SIAM J. IMAGING SCIENCES
Vol. 13, No. 3, pp. 1446–1466

A Discriminative Projection and Representation-Based Classification Framework for Face Recognition*

Kangkang Deng†, Zheng Peng‡, and Wenxing Zhu§

Abstract. The sparse representation-based classifier (SRC) has been developed and verified as having great potential for real-world face recognition. In this paper, we propose a discriminative projection and representation-based classification (DPRC) method to enhance the discriminant ability of the SRC. The proposed method first obtains a discriminative projection matrix not only maximizing the ratio of the distance within interclass over the distance within intraclass, but also minimizing the linear approximation error within intraclass. Then it maps the original data onto the discriminative space, and adopts an SRC method to obtain the final solution. An inexact augmented Lagrangian method of multiplier is proposed for finding the optimal representation vector in our framework, and a proximal alternating minimization method is adopted to the iteration subproblems of the proposed method. The proposed method is proven to have the subsequence convergence property. Experimental results on Yale, ORL, and AR face image databases demonstrate that, compared with some existing feature extraction methods based on the SRC, the proposed DPRC method is more efficient.

Key words. face recognition, sparse representation, discriminative projection, augmented Lagrangian method of multipliers, subsequence convergence

AMS subject classifications. 68A45, 68G10, 65K10, 90C50

DOI. 10.1137/19M1253873

1. Introduction and related works.

Face recognition has aroused considerable research interests in pattern recognition and computer vision areas during the last few decades [1, 25], and numerous methods were developed. Among these methods, representation-based classification (RC) has drawn intensive interests because of the noticeable performance.

The RC method proposed by Wright et al. [26] is indeed a sparse representation classification (SRC). Given a testing image, the SRC obtains a sparse coding over the whole training image set, then a classification method is performed by checking which class yields the least coding error. Specifically, the SRC utilizes the training samples of all classes to sparsely represent a testing sample by imposing ℓ_1-regularization. Zhang, Yang, and Feng [31] put forward a collaborative RC (CRC) by introducing an ℓ_2-regularization instead of the ℓ_1-regularization for efficient face recognition. Other regularization RC methods, including block sparse RC

*Received by the editors April 3, 2019; accepted for publication (in revised form) June 5, 2020; published electronically August 31, 2020.
https://doi.org/10.1137/19M1253873
Funding: The work of the authors was supported by the National Natural Science Foundation of China grant 11571074.

†College of Mathematics and Computer Science, Fuzhou University, Fuzhou 350108, China (freedeng@126.com).
‡Corresponding author. School of Mathematics and Computational Science, Xiangtan University, Xiangtan 411105, China (pzheng@xtu.edu.cn); College of Mathematics and Computer Science, Fuzhou University, Fuzhou 350108, China (pzheng@fzu.edu.cn).
§College of Mathematics and Computer Science, Fuzhou University, Fuzhou 350108, China (wxzhu@fzu.edu.cn).

(BSRC) [7], joint sparse RC [21], low rank RC [12], were also presented in literature. To improve the robustness of the SRC to noised circumstances, Qian et al. [17] proposed a novel robust low-rank regularized regression method, and for the other robust RC methods, the interested readers are also referred to [30, 7, 22] and [27]. In [24], the authors developed a unified framework termed atomic RC for some popular RC methods.

A key assumption of the RC methods is that, the training samples of a single class do lie in a subspace. However, this assumption is invalid if the samples in each class are insufficient or ill-conditioned. Such cases are common in practice, e.g., the samples in each class are few, or the samples within the same class do not have a good linear representation, or the samples within the different classes do not have a good exclusive effect in the linear representation view. To handle those cases, in the extended SRC (ESRC), Deng, Hu, and Guo [6] assumed that a testing image equals a prototype image plus some (linear) variations, and adopted an auxiliary intraclass variant dictionary to represent the possible variation between the training samples and testing image.

Many discriminative representation methods have also been proposed to achieve better performance for face recognition [29, 32, 9]. Yang et al. [28] proposed a sparse representation classifier steered discriminative projection method, which maximizes the ratio of between-class reconstruction residual to within-class reconstruction residual in the projected space. Linear discriminant regression classification (LDRC) [10] is a discriminant regression analysis method which embeds the Fisher criterion into the linear regression classification [16]. Fang et al. [8] combined the feature learning with classification to learn a robust latent subspace. A disadvantage of these mentioned methods is that, they did not full capture the discriminative information.

In this paper, we propose a discriminative projection and RC framework by taking full advantage of discriminant information. The proposed framework is divided into two stages: learning a discriminative projection and performing RC in the projection space. Our main contributions can be summarized as follows:

1. In the first stage, we obtain a latent representation by learning a discriminative projection, which can improve the class separability and the degree of linear reconstruction intraclass. In the second stage, we perform the SRC method [25] in the projection space. Our framework is compatible with the existing RC methods, and the SRC method in the second stage can be replaced by any other RC methods.
2. We propose an inexact augmented Lagrangian method of multipliers (ALM) to solve the optimization model in the first stage, and prove that the proposed inexact ALM algorithm has a subsequence convergence property.
3. We perform a lot of experiment on Yale, ORL, and AR face image databases, and demonstrate that the proposed discriminative projection and RC (DPRC) method is more efficient compared with some existing RC methods.

1.1. Notations. In the rest of this paper, for any matrix $X \in \mathbb{R}^{m\times n}$, we denote the Frobenius norm by $\|X\|_F$, and define $\|X\|_\infty := \max_{1\le i\le m, 1\le j\le n}\{|X_{i,j}|\}$. Let $\mathrm{diag}(Z)$ be a vector where the ith entry is Z_{ii}, and $\mathrm{Diag}([Z_1,\ldots,Z_m])$ be a block diagonal matrix, where the ith block matrix is Z_i. Let $f : \mathbb{R}^d \to (-\infty, +\infty]$ be a proper and lower semicontinuous function, the Fréchet subdifferential of f is defined as $\hat{\partial} f(x) := \{u | \lim_{y\neq x} \inf_{y\to x} \frac{f(y)-f(x)-<u,y-x>}{\|y-x\|} \ge 0\}$,

and the limiting subdifferential of f is defined as $\partial f(x) := \{u \in \mathbb{R}^d | \exists x^k \to x, f(x^k) \to f(x)$ and $u^k \in \hat{\partial} f(x^k) \to u$ as $k \to \infty\}$. Other notations will be defined when they occur.

1.2. Organization. The rest of this paper is arranged as follows. In section 2 we give a brief review on the RC methods. Then, section 3 presents the DPRC framework. In section 4, an inexact ALM algorithm is proposed to find the projection matrix in our framework, and we prove that the inexact ALM has a subsequence convergence property. Section 5 provides some numerical results to show that, compared with some state-of-the-art methods, our framework is more efficient. Section 6 concludes this paper with some final remarks.

2. Representation-based classification. Give a training set $X = [X_1, \ldots, X_m] \in \mathbb{R}^{r \times n}$, where the submatrix $X_i \in \mathbb{R}^{r \times n_i}$ represents the training samples in class i, and $n = \sum_{i=1}^{m} n_i$ is the sample size, m is the number of classes in the training sample set. Assume that the noise-free testing sample y will approximately lie in the linear span of the training samples with the same class label of y. Then, the RC finds the optimal representation vector via

$$\begin{cases} c^* & = \arg\min_{c \in \mathbb{R}^n} \Omega(c) \\ \text{s.t.} & y = Xc, \end{cases} \tag{2.1}$$

where $\Omega(c)$ is a regularization function. However, data may be noised in real-world applications. In that case, it is impossible to exactly express y as a sparse linear superposition of the training samples. For noisy data y, we consider the model

$$c^* = \arg\min_{c \in \mathbb{R}^n} \lambda \Omega(c) + \|y - Xc\|_2^2, \tag{2.2}$$

where λ is a regularization parameter. Different methods may use different regularization functions. In SRC [26], the authors considered an ℓ_1 sparse RC model, i.e., $\Omega(c) = \|c\|_1$. The BSRC [7] generalized SRC and took into account a block structure of training samples, and obtained a group sparse representation, where the samples with the same class label form a group. In this case, $\Omega(c) = \sum_{i=1}^{m} \|c_{\mathcal{I}_i}\|$, where $\mathcal{I}_i$ is the index set of class i. If $\Omega(c) = \|c\|_2^2$, the model (2.2) reduces to the CRC [31].

For a testing sample y, we first get the representation vector c^* via model (2.1) for clean data, or model (2.2) for noisy data. Then, we calculate the class-dependent residual for each class $r_k(y) = \|y - X\delta_k(c^*)\|_2, k = 1, \ldots, m$, where $\delta_k(c^*) \in R^n$ is a new vector whose nonzero entries are the components of c^* associated with the class k. The testing sample y is assigned to the class corresponding to the minimal residual.

3. The DRPC framework. In this section, we propose a DPRC framework for face recognition. The proposed framework has two stages. In the first stage, we learn a discriminative projection matrix that enhances simultaneously the class separability and the degree of linear reconstruction intraclass. In the second stage, the samples are projected into the discriminative space via the projection matrix learned in the first stage, and then the SRC [26] is adopted to the latent samples in the discriminative projection space. Figure 1 gives a conceptual illustration of discriminative projection.

