating direction method of multiplier (ADMM) and eigendecomposition is designed to solve the optimization problem. Experiments show that our method is more discriminative and robust than some existing methods.

Keywords: Dimensionality reduction · Low-rank representation · Dictionary learning

1 Introduction

Many real-world data, such as face image, natural language and text, usually have very high dimensionality. To avoid the curse of dimensionality problem, a commonly used technique is linear dimensionality reduction, which aims to project the original high-dimensional data to a low-dimensional subspace with some desired intrinsic structure preservation. As stated in [2], a compact and discriminative embedding is vitally crucial for pattern recognition.

Principal Component Analysis (PCA) [23] is the most classical subspace learning method, which can preserve the global structure. Besides, some manifold-learning-based methods, such as Locality Preserving Projections (LPP) [11] and Neighborhood Preserving Embedding (NPE) [10], are recently proposed to learn a subspace with local geometric structure preservation. However, these unsupervised dimensionality reduction methods only assume that the data lies

C. Wallraven et al. (Eds.): ACPR 2021, LNCS 13189, pp. 87–100, 2022.
https://doi.org/10.1007/978-3-031-02444-3_7

on a single low-dimensional subspace or manifold [16]. Therefore, they cannot well capture the global class-special structure for multi-class data, which greatly degrades the discriminative ability of the learned subspace.

The key to improve the discriminative ability is to preserve the multi-subspace structure of original data [15,22]. As studied in [24], many real-world data can be seen as lying near multiple subspaces with each subspace corresponding to one class, e.g. face image [1]. Such a discriminative structure is crucial for data clustering, as studied in Sparse Subspace Clustering (SCC) [7] and Low-Rank Representation (LRR) [17].

Recently, some innovative works further demonstrate the great potential of the multi-subspace structure prior in representation learning. In [12,13,20,29], a deep convolutional auto-encoder is first initialized to learn a compact representation, and then the multi-subspace structure is enforced into the deep feature space, resulting the so-call self-expression layer. Besides, a recent work [28] shows that the objective of maximal coding rate reduction (MCR^2) will lead to the multiple orthogonal subspaces structure, which is significantly more robust than maximal cross-entropy in deep learning. Moreover, in [5,25], MCR^2 is further considered as the first principle of deep networks, for the unrolling algorithm to optimize MCR^2 is actually a modern residual network.

Motivated by these encouraging researches, this paper focus on enforcing the multi-subspace structure into the linear embedding space to improve the discriminative ability. To this end, we propose a novel dimensionality reduction method called Low-Rank Correlation Analysis (LRCA). The main contributions of this paper are listed as follows.

- We design a seamless formulation to optimize LRR and the linear embedding of data jointly. As such, the discriminative multi-subspace structure uncovered by LRR can be enforced into the embedding space.
- An iterative algorithm is designed to solve the proposed joint model. Technically, the designed algorithm optimizes the objective function by alternating direction method of multiplier (ADMM) and eigendecomposition iteratively.
- Theoretically, we prove the convergence of the designed algorithm. Experiments on several real-world datasets show that the proposed LRCA is more discriminative and robust than some existing methods.

2 Related Work

In this paper, the column-centered data matrix is denoted by $\mathbf{X} \in \mathbb{R}^{d\times n}$ where each column $\mathbf{x}_i$ is a d-dimensional sample. The projection matrix is denoted by $\mathbf{P} \in \mathbb{R}^{d\times k}$. The trace function of a square matrix is defined by $\text{tr}(\cdot)$ and the inner product of two matrices is denoted by $\langle \cdot, \cdot \rangle$.

2.1 Canonical Correlation Analysis

Canonical correlation analysis (CCA) [9] is the analog to PCA for two data matrices. The correlation coefficients of two random variables $\mathbf{x}, \mathbf{y}$ is defined as follows:

$$\mathrm{Corr}(\mathbf{x}, \mathbf{y}) = \frac{\mathbb{E}[(\mathbf{x} - \mathbb{E}(\mathbf{x}))(\mathbf{y} - \mathbb{E}(\mathbf{y}))]}{\sqrt{\mathrm{var}(\mathbf{x})\mathrm{var}(\mathbf{y})}}. \tag{1}$$

Now suppose that we have two data matrices, denoted by $\mathbf{X}$ and $\mathbf{Y}$, with centered columns. To find a most correlated latent subspace, CCA maximizes the empirical correlation criterion between two projected data matrices $\mathbf{P}^T\mathbf{X}$ and $\mathbf{Q}^T\mathbf{Y}$, which can be mathematically expressed as:

$$\max_{\mathbf{P},\mathbf{Q}} \frac{\langle \mathbf{P}^T\mathbf{X}, \mathbf{Q}^T\mathbf{Y} \rangle}{\sqrt{\langle \mathbf{P}^T\mathbf{X}, \mathbf{P}^T\mathbf{X} \rangle \langle \mathbf{Q}^T\mathbf{Y}, \mathbf{Q}^T\mathbf{Y} \rangle}}, \tag{2}$$

where the denominator is used to control the variance of $\mathbf{P}^T\mathbf{X}$ and $\mathbf{Q}^T\mathbf{Y}$.

