Best Pair Formulation & Accelerated Scheme for Non-convex Principal Component Pursuit

Aritra Dutta, Filip Hanzely, Jingwei Liang, and Peter Richtárik

***Abstract*—Given two disjoint sets, the *best pair* problem aims to find a point in one set and another point in the other set with minimal distance between them. In this paper, we formulate the classical robust principal component analysis (RPCA) problem as a best pair problem and design an accelerated proximal gradient algorithm to solve it. We prove that the method enjoys global convergence with a local linear rate. Our extensive numerical experiments on both real and synthetic data sets suggest that our proposed algorithm outperforms relevant baseline algorithms in the literature.**

***Index Terms*—Principal component analysis, Principal component pursuit, Robust PCA, Low-rank and sparse matrix decomposition, Matrix completion, Video segmentation, Optimization, Non-convex robust PCA, Accelerated proximal gradient.**

I. Introduction

LET $A \in \mathbb{R}^{m\times n}$ be a given matrix, the generalized low-rank recovery model can be written as [1]:

$$\min_{L\in\mathbb{R}^{m\times n}} \mathcal{F}(A,L) + \lambda\mathcal{R}(L), \tag{1}$$

where $\mathcal{F}(A,L)$ is a loss function, $\mathcal{R}(L) \stackrel{\text{def}}{=} \sum_{i=1}^{n}\mathcal{R}_i(L)$ is a suitable regularizer, and $\lambda > 0$ is a balancing parameter. By an appropriate choice of the loss function and the regularizer, (1) can express a wide range of low-rank matrix approximation problems. For example, by setting $\mathcal{F}(A,L) = \|A-L\|^2, \lambda = 1$, where $\|\cdot\|$ denotes the Frobenius norm, and $\mathcal{R}(L) = \iota_{\text{rank}(L)\le r}(L)$ — the characteristic function (7) of the set $\{L \in \mathbb{R}^{m\times n} : \text{rank}(L) \le r\}$, (1), specializes to:

$$\min_{L\in\mathbb{R}^{m\times n}} \|A-L\|^2 + \iota_{\text{rank}(L)\le r}(L), \tag{2}$$

which is a *best approximation* formulation of the classical principal component analysis (PCA). The solution of (2) is given by: $\hat{L} = U\mathbf{H}_r(\Sigma)V^\top$, where $U\Sigma V^\top = A$ is a singular value decomposition (SVD) of A and $\mathbf{H}_r(\cdot)$ is the hard-thresholding operator that keeps the r largest singular values. Although PCA is vastly used in different engineering applications, it can only handle the presence of uniformly distributed noise and is rather sensitive to sparse outliers in the data matrix [2], [3], [4]. To overcome this shortcoming and deal with sparse errors, [5], [4] replaced the Frobenius norm in (2) by ℓ_0 pseudo norm, and introduced the celebrated *principal component pursuit* (PCP) problem:

$$\min_{L\in\mathbb{R}^{m\times n}} \|A-L\|_{\ell_0} + \lambda\text{rank}(L). \tag{3}$$

However, the above problem is non-convex and NP-hard [6].

One of the most commonly used, tractable surrogate reformulations of (3) is replacing the rank function with nuclear norm $\|L\|_\star$ and ℓ_0 pseudo norm with ℓ_1-norm $\|A-L\|_{\ell_1}$ [7], [8]. Exploiting this idea, *Robust PCA* (RPCA) was introduced as a convex surrogate of the PCP problem [3], [2], [4]:

$$\min_{L\in\mathbb{R}^{m\times n}} \|A-L\|_{\ell_1} + \lambda\|L\|_\star. \tag{4}$$

It was shown in [5], [4] that under a rank-sparsity incoherence assumption, problem (3) can be provably solved via (4), as the solutions of them lie close to each other with high probability.

Recently, [9] reformulated (3) as a non-convex feasibility problem, which does not require any objective function, convex relaxation, or surrogate convex constraints. Instead, it exploits the idea that the solution to the PCP problem lies in the intersection of two sets — one convex and one non-convex if one considers both the target rank r and the target sparsity α as hyperparameters. Let $X = \begin{pmatrix} S \\ L \end{pmatrix} \in \mathbb{R}^{2m\times n}$ and $K = [\text{Id}, \text{Id}]$ where Id is the identity operator of the space $\mathbb{R}^{m\times n}$, define

$$\mathcal{X} \stackrel{\text{def}}{=} \{X : KX = A\} \text{ and}$$
$$\mathcal{Y} \stackrel{\text{def}}{=} \{X : \text{rank}(L) \le r, \|S_{i,\cdot}\|_0 \le \alpha m, \|S_{\cdot,j}\|_0 \le \alpha n\}$$

where $i \in [m],\ j \in [n]$. Note that $\mathcal{X}$ is convex and $\mathcal{Y}$ is non-convex[1]. Given the sets, Dutta et al. [9] reformulated (3) as non-convex feasibility problem:

$$\text{find} \quad X \in \mathbb{R}^{2m\times n} \text{ such that } X \in \mathcal{X}\cap\mathcal{Y}. \tag{5}$$

If we replace Id in K with Bernoulli binary matrix, (5) becomes the reformulation of PCP with partial observation.

A. Formulation and Contributions

In this paper we consider reformulating the feasibility problem (5) as a *best pair* problem. Given two sets $\mathcal{X}, \mathcal{Y} \subset \mathbb{R}^{2m\times n}$, the best pair problem aims to find a pair of points $(X^\star, Y^\star) \in \mathcal{X}\times\mathcal{Y}$ such that they have the closest distance, *i.e.* $(X^\star, Y^\star)$ is a solution of the problem below:

$$\min_{X\in\mathcal{X}, Y\in\mathcal{Y}} \ \tfrac{1}{2}\|X-Y\|^2. \tag{6}$$

A. Dutta is with the Division of Computer, Electrical and Management Sciences & Engineering, King Abdullah University of Science and Technology, Saudi Arabia, e-mail:aritra.dutta@kaust.edu.sa (see www.aritradutta.com).

F. Hanzely is with the Division of Computer, Electrical and Management Sciences & Engineering, King Abdullah University of Science and Technology, Saudi Arabia, e-mail:filip.hanzely@kaust.edu.sa.

J. Liang is with the Department of Applied Mathematics and Theoretical Physics, University of Cambridge, Cambridge UK, e-mail: jl993@cam.ac.uk.

P. Richtárik is with the Division of Computer, Electrical and Management Sciences & Engineering, King Abdullah University of Science and Technology, Saudi Arabia, e-mail:peter.Richtárik@kaust.edu.sa (see Richtárik.org).

[1]The α-sparsity constraint on S means that for $\alpha \in (0,1)$, each row and column of S contains no more than αn and αm number of non-zero entries, respectively. This is slightly more complicated than directly applying $\|\cdot\|_0$ constraint. However, it often works better in practice.

When the intersection of $\mathcal{X} \cap \mathcal{Y} \neq \emptyset$ is non-empty, (6) reduces to the feasibility problem, with $X^\star = Y^\star \in \mathcal{X} \cap \mathcal{Y}$. Given a set $\mathcal{X}$, define its characteristic function by

$$\iota_{\mathcal{X}}(X) \stackrel{\text{def}}{=} \begin{cases} 0 : X \in \mathcal{X}, \\ +\infty : \text{otherwise.} \end{cases} \tag{7}$$

Then (6) can be equivalently written as

$$\min_{X,Y \in \mathbb{R}^{2m \times n}} \iota_{\mathcal{X}}(X) + \tfrac{1}{2}\|X - Y\|^2 + \iota_{\mathcal{Y}}(Y). \tag{8}$$

Observe that for a given Y, problem (8) becomes $\min_{X \in \mathbb{R}^{2m \times n}} \iota_{\mathcal{X}}(X) + \frac{1}{2}\|X - Y\|^2$ which is the Moreau envelope [10] of $\iota_{\mathcal{X}}(X)$ of index 1:

$${}^1\big(\iota_{\mathcal{X}}(Y)\big) \stackrel{\text{def}}{=} \min_{X \in \mathbb{R}^{2m \times n}} \tfrac{1}{2}\|X - Y\|^2 + \iota_{\mathcal{X}}(X).$$

As a result, we can simplify (8) to the case of only Y,

$$\boxed{\min_{Y \in \mathbb{R}^{2m \times n}} \iota_{\mathcal{Y}}(Y) + {}^1\big(\iota_{\mathcal{X}}(Y)\big).} \tag{9}$$

For the rest of the paper, we focus on (9) and our main contributions are summarized below:

- **New formulation and a new algorithm for non-convex PCP.** We reformulate the non-convex set feasibility formulation of RPCA to a *best pair* problem. Although our formulation was inspired by formulation (5) from [9], to the best of our knowledge, we are the first to formulate and solve RPCA via the best pair. To this end, we design a fast and efficient algorithm — an accelerated proximal gradient method — to solve it.
- **Theoretical convergence guarantees.** Both global and local convergence analysis of the accelerated proximal gradient method are provided. Globally, we show that our algorithm converges to a critical point. Locally, our algorithm enjoys a fast linear rate of convergence, which we can sharply estimate. We owe this novelty to our best pair formulation. In contrast, the non-convex projection RPCA from [9] or GoDec [11] are not able to estimate the local convergence rate. Further, as we tackle the non-convex objective directly (i.e., we do not minimize a surrogate function), we do not require the rank-sparsity incoherence assumption [4], [12].
- **Numerical experiments and applications to real-world problems.** We apply the proposed method to several well-tested applications in computer vision. Our extensive experiments on both real and synthetic data suggest that our algorithm matches or outperforms relevant baseline algorithms in *fractions* of their execution time. Additionally, in the appendix, we provide empirical validity of the hyperparameter sensitivity of our proposed approach.

B. Related Work

In this scope, for completeness, we quote a few seminal works on RPCA, matrix completion and related problems. Besides (4), there are other formulations of RPCA. One of the most popular ways is to introduce an auxiliary variable, S, and add an additional constraint $L + S = A$, which yields

$$\min_{L,S \in \mathbb{R}^{m \times n}} \|S\|_{\ell_1} + \lambda \|L\|_\star \;\text{ subject to }\; L + S = A. \tag{10}$$

This *constrained* formulation enables several avenues to solve RPCA, such as the exact and inexact augmented Lagrangian method of multipliers by Lin et al. [2], accelerated proximal gradient method [3], alternating direction method [13], alternating projection with intermediate denoising [14], dual approach [15], and SpaRCS [16], manifold optimization by Yi et al. [17] and Zhang & Yang [18], are among the popular ones. We refer to [19] for a comprehensive review.

For the discussion above, A is fully observed with no data missing. One can also consider the case that A is partially observed, that is, there exists a projection operator (or simply a Bernoulli binary mask) P_Ω on the set of observed data entries $\Omega \subseteq [m] \times [n]$ that is defined by

$$\big(\mathrm{P}_\Omega[A]\big)_{ij} = \begin{cases} A_{ij} : (i,j) \in \Omega, \\ 0 : \text{otherwise.} \end{cases} \tag{11}$$

The partially observed version of (10) reads

$$\begin{aligned} &\min_{L,S \in \mathbb{R}^{m \times n}} \|S\|_{\ell_1} + \lambda \|L\|_\star \\ &\text{subject to} \quad \mathrm{P}_\Omega(L + S) = \mathrm{P}_\Omega(A). \end{aligned} \tag{12}$$

Besides (10) and (12), other tractable reformulations of (3) also exist. For example, if the rank and target sparsity are user-defined, then it is common practice to relax the equality constraint in (10) and consider it in the objective function as a penalty. This, together with explicit constraints on the target rank, r, and target sparsity level, α, (*user-defined* hyperparameter), leads to the GoDec formulation [11]. One can also extend the above model to the case of partially observed data that leads to a more general class of problems that is commonly known as the *robust matrix completion (RMC)* problem [20], [21], [22], [23], [24], [25], [26], [27], [28], [29], [30], [31] that contains the variant proposed in [11] as a special case. With $S = 0$, the matrix completion (MC) problem is also a special case of the RMC problem [32], [33], [7], [34], [35], [36], [37], [38], [39]. Lastly, when the whole matrix is observed, the RMC problem is nothing but (10). To find a *unified treatment* of several low-rank and sparsity inducing formulations in one essentially single form, we refer the readers to [40] and the survey [41].

a) Incoherence: The incoherence condition is used when there is an "identifiability" issue between the sparse and the low-rank decomposition. Chandrasekaran et al. [12] considered a deterministic sparsity model, where each row and column of the $m \times n$ matrix S has at most α-fraction of non-zero entries. In contrast, [4] considered a random-sparsity model with an additional incoherence condition. Consequently, [4] obtained an exact recovery with a constant fraction of corrupted (sparse) entries, even when L is almost full-rank. Indeed, incoherence is required for the following problem: *Can we recover $\hat{L}, \hat{S}$ exactly or uniquely given that we observe only $A = \hat{L} + \hat{S}$ and we know that$\hat{L}$ is of low-rank and $\hat{S}$ is sparse?* In contrast, we ask the following question: *Given a matrix A, can we find $L^\star, S^\star$ such that A is close to $L^\star + S^\star$, while $L^\star$ is low rank and $S^\star$ is sparse?* As a consequence of asking a slightly different question, incoherence is not required anymore for our case. However, our approach allows answering the first question under incoherence as well: Given that

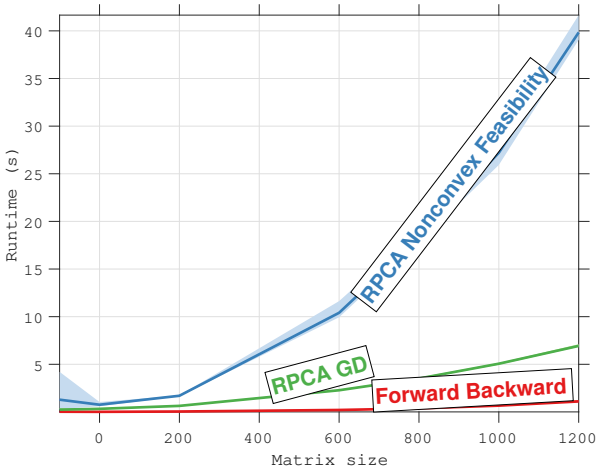


Fig. 1: Comparison of matrix size (original rank $m/10$) and runtime to recover a $m/10$ low-rank matrix by RPCA F and best pair. The shaded regions represent the variation in computational-time across a wide number of running instances.

$A = \hat{L} + \hat{S}$ and we know target rank of $\hat{L}$ and target sparsity of $\hat{S}$, the optimal value of best pair objective becomes 0. In such a case, the incoherence immediately yields the uniqueness of the solution [4], which means that $L^\star = \hat{L}, S^\star = \hat{S}$ and thus the recovery is exact. A similar argument applies to the RMC problem.

