Accelerated Low-Rank Visual Recovery by Random Projection

Yadong Mu, Jian Dong, Xiaotong Yuan, Shuicheng Yan
Department of Electrical and Computer Engineering
National University of Singapore, 117576, Singapore

Abstract

Exact recovery from contaminated visual data plays an important role in various tasks. By assuming the observed data matrix as the addition of a low-rank matrix and a sparse matrix, theoretic guarantee exists under mild conditions for exact data recovery. Practically matrix nuclear norm is adopted as a convex surrogate of the non-convex matrix rank function to encourage low-rank property and serves as the major component of recently-proposed Robust Principal Component Analysis (R-PCA). Recent endeavors have focused on enhancing the scalability of R-PCA to large-scale datasets, especially mitigating the computational burden of frequent large-scale Singular Value Decomposition (SVD) inherent with the nuclear norm optimization. In our proposed scheme, the nuclear norm of an auxiliary matrix is minimized instead, which is related to the original low-rank matrix by random projection. By design, the modified optimization entails SVD on matrices of much smaller scale, as compared to the original optimization problem. Theoretic analysis well justifies the proposed scheme, along with greatly reduced optimization complexity. Both qualitative and quantitative studies are provided on various computer vision benchmarks to validate its effectiveness, including facial shadow removal, surveillance background modeling and large-scale image tag transduction. It is also highlighted that the proposed solution can serve as a general principal to accelerate many other nuclear norm oriented problems in numerous tasks.

1. Introduction

Visual data which are corrupted either by sensory noises or interferential outliers during data acquisition frustrate many computer vision algorithms, which motivates many robust estimation methods such as RANSAC [5] in the past decades. Recent years have witnessed a surge of sparsity-oriented robust learning. Early study in machine learning revealed empirical superiority of ℓ_1-norm regularization over that based on ℓ_2-norm. This rough message soon spread to other related research fields and inspired the investigation of other forms of sparsity. The sparse learning methods have reshaped a large part of the computer vision research and are undergoing a continuing progress from both theoretic and practical aspects.

In this paper we focus on visual recovery that can be cast as a low-rank matrix recovery problem, which favors sparse singular value structures for the recovered data matrix. Such a sparsity prior gains notable successes in various applications, including face processing [16], movie recommendation [11] and even photometric stereo [19]. Specifically, matrix nuclear norm was employed to enforce such sparsity. We investigate an important sparse learning framework named Robust Principal Component Analysis (R-PCA) [16][1] in the literature, which restores the true subspace structure by identifying sparse residuals from the observed data matrix.

Our main contribution is the exposition of a principled accelerated R-PCA algorithm for visual recovery, which radically differs from previous efforts on enhancing the scalability of R-PCA. One major source of previous low optimization efficacy stems from estimating the singular structures of large-size matrices. To address this issue, we advocate utilizing the recent idea of "compressed optimization" [2, 1, 15]. In detail, the data matrices are beforehand compressed by projecting onto random matrices and the optimization proceeds intelligently on either the original or compressed data, balancing accuracy and computational expense. In Section 4 we employ such an idea to accelerate the time-consuming nuclear norm regularized optimization. Theoretic analysis regarding efficiency gain and performance loss is later provided in Section 5 and empirical validation on several vision benchmarks is presented in Section 6. It is also highlighted that this principled idea can be adapted to other nuclear norm oriented problems [11].

2. Related Work

The relevant literature to our work mainly originates from the following lines of work:

1) Compressed Sensing: Namely it studies the process of acquiring and reconstructing a signal utilizing the prior

[1] Although being spiritually similar, another line of work (*e.g.*, [7]) which also used the term "R-PCA" has fundamental difference to the work discussed here. We stick to the notation without expected confusion.

knowledge that it is sparse or compressible. Under mild conditions a sparse signal can be reconstructed from limited number of observations, even nearly all of which are corrupted [17]. Equipped with reasonable visual dictionaries or bases, many tasks in computer vision can be similarly cast as seeking a sparse coefficients on the bases and solved by techniques borrowed from compressed sensing. Several successful applications have been presented, including face recognition [18, 14] and image super-resolution [20].

2) Nuclear Norm Oriented Learning: The idea of low-rank matrix can be regarded as extending the concept of "sparse vector" to the matrix field. However, matrix rank is neither continuous nor convex, which complicates the pursuit of global optimum. A popular surrogate function is the matrix nuclear norm (also known as trace norm or Ky Fan k-norm in the literature), which is defined as the sum of all singular values and a convex function. Srebro et al. proposed max-margin matrix factorization (MMMF) [12] for collaborative prediction, using matrix nuclear norm and hinge loss to obtain a low-rank representation. Wright et al. [16] demonstrated its success in surveillance background estimation and facial shadow removal.

3) Compressed Optimization: It represents very recent progress towards large-scale optimization. The basic idea is to sketch large data matrix using random projection. Representative work includes compressed least-squares for over-determined linear system [2, 1] and compressed non-negative matrix factorization [15]. In these aforementioned works, the random matrices used for compression are mostly produced from standard Gaussian distribution or randomized Fourier transform [9]. One recent work by Shi et al. [10] discussed the random projection using hashing, and successfully accelerated sparse coding based face recognition. To our best knowledge, all of previous work along this line are focusing on the acceleration for linear systems. Our work in this paper is the first one to investigate compressed optimization for matrix nuclear norm based problems.