2.2 Low-Rank Representation

LRR [16] aims to uncover the latent multi-subspace structure of data by optimizing a low-rank coefficient matrix $\mathbf{Z}$, leading to the following problem:

$$\min_{\mathbf{Z}} \ \mathrm{rank}(\mathbf{Z}), \quad \text{s.t. } \mathbf{X} = \mathbf{X}\mathbf{Z}. \tag{3}$$

Considering the NP-hard property of rank minimization problem and the existing outliers, LRR practically solves the following relaxed convex optimization model:

$$\min_{\mathbf{Z}} \ \|\mathbf{E}\|_{2,1} + \lambda\|\mathbf{Z}\|_*, \quad \text{s.t. } \mathbf{X} = \mathbf{X}\mathbf{Z} + \mathbf{E}, \tag{4}$$

where $\lambda \geq 0$ is a balance parameter, $\|\cdot\|_*$ represents the nuclear norm and $\|\cdot\|_{2,1}$ denotes the $\ell_{2,1}$-norm. The relaxed convex model (4) can be efficiently solved by ADMM. Such a low-rank coefficient matrix $\mathbf{Z}$ can well describe the class-special structure and thus is expected to be enforced into the embedding space.

3 Low-Rank Correlation Analysis

Many multi-class data can be modeled as samples drawn from multiple subspaces with each subspace corresponding to one class. However, most existing unsupervised dimensionality reduction methods assume that the data lies on a single subspace or manifold, therefore they fail to capture the global class-special structure, which greatly degrades the discriminative ability.

Motivated by this observation, our key insight is to enforce the multi-subspace structure into the embedding space $\mathbb{R}^k$. To this end, we first utilize LRR for robust subspace segmentation. Then, we propose a novel correlation criterion to establish the connection between the low rank coefficient matrix $\mathbf{Z}$ and the linear embedding $\mathbf{P}^T\mathbf{X}$.

Correlation Criterion: Suppose we have an over-complete dictionary $\mathbf{D} \in \mathbb{R}^{k\times n}$ with each column $\mathbf{d}_i$ being an atomic vector in embedding space $\mathbb{R}^k$, we then construct the so-called self-expressive data matrix $\mathbf{DZ}$. As the coefficient matrix $\mathbf{Z}$ captures the multi-subspace structure of the original data, self-expressive data matrix $\mathbf{DZ}$ then have the potential to preserve such a structure into the embedding space. Similar to CCA (2), we naturally maximize the correlation criterion between the linear embedding $\mathbf{P}^T\mathbf{X}$ and the self-expressive data matrix $\mathbf{DZ}$ to find a latent correlated subspace:

$$\mathrm{Corr}(\mathbf{D},\mathbf{P}) = \frac{\langle \mathbf{DZ}, \mathbf{P}^T\mathbf{X}\rangle}{\sqrt{\langle \mathbf{DZ},\mathbf{DZ}\rangle \langle \mathbf{P}^T\mathbf{X},\mathbf{P}^T\mathbf{X}\rangle}}. \tag{5}$$

To optimize a compact and discriminative embedding, we need to impose some regularizations into the over-complete dictionary $\mathbf{D}$ and the projection matrix $\mathbf{P}$. However, correlation criterion (5) does not facilitate the use of regularization due to the existence of the denominator. Therefore, we have to transform this formulation.

Regularization: For the projection matrix $\mathbf{P}$, parsimony have been widely used as a guiding principle. Therefore, the formulation (5) should equip with a regularization $\mathcal{R}(\mathbf{P})$. Motivated by MMC [14] where a penalty term is used to replace the denominator of the classical Fisher's criterion, we redefine a modified correlation criterion as follows:

$$\mathcal{J}(\mathbf{D},\mathbf{P}) = 2\langle \mathbf{DZ}, \mathbf{P}^T\mathbf{X}\rangle - \mathrm{tr}(\mathbf{P}^T\mathbf{X}\mathbf{X}^T\mathbf{P}) - \alpha\mathcal{R}(\mathbf{P}), \tag{6}$$

where $\alpha \geq 0$ is a balance parameter. As we can see, $\mathcal{J}(\mathbf{D},\mathbf{P})$ uses the penalty term $\mathrm{tr}(\mathbf{P}^T\mathbf{X}\mathbf{X}^T\mathbf{P})$ to control the variance of $\mathbf{P}^T\mathbf{X}$. In our model, we only consider a simple regularization $\mathcal{R}(\mathbf{P}) = \|\mathbf{P}\|_F^2$. Now the correlation criterion (5) becomes:

$$\max_{\mathbf{D},\mathbf{P}} \frac{\mathcal{J}(\mathbf{D},\mathbf{P})}{\sqrt{\mathrm{tr}\left(\mathbf{D}\boldsymbol{\Omega}\mathbf{D}^T\right)}}, \tag{7}$$

where $\boldsymbol{\Omega} = \mathbf{Z}\mathbf{Z}^T$.