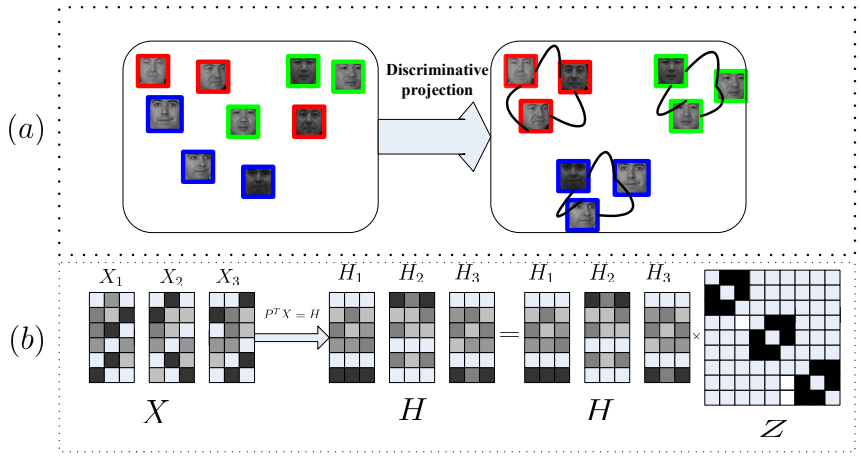

Figure 1. *Conceptual illustration for discriminative projection.* (a) *different colors denote different subjects. The left side is the image set in origin space, the right side is the image set in discriminative projection space.* (b) *An illustration in matrix form.*

3.1. Learning a discriminative projection. Given a training set $X = [X_1, \ldots, X_m] \in \mathbb{R}^{r\times n}$, our goal is to infer a latent sample space $H = [H_1, \ldots, H_m] \in \mathbb{R}^{d\times n}$ which preserves the better performance for some classical RC methods. Specifically, we seek a projection matrix $P \in \mathbb{R}^{r\times d}$ such that $P^TP = I_d$ and $P^TX = H$, where $P^TP = I_d$ is the property of the projection matrix; $P^TX = H$ means that H is a projection of X in the projection space via P. For this purpose, we have

$$\begin{cases} \min\limits_{P,H,Z} & \theta(H,Z) + \phi(P) \\ \text{s.t.} & P^TP = I_d,\ P^TX = H,\ \operatorname{diag}(Z) = 0, \end{cases} \tag{3.1}$$

where Z is the reconstruction coefficient matrix for the intraclass samples of data matrix H in the latent space, and $\theta(H,Z)$ is the corresponding reconstruction residual, $\phi(P)$ is a merit function for the class separability.

The linear reconstruction residual $\theta(H,Z)$ enforces that the intraclass samples in the latent space have a better linear reconstruction. Hence, for each H_i $(i = 1, 2, \ldots, m)$ we need to minimize

$$\|H_i - H_iZ_i\|_F^2,$$

where Z_i satisfying $\operatorname{diag}(Z_i) = 0$ is the regression coefficient of the ith sample H_i. To avoid the trivial solution, a regularization term is needed. By the regularization we obtain

$$\|H_i - H_iZ_i\|_F^2 + \eta\|Z_i\|_F^2.$$

In summary, the regularized linear reconstruction residual $\theta(H,Z)$ has the form

$$\theta(H,Z) = \sum_{i=1}^{m}\{\|H_i - H_iZ_i\|_F^2 + \eta\|Z_i\|_F^2\}.$$

Let $Z = \mathrm{Diag}([Z_1, \ldots, Z_m])$. Then, $\theta(H, Z)$ has a compact form:

$$\theta(H, Z) = \|H - HZ\|_F^2 + \eta\|Z\|_F^2, \tag{3.2}$$

where Z satisfies $\mathrm{diag}(Z) = 0$.

The role of $\phi(P)$ is to push the samples in different classes far away from each other, and pull those within the same class closeer to each other. Similarly to linear discriminant analysis [3], we give the between-class scatter matrix S_b and the within-class scatter matrix S_w as follows:

$$\begin{cases} S_b = \sum_{i=1}^m n_i(\mu^i - \mu)(\mu^i - \mu)^T, \\ S_w = \sum_{i=1}^m \sum_{j=1}^{n_i}(x_j^i - \mu^i)(x_j^i - \mu^i)^T, \end{cases} \tag{3.3}$$

where x_j^i is the jth sample in class i, μ^i is the sample mean of class i, and μ is the total sample mean. By (3.3), we have

$$\phi(P) = P^T(S_b - \nu S_w)P, \tag{3.4}$$

where $\nu > 0$ is a trade-off parameter.

In summary, the optimization model for finding the discriminative projection matrix P has the form

$$\begin{cases} \min\limits_{H,P,Z} & \|H - HZ\|_F^2 + \eta\|Z\|_F^2 + P^T(S_b - \nu S_w)P \\ \text{s.t.} & P^TP = I_d,\ P^TX = H, \mathrm{diag}(Z) = 0. \end{cases} \tag{3.5}$$

3.2. Performing the SRC method in the projection space. After obtaining the latent sample representation $H = P^TX$, we adopt the SRC [26] in the latent sample space. Specifically, given an unlabeled testing sample y, we first obtain the latent representation $u = P^Ty$. Then, we obtain the representation vector c^* in the latent space via

$$c^* = \arg\min_{c \in \mathbb{R}^n} \lambda\|c\|_1 + \|u - Hc\|_2^2, \tag{3.6}$$

and compute the residual $r_k(u) = \|u - H\delta_k(c^*)\|_2^2$. Finally, we predict the class of y via $\mathrm{identity}(y) = \arg\min_{k \in \mathcal{K}} r_k(z)$. Of course, we can also apply the other representation-based methods in this stage, and the SRC method is a better choice.

Algorithm 3.1 summarizes the proposed framework. First, we obtain a discriminative projection matrix P by solving problem (3.5). Then, the training samples and testing samples are projected onto a low-dimensional space (the discriminative space) via projection matrix P, and get their low-dimensional representations. Finally, the SRC [26] is adopted to the resulting low-dimensional representation data, and gives the classification result.

4. Optimization method and convergence. In this section, an inexact ALM algorithm is proposed to solve problem (3.5) for getting discriminative projection matrix P, and a proximal alternating minimization (PAM) method [2] is used to solve the subproblem in inexact ALM. The convergence of the proposed inexact ALM is established.

Algorithm 3.1 The discriminative projection and sparse RC framework, DPRC.

Input: A training set $X = [X_1, \ldots, X_m] \in \mathbb{R}^{r\times n}$ with m classes, an unlabeled testing sample $y \in \mathbb{R}^n$, regularization parameter $\lambda > 0$.

Output: Class-label of testing sample y

s1. Normalize the columns of X and y such that they have unit ℓ_2-norm.

s2. Obtain matrices P and H via solving problem (3.5).

s3. Compute the representation of training samples and testing sample in the latent space

$$H = P^T X, u = P^T y.$$

s4. Obtain the representation vector c^* for u with respect to H in the latent space, where

$$c^* = \arg\min_{c\in\mathbb{R}^n} \lambda\|c\|_1 + \|u - Hc\|_2^2.$$

s5. Compute residual

$$r_k(u) = \|u - H\delta_k(c^*)\|_2^2,\ k \in \mathcal{K} := \{1, \ldots, m\},$$

where $\delta_k(c^*) \in R^n$ is a vector whose nonzero entries are those of the components of c^* associated with class k.

s6. Predict the class of y via identity$(y) = \arg\min_{k\in\mathcal{K}} r_k(z)$.