3. Preliminaries

3.1. Robust Principal Component Analysis

The generating procedure of real-world observations is always suffering from noises and outliers. Assume the collected data matrix $D \in \mathbb{R}^{m\times n}$ has an underlying low-rank structure, yet corrupted by sparse additive noises. Denote these two ingredients as $A, E \in \mathbb{R}^{m\times n}$ respectively. To be immune to arbitrarily large errors, the ideal penalty for error matrix E is the matrix zero-norm, *i.e.*, counting non-zero elements in the matrix. The initial formulation [16] can be described as below:

$$\min_{A,E} \ rank(A) + \lambda\|E\|_0, \quad s.t. \ D = A + E. \tag{1}$$

Unfortunately, Problem in (1) is difficult to solve owing to the non-convexity and high non-smoothness from the rank measure and zero-norm penalty. Similar to the trick used in the vector-based optimization, we practically solve the following relaxed form:

$$\min_{A,E} \ \|A\|_* + \lambda\|E\|_1, \ \ s.t. \ D = A + E, \tag{2}$$

where $\|\cdot\|_1$ denotes the matrix ℓ_1-norm (*i.e.*, the sum of absolute matrix entries), $\|\cdot\|_*$ denotes the matrix nuclear norm and λ is a positive parameter for balancing. In detail, suppose by Singular Value Decomposition (SVD), $A_{m\times n} = U_{m\times m}\Sigma_{m\times n}V^T_{n\times n}$, where the subscripts describe the matrix sizes. Both U and V are known to be orthogonal and the (i,i)-th entry of Σ is equal to the singalur value σ_i for $i = 1 \ldots \min(m,n)$, otherwise 0. Nuclear norm is defined as the sum of all singular values. It has been suggested as a convex surrogate to the rank function, and proves to be the convex envelope (smallest bounding convex function) of the rank function on matrices with unit spectral norm [16].

3.2. Augmented Lagrange Multiplier Method

The optimization of R-PCA in (2) is straightforward by observing that both the constraints and objectives are convex. Normally it can be recast as a Semi-Definite Program (SDP) and optimized by off-the-shelf interior-point solvers. However, the Newton step in each iteration is computationally expensive, which makes the scalability to large matrices problematic. A naive implementation even consumes about 10^4 iterations to converge and runs 8 hours on a common PC for a matrix of size 800×800. Recent endeavor on accelerated R-PCA [8] has employed the techniques such as Augmented Lagrange Multipliers (ALM)[2], which guarantees a better convergence rate. In this subsection we elaborate on a brief description for ALM in solving R-PCA. Particularly, the general ALM method targets constrained optimization problems as below:

$$\min_X f(X), \ \ s.t. \ h(X) = 0, \tag{3}$$

where $f(X)$ and $h(X)$ are both convex. We can get the augmented Lagrangian function:

$$\mathcal{L}(X, Y, \mu) = f(X) + \langle Y, h(X)\rangle + \frac{\mu}{2}\|h(X)\|_F^2, \tag{4}$$

where μ is a parameter that is increased in iterations. $\|\cdot\|_F$ denotes the Frobunius norm of a matrix. We can relate the

[2]Note that ALM method has two popular variants, *i.e.*, exact ALM and inexact ALM [8]. The latter is faster yet sometimes converges to incorrect solutions, thus by default the notation ALM refers to the exact version both in algorithmic discussion and experimental design until otherwise notified. Extensions of the proposed algorithm to the inexact ALM and other relevant methods (*e.g.*, Accelerated Proximal Gradient [8]) are straightforward.

general ALM method to R-PCA problem by identifying the variable $X = (A, E)$ and

$$f(X) = \|A\|_* + \lambda\|E\|_1, \;\; h(X) = D - A - E.$$

The overall ALM optimization of R-PCA proceeds by iteratively increasing μ (*e.g.*, $\mu_{k+1} = 2\mu_k$) in each iteration and minimizing the resultant $\mathcal{L}(X, Y, \mu_k)$ in Equation (4). In each iteration, a global optimum is feasible by alternatively optimizing either A or E with the other variable fixed, *i.e.*, the following two steps

Step-1: Update A with E fixed

$$A_{k+1} = \arg\min_A \frac{1}{\mu_k}\|A\|_* + \frac{1}{2}\|A - (D - E_k + \frac{1}{\mu_k}Y_k)\|_F^2.$$

Step-2: Update E with A fixed

$$E_{k+1} = \arg\min_E \frac{\lambda}{\mu_k}\|E\|_1 + \frac{1}{2}\|E - (D - A_{k+1} + \frac{1}{\mu_k}Y_k)\|_F^2.$$

Both of above subproblems are verified to have closed-form solutions. In detail, we define the following soft-thresholding (ST) operator [3] for shrinkage purpose, *i.e.*, $\mathcal{S}_\epsilon[x] = \text{sign}(x) \cdot \max(|x| - \epsilon, 0)$. Assume $W = U\Sigma V^T$ by SVD, the optima for the subproblems can be directly obtained from the ST operator:

$$U\mathcal{S}_\epsilon[\Sigma]V^T = \arg\min_X \epsilon\|X\|_* + \frac{1}{2}\|X - W\|_F^2, \quad (5)$$

$$\mathcal{S}_\epsilon[W] = \arg\min_X \epsilon\|X\|_1 + \frac{1}{2}\|X - W\|_F^2. \quad (6)$$

4. Optimization by Projected Nuclear Norm

In the algorithm described in Section 3.2, the computational bottleneck in each iteration is the SVD computation in Step-1, while other steps are amenable to parallelized accelerating tricks. The full SVD of a 2000×2000 matrix takes more than one minute on a common PC. Although previous study proposes partial SVD for saving unnecessary computations [8] under some scenarios, generally dense SVD-related computations are unavoidable. The mainstream of previous efforts on R-PCA acceleration mainly focus on reducing iteration count before convergence. Instead, here we revisit this problem from a novel aspect. For very-large data matrices (*e.g.*, on the order of $10^4 \times 10^4$), a plausible solution is to devise a surrogate nuclear norm defined on a size-reduced matrix. Here we demonstrate this possibility by utilizing the isometry-preserving random projection or hashing. Suppose the problem scale is $m \times n$ and $P = \frac{1}{\sqrt{p}} \cdot \widetilde{P} \in \mathbb{R}^{m \times p}$ ($p \ll m$), where $\widetilde{P}$ is a random matrix drawn from the Gaussian distribution with zero mean and unit standard deviation. The original R-PCA is reformulated as below:

$$\begin{aligned} \min_{A',A,E} \quad & \|A'\|_* + \lambda\|E\|_1 \\ s.t. \quad & D = A + E, \; A' = P^T A. \end{aligned} \quad (7)$$