For the over-complete dictionary $\mathbf{D}$, the atomic vectors $\mathbf{d}_i$ corresponding to the same subspace should be kept as compact as possible in the embedding space. Therefore, we need to optimize a locality-preserving dictionary $\mathbf{D}$:

$$\min_{\mathbf{D}} \sum_j \sum_i \|\mathbf{d}_i - \mathbf{d}_j\|_2^2 \mathbf{W}_{ij} \Leftrightarrow \min_{\mathbf{D}} \mathrm{tr}(\mathbf{D}\mathbf{L}\mathbf{D}^T), \tag{8}$$

where $\mathbf{W}$ is an affinity matrix and $\mathbf{L}$ is the corresponding Laplacian matrix. Note that, $\mathbf{W}$ can be simply constructed from the low rank coefficient matrix $\mathbf{Z}$ as follows:

$$\mathbf{W}_{ij} = \begin{cases} 1 & \text{if } \mathbf{x}_i \in C_t(\mathbf{x}_j) \text{ or } \mathbf{x}_j \in C_t(\mathbf{x}_i) \\ 0 & \text{otherwise,} \end{cases} \tag{9}$$

where $C_t(\mathbf{x}_i)$ is the set of t nearest data points of $\mathbf{x}_i$ computed by $|\mathbf{Z}|$. By replace $\mathbf{\Omega}$ in (7) with the Laplacian matrix $\mathbf{L}$, we then derive the following problem:

$$\max_{\mathbf{D},\mathbf{P}} \frac{\mathcal{J}(\mathbf{D},\mathbf{P})}{\sqrt{\mathrm{tr}(\mathbf{DLD}^T)}}. \tag{10}$$

To facilitate optimization, by fixing the denominator of (10), we can optimize the constraint correlation criterion as follows:

$$\begin{aligned} &\max_{\mathbf{D},\mathbf{P}} \ \mathcal{J}(\mathbf{D},\mathbf{P}) \\ &\text{s.t. } \mathbf{DLD}^T = \mathbf{I_k}. \end{aligned} \tag{11}$$

LRCA Model: Combining the model (4) with the model (11), a two-step model can be directly developed to learn a discriminative subspace. However, in such two-step scheme, the coefficient matrix $\mathbf{Z}$ is learned from (4) independently, and thus cannot be adaptive during the iterative process. To address this drawback, we design a joint formulation, called Low-Rank Correlation Analysis (LRCA), as follows:

$$\begin{aligned} &\min_{\mathbf{E},\mathbf{Z},\mathbf{D},\mathbf{P}} \ \|\mathbf{E}\|_{2,1} + \lambda\|\mathbf{Z}\|_* - \gamma\mathcal{J}(\mathbf{D},\mathbf{P}) \\ &\text{s.t. } \mathbf{X} = \mathbf{XZ} + \mathbf{E}, \ \mathbf{DLD}^T = \mathbf{I_k}, \end{aligned} \tag{12}$$

where $\gamma \geq 0$ is a trade-off parameter.

The proposed LRCA integrates the robust subspace segmentation (4) and the constraint correlation criterion (11) together with a seamless framework. As a result, the robustness and discriminative ability can be greatly improved.

4 Optimization and Convergence

Since the proposed LRCA contains four matrix variables, we design an algorithm to iteratively optimize variable pairs $(\mathbf{E},\mathbf{Z})$ and $(\mathbf{D},\mathbf{P})$ one by one.

Low-Rank Representation Step: Fixed $(\mathbf{D},\mathbf{P})$, with some simple transformation on (12), we can easily obtain the following equivalent problem:

$$\begin{aligned} &\min_{\mathbf{E},\mathbf{Z}} \ \|\mathbf{E}\|_{2,1} + \lambda\|\mathbf{Z}\|_* - 2\gamma\langle\mathbf{Z},\mathbf{M}\rangle \\ &\text{s.t. } \mathbf{X} = \mathbf{XZ} + \mathbf{E}, \end{aligned} \tag{13}$$

where $\mathbf{M} = \mathbf{D}^T\mathbf{P}^T\mathbf{X}$ is a constant matrix. Now we present how to solve this modified LRR problem by alternating direction method of multiplier (ADMM) algorithm. First, by introducing an auxiliary variable $\mathbf{J}$, we convert (13) to the following equivalent problem:

$$\begin{aligned} &\min_{\mathbf{E},\mathbf{Z},\mathbf{J}} \ \|\mathbf{E}\|_{2,1} + \lambda\|\mathbf{J}\|_* - 2\gamma\langle\mathbf{Z},\mathbf{M}\rangle \\ &\text{s.t. } \mathbf{X} = \mathbf{XZ} + \mathbf{E}, \ \mathbf{Z} = \mathbf{J}, \end{aligned} \tag{14}$$

Algorithm 1: Solving Subproblem (13) by ADMM.

Input: Data matrix $\mathbf{X}$, matrix $\mathbf{M}$, parameter λ, γ, μ
Output: Low rank coefficient matrix $\mathbf{Z}$