4.1. The inexact ALM of multipliers.

For simplicity, let

$$\Phi(H, Z, P) := \|H - HZ\|_F^2 + \eta\|Z\|_F^2 + P^T(S_b - \nu S_w)P,$$

and $\delta_{\mathcal{M}}$ be an indicator function defined by

$$\delta_{\mathcal{M}}(x) = \begin{cases} 0 & \text{if } x \in \mathcal{M}, \\ +\infty & \text{otherwise,} \end{cases}$$

where $\mathcal{M} = \{J \in \mathbb{R}^{r\times d} | J^T J = I_d\}$ is a Stiefel manifold. Let $\mathcal{W} = \{Z \in \mathbb{R}^{n\times n} | \mathcal{A}Z = 0\}$, where $\mathcal{A}Z = \text{diag}(Z)$ is a linear operator. By introducing an auxiliary variable with constraint $J = P$, problem (3.5) can be rewritten as

$$\begin{cases} \min & \Phi(H, Z, P) + \delta_{\mathcal{M}}(J) + \delta_{\mathcal{W}}(Z) \\ \text{s.t.} & P^T X = H,\ P = J. \end{cases} \tag{4.1}$$

The augmented Lagrangian function associated with (4.1) is

$$\begin{aligned} \widetilde{L}_\rho(H, Z, P, J; \Lambda_1, \Lambda_2) = \Phi(H, Z, P) + \delta_{\mathcal{M}}(J) + \delta_{\mathcal{W}}(Z) + \langle \Lambda_1, P^T X - H\rangle \\ + \frac{\rho}{2}\|P^T X - H\|_F^2 + \langle \Lambda_2, P - J\rangle + \frac{\rho}{2}\|P - J\|_F^2, \end{aligned} \tag{4.2}$$

where $\Lambda_1 \in \mathbb{R}^{d\times n}, \Lambda_2 \in \mathbb{R}^{r\times d}$ are multipliers, and ρ is a penalty parameter. Inspired by [33], we consider the scaled form

$$L_\rho(H, Z, P, J; \Lambda_1, \Lambda_2) = \frac{1}{\rho}\tilde{L}_\rho(H, Z, P, J; \Lambda_1, \Lambda_2). \tag{4.3}$$

Then, at the kth iteration of the proposed inexact ALM, the jointed variable (H, Z, P, J) is first updated by fixing multipliers (Λ_1, Λ_2) and penalty parameter ρ, to obtain an approximate solution (H^k, Z^k, P^k, J^k) with tolerance ϵ_k. For this purpose, we solve the subproblem

$$(H^k, Z^k, P^k, J^k) := \arg \min_{J^T J = I} L_{\rho_{k-1}}(H, Z, P, J; \Lambda_1^{k-1}, \Lambda_2^{k-1}) \tag{4.4}$$

by an iteration method with a stopping criterion in the following:

$$A^k \in \partial L_{\rho_{k-1}}(H^k, Z^k, P^k, J^k; \Lambda_1^{k-1}, \Lambda_2^{k-1}) \ \text{ and } \|A^k\|_\infty \le \frac{\epsilon_{k-1}}{\rho_{k-1}}, \tag{4.5}$$

where $\epsilon_k \downarrow 0$ as $k \to \infty$. Then, we update the multipliers Λ_1, Λ_2 and penalty parameter ρ by a suitable mechanism. The outline of the inexact ALM is summarized in Algorithm 4.1.

Algorithm 4.1 The inexact ALM of multipliers for (4.1).

Input: Training sample matrix $X = [X_1, \ldots, X_m] \in \mathbb{R}^{r\times n}$, the dimensions of latent space d.
Output: Projection matrix $P \in \mathbb{R}^{r\times d}$.

Initialization: Given $\{Z^0, P^0, H^0, J^0; \Lambda_1^0, \Lambda_2^0\}$. Let $\{\epsilon_k\}_{k\in\mathbb{N}} \downarrow 0$, $\tau \in (0,1), \mu > 1, \rho_0 > 0, \varepsilon > 0$, $-\infty < \bar{\Lambda}_{i,min} \le \bar{\Lambda}_{i,max} < +\infty, i = 1, 2$.

While $k \ge 1$ **do**

s1. Obtain (H^k, Z^k, P^k, J^k) by solving subproblem (4.4) with termination criterion (4.5).

s2. Update the multipliers by

$$\Lambda_1^k = \bar{\Lambda}_1^{k-1} + \rho_{k-1} \cdot ((P^k)^T X - H^k),$$
$$\Lambda_2^k = \bar{\Lambda}_2^{k-1} + \rho_{k-1} \cdot (P^k - J^k),$$

where $\bar{\Lambda}_i^{k-1}$ is a projection of Λ_i^{k-1} on the set $\{\Lambda_i |\ \bar{\Lambda}_{i,min} \le \Lambda_i \le \bar{\Lambda}_{i,max}\}, i = 1, 2$.

s3. Update the penalty parameter by

$$\rho_k = \begin{cases} \rho_{k-1} & \text{if } \|R_i^k\|_\infty \le \tau \|R_i^{k-1}\|_\infty, i = 1, 2, \\ \mu\rho_{k-1} & \text{otherwise,} \end{cases}$$

where $R_1^k := (P^k)^T X - H^k, R_2^k := P^k - J^k$.

s4. If

$$\max_{i=1,2} \{\|R_i^k\|_\infty\} \le \varepsilon,$$

then stop. Otherwise, let $k := k + 1$ and goto s1.

Return: $P := P^k$

4.2. The PAM method for subproblem (4.4). The main computational cost of Algorithm 4.1 is in step s1, which finds an approximating minimizer of $L_{\rho_{k-1}}$ with respect to (H, Z, P, J) for fixed $(\Lambda_1^{k-1}, \Lambda_2^{k-1})$. We achieve it by adopting a PAM method. Let

$$G_k(H, Z, P, J) = \frac{1}{\rho_{k-1}}\Phi(H, Z, P) + \frac{1}{\rho_{k-1}}\left\langle \Lambda_1^{k-1}, P^T X - H \right\rangle + \frac{1}{2}\|P^T X - H\|_F^2$$
$$+ \frac{1}{\rho_{k-1}}\left\langle \Lambda_2^{k-1}, P - J \right\rangle + \frac{1}{2}\|P - J\|_F^2,$$

$F_k(J) = \frac{1}{\rho_{k-1}}\delta_{\mathcal{M}}(J)$ and $Q_k(Z) = \frac{1}{\rho_{k-1}}\delta_{\mathcal{W}}(Z)$. By (4.2) and (4.3) we have

$$L_{\rho_{k-1}}(H, Z, P, J; \Lambda_1^{k-1}, \Lambda_2^{k-1}) = G_k(H, Z, P, J) + F_k(J) + Q_k(Z).$$

Then, at the jth iteration of the PAM method for minimizing $L_{\rho_{k-1}}(H, Z, P, J; \Lambda_1^{k-1}, \Lambda_2^{k-1})$, the following minimization problems are solved:

$$\text{(4.6)} \quad \begin{cases} H^{k,j} = \arg\min\limits_{H} & G_k(H, Z^{k,j-1}, P^{k,j-1}, J^{k,j-1}) + \frac{s_1^{k,j-1}}{2}\|H - H^{k,j-1}\|_F^2, \\ Z^{k,j} = \arg\min\limits_{Z} & G_k(H^{k,j}, Z, P^{k,j-1}, J^{k,j-1}) + Q_k(Z) + \frac{s_2^{k,j-1}}{2}\|Z - Z^{k,j-1}\|_F^2, \\ P^{k,j} = \arg\min\limits_{P} & G_k(H^{k,j}, Z^{k,j}, P, J^{k,j-1}) + \frac{s_3^{k,j-1}}{2}\|P - P^{k,j-1}\|_F^2, \\ J^{k,j} = \arg\min\limits_{J^TJ=I_d} & G_k(H^{k,j}, Z^{k,j}, P^{k,j}, J) + F_k(J) + \frac{s_4^{k,j-1}}{2}\|J - J^{k,j-1}\|_F^2. \end{cases}$$

In iteration subproblems (4.6), the variables H, Z, P can be updated by one iteration of a gradient descent method. The subproblem with regards to J can be reformulated to

$$J^{k,j} = \arg\min_{J^TJ=I_d} \left\{ \frac{1}{\rho_{k-1}} \left\langle \Lambda_2^{k-1}, P^{k,j} - J \right\rangle + \frac{1}{2}\|P^{k,j} - J\|_F^2 + \frac{s_4^{k,j-1}}{2}\|J - J^{k,j-1}\|_F^2 \right\},$$

and it has a closed solution in the analytic form via a projection operator on the orthogonal constraint $J^TJ = I_d$. Algorithm 4.2 summarizes the PAM method for problem (4.4).

Remark 4.1. In Algorithm 4.2, the initial point is given by

$$(H^{k,0}, Z^{k,0}, P^{k,0}, J^{k,0}) = (H^{k-1}, Z^{k-1}, P^{k-1}, J^{k-1})$$

for all $k > 1$. To check stoping criterion (4.5), we set $A^k = [A_H^{k,j}, A_Z^{k,j}, A_P^{k,j}, A_J^{k,j}]$, where

$$\text{(4.7)} \quad \begin{cases} A_H^{k,j} = \nabla_H G_k(H^{k,j}, Z^{k,j}, P^{k,j}, J^{k,j}) - \nabla_H G_k(H^{k,j}, Z^{k,j-1}, P^{k,j-1}, J^{k,j-1}) \\ \qquad + s_1^{k,j-1}(H^{k,j-1} - H^{k,j}), \\ A_Z^{k,j} = \nabla_Z G_k(H^{k,j}, Z^{k,j}, P^{k,j}, J^{k,j}) - \nabla_Z G_k(H^{k,j}, Z^{k,j}, P^{k,j-1}, J^{k,j-1}) \\ \qquad + s_2^{k,j-1}(Z^{k,j-1} - Z^{k,j}), \\ A_P^{k,j} = \nabla_P G_k(H^{k,j}, Z^{k,j}, P^{k,j}, J^{k,j}) - \nabla_P G_k(H^{k,j}, Z^{k,j}, P^{k,j}, J^{k,j-1}) \\ \qquad + s_3^{k,j-1}(P^{k,j-1} - P^{k,j}), \\ A_J^{k,j} = s_4^{k,j-1}(J^{k,j-1} - J^{k,j}). \end{cases}$$

4.3. The convergence. Consider a general nonconvex and nonsmooth problem of the form

$$\text{(4.9)} \quad \min_{x,y} f(x) + g(y) + h(x, y),$$

where f and g are extended-real-valued (not necessary convex or smooth) functions, and h is a smooth (possibly nonconvex) function. Attouch, Bolte, and Svaiter [2] proposed a PAM

Algorithm 4.2 The PAM for subproblem (4.4).