Algorithm 1: Accelerated R-PCA using linear projection and the exact ALM method

0 **input**: Observation matrix $D \in \mathbb{R}^{m \times n}$.
1 $Y_0^* = 0$; $E_0^* = 0$; $A_0^* = 0$; $\mu_1^0 > 0$; $\mu_2^0 > 0$; $\rho > 1$; $\lambda = 1/\sqrt{m}$; $k = 0$;
while *not converged* **do**
2 $\quad A_{k+1}^0 = A_k^*$; $E_{k+1}^0 = E_k^*$; $j = 0$;
$\quad$ **while** *not converged* **do**
3 $\quad\quad E_{k+1}^{j+1} = \mathcal{S}_{\frac{\lambda}{\mu_1^k}}[D - A_{k+1}^j + (\mu_1^k)^{-1}Y_k^*]$;
4 $\quad\quad (U, \Sigma, V) = \text{SVD}(P^T A_{k+1}^j)$;
5 $\quad\quad (A')_{k+1}^{j+1} = U\mathcal{S}_{(\mu_2^k)^{-1}}[\Sigma]V^T$;
6 $\quad\quad A_{k+1}^{j+1} = \frac{p}{m+\mu p}P(W_2 - P^T W_1) + W_1$;
7 $\quad\quad j = j + 1$;
$\quad$ **end**
8 $\quad Y_{k+1}^* = Y_k^* + \mu_1^k(D - A_{k+1}^* - E_{k+1}^*)$;
9 $\quad \mu_1^{k+1} = \rho\mu_1^k$; $\mu_2^{k+1} = \rho\mu_2^k$; $k = k + 1$;
end

where $A' = P^T A \in \mathbb{R}^{p \times n}$ is a projected low-rank matrix, with smaller size than original $m \times n$. The augmented Lagrange function for Problem (7) can be presented as below:

$$\begin{aligned} \mathcal{L}(A', A, E, Y, \mu_1, \mu_2) \triangleq \|A'\|_* + \lambda\|E\|_1 + \langle Y, D - A - E\rangle \\ + \frac{\mu_1}{2}\|D - A - E\|_F^2 + \frac{\mu_2}{2}\|A' - P^T A\|_F^2, \end{aligned}$$

where the constraint $A' = P^T A$ is actually used as penalty function unlike in standard ALM trick. It is mainly designed to avoid the optimal matrix A overfitting to specific random matrix P, as explained in Subsection 5.2 later.

Following the same spirit described in Section 3.2, optimizing the transformed problem with linear random projection in (7) involves the following three subproblems during optimizing $\mathcal{L}(A', A, E, Y, \mu_1, \mu_2)$ in each iteration:

$$\begin{aligned} &\text{(P1)} \quad \min_{A'} \frac{1}{\mu_2}\|A'\|_* + \frac{1}{2}\|A' - P^T A\|_F^2, \\ &\text{(P2)} \quad \min_E \frac{\lambda}{\mu_1}\|E\|_1 + \frac{1}{2}\|E - (D - A + \frac{1}{\mu_1}Y)\|_F^2, \\ &\text{(P3)} \quad \min_A \frac{\mu_1}{2}\|A - (D - E + \frac{1}{\mu_1}Y)\|_F^2 + \frac{\mu_2}{2}\|P^T A - A'\|_F^2. \end{aligned}$$

The algorithmic pseudo-code is found in Algorithm 1. Note that both (P1) and (P2) are solvable using the ST operator presented in (5) or (6) (lines 4-5 and line 3 respectively), with reduced problem size. (P3) proves to be a quadratic program with closed-form optimal solutions. Specifically, use the abbreviations $W_1 = D - E + Y/\mu_1$ and $W_2 = A'$. By setting the first-order derivative of the objective function in (P3) to be zero, we get

$$(\mu_1 I + \mu_2 PP^T)A = \mu_1 W_1 + \mu_2 PW_2, \quad (8)$$

which has efficient approximate optimum by observing $P^T P \approx \frac{m}{p}I$ (I is the identify matrix with its size inferred in

the context) derived from the property of Gaussian random matrices [13]. Based on the Sherman-Morrison-Woodbury formula[3], it can be verified that

$$(\mu I + PP^T)^{-1} \approx \mu^{-1} I - \mu^{-1} \cdot \frac{p}{\mu p + m} \cdot PP^T, \quad (9)$$

where $\mu = \mu_1/\mu_2$ and the approximation comes from $P^T P \approx \frac{m}{p} I$ as aforementioned. By multiplying the above matrix inverse to the right hand in (8), finally we obtain the optimum by $A^* \approx \frac{p}{m+\mu p} P(W_2 - P^T W_1) + W_1$.

For the case that both m, n are very large as seen in many collaborative filtering problems, we employ the bilinear random projection as following

$$\begin{array}{ll} \min_{A',A,E} & \|A'\|_* + \lambda \|E\|_1 \\ s.t. & D = A + E, \;\; A' = P^T A Q, \end{array} \quad (10)$$

where $Q = 1/\sqrt{q} \cdot \widetilde{Q} \in \mathbb{R}^{n \times q}$ with $\widetilde{Q}$ drawn from standard Gaussian distribution ($q \ll n$). Likewise we can get the following updating rule for matrix A:

$$A^* \approx \frac{pq}{mn + \mu pq} \left(PW_2 Q^T + \mu W_1 \right). \quad (11)$$

5. Theoretic Analysis

The proposed algorithm in Section 4 provides the possibility of breaking the curse of large-scale R-PCA optimization. This speedup brought by size-reduced nuclear norm optimization is at the cost of degraded estimation accuracy. It is meaningful to investigate the relationship between $\|A\|_*$ and $\|A'\|_*$ under randomly-generated projection matrix (since their nearness implies the recovery condition of R-PCA can be approximately applied), and other issues including the computational and storage complexities for original R-PCA and the proposed method.