Initialize $\mathbf{Z} = \mathbf{J} = \mathbf{0}$,$\mathbf{Y}_1 = \mathbf{0}$,$\mathbf{Y}_2 = \mathbf{0}$
while *no converged* **do**
  Update $\mathbf{J}$ by $\mathbf{J} = \arg\min \frac{\lambda}{\mu}\|\mathbf{J}\|_* + \frac{1}{2}\|\mathbf{J} - (\mathbf{Z} + \mathbf{Y}_2/\mu)\|_F^2$.
  Update $\mathbf{Z}$ by $\mathbf{Z} = (\mathbf{X}^T\mathbf{X} + \mathbf{I})^{-1}[\mathbf{X}^T\mathbf{X} - \mathbf{X}^T\mathbf{E} + \mathbf{J} + (\mathbf{X}^T\mathbf{Y}_1 - \mathbf{Y}_2)/\mu + 2\gamma\mathbf{M}]$.
  Update $\mathbf{E}$ by $\mathbf{E} = \arg\min \frac{1}{2}\|\mathbf{E} - (\mathbf{X} - \mathbf{X}\mathbf{Z} + \mathbf{Y}_1/\mu)\|_F^2 + \frac{1}{\mu}\|\mathbf{E}\|_{2,1}$.
  Update $\mathbf{Y}_1, \mathbf{Y}_2$ by $\mathbf{Y}_1 = \mathbf{Y}_1 + \mu(\mathbf{X} - \mathbf{X}\mathbf{Z} - \mathbf{E})$ and $\mathbf{Y}_2 = \mathbf{Y}_2 + \mu(\mathbf{Z} - \mathbf{J})$.
  Check the convergence conditions $\|\mathbf{X} - \mathbf{X}\mathbf{Z} - \mathbf{E}\|_\infty < \epsilon$ and $\|\mathbf{Z} - \mathbf{J}\|_\infty < \epsilon$.

Algorithm 2: Low Rank Correlation Analysis.

Input: Data matrix $\mathbf{X}$, parameter λ, γ, α, iteration times T
Output: Dictionary $\mathbf{D}$, Projection matrix $\mathbf{P}$

Initialize coefficient matrix $\mathbf{Z}$ by LRR
for $i = 1 : T$ **do**
  Update $\mathbf{D}$ by (20).
  Update $\mathbf{P}$ by (18).
  Compute $\mathbf{M} = \mathbf{D}^T\mathbf{P}^T\mathbf{X}$.
  Update $(\mathbf{Z}, \mathbf{E})$ by Algorithm 1.

which can be solved by minimizing the augmented Lagrangian function as follows:

$$\begin{aligned}\mathcal{L} =& \|\mathbf{E}\|_{2,1} + \lambda\|\mathbf{J}\|_* - 2\gamma\langle\mathbf{Z}, \mathbf{M}\rangle + \mathrm{tr}\left[\mathbf{Y}_1^T(\mathbf{X} - \mathbf{X}\mathbf{Z} - \mathbf{E})\right] \\ &+ \mathrm{tr}\left[\mathbf{Y}_2^T(\mathbf{Z} - \mathbf{J})\right] + \frac{\mu}{2}\left(\|\mathbf{X} - \mathbf{X}\mathbf{Z} - \mathbf{E}\|_F^2 + \|\mathbf{Z} - \mathbf{J}\|_F^2\right),\end{aligned} \tag{15}$$

where $\mathbf{Y}_1$ and $\mathbf{Y}_2$ are Lagrange multipliers and $\mu > 0$ is a penalty parameter. The detailed ADMM algorithm is presented in Algorithm 1.

Maximal Correlation Step: Fixed $(\mathbf{Z}, \mathbf{E})$, by discarding some constant terms in (12), the optimization problem becomes:

$$\begin{aligned}&\max_{\mathbf{D},\mathbf{P}} \ 2\langle\mathbf{D}\mathbf{Z}, \mathbf{P}^T\mathbf{X}\rangle - \mathrm{tr}(\mathbf{P}^T\mathbf{X}\mathbf{X}^T\mathbf{P}) - \alpha\|\mathbf{P}\|_F^2 \\ &\text{s.t. } \mathbf{D}\mathbf{L}\mathbf{D}^T = \mathbf{I}_\mathbf{k}.\end{aligned} \tag{16}$$

Fortunately, problem (16) have a close form of solution by eigendecomposition. Mathematically, fixed $\mathbf{D}$, we can derive a equivalent regression formulation as follows:

$$\min_{\mathbf{P}} \ \|\mathbf{D}\mathbf{Z} - \mathbf{P}^T\mathbf{X}\|_F^2 + \alpha\|\mathbf{P}\|_F^2. \tag{17}$$

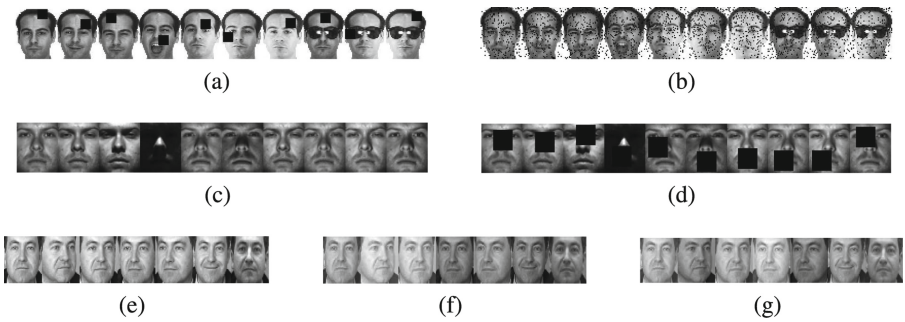


Fig. 1. The corrupted image samples of different datasets. AR dataset corrupted by (a) 10×10 block, (b) 10% salt-and-pepper noise. Extended Yale B dataset corrupted by (c) 0×0 block and (d) 20×20 block. FERET dataset corrupted by gray image with intensity in (e) [0, 0], (f) [0, 100] and (g) [0, 200].