C. Notations

Throughout the paper, $\mathbb{N}$ is the set of non-negative integers. For a nonempty closed convex set $\Omega \subset \mathbb{R}^n$, denote P_Ω the orthogonal projector onto Ω. Let $\mathcal{R} : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be a lower semi-continuous (lsc) function, its domain is defined as $\mathrm{dom}(\mathcal{R}) \stackrel{\mathrm{def}}{=} \{x \in \mathbb{R}^n : \mathcal{R}(x) < +\infty\}$, and it is said to be proper if $\mathrm{dom}(\mathcal{R}) \neq \emptyset$. We need the following notions from variational analysis, see e.g. [42] for details. Given $x \in \mathrm{dom}(\mathcal{R})$, the Fréchet subdifferential $\partial^F \mathcal{R}(x)$ of $\mathcal{R}$ at x, is the set of vectors $v \in \mathbb{R}^n$ that satisfies $\liminf_{z \to x,\, z \neq x} \frac{1}{\|x-z\|}(\mathcal{R}(z) - \mathcal{R}(x) - \langle v, z - x\rangle) \geq 0$. If $x \notin \mathrm{dom}(\mathcal{R})$, then $\partial^F \mathcal{R}(x) = \emptyset$. The limiting-subdifferential (or simply subdifferential) of $\mathcal{R}$ at x, written as $\partial \mathcal{R}(x)$, is defined as $\partial \mathcal{R}(x) \stackrel{\mathrm{def}}{=} \{v \in \mathbb{R}^n : \exists x_k \to x, \mathcal{R}(x_k) \to \mathcal{R}(x), v_k \in \partial^F \mathcal{R}(x_k) \to v\}$. Denote $\mathrm{dom}(\partial \mathcal{R}) \stackrel{\mathrm{def}}{=} \{x \in \mathbb{R}^n : \partial \mathcal{R}(x) \neq \emptyset\}$. Both $\partial^F \mathcal{R}(x)$ and $\partial \mathcal{R}(x)$ are closed, with $\partial^F \mathcal{R}(x)$ convex and $\partial^F \mathcal{R}(x) \subset \partial \mathcal{R}(x)$ [42, Proposition 8.5]. Since $\mathcal{R}$ is lsc, it is (subdifferentially) regular at x if and only if $\partial^F \mathcal{R}(x) = \partial \mathcal{R}(x)$ [42, Corollary 8.11]. A necessary condition for x to be a minimizer of $\mathcal{R}$ is $0 \in \partial \mathcal{R}(x)$. The set of critical points of $\mathcal{R}$ is $\mathrm{crit}(\mathcal{R}) = \{x \in \mathbb{R}^n : 0 \in \partial \mathcal{R}(x)\}$.

II. An accelerated proximal gradient algorithm

In this section, we describe a gradient-based optimization method for solving (9). Denote $\mathrm{P}_\mathcal{X}, \mathrm{P}_\mathcal{Y}$ as the projection operators onto $\mathcal{X}$ and $\mathcal{Y}$, respectively. Since $\mathcal{X}$ is a non-empty closed convex set, its characteristic function $\iota_\mathcal{X}$ is proper convex and lower semi-continuous. Owing to [10], the Moreau envelope is a convex differentiable function with a gradient of the form

$$\nabla\big({}^1(\iota_\mathcal{X}(Y))\big) = (\mathrm{Id} - \mathrm{P}_\mathcal{X})(Y)$$

which is 1-Lipschitz continuous. Clearly, (9) admits a "non-smooth + smooth" structure, and in literature one prevailing algorithm to apply is the proximal gradient method [43], a.k.a. Forward-Backward splitting. In this paper, we consider an accelerated version of the method, see Algorithm 1 below.

Algorithm 1: An accelerated proximal gradient method

Initial: Parameters $\gamma \in]0, 2]$, $Y_0 \in \mathbb{R}^{2m \times n}$, $Y_{-1} = Y_0$, $\{a_k\}_{k \geq 0}$, $\{b_k\}_{k \geq 0}$.

repeat

$$\begin{aligned} U_k &= Y_k + a_k(Y_k - Y_{k-1}), \\ V_k &= Y_k + b_k(Y_k - Y_{k-1}), \\ Y_{k+1} &= \mathrm{P}_\mathcal{Y}\big(U_k - \gamma(V_k - \mathrm{P}_\mathcal{X}(V_k))\big). \end{aligned} \tag{13}$$

$k = k + 1$;

until *convergence*;

Remark 1.
- In Algorithm 1, U_k, V_k are the so-called inertial points, which aim to provide acceleration under properly chosen a_k and b_k (see (16) later). The presence of such auxiliary points is stadard for first-order momentum based acceleration schemes [44].
- If we set $\gamma = 1$ and $a_k, b_k \equiv 0$, Algorithm 1 becomes the alternating projection method for the feasibility problem (5), which covers the method from [9] as a special case.
- From (5) to (9), we can also consider Moreau envelope of the non-convex set $\mathcal{Y}$, that is $\min_{X \in \mathbb{R}^{2m \times n}} \iota_\mathcal{X}(X) + {}^1(\iota_\mathcal{Y}(X))$ which also works well in practice.
- Algorithm 1 is a special case of the multi-step inertial proximal gradient descent method considered in [45] for general non-smooth non-convex optimization.

A. The projection operators

The projection operators $\mathrm{P}_\mathcal{X}, \mathrm{P}_\mathcal{Y}$ are easy to compute. Given $X = \begin{pmatrix} S \\ L \end{pmatrix}$, since $\mathcal{X}$ is an affine subspace, projecting X onto $\mathcal{X}$ is $\mathrm{P}_\mathcal{X}(X) = \frac{1}{2}\begin{pmatrix} A + S - L \\ A - S + L \end{pmatrix}$. For partially observed case with $K = [\mathrm{P}_\Omega, \mathrm{P}_\Omega]$ where P_Ω is the binary mask defined in (11), we have $\mathrm{P}_\mathcal{X}(X) = \begin{pmatrix} S \\ L \end{pmatrix} + \frac{1}{2}\begin{pmatrix} \mathrm{P}_\Omega[A - S - L] \\ \mathrm{P}_\Omega[A - S - L] \end{pmatrix}$. Projection $\mathrm{P}_\mathcal{Y}$ consists of separately applying the projection on the set of low-rank matrices for L and projection on the set of sparse matrices for S. The low rank projection can be performed via SVD — if the target rank is r, we perform a SVD and consider the first r singular values (arranged in an non-increasing order) and the corresponding singular vectors and set the rest $(\min\{m, n\} - r)$ singular values to 0. The projection on sparse set with sparsity level, α, as stated before, is hard and expensive to compute. Therefore, we project onto an approximate support set of sparse set where the support is characterized by Ω_α. In other words, for each column, we set zeros for $\lfloor (1 - \alpha)m \rfloor$ smallest elements in absolute value (similarly, for each row, we set zeros for $\lfloor (1 - \alpha)n \rfloor$

smallest elements). This approach is consistent with the current literature [17], [18], [9]. Formally, we use the approximate projection via the following operator $\mathcal{T}_\alpha[S]$ (also see [9], [17], [18]) onto $\mathcal{Y}$:

$$\mathcal{T}_\alpha[S] \stackrel{\text{def}}{=} \{\mathrm{P}_{\Omega_\alpha}(S) \in \mathbb{R}^{m\times n} : (i,j) \in \Omega_\alpha \text{ if } |S_{ij}| \geq |S^{(\alpha n)}_{(i,.)}| \text{ and } |S_{ij}| \geq |S^{(\alpha m)}_{(.,j)}|\}, \tag{14}$$

where $S^{(\alpha n)}_{(i,.)}$ and $S^{(\alpha m)}_{(.,j)}$ denote the α fraction of largest entries of S along the i^{th} row and j^{th} column, respectively. We note that this approximation is not necessary if we consider the sparsity constraint over the whole matrix (i.e., not column-wise and row-wise), which is cheap to perform but slightly inferior in practice. Since the approximate projection produces better results, we use it throughout the paper. For more details on both low-rank and sparse projections, see [9].

B. Global convergence

Since set $\mathcal{Y}$ is semi-algebraic [46], our global convergence guarantees of Algorithm 1 is established via Kurdyka-Łojasiewicz property.

1) Kurdyka-Łojasiewicz property: Let $R : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be a proper lower semi-continuous function. For η_1, η_2 such that $-\infty < \eta_1 < \eta_2 < +\infty$, define the set

$$[\eta_1 < R < \eta_2] \stackrel{\text{def}}{=} \{Y \in \mathbb{R}^n : \eta_1 < R(Y) < \eta_2\}.$$

Definition 2. Function R is said to have the Kurdyka-Łojasiewicz property at $\overline{Y} \in \mathrm{dom}(R)$ if there exists $\eta \in]0, +\infty]$, a neighbourhood U of $\overline{Y}$ and a continuous concave function $\varphi : [0, \eta[\to \mathbb{R}_+$ such that

(i) $\varphi(0) = 0$, φ is C^1 on $]0, \eta[$, and $\forall s \in]0, \eta[$, $\varphi'(s) > 0$;
(ii) for all $Y \in U \cap [R(\overline{Y}) < R < R(\overline{Y}) + \eta]$, there holds the Kurdyka-Łojasiewicz inequality $\varphi'(R(Y) - R(\overline{Y}))\mathrm{dist}\big(0, \partial R(Y)\big) \geq 1$.

Proper lower semi-continuous functions which satisfy the Kurdyka-Łojasiewicz property at each point of $\mathrm{dom}(\partial R)$ are called KL functions.

KL functions include the class of semi-algebraic functions, see [47], [46]. For instance, the ℓ_0 pseudo-norm and the rank function are KL.

2) Global convergence: To deliver the convergence result, we rewrite (9) into the following generic form

$$\min_{Y\in\mathbb{R}^{2m\times n}} \big\{\Phi(Y) \stackrel{\text{def}}{=} \mathcal{R}(Y) + \mathcal{F}(Y)\big\}, \tag{15}$$

where we assume that

(A.1) $\mathcal{R} : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is proper lower semi-continuous, and bounded from below;
(A.2) $\mathcal{F} : \mathbb{R}^n \to \mathbb{R}$ is convex differentiable and its gradient $\nabla\mathcal{F}$ is L-Lipschitz continuous.

Let $\nu > 0$ be a constant. Define the following quantities,

$$\begin{aligned} \beta_k &\stackrel{\text{def}}{=} \tfrac{1-\gamma L - a_k - \nu}{2\gamma}, \quad \underline{\beta} \stackrel{\text{def}}{=} \liminf_{k\in\mathbb{N}} \beta_k \\ \alpha_k &\stackrel{\text{def}}{=} \tfrac{\gamma b_k^2 L^2 + \nu a_k}{2\nu\gamma}, \quad \overline{\alpha} \stackrel{\text{def}}{=} \limsup_{k\in\mathbb{N}} \alpha_k. \end{aligned} \tag{16}$$

Theorem 3 (Global convergence). *For problem* (15)*, assume* (**A.1**)-(**A.2**) *hold, and that* Φ *is a proper lower semi-continuous KL function which is bounded from below. For Algorithm 1, choose* ν, γ, a_k, b_k *such that*

$$\delta \stackrel{\text{def}}{=} \underline{\beta} - \overline{\alpha} > 0. \tag{17}$$

Then each bounded sequence $\{Y_k\}_{k\in\mathbb{N}}$ *satisfies*

(a) $\{Y_k\}_{k\in\mathbb{N}}$ *has finite length, i.e.* $\sum_k \|Y_k - Y_{k-1}\| < +\infty$*;*
(b) *There exists a critical point* $Y^\star \in \mathrm{crit}(\Phi)$ *such that* $\lim_{k\to+\infty} Y_k = Y^\star$.

The proof of the above theorem can be found in the appendix. We also refer to [45] and the reference therein for more results on the non-convex proximal gradient method.

C. Local linear convergence

Now we turn to the local perspective and present a local linear convergence analysis for Algorithm 1. For the constraint set $\mathcal{Y}$ defined in (5), consider the following decomposition

$$\begin{aligned} \mathcal{Y}_L &\stackrel{\text{def}}{=} \left\{Y = \begin{pmatrix} S \\ L \end{pmatrix} : \mathrm{rank}(L) \leq r\right\}, \\ \mathcal{Y}_S &\stackrel{\text{def}}{=} \left\{Y = \begin{pmatrix} S \\ L \end{pmatrix} : S \text{ is } \alpha\text{-sparse}\right\}. \end{aligned}$$

For the sequence Y_k generated by (13), suppose $Y_k = \begin{pmatrix} S_k \\ L_k \end{pmatrix}$. It is immediate that $\mathrm{rank}(L_k) \leq r$ holds for all k. For S_k, though it is always α-sparse, the locations of non-zero elements change along the iteration. In the following, we first show that after a finite number of iterations the locations of non-zero elements of S_k stop changing, that is S_k will have the same support as that of $S^\star$ to which S_k converges, and Algorithm 1 enters a linear convergence regime.

1) Support identification of S_k*:* Let $Y^\star = \begin{pmatrix} S^\star \\ L^\star \end{pmatrix}$ be a critical point of (9) to which Y_k converges. Denote the set $\mathcal{S} \stackrel{\text{def}}{=} \{S \in \mathbb{R}^{m\times n} : \mathrm{supp}(S^\star) \subseteq \mathrm{supp}(S^\star)\}$. Clearly, $S^\star \in \mathcal{S}$ and we have the result below regarding S_k and $\mathcal{S}$.

Theorem 4 (Support identification). *For Algorithm 1, suppose Theorem 3 holds. Then* Y_k *converges to a critical point* $Y^\star$ *of* (9)*. For all* k *large enough, we have* $S_k \in \mathcal{S}$.

The above result simply means that after finite number of iterations, $\mathrm{supp}(S_k) = \mathrm{supp}(S^\star)$ holds for all k large enough.

2) Local linear convergence: Given a critical point $Y^\star$, let $X^\star = \mathrm{P}_{\mathcal{X}}(Y^\star)$, we have

$$X^\star \in \mathcal{X} \text{ and } S^\star \in \mathcal{S}, \quad L^\star \in \mathcal{Y}_L.$$

Note that the first two sets, $\mathcal{X}, \mathcal{S}$ are (affine) subspaces, hence smooth, and $\mathcal{Y}_L$ is the set of fixed-rank matrices which is C^2-smooth manifold [48]. To derive the local linear rate, we need to utilize the smoothness of these sets. Let $\mathcal{M}$ be a C^2-smooth manifold and let $\mathcal{T}_{\mathcal{M}}(X)$ the tangent space of $\mathcal{M}$ at $X \in \mathcal{M}$. Next, we introduce a lemma, which is crucial for the local linear convergence analysis.

Lemma 5 ([49, Lemma 5.1]). *Let* $\mathcal{M}$ *be a* C^2*-smooth manifold around* X*. Then for any* $X' \in \mathcal{M} \cap \mathcal{N}$*, where* $\mathcal{N}$ *is a neighbourhood of* X*, the projection operator* $\mathrm{P}_{\mathcal{M}}(X')$ *is single-valued and* C^1 *around* X*, and* $X' - X = \mathrm{P}_{\mathcal{T}_{\mathcal{M}}(X)}(X' - X) + o(\|X' - X\|)$*. If moreover,* $\mathcal{M} = X + \mathcal{T}_{\mathcal{M}}(X)$ *is an affine subspace, then* $X' - X = \mathrm{P}_{\mathcal{T}_{\mathcal{M}}(X)}(X' - X)$.

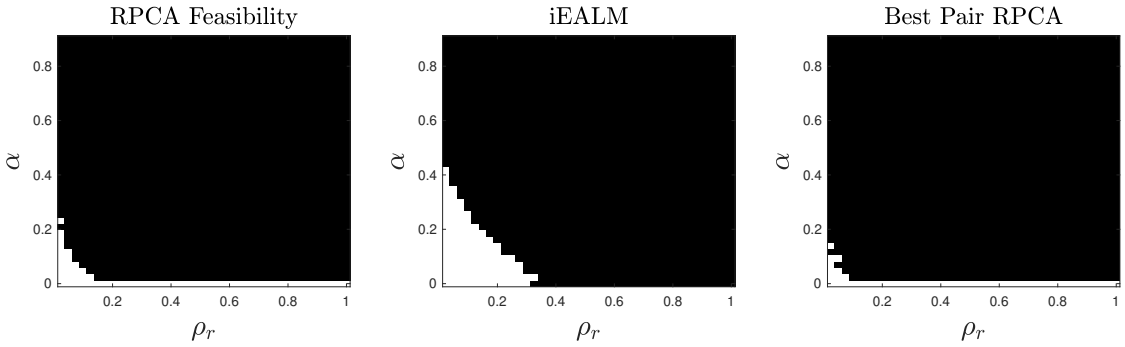


Fig. 2: Phase transition diagram for RPCA F, iEALM, and APG with respect to rank and error sparsity. Here, $\rho_r = \mathrm{rank}(L)/m$ and α is the sparsity measure. We have $(\rho_r, \alpha) \in (0.025, 1] \times (0, 1)$ with $r = 5 : 5 : 200$ and $\alpha = \texttt{linspace}(0, 0.99, 40)$. We perform 10 runs of each algorithm.