Input: $(H^{k,0}, Z^{k,0}, P^{k,0}, J^{k,0}), \epsilon_{k-1}, \rho_{k-1}$.
Output: (H^k, Z^k, P^k, J^k)

s1. Let $j = 0$.
s2. Compute $A^{k,j}$ by (4.7).
s3. If

$$\|A^{k,j}\|_\infty \geq \epsilon_{k-1}/\rho_{k-1}, \tag{4.8}$$

then update $(H^{k,j}, Z^{k,j}, P^{k,j}, J^{k,j})$ via one iteration for problem (4.6).
s4. Let $j := j + 1$, go to s2.

Return: $(H^k, Z^k, P^k, J^k) = (H^{k,j}, Z^{k,j}, P^{k,j}, J^{k,j})$

method for problem (4.9). Given a pair (x^k, y^k), the PAM method updates x and y alternately by

$$\begin{cases} x^{k+1} & \in \arg\min_x f(x) + h(x, y^k) + \dfrac{c_k}{2}\|x - x^k\|^2, \\ y^{k+1} & \in \arg\min_y g(y) + h(x^{k+1}, y) + \dfrac{d_k}{2}\|y - y^k\|^2. \end{cases} \tag{4.10}$$

Under some suitable assumptions, Attouch, Bolte, and Svaiter [2] proved that each bounded sequence generated by the PAM method converges to a critical point of problem (4.9). The convergence still holds for the extension of the PAM method to solve more general settings involving $p > 2$ blocks. Specifically, in the case that the objective function has p-blocks variables of the form

$$\psi(x_1, \ldots, x_p) := h(x_1, \ldots, x_p) + \sum_{i=1}^{p} f_i(x_i) \quad , x_i \in \mathbb{R}^{n_i}, \tag{4.11}$$

where $h : \mathbb{R}^{n_1} \times \cdots \times \mathbb{R}^{n_p} \to \mathbb{R}$ is continuously differentiable, each $f_i :: \mathbb{R}^{n_1} \to \mathbb{R}(i = 1, \ldots, p)$ is a proper and low-semicontinuous function, the PAM method solves at each iteration the p subproblems

$$x_i^{k+1} \in \arg\min_{x_i} \left\{ h(x_1^{k+1}, \ldots, x_{i-1}^{k+1}, x_i, x_i^k, \ldots, x_p^k) + f_i(x_i) + \frac{\rho_i^k}{2}\|x_i - x_i^k\|^2 \right\} \tag{4.12}$$

for $i = 1, \ldots, p$.

For the convergence of the PAM method (4.12), the Kurdyka–Łojasiewicz property plays a critical role.

Definition 4.2 (see Kurdyka–Łojasiewicz property [19]). *Let $\sigma : \mathbb{R}^d \to (-\infty, +\infty]$ be proper and lower semicontinuous. The function σ is said to have the* Kurdyka–Łojasiewicz (K-L) property *at $\bar{u} \in dom(\partial\sigma) := \{u \in \mathbb{R}^d | \partial\sigma(u) \neq \emptyset\}$, if there exist $\eta \in [0, +\infty)$, a neighborhood U of $\bar{u}$, and a concave function $\varphi : [0, \eta) \to \mathbb{R}_+$ satisfies $\varphi(0) = 0$, φ is continuously differentiable on $(0, \eta)$ and continuous at 0, and $\varphi'(s) > 0$ for all $s \in (0, \eta)$, such that for all*

$$x \in U \cap \{x | \sigma(\bar{x}) < \sigma(x) < \sigma(\bar{x}) + \eta\} \tag{4.13}$$

the following inequality holds:

$$\varphi'\Big(\sigma(x)-\sigma(\bar{x})dist(0,\partial\sigma(x))\Big)\geq 1. \tag{4.14}$$

If σ satisfies the K-L property at each point of $dom(\partial\sigma)$, then σ is called a K-L function.

To guarantee global convergence of the PAM method, we need the following assumptions.

Assumption 4.1.

(i) For each $i=1,\ldots,p$, f_i is a proper and low-semicontinuous function, and $h\in C^1$ has locally Lipschitz continuous gradient.

(ii) the sequence $\{\rho_i^k : k\in\mathbb{N}\}_{i=1,\ldots,p}$ is bounded, and $\inf_{\mathbb{R}^{n_1}\times\cdots\times\mathbb{R}^{n_p}} f > -\infty$.

Under Assumption 4.1, the convergence of the PAM method is established in [2].

Theorem 4.3 (see [2, Theorem 6.2]). *Let $\{z^k := (x_1^k,\ldots,x_p^k)\}_{k\in\mathbb{N}}$ be a bounded sequence generated by the PAM method (4.12), and suppose that* Assumption 4.1 *holds. Then we have*

(i) *the sequence $\{z^k\}_{k\in\mathbb{N}}$ has finite length, i.e.,*

$$\sum_{i=1}^{\infty}\|z^{k+1}-z^k\|<\infty;$$

(ii) *if ψ is a K-L function, then the sequence $\{z^k\}_{k\in\mathbb{N}}$ converges to a critical point of ψ.*

The following assertions will be utilized in the convergence analysis.

Proposition 4.4.

(1) *For the limiting subdifferential of indicator function $\delta_{\mathscr{X}}$, where $\mathscr{X}$ is a closed set, we have* [15]

$$\partial\delta_{\mathscr{X}}(x)=\mathcal{N}_{\mathscr{X}}(x),$$

where $N_{\mathscr{X}}(x)$ is the normal cone of $\mathscr{X}$ at x.

(2) *Particularly, let $\mathcal{A}:\mathbb{R}^{r\times n}\to\mathbb{R}^{d\times n}$ be a linear mapping. If $\mathscr{X}_1:=\{X|\ \mathcal{A}X=0\}$ and $\mathscr{X}_2:=\{X|\ X^TX=I\}$, then*

$$\begin{aligned}\mathcal{N}_{\mathscr{X}_1}(X)&=\{\mathcal{A}^*Y|\ Y\in\mathbb{R}^{d\times n}\},\\ \mathcal{N}_{\mathscr{X}_2}(X)&=\{XS|\ S=S^T\}.\end{aligned} \tag{4.15}$$

Essentially, Algorithm 4.2 is the PAM method (4.12) adopted to iteration subproblem (4.4) and, correspondingly,

$$h(H,Z,P,J)=G_k(H,Z,P,J),\quad f_1(J)=F_k(J),\quad f_2(Z)=Q_k(Z).$$

Hence, similarly to the convergence analysis in [5], the sequence generated by Algorithm 4.2 is strongly convergent.

Theorem 4.5. *Let $\{(H^{k,j},Z^{k,j},P^{k,j},J^{k,j})\}_{j\in\mathbb{N}}$ be a sequence generated by* Algorithm 4.2 *for fixed $k\geq 1$, and choose the parameters $s_i^{k,j}(i=1,\ldots,4)$ such that $\{s_i^{k,j} : j\in\mathbb{N}\}$ is bounded. Then we have the following.*

(1) A^k *defined by* (4.7) *satisfies that for all* $j \in \mathbb{N}$,

$$A^k \in \partial L_{\rho_{k-1}}(H^{k,j}, Z^{k,j}, P^{k,j}, J^{k,j}; \Lambda_1^{k-1}, \Lambda_2^{k-1})$$

and

$$\|A^k\|_\infty \to 0 \text{ as } j \to \infty.$$

(2) *The sequence* $\{W^{k,j} := (H^{k,j}, Z^{k,j}, P^{k,j}, J^{k,j})\}_{j\in\mathbb{N}}$ *has finite length, i.e.,*

$$\sum_{j=1}^{\infty} \left\| (H^{k,j+1}, Z^{k,j+1}, P^{k,j+1}, J^{k,j+1}) - (H^{k,j}, Z^{k,j}, P^{k,j}, J^{k,j}) \right\| \le \infty. \tag{4.16}$$

(3) *Any accumulation point of the sequence* $\{W^{k,j}\}_{j\in\mathbb{N}}$, *denoted by* $W^{k,*}$, *is a critical point of function* $L_{\rho_{k-1}}(H, Z, P, J; \Lambda_1^{k-1}, \Lambda_2^{k-1})$.

Proof. For simplicity, let $W := (H, Z, P, J)$ and $L_k(W) := L_{\rho_{k-1}}(W; \Lambda_1^{k-1}, \Lambda_2^{k-1})$.