Setting the partial derivative of the objective function in (17) with respect to $\mathbf{P}$ equal to zero gives the optimal solution of $\mathbf{P}$:

$$\mathbf{P} = (\mathbf{X}\mathbf{X}^T + \alpha\mathbf{I})^{-1}\mathbf{X}\mathbf{Z}^T\mathbf{D}^T. \tag{18}$$

Then, by substituting (18) back into (16), we derive the following trace maximization problem for $\mathbf{D}$:

$$\begin{aligned} &\max_{\mathbf{D}} \ \mathrm{tr}[\mathbf{D}\mathbf{Z}\mathbf{X}^T(\mathbf{X}\mathbf{X}^T + \alpha\mathbf{I})^{-1}\mathbf{X}\mathbf{Z}^T\mathbf{D}^T] \\ &\text{s.t. } \mathbf{D}\mathbf{L}\mathbf{D}^T = \mathbf{I_k}. \end{aligned} \tag{19}$$

The optimal solution of (19) is given from solving the generalized eigen equation as follows:

$$[\mathbf{Z}\mathbf{X}^T(\mathbf{X}\mathbf{X}^T + \alpha\mathbf{I})^{-1}\mathbf{X}\mathbf{Z}^T]\ \mathbf{v} = \eta\ \mathbf{L}\mathbf{v}, \tag{20}$$

where η is the eigenvalue and $\mathbf{v}$ is the corresponding eigenvector. Then, we can obtain the optimal solution of $\mathbf{D}^*$ with the i-th row being the eigenvectors corresponding to the i-th largest eigenvalues. Substituting the optimal $\mathbf{D}^*$ into (17) further gives the optimal solution of $\mathbf{P}^*$.

Overall, the total scheme is easy to implement. We describe the iterative process in Algorithm 2. Since the designed algorithm is an iterative method, we also give the Theorem 1 for its convergence.

Theorem 1. *The Algorithm 2 monotonically decreases the objective function value of (12) in each iteration and gives a local optimal solution.*

Proof. Boyd *et al.* [3] have comprehensively proven the convergence of the ADMM algorithm. Since the subproblem (13) is a convex optimization problem with two block variables and a linear constraint, as a direct result, the proposed Algorithm 1 will approach an optimal value of the objective function

(13). Besides, for the subproblem (16), the close form (18) and (20) give its optimal solution. Therefore, the total objective function of (12) will be decreased in each iteration. As the objective function (12) is bounded, the iterative scheme in Algorithm 2 will finally converge to a local optimal solution.

5 Experiments

We evaluate the robustness and discriminative ability of the proposed LRCA on three standard face datasets, e.g., AR, FERET and Extended Yale B datasets, where each face subject correspond to a subspace. Some most related methods are selected for comparison, such as PCA [23] and JSPCA [27], manifold-learning-based methods (NPE [10], LPP [11], OLPP [4]), and some low-rank methods LRR-Graph embedding [26], LRLE [6].

Table 1. The performance (recognition accuracy (%), dimension) of different algorithms on AR dataset different kinds of noise.

Method	AR		$AR_{10\times10}$		$AR_{20\times20}$		$AR_{\text{salt-and-pepper}}$	
	$T=5$	$T=7$	$T=5$	$T=7$	$T=5$	$T=7$	$T=5$	$T=7$
PCA	85.83 (90)	87.59 (100)	63.48 (100)	67.49 (100)	36.67 (100)	38.44 (100)	77.54 (83)	82.92 (96)
JSPCA	80.78 (100)	85.54 (100)	60.14 (100)	65.05 (100)	29.90 (100)	33.64 (100)	75.67 (100)	81.16 (100)
LPP	88.11 (100)	91.54 (100)	61.13 (100)	66.72 (100)	34.37 (98)	40.03 (100)	73.22 (43)	80.42 (68)
OLPP	78.67 (100)	84.24 (98)	53.08 (100)	59.81 (100)	25.81 (100)	30.95 (99)	66.89 (51)	75.81 (84)
NPE	93.50 (90)	96.25 (96)	77.24 (99)	83.60 (96)	46.51 (100)	52.65 (100)	65.00 (32)	79.35 (44)
LRLE	84.36 (100)	88.21 (100)	66.99 (100)	70.75 (100)	34.99 (100)	38.74 (100)	78.63 (81)	83.78 (98)
LRR-Graph	91.27 (100)	93.70 (100)	65.26 (100)	70.69 (100)	44.68 (100)	48.99 (100)	76.92 (47)	83.79 (55)
LRCA	**95.82** (**99**)	**97.52** (**100**)	**83.56** (**98**)	**90.63** (**100**)	**53.88** (**100**)	**62.87** (**100**)	**88.52** (**54**)	**92.70** (**55**)

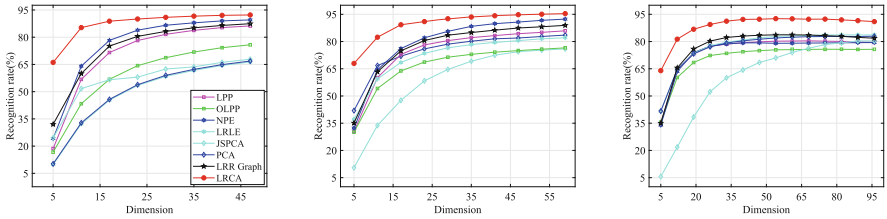


Fig. 2. Recognition on the (a) Extended Yale B (T = 30) (b) AR (T = 5) and (c) AR (10% salt-and-pepper noise, T = 7).