Denote the tangent spaces of $\mathcal{X}, \mathcal{Y}$ at $X^\star, Y^\star$ as $T_{\mathcal{X}}^{X^\star}$ and $T_{\mathcal{Y}}^{Y^\star}$, respectively. We refer to the appendix for detailed expressions of these tangent spaces. Denote $\mathrm{P}_{T_{\mathcal{X}}^{X^\star}}$ and $\mathrm{P}_{T_{\mathcal{Y}}^{Y^\star}}$ the projections onto the tangent spaces. Define $\mathcal{P} \stackrel{\text{def}}{=} \mathrm{P}_{T_{\mathcal{Y}}^{Y^\star}}\big((1-\gamma)\mathrm{Id} + \gamma \mathrm{P}_{T_{\mathcal{X}}^{X^\star}}\big)\mathrm{P}_{T_{\mathcal{Y}}^{Y^\star}}$, and

$$D_k \stackrel{\text{def}}{=} \begin{pmatrix} Y_k - Y^\star \\ Y_{k-1} - Y^\star \end{pmatrix} \text{ and } \mathcal{Q} \stackrel{\text{def}}{=} \begin{bmatrix} (1+a)\mathcal{P} & -a\mathcal{P} \\ \mathrm{Id} & 0 \end{bmatrix}$$

with $a \in [0, 1]$. Denote $\rho_{\mathcal{P}}, \rho_{\mathcal{Q}}$ the spectral radiuses of $\mathcal{P}, \mathcal{Q}$.

Theorem 6 (Local linear convergence). *For Algorithm 1, suppose Theorem 4 holds. Then Y_k converges to a critical point $Y^\star$ of* (9)*. Suppose $b_k = a_k \equiv a \in [0, 1]$, there exists a $K > 0$ such that for all $k \geq K$, $D_{k+1} = \mathcal{Q}D_k + o(\|D_k\|)$. Moreover, if $\rho_{\mathcal{P}} < 1$, then so is $\rho_{\mathcal{Q}}$, and for all k large enough we have $\|Y_k - Y^\star\| = O(\rho_{\mathcal{Q}}^k)$.*

Remark 7. If $T_{\mathcal{X}}^{X^\star} \cap T_{\mathcal{Y}}^{Y^\star} = \{0\}$, then it can be shown that $\rho_{\mathcal{P}} < 1$. Given $\rho_{\mathcal{P}}$, $\rho_{\mathcal{Q}}$ can be expressed explicitly in terms of a and $\rho_{\mathcal{P}}$. For the general setting of $a_k \to a \in [0, 1], b_k \to b \in [0, 1]$ and how inertial speeds up local linear convergence, we refer to [50, Chapter 6] for detailed discussions.

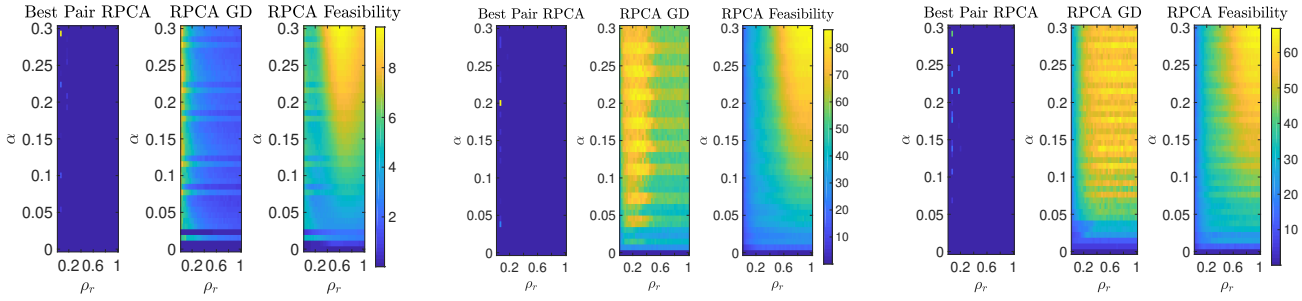
We provide the proofs of global convergence in Appendix, Section B, and the proofs of local linear convergence are provided in Appendix, Section C. An example comparing our theoretical rate estimation of local linear convergence rate and practical observation is provided in Appendix C-Figure 10.

III. Numerical experiments

In this section, we extensively tested our best-pair formulation on both real and synthetic data against a vast class of PCP algorithms. The first set of algorithms that we tested against, e.g., iEALM and APG, determine the target rank and sparsity *robustly* from the given set of hyperparameters. On the other hand, for the second set of algorithms, e.g., RPCA gradient descent (RPCA GD), Go decomposition (GoDec), and RPCA non-convex feasibility (RPCA NCF), the target rank and sparsity are user-defined. While our accelerated proximal gradient method belongs to the second class, to show its effectiveness, we compare it with both classes of state-of-the-art robust PCP algorithms (see Table I in the appendix Section D) on several computer vision applications — removal of shadows and specularities from face images, background estimation or tracking from full and partially observed video sequences, inlier detection from a grossly corrupted dataset (see Supplementary material, Appendix F in the appendix). In all experiments, we use the approximate projection [9], [17], [18] onto $\mathcal{Y}$ as mentioned in (14).

1) Results on synthetic data: The primary goal of this set of experiments is to understand the behaviour of our proposed method on some well-understood data and to test against some state-of-the-art algorithms. To construct our test matrix A, for these experiments, we used the idea proposed by Wright et al. [3]. First, we generate the low-rank matrix, L, as a product of two independent full-rank matrices of size $m \times r$ with $r < m$ such that elements are independent and identically distributed (i.i.d.) and sampled from a normal distribution — $\mathcal{N}(0, 1)$. We generate the sparse matrix, S, such that its elements are chosen uniformly from the interval $[-500, 500]$. We create the sparse support set by using the operator (14). Finally, we write A as $A = L + S$. We fix $m = 200$ and define $\rho_r = \mathrm{rank}(L)/m$, where $\mathrm{rank}(L)$ varies. We choose the sparsity level $\alpha \in (0, 1)$ by using standard MATLAB command $\texttt{linspace}(0, 0.99, 40)$ by dividing the interval $[0, 0.99]$ into 40 subintervals.

Phase transition experiments. For each pair of (ρ_r, α), we apply iEALM, RPCA NCF, and our algorithm to recover the pair $(\hat{L}, \hat{S})$. For iEALM, we set $\lambda = 1/\sqrt{m}$ and use $\mu = 1.25/\|A\|_2$ and $\rho = 1.5$, where $\|A\|_2$ is the spectral norm (maximum singular value) of A. For a given $\epsilon > 0$, if the recovered matrix pair $(\hat{L}, \hat{S})$, satisfies the relative error $\frac{\|A - \hat{L} - \hat{S}\|}{\|A\|} < \epsilon$ then we consider the construction is viable. We set $\epsilon = 0.01$. In Figure 2, we produce the *phase transition diagrams* to show the fraction of perfect recovery of A, where white denotes *success* and black denotes *failure*. We run the experiments five times and plot the results. The success of iEALM is approximately below the line $\rho_r + \alpha \approx 0.3$. This is consistent with what reported in [3]. On the other hand, we note that the performance of our best pair RPCA is almost similar to that of [9] when the sparsity level α is small, and both approaches can efficiently provide a feasible reconstruction for any ρ_r in that case. We also note that for low sparsity level, iEALM can only provide a feasible reconstruction for



(a) (b) (c)

Fig. 3: RMSE to compare between RPCA GD, RPCA feasibility, and best-pair RPCA with respect to rank and error sparsity for (a) $\Omega = 0.9(mn)$, (b) $0.5(mn)$, and (b) $0.25(mn)$. We set $\rho_r = \text{rank}(L)/m$ and α is the sparsity measure. We have $(\rho_r, \alpha) \in (0.025, 1] \times (0, 0.3]$ with $r = 5 : 5 : 200$ and $\alpha = \texttt{linspace}(0, 0.3, 40)$.

$\rho_r \leq 0.3$. Due to their robustness to any low-rank structure when α is low, RPCA NCF and best pair RPCA can be proved to be effective in many real-world applications. In many real-world problems, involving the video/image data can ideally have any inherent low-rank structure and are generally corrupted by sparse outliers of arbitrary large magnitudes. In those instances, RPCA NCF and our best pair RPCA could be very useful. We justify this more in the later section.

Root mean square error measure. To validate our performance against RPCA GD of [17] and RPCA NCF on missing data case, we use a different metric — root mean square error (RMSE) and relative error in recovering the low-rank matrix on the set of observable entries Ω. Since RPCA GD does not explicitly recover a sparse matrix, S, it is unjustified to test it against the same relative error as in the previous section. Therefore, for the true low-rank, L, and a low-rank recovery, $\hat{L}$, we use the metric $\frac{\|P_\Omega(L-\hat{L})\|}{\sqrt{mn}}$ as the measure of RMSE and $\frac{\|P_\Omega(L-\hat{L})\|}{\|P_\Omega(L)\|}$ as relative error. From Figure 3, we can conclude that our best pair RPCA has least RMSE compare to that of RPCA GD and RPCA NCF. Moreover, the RMSE remains *mostly* unaltered as the cardinality of support set, Ω decreases. See Figure 6 in the appendix where we produce the *phase transition diagrams* for relative error for all three methods. We consider for each pair of (ρ_r, α), the construction is viable (and hence white in the diagram) if $\frac{\|P_\Omega(L-\hat{L})\|}{\|P_\Omega(L)\|} < 0.01$, otherwise the black denotes *failure*.

Computational time. In Figure 1, we compare the wall-clock execution time between RPCA NCF, RPCA GD, and RPCA Best pair. For simplicity, the test matrices are square matrices of size $m \times m$ with original rank $10\%m$, and we perform ten independent runs of each method on each test instance and plot the average time. As the matrix size increases, the execution time of RPCA NCF grows almost quadratically. However, in all cases, the execution time of our best pair formulation is very modest.

Hyperparameter settings. We provide numerical experiments on the sensitivity of Algorithm 1 to the choice of hyperparameters γ, a_k, b_k, r, and α in Appendix E. Additionally, we also discussed the sensitivity of Algorithm 1 on different choice of the starting point.

2) Background estimation from video sequences: Background estimation or moving object tracking [51], [52], [53], [54], [55], [56], [57] is considered as one of the classic problems in computer vision and is used as a crucial component in human activity recognition, tracking, and video analysis from surveillance cameras. When the video is captured by a static camera, minimizing the rank of the matrix $A \in \mathbb{R}^{m\times n}$, that concatenates n video frames (after converting them into vectors) represents the structure of the linear subspace, L, that contains the background and an error, S, that emphasizes the foreground components. However, the exact desired rank is often tuned empirically, as the ideal rank-one background is often unrealistic as the changing illumination, occluded foreground/background objects, reflection, and noise are typically also a part of the video frames. Based on the above observation, we note that the problem can be cast typically as (4). However, as we explained in some cases when the target rank and the sparsity level are user-defined hyperparameters, one might use a different approach as in [11], [9], [17] as well. Additionally, there might be missing/unobserved pixels in the video, and that makes the problem more complex and only a few methods, such as RPCA NCF, GRASTA [58], RPCA GD remedy to that. Therefore, we tested our best pair RPCA to a wide range of methods. In our experiments, we use *two* different video sequences: (i) the `Basic` sequence from Stuttgart synthetic dataset [59], (ii) the `waving tree` sequence [60]. We extensively use the Stuttgart video sequence as it is a challenging sequence that comprises both static and dynamic foreground objects and varying illumination in the background. Additionally, it comes with foreground ground truth for each frame. For iEALM and APG, we set the parameters the same as in the previous sections. For Best pair RPCA, RPCA GD, RPCA NCF, and GoDec, we set $r = 2$, target sparsity 10% and additionally, for GoDec, we set $q = 2$. For GRATSA, we set the parameters the same as those mentioned in the authors' website[2]. The qualitative analysis on the background and foreground recovered on both, full observation (in Figure 4) and partial observation (in Figure 5), suggest that our method recovers a visually better quality background and foreground compare to the other methods.

[2]https://sites.google.com/site/hejunzz/grasta

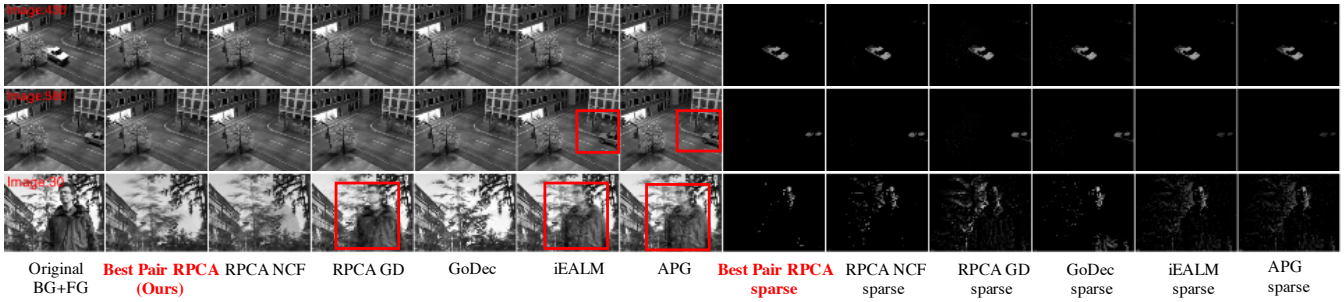


Fig. 4: Background estimation from video sequences. Except Best pair RPCA, RPCA NCF, and GoDec all other methods struggle to remove the static foreground object.

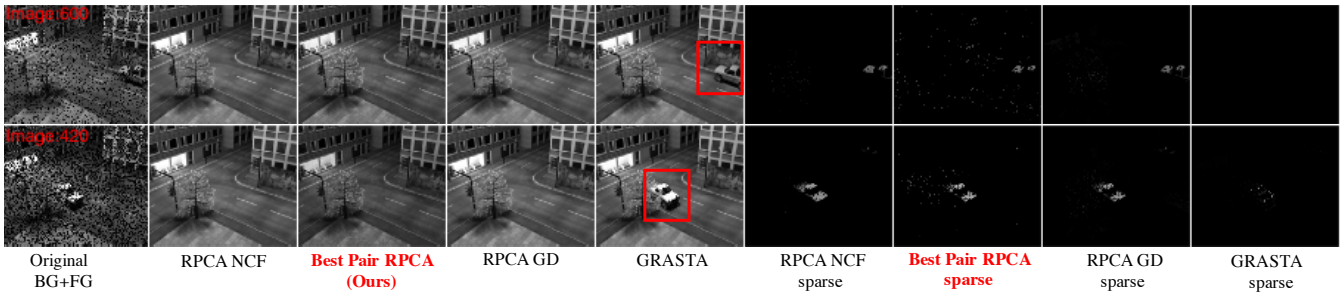


Fig. 5: Background estimation on subsampled Stuttgart `Basic` video sequence. We use $\Omega = 0.9(m.n)$ and $\Omega = 0.8(m.n)$, respectively.

