Bilinear Factor Matrix Norm Minimization for Robust PCA: Algorithms and Applications

Fanhua Shang, *Member, IEEE*, James Cheng, Yuanyuan Liu, Zhi-Quan Luo, *Fellow, IEEE*, and Zhouchen Lin, *Senior Member, IEEE*

Abstract—The heavy-tailed distributions of corrupted outliers and singular values of all channels in low-level vision have proven effective priors for many applications such as background modeling, photometric stereo and image alignment. And they can be well modeled by a hyper-Laplacian. However, the use of such distributions generally leads to challenging non-convex, non-smooth and non-Lipschitz problems, and makes existing algorithms very slow for large-scale applications. Together with the analytic solutions to ℓ_p-norm minimization with two specific values of p, i.e., $p=1/2$ and $p=2/3$, we propose two novel bilinear factor matrix norm minimization models for robust principal component analysis. We first define the double nuclear norm and Frobenius/nuclear hybrid norm penalties, and then prove that they are in essence the Schatten-$1/2$ and $2/3$ quasi-norms, respectively, which lead to much more tractable and scalable Lipschitz optimization problems. Our experimental analysis shows that both our methods yield more accurate solutions than original Schatten quasi-norm minimization, even when the number of observations is very limited. Finally, we apply our penalties to various low-level vision problems, e.g., text removal, moving object detection, image alignment and inpainting, and show that our methods usually outperform the state-of-the-art methods.

Index Terms—Robust principal component analysis, rank minimization, Schatten-p quasi-norm, ℓ_p-norm, double nuclear norm penalty, Frobenius/nuclear norm penalty, alternating direction method of multipliers (ADMM)

1 Introduction

The sparse and low-rank priors have been widely used in many real-world applications in computer vision and pattern recognition, such as image restoration [1], face recognition [2], subspace clustering [3], [4], [5] and robust principal component analysis [6] (RPCA, also called low-rank and sparse matrix decomposition in [7], [8] or robust matrix completion in [9]). Sparsity plays an important role in various low-level vision tasks. For instance, it has been observed that the gradient of natural scene images can be better modeled with a heavy-tailed distribution such as hyper-Laplacian distributions ($p(x)\propto e^{-k|x|^{\alpha}}$, typically with $0.5\leq\alpha\leq 0.8$, which correspond to non-convex ℓ_p-norms) [10], [11], as exhibited by the sparse noise/outliers in low-level vision problems [12] shown in Fig. 1. To induce sparsity, a principled way is to use the convex ℓ_1-norm [6], [13], [14], [15], [16], [17], which is the closest convex relaxation of the sparser ℓ_p-norm, with compressed sensing being a prominent example. However, it has been shown in [18] that the ℓ_1-norm over-penalizes large entries of vectors and results in a biased solution. Compared with the ℓ_1-norm, many non-convex surrogates of the ℓ_0-norm listed in [19] give a closer approximation, e.g., SCAD [18] and MCP [20]. Although the use of hyper-Laplacian distributions makes the problems non-convex, fortunately an analytic solution can be derived for two specific values of p, $1/2$ and $2/3$, by finding the roots of a cubic and quartic polynomial, respectively [10], [21], [22]. The resulting algorithm can be several orders of magnitude faster than existing algorithms [10].

As an extension from vectors to matrices, the low-rank structure is the sparsity of the singular values of a matrix. Rank minimization is a crucial regularizer to induce a low-rank solution. To solve such a problem, the rank function is usually relaxed by its convex envelope [5], [23], [24], [25], [26], [27], [28], [29], the nuclear norm (i.e., the sum of the singular values, also known as the trace norm or Schatten-1 norm [26], [30]). By realizing the intimate relationship between the ℓ_1-norm and nuclear norm, the latter also over-penalizes large singular values, that is, it may make the solution deviate from the original solution as the ℓ_1-norm does [19], [31]. Compared with the nuclear norm, the Schatten-q norm for $0<q<1$ is equivalent to the ℓ_q-norm on singular values and makes a closer approximation to the rank function [32], [33]. Nie et al. [31] presented an efficient augmented Lagrange multiplier (ALM) method to solve the joint ℓ_p-norm and Schatten-q norm (LpSq) minimization. Lai et al. [32] and Lu et al. [33] proposed iteratively reweighted least squares methods for solving Schatten quasi-norm minimization problems. However, all these algorithms have to

- F. Shang, J. Cheng, and Y. Liu are with the Department of Computer Science and Engineering, The Chinese University of Hong Kong, Hong Kong. E-mail: {fhshang, jcheng, yyliu}@cse.cuhk.edu.hk.
- Z.-Q. Luo is with Shenzhen Research Institute of Big Data, the Chinese University of Hong Kong, Shenzhen, China and Department of Electrical and Computer Engineering, University of Minnesota, Minneapolis, MN 55455. E-mail: luozq@cuhk.edu.cn.
- Z. Lin is with Key Laboratory of Machine Perception (MOE), School of EECS, Peking University, Beijing 100871, P.R. China and with the Cooperative Medianet Innovation Center, Shanghai Jiao Tong University, Shanghai 200240, P.R. China. E-mail: zlin@pku.edu.cn.

Manuscript received 10 June 2016; revised 6 June 2017; accepted 24 July 2017. Date of publication 3 Sept. 2017; date of current version 13 Aug. 2018.
(Corresponding authors: Fanhua Shang and Yuanyuan Liu.)
Recommended for acceptance by D. Xu.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TPAMI.2017.2748590

Fig. 1. Sparsity priors. Left: A typical image of video sequences. Middle: the sparse component recovered by [6]. Right: the empirical distribution of the sparse component (blue solid line), along with a hyper-Laplacian fit with $\alpha = 1/2$ (green dashdot line) and $\alpha = 2/3$ (red dotted line).

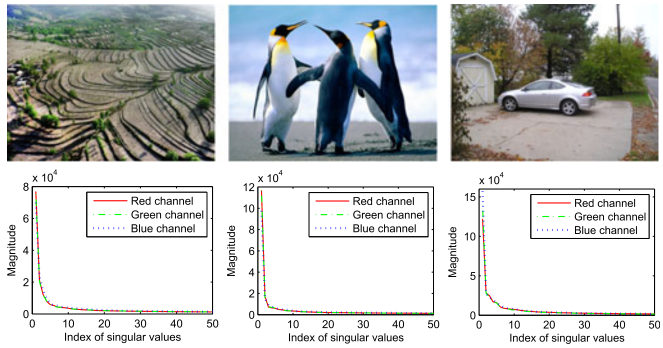


Fig. 2. The heavy-tailed empirical distributions (bottom) of the singular values of the three channels of these three images (top).

be solved iteratively, and involve singular value decomposition (SVD) in each iteration, which occupies the largest computation cost, $\mathcal{O}(\min(m,n)mn)$ [34], [35].

It has been shown in [36] that the singular values of non-local matrices in natural images usually exhibit a heavy-tailed distribution ($p(\sigma) \propto e^{-k|\sigma|^{\alpha}}$), as well as the similar phenomena in natural scenes [37], [38], as shown in Fig. 2. Similar to the case of heavy-tailed distributions of sparse outliers, the analytic solutions can be derived for the two specific cases of α, $1/2$ and $2/3$. However, such algorithms have high per-iteration complexity $\mathcal{O}(\min(m,n)mn)$. Thus, we naturally want to design equivalent, tractable and scalable forms for the two cases of the Schatten-q quasi-norm, $q = 1/2$ and $2/3$, which can fit the heavy-tailed distribution of singular values closer than the nuclear norm, as analogous to the superiority of the ℓ_p quasi-norm (hyper-Laplacian priors) to the ℓ_1-norm (Laplacian priors).