(i) Given $W^{k,j-1}$, by the first-order optimality condition for the subproblems of the PAM method we have

$$\begin{cases} \nabla_H G_k(H^{k,j}, Z^{k,j-1}, P^{k,j-1}, J^{k,j-1}) + s_1^{k,j}(H^{k,j} - H^{k,j-1}) = 0, \\ \omega^{k,j} + \nabla_Z G_k(H^{k,j}, Z^{k,j}, P^{k,j-1}, J^{k,j-1}) + s_2^{k,j}(Z^{k,j} - Z^{k,j-1}) = 0, \\ \nabla_P G_k(H^{k,j}, Z^{k,j}, P^{k,j}, J^{k,j-1}) + s_3^{k,j}(P^{k,j} - P^{k,j-1}) = 0, \\ \nu^{k,j} - \Lambda_2^{k-1}/\rho_{k-1} + (P^{k,j} - J^{k,j}) + s_4^{k,j}(J^{k,j} - J^{k,j-1}) = 0, \end{cases}$$

where $\omega^{k,j} \in \partial Q_k(Z^{k,j})$ and $\nu^{k,j} \in \partial F_k(J^{k,j})$. It follows that

$$\begin{cases} A_H^{k,j} = \nabla_H G_k(H^{k,j}, Z^{k,j}, P^{k,j}, J^{k,j}) \in \partial_H L_k(W^{k,j}), \\ A_Z^{k,j} = \omega^{k,j} + \nabla_Z G_k(H^{k,j}, Z^{k,j}, P^{k,j}, J^{k,j}) \in \partial_Z L_k(W^{k,j}), \\ A_P^{k,j} = \nabla_P G_k(H^{k,j}, Z^{k,j}, P^{k,j}, J^{k,j}) \in \partial_P L_k(W^{k,j}), \\ A_J^{k,j} = \nu^{k,j} + \nabla_J G_k(H^{k,j}, Z^{k,j}, P^{k,j}, J^{k,j}) \in \partial_J L_k(W^{k,j}). \end{cases}$$

By the subdifferentiability [15], we get

$$\partial L_k(W) = \partial_H L_k(W^{k,j}) \times \partial_Z L_k(W^{k,j}) \times \partial_P L_k(W^{k,j}) \times \partial_J L_k(W^{k,j}),$$

which implies that

$$A^k \in \partial L_k(W^{k,j}).$$

Now we need only to verify that, for each fixed $k \in \mathbb{N}$, $\|A^k\|_\infty \to 0$ as $j \to \infty$. We achieve this by showing that the function $L_k(W)$ satisfies the conditions of Theorem 4.3, hence it is available.

First, $L_k(W)$ satisfies Assumption 4.1. To see this, for given Λ^{k-1} and ρ_{k-1}, one can check that

(a) $F_k(J)$ is the indicator function of the Stiefel manifold, which is a proper and low-semicontinuous function satisfying $\inf F_k(J) > -\infty$, $Q_k(J)$ is the indicator function of the convex set (linear constraint set), which also is a proper and low-semicontinuous function satisfying $\inf F_k(J) > -\infty$, and G_k is a C^1 function. Hence, Assumption 4.1(i) holds;

(b) since G_k is a quadratic function with respect to (H, Z, P, J), and the sequences $\{s_i^{k,j} : j \in \mathbb{N}\}$ are bounded and fixed $k \geq 1$, where $i = 1, 2, 3, 4$. Assumption 4.1(ii) holds.

Second, for all index $k \geq 1$, the sequence $\{W^{k,j}\}_{j\in\mathbb{N}}$ is bounded. Otherwise, suppose by contradiction that $\lim_{j\to\infty} \|W^{k,j}\|_\infty = +\infty$. Then, on the one hand, since $\{\rho_k\}_{k\in\mathbb{N}}$ is nondecreasing, and $L_k(W)$ is a coercive function, it holds that $\lim_{j\to\infty} L_k(W^{k,j}) = +\infty$; on the other hand, by [19, Lemma 3], it follows that

$$L_k(W^{k,j}) \leq L_k(W^{k,j-1}) \leq L_k(W^{k,0})$$

which leads to a contradiction.

(ii) Follows from (i), $\{W^{k,j}\}_{j\in\mathbb{N}}$ is bounded. Recall that

$$L_k(W) = \Phi(H, Z, P) + \delta_{\mathcal{M}}(J) + \delta_{\mathcal{W}}(Z) + \left\langle \Lambda_1^{k-1}, P^T X - H \right\rangle + \frac{\rho_{k-1}}{2}\|P^T X - H\|_F^2 + \left\langle \Lambda_2^{k-1}, P - J \right\rangle + \frac{\rho_{k-1}}{2}\|P - J\|_F^2.$$

Since polynomial functions are semialgebraic, and a finite sum of semialgebraic functions is also a semialgebraic function, we have that $L_k(W)$ is a semialgebraic function. In summary, all conditions of Theorem 4.3 are satisfied. The assertion of Theorem 4.5 follows directly from Theorem 4.3 and the proof is completed. ∎

Lemma 4.6. *Suppose that (H^*, Z^*, P^*, J^*) is a local minimizer of problem (4.1). Then, there exist $(\Lambda_1^*, \Lambda_2^*, \Lambda_3^*)$ with proper dimension such that*

$$\begin{pmatrix} (\partial_H \Phi(H^*, Z^*, P^*))^T \\ \partial\delta_{\mathcal{W}}(Z^*) + \partial_Z \Phi(H^*, Z^*, P^*) \\ \partial_P \Phi(H^*, Z^*, P^*) \\ 0 \end{pmatrix} + \begin{pmatrix} -I & 0 & 0 \\ 0 & 0 & 0 \\ X & I & 0 \\ 0 & -I & 2J^* \end{pmatrix} \begin{pmatrix} (\Lambda_1^*)^T \\ \Lambda_2^* \\ \Lambda_3^* \end{pmatrix} = 0 \tag{4.17}$$

and

$$(P^*)^T X - H^* = 0,\ P^* = J^*,\ (J^*)^T J^* = I_d.$$

Correspondingly, we have

$$\begin{pmatrix} (\partial_H \Phi(H^*, Z^*, P^*))^T \\ \partial\delta_{\mathcal{W}}(Z^*) + \partial_Z \Phi(H^*, Z^*, P^*) \\ \partial_P \Phi(H^*, Z^*, P^*) \end{pmatrix} + \begin{pmatrix} -I & 0 \\ 0 & 0 \\ X & 2P^* \end{pmatrix} \begin{pmatrix} (\Lambda_1^*)^T \\ \Lambda_3^* \end{pmatrix} = 0. \tag{4.18}$$

Proof. Since (H^*, Z^*, P^*, J^*) is a local minimizer, it is feasible. It is clear that

$$(P^*)^T X - H^* = 0,\ P^* = J^*,\ (J^*)^T J^* = I_d. \tag{4.19}$$

For convenience, let $W := (H^T; Z; P; J)$ and

$$g(W) = \begin{pmatrix} -I & 0 & X^T & 0 \\ 0 & 0 & I & -I \end{pmatrix} W.$$

Letting $\Omega = \{W|\ g(W) = 0\}$, problem (4.1) is equivalent to

$$\min_{W}\ \Phi(H, Z, P) + \delta_{\mathcal{M}}(J) + \delta_{\Omega}(W) + \delta_{\mathcal{W}}(Z). \tag{4.20}$$

Hence, by the generalized Fermat's rule and subdifferentiability property, we get

$$0 \in \begin{pmatrix} (\partial_H \Phi(H^*, Z^*, P^*))^T \\ \partial_Z \Phi(H^*, Z^*, P^*) \\ \partial_P \Phi(H^*, Z^*, P^*) \\ \partial_J \delta_{\mathcal{M}}(J^*) \end{pmatrix} + \partial\delta_{\Omega}(W^*) + \partial\delta_{\mathcal{W}}(Z^*).$$

It follows from (4.15) that

$$\partial\delta_{\Omega}(W^*) = \mathcal{N}_{\Omega}(W^*) = \left\{ \begin{pmatrix} -I & 0 \\ 0 & 0 \\ X & I \\ 0 & -I \end{pmatrix} \begin{pmatrix} \Lambda_1^T \\ \Lambda_2 \end{pmatrix} \Big| \Lambda_1 \in \mathbb{R}^{d\times n}, \Lambda_2 \in \mathbb{R}^{r\times n} \right\} \tag{4.21}$$

and

$$\partial_J \delta_{\mathcal{M}}(J^*) = \mathcal{N}_{\mathcal{M}}(J^*) = \{J^* S|\ S = S^T\}. \tag{4.22}$$

Hence, there are Λ_1^*, Λ_2^*, and $\Lambda_3^* \in \{S|\ S = S^T\}$ such that

$$\begin{aligned} 0 &\in \begin{pmatrix} (\partial_H \Phi(H^*, Z^*, P^*))^T \\ \partial\delta_{\mathcal{W}}(Z^*) + \partial_Z \Phi(H^*, Z^*, P^*) \\ \partial_P \Phi(H^*, Z^*, P^*) \\ 2J^*\Lambda_3^* \end{pmatrix} + \begin{pmatrix} -I & 0 \\ 0 & 0 \\ X & I \\ 0 & -I \end{pmatrix} \begin{pmatrix} (\Lambda_1^*)^T \\ \Lambda_2^* \end{pmatrix} \\ &= \begin{pmatrix} (\partial_H \Phi(H^*, Z^*, P^*))^T \\ \partial\delta_{\mathcal{W}}(Z^*) + \partial_Z \Phi(H^*, Z^*, P^*) \\ \partial_P \Phi(H^*, Z^*, P^*) \\ 0 \end{pmatrix} + \begin{pmatrix} -I & 0 & 0 \\ 0 & 0 & 0 \\ X & I & 0 \\ 0 & -I & 2J^* \end{pmatrix} \begin{pmatrix} (\Lambda_1^*)^T \\ \Lambda_2^* \\ \Lambda_3^* \end{pmatrix} \end{aligned} \tag{4.23}$$

which proves (4.17). Moreover, it yields that $\Lambda_2^* = 2J^*\Lambda_3^*$. Substituting it into (4.23), we get (4.18) and complete the proof. ∎

By Lemma 4.6, we have the following convergence theorem.