Note that, RPCA GD recovers a fragmentary foreground with more false positives compare to our method; moreover, RPCA GD, GRASTA, iEALM, and APG cannot remove the static foreground object. Additionally, we note that to reach a 10^{-4} accurate solution while processing 600 frames of resolution 144×176 of the `Basic` sequence, APG and iEALM require an average of 83.91 seconds and 50.45 seconds, respectively. While on the same task, RPCA NCF takes about 32.52 seconds, GoDec takes 24.98 seconds, RPCA GD takes 27.43 seconds, and our best pair takes 34.8 seconds.

a) Quantitative measures: We used a metric called ϵ-proximity metric or $d_\epsilon(X, Y)$ to determine the quality of the n recovered foreground frames, $X = (X_1\ X_2 \cdots X_n)$, and the available foreground ground-truth, $Y = (Y_1\ Y_2 \cdots Y_n)$, where $X_i, Y_i \in \mathbb{R}^m$ are vectorized frames. To calculate the ϵ-proximity metric, the user can vary the threshold ϵ in the interval $[0, 1]$ and compare between the set of n normalized frames X and Y. Each point in $d_\epsilon(X, Y)$ represents the percentage of the pixels in the recovered frames, X, that is ϵ-close to the ground truth frames, Y. Therefore, $\epsilon = 0$ denotes exact recovery and $0 \leq d_\epsilon(X, Y) \leq 1$. We refer the readers to [9] for a detailed overview of the metric. Additionally, we used another state-of-the-art measure called the mean structural similarity index measure (SSIM) by [61]. We provide a quantitative evaluation of our best pair RPCA with respect to the ϵ-proximity metric–$d_\epsilon(X, Y)$ and the mean SSIM in recovering the foreground objects in Figures 7 and 8 in the appendix. From 7 (a), we observe that as ϵ increases, all the methods perform almost similarly. However, for $\epsilon = 0$, RPCA-GD is unable to obtain an exact recovery of the foreground pixel values and hence obtains $d_\epsilon(X, Y) = 0$ when $\epsilon = 0$. Except for RPCA GD, all the other methods recover more-or-less the same exact value of the foreground pixels for all $\epsilon \in [0, 1]$. However, with respect to the mean SSIM metric as in 7 (b), all methods are performing equally well by yielding similar mean SSIM across 600 frames of Stuttgart `Basic` sequence. In Figure 8, we compared the performance of our best pair RPCA on partially observed data on `Basic` sequence. We note that, as the size of the support set Ω decreases, the performance of RPCA GD suffers. Although, both RPCA best pair and RPCA NCF maintain a stable performance across all Ω.

3) Removal of shadows and specularities.: Set of images of an object under unknown pose and arbitrary lighting conditions, form a convex cone in the space of all possible images which may have *unbounded dimension* [62], [63]. However, the images under distant, isotropic lighting can be approximated by a 9-dimensional linear subspace, which is popularly referred to as the *harmonic plane*. We used three subjects `B11`,`B12`, and `B13` from the `Extended Yale Face Database` [64] for our simulations. We used 63 downsampled images of resolution of 120×160 of each subject. For APG and iEALM, we set the parameters the same as in the previous section. For RPCA GD, RPCA NCF, and our method, we set target rank $r = 9$ and sparsity level to 0.1. The qualitative analysis on the recovered images from Figure 9 shows while RPCA GD recovers patchy and granular face images, our best pair reformulation provides comparable reconstruction to that of iEALM, APG, and RPCA NCF.

IV. Conclusion and future work

In this paper, we presented the classical robust PCA problem as a non-convex best pair problem and proposed an accelerated proximal gradient algorithm to solve it. Additionally, we established a global convergence guarantee of our algorithm, which is generally not feasible for non-convex set projection-type algorithms. However, the approach and theory we develop extends beyond robust PCA and can be considered as a standalone model — any set feasibility problem that consists of two sets can be formulated as the best pair problem, following our model. Moreover, for any best pair problem that consists of convex and non-convex feasibility set, Algorithm 1 applies along with its theory. Together with our theoretical contribution, we validated the performance of our proposed algorithm on synthetic and real-world datasets and compared against several state-of-the-art robust PCP algorithms. We note that, in addition to our extensive experiments on some classical problems in computer vision presented in this paper, one can use our approach in high dynamic range imaging, inpainting, outlier rejection for photometric stereo, web data ranking, bioinformatic data analysis, fMRI imaging, hyperspectral image denoising, spectral clustering ([2], [5], [21], [9], [65], [7], [66], [67]), and only a few to mention.

Appendix A
Organization

The organization of the appendix is: Proofs for the global convergence result of Algorithm 1 is provided in Appendix, Section B; The proof of local linear convergence and a numerical example are provided in Appendix, Section C.

Appendix B
Proof of the global convergence

For convenience, define $\Delta_k \overset{\text{def}}{=} \|Y_k - Y_{k-1}\|$.

Lemma 8. *For the update of Y_{k+1} in* (13)*, let $G_{k+1} \overset{\text{def}}{=} \frac{1}{\gamma}(U_k - Y_{k+1}) - \nabla\mathcal{F}(V_k) + \nabla\mathcal{F}(Y_{k+1})$. Then $G_{k+1} \in \partial\Phi(Y_{k+1})$ and $\|G_{k+1}\| \leq \big(\frac{1}{\gamma} + L\big)\Delta_{k+1} + (\frac{a_k}{\gamma} + b_k L)\Delta_k$.*

Proof. From the definition of proximity operator and the update of Y_{k+1} (13), we have $U_k - \gamma\nabla\mathcal{F}(V_k) - Y_{k+1} \in \gamma\partial\mathcal{R}(Y_{k+1})$. Adding $\gamma\nabla\mathcal{F}(Y_{k+1})$ to both sides, we obtain $G_{k+1} = \frac{U_k - \gamma\nabla\mathcal{F}(V_k) - Y_{k+1} + \gamma\nabla\mathcal{F}(Y_{k+1})}{\gamma} \in \partial\Phi(Y_{k+1})$. Applying further the triangle inequality together with the Lipschitz continuity of $\nabla\mathcal{F}$, we get

$$\begin{aligned}\|G_{k+1}\| &\leq \tfrac{1}{\gamma}\|U_k - Y_{k+1}\| + L\|V_k - Y_{k+1}\| \\ &\leq \tfrac{1}{\gamma}(\Delta_{k+1} + a_k\Delta_k) + L(\Delta_{k+1} + b_k\Delta_k),\end{aligned}$$

which concludes the proof. □

Lemma 9. *For Algorithm 1, given the parameters γ, a_k, b_k, there holds $\Phi(Y_{k+1}) + \underline{\beta}\Delta_{k+1}^2 \leq \Phi(Y_k) + \bar{\alpha}\Delta_k^2$.*

Proof. Let $\mathcal{L}_k(Y) \overset{\text{def}}{=} \gamma\mathcal{R}(Y) + \frac{1}{2}\|Y - U_k\|^2 + \gamma\langle Y, \nabla\mathcal{F}(V_k)\rangle$, then there holds

$$\min_{Y\in\mathbb{R}^n} \mathcal{L}_k(Y) \Leftrightarrow \min_{Y\in\mathbb{R}^n} \gamma\mathcal{R}(Y) + \tfrac{1}{2}\|Y - (U_k - \gamma\nabla\mathcal{F}(V_k))\|^2.$$

Owing to the definition of proximity operator, for the update of Y_{k+1} in (13) we have

$$Y_{k+1} \in \operatorname{argmin}_{Y\in\mathbb{R}^n} \mathcal{L}_k(Y), \tag{18}$$

which means that $\mathcal{L}_k(Y_{k+1}) \leq \mathcal{L}_k(Y_k)$. That is

$$\begin{aligned}&\mathcal{R}(Y_{k+1}) + \tfrac{1}{2\gamma}\|Y_{k+1} - U_k\|^2 + \langle Y_{k+1}, \nabla\mathcal{F}(V_k)\rangle \\ &\leq \mathcal{R}(Y_k) + \tfrac{1}{2\gamma}\|Y_k - U_k\|^2 + \langle Y_k, \nabla\mathcal{F}(V_k)\rangle.\end{aligned}$$

Therefore, we get

$$\begin{aligned}\mathcal{R}(Y_k) &\geq \mathcal{R}(Y_{k+1}) + \tfrac{1}{2\gamma}\|Y_{k+1} - Y_k + Y_k - U_k\|^2 \\ &\quad + \langle Y_{k+1} - Y_k, \nabla\mathcal{F}(V_k)\rangle - \tfrac{1}{2\gamma}\|Y_k - U_k\|^2 \\ &= \mathcal{R}(Y_{k+1}) + \langle Y_{k+1} - Y_k, \nabla\mathcal{F}(Y_k)\rangle + \tfrac{1}{2\gamma}\Delta_{k+1}^2 \\ &\quad - \tfrac{a_k}{\gamma}\langle Y_k - Y_{k+1}, Y_k - Y_{k-1}\rangle \\ &\quad + \langle Y_{k+1} - Y_k, \nabla\mathcal{F}(V_k) - \nabla\mathcal{F}(Y_k)\rangle.\end{aligned} \tag{19}$$

Since $\nabla\mathcal{F}$ is L-Lipschitz, then

$$\langle\nabla\mathcal{F}(Y_k), Y_{k+1} - Y_k\rangle \geq \mathcal{F}(Y_{k+1}) - \mathcal{F}(Y_k) - \tfrac{L}{2}\Delta_{k+1}^2.$$

For the inner product $\langle Y_k - Y_{k+1}, Y_k - Y_{k-1}\rangle$, applying the Pythagorean relation $2\langle c_1 - c_2, c_1 - c_3\rangle = \|c_1 - c_2\|^2 + \|c_1 - c_3\|^2 - \|c_2 - c_3\|^2$, we get $\langle Y_k - Y_{k+1}, Y_k - Y_{k-1}\rangle \leq \frac{1}{2}\big(\|Y_k - Y_{k+1}\|^2 + \|Y_k - Y_{k-1}\|^2\big)$. Using further Young's inequality with $\nu > 0$ we obtain

$$\begin{aligned}&\langle Y_{k+1} - Y_k, \nabla\mathcal{F}(V_k) - \nabla\mathcal{F}(Y_k)\rangle \\ &\geq -(\tfrac{\nu}{2}\Delta_{k+1}^2 + \tfrac{1}{2\nu}\|\nabla\mathcal{F}(V_k) - \nabla\mathcal{F}(Y_k)\|^2) \geq -(\tfrac{\nu}{2}\Delta_{k+1}^2 + \tfrac{b_k^2L^2}{2\nu}\Delta_k^2).\end{aligned} \tag{20}$$

Combining the above 3 inequalities with (19) leads to

$$\begin{aligned}\mathcal{R}(Y_k) &\geq \mathcal{R}(Y_{k+1}) + \mathcal{F}(Y_{k+1}) - \mathcal{F}(Y_k) - \tfrac{L}{2}\Delta_{k+1}^2 \\ &\quad + \tfrac{1}{2\gamma}\Delta_{k+1}^2 - \tfrac{a_k}{2\gamma}\|Y_k - Y_{k+1}\|^2 \\ &\quad - \tfrac{a_k}{2\gamma}\|Y_k - Y_{k-1}\|^2 - \tfrac{\nu}{2}\Delta_{k+1}^2 - \tfrac{b_k^2L^2}{2\nu}\Delta_k^2,\end{aligned} \tag{21}$$

hence $\Phi(Y_{k+1}) + \frac{1-\gamma L - a_k - \nu}{2\gamma}\Delta_{k+1}^2 \leq \Phi(Y_k) + \frac{\gamma b_k^2L^2 + \nu a_k}{2\nu\gamma}\Delta_k^2$. Recalling the definition of $\underline{\beta}, \bar{\alpha}$ concludes the proof. □

Define $\mathcal{H}$ the product space $\mathcal{H} \overset{\text{def}}{=} \mathbb{R}^n \times \mathbb{R}^n$ and $Z_k = (Y_k, Y_{k-1}) \in \mathcal{H}$. Given Z_k, define the function

$$\Psi(Z_k) \overset{\text{def}}{=} \Phi(Y_k) + \bar{\alpha}\Delta_k^2,$$

which is is a KL function if Φ is. Denote $\mathcal{C}_{Y_k}, \mathcal{C}_{Z_k}$ the set of cluster points of sequences $\{Y_k\}_{k\in\mathbb{N}}$ and $\{Z_k\}_{k\in\mathbb{N}}$ respectively, and $\operatorname{crit}(\Psi) = \{Z = (Y, Y) \in \mathcal{H} : Y \in \operatorname{crit}(\Phi)\}$.

Lemma 10. *For Algorithm 1, choose ν, γ, a_k, b_k such that* (17) *holds. If Φ is bounded from below, then*

(i) $\sum_{k\in\mathbb{N}} \Delta_k^2 < +\infty$;
(ii) $\Psi(Z_k)$ *is monotonically decreasing and convergent;*
(iii) *The sequence $\Phi(Y_k)$ is convergent.*

Proof. Define $\delta = \underline{\beta} - \bar{\alpha} > 0$, from Lemma 9, we have $\delta\Delta_{k+1}^2 \leq \big(\Phi(Y_k) - \Phi(Y_{k+1})\big) + \bar{\alpha}(\Delta_k^2 - \Delta_{k+1}^2)$. Let $Y_{-1} = Y_0$ and the above inequality over k:

$$\begin{aligned}\delta\textstyle\sum_{k\in\mathbb{N}} \Delta_{k+1}^2 &\leq \textstyle\sum_{k\in\mathbb{N}}\big(\Phi(Y_k) - \Phi(Y_{k+1})\big) + \sum_{k\in\mathbb{N}}\bar{\alpha}(\Delta_k^2 - \Delta_{k+1}^2) \\ &\leq \Phi(Y_0) + \bar{\alpha}\textstyle\sum_{k\in\mathbb{N}}(\Delta_k^2 - \Delta_{k+1}^2) \\ &= \Phi(Y_0) + \bar{\alpha}\Delta_0^2 = \Phi(Y_0),\end{aligned}$$

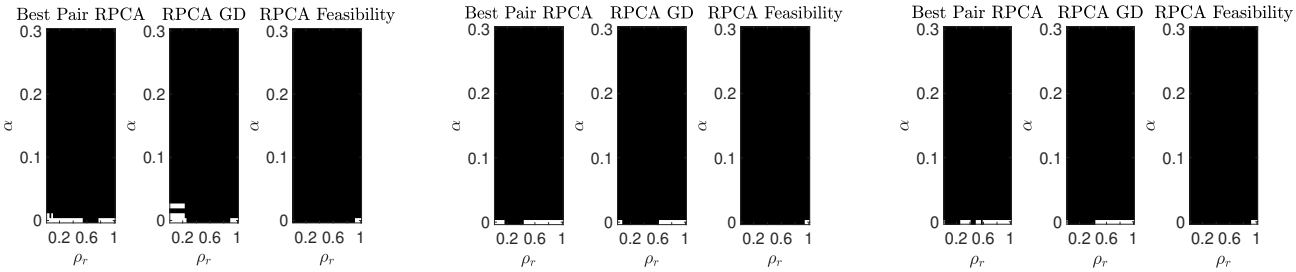


Fig. 6: Phase transition for $\frac{\|P_\Omega(L-\hat{L})\|}{\|P_\Omega(L)\|} < 10^{-2}$ to compare between RPCA GD, RPCA feasibility, and best-pair RPCA with respect to rank and error sparsity for (a) $\Omega = 0.9(mn)$, (b) $0.5(mn)$, and (c) $0.25(mn)$. We set $\rho_r = \mathrm{rank}(L)/m$ and α is the sparsity measure. We have $(\rho_r, \alpha) \in (0.025, 1] \times (0, 0.3]$ with $r = 5:5:200$ and $\alpha = \texttt{linspace}(0, 0.3, 40)$.