5.1 Datasets

The AR dataset [19] contains 2400 face images of 120 people. To test the robustness, all the face images are randomly corrupted by 10×10 block, 20×20 block occlusion and 10% salt-and-pepper noise, respectively.

The Extended Yale B dataset [8] contains 64 frontal face images per subject taken under different illumination conditions. Face images from 38 subjects are included, all the face images are randomly corrupted by 10×10 block and 20×20 block occlusion, respectively.

The FERET dataset [21] contains the 200 classes while there are only 7 samples for each person. All the face images are corrupted by gray image with random intensity generated from uniform distribution in [0, 100] and [0, 200], respectively.

Figure 1 show some sample images of these datasets.

Table 2. The performance (recognition accuracy (%), dimension) of different algorithms on FERET dataset with different intensity gray occlusion.

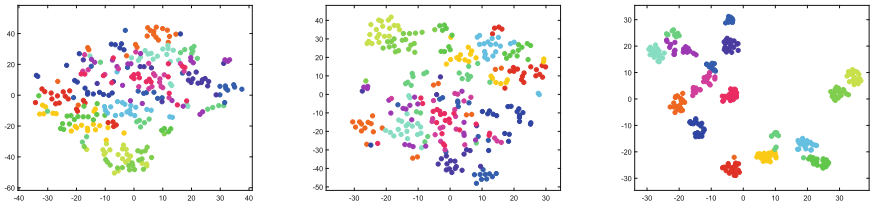
Intensity	PCA	JSPCA	LPP	OLPP	NPE	LRLE	LRR-Graph	**LRCA**
[0, 0]	53.87 (100)	56.23 (89)	55.05 (100)	54.50 (99)	51.90 (78)	53.10 (100)	61.23 (100)	**61.70 (70)**
[0, 100]	29.55 (100)	30.35 (85)	53.33 (100)	49.53 (99)	51.88 (52)	29.38 (90)	55.87 (94)	**61.43 (70)**
[0, 200]	21.83 (91	21.63 (97)	55.13 (98)	52.17 (100)	55.48 (66)	21.45 (100)	52.70 (98)	**59.07 (63)**

Table 3. The performance (recognition accuracy (%), dimension) of different algorithms on Extended Yale B dataset with different sizes of block occlusion.

Block size (T)	PCA	JSPCA	LPP	OLPP	NPE	LRLE	LRR-Graph	**LRCA**
0×0 (10)	55.40	53.18	74.96	54.62	82.44	55.15	79.17	**83.03**
	(100)	(63)	(100)	(100)	(97)	(100)	(100)	(**100**)
10×10 (10)	37.23	37.73	44.77	32.99	53.63	36.89	52.88	**58.20**
	(100)	(100)	(100)	(100)	(100)	(100)	(97)	(**100**)
20×20 (10)	15.23	17.06	18.22	14.19	21.14	14.65	20.26	**26.71**
	(100)	(60)	(100)	(100)	(97)	(100)	(100)	(**100**)
0×0 (20)	66.75	65.32	85.66	73.36	**89.35**	66.78	86.92	89.32
	(100)	(64)	(100)	(100)	(**100**)	(100)	(100)	(100)
10×10 (20)	46.84	47.80	59.63	48.41	70.60	46.88	64.10	**73.55**
	(100)	(64)	(100)	(100)	(100)	(100)	(100)	(**100**)
20×20 (20)	20.39	23.21	26.50	21.14	28.76	19.66	28.61	**36.02**
	(100)	(60)	(100)	(100)	(100)	(100)	(100)	(**100**)
0×0 (30)	74.61	73.00	89.37	81.93	91.84	74.50	90.43	**93.10**
	(100)	(64)	(100)	(100)	(100)	(100)	(100)	(**100**)
10×10 (30)	53.86	55.34	67.86	58.08	79.05	54.51	71.33	**83.80**
	(100)	(100)	(100)	(100)	(100)	(100)	(100)	(**100**)
20×20 (30)	23.73	27.33	32.03	24.95	33.84	22.76	33.53	**41.43**
	(99)	(60)	(100)	(100)	(100)	(100)	(100)	(**100**)

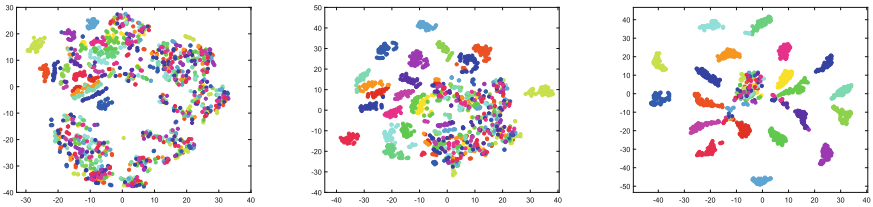
5.2 Experimental Settings

We randomly select T training samples per class and the rest samples for test. According to the different sizes of each dataset, we choose T for the AR, FERET and Extended Yale B datasets as $T = 5, 7$, $T = 4$ and $T = 10, 20, 30$, respectively. For simplicity, we directly set the balance parameter $\gamma = 1$ for the proposed LRCA. The dimensions of the low-dimensional subspace ranged from 1 to 100. The nearest neighbor classifier (1NN) with Euclidean distance is used to compute the average recognition rates of the learned representation on the test set. All algorithms are independently run 10 times on each dataset to provide reliable experimental results, and the average recognition rate is computed and reported.