5.1. Bounds of the Projected Nuclear Norm

Before continuing, we first point out the isometry property of Gaussian random projection:

Lemma 1 ***(Norm-Preserving Property [13]):*** *Let each entry of an $n \times p$ matrix R be chosen independently from standard Gaussian distribution (*i.e.*, with zero mean and unit standard deviation). Denote $v = \frac{1}{\sqrt{p}} R^T u$ for any $u \in \mathbb{R}^n$. Then for $\epsilon \in (0, 1)$, there are $E(\|v\|^2) = \|u\|^2$ and $Prob\left(\left| \|v\|^2 - \|u\|^2 \right| \geq \epsilon \|u\|^2 \right) < 2 \exp\left(-\frac{p}{4} \left(\epsilon^2 - \epsilon^3 \right) \right)$.*

From Lemma 1 we get the bounds of projected nuclear norm from both below and above:

Theorem 1 ***(Bounds for Projected Nuclear Norm)*** *Consider a low-rank data matrix $A \in \mathbb{R}^{m \times n}$ and the projected matrix $B = \frac{1}{\sqrt{p}} R^T A \in \mathbb{R}^{p \times n}$, where $R \in \mathbb{R}^{m \times p}$ ($p \ll m$) is the projection matrix drawn from standard Gaussian distribution. Assume $rank(A) = r$ and $p > r$. With high probability and small ϵ the following relations hold*

$$\begin{array}{ll} (K1) & (1-\epsilon)\|A\|_F^2 \leq \|1/\sqrt{p} \cdot R^T A\|_F^2 \leq (1+\epsilon)\|A\|_F^2, \\ (K2) & \|1/\sqrt{p} \cdot R^T A\|_* \geq \sqrt{1-\epsilon}/\sqrt{r} \cdot \|A\|_*, \\ (K3) & \|1/\sqrt{p} \cdot R^T A\|_* \leq \sqrt{1+\epsilon} \cdot \|A\|_*. \end{array}$$

Proof of (K1): *Denote matrix A as concatenated column vectors,* i.e.*, $A = (a_1, \ldots, a_n)$. From Lemma 1, with high probability, $\forall\, i \in \{1, \ldots, n\}$ we have*

$$(1-\epsilon)\|a_i\|_2^2 \leq \|1/\sqrt{p} \cdot R^T a_i\|_2^2 \leq (1+\epsilon)\|a_i\|_2^2, \quad (12)$$

which can be accumulated over all i to calculate $\|A\|_F^2 = \|a_1\|^2 + \ldots + \|a_m\|^2$. Obviously (K1) holds.

Proof of (K2): *Performing SVD on both A and B, we get*

$$A = \sum_{i=1}^{r} \sigma_i u_i v_i^T, \quad B = \sum_{i=1}^{r} \lambda_i a_i b_i^T,$$

where we assume all singular values have been sorted in descending order, i.e.*, $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_r$ and $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_r$. To bound $\|B\|_*^2$ from below, first we observe that for any $1 \leq k \leq r$, with high probability*

$$\begin{aligned} \sum_{i=1\ldots k} \lambda_i^2 &\geq \sum_{i=1\ldots k} v_i^T B^T B v_i = \sum_{i=1\ldots k} \frac{1}{p} v_i^T A^T R R^T A v_i \\ &= \sum_{i=1\ldots k} \sigma_i^2 \left\| \frac{1}{\sqrt{p}} R^T u_i \right\|^2 \geq (1-\epsilon) \sum_{i=1\ldots k} \sigma_i^2, \end{aligned}$$

where the first inequality follows from the variational characterization of the Ky Fan k-norms [6], i.e.*,*

$$\begin{array}{l} \forall k \in \{1, rank(A)\}, \;\; \lambda_1(A) + \ldots + \lambda_k(A) \\ = \max\left\{ tr(U_k^T A U_k) \;:\; U_k \in \mathbb{R}^{n \times k} \text{ and } U_k^T U_k = I \right\}. \end{array}$$

For finite-dimensional real vectors, $\|x\|_2 \leq \|x\|_1 \leq \sqrt{d}\|x\|_2$, where d is its dimensionality. Therefore we have

$$\sum_{i=1\ldots r} \lambda_i \geq \sqrt{\sum_{i=1\ldots r} \lambda_i^2} \geq \sqrt{(1-\epsilon) \sum_{i=1\ldots r} \sigma_i^2} \geq \sqrt{\frac{1-\epsilon}{r}} \sum_{i=1\ldots r} \sigma_i,$$

which accomplishes the proof of (K2).

Proof of (K3): *The nuclear norm can be recast as the minimum of an optimization problem about the Frobenius norm [11],* i.e.*, it can be equivalently defined as*

$$\|X\|_* = \min_{X = UV^T} \|U\|_F \|V\|_F, \quad (13)$$

where (U, V) is arbitrary decomposition of X without bounded dimensionalities. Denote the SVD of A is $U \Sigma V^T$.

[3] Visit http://en.wikipedia.org/wiki/Woodbury_matrix_identity for a quick reference.