We summarize the main contributions of this work as follows. 1) By taking into account the heavy-tailed distributions of both sparse noise/outliers and singular values of matrices, we propose two novel tractable bilinear factor matrix norm minimization models for RPCA, which can fit empirical distributions very well to corrupted data. 2) Different from the definitions in our previous work [34], we define the double nuclear norm and Frobenius/nuclear hybrid norm penalties as tractable low-rank regularizers. Then we prove that they are in essence the Schatten-$1/2$ and $2/3$ quasi-norms, respectively. The solution of the resulting minimization problems only requires SVDs on two much smaller factor matrices as compared with the much larger ones required by existing algorithms. Therefore, our algorithms can reduce the per-iteration complexity from $\mathcal{O}(\min(m,n)mn)$ to $\mathcal{O}(mnd)$, where $d \ll m, n$ in general. In particular, our penalties are Lipschitz, and more tractable and scalable than original Schatten quasi-norm minimization, which is non-Lipschitz and generally NP-hard [32], [39]. 3) Moreover, we present the convergence property of the proposed algorithms for minimizing our RPCA models and provide their proofs. We also extend our algorithms to solve matrix completion problems, e.g., image inpainting. 4) We empirically study both of our bilinear factor matrix norm minimizations and show that they outperform original Schatten norm minimization, even with only a few observations. Finally, we apply the defined low-rank regularizers to address various low-level vision problems, e.g., text removal, moving object detection, and image alignment and inpainting, and obtain superior results than existing methods.

2 Related Work

In this section, we mainly discuss some recent advances in RPCA, and briefly review some existing work on RPCA and its applications in computer vision (readers may see [6] for a review). RPCA [24], [40] aims to recover a low-rank matrix $L \in \mathbb{R}^{m\times n}$ ($m \geq n$) and a sparse matrix $S \in \mathbb{R}^{m\times n}$ from corrupted observations $D = L^* + S^* \in \mathbb{R}^{m\times n}$ as follows:

$$\min_{L,S}\ \lambda\,\mathrm{rank}(L) + \|S\|_{\ell_0},\ \mathrm{s.t.}, L + S = D, \tag{1}$$

where $\|\cdot\|_{\ell_0}$ denotes the ℓ_0-norm[1] and $\lambda > 0$ is a regularization parameter. Unfortunately, solving (1) is NP-hard. Thus, we usually use the convex or non-convex surrogates to replace both of the terms in (1), and formulate this problem into the following more general form

$$\min_{L,S}\ \lambda\|L\|_{S_q}^q + \|S\|_{\ell_p}^p,\ \mathrm{s.t.}, \mathcal{P}_\Omega(L+S) = \mathcal{P}_\Omega(D), \tag{2}$$

where in general $p, q \in [0, 2]$, $\|S\|_{\ell_p}$ and $\|L\|_{S_q}$ are depicted in Table 1 and can be seen as the loss term and regularized term, respectively, and $\mathcal{P}_\Omega$ is the orthogonal projection onto the linear subspace of matrices supported on $\Omega := \{(i,j) | D_{ij}\ \text{is observed}\}$: $\mathcal{P}_\Omega(D)_{ij} = D_{ij}$ if $(i,j) \in \Omega$ and $\mathcal{P}_\Omega(D)_{ij} = 0$ otherwise. If Ω is a small subset of the entries of the matrix, (2) is also known as the robust matrix completion problem as in [9], and it is impossible to exactly recover S^* [41]. As analyzed in [42], we can easily see that the optimal solution $S_{\Omega^c} = \mathbf{0}$, where Ω^c is the complement of Ω, i.e., the index set of unobserved entries. When $p = 2$ and $q = 1$, (2) becomes a nuclear norm regularized least squares problem as in [43] (e.g., image inpainting in Section 6.3.4).

2.1 Convex Nuclear Norm Minimization

In [6], [40], [44], both of the non-convex terms in (1) are replaced by their convex envelopes, i.e., the nuclear norm ($q = 1$) and the ℓ_1-norm ($p = 1$), respectively.

$$\min_{L,S}\ \lambda\|L\|_* + \|S\|_{\ell_1},\ \mathrm{s.t.}, L + S = D. \tag{3}$$

Wright et al. [40] and Candès et al. [6] proved that, under some mild conditions, the convex relaxation formulation (3) can exactly recover the low-rank and sparse matrices (L^*, S^*) with high probability. The formulation (3) has been widely used in many computer vision applications, such as

1. Strictly speaking, the ℓ_0-norm is not actually a norm, and is defined as the number of non-zero elements. When $p \geq 1$, $\|S\|_{\ell_p}$ strictly defines a norm which satisfies the three norm conditions, while it defines a quasi-norm when $0 < p < 1$. Due to the relationship between $\|S\|_{\ell_p}$ and $\|L\|_{S_q}$, the latter has the same cases as the former.

TABLE 1
The Norms of Sparse and Low-Rank Matrices

p, q	Sparsity		Low-rankness	
0	$\|S\|_{\ell_0}$	ℓ_0-norm	$\|L\|_{S_0}=\|\sigma\|_{\ell_0}$	rank
$(0,1)$	$\|S\|_{\ell_p}=(\sum\|S_{ij}\|^p)^{1/p}$	ℓ_p-norm	$\|L\|_{S_q}=(\sum\sigma_k^q)^{1/q}$	Schatten-q norm
1	$\|S\|_{\ell_1}=\sum\|S_{ij}\|$	ℓ_1-norm	$\|L\|_*=\sum\sigma_k$	nuclear norm
2	$\|S\|_F=\sqrt{\sum S_{ij}^2}$	Frobenius norm	$\|L\|_F=\sqrt{\sum\sigma_k^2}$	Frobenius norm

Let $\sigma=(\sigma_1,\ldots,\sigma_r)\in\mathbb{R}^r$ be the non-zero singular values of $L\in\mathbb{R}^{m\times n}$.

object detection and background subtraction [17], image alignment [50], low-rank texture analysis [29], image and video restoration [51], and subspace clustering [27]. This is mainly because the optimal solutions of the sub-problems involving both terms in (3) can be obtained by two well-known proximal operators: the singular value thresholding (SVT) operator [23] and the soft-thresholding operator [52]. The ℓ_1-norm penalty in (3) can also be replaced by the $\ell_{1,2}$-norm as in outlier pursuit [28], [53], [54], [55] and subspace learning [5], [56], [57].

To efficiently solve the popular convex problem (3), various first-order optimization algorithms have been proposed, especially the alternating direction method of multipliers [58] (ADMM, or also called inexact ALM in [44]). However, they all involve computing the SVD of a large matrix of size $m\times n$ in each iteration, and thus suffer from high computational cost, which severely limits their applicability to large-scale problems [59], as well as existing Schatten-q quasi-norm $(0<q<1)$ minimization algorithms such as LpSq [31]. While there have been many efforts towards fast SVD computation such as partial SVD [60], the performance of those methods is still unsatisfactory for many real applications [59], [61].