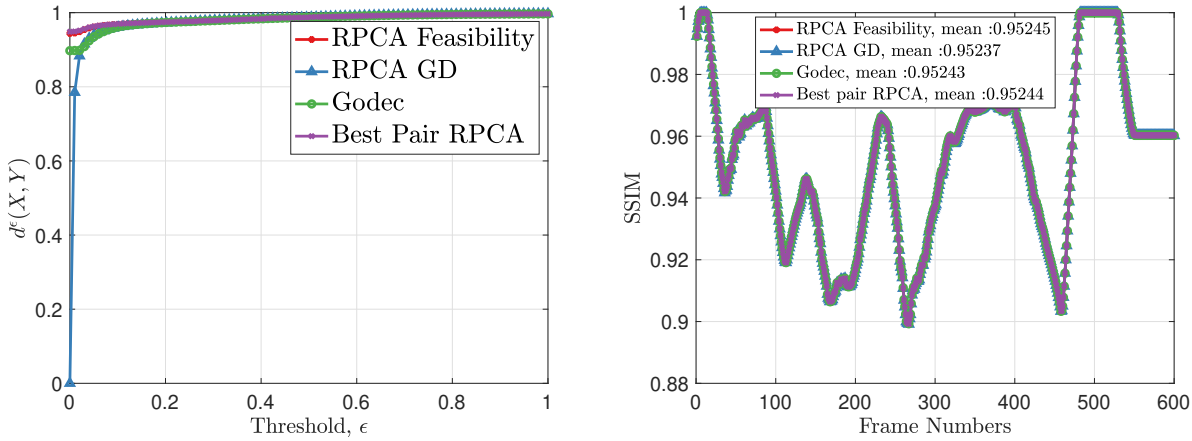


Fig. 7: Quantitative comparison between different algorithms on Stuttgart `Basic` sequence. We compare the recovered foreground by different methods with respect to the foreground GT available for each frame on two different metrics: ϵ-proximity metric–$d_\epsilon(X,Y)$ as in [9] and structural similarity index measure by [61]. In recovering the foreground objects, our best pair RPCA is as robust as the other baseline methods with respect to both metrics.

which means, as $\Phi(Y_0)$ is bounded, $\sum_{k\in\mathbb{N}}\Delta_{k+1}^2 \le \frac{\Phi(Y_0)}{\delta} < +\infty$. From Lemma 9, by pairing terms on both sides in Lemma 9, we get $\Psi(Z_{k+1})+(\underline{\beta}-\overline{\alpha})\Delta_{k+1}^2 \le \Psi(Z_k)$. Since we assume $\underline{\beta}-\overline{\alpha} > 0$, hence $\Psi(Z_k)$ is monotonically non-increasing. The convergence of $\Phi(Y_k)$ is straightforward. □

Lemma 11. *For Algorithm 1, choose ν, γ, a_k, b_k such that* (17) *holds. If Φ is bounded from below and $\{Y_k\}_{k\in\mathbb{N}}$ is bounded, then Y_k converges to a critical point of Φ.*

Proof. Since $\{Y_k\}_{k\in\mathbb{N}}$ is bounded, there exists a subsequence $\{Y_{k_j}\}_{k_j\in\mathbb{N}}$ and cluster point $\overline{Y}$ s.t. $Y_{k_j} \to \overline{Y}$ as $j \to +\infty$. Next we show $\Phi(Y_{k_j}) \to \Phi(\overline{Y})$ and $\overline{Y}$ is a critical point of Φ.

Since $\mathcal{R}$ is lsc, we have $\liminf_{j\to+\infty}\mathcal{R}(Y_{k_j}) \ge \mathcal{R}(\overline{Y})$. From (18), we have $\mathcal{L}_{k_j-1}(Y_{k_j}) \le \mathcal{L}_{k_j-1}(\overline{Y})$ and thus

$$\begin{aligned}\mathcal{R}(\overline{Y}) &\ge \mathcal{R}(Y_{k_j}) + \tfrac{1}{2\gamma}\|Y_{k_j} - U_{k_j-1}\|^2 \\ &\quad + \langle Y_{k_j} - \overline{Y}, \nabla\mathcal{F}(V_{k_j-1})\rangle - \tfrac{1}{2\gamma}\|\overline{Y} - U_{k_j-1}\|^2 \\ &= \mathcal{R}(Y_{k_j}) + \tfrac{1}{2\gamma}(\|Y_{k_j} - \overline{Y}\|^2 + 2\langle Y_{k_j} - \overline{Y}, \overline{Y} - U_{k_j-1}\rangle) \\ &\quad + \langle Y_{k_j} - \overline{Y}, \nabla\mathcal{F}(V_{k_j-1})\rangle.\end{aligned}$$

Taking the limit of the above inequality and using $\Delta_k^2 \to 0$, $Y_{k_j} \to \overline{Y}$, we get $\limsup_{j\to+\infty}\mathcal{R}(Y_{k_j}) \le \mathcal{R}(\overline{Y})$. As a result, $\lim_{k\to+\infty}\mathcal{R}(Y_{k_j}) = \mathcal{R}(\overline{Y})$. Since $\mathcal{F}$ is continuous, then $\mathcal{F}(Y_{k_j}) \to \mathcal{F}(\overline{Y})$, hence $\Phi(Y_{k_j}) \to \Phi(\overline{Y})$.

Furthermore, owing to Lemma 8, $G_{k_j} \in \partial\Phi(Y_{k_j})$, and (i) of Lemma 10 we have $G_{k_j} \to 0$ as $k \to +\infty$. Therefore, as $j \to +\infty$, we have

$$G_{k_j} \in \partial\Phi(Y_{k_j}),\ \ (Y_{k_j}, G_{k_j}) \to (\overline{Y}, 0) \ \text{ and } \ \Phi(Y_{k_j}) \to \Phi(\overline{Y}).$$

Hence $0 \in \partial\Phi(\overline{Y})$, *i.e.* $\overline{Y}$ is a critical point. □

Proof of Theorem 3. Putting together the above lemmas, we draw the following useful conclusions:

- **(C.1)** Denote $\delta = \underline{\beta}-\overline{\alpha}$, then $\Psi(Z_{k+1})+\delta\Delta_{k+1}^2 \le \Psi(Z_k)$;
- **(C.2)** Define $w_k \overset{\text{def}}{=} \begin{pmatrix} G_k + 2\overline{\alpha}(Y_k - Y_{k-1}) \\ 2\overline{\alpha}(Y_{k-1} - Y_k)\end{pmatrix}$, then $w_k \in \partial\Psi(Z_k)$. Owing to Lemma 8, there exists a $\sigma > 0$ such that $\|w_k\| \le \sigma(\Delta_k + \Delta_{k-1})$;
- **(C.3)** if Y_{k_j} is a subsequence such that $Y_{k_j} \to \overline{Y}$, then $\Psi(Z_k) \to \Psi(\overline{Z})$ where $\overline{Z} = (\overline{Y}, \overline{Y})$.
- **(C.4)** $\mathcal{C}_{Z_k} \subseteq \mathrm{crit}(\Psi)$;
- **(C.5)** $\lim_{k\to+\infty}\mathrm{dist}(Z_k, \mathcal{C}_{Z_k}) = 0$;
- **(C.6)** $\mathcal{C}_{Z_k}$ is non-empty, compact and connected;
- **(C.7)** Ψ is finite and constant on $\mathcal{C}_{Z_k}$.

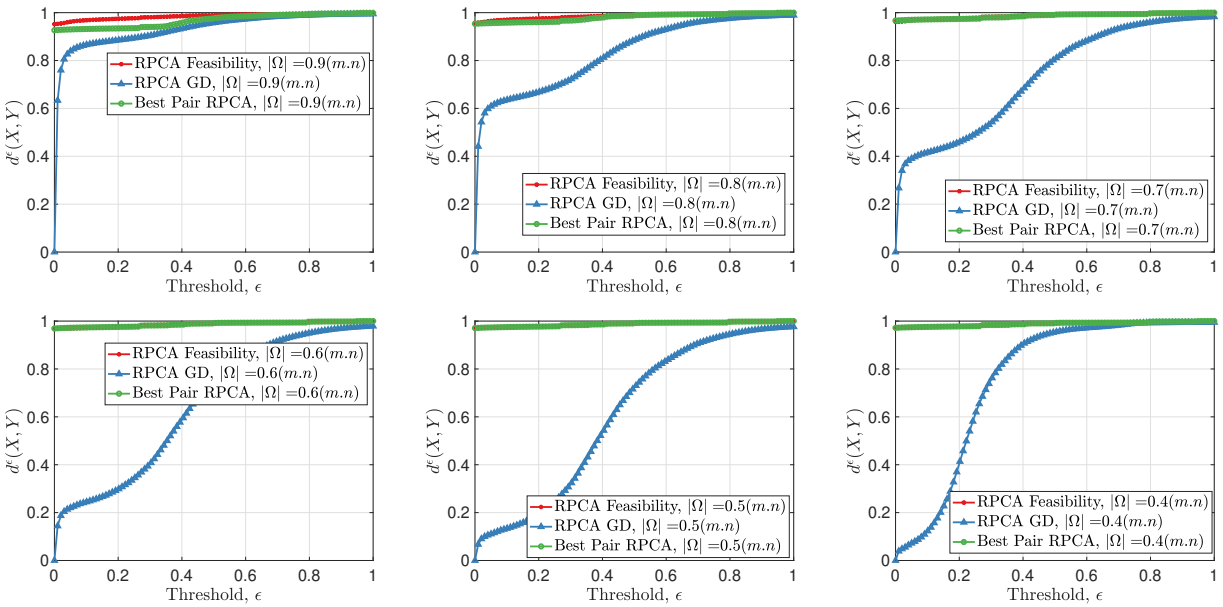


Fig. 8: Quantitative comparison of foreground recovered by best pair RPCA, RPCA GD, and RPCA feasibility. Stuttgart `Basic` sequence, frame size 144×176 with observable entries: (a) $|\Omega| = 0.9(m.n)$, (b) $|\Omega| = 0.8(m.n)$, (c) $|\Omega| = 0.7(m.n)$, (d) $|\Omega| = 0.6(m.n)$, (e) $|\Omega| = 0.5(m.n)$, and (f) $|\Omega| = 0.4(m.n)$. The performance of RPCA GD drops significantly as $|\Omega|$ decreases. In contrast, the performance of our best pair RPCA and RPCA NCF stay stable irrespective of the size of $|\Omega|$.

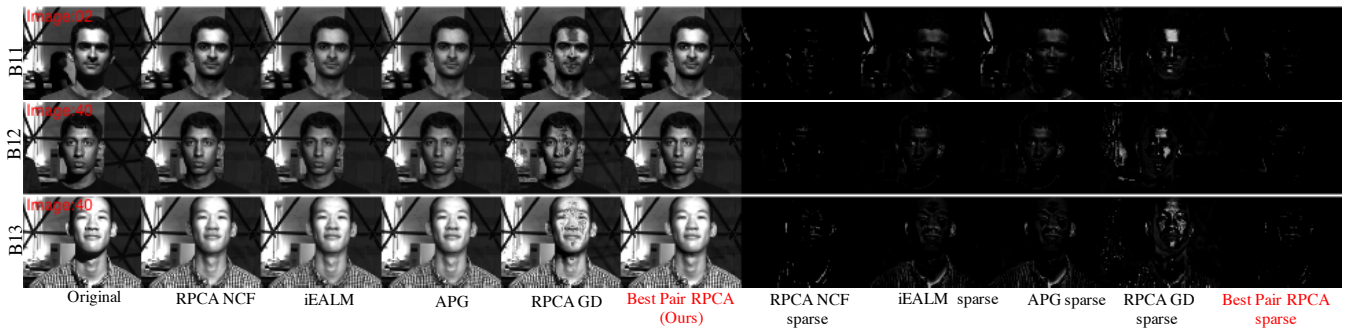


Fig. 9: Shadow and specularities removal from face images captured under varying illumination and camera position. Our feasibility approach provides comparable reconstruction to that of iEALM, APG, and RPCA NCF.

Next we prove the claims of Theorem 3.

(a) Let $\overline{Y} \in \mathrm{crit}(\Phi)$ be a critical point, then $\overline{Z} = (\overline{Y}, \overline{Y}) \in \mathcal{C}_{Z_k}$. Then owing to (**C.3**), we have $\Psi(Z_k) \to \Psi(\overline{Z})$.

Suppose there exists K such that $\Psi(Z_K) = \Psi(\overline{Z})$. Then, the descent property (**C.1**) implies that $\Psi(Z_k) = \Psi(\overline{Z})$ holds for all $k \geq K$. Thus, Z_k is constant for $k \geq K$, hence has finite length.

On the other hand, suppose that $\psi_k \stackrel{\text{def}}{=} \Psi(Z_k) - \Psi(\overline{Z}) > 0$. Owing to (**C.6**), (**C.7**) and Definition 2, the KL property of Ψ implies that there exist ϵ, η and a concave function φ, and $\mathcal{U} \stackrel{\text{def}}{=} \big\{S \in \mathcal{H} : \mathrm{dist}(S, \mathcal{C}_{Z_k}) < \epsilon\big\} \cap \big[\Psi(\overline{Z}) < \Psi(S) < \Psi(\overline{Z}) + \eta\big]$ such that for all $Z \in \mathcal{U}$:

$$\varphi'\big(\Psi(Z) - \Psi(\overline{Z})\big)\mathrm{dist}\big(0, \partial\Psi(Z)\big) \geq 1. \tag{22}$$

Let $k_1 \in \mathbb{N}$ be such that $\Psi(Z_k) < \Psi(\overline{Z}) + \eta$ holds for all $k \geq k_1$. Owing to (**C.5**), there exists another $k_2 \in \mathbb{N}$ such that $\mathrm{dist}(Z_k, \mathcal{C}_{Z_k}) < \epsilon$ holds for all $k \geq k_2$. Let $K = \max\{k_1, k_2\}$. Then $Z_k \in \mathcal{U}$ holds for all $k \geq K$. Furthermore using (22), we have for $k \geq K$

$$\varphi'(\psi_k)\mathrm{dist}\big(0, \partial\Psi(Z_k)\big) \geq 1.$$

Note that φ is concave, hence φ' is decreasing. As $\Psi(Z_k)$ is decreasing too, we have

$$\begin{aligned}\varphi(\psi_k) - \varphi(\psi_{k+1}) &\geq \varphi'(\psi_k)\big(\Psi(Z_k) - \Psi(Z_{k+1})\big) \\ &\geq \tfrac{\Psi(Z_k) - \Psi(Z_{k+1})}{\mathrm{dist}(0, \partial\Psi(Z_k))}.\end{aligned}$$

From (**C.2**), since $\mathrm{dist}(0, \partial\Psi(Z_k)) \leq \|w_k\|$, we get

$$\varphi(\psi_k) - \varphi(\psi_{k+1}) \geq \tfrac{\Psi(Z_k) - \Psi(Z_{k+1})}{\|w_k\|} \geq \tfrac{\Psi(Z_k) - \Psi(Z_{k+1})}{\sigma(\Delta_k + \Delta_{k-1})}.$$

Moreover, (**C.1**) yields $\Psi(Z_k) - \Psi(Z_{k+1}) \geq \delta\Delta_{k+1}^2$ and thus $\varphi(\psi_k) - \varphi(\psi_{k+1}) \geq \frac{\delta\Delta_{k+1}^2}{\sigma(\Delta_k + \Delta_{k-1})}$ which yields

$$\Delta_{k+1}^2 \leq \big(\tfrac{\sigma}{\delta}(\varphi(\psi_k) - \varphi(\psi_{k+1}))\big)(\Delta_k + \Delta_{k-1}). \tag{23}$$

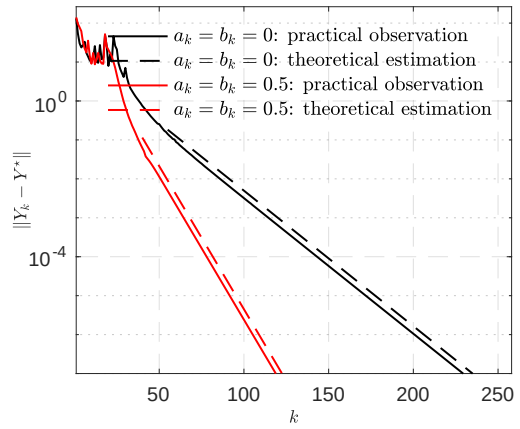


Fig. 10: Local linear convergence of Algorithm 1.

Taking the square root of both sides and applying Young's inequality we further obtain

$$2\Delta_{k+1} \leq \tfrac{1}{2}(\Delta_k + \Delta_{k-1}) + \tfrac{2\sigma}{\delta}\big(\varphi(\psi_k) - \varphi(\psi_{k+1})\big). \tag{24}$$

Summing up both sides over k and using $Y_0 = Y_{-1}$ yield

$$\ell \overset{\text{def}}{=} \sum_{k\in\mathbb{N}} \Delta_k \leq \Delta_1 + \tfrac{2\sigma}{\delta}\varphi(\psi_0) < +\infty,$$

which concludes the finite length property of Y_k.