Theorem 4.7. *Suppose that* Assumption 4.1 *holds, and* $\{(H^k, Z^k, P^k, J^k)\}_{k\in\mathbb{N}}$ *is a sequence generated by* Algorithm 4.1. *If* $\{(H^k, Z^k, P^k, J^k)\}_{k\in\mathbb{N}}$ *is bounded, then any accumulation point of the sequence* $\{(H^k, Z^k, P^k, J^k)\}_{k\in\mathbb{N}}$, *denoted by* (H^*, Z^*, P^*, J^*), *is the first-order critical point of problem* (4.1) *and, correspondingly,* (H^*, Z^*, P^*) *is the first-order critical point of problem* (3.5).

Proof. For any accumulation point (H^*, Z^*, P^*, J^*) of sequence $\{(H^k, Z^k, P^k, J^k)\}_{k\in\mathbb{N}}$ generated by the proposed method, there exists a subsequence $\{(H^k, Z^k, P^k, J^k)\}_{k\in\mathcal{K}}$ converging to (H^*, Z^*, P^*, J^*).

To prove that (H^*, Z^*, P^*, J^*) is the first-order critical point, we first show it is feasible. The feasibility condition $(J^*)^T J^* = I_d$ is trivial since $(J^k)^T J^k = I_d$ holds for all $k \in \mathbb{N}$.

(i) If $\{\rho_k\}$ is bounded, then by the updating rule of ρ_k in Algorithm 4.1, there exists a $k_0 \in \mathbb{N}$ such that

$$\|R_j^k\|_\infty \le \tau \|R_j^{k-1}\|_\infty \ \forall k \ge k_0, j = 1, 2.$$

By the definition of R_j^k, it follows that

$$\begin{cases} \|(P^k)^T X - H^k\|_\infty \le \tau \|(P^{k-1})^T X - H^{k-1}\|_\infty, \\ \|P^k - J^k\|_\infty \le \tau \|P^{k-1} - J^{k-1}\|_\infty \end{cases} \tag{4.24}$$

for all $k \ge k_0$. Taking the limit as $k \to \infty$ on both sides of (4.24), we get

$$\begin{cases} (P^*)^T X - H^* = 0, \\ P^* = J^*. \end{cases} \tag{4.25}$$

(ii) If $\{\rho_k\}$ is unbounded, by the generalized Fermat's rule, finding a solution satisfying the condition (4.8) is equivalent to calculating a point (H^k, Z^k, P^k, J^k) such that

$$\begin{aligned} &\left\| \frac{1}{\rho_{k-1}} \begin{pmatrix} (\partial_H \Phi(H^k, Z^k, P^k))^T \\ \partial \delta_{\mathcal{W}}(Z^k) + \partial_Z \Phi(H^k, Z^k, P^k) \\ \partial_P \Phi(H^k, Z^k, P^k) \\ 0 \end{pmatrix} \right. \\ &\quad \left. + \begin{pmatrix} -I & 0 & 0 \\ 0 & 0 & 0 \\ X & I & 0 \\ 0 & -I & 2J^k \end{pmatrix} \begin{pmatrix} ((P^k)^T X - H^k) + \frac{1}{\rho_{k-1}} \bar{\Lambda}_1^{k-1})^T \\ (P^k - J^k) + \frac{1}{\rho_{k-1}} \bar{\Lambda}_2^{k-1}) \\ \frac{1}{\rho_{k-1}} \Lambda_3^k \end{pmatrix} \right\|_\infty \\ &\le \frac{\epsilon_{k-1}}{\rho_{k-1}}, \end{aligned}$$

where $\lim_{k\to\infty} \epsilon_k = 0$. Notice that $\{\bar{\Lambda}_1^k\}$ and $\{\bar{\Lambda}_2^k\}$ are bounded; it is easy to verify that $\{\partial \delta_{\mathcal{W}}(Z^k)\}$, $\{\partial_H \Phi(H^k, Z^k, P^k)\}$, $\{\partial_Z \Phi(H^k, Z^k, P^k)\}$, and $\{\partial_P \Phi(H^k, Z^k, P^k)\}$ are also bounded. Hence we have a convergent subsequence. Letting $k \in \mathcal{K} \subset \mathbb{N}$ be the index of the convergent subsequence and $k \to \infty$, it follows from the above inequality that

$$\begin{pmatrix} -I & 0 \\ X & I \end{pmatrix} \begin{pmatrix} ((P^*)^T X - H^*)^T \\ P^* - J^* \end{pmatrix} = 0. \tag{4.26}$$

In summary, (H^*, Z^*, P^*, J^*) is feasible.

Second we prove that there exist Λ_1^*, Λ_2^*, and Λ_3^* such that $(H^*, Z^*, P^*, J^*; \Lambda_1^*, \Lambda_2^*, \Lambda_3^*)$ satisfies the first-order critical condition (4.17).

Since $\{(H^k, Z^k, P^k, J^k)\}_{k\in\mathbb{N}}$ is bounded, it converges. There exists an index subset $\mathcal{K} \subset \mathbb{N}$ such that

$$\lim_{k\in\mathcal{K}, k\to\infty} (H^k, Z^k, P^k, J^k) = (H^*, Z^*, P^*, J^*).$$

Combining this with the updating rule of Λ_1^k and Λ_2^k in Algorithm 4.1, we have from (4.26) that there exists a ξ^k such that $\left\|\xi^k\right\|_\infty \leq \frac{\epsilon_{k-1}}{\rho_{k-1}}$, where

$$\xi^k \in \begin{pmatrix} (\partial_H \Phi(H^k, Z^k, P^k))^T \\ \partial \delta_{\mathcal{W}}(Z^k) + \partial_Z \Phi(H^k, Z^k, P^k) \\ \partial_P \Phi(H^k, Z^k, P^k) \\ 0 \end{pmatrix} / \rho_{k-1} + \begin{pmatrix} -I & 0 & 0 \\ 0 & 0 & 0 \\ X & I & 0 \\ 0 & -I & 2J^k \end{pmatrix} \begin{pmatrix} (\Lambda_1^k)^T \\ \Lambda_2^k \\ \Lambda_3^k \end{pmatrix} / \rho_{k-1}. \tag{4.27}$$

Let

$$\Xi^k := \begin{pmatrix} -I & 0 & 0 \\ 0 & 0 & 0 \\ X & I & 0 \\ 0 & -I & 2J^k \end{pmatrix}, \ \Gamma^k := \begin{pmatrix} (\Lambda_1^k)^T \\ \Lambda_2^k \\ \Lambda_3^k \end{pmatrix},$$

then

$$\Xi^k \Gamma^k = \rho_{k-1} \xi^k - \begin{pmatrix} (\partial_H \Phi(H^k, Z^k, P^k))^T \\ \partial \delta_{\mathcal{W}}(Z^k) + \partial_Z \Phi(H^k, Z^k, P^k) \\ \partial_P \Phi(H^k, Z^k, P^k) \\ 0 \end{pmatrix}. \tag{4.28}$$

Since Ξ^k is full column rank, $(\Xi^k)^T \Xi^k$ is nonsingular. By (4.28) we get

$$\Gamma^k = \left((\Xi^k)^T \Xi^k\right)^{-1} (\Xi^k)^T \left[\rho_{k-1} \xi^k - \begin{pmatrix} (\partial_H \Phi(H^k, Z^k, P^k))^T \\ \partial \delta_{\mathcal{W}}(Z^k) + \partial_Z \Phi(H^k, Z^k, P^k) \\ \partial_P \Phi(H^k, Z^k, P^k) \\ 0 \end{pmatrix} \right]. \tag{4.29}$$