2.2 Non-Convex Formulations

To address this issue, Shen et al. [47] efficiently solved the RPCA problem by factorizing the low-rank component into two smaller factor matrices, i.e., $L=UV^T$ as in [62], where $U\in\mathbb{R}^{m\times d}$, $V\in\mathbb{R}^{n\times d}$, and usually $d\ll\min(m,n)$, as well as the matrix tri-factorization (MTF) [48] and factorized data [63] cases. In [16], [42], [64], [65], [66], the column-orthonormal constraint is imposed on the first factor matrix U. According to the following matrix property, the original convex problem (3) can be reformulated as a smaller matrix nuclear norm minimization problem.

Property 1. For any matrix $L\in\mathbb{R}^{m\times n}$ and $L=UV^T$. If U has orthonormal columns, i.e., $U^TU=I_d$, then $\|L\|_*=\|V\|_*$.

Cabral et al. [49] and Kim et al. [15] replaced the nuclear norm regularizer in (3) with the equivalent non-convex formulation stated in Lemma 1, and proposed scalable bilinear spectral regularized models, similar to collaborative filtering applications in [26], [67]. In [15], an elastic-net regularized matrix factorization model was proposed for subspace learning and low-level vision problems.

Besides the popular nuclear norm, some variants of the nuclear norm were presented to yield better performance. Gu et al. [37] proposed a weighted nuclear norm (i.e., $\|L\|_{w,*}=\sum_i w_i\sigma_i$, where $w=[w_1,\ldots,w_n]^T$), and assigned different weights w_i to different singular values such that the shrinkage operator becomes more effective. Hu et al. [38] first used the truncated nuclear norm (i.e., $\|L\|_d=\sum_{i=d+1}^n\sigma_i$) to address image recovery problems. Subsequently, Oh et al. [45] proposed an efficient partial singular value thresholding (PSVT) algorithm to solve many RPCA problems of low-level vision. The formulations mentioned above are summarized in Table 2.

3 Bilinear Factor Matrix Norm Minimization

In this section, we first define the double nuclear norm and Frobenius/nuclear hybrid norm penalties, and then prove the equivalence relationships between them and the Schatten quasi-norms. Incorporating with hyper-Laplacian priors of both sparse noise/outliers and singular values, we propose two novel bilinear factor matrix norm regularized models for RPCA. Although the two models are still non-convex and even non-smooth, they are more tractable and scalable optimization problems, and their each factor matrix term is convex. On the contrary, the original Schatten quasi-norm minimization problem is very difficult to solve

TABLE 2
Comparison of Various RPCA Models and Their Properties

Model	Objective function	Constraints	Parameters	Convex?	Per-iteration Complexity
RPCA [6], [44]	$\lambda\|L\|_*+\|S\|_{\ell_1}$	$L+S=D$	λ	Yes	$\mathcal{O}(mn^2)$
PSVT [45]	$\lambda\|L\|_d+\|S\|_{\ell_1}$	$L+S=D$	λ, d	No	$\mathcal{O}(mn^2)$
WNNM [46]	$\lambda\|L\|_{w,*}+\|S\|_{\ell_1}$		λ		
LMaFit [47]	$\|D-L\|_{\ell_1}$	$UV^T=L$	d	No	$\mathcal{O}(mnd)$
MTF [48]	$\lambda\|W\|_*+\|S\|_{\ell_1}$	$UWV^T+S=D, U^TU=V^TV=I_d$	λ, d		
RegL1 [16]	$\lambda\|V\|_*+\|\mathcal{P}_\Omega(D-UV^T)\|_{\ell_1}$	$U^TU=I_d$	λ, d		
Unifying [49]	$\frac{\lambda}{2}(\|U\|_F^2+\|V\|_F^2)+\|\mathcal{P}_\Omega(D-L)\|_{\ell_1}$	$UV^T=L$	λ, d		
factEN [15]	$\frac{\lambda_1}{2}(\|U\|_F^2+\|V\|_F^2)+\frac{\lambda_2}{2}\|L\|_F^2+\|\mathcal{P}_\Omega(D-L)\|_{\ell_1}$	$UV^T=L$	λ_1,λ_2,d		
LpSq [31]	$\lambda\|L\|_{S_q}^q+\|\mathcal{P}_\Omega(D-L)\|_{\ell_p}^p\ (0<p,q\leq 1)$	$L+S=D$	λ	No	$\mathcal{O}(mn^2)$
$(\text{S+L})_{1/2}$	$\frac{\lambda}{2}(\|U\|_*+\|V\|_*)+\|\mathcal{P}_\Omega(S)\|_{\ell_{1/2}}^{1/2}$	$L+S=D, UV^T=L$	λ, d	No	$\mathcal{O}(mnd)$
$(\text{S+L})_{2/3}$	$\frac{\lambda}{3}(\|U\|_F^2+2\|V\|_*)+\|\mathcal{P}_\Omega(S)\|_{\ell_{2/3}}^{2/3}$				

Note that $U\in\mathbb{R}^{m\times d}$ and $V\in\mathbb{R}^{n\times d}$ are the factor matrices of L, i.e., $L=UV^T$.

because it is generally non-convex, non-smooth, and non-Lipschitz, as well as the ℓ_q quasi-norm [39].

As in some collaborative filtering applications [26], [67], the nuclear norm has the following alternative non-convex formulations.

Lemma 1. *For any matrix $X \in \mathbb{R}^{m\times n}$ of rank at most $r \le d$, the following equalities hold*

$$\begin{aligned}\|X\|_* &= \min_{U\in\mathbb{R}^{m\times d},V\in\mathbb{R}^{n\times d}:X=UV^T} \frac{1}{2}(\|U\|_F^2+\|V\|_F^2)\\ &= \min_{U,V:X=UV^T}\|U\|_F\|V\|_F.\end{aligned} \tag{4}$$

The bilinear spectral penalty in the first equality of (4) has been widely used in low-rank matrix completion and recovery problems, such as RPCA [15], [49], online RPCA [68], matrix completion [69], and image inpainting [25].

3.1 Double Nuclear Norm Penalty

Inspired by the equivalence relation between the nuclear norm and the bilinear spectral penalty, our double nuclear norm (D-N) penalty is defined as follows.

Definition 1. *For any matrix $X \in \mathbb{R}^{m\times n}$ of rank at most $r \le d$, we decompose it into two factor matrices $U \in \mathbb{R}^{m\times d}$ and $V \in \mathbb{R}^{n\times d}$ such that $X = UV^T$. Then the double nuclear norm penalty of X is defined as*

$$\|X\|_{\text{D-N}} = \min_{U,V:X=UV^T}\frac{1}{4}(\|U\|_*+\|V\|_*)^2. \tag{5}$$

Different from the definition in [34], [70], i.e., $\min_{U,V:X=UV^T}\|U\|_*\|V\|_*$, which cannot be used directly to solve practical problems, Definition 1 can be directly used in practical low-rank matrix completion and recovery problems, e.g., RPCA and image recovery. Analogous to the well-known Schatten-q quasi-norm [31], [32], [33], the double nuclear norm penalty is also a quasi-norm, and their relationship is stated in the following theorem.

Theorem 1. *The double nuclear norm penalty $\|\cdot\|_{\text{D-N}}$ is a quasi-norm, and also the Schatten-1/2 quasi-norm, i.e.,*

$$\|X\|_{\text{D-N}} = \|X\|_{S_{1/2}}. \tag{6}$$

To prove Theorem 1, we first give the following lemma, which is mainly used to extend the well-known trace inequality of John von Neumann [71], [72].