(b) Then the convergence of the sequence follows from the fact that $\{Y_k\}_{k\in\mathbb{N}}$ is a Cauchy sequence, hence convergent. Owing to Lemma 11, there exists a critical point $Y^\star \in \mathrm{crit}(\Phi)$ such that $\lim_{k\to+\infty} Y_k = Y^\star$. □

APPENDIX C
PROOF OF LOCAL LINEAR CONVERGENCE

Before presenting the proof for local linear convergence, in Figure 10, we provide the comparison between the theoretical estimation of the convergence rate of Algorithm 1 and practical observation of the convergence speed of $\|Y_k - Y^\star\|$ when applied to solve a synthetic problem. The size of the problem is $\mathbb{R}^{32\times 32}$, which is small as larger size will make the rate estimation very difficult to compute. It can be observed that our theoretical rate estimation is very tight, given that the *dashed* lines and the *solid* ones are parallel to each other. We also compare two different settings of (a_k, b_k) that $a_k = b_k = 0$ and $a_k = b_k = 0.5$, it can be observed that the latter choice is faster than the former one. We refer to [50, Chapter 6] for detailed discussions on the benefits of non-zero a_k, b_k.

Since we are in the non-convex setting, we need the prox-regularity of the non-convexity. A lower semi-continuous function $\mathcal{R}$ is r-prox-regular at $\bar{x} \in \mathrm{dom}(\mathcal{R})$ for $\bar{v} \in \partial\mathcal{R}(\bar{x})$ if $\exists r > 0$ such that $\mathcal{R}(x') > \mathcal{R}(x) + \langle v,\, x' - x\rangle - \frac{1}{2r}\|x - x'\|^2$ $\forall x, x'$ near $\bar{x}$, $\mathcal{R}(x)$ near $\mathcal{R}(\bar{x})$ and $v \in \partial\mathcal{R}(x)$ near $\bar{v}$.

To prove Theorem 4, we rely on a so-called partial smoothness concept. Let $\mathcal{M} \subset \mathbb{R}^n$ be a C^2-smooth submanifold, let $\mathcal{T}_{\mathcal{M}}(x)$ the tangent space of $\mathcal{M}$ at any point $x \in \mathcal{M}$.

Definition 12. The function $\mathcal{R} : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is C^2-*partly smooth at* $\bar{x} \in \mathcal{M}$ *relative to* $\mathcal{M}$ *for* $\bar{v} \in \partial\mathcal{R}(\bar{x}) \neq \emptyset$ if $\mathcal{M}$ is a C^2-submanifold around $\bar{x}$, and

(i) (Smoothness): $\mathcal{R}$ restricted to $\mathcal{M}$ is C^2 around $\bar{x}$;
(ii) (Regularity): $\mathcal{R}$ is regular at all $x \in \mathcal{M}$ near $\bar{x}$ and $\mathcal{R}$ is r-prox-regular at $\bar{x}$ for $\bar{v}$;
(iii) (Sharpness): $\mathcal{T}_{\mathcal{M}}(\bar{x}) = \mathrm{par}(\partial\mathcal{R}(x))^\perp$;
(iv) (Continuity): The set-valued mapping $\partial\mathcal{R}$ is continuous at $\bar{x}$ relative to $\mathcal{M}$.

We denote the class of partly smooth functions at x relative to $\mathcal{M}$ for v as $\mathrm{PSF}_{x,v}(\mathcal{M})$. Partial smoothness was first introduced in [68] and its directional version stated here is due to [69], [70]. Prox-regularity is sufficient to ensure that the partly smooth submanifolds are locally unique [69, Corollary 4.12], [70, Lemma 2.3 and Proposition 10.12].

Proof of Theorem 4. First we have

- $\mathcal{Y}_L$ is the set of fixed-rank matrices, hence partly smooth.
- Since $\mathcal{S}$ is a subspace, hence it is partly smooth at $S^\star$ relative to any $W \in (\mathcal{S})^\perp$.

Under the conditions of Theorem 3, there exists a critical point $Y^\star$ such that $Y_k \to Y^\star$ and $\Phi(Y_k) \to \Phi(Y^\star)$.

Convergence properties of $\{Y_k\}_{k\in\mathbb{N}}$ (Theorem 3) entails $\|U_k - Y_k\| \to 0$ and $\|V_k - Y^\star\| \to 0$. In turn,

$$\mathrm{dist}\big(-\nabla\mathcal{F}(x^\star), \partial\mathcal{R}(Y_{k+1})\big) \leq \tfrac{1}{\gamma}\|U_k - Y_{k+1}\| + \|V_k - Y^\star\| \to 0.$$

Altogether, this shows that the conditions of [69, Theorem 4.10] or [70, Proposition 10.12] are fulfilled on $\mathcal{R}$ at $Y^\star$ for $-\nabla\mathcal{F}(Y^\star)$, and the identification result follows, that is $(\mathcal{Y}_{r,k}, \mathcal{Y}_{\alpha,k}) \in \mathcal{Y}_L \times \mathcal{S}$ for all k large enough, and we conclude the proof. □

a) Tangent space $T_{\mathcal{X}}^{X^\star}$: Given $X^\star \in \mathcal{X}$, the tangent space simply reads $KX = 0$. Let E be the kernel of K. Then we have the projection operator onto $KX = 0$ reads $\mathrm{P}_{T_{\mathcal{X}}^{X^\star}} = E(E^TE)^{-1}E^T$.

b) Tangent space of $\mathcal{Y}_L$: Let $\mathbb{M} = M_{m,n}(\mathbb{R})$ be the space of $m \times n$ matrices with the classical inner product $\langle A, B\rangle = \mathrm{Trace}(A^TB)$. The set of matrices with fixed rank r, $\mathcal{Y}_L = \big\{X \in \mathbb{M} : \mathrm{rank}(X) = r\big\}$ is a smooth manifold around any matrix $L \in \mathcal{Y}_L$. Given $L^\star$, with the help of the singular value decomposition $L = U\Sigma V^T$, the tagent space at L to $\mathcal{Y}_L$ is

$$T_{\mathcal{Y}_L}^{L^\star} = \big\{H \in \mathbb{M} \colon u_i^T H v_j = 0, \text{ for all } r < i \leq m, r < j \leq n\big\}.$$

Let $U = [u_1, u_2, \cdots, u_m], V = [v_1, v_2, \cdots, v_n]$ and Σ the diagonal matrix with singular value in decending order. Denote

$$\mathcal{L} = \big\{L \in \mathbb{M} : X = u_i^T v_j, \quad \text{for all } \{i,j\}_{1\leq i\leq m, 1\leq j\leq n} \setminus \{i,j\}_{r<i\leq m, r<j\leq n}\big\},$$

then $\mathcal{L}$ forms the basis of $\mathcal{T}$ and $\dim(\mathcal{L}) = mn - r^2$, there for define $Z \overset{\text{def}}{=} [L_1(:); L_2(:); \cdots; L_{mn-r^2}(:)]$, $L_i \in \mathcal{L}$ and $\mathrm{P}_{T_{\mathcal{Y}_L}^{L^\star}} \overset{\text{def}}{=} Z(Z^TZ)^{-1}Z^T$, then $\mathrm{P}_{T_{\mathcal{Y}_L}^{L^\star}}$ is the explicit form of the projection operator of projecting onto subspace $T_{\mathcal{Y}_L}^{L^\star}$.

c) Tangent space of $\mathcal{S}$: Given $S^\star \in \mathcal{S}$, denote the tangent space as $T_{\mathcal{S}}^{S^\star}$. Let $\mathrm{vec}(S^\star)$ be the vector form of $S^\star$, then $\mathrm{P}_{T_{\mathcal{S}}^{S^\star}} = \mathrm{diag}\big(|\mathrm{vec}(S^\star)| > 0\big)$. Finally, we have $\mathrm{P}_{T_{\mathcal{Y}}^{Y^\star}} = \begin{bmatrix} \mathrm{P}_{T_{\mathcal{S}}^{S^\star}} & \\ & \mathrm{P}_{T_{\mathcal{Y}_L}^{L^\star}} \end{bmatrix}$.

Proof of Theorem 6. From (13), when $a_k, b_k \equiv 0$, we have that $Y_{k+1} = \mathrm{P}_{\mathcal{Y}}\big(Y_k - \gamma(Y_k - \mathrm{P}_{\mathcal{X}}(Y_k))\big)$. Let $Y^\star$ be a critical point that Y_k converges to. Then, we get $Y^\star = \mathrm{P}_{\mathcal{Y}}\big(Y^\star - \gamma(Y^\star - \mathrm{P}_{\mathcal{X}}(Y^\star))\big)$. Denote $X_k = \mathrm{P}_{\mathcal{X}}(Y_k)$ and $X^\star = \mathrm{P}_{\mathcal{X}}(Y^\star)$, we have

$$\begin{aligned} X_k - X^\star &= \mathrm{P}_{T_{\mathcal{X}}^{X^\star}}(X_k - X^\star) = \mathrm{P}_{T_{\mathcal{X}}^{X^\star}}\mathrm{P}_{\mathcal{X}}(Y_k - Y^\star) \\ &= \mathrm{P}_{T_{\mathcal{X}}^{X^\star}}(Y_k - Y^\star) \\ &= \mathrm{P}_{T_{\mathcal{X}}^{X^\star}}\mathrm{P}_{T_{\mathcal{Y}}^{Y^\star}}(Y_k - Y^\star) + o(\|Y_k - Y^\star\|). \end{aligned}$$

Consider the difference of the above two equations, owing to Lemma 5, we get

$$\begin{aligned} &Y_{k+1} - Y^\star \\ &= \mathrm{P}_{\mathcal{Y}}\big((1-\gamma)Y_k + \gamma\mathrm{P}_{\mathcal{X}}(Y_k)\big) - \mathrm{P}_{\mathcal{Y}}\big((1-\gamma)Y^\star + \gamma\mathrm{P}_{\mathcal{X}}(Y^\star)\big) \\ &= \mathrm{P}_{T_{\mathcal{Y}}^{Y^\star}}\big((1-\gamma)Y_k + \gamma\mathrm{P}_{\mathcal{X}}(Y_k) - (1-\gamma)Y^\star - \gamma\mathrm{P}_{\mathcal{X}}(Y^\star)\big) \\ &\quad + o(\|Y_k - Y^\star\|) \\ &= \mathrm{P}_{T_{\mathcal{Y}}^{Y^\star}}\big((1-\gamma)(Y_k - Y^\star) + \gamma(X_k - X^\star)\big) + o(\|Y_k - Y^\star\|) \\ &= \mathrm{P}_{T_{\mathcal{Y}}^{Y^\star}}\big((1-\gamma)\mathrm{Id} + \gamma\mathrm{P}_{T_{\mathcal{X}}^{X^\star}}\big)\mathrm{P}_{T_{\mathcal{Y}}^{Y^\star}}(Y_k - Y^\star) + o(\|Y_k - Y^\star\|), \end{aligned}$$

which means $Y_{k+1} - Y^\star = \mathcal{P}(Y_k - Y^\star) + o(\|Y_k - Y^\star\|)$. Note that $\mathcal{P}$ is symmetric positive semi-definite, hence all its eigenvalues are real and lie in $[0, 1]$.

Now, assume that $b_k = a_k \equiv a$, then we have from (13)

$$\begin{aligned} Z_k &= (1+a)Y_k - aY_{k-1}, \\ Y_{k+1} &= \mathrm{P}_{\mathcal{Y}}\big(Z_k - \gamma(Z_k - \mathrm{P}_{\mathcal{X}}(Z_k))\big). \end{aligned}$$

Follow the derivation of $Y_{k+1} - Y^\star$ above, we get

$$\begin{aligned} &Y_{k+1} - Y^\star \\ &= (1+a)\mathcal{P}(Y_k - Y^\star) - a\mathcal{P}(Y_{k-1} - Y^\star) + o(\|Y_k - Y^\star\|) \\ &= \begin{bmatrix}(1+a)\mathcal{P} & -a\mathcal{P}\end{bmatrix}\begin{pmatrix} Y_k - Y^\star \\ Y_{k-1} - Y^\star \end{pmatrix} + o(\|Y_k - Y^\star\|). \end{aligned}$$

Plus the definition of D_k and the fact that $o(\|Y_k - Y^\star\|) = o(\|D_k\|)$, we obtain $D_{k+1} = \mathcal{Q}D_k + o(\|D_k\|)$. Owing to [50, Chapter 6], if $\rho_{\mathcal{P}} < 1$, then so is $\rho_{\mathcal{Q}} < 1$, and the linear convergence result follows. □

REFERENCES

[1] M. Udell, C. Horn, R. Zadeh, and S. Boyd, "Generalized low rank models," *Foundations and Trends in Machine Learning*, vol. 9, no. 1, pp. 1–118, 2016.

[2] Z. Lin, M. Chen, and Y. Ma, "The augmented Lagrange multiplier method for exact recovery of corrupted low-rank matrices," 2010, arXiv1009.5055.

[3] J. Wright, Y. Peng, Y. Ma, A. Ganseh, and S. Rao, "Robust principal component analysis: Exact recovery of corrupted low-rank matrices by convex optimization," *Proceedings of 22nd Advances in Neural Information Processing systems*, pp. 2080–2088, 2009.

[4] E. J. Candès, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis?" *Journal of the Association for Computing Machinery*, vol. 58, no. 3, pp. 11:1–11:37, 2011.

[5] V. Chandrasekaran, S. Sanghavi, P. A. Parrilo, and A. S. Willsky, "Rank-sparsity incoherence for matrix decomposition," *SIAM Journal on Optimization*, vol. 21, no. 2, pp. 572–596, 2011.

[6] L. Vandenberghe and S. Boyd, "Semidefinite programming," *SIAM review*, vol. 38, no. 1, pp. 49–95, 1996.

[7] J. F. Cai, E. J. Candès, and Z. Shen, "A singular value thresholding algorithm for matrix completion," *SIAM Journal on Optimization*, vol. 20, no. 4, pp. 1956–1982, 2010.

[8] B. Recht, M. Fazel, and P. A. Parrilo, "Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization," *SIAM review*, vol. 52, no. 3, pp. 471–501, 2010.

[9] A. Dutta, F. Hanzely, and P. Richtárik, "A nonconvex projection method for robust PCA," in *Thirty-Third AAAI Conference on Artificial Intelligence (AAAI-19)*, 2019, pp. 1468–1476.

[10] H. H. Bauschke and P. L. Combettes, *Convex analysis and monotone operator theory in Hilbert spaces*. Springer, 2011.

[11] T. Zhou and D. Tao, "Godec: Randomized low-rank and sparse matrix decomposition in noisy case," in *Proceedings of the 28th International Conference on Machine Learning (ICML)*, 2011, pp. 33–40.

[12] V. Chandrasekaran, S. Sanghavi, P. A. Parrilo, and A. S. Willsky, "Rank-sparsity incoherence for matrix decomposition," *SIAM Journal on Optimization*, vol. 21, no. 2, pp. 572–596, 2011.

[13] X. Yuan and J. Yang, "Sparse and low-rank matrix decomposition via alternating direction methods," *Pacific Journal of Optimization*, vol. 9, no. 1, pp. 167–180, 2013.

[14] P. Netrapalli, U. N. Niranjan, S. Sanghavi, A. Anandkumar, and P. Jain, "Non-convex robust PCA," in *Advances in Neural Information Processing Systems 27*, 2014, pp. 1107–1115.

[15] Z. Lin, A. Ganesh, J. Wright, L. Wu, M. Chen, and Y. Ma, "Fast convex optimization algorithms for exact recovery of a corrupted low-rank matrix," *UIUC Technical Report UILU-ENG-09-2214*, 2009.

[16] A. E. Waters, A. C. Sankaranarayanan, and R. Baraniuk, "SpaRCS: Recovering low-rank and sparse matrices from compressive measurements," *Proceedings of 24nd Advances in Neural Information Processing systems*, pp. 1089–1097, 2011.

[17] X. Yi, D. Park, Y. Chen, and C. Caramanis, "Fast algorithms for robust PCA via gradient descent," *Advances in Neural Information Processing systems*, pp. 361–369, 2016.

[18] T. Zhang and Y. Yang, "Robust PCA by manifold optimization," *Journal of Machine Learning Research*, vol. 19, pp. 1–39, 2018.