Taking the limit on (4.29) as $k \in \mathcal{K}$ is trending to ∞, and utilizing $\|\xi^k\|_\infty \leq \frac{\epsilon_{k-1}}{\rho_{k-1}}$ and $\lim_{k\to\infty} \epsilon_k = 0$, we have

$$\Gamma^* := \lim_{k \in \mathcal{I}, k \to \infty} \Gamma^k = -\left((\Xi^*)^T \Xi^*\right)^{-1} (\Xi^*)^T \begin{pmatrix} (\partial_H \Phi(H^*, Z^*, P^*))^T \\ \partial \delta_{\mathcal{W}}(Z^*) + \partial_Z \Phi(H^*, Z^*, P^*) \\ \partial_P \Phi(H^*, Z^*, P^*) \\ 0 \end{pmatrix}, \tag{4.30}$$

where

$$\Xi^* = \begin{pmatrix} -I & 0 & 0 \\ 0 & 0 & 0 \\ X & I & 0 \\ 0 & -I & 2J^* \end{pmatrix}$$

is full column rank. It follows that

$$\begin{pmatrix} (\partial_H \Phi(H^*, Z^*, P^*))^T \\ \partial \delta_{\mathcal{W}}(Z^*) + \partial_Z \Phi(H^*, Z^*, P^*) \\ \partial_P \Phi(H^*, Z^*, P^*) \\ 0 \end{pmatrix} + \begin{pmatrix} -I & 0 & 0 \\ X & I & 0 \\ 0 & -I & 2J^* \end{pmatrix} \begin{pmatrix} (\Lambda_1^*)^T \\ \Lambda_2^* \\ \Lambda_3^* \end{pmatrix} = 0.$$

Figure 2. *Sample images in the Yale database with variations.*

The last equation implies that, (H^*, Z^*, P^*, J^*) is a critical point of problem (4.1). Correspondingly, by Lemma 4.6, (H^*, Z^*, P^*) is a critical point of problem (3.5). ∎

5. Experiments. In this section, we present some experimental results on publicly available databases to demonstrate the efficiency of the proposed framework. Three face database sets, including Yale [3], ORL [20], and AR [13], are used in our experiments. For each database, we adopt k-fold cross validation to obtain a training set and a testing set. The proposed framework in this paper (DPRC) is compared with some state-of-the-art RC methods; they are SRC [26], CRC [31], LRC [16], ESRC [6], LDRC [10], linear collaborative discriminant regression classification (LCDRC) [18], SR-SLR [11], the least squares regression method (LSR) [23], and robust low-rank regularized regression (RLR^3) [17]. We list the recognition rate (in percentage) obtained by all methods for comparisons. The recognition rate is a rate of the number of the test samples being correctly identified to total test samples.

5.1. Results on the Yale database. The Yale database consists of 165 32×32 pixel cropped grayscale face images in GIF format of 15 individuals. There are 11 images per person, one per different facial expression or configuration: center-light, w/glasses, happy; left-light, w/no glasses, normal; right-light, sad, sleepy, surprised; and wink, as shown in Figure 2. We randomly split the database into two parts: the first part is used as the training sample set, and the other is used for testing. We will indicate the significance of the proposed method by two experiments.

Experiment 1: We investigate the effect of different training sample sizes on the Yale database. We choose the processed data sets in Deng[1] [4]. For each subject, $t(= 2, 3, 4, 5, 6)$ samples are selected for training and the rest are used for testing. For a given t, there are 50 random splits. Set $d = 200$ for low-dimensional space. Table 1 summarizes the average results over 50 runs. It is obvious that the DPRC consistently and visibly performs the best for all selected values of t.

Experiment 2: We test the performance of the DPRC by increasing the dimension d from 20 to 800, Figure 3 shows the average recognition rates of 50 runs on the Yale database. The result shows that the recognition rate increases as the dimension increases. When the dimension exceeds 200, the recognition rate tends to stabilize.

5.2. Results on the ORL database. The ORL database contains face images of 40 distinct subjects captured at different time with variations in illumination, facial expression, and

[1] http://www.cad.zju.edu.cn/home/dengcai/Data/FaceData.html

Table 1
The face recognition rates (%) on the Yale database with t training samples per person.

Methods	$t=2$	$t=3$	$t=4$	$t=5$	$t=6$
SRC	53.17	62.00	67.50	71.97	74.85
CRC	55.67	67.03	72.95	78.53	81.01
LSR	59.40	70.30	76.13	80.73	82.77
LRC	46.77	55.78	60.24	64.88	67.49
LCDRC	58.66	69.36	74.80	77.84	78.53
LDRC	56.35	69.33	76.40	80.95	83.60
SR-SLR	56.81	67.25	74.57	79.66	81.41
ESRC	57.79	70.18	76.15	81.00	81.88
RLR^3	55.61	66.98	73.10	78.56	81.01
DPRC	**60.03**	**70.83**	**77.81**	**82.22**	**84.96**

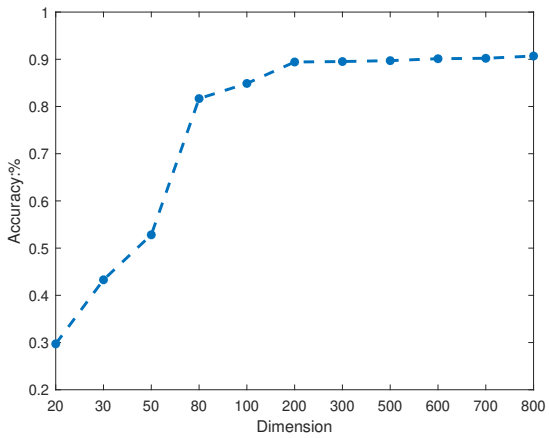

Figure 3. *The recognition performance on dimensions of the low-dimensional subspace on the Yale database.*

details (glasses). There are no restrictions imposed on the expression but the side movement or tilt is controlled within 20 degrees, as shown in Figure 4. For each subject, we select $t(=2,3,4,5,6)$ images for training, and the rest are used for testing. Table 2 shows the average results over 50 runs. The methods ESRC, SLR, and SR-SLR significantly outperform SRC, while our method (DPRC) consistently and visibly performs the best for all selected values of t.

5.3. Results on the AR database. The AR database consists of over 3,000 frontal images of 126 individuals. There are 26 images of each individual, taken at two different occasions [13]. The faces in the AR database contain variations such as illumination change, expressions, and facial disguises (i.e., sunglasses or scarf). In each of two separate sessions, seven undisguised images with expression or illumination variation, three images in sunglasses, and three images in scarf disguise are taken from each subject. As suggested in [14], we select a subset consisting of 2,600 images from 100 subjects (50 male and 50 female) in our experiments, and the images are cropped with dimension 32×32, as shown in Figure 5. For each subject, we randomly select $t(=2,3,4,5,6,7,8)$ images for training, and the rest for testing. The average results over 10

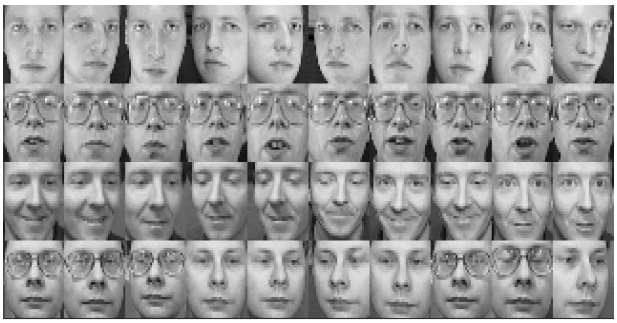

Figure 4. *Sample images in the ORL database with variations.*

Table 2
The face recognition rate (%) on the ORL database with t train samples per person.

Methods	$t = 2$	$t = 3$	$t = 4$	$t = 5$	$t = 6$
SRC	78.16	85.99	90.45	92.76	94.30
CRC	80.47	87.15	91.39	93.46	94.85
LSR	81.21	87.50	91.56	93.91	95.18
LRC	70.70	81.52	88.05	91.81	93.86
LCDRC	78.58	87.10	91.55	93.54	96.62
LDRC	78.13	88.34	93.26	95.91	96.78
SR-SLR	81.28	88.81	93.33	95.52	96.46
ESRC	82.64	89.91	93.59	95.63	96.48
RLR^3	81.19	87.61	91.78	93.94	95.26
DPRC	**82.72**	**90.45**	**94.26**	**96.07**	**97.20**

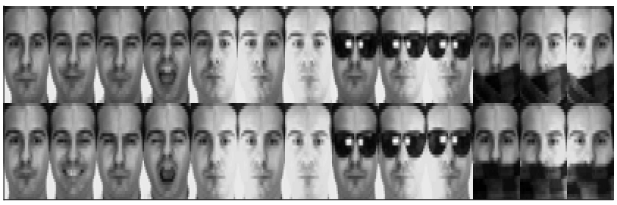

Figure 5. *Sample images in the AR database with variations of expressions, illumination, and occlusions. There are two rows of images corresponding to two sessions.*

runs obtained by using different classification methods are shown in Table 3. Apparently, our DPRC method achieves the best classification results in most cases, which also verifies that the proposed method outperforms all the other classification methods under different training conditions.

6. Conclusions. In this paper, we presented a DPRC framework for face recognition to enhance the discriminant ability of the RC methods. The proposed framework first obtains a

Table 3
The face recognition rate (%) on the AR database with t training samples per person.