Define $U_0 = U\Sigma^{\frac{1}{2}}$ *and* $V_0 = V\Sigma^{\frac{1}{2}}$*, it can be verified* $\|A\|_* = \|U_0\|_F \cdot \|V_0\|_F$ *and*

$$\begin{aligned}\left\|\frac{1}{\sqrt{p}}R^TA\right\|_* &\leq \left\|\frac{1}{\sqrt{p}}R^TU_0\right\|_F \|V_0\|_F \\ &= \left\|\frac{1}{\sqrt{p}}R^TU\Sigma^{\frac{1}{2}}\right\|_F \cdot \sqrt{tr(V\Sigma^{\frac{1}{2}}\Sigma^{\frac{1}{2}}V^T)} \\ &\leq \sqrt{1+\epsilon}\|U\Sigma^{\frac{1}{2}}\|_F \cdot \|\Sigma^{\frac{1}{2}}V^T\|_F \qquad (14) \\ &= \sqrt{1+\epsilon}\|U_0\|_F \cdot \|V_0\|_F = \sqrt{1+\epsilon}\|A\|_*,\end{aligned}$$

where Inequality (14) follows from (K1).

5.2. Break the Curse of Large Nullspace

Both Problems (7) and (10) rely on random matrices P or Q. Unfortunately, directly optimizing with fixed P or Q will get problematic results. Recall the facts $p \ll m$ and $q \ll n$ for dimensionality reduction purpose. It immediately indicates that the row nullspace of P^T is of far larger dimension than its range space, thus optimal A^* tends to hide most of its energy into the nullspace and causes the undesired $\|P^TA^*\|_* \approx 0$. To avoid this issue, the random matrix in each iteration of Algorithm 1 is independently generated and promptly changed at the inner loop. In practice we can beforehand produce a larger random matrix M with size $sm \times sn$ with s slightly bigger than 1 (*e.g.*, 1.2 in our implementation). To sample $m \times n$ matrix P is equivalent to randomly shifting an $m \times n$ sub-window across M.

5.3. Beyond Gaussian Random Matrix

The prior analysis focuses on Gaussian random matrix case. Actually the above analysis can be straightforwardly extended to any random matrices with zero mean and unit variance, so that Lemma 1 still holds with slightly different decaying probability based on the Johnson-Lindenstrauss lemma [13] and Theorem 1 thus follows to hold. Specifically, we can employ other hashing-based sparse random matrices for the same purpose. An example is the following very-sparse matrix R with its (i, j)-th entry as

$$r_{ij} = \sqrt{k} \times \begin{cases} 1, & \text{with prob } \frac{1}{2k} \\ 0, & \text{with prob } 1 - \frac{1}{k} \\ -1, & \text{with prob } \frac{1}{2k} \end{cases}$$

where k is the parameter to control the matrix sparsity. It is easily verified $E(r_{ij}) = 0$ and $E(r_{ij}^2) = 1$.

5.4. Complexity Analysis

The original ALM based R-PCA (Section 3.2) has its computational complexity hinged on the SVD of $m \times n$ matrix, whose complexity is known to be $\mathcal{O}(m \cdot n \cdot \min(m, n))$, roughly a cubic function for near-square matrices. For our linear projection based R-PCA, the complexities of (P1)-(P3) in Section 4 are $pmn + \mathcal{O}(np^2)$, $\mathcal{O}(mn)$, and $\mathcal{O}(pmn)$ respectively, resulting an overall complexity of $\mathcal{O}(pmn)$. Recall that $p \ll m$, this complexity can be regarded as a quadratic function for near-square matrices. As discussed in Subsection 5.2, an $sm \times sn$ ($s > 1$) random matrix need to be generated beforehand, which triggers $\mathcal{O}(mn)$ extra memory compared with original R-PCA. Analysis of our proposed bilinear projection version can be likewise done, and omitted here for space limitation.

6. Experiments

Several corroborating experiments are presented, including 1) simulation on synthetic data with varying dimensionality, which emphasizes the boosted efficacy brought by our proposed projected nuclear norm, along with the accompanying accuracy loss; 2) two qualitative demonstrations on facial artifacts (shadows and specularities) removal and video background modeling; and 3) large-scale image tag transduction conducted on the benchmark of NUS-WIDE-270K, where we re-interprets the "error matrix" E to enhance image tag quality. All algorithms are implemented in pure Matlab language without unmentioned accelerating tricks and all statistics are collected based on a common PC equipped with Intel Q9550 CPU and 8GB physical memory.

6.1. Simulation on Synthetic Data

We synthesize a low-rank matrix $A \in \mathbb{R}^{m \times m}$ as a product of two $m \times r$ (r is set to be $0.05m$) matrices both drawn from the normal distribution, and additively corrupt it with sparse matrix $E \in \mathbb{R}^{m \times m}$, whose non-zero entries (10% in proportion) are uniformly distributed in $[-500, 500]$. We apply the original R-PCA and our proposed algorithms (with either linear or bilinear projections) to recover these two heterogenous sub-structures underlying $D = A + E$. The detailed comparative results are reported in Table 1.

The reported time is the seconds of a single pass during optimization (corresponding to lines 3-7 in Algorithm 1). From Table 1 we can observe a dominating superiority of projection-accelerated optimization in terms of efficacy. For the linear projection version, the overall optimization for $m = 6000$ is accomplished within 1.5 hours (CPU times), while the original R-PCA is impractical due to its slow optimization (estimated to be finished in more than two days). As suggested in [8], stopping criterion is $\|D - A_k^* - E_k^*\|_F / \|D\|_F \leq \epsilon$, where ϵ is a small positive number (*e.g.*, 10^{-7}), and it typically requires hundreds of passes until final convergence. In Figure 1 two curves are plotted to depict the convergence tendencies of the original and linear-projection based algorithms respectively, both of which demonstrate a log-linear relationship coincided with the analysis in [8] (Theorem 1 therein). The slightly slow convergence for our proposed algorithm is supposed to stem from the stochastic optimization with random matrices P and Q.