(a) (b) (c)

Fig. 3. t-SNE representation of the first 20 dimensions obtained by (a) PCA, (b) LPP and (c) LRCA on AR dataset with 10×10 block ($T = 7$).



(a) (b) (c)

Fig. 4. t-SNE representation of the first 40 dimensions obtained by (a) PCA, (b) LPP and (c) LRCA on Extended Yale B dataset ($T = 30$).

5.3 Results and Discussions

Tables 1, 2 and 3 report the average recognition rate and the corresponding dimension of these datasets. As can be seen, LRCA achieves the best recognition rate in most cases. The key reason is that LRCA can uncover the multi-subspace structure, thus the face representation learned by the proposed LRCA is more discriminative.

LRCA is also significantly robust for different kinds of noises. When the datasets are corrupted by block occlusion, salt-and-pepper noise and illumination variation, the recognition rate of LRCA is higher than other methods. The key reason is that the proposed LRCA imposes the global low rank constraint on coefficient matrix, therefore, some corrupted data can be automatically corrected.

LRCA derives a more compact face representation. As can be seen in Fig. 2, the recognition rate curve of LRCA is steeper than the others, which means LRCA can learn a sufficiently discriminative face representation with relatively low dimensionality.

5.4 Visualization of Discriminative Ability

Figure 3 and Fig. 4 display the t-SNE representation [18] learned by PCA, LPP and LRCA, where each color denotes a subject of face images. Specifically, for

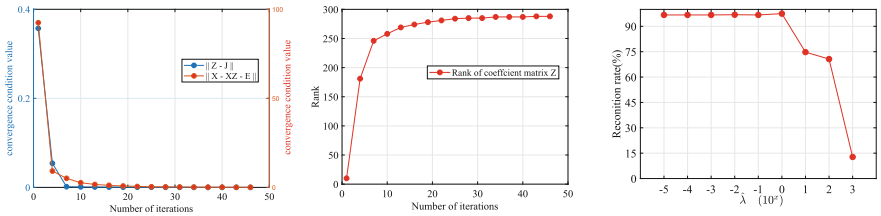


Fig. 5. The variation of (a) two convergence conditions of ADMM Algorithm 1, (b) the corresponding coefficient matrix $\mathbf{Z}$'s rank on Extend Yale B dataset. (c) Recognition rates versus the value of parameter $\hat{\lambda}$ on AR dataset.

corrupted AR dataset, we first project 15 subjects of face images to the first 20-dimensional subspace, and then visualize them on 2-dimensional spaces by t-SNE algorithm. For the Extended Yale B dataset, we project 20 subjects of face images to the first 40-dimensional subspace and then perform t-SNE algorithm. As can be seen, the face images of different classes are well separated and the intra-class distances are simultaneously compressed by the proposed LRCA. Therefore, LRCA can intuitively derive a more compact and discriminative face representation.

5.5 Convergence and Parameter Sensitivity Study

As the designed Algorithm 1 is convergent, we explore the convergence speed and the parameter sensitivity. Figure 5(a) and 5(b) show the convergent properties of the Algorithm 1 and the corresponding rank of the coefficient matrix $\mathbf{Z}$ on Extended Yale B dataset, respectively. As can be seen, the convergence speed of ADMM Algorithm 1 is very fast.

We set parameter λ as $\lambda = \hat{\lambda} E_\lambda$ to control the low rank property, in which $E_\lambda = \sum_{i=1}^{n} \|\mathbf{x}_i\|_2 / n$ is used to eliminate influence of dimension. From Fig. 5(c), when $\hat{\lambda} > 1$, the recognition rate declined rapidly. The reason is that the rank of the coefficient matrix $\mathbf{Z}$ is too low to capture the multi-subspace structure.

6 Conclusion

In this paper, we propose a novel unsupervised dimensionality reduction method. Since the proposed LRCA utilize LRR for robust subspace segmentation and preserve such a class-special structure, the robustness and discriminative ability of the embedding space can be greatly improved together. We design an iterative algorithm with convergence guarantee to optimize the objective function by using ADMM and eigendecomposition alternatively. Extensive experiments show the superior robustness and discriminative ability of the proposed LRCA.

Acknowledgement. This work was supported in part by the Natural Science Foundation of China under Grant 61976145, Grant 62076164 and Grant 61802267, in part by the Guangdong Basic and Applied Basic Research Foundation (No. 2021A1515011861), and in part by the Shenzhen Municipal Science and Technology Innovation Council under Grants JCYJ20180305124834854 and JCYJ20190813100801664.