Methods	$t=2$	$t=3$	$t=4$	$t=5$	$t=6$	$t=7$	$t=8$
SRC	36.93	47.53	54.54	61.17	65.53	89.85	72.30
CRC	63.65	75.05	80.50	84.62	88.50	90.36	91.75
LSR	69.37	80.55	85.62	89.04	91.75	93.22	94.25
LRC	19.20	28.63	36.51	46.45	53.41	58.49	64.95
LCDRC	56.65	64.66	66.80	71.02	72.35	74.50	77.40
LDRC	56.48	74.67	82.91	88.79	91.73	93.82	94.91
SR-SLR	56.48	72.27	81.28	87.89	72.35	93.64	94.95
ESRC	56.15	69.79	75.43	81.60	85.51	87.29	89.38
RLR^3	60.60	75.06	83.53	88.22	91.56	93.73	94.85
DPRC	**69.71**	**82.27**	**87.54**	**91.37**	**93.51**	**95.37**	**95.50**

discriminative projection matrix, which not only maximizes the ratio of the distance within interclass over the distance within intraclass, but also minimizes the linear reconstruction error within intraclass. Then the original data are projected onto the discriminative projection space, and the SRC method is adopted to obtain a final solution. An inexact ALM algorithm has been proposed for solving the resulting optimization problem in our framework, and a PAM method is adopted to the iteration subproblem of the inexact ALM. We proved that the proposed inexact ALM algorithm has a subsequence convergence property. Extensive experiments on publicly available face image databases showed that, compared to some state-of-the-art representation-based classification methods, the proposed framework and optimization algorithm are advancement with high recognition rate.

Acknowledgments. The authors are very grateful to the referees and editor for their helpful and constructive comments, which played an important role in improving this paper, and also thanks so much to the authors of [17] for providing the source code of the RLR^3 method.

REFERENCES

[1] T. Ahonen, A. Hadid, and M. Pietikäinen, *Face recognition with local binary patterns*, in European Conference on Computer Vision, Springer, Berlin, 2004, pp. 469–481.
[2] H. Attouch, J. Bolte, and B. F. Svaiter, *Convergence of descent methods for semi-algebraic and tame problems: Proximal algorithms, forward-backward splitting, and regularized Gauss–Seidel methods*, Math. Program. Series B, 137 (2013), pp. 91–129.
[3] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, *Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection*, IEEE Trans. Pattern Anal. Mach. Intell., 19 (1997), pp. 711–720.
[4] D. Cai, X. He, Y. Hu, J. Han, and T. Huang, *Learning a spatially smooth subspace for face recognition*, in 2007 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, Piscataway, NJ, 2007, pp. 1–7.
[5] W. Chen, H. Ji, and Y. You, *An augmented Lagrangian method for ℓ_1-regularized optimization problems with orthogonality constraints*, SIAM J. Sci. Comput., 38 (2016), pp. B570–B592.
[6] W. Deng, J. Hu, and J. Guo, *Extended SRC: Undersampled face recognition via intraclass variant dictionary*, IEEE Trans. Pattern Anal. Mach. Intell., 34 (2012), pp. 1864–1870.
[7] E. Elhamifar and R. Vidal, *Robust classification using structured sparse representation*, in CVPR 2011, IEEE, Piscataway, NJ, 2011, pp. 1873–1879.

[8] X. Fang, S. Teng, Z. Lai, Z. He, S. Xie, and W. K. Wong, *Robust latent subspace learning for image classification*, IEEE Trans. Neural Netw. Learn. Syst., 29 (2018), pp. 2502–2515.
[9] P. Huang, G. Gao, C. Qian, G. Yang, and Z. Yang, *Fuzzy linear regression discriminant projection for face recognition*, IEEE Access, 5 (2017), pp. 4340–4349.
[10] S. M. Huang and J. F. Yang, *Linear discriminant regression classification for face recognition*, IEEE Signal Process. Lett., 20 (2013), pp. 91–94.
[11] M. Iliadis, L. Spinoulas, A. S. Berahas, H. Wang, and A. K. Katsaggelos, *Sparse representation and least squares-based classification in face recognition*, in 2014 22nd European Signal Processing Conference (EUSIPCO), IEEE, Piscataway, NJ, 2014, pp. 526–530.
[12] G. Liu, Z. Lin, S. Yan, J. Sun, Y. Yu, and Y. Ma, *Robust recovery of subspace structures by low-rank representation*, IEEE Trans. Pattern Anal. Mach. Intell., 35 (2012), pp. 171–184.
[13] A. M. Martinez, *The AR Face Database*, Technical report 24, Computer Vision Center, Universitat Autònoma de Barcelona, Barcelona, 1998.
[14] A. M. Martinez and A. C. Kak, *PCA versus lDA*, IEEE Trans. Pattern Anal. Mach. Intell., 23 (2009), pp. 228–233.
[15] B. S. Mordukhovich and Y. H. Shao, *On nonconvex subdifferential calculus in Banach spaces*, J. Convex Analysis, 2 (1995), pp. 211–227.
[16] I. Naseem, R. Togneri, and M. Bennamoun, *Linear regression for face recognition*, IEEE Trans. Pattern Anal. Mach. Intell., 32 (2010), pp. 2106–2112.
[17] J. Qian, J. Yang, F. Zhang, and Z. Lin, *Robust low-rank regularized regression for face recognition with occlusion*, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, IEEE, Piscataway, NJ, 2014, pp. 21–26.
[18] X. Qu, S. Kim, R. Cui, and H. J. Kim, *Linear collaborative discriminant regression classification for face recognition*, J. Vis. Comm. Image Represent., 31 (2015), pp. 312–319.
[19] P. Redont and A. Soubeyran, *Proximal alternating minimization and projection methods for nonconvex problems: An approach based on the Kurdyka–Lojasiewicz inequality*, Math. Oper. Res., 35 (2010), pp. 438–457.
[20] F. S. Samaria and A. C. Harter, *Parameterization of a stochastic model for human face identification*, in Proceedings of the Second IEEE Workshop on Applications of Computer Vision, IEEE Computer Society, Los Alamitos, CA, 1994, pp. 138–142.
[21] S. Shekhar, V. M. Patel, N. M. Nasrabadi, and R. Chellappa, *Joint sparse representation for robust multimodal biometrics recognition*, IEEE Trans. Pattern Anal. Mach. Intell., 36 (2014), pp. 113–126.
[22] J. Sun, Y. Zhang, and J. Wright, *Efficient point-to-subspace query in ℓ_1 with application to robust object instance recognition*, SIAM J. Imaging Sci., 7 (2014), pp. 2105–2138.
[23] R. Tibshirani, *Regression shrinkage and selection via the lasso*, J. Roy. Statist. Soc., 73 (2011), pp. 267–288.
[24] Y. Wang, Y. Tang, L. Li, H. Chen, and J. Pan, *Atomic representation-based classification: Theory, algorithm, and applications*, IEEE Trans. Pattern Anal. Mach. Intell., 41 (2019), pp. 6–19.
[25] J. Wright, Y. Ma, Y. Tao, Z. Lin, and H.-Y. Shum, *Classification via minimum incremental coding length*, SIAM J. Imaging Sci., 2 (2009), pp. 367–395.
[26] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, *Robust face recognition via sparse representation*, IEEE Trans. Pattern Anal. Mach. Intell., 31 (2009), pp. 210–227.
[27] Y. Xu, Z. Zhong, J. Yang, J. You, and D. Zhang, *A new discriminative sparse representation method for robust face recognition via l_2 regularization*, IEEE Trans. Neural Netw. Learn. Syst., 28 (2017), pp. 2233–2242.
[28] J. Yang, D. Chu, L. Zhang, Y. Xu, and J. Yang, *Sparse representation classifier steered discriminative projection with applications to face recognition*, IEEE Trans. Neural Netw. Learn. Syst., 24 (2013), pp. 1023–1035.
[29] M. Yang, L. Zhang, X. Feng, and D. Zhang, *Sparse representation based Fisher discrimination dictionary learning for image classification*, Int. J. Comput. Vis., 109 (2014), pp. 209–232.
[30] M. Yang, L. Zhang, J. Yang, and D. Zhang, *Robust sparse coding for face recognition*, in CVPR 2011, IEEE, Piscataway, NJ, 2011, pp. 625–632.

[31] L. Zhang, M. Yang, and X. Feng, *Sparse representation or collaborative representation: Which helps face recognition?*, in 2011 International Conference on Computer Vision, IEEE, Piscataway, NJ, 2011, pp. 471–478.

[32] Z. Zhang, Z. Lai, Y. Xu, L. Shao, J. Wu, and G. Xie, *Discriminative elastic-net regularized linear regression*, IEEE Trans. Image Process., 26 (2017), pp. 1466–1481.

[33] H. Zhu, X. Zhang, D. Chu, and L. Z. Liao, *Nonconvex and nonsmooth optimization with generalized orthogonality constraints: An approximate augmented Lagrangian method*, J. Sci. Comput., 72 (2017), pp. 1–42.