Table 1. Investigation on the simulated data. The variable m corresponds to the size of square matrices. For our proposed projection-based algorithms, the reduced dimensionalities by random matrices P and Q are both set to be $\min(0.1 \times m, 1000)$. Parameter λ is set to be $\frac{1}{\sqrt{m}}$, $\frac{1}{4\sqrt{m}}$ and $\frac{1}{4\sqrt{m}}$ in these three variants without further tuning. In the accelerated versions, to gauge the quality of estimated matrix E, small absolute entries are set to zero so that only those with top 10% magnitudes are kept. The value of $\text{ACC}(E)$ denotes the number of correctly-predicted matrix elements given the ground-truth of matrix E. Note that for $m = 6000$, the performance of original R-PCA is not reported since it is estimated to consume more than two days to converge and thus beyond the scope of the practitioners. The reported time is measured using the Matlab built-in "cputime" function in the unit of second for a single pass during the optimization (note that it is not equal to the physical time on our quad-core CPU). See text for more explanation.

m	Data Description		Original R-PCA			Linear Projection		Bilinear Projection	
	$\|A\|_*$	$\|E\|_0$	$\|A\|_*$	$\|E\|_0$	Time	$\text{ACC}(E)$	Time	$\text{ACC}(E)$	Time
500	1.24×10^4	25,000	1.24×10^4	25,013	0.84	24,283	0.11	24,254	0.10
1000	4.92×10^4	100,000	4.92×10^4	100,022	4.01	96,115	0.44	95,958	0.35
2000	1.97×10^5	400,000	1.97×10^5	400,040	21.94	378,428	1.82	377,895	1.28
4000	7.90×10^5	1,600,000	7.90×10^5	1,600,111	171.49	1,479,590	8.19	1,479,507	5.83
6000	1.78×10^6	3,600,000	N/A	N/A	N/A	3,275,617	23.01	3,276,579	16.16

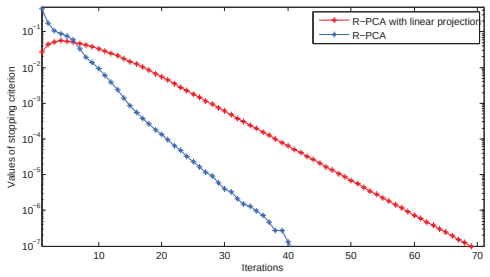

Figure 1. Comparison of convergence speed between original R-PCA and our linear projection based version. The optimization terminates when the value of stopping criterion is below 10^{-7}.

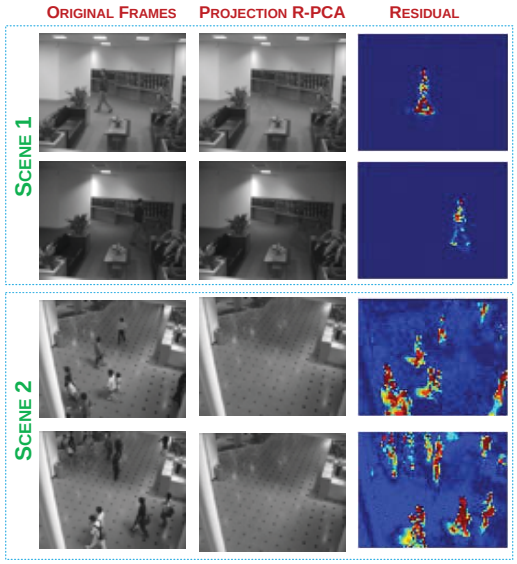

Figure 2. Background modeling on two surveillance scenes, *i.e.*, Lobby (top) and ShoppingCenter (bottom). Each frame is of 64×80 pixels and totally 500 frames are used for both tasks. The middle and right columns present the restored background from the proposed linear-projection R-PCA (reduced to 500-D) and corresponding sparse residuals.

6.2. Background Modeling and Facial Recovery

This subsection elaborates on qualitative demonstrations on structured data such as aligned faces and surveillance videos. Background modeling is a crucial operation for activity detection in surveillance video. The problem is complicated by both multiple moving objects and background variations caused by illumination etc. Following the formulation of R-PCA, we can reasonably postulate the background is controlled by few factors and hence exhibits low-rank property. Foreground activity is detected by identifying spatially localized sparse residuals. The idea is validated on a public surveillance database[4]. Figure 2 illustrates the restored background for two different scenes using the proposed linear-projection R-PCA. The results are obtained by projecting original 5120-D frame onto 500-D Gaussian random matrices.

A key observation in face recognition is that faces of the same person captured under varying illumination approximately reside in a low-rank linear subspace known as the *harmonic plane* [16]. However, real face data often suffer from self-shadowing and specularity when captured under directional illumination. These aforementioned factors can be naturally mapped to the ingredients in R-PCA, which enables shadow-free facial recovery by separating low-rank matrix A from the original data matrix D. The idea is validated on Extended Yale-B database, as shown at the left of Figure 3. The original 32256-D face feature vectors (*i.e.*, 192×168 pixels) are reduced to 2000-D in the implementation of our proposed algorithm, which produces comparable results to the original R-PCA. We also investigate the influence of parameter p (the reduced dimensionality after random projection) on the right panel of Figure 3, from which it is observed that increasing the value of p continuously approaches the "ground truth" generated by original R-PCA.