Lemma 2. *Let $X \in \mathbb{R}^{n\times n}$ be a symmetric positive semi-definite (PSD) matrix and its full SVD be $U_X\Sigma_X U_X^T$ with $\Sigma_X = \mathrm{diag}(\lambda_1,\ldots,\lambda_n)$. Suppose Y is a diagonal matrix (i.e., $Y = \mathrm{diag}(\tau_1,\ldots,\tau_n)$), and if $\lambda_1 \ge \cdots \ge \lambda_n \ge 0$ and $0 \le \tau_1 \le \cdots \le \tau_n$, then*

$$\mathrm{Tr}(X^TY) \ge \sum_{i=1}^n \lambda_i\tau_i.$$

Lemma 2 can be seen as a special case of the well-known von Neumann's trace inequality, and its proof is provided in the Supplementary Materials, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TPAMI.2017.2748590.

Lemma 3. *For any matrix $X = UV^T \in \mathbb{R}^{m\times n}$, $U \in \mathbb{R}^{m\times d}$ and $V \in \mathbb{R}^{n\times d}$, the following inequality holds*

$$\frac{1}{2}(\|U\|_*+\|V\|_*) \ge \|X\|_{S_{1/2}}^{1/2}.$$

Proof. Let $U = L_U\Sigma_U R_U^T$ and $V = L_V\Sigma_V R_V^T$ be the thin SVDs of U and V, where $L_U \in \mathbb{R}^{m\times d}$, $L_V \in \mathbb{R}^{n\times d}$, and $R_U, \Sigma_U, R_V, \Sigma_V \in \mathbb{R}^{d\times d}$. Let $X = L_X\Sigma_X R_X^T$, where the columns of $L_X \in \mathbb{R}^{m\times d}$ and $R_X \in \mathbb{R}^{n\times d}$ are the left and right singular vectors associated with the top d singular values of X with rank at most r $(r \le d)$, and $\Sigma_X = \mathrm{diag}([\sigma_1(X),\ldots,\sigma_r(X),0,\ldots,0]) \in \mathbb{R}^{d\times d}$. Suppose $W_1 = L_X\Sigma_U L_X^T$, $W_2 = L_X\Sigma_U^{\frac{1}{2}}\Sigma_X^{\frac{1}{2}}L_X^T$ and $W_3 = L_X\Sigma_X L_X^T$, we first construct the following PSD matrices $\mathcal{M}_1 \in \mathbb{R}^{2m\times 2m}$ and $\mathcal{S}_1 \in \mathbb{R}^{2m\times 2m}$

$$\mathcal{M}_1 = \begin{bmatrix} -L_X\Sigma_U^{\frac{1}{2}} \\ L_X\Sigma_X^{\frac{1}{2}}\end{bmatrix}\begin{bmatrix} -\Sigma_U^{\frac{1}{2}}L_X^T & \Sigma_X^{\frac{1}{2}}L_X^T\end{bmatrix} = \begin{bmatrix} W_1 & -W_2 \\ -W_2^T & W_3\end{bmatrix} \succeq 0,$$

$$\mathcal{S}_1 = \begin{bmatrix} I_n \\ L_U\Sigma_U^{-\frac{1}{2}}L_U^T\end{bmatrix}\begin{bmatrix} I_n & L_U\Sigma_U^{-\frac{1}{2}}L_U^T\end{bmatrix} = \begin{bmatrix} I_n & L_U\Sigma_U^{-\frac{1}{2}}L_U^T \\ L_U\Sigma_U^{-\frac{1}{2}}L_U^T & L_U\Sigma_U^{-1}L_U^T\end{bmatrix} \succeq 0.$$

Because the trace of the product of two PSD matrices is always non-negative (see the proof of Lemma 6 in [73]), we have

$$\mathrm{Tr}\left(\begin{bmatrix} I_n & L_U\Sigma_U^{-\frac{1}{2}}L_U^T \\ L_U\Sigma_U^{-\frac{1}{2}}L_U^T & L_U\Sigma_U^{-1}L_U^T\end{bmatrix}\begin{bmatrix} W_1 & -W_2 \\ -W_2^T & W_3\end{bmatrix}\right) \ge 0.$$

By further simplifying the above expression, we obtain

$$\mathrm{Tr}(W_1) - 2\mathrm{Tr}(L_U\Sigma_U^{-\frac{1}{2}}L_U^TW_2^T) + \mathrm{Tr}(L_U\Sigma_U^{-1}L_U^TW_3) \ge 0. \tag{7}$$

Recalling $X = UV^T$ and $L_X\Sigma_X R_X^T = L_U\Sigma_U R_U^T R_V\Sigma_V L_V^T$, thus $L_U^TL_X\Sigma_X = \Sigma_U R_U^T R_V\Sigma_V L_V^T R_X$. Due to the orthonormality of the columns of L_U, R_U, L_V, R_V, L_X and R_X, and using the well-known von Neumann's trace inequality [71], [72], we obtain

$$\begin{aligned}\mathrm{Tr}(L_U\Sigma_U^{-1}L_U^TW_3) &= \mathrm{Tr}(L_U\Sigma_U^{-1}L_U^TL_X\Sigma_X L_X^T)\\ &= \mathrm{Tr}(L_U\Sigma_U^{-1}\Sigma_U R_U^TR_V\Sigma_V L_V^TR_XL_X^T)\\ &= \mathrm{Tr}(L_UR_U^TR_V\Sigma_VL_V^TR_XL_X^T)\\ &\le \mathrm{Tr}(\Sigma_V) = \|V\|_*.\end{aligned} \tag{8}$$

Using Lemma 2, we have

$$\begin{aligned}\mathrm{Tr}(L_U\Sigma_U^{-\frac{1}{2}}L_U^TW_2^T) &= \mathrm{Tr}(L_U\Sigma_U^{-\frac{1}{2}}L_U^TL_X\Sigma_U^{\frac{1}{2}}\Sigma_X^{\frac{1}{2}}L_X^T)\\ &= \mathrm{Tr}(\Sigma_U^{-\frac{1}{2}}L_U^TL_X\Sigma_U^{\frac{1}{2}}\Sigma_X^{\frac{1}{2}}L_X^TL_U)\\ &= \mathrm{Tr}(\Sigma_U^{-\frac{1}{2}}O_1\Sigma_U^{\frac{1}{2}}\Sigma_X^{\frac{1}{2}}O_1^T)\\ &\ge \mathrm{Tr}(\Sigma_U^{-\frac{1}{2}}\Sigma_U^{\frac{1}{2}}\Sigma_X^{\frac{1}{2}}) = \mathrm{Tr}(\Sigma_X^{\frac{1}{2}}) = \|X\|_{S_{1/2}}^{1/2},\end{aligned} \tag{9}$$

where $O_1 = L_U^TL_X \in \mathbb{R}^{d\times d}$, and it is easy to verify that $O_1\Sigma_U^{\frac{1}{2}}\Sigma_X^{\frac{1}{2}}O_1^T$ is a symmetric PSD matrix. Using (7), (8), (9) and $\mathrm{Tr}(W_1) = \|U\|_*$, we have $\|U\|_* + \|V\|_* \ge 2\|X\|_{S_{1/2}}^{1/2}$. □

The detailed proof of Theorem 1 is provided in the Supplementary Materials, available online. According to

Theorem 1, it is easy to verify that the double nuclear norm penalty possesses the following property [34].