References

1. Basri, R., Jacobs, D.W.: Lambertian reflectance and linear subspaces. IEEE Trans. Pattern Anal. Mach. Intell. **25**(2), 218–233 (2003)
2. Bengio, Y., Courville, A., Vincent, P.: Representation learning: a review and new perspectives. IEEE Trans. Pattern Anal. Mach. Intell. **35**(8), 1798–1828 (2013). https://doi.org/10.1109/TPAMI.2013.50
3. Boyd, S.P., Parikh, N., Chu, E., Peleato, B., Eckstein, J.: Distributed optimization and statistical learning via the alternating direction method of multipliers. Found. Trends Mach. Learn. **3**(1), 1–122 (2011)
4. Cai, D., He, X., Han, J., Zhang, H.: Orthogonal Laplacianfaces for face recognition. IEEE Trans. Image Process. **15**(11), 3608–3614 (2006)
5. Chan, K.H.R., Yu, Y., You, C., Qi, H., Wright, J., Ma, Y.: Deep networks from the principle of rate reduction. arXiv preprint arXiv:2010.14765 (2020)
6. Chen, Y., Lai, Z., Wong, W.K., Shen, L., Hu, Q.: Low-rank linear embedding for image recognition. IEEE Trans. Multimedia **20**(12), 3212–3222 (2018)
7. Elhamifar, E., Vidal, R.: Sparse subspace clustering: algorithm, theory, and applications. IEEE Trans. Pattern Anal. Mach. Intell. **35**(11), 2765–2781 (2013)
8. Georghiades, A.S., Belhumeur, P.N., Kriegman, D.J.: From few to many: illumination cone models for face recognition under variable lighting and pose. IEEE Trans. Pattern Anal. Mach. Intell. **23**(6), 643–660 (2001)
9. Härdle, W.K., Simar, L.: Canonical Correlation Analysis, pp. 443–454. Springer, Heidelberg (2015)
10. He, X., Cai, D., Yan, S., Zhang, H.J.: Neighborhood preserving embedding. In: IEEE International Conference on Computer Vision, vol. 2, pp. 1208–1213 (2005)
11. He, X., Niyogi, P.: Locality preserving projections. IEEE Trans. Knowl. Data Eng., 153–160 (2003)
12. Ji, P., Zhang, T., Li, H., Salzmann, M., Reid, I.D.: Deep subspace clustering networks. In: Conference on Neural Information Processing Systems, pp. 24–33 (2017)
13. Kheirandishfard, M., Zohrizadeh, F., Kamangar, F.: Deep low-rank subspace clustering. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 3776–3781. IEEE (2020)
14. Li, H., Jiang, T., Zhang, K.: Efficient and robust feature extraction by maximum margin criterion. In: Conference and Workshop on Neural Information Processing Systems, pp. 97–104. MIT Press (2003)
15. Li, J., Wu, Y., Zhao, J., Lu, K.: Low-rank discriminant embedding for multiview learning. IEEE Trans. Cybern. **47**(11), 3516–3529 (2017)
16. Liu, G., Lin, Z., Yan, S., Sun, J., Yu, Y., Ma, Y.: Robust recovery of subspace structures by low-rank representation. IEEE Trans. Pattern Anal. Mach. Intell. **35**(1), 171–184 (2013)
17. Liu, G., Lin, Z., Yu, Y.: Robust subspace segmentation by low-rank representation. In: International Conference on Machine Learning, pp. 663–670. Omnipress (2010)
18. van der Maaten, L., Hinton, G.: Visualizing data using t-SNE. J. Mach. Learn. Res. **9**(86), 2579–2605 (2008)

19. Martinez, A., Benavente, R.: The AR face database. CVC Technical Report 24 (1998)
20. Peng, X., Xiao, S., Feng, J., Yau, W., Yi, Z.: Deep subspace clustering with sparsity prior. In: International Joint Conferences on Artificial Intelligence, pp. 1925–1931. IJCAI/AAAI Press (2016)
21. Phillips, P.J., Moon, H., Rizvi, S.A., Rauss, P.J.: The FERET evaluation methodology for face-recognition algorithms. IEEE Trans. Pattern Anal. Mach. Intell. **22**(10), 1090–1104 (2000)
22. Sørensen, M., Sidiropoulos, N.D.: Multi-set low-rank factorizations with shared and unshared components. IEEE Trans. Sig. Process. **68**, 5122–5137 (2020)
23. Turk, M., Pentland, A.: Eigenfaces for recognition. J. Cogn. Neurosci. **3**(1), 71–86 (1991)
24. Vidal, R., Ma, Y., Sastry, S.: Generalized principal component analysis (GPCA). IEEE Trans. Pattern Anal. Mach. Intell. **27**(12), 1945–1959 (2005)
25. Wu, Z., Baek, C., You, C., Ma, Y.: Incremental learning via rate reduction. CoRR abs/2011.14593 (2020)
26. Yan, S., Xu, D., Zhang, B., Zhang, H., Yang, Q., Lin, S.: Graph embedding and extensions: a general framework for dimensionality reduction. IEEE Trans. Pattern Anal. Mach. Intell. **29**(1), 40–51 (2007)
27. Yi, S., Lai, Z., He, Z., Cheung, Y., Liu, Y.: Joint sparse principal component analysis. Pattern Recogn. **61**, 524–536 (2017)
28. Yu, Y., Chan, K.H.R., You, C., Song, C., Ma, Y.: Learning diverse and discriminative representations via the principle of maximal coding rate reduction. CoRR abs/2006.08558 (2020)
29. Zhang, J., et al.: Self-supervised convolutional subspace clustering network. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 5473–5482. Computer Vision Foundation/IEEE (2019)