6.3. Large-Scale Image Tag Transduction

Finally we demonstrate the power of the proposed method on the task of image tag transduction. Recent tremendous accumulation of socially-sharing images and personal albums has raised new challenges to tag-based semantic image indexing and retrieval. Unfortunately, most of Web images have no tags or are casually assigned noisy tags. Image tag transduction refers to the effort of propagating known tags (usually annotated by experienced volunteers) on selected images to the rest un-annotated image collections. We experiment on the large-scale dataset NUS-

[4] http://perception.i2r.a-star.edu.sg/bk_model/bk_index.html

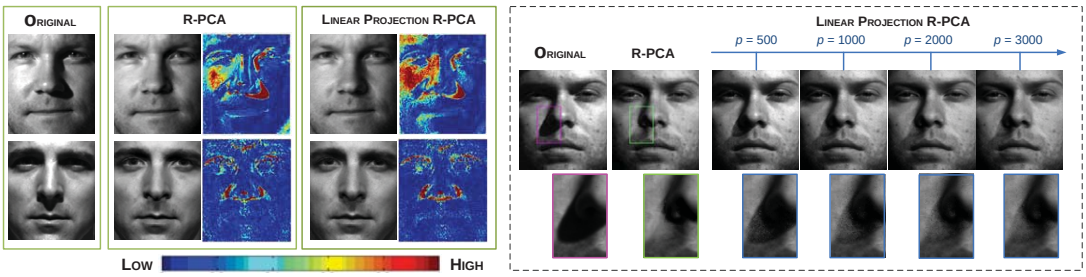

Figure 3. **Left**: Removal of facial specularities and shadows on Extended Yale-B database. The original face size is 192×168 pixels. Roughly 30 images are used for each individual. **Right**: Study on the effect of reduced dimensionality in our proposed projection-based R-PCA. Close-up views of specific patches are displayed to highlight the details.

WIDE-270K[5] collected from Flickr, which is comprised of 269,648 images accompanied with around 1K tags. A group of volunteers are solicited to manually annotate the groundtruth tags of each image on a subset of 81 tags, averagely roughly 2 tags per image, which are used here for evaluation. We randomly split the whole images into two even subsets for training or testing respectively.

A large part of existing tag transduction algorithms fall in a two-step paradigm, *i.e.*, first tag propagation to obtain the tag-image association confidences, and then tag inference from the real-valued confidence values. The former step has gained extensive studies in past years, especially in the context of graph-based modeling. Instead of contributing another tag propagation algorithm, our study focuses on revisiting the strategies in the second step, where the idea of R-PCA is used to promote adaptive tag inference. For completeness, we present some details in implementing tag propagation. We extract three types of features, including 144-D color correlogram, 73-D edge direction histogram, and 225-D block-wise color moments, forming a 442-D global feature for each image, from which hashing-accelerated ℓ_1-graph is efficiently built [4]. Known tags on the training set are then disseminated to the rest images based on the belief of inter-connection of visual appearance and semantics. Specifically, Markov-style tag propagation is employed. Let $f(p)$ be the tag vector for image p, the updating rule is $f^{(t)}(p) \leftarrow \sum_{q=1\ldots k} \omega_{pq} \cdot f^{(t-1)}(q)$, where ω_{pq} denotes the pairwise affinity value in the ℓ_1-graph.

The tag-propagation performance on the ℓ_1-graph measured by Mean Accuracy Precision (MAP) is 0.126 across all tags, compared with 0.072 on traditional k-NN graph. Although low performance at first glance, it approaches the state-of-the-art [4] on this notoriously challenging dataset. Standard tag inference from real-valued confidences is done by selecting top-k confident tags for each image, denoted as *NaiveThres* hereafter. Such a scheme is vulnerable to unbalanced annotations between different tags and tends to keep abundantly-annotated tags. The continuous distributions of propagated confidences (see the right one of Figure 4) further complicate determining the optimal k, and arbitrary k uniformly applied to all images ignores the fact of uneven tag assignment numbers across image collections. See Figure 4 for corroborating statistics.

[5] http://lms.comp.nus.edu.sg/research/NUS-WIDE.htm

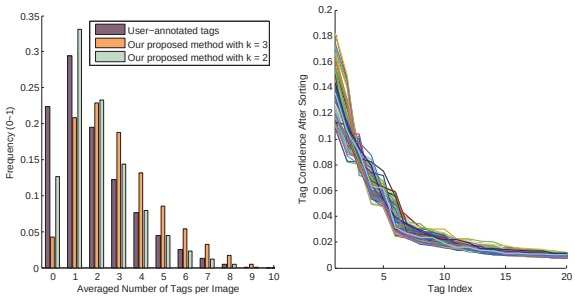

Figure 4. Left: Distributions of tag numbers per image on NUS-WIDE-270K. Note that our proposed algorithm results in adaptive distributions unlike *NaiveThres*. Right: Top-20 tag confidences after propagation on the ℓ_1-graph for randomly-sampled 100 images.

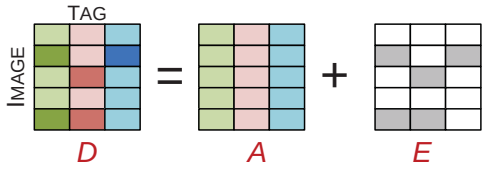

Figure 5. Illustrative sketch of the proposed image tag assignment scheme. The initial tag-confidence matrix is supposed to contain a low-rank component and a sparse image-tag assignment matrix. Each tag is displayed in different colors and darker color indicates higher confidence. Better viewing in color mode.

We propose a novel scheme that adaptively determines the tag number and assignment for each image. Denote the tag matrix after propagation as $D \in \mathbb{R}^{n \times t}$, where n, t correspond to the image and tag numbers. The key idea of our proposed scheme is to decompose it into a low-rank matrix A and a sparse matrix E using R-PCA techniques. The intuition behind the decomposition is illustrated in Figure 5. Aforementioned combinative propagation for each tag results in a near-flat confidence distribution across all images along with a bunch of spiky ones (optimal candidates for final tag assignment). As shown in Figure 5, the low-rank matrix A can be useful for mode identification, and matrix E conveys sparse tags violating the low-rank assumption, rather than destructive noises as in the prior applications.