Property 2. *Given a matrix $X \in \mathbb{R}^{m\times n}$ with $\operatorname{rank}(X) \le d$, the following equalities hold*

$$\|X\|_{\mathrm{D-N}} = \min_{U\in\mathbb{R}^{m\times d}, V\in\mathbb{R}^{n\times d}: X=UV^T} \left(\frac{\|U\|_* + \|V\|_*}{2}\right)^2 = \min_{U,V:X=UV^T} \|U\|_*\|V\|_*.$$

3.2 Frobenius/Nuclear Norm Penalty

Inspired by the definitions of the bilinear spectral and double nuclear norm penalties mentioned above, we define a Frobenius/nuclear hybrid norm (F-N) penalty as follows.

Definition 2. *For any matrix $X \in \mathbb{R}^{m\times n}$ of rank at most $r \le d$, we decompose it into two factor matrices $U \in \mathbb{R}^{m\times d}$ and $V \in \mathbb{R}^{n\times d}$ such that $X = UV^T$. Then the Frobenius/nuclear hybrid norm penalty of X is defined as*

$$\|X\|_{\mathrm{F-N}} = \min_{U,V:X=UV^T} \left[(\|U\|_F^2 + 2\|V\|_*)/3\right]^{3/2}. \tag{10}$$

Different from the definition in [34], i.e., $\min_{U,V:X=UV^T} \|U\|_F\|V\|_*$, Definition 2 can also be directly used in practical problems. Analogous to the double nuclear norm penalty, the Frobenius/nuclear hybrid norm penalty is also a quasi-norm, as stated in the following theorem.

Theorem 2. *The Frobenius/nuclear hybrid norm penalty, $\|\cdot\|_{\mathrm{F-N}}$, is a quasi-norm, and is also the Schatten-$2/3$ quasi-norm, i.e.,*

$$\|X\|_{\mathrm{F-N}} = \|X\|_{S_{2/3}}. \tag{11}$$

To prove Theorem 2, we first give the following lemma.

Lemma 4. *For any matrix $X = UV^T \in \mathbb{R}^{m\times n}$, $U \in \mathbb{R}^{m\times d}$ and $V \in \mathbb{R}^{n\times d}$, the following inequality holds*

$$\frac{1}{3}\left(\|U\|_F^2 + 2\|V\|_*\right) \ge \|X\|_{S_{2/3}}^{2/3}.$$

Proof. To prove this lemma, we use the same notations as in the proof of Lemma 3, e.g., $X = L_X\Sigma_X R_X^T$, $U = L_U\Sigma_U R_U^T$ and $V = L_V\Sigma_V R_V^T$ denote the thin SVDs of U and V, respectively. Suppose $\overline{W}_1 = R_X\Sigma_V R_X^T$, $\overline{W}_2 = R_X\Sigma_V^{\frac{1}{2}}\Sigma_X^{\frac{2}{3}}R_X^T$ and $\overline{W}_3 = R_X\Sigma_X^{\frac{4}{3}}R_X^T$, we first construct the following PSD matrices $\mathcal{M}_2 \in \mathbb{R}^{2m\times 2m}$ and $\mathcal{S}_2 \in \mathbb{R}^{2m\times 2m}$

$$\begin{aligned}
\mathcal{M}_2 &= \begin{bmatrix} -R_X\Sigma_V^{\frac{1}{2}} \\ R_X\Sigma_X^{\frac{2}{3}} \end{bmatrix}\begin{bmatrix} -\Sigma_V^{\frac{1}{2}}R_X^T & \Sigma_X^{\frac{2}{3}}R_X^T \end{bmatrix} \\
&= \begin{bmatrix} \overline{W}_1 & -\overline{W}_2 \\ -\overline{W}_2^T & \overline{W}_3 \end{bmatrix} \succeq 0, \\
\mathcal{S}_2 &= \begin{bmatrix} I_n \\ L_V\Sigma_V^{-\frac{1}{2}}L_V^T \end{bmatrix}\begin{bmatrix} I_n & L_V\Sigma_V^{-\frac{1}{2}}L_V^T \end{bmatrix} \\
&= \begin{bmatrix} I_n & L_V\Sigma_V^{-\frac{1}{2}}L_V^T \\ L_V\Sigma_V^{-\frac{1}{2}}L_V^T & L_V\Sigma_V^{-1}L_V^T \end{bmatrix} \succeq 0.
\end{aligned}$$

Similar to Lemma 3, we have the following inequality:

$$\operatorname{Tr}\left(\begin{bmatrix} I_n & L_V\Sigma_V^{-\frac{1}{2}}L_V^T \\ L_V\Sigma_V^{-\frac{1}{2}}L_V^T & L_V\Sigma_V^{-1}L_V^T \end{bmatrix}\begin{bmatrix} \overline{W}_1 & -\overline{W}_2 \\ -\overline{W}_2^T & \overline{W}_3 \end{bmatrix}\right) \ge 0.$$

By further simplifying the above expression, we also obtain

$$\operatorname{Tr}(\overline{W}_1) - 2\operatorname{Tr}(L_V\Sigma_V^{-\frac{1}{2}}L_V^T\overline{W}_2^T) + \operatorname{Tr}(L_V\Sigma_V^{-1}L_V^T\overline{W}_3) \ge 0. \tag{12}$$

Since $L_X\Sigma_X R_X^T = L_U\Sigma_U R_U^T R_V\Sigma_V L_V^T$, $L_V^T R_X\Sigma_X = (L_X^T L_U \Sigma_U R_U^T R_V\Sigma_V)^T = \Sigma_V R_V^T R_U\Sigma_U L_U^T L_X$. Due to the orthonormality of the columns of L_U, R_U, L_V, R_V, L_X and R_X, and using the von Neumann's trace inequality [71], [72], we have

$$\begin{aligned}
\operatorname{Tr}(L_V\Sigma_V^{-1}L_V^T\overline{W}_3) &= \operatorname{Tr}(L_V\Sigma_V^{-1}L_V^T R_X\Sigma_X\Sigma_X^{\frac{1}{3}}R_X^T) \\
&= \operatorname{Tr}(L_V\Sigma_V^{-1}\Sigma_V R_V^T R_U\Sigma_U L_U^T L_X\Sigma_X^{\frac{1}{3}}R_X^T) \\
&= \operatorname{Tr}(R_X^T L_V R_V^T R_U\Sigma_U L_U^T L_X\Sigma_X^{\frac{1}{3}}) \\
&\le \operatorname{Tr}(\Sigma_U\Sigma_X^{\frac{1}{3}}) \le \frac{1}{2}\left(\|U\|_F^2 + \|X\|_{S_{2/3}}^{2/3}\right).
\end{aligned} \tag{13}$$