[19] T. Bouwmans and E.-H. Zahzah, "Robust PCA via principal component pursuit: A review for a comparative evaluation in video surveillance," *Computer Vision and Image Understanding*, vol. 122, pp. 22–34, 2014.

[20] Y. Chen, H. Xu, C. Caramanis, and S. Sanghavi, "Robust matrix completion and corrupted columns," in *Proceedings of the 28th International Conference on International Conference on Machine Learning*, 2011, pp. 873–880.

[21] M. Tao and J. Yang, "Recovering low-rank and sparse components of matrices from incomplete and noisy observations," *SIAM Journal on Optimization*, vol. 21, no. 1, pp. 57–81, 2011.

[22] D. Hsu, S. M. Kakade, and T. Zhang, "Robust matrix decomposition with sparse corruptions," *IEEE Transactions on Information Theory*, vol. 57, no. 11, pp. 7221–7234, 2011.

[23] F. Nie, H. Wang, X. Cai, H. Huang, and C. Ding, "Robust matrix completion via joint schatten p-norm and lp-norm minimization," in *2012 IEEE 12th International Conference on Data Mining*. IEEE, 2012, pp. 566–574.

[24] X. Li, "Compressed sensing and matrix completion with constant proportion of corruptions," *Constructive Approximation*, vol. 37, no. 1, pp. 73–99, 2013.

[25] L. Cambier and P.-A. Absil, "Robust low-rank matrix completion by riemannian optimization," *SIAM Journal on Scientific Computing*, vol. 38, no. 5, pp. S440–S460, 2016.

[26] O. Klopp, K. Lounici, and A. B. Tsybakov, "Robust matrix completion," *Probability Theory and Related Fields*, vol. 169, no. 1-2, pp. 523–564, 2017.

[27] X. Jiang, Z. Zhong, X. Liu, and H. C. So, "Robust matrix completion via alternating projection," *IEEE Signal Processing Letters*, vol. 24, no. 5, pp. 579–583, 2017.

[28] W.-J. Zeng and H. C. So, "Outlier-robust matrix completion via ℓ_p-minimization," *IEEE Transactions on Signal Processing*, vol. 66, no. 5, pp. 1125–1140, 2017.

[29] Y. Cherapanamjeri, P. Jain, and P. Netrapalli, "Thresholding based outlier robust PCA," in *Proceedings of the 30th Conference on Learning Theory (COLT)*, 2017, pp. 593–628.

[30] Y. Cherapanamjeri, K. Gupta, and P. Jain, "Nearly optimal robust matrix completion," in *Proceedings of the 34th International Conference on Machine Learning (ICML)*, 2017, pp. 797–805.

[31] A. Dutta and X. Li, "A fast weighted SVT algorithm," in *2018 5th International Conference on Systems and Informatics (ICSAI)*, 2018, pp. 1022–1026.

[32] E. J. Candès and Y. Plan, "Matrix completion with noise," *Proceedings of the IEEE*, vol. 98, no. 6, pp. 925–936, 2009.

[33] P. Jain, P. Netrapalli, and S. Sanghavi, "Low-rank matrix completion using alternating minimization," in *Proceedings of the Forty-fifth Annual ACM Symposium on Theory of Computing*, 2013, pp. 665–674.

[34] P. Jain and P. Netrapalli, "Fast exact matrix completion with finite samples," in *Proceedings of The 28th Conference on Learning Theory (COLT)*, 2015, pp. 1007–1034.
[35] E. J. Candès and B. Recht, "Exact matrix completion via convex optimization," *Foundations of Computational Mathematics*, vol. 9, no. 6, pp. 717–772, 2009.
[36] R. Keshavan, A. Montanari, and S. Oh, "Matrix completion from a few entries," *IEEE Transactions on Information Theory*, vol. 56, no. 6, pp. 2980–2998, 2010.
[37] E. J. Candès and T. Tao, "The power of convex relaxation: Near-optimal matrix completion," *IEEE Transactions on Information Theory*, vol. 56, no. 5, pp. 2053–2080, 2010.
[38] J. Mareček, P. Richtárik, and M. Takáč, "Matrix completion under interval uncertainty," *European Journal of Operational Research*, vol. 256, no. 1, pp. 35 – 43, 2017.
[39] Z. Wen, W. Yin, and Y. Zhang, "Solving a low-rank factorization model for matrix completion by a nonlinear successive over-relaxation algorithm," *Mathematical Programming Computation*, vol. 4, no. 4, pp. 333–361, 2012.
[40] F. Wen, R. Ying, P. Liu, and T.-K. Truong, "Nonconvex regularized robust pca using the proximal block coordinate descent algorithm," *IEEE Transactions on Signal Processing*, vol. 67, no. 20, pp. 5402–5416, 2019.
[41] F. Wen, L. Chu, P. Liu, and R. C. Qiu, "A survey on nonconvex regularization-based sparse and low-rank recovery in signal processing, statistics, and machine learning," *IEEE Access*, vol. 6, pp. 69 883–69 906, 2018.
[42] R. T. Rockafellar and R. Wets, *Variational analysis*. Springer Verlag, 1998, vol. 317.
[43] P. L. Lions and B. Mercier, "Splitting algorithms for the sum of two nonlinear operators," *SIAM Journal on Numerical Analysis*, vol. 16, no. 6, pp. 964–979, 1979.
[44] Y. Nesterov, *Introductory lectures on convex optimization: A basic course*. Springer, 2004, vol. 87.
[45] J. Liang, J. Fadili, and G. Peyré, "A multi-step inertial Forward–Backward splitting method for non-convex optimization," in *Advances in Neural Information Processing Systems*, 2016.
[46] J. Bolte, A. Daniilidis, O. Ley, and L. Mazet, "Characterizations of Łojasiewicz inequalities: subgradient flows, talweg, convexity," *Transactions of the American Mathematical Society*, vol. 362, no. 6, pp. 3319–3363, 2010.
[47] J. Bolte, A. Daniilidis, and A. Lewis, "The Łojasiewicz inequality for nonsmooth subanalytic functions with applications to subgradient dynamical systems," *SIAM Journal on Optimization*, vol. 17, no. 4, pp. 1205–1223, 2007.
[48] J. M. Lee, *Smooth manifolds*. Springer, 2003.
[49] J. Liang, J. Fadili, and G. Peyré, "Local linear convergence of Forward–Backward under partial smoothness," in *Advances in Neural Information Processing Systems*, 2014, pp. 1970–1978.
[50] J. Liang, "Convergence rates of first-order operator splitting methods," Ph.D. dissertation, Normandie Université; GREYC CNRS UMR 6072, 2016.
[51] T. Bouwmans, L. Maddalena, and A. Petrosino, "Scene background initialization: A taxonomy," *Pattern Recognition Letters*, vol. 96, pp. 3–11, 2017.
[52] A. Dutta, X. Li, and P. Richtárik, "Weighted low-rank approximation of matrices and background modeling," 2018, arXiv:1804.06252.
[53] T. Bouwmans, A. Sobral, S. Javed, S. K. Jung, and E.-H. Zahzah, "Decomposition into low-rank plus additive matrices for background/foreground separation: A review for a comparative evaluation with a large-scale dataset," *Computer Science Review*, vol. 23, pp. 1–71, 2017.
[54] T. Bouwmans, "Traditional and recent approaches in background modefor foreground detection: An overview," *Computer Science Review*, vol. 11–12, pp. 31 – 66, 2014.
[55] A. Dutta, X. Li, and P. Richtárik, "A batch-incremental video background estimation model using weighted low-rank approximation of matrices," in *The IEEE International Conference on Computer Vision Workshops (ICCVW)*, 2017, pp. 1835–1843.
[56] A. Dutta and P. Richtárik, "Online and batch supervised background estimation via l1 regression," in *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2019, pp. 541–550.
[57] A. Dutta and X. Li, "Weighted low rank approximation for background estimation problems," in *The IEEE International Conference on Computer Vision Workshops (ICCVW)*, 2017, pp. 1853–1861.
[58] J. He, L. Balzano, and A. Szlam, "Incremental gradient on the Grassmannian for online foreground and background separation in subsampled video," *IEEE Computer Vision and Pattern Recognition*, pp. 1937–1944, 2012.
[59] S. Brutzer, B. Höfer, and G. Heidemann, "Evaluation of background subtraction techniques for video surveillance," *IEEE Computer Vision and Pattern Recognition*, pp. 1568–1575, 2012.
[60] K. Toyama, J. Krumm, B. Brumitt, and B. Meyers, "Wallflower: Principles and practice of background maintainance," *Seventh International Conference on Computer Vision*, pp. 255–261, 1999.
[61] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transaction on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.
[62] R. Basri and D. Jacobs, "Lambertian reflection and linear subspaces," *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 25, no. 3, pp. 218–233, 2003.
[63] P. N. Belhumeur and D. J. Kriegman, "What is the set of images of an object under all possible illumination conditions?" *International Journal of Computer Vision*, vol. 28, no. 3, pp. 245–260, 1998.
[64] A. Georghiades, P. Belhumeur, and D. Kriegman, "From few to many: Illumination cone models for face recognition under variable lighting and pose," *IEEE Transactions on PAMI*, vol. 23, no. 6, pp. 643–660, 2001.
[65] J. Goes, T. Zhang, R. Arora, and G. Lerman, "Robust stochastic principal component analysis," in *Proceedings of the 17th International Conference on Articial Intelligence and Statistics*, 2014, pp. 266–274.
[66] T. Oh, Y. Tai, J. Bazin, H. Kim, and I. S. Kweon, "Partial sum minimization of singular values in robust PCA: Algorithm and applications," *IEEE Transactions on 'Pattern Analysis and Machine Intelligence*, vol. 38, no. 4, pp. 744–758, 2016.
[67] S. Wang, Y. Wang, Y. Chen, P. Pan, Z. Sun, and G. He, "Robust PCA using matrix factorization for background/foreground separation," *IEEE Access*, vol. 6, pp. 18 945–18 953, 2018.
[68] A. S. Lewis, "Active sets, non-smoothness, and sensitivity," *SIAM Journal on Optimization*, vol. 13, no. 3, pp. 702–725, 2003.
[69] A. S. Lewis and S. Zhang, "Partial smoothness, tilt stability, and generalized Hessians," *SIAM Journal on Optimization*, vol. 23, no. 1, pp. 74–94, 2013.
[70] D. Drusvyatskiy and A. S. Lewis, "Optimality, identifiability, and sensitivity," *Mathematical Programming*, pp. 1–32, 2013.
[71] L. Fei-Fei, R. Fergus, and P. Perona, "Learning generative visual models from few training examples: An incremental Bayesian approach tested on 101 object categories," *Computer Vision and Image Understanding*, vol. 106, no. 1, pp. 59–70, 2007.

Supplementary Material

We provide a comprehensive table to list all baselines we compare to in Appendix D. Extra supporting numerical experiments are reported in Appendix E. Additionally, we provide another real-world example — "Inlier detection from noisy observation" in Appendix F.

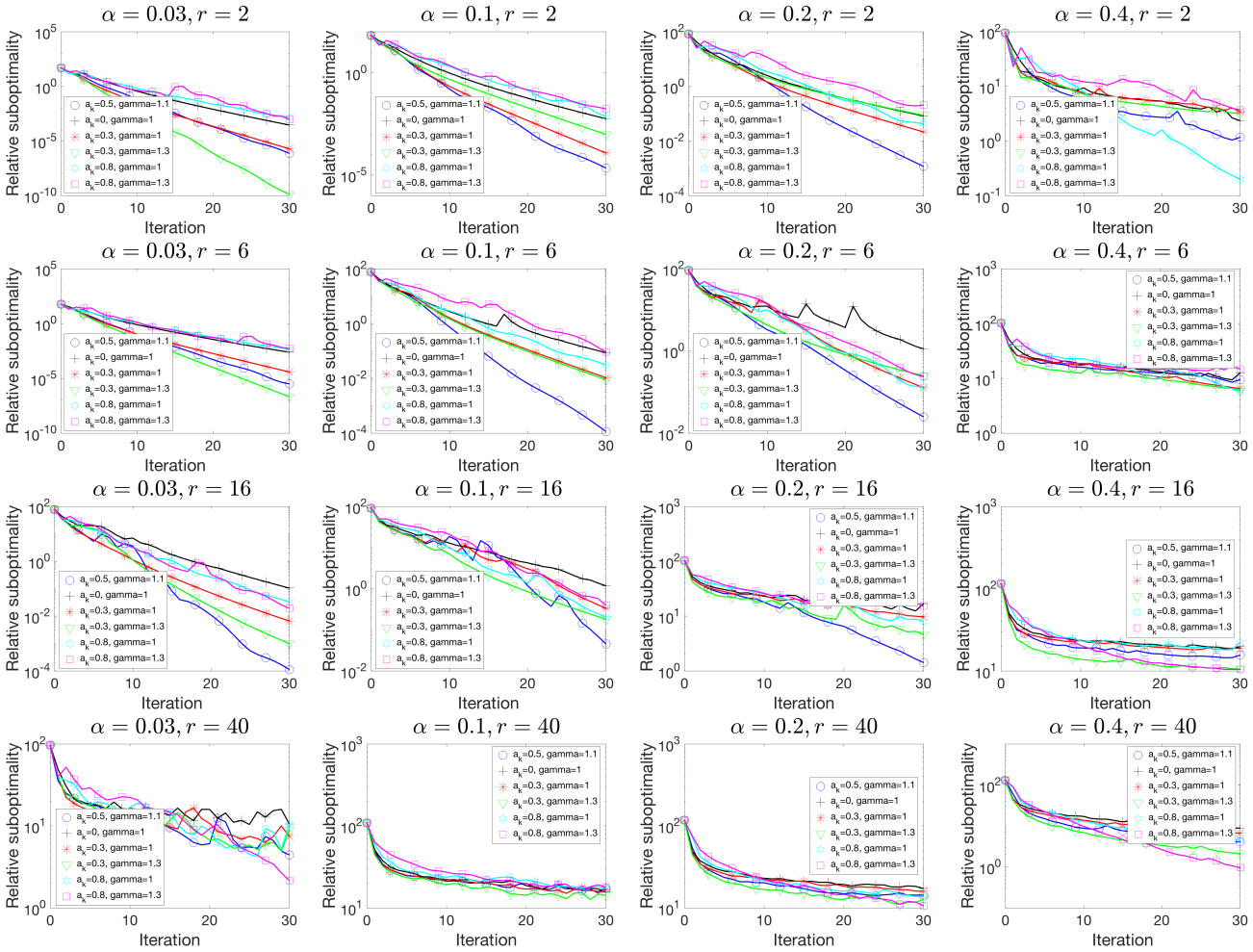


Fig. 11: Sensitivity of Algorithm 1 with respect to choice of $\gamma, a_k = b_k$.

APPENDIX D
TABLE OF BASELINE METHODS

In Table I below, we summarize all the methods compared in our paper.

Algorithm	Abbreviation	Appearing in Experiment	Reference
Inexact Augmented Lagrange Method of Multipliers	iEALM	Fig. 2, 9, 4	[2]
Accelerated Proximal Gradient	APG	Fig. 9, 4	[3]
Singular Value Thresholding	SVT	Table II	[7]
Grassmannian Robust Adaptive Subspace Tracking Algorithm	GRASTA	Fig. 5	[58]
Go Decomposition	GoDec	Fig. 4, 7	[11]
Robust PCA Gradient Descent	RPCA GD	Fig. 3, 9, 4, 5, 6, 7, 8	[17]
Robust PCA Nonconvex Feasibility	RPCA NCF	Fig. 2, 9, 4, 5, 7, 8 , 14	[9]
Robust stochastic PCA Algorithms	SGD, R-SGD1, R-SGD2 Inc, R-Inc, MD, R-MD	Fig. 14, Table II	[65]

TABLE I: Algorithms compared in this paper.