The final tag assignments are decided by setting the en-

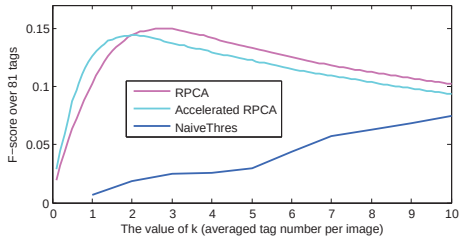

Figure 6. Performance comparison in terms of F-score for *NaiveThres*, RPCA and our linear projection based version in $k \in [1, 10]$. Recall that the averaged tag number per image is roughly 2 on NUS-WIDE-270K. In the latter two algorithms $\lambda = 1/\sqrt{n}$ (n is the image number). The peak performance of *NaiveThres* is 0.094 achieved at $k = 17$ over $k \in [1, 81]$, however, such a k value is unreasonable for real-world image tags.

Figure 7. Exemplar tag transduction results on NUS-WIDE-270K.

tries of matrix E above a specific threshold to be 1, otherwise 0. To learn this threshold, all non-zero entries in E are sorted in descending order. The threshold is chosen so that averaged tag assignments per image equal k, therefore we equivalently use parameter k to balance precision and recall rate (the implication of k is intrinsically different from that in *NaiveThres*). Figure 6 illustrates the performance under varying k for both *NaiveThres* and R-PCA based techniques, in terms of F-score. The peak performances of original R-PCA and our linear-projection accelerated version are 0.150 (achieved at $k = 2.6$) and 0.144 (achieved at $k = 1.7$) respectively, compared with 0.094 by *NaiveThres* (achieved at $k = 17$). Such a simple scheme enhances the tag quality to a salient extent compared with the standard *NaiveThres*. Considering the popularity of propagation-based method in tag transduction, we believe the proposed scheme could be plugged in and promote other relevant algorithms. Proof of image-adaptive selection of tag numbers can be found on the left of Figure 4 and exemplar image tagging results are found in Figure 7.

7. Conclusion and Future Work

This paper demonstrates the power of projected matrix nuclear norm by reformulating R-PCA. The proposed method brings tremendous speedup in optimization by avoiding large-scale SVD, meanwhile the performance loss is controllable. Theoretic analysis on its bounding property, extensions to other random and hashing matrices and complexities is also provided. Finally we present both qualitative and quantitative evaluations on various datasets, including one large-scale image tagging databases. Here it is worthy to highlight that the projected nuclear norm is a general tool and immediately applicable to many other nuclear norm oriented formulations.

Acknowledgement

This research is done for CSIDM Project No. CSIDM-200803 partially funded by a grant from the National Research Foundation (NRF) administered by the Media Development Authority (MDA) of Singapore.

References

[1] O. ambrym Maillard and R. Munos. Compressed least-squares regression. In *NIPS*, 2009. 2609, 2610

[2] C. Boutsidis and P. Drineas. Random projections for the nonnegative least-squares problem. *Linear Algebra and Its Applications*, 431:760–771, 2009. 2609, 2610

[3] J.-F. Cai, E. Candes, and Z. Shen. A singular value thresholding algorithm for matrix completion. *SIAM Journal of Optimization*, 20(4):1956–1982, 2010. 2611

[4] X. Chen, Y. Mu, S. Yan, and T.-S. Chua. Efficient large-scale image annotation by probabilistic collaborative multi-label propagation. In *ACM Multimedia*, 2010. 2615

[5] D. Forsyth and J. Ponce. *Computer Vision: A Modern Approach*. Prentice Hall Professional Technical Reference, 2002. 2609

[6] R. Horn. *Topics in matrix analysis*. Cambridge University Press, New York, NY, USA, 1986. 2612

[7] F. D. la Torre and M. Black. Robust principal component analysis for computer vision. In *ICCV*, 2001. 2609

[8] 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. 2610, 2611, 2613

[9] V. Rokhlin and M. Tygert. A fast randomized algorithm for overdetermined linear least-squares regression. *Proceedings of the National Academy of Sciences*, 105(36):13212–13217, 2008. 2610

[10] Q. Shi, C. Shen, and H. Li. Rapid face recognition using hashing. In *CVPR*, 2010. 2610

[11] N. Srebro. *Learing with Matrix Factorizations*. PhD thesis, Department of Computer Science, MIT, 2004. 2609, 2612

[12] N. Srebro, J. Rennie, and T. Jaakkola. Maximum margin matrix factorizations. In *NIPS*, 2005. 2610

[13] S. Vempala. *The Random Projection Method*. Dimacs Series in Discrete Math, 2005. 2612, 2613

[14] A. Wagner, J. Wright, A. Ganesh, Z. Zhou, and Y. Ma. Towards a practical face recognition system: Robust registration and illumination by sparse representation. In *CVPR*, 2009. 2610

[15] F. Wang and P. Li. Efficient nonnegative matrix factorization with random projections. In *SDM*, 2010. 2609, 2610

[16] J. Wright, A. Ganesh, S. Rao, and Y. Ma. Robust principal component analysis: Exact recovery of corrupted low-rank matrices. In *NIPS*, 2009. 2609, 2610, 2614

[17] J. Wright and Y. Ma. Dense error correction via l1-minimization. In *ICASSP*, 2009. 2610

[18] J. Wright, A. Yang, A. Ganesh, S. Sastry, and Y. Ma. Robust face recognition via sparse representation. *IEEE Transactions Pattern Analysis Machine Intelligence*, 31(2):210–227, 2009. 2610

[19] L. Wu, A. Ganesh, B. Shi, Y. Matsushita, Y. Wang, and Y. Ma. Robust photometric stereo via low-rank matrix completion and recovery. In *ACCV*, 2010. 2609

[20] J. Yang, J. Wright, T. Huang, and Y. Ma. Image super-resolution as sparse representation of raw image patches. In *CVPR*, 2008. 2610