By Lemma 2, we also have

$$\begin{aligned}
\operatorname{Tr}(L_V\Sigma_V^{-\frac{1}{2}}L_V^T\overline{W}_2^T) &= \operatorname{Tr}(L_V\Sigma_V^{-\frac{1}{2}}L_V^T R_X\Sigma_V^{\frac{1}{2}}\Sigma_X^{\frac{2}{3}}R_X^T) \\
&= \operatorname{Tr}(\Sigma_V^{-\frac{1}{2}}L_V^T R_X\Sigma_V^{\frac{1}{2}}\Sigma_X^{\frac{2}{3}}R_X^T L_V) \\
&= \operatorname{Tr}(\Sigma_V^{-\frac{1}{2}}O_2\Sigma_V^{\frac{1}{2}}\Sigma_X^{\frac{2}{3}}O_2^T) \\
&\ge \operatorname{Tr}(\Sigma_V^{-\frac{1}{2}}\Sigma_V^{\frac{1}{2}}\Sigma_X^{\frac{2}{3}}) = \operatorname{Tr}(\Sigma_X^{\frac{2}{3}}) = \|X\|_{S_{2/3}}^{2/3},
\end{aligned} \tag{14}$$

where $O_2 = L_V^T R_X \in \mathbb{R}^{d\times d}$, and it is easy to verify that $O_2\Sigma_V^{\frac{1}{2}}\Sigma_X^{\frac{2}{3}}O_2^T$ is a symmetric PSD matrix. Using (12), (13), (14), and $\operatorname{Tr}(\overline{W}_1) = \|V\|_*$, then we have $\|U\|_F^2 + 2\|V\|_* \ge 3\|X\|_{S_{2/3}}^{2/3}$. □

According to Theorem 2 (see the Supplementary Materials, available online, for its detailed proof), it is easy to verify that the Frobenius/nuclear hybrid norm penalty possesses the following property [34].

Property 3. *For any matrix $X \in \mathbb{R}^{m\times n}$ with $\operatorname{rank}(X) = r \le d$, the following equalities hold*

$$\|X\|_{\mathrm{F-N}} = \min_{U\in\mathbb{R}^{m\times d}, V\in\mathbb{R}^{n\times d}: X=UV^T} \left(\frac{\|U\|_F^2 + 2\|V\|_*}{3}\right)^{3/2} = \min_{U,V:X=UV^T} \|U\|_F\|V\|_*.$$

Similar to the relationship between the Frobenius norm and nuclear norm, i.e., $\|X\|_F \le \|X\|_* \le \operatorname{rank}(X)\|X\|_F$ [26], the bounds hold for between both the double nuclear norm and Frobenius/nuclear hybrid norm penalties and the nuclear norm, as stated in the following property.

Property 4. *For any matrix $X \in \mathbb{R}^{m\times n}$, the following inequalities hold*

$$\|X\|_* \le \|X\|_{\mathrm{F-N}} \le \|X\|_{\mathrm{D-N}} \le \operatorname{rank}(X)\|X\|_*.$$

The proof of Property 4 is similar to that in [34]. Moreover, both the double nuclear norm and Frobenius/nuclear hybrid norm penalties naturally satisfy many properties of quasi-norms, e.g., the unitary-invariant property. Obviously, we can find that Property 4 in turn implies that any low F-N or D-N penalty is also a low nuclear norm approximation.

3.3 Problem Formulations

Without loss of generality, we can assume that the unknown entries of D are set to zero (i.e., $\mathcal{P}_{\Omega^c}(D)=\mathbf{0}$), and S_{Ω^c} may be any values[2] (i.e., $S_{\Omega^c}\in(-\infty,+\infty)$) such that $\mathcal{P}_{\Omega^c}(L+S)=\mathcal{P}_{\Omega^c}(D)$. Thus, the constraint with the projection operator $\mathcal{P}_\Omega$ in (2) is considered instead of just $L+S=D$ as in [42]. Together with hyper-Laplacian priors of sparse components, we use $\|L\|_{\text{D-N}}^{1/2}$ and $\|L\|_{\text{F-N}}^{2/3}$ defined above to replace $\|L\|_{S_q}^q$ in (2), and present the following double nuclear norm and Frobenius/nuclear hybrid norm penalized RPCA models:

$$\begin{aligned}&\min_{U,V,L,S}\frac{\lambda}{2}(\|U\|_*+\|V\|_*)+\|\mathcal{P}_\Omega(S)\|_{\ell_{1/2}}^{1/2},\\&\text{s.t., } UV^T=L, L+S=D\end{aligned}\tag{15}$$

$$\begin{aligned}&\min_{U,V,L,S}\frac{\lambda}{3}\left(\|U\|_F^2+2\|V\|_*\right)+\|\mathcal{P}_\Omega(S)\|_{\ell_{2/3}}^{2/3},\\&\text{s.t., } UV^T=L, L+S=D.\end{aligned}\tag{16}$$

From the two proposed models (15) and (16), one can easily see that the norm of each bilinear factor matrix is convex, and they are much more tractable and scalable optimization problems than the original Schatten quasi-norm minimization problem as in (2).

4 Optimization Algorithms

To efficiently solve both our challenging problems (15) and (16), we need to introduce the auxiliary variables $\widehat{U}$ and $\widehat{V}$, or only $\widehat{V}$ to split the interdependent terms such that they can be solved independently. Thus, we can reformulate Problems (15) and (16) into the following equivalent forms

$$\begin{aligned}&\min_{U,V,L,S,\widehat{U},\widehat{V}}\frac{\lambda}{2}\left(\|\widehat{U}\|_*+\|\widehat{V}\|_*\right)+\|\mathcal{P}_\Omega(S)\|_{\ell_{1/2}}^{1/2},\\&\text{s.t., } \widehat{U}=U,\ \widehat{V}=V,\ UV^T=L,\ L+S=D\end{aligned}\tag{17}$$

$$\begin{aligned}&\min_{U,V,L,S,\widehat{V}}\frac{\lambda}{3}\left(\|U\|_F^2+2\|\widehat{V}\|_*\right)+\|\mathcal{P}_\Omega(S)\|_{\ell_{2/3}}^{2/3},\\&\text{s.t., } \widehat{V}=V,\ UV^T=L,\ L+S=D.\end{aligned}\tag{18}$$

4.1 Solving (17) via ADMM

Inspired by recent progress on alternating direction methods [44], [58], we mainly propose an efficient algorithm based on the alternating direction method of multipliers [58] (ADMM, also known as the inexact ALM [44]) to solve the more complex problem (17), whose augmented Lagrangian function is given by

2. Considering the optimal solution $S_{\Omega^c}=\mathbf{0}$, S_{Ω^c} must be set to 0 for the expected output S.

$$\begin{aligned}\mathcal{L}_\mu(U,V,L,S,\widehat{U},\widehat{V},\{Y_i\})&=\frac{\lambda}{2}(\|\widehat{U}\|_*+\|\widehat{V}\|_*)+\|\mathcal{P}_\Omega(S)\|_{\ell_{\frac{1}{2}}}^{\frac{1}{2}}\\&+\langle Y_1,\widehat{U}-U\rangle+\langle Y_2,\widehat{V}-V\rangle+\langle Y_3,UV^T-L\rangle+\langle Y_4,L+S-D\rangle\\&+\frac{\mu}{2}\left(\|\widehat{U}-U\|_F^2+\|\widehat{V}-V\|_F^2+\|UV^T-L\|_F^2+\|L+S-D\|_F^2\right),\end{aligned}$$

where $\mu>0$ is the penalty parameter, $\langle\cdot,\cdot\rangle$ represents the inner product operator, and $Y_1\in\mathbb{R}^{m\times d}$, $Y_2\in\mathbb{R}^{n\times d}$ and $Y_3,Y_4\in\mathbb{R}^{m\times n}$ are Lagrange multipliers.