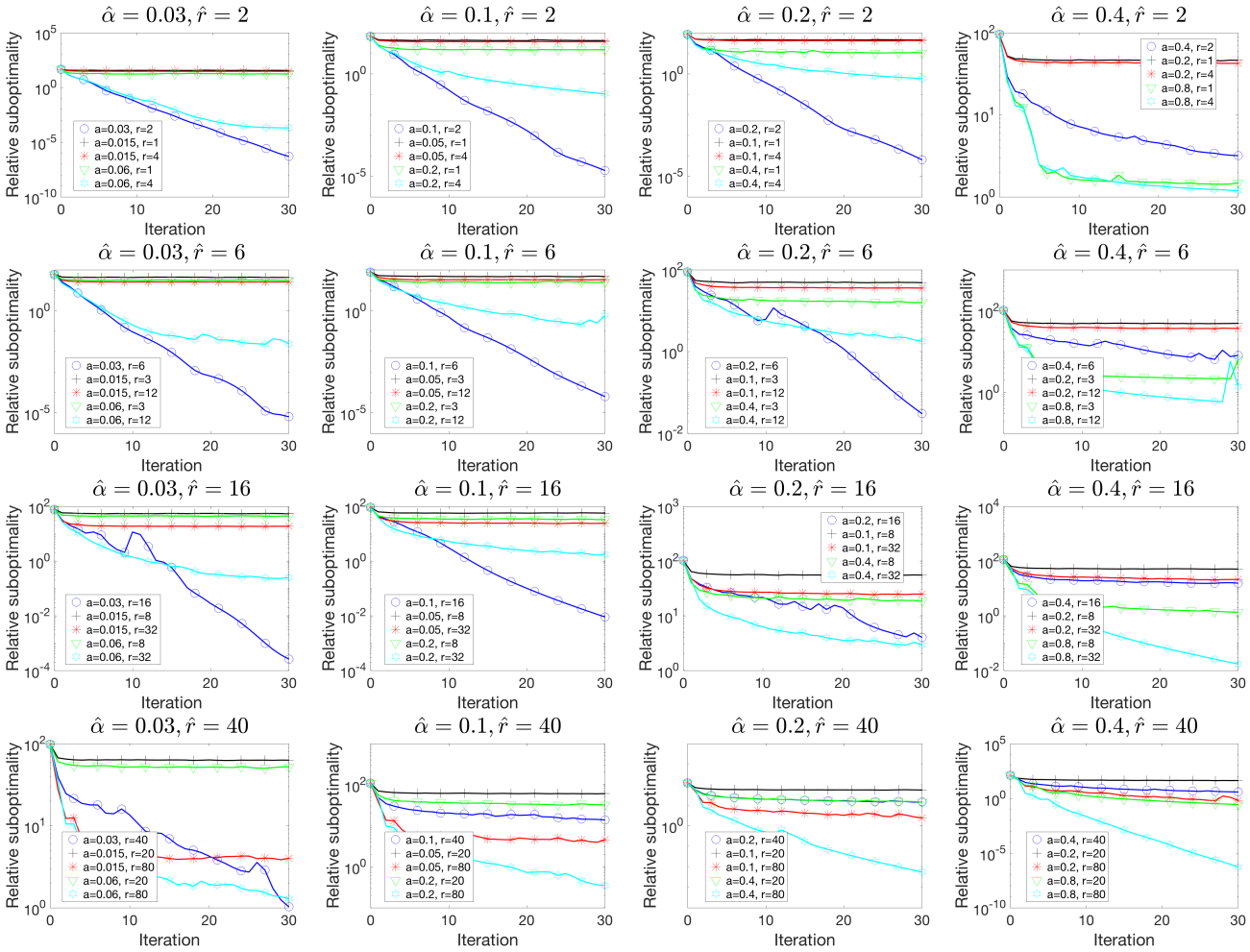


Fig. 12: Sensitivity of Algorithm 1 with respect to the correct choice of target rank and target sparsity.

APPENDIX E
EXTRA EXPERIMENTS

In this section, we empirically study convergence properties of Algorithm 1 on synthetic, well-understood data. In particular, we examine its sensitivity to user-specified parameters γ, a_k, b_k, target sparsity level α, target rank r, and lastly, the sensitivity to initialization. Moreover, we provide extra phase transition diagrams and both quantitative and qualitative results on the inlier detection problem.

A. Sensitivity to the choice of γ, a_k, b_k

In this experiment, we compare different choices of algorithm parameters γ, a_k, b_k on instances of (6) with various target sparsity level α and target rank r. In each experiment, we make sure that the solution exists; we generate random matrices $\tilde{L}, \tilde{S}$ (with independent entries $\mathcal{N}(0,1)$), project them onto low rank and sparse constraint set respectively to obtain $\hat{L}, \hat{S}$ and set $A = \hat{L} + \hat{S}$. For simplicity we consider only $a_k = b_k = a$ and $m = n = 100$. Figure 11 shows the result. We see that parameter choice $\gamma = 1.1, a_k = b_k = \frac{1}{2}$ is the most reliable.

B. Sensitivity to the choice of r, α

In this experiment, we examine how sensitive is Algorithm 1 on the correct choice of the target sparsity level α and rank r. In each experiment, we generate random matrices $\tilde{L}, \tilde{S}$ (with independent entries $\mathcal{N}(0,1)$), project them onto $\hat{r}$-low rank and $\hat{\alpha}$-sparse constraint set respectively to obtain $\hat{L}, \hat{S}$ and set $A = \hat{L} + \hat{S}$. Then, we run Algorithm 1 with various choices of r, α and report the results. For simplicity we consider only $\gamma = 1.1, a_k = b_k = \frac{1}{2}$ (from the previous experiment) and $m = n = 100$. Figure 12 shows the result. We can see that if sparsity level is underestimated, the method converges very slowly. Moreover, the method is more sensitive to the correct choices of target sparsity than target rank. The last take-away from this experiment is that over-estimation of target parameters usually leads to slightly slower convergence.

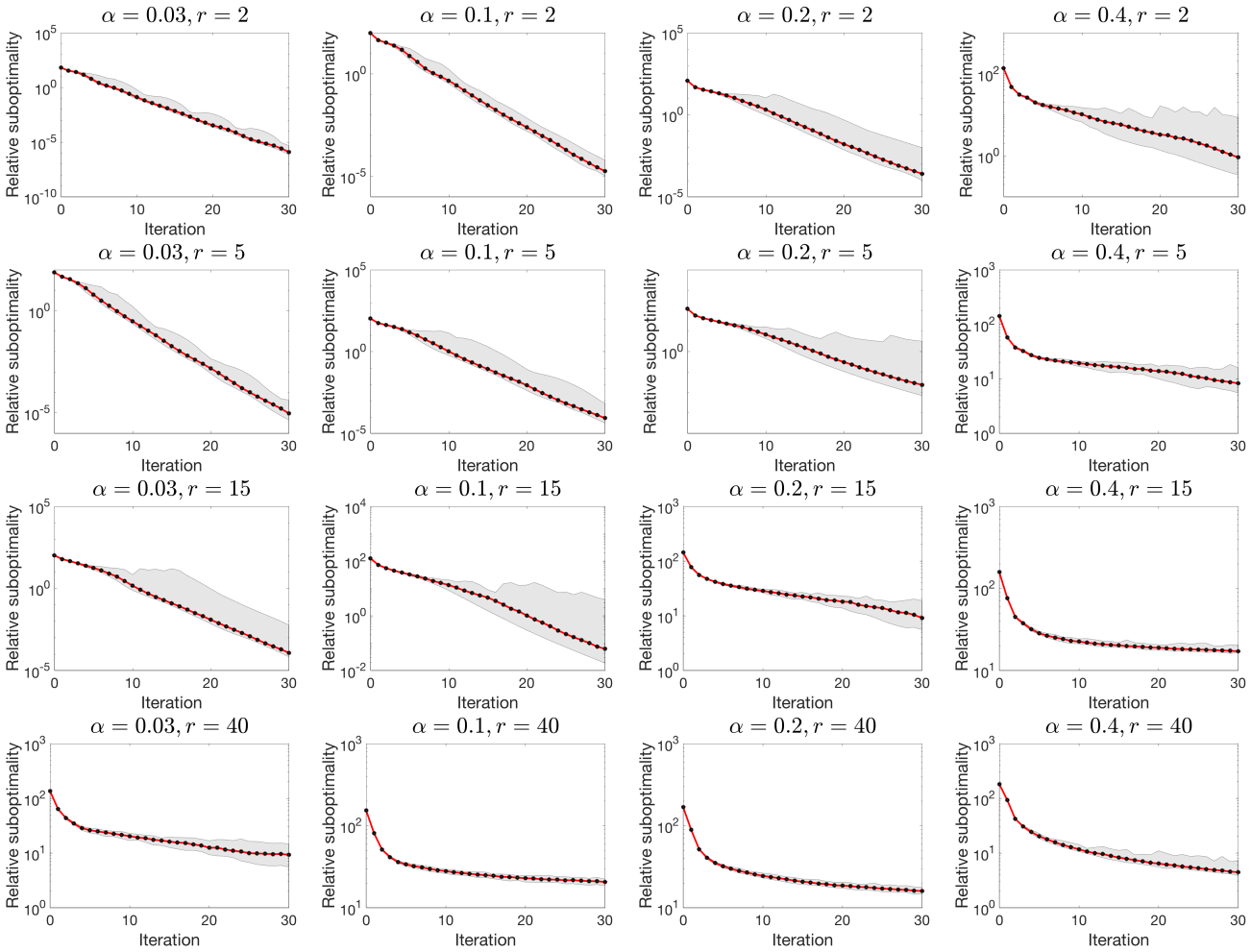


Fig. 13: Sensitivity of Algorithm 1 with respect to the starting point.

C. *Sensitivity to the choice of the starting point*

In the last experiment, we examine how the starting point influences the convergence rate. For each problem instance, we perform 50 independent runs of Algorithm 1 and report the best, worst and median performance.

For simplicity, we consider only problems with known target rank and sparsity – we generate random matrices $\tilde{L}, \tilde{S}$ (with independent entries $\mathcal{N}(0,1)$), project them onto low rank and sparse constraint set respectively to obtain $\hat{L}, \hat{S}$ and set $A = \hat{L} + \hat{S}$. Further, we set $a_k = b_k = 0.5$, $\gamma = 1.1$ and $m = n = 100$. Figure 13 shows the result. We can see that the convergence speed of Algorithm 1 is, in most cases, not influenced significantly by the starting point. Thus, the non-convex nature of the problem is surprisingly not causing any issues. Lastly, the convergence rate of Algorithm 1 is faster for small values of α, r, which is often the most interesting case in terms of the practical application.

APPENDIX F
INLIER DETECTION

Historically, PCA and RPCA are used in detecting the inliers and the outliers from a composite dataset. We infused 400 random, grayscale, downsampled (20×20 pixels) natural images from the BACKGROUND/Google folder of the Caltech101 database [71] with the `Yale Extended Face Database` to construct the data set. The inliers are the grayscale images of faces (of the same resolution) under different illuminations, while the 400 random natural images serve as outliers. The goal is to consider a low-dimensional model and to project the inliers to a 9-dimensional linear subspace where the images of the same face lie. Goes et al. in [65] designed seven algorithms to explicitly find a low-rank subspace. To this end, Goes et al. used the classical SGD, an incremental approach, and mirror descent algorithms to find the 9-dimensional subspace. However, we split the dataset, A, into a 9-dimensional low-rank subspace L and expect the outliers to be in the sparse set, S. Once we find L, we find the basis of L via orthogonalization and project the faces on it. In Figure 14, we show the qualitative results of our experiments.[3]

[3]The codes and datasets for experiments in Section F are obtained from https://github.com/jwgoes/RSPCA

Metric	SGD	R-SGD1	R-SGD2	Inc	R-Inc	MD	R-MD	RPCA-F	Best pair	SVT
$\frac{\|P_L - P_{L^\star}\|_F}{3\sqrt{2}}$	0.7	0.86	4.66	0.77	0.72	0.67	0.67	0.78	0.76	0.79

TABLE II: Quantitative performance of different algorithms in inlier detection experiment. Except R-SGD2 all methods are competitive.

As proposed in [65], we use the normalized error term $\|P_L - P_{L^\star}\|_F/3\sqrt{2}$, where L is subspace fitted by the PCA to the set of inliers and $L^\star$ be the subspace fitted by different algorithms. Note that the metric is expected to lie between 0 and 1 where the smaller is, the better. We refer to Table II for our quantitative results.

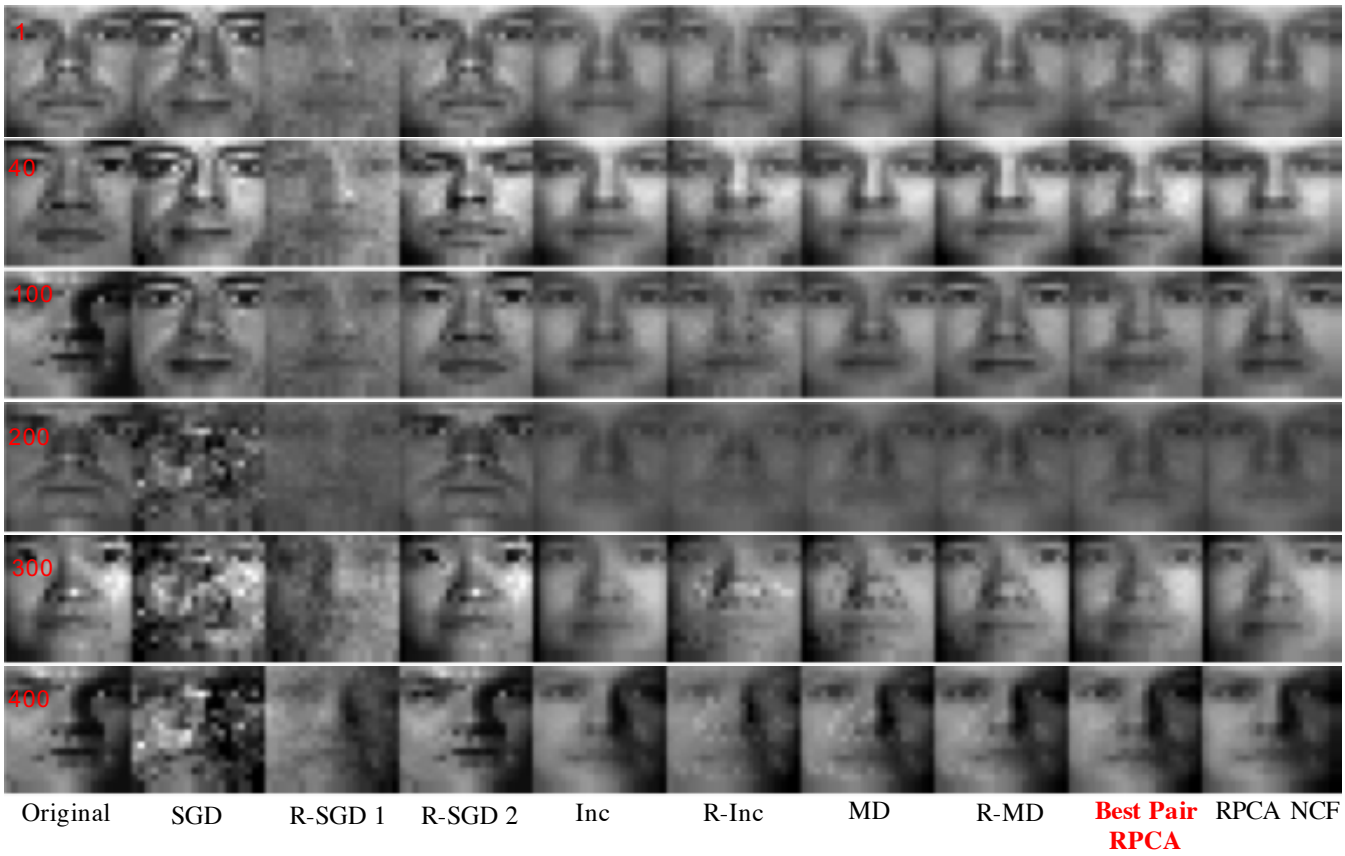


Fig. 14: Inliers and outliers detection. We project the face images (inliers) to 9 dimensional subspaces found by different methods.