4.1.1 *Updating* U_{k+1} *and* V_{k+1}

To update U_{k+1} and V_{k+1}, we consider the following optimization problems

$$\min_U\|\widehat{U}_k-U+\mu_k^{-1}Y_1^k\|_F^2+\|UV_k^T-L_k+\mu_k^{-1}Y_3^k\|_F^2,\tag{19}$$

$$\min_V\|\widehat{V}_k-V+\mu_k^{-1}Y_2^k\|_F^2+\|U_{k+1}V^T-L_k+\mu_k^{-1}Y_3^k\|_F^2.\tag{20}$$

Both (19) and (20) are least squares problems, and their optimal solutions are given by

$$U_{k+1}=(\widehat{U}_k+\mu_k^{-1}Y_1^k+M_kV_k)(I_d+V_k^TV_k)^{-1},\tag{21}$$

$$V_{k+1}=(\widehat{V}_k+\mu_k^{-1}Y_2^k+M_k^TU_{k+1})(I_d+U_{k+1}^TU_{k+1})^{-1},\tag{22}$$

where $M_k=L_k-\mu_k^{-1}Y_3^k$, and I_d denotes an identity matrix of size $d\times d$.

Algorithm 1. ADMM for Solving $(\text{S+L})_{1/2}$ Problem (17)

Input: $\mathcal{P}_\Omega(D)\in\mathbb{R}^{m\times n}$, the given rank d, and λ.
Initialize: μ_0, $\rho>1$, $k=0$, and ϵ.
1: **while** not converged **do**
2: **while** not converged **do**
3: Update U_{k+1} and V_{k+1} by (21) and (22).
4: Compute $\widehat{U}_{k+1}$ and $\widehat{V}_{k+1}$ via the SVT operator [23].
5: Update L_{k+1} and S_{k+1} by (26) and (29).
6: **end while**// Inner loop
7: Update the multipliers by
$Y_1^{k+1}=Y_1^k+\mu_k(\widehat{U}_{k+1}-U_{k+1})$,
$Y_2^{k+1}=Y_2^k+\mu_k(\widehat{V}_{k+1}-V_{k+1})$,
$Y_3^{k+1}=Y_3^k+\mu_k(U_{k+1}V_{k+1}^T-L_{k+1})$,
$Y_4^{k+1}=Y_4^k+\mu_k(L_{k+1}+S_{k+1}-D)$.
8: Update μ_{k+1} by $\mu_{k+1}=\rho\mu_k$.
9: $k\leftarrow k+1$.
10: **end while**// Outer loop
Output U_{k+1} and V_{k+1}.

4.1.2 *Updating* $\widehat{U}_{k+1}$ *and* $\widehat{V}_{k+1}$

To solve $\widehat{U}_{k+1}$ and $\widehat{V}_{k+1}$, we fix the other variables and solve the following optimization problems

$$\min_{\widehat{U}}\frac{\lambda}{2}\|\widehat{U}\|_*+\frac{\mu_k}{2}\|\widehat{U}-U_{k+1}+Y_1^k/\mu_k\|_F^2,\tag{23}$$

$$\min_{\widehat{V}}\frac{\lambda}{2}\|\widehat{V}\|_*+\frac{\mu_k}{2}\|\widehat{V}-V_{k+1}+Y_2^k/\mu_k\|_F^2.\tag{24}$$

Both (23) and (24) are nuclear norm regularized least squares problems, and their closed-form solutions can be given by the so-called SVT operator [23], respectively.

4.1.3 Updating L_{k+1}

To update L_{k+1}, we can obtain the following optimization problem

$$\min_{L} \|U_{k+1}V_{k+1}^T - L + \mu_k^{-1}Y_3^k\|_F^2 + \|L + S_k - D + \mu_k^{-1}Y_4^k\|_F^2. \quad (25)$$

Since (25) is a least squares problem, and thus its closed-form solution is given by

$$L_{k+1} = \frac{1}{2}\left(U_{k+1}V_{k+1}^T + \mu_k^{-1}Y_3^k - S_k + D - \mu_k^{-1}Y_4^k\right). \quad (26)$$

4.1.4 Updating S_{k+1}

By keeping all other variables fixed, S_{k+1} can be updated by solving the following problem

$$\min_{S} \|\mathcal{P}_\Omega(S)\|_{\ell_{1/2}}^{1/2} + \frac{\mu_k}{2}\|S + L_{k+1} - D + \mu_k^{-1}Y_4^k\|_F^2. \quad (27)$$

Generally, the ℓ_p-norm ($0 < p < 1$) leads to a non-convex, non-smooth, and non-Lipschitz optimization problem [39]. Fortunately, we can efficiently solve (27) by introducing the following half-thresholding operator [21].

Proposition 1. *For any matrix $A \in \mathbb{R}^{m\times n}$, and $X^* \in \mathbb{R}^{m\times n}$ is an $\ell_{1/2}$ quasi-norm solution of the following minimization*

$$\min_{X} \|X - A\|_F^2 + \gamma\|X\|_{\ell_{1/2}}^{1/2}, \quad (28)$$

then the solution X^* can be given by $X^* = \mathcal{H}_\gamma(A)$, where the half-thresholding operator $\mathcal{H}_\gamma(\cdot)$ is defined as

$$\mathcal{H}_\gamma(a_{ij}) = \begin{cases} \frac{2}{3}a_{ij}\left[1 + \cos\left(\frac{2\pi - 2\phi_\gamma(a_{ij})}{3}\right)\right], & |a_{ij}| > \frac{\sqrt[3]{54\gamma^2}}{4}, \\ 0, & \text{otherwise,} \end{cases}$$

where $\phi_\gamma(a_{ij}) = \arccos(\frac{\gamma}{8}(|a_{ij}|/3)^{-3/2})$.

Before giving the proof of Proposition 1, we first give the following lemma [21].

Lemma 5. *Let $y \in \mathbb{R}^{l\times 1}$ be a given vector, and $\tau > 0$. Suppose that $x^* \in \mathbb{R}^{l\times 1}$ is a solution of the following problem,*

$$\min_{x} \|Bx - y\|_2^2 + \tau\|x\|_{\ell_{1/2}}^{1/2}.$$

Then for any real parameter $\mu \in (0, \infty)$, x^* can be expressed as $x^* = \mathcal{H}_{\tau\mu}(\varphi_\mu(x^*))$, where $\varphi_\mu(x^*) = x^* + \mu B^T(y - Bx^*)$.

Proof. The formulation (28) can be reformulated as the following equivalent form

$$\min_{\mathrm{vec}(X)} \|\mathrm{vec}(X) - \mathrm{vec}(A)\|_2^2 + \gamma\|\mathrm{vec}(X)\|_{\ell_{1/2}}^{1/2}.$$

Let $B = I_{mn}$, $\mu = 1$, and using Lemma 5, the closed-form solution of (28) is given by $\mathrm{vec}(X^*) = \mathcal{H}_\gamma(\mathrm{vec}(A))$. □

Using Proposition 1, the closed-form solution of (27) is

$$\begin{aligned} S_{k+1} =& \mathcal{P}_\Omega\left(\mathcal{H}_{2/\mu_k}(D - L_{k+1} - \mu_k^{-1}Y_4^k)\right) \\ &+ \mathcal{P}_\Omega^\perp\left(D - L_{k+1} - \mu_k^{-1}Y_4^k\right) \end{aligned} \quad (29)$$

where $\mathcal{P}_\Omega^\perp$ is the complementary operator of $\mathcal{P}_\Omega$. Alternatively, Zuo et al. [22] proposed a generalized shrinkage-thresholding operator to iteratively solve ℓ_p-norm minimization with arbitrary p values, i.e., $0 \leq p < 1$, and achieve a higher efficiency.

Based on the description above, we develop an efficient ADMM algorithm to solve the double nuclear norm penalized problem (17), as outlined in **Algorithm** 1. To further accelerate the convergence of the algorithm, the penalty parameter μ, as well as ρ, are adaptively updated by the strategy as in [44]. The varying μ, together with shrinkage-thresholding operators such as the SVT operator, sometimes play the role of adaptive selection on the rank of matrices or the number of non-zeros elements. We found that updating $\{U_k, V_k\}$, $\{\widehat{U}_k, \widehat{V}_k\}$, L_k and S_k just once in the inner loop is sufficient to generate a satisfying accurate solution of (17), so also called inexact ALM, which is used for computational efficiency. In addition, we initialize all the elements of the Lagrange multipliers Y_1, Y_2 and Y_3 to 0, while all elements in Y_4 are initialized by the same way as in [44].

4.2 Solving (18) via ADMM

Similar to Algorithm 1, we also propose an efficient ADMM algorithm to solve (18) (i.e., Algorithm 2), and provide the details in the Supplementary Materials, available online. Since the update schemes of U_k, V_k, $\widehat{V}_k$ and L_k are very similar to that of Algorithm 1, we discuss their major differences below.

By keeping all other variables fixed, S_{k+1} can be updated by solving the following problem

$$\min_{S} \left\|\mathcal{P}_\Omega(S)\right\|_{\ell_{2/3}}^{2/3} + \frac{\mu_k}{2}\left\|S + L_{k+1} - D + \mu_k^{-1}Y_3^k\right\|_F^2. \quad (30)$$

Inspired by [10], [22], [74], we introduce the following two-thirds-thresholding operator to efficiently solve (30).

Proposition 2. *For any matrix $C \in \mathbb{R}^{m\times n}$, and $X^* \in \mathbb{R}^{m\times n}$ is an $\ell_{2/3}$ quasi-norm solution of the following minimization*

$$\min_X \|X - C\|_F^2 + \gamma\|X\|_{\ell_{2/3}}^{2/3}, \quad (31)$$

then the solution X^* can be given by $X^* = \mathcal{T}_\gamma(C)$, where the two-thirds-thresholding operator $\mathcal{T}_\gamma(\cdot)$ is defined as

$$\mathcal{T}_\gamma(c_{ij}) = \begin{cases} \frac{\mathrm{sgn}(c_{ij})\left(\psi_\gamma(c_{ij}) + \sqrt{\frac{2|c_{ij}|}{\psi_\gamma(c_{ij})} - \psi_\gamma^2(c_{ij})}\right)^3}{8}, & |c_{ij}| > \frac{2\sqrt[4]{3\gamma^3}}{3}, \\ 0, & \text{otherwise,} \end{cases}$$

where $\psi_\gamma(c_{ij}) = \frac{2}{\sqrt{3}}\sqrt{\sqrt{\gamma}\cosh(\mathrm{arccosh}(\frac{27c_{ij}^2}{16}\gamma^{-3/2})/3)}$, and $\mathrm{sgn}(\cdot)$ is the sign function.

Before giving the proof of Proposition 2, we first give the following lemma [22], [74].

Lemma 6. *Let $y \in \mathbb{R}$ be a given real number, and $\tau > 0$. Suppose that $x^* \in \mathbb{R}$ is a solution of the following problem,*

$$\min_{x} x^2 - 2xy + \tau|x|^{2/3}. \quad (32)$$

Then x^* has the following closed-form thresholding formula

$$x^* = \begin{cases} \frac{\mathrm{sgn}(y)\left(\psi_\tau(y) + \sqrt{\frac{2|y|}{\psi_\tau(y)} - \psi_\tau^2(y)}\right)^3}{8}, & |y| > \frac{2\sqrt[4]{3\tau^3}}{3}, \\ 0, & \text{otherwise.} \end{cases}$$

Proof. It is clear that the operator in Lemma 6 can be extended to vectors and matrices by applying it element-wise. Using Lemma 6, the closed-form thresholding formula of (31) is given by $X^* = \mathcal{T}_\gamma(C)$. □

By Proposition 2, the closed-form solution of (30) is

$$S_{k+1} = \mathcal{P}_{\Omega}\big(\mathcal{T}_{2/\mu_k}(D - L_{k+1} - \mu_k^{-1} Y_3^k)\big) + \mathcal{P}_{\Omega}^{\perp}\big(D - L_{k+1} - \mu_k^{-1} Y_3^k\big). \tag{33}$$

5 Algorithm Analysis

We mainly analyze the convergence property of Algorithm 1. Naturally, the convergence of Algorithm 2 can also be guaranteed in a similar way. Moreover, we also analyze their per-iteration complexity.

5.1 Convergence Analysis

Before analyzing the convergence of Algorithm 1, we first introduce the definition of the critical points (or stationary points) of a non-convex function given in [75].

Definition 3. *Given a function $f : \mathbb{R}^n \to (-\infty, +\infty]$, we denote the domain of f by $\mathrm{dom} f$, i.e., $\mathrm{dom} f := \{x \in \mathbb{R}^n : f(x) \leq +\infty\}$. A function f is said to be proper if $\mathrm{dom} f \neq \emptyset$; lower semicontinuous at the point x_0 if $\lim_{x \to x_0} \inf f(x) \geq f(x_0)$. If f is lower semicontinuous at every point of its domain, then it is called a lower semicontinuous function.*

Definition 4. *Let a non-convex function $f : \mathbb{R}^n \to (-\infty, +\infty]$ be a proper and lower semi-continuous function. x is a critical point of f if $0 \in \partial f(x)$, where $\partial f(x)$ is the limiting sub-differential of f at x, i.e., $\partial f(x) = \{u \in \mathbb{R}^n : \exists x^k \to x, f(x^k) \to f(x) \text{ and } u^k \in \widehat{\partial} f(x^k) \to u \text{ as } k \to \infty\}$, and $\widehat{\partial} f(x^k)$ is the Frèchet sub-differential of f at x (see [75] for more details).*

As stated in [45], [76], the general convergence property of the ADMM for non-convex problems has not been answered yet, especially for multi-block cases [76], [77]. For such challenging problems (15) and (16), although it is difficult to guarantee the convergence to a local minimum, our empirical convergence tests showed that our algorithms have strong convergence behavior (see the Supplementary Materials, available online, for details). Besides the empirical behavior, we also provide the convergence property for Algorithm 1 in the following theorem.

Theorem 3. *Let $\{(U_k, V_k, \widehat{U}_k, \widehat{V}_k, L_k, S_k, \{Y_i^k\})\}$ be the sequence generated by Algorithm 1. Suppose that the sequence $\{Y_3^k\}$ is bounded, and μ_k is non-decreasing and $\sum_{k=0}^{\infty}(\mu_{k+1}/\mu_k^{4/3}) < \infty$, then*

(I) $\{(U_k, V_k)\}$, $\{(\widehat{U}_k, \widehat{V}_k)\}$, $\{L_k\}$ and $\{S_k\}$ are all Cauchy sequences;

(II) Any accumulation point of the sequence $\{(U_k, V_k, \widehat{U}_k, \widehat{V}_k, L_k, S_k)\}$ satisfies the Karush-Kuhn-Tucker (KKT) conditions for Problem (17).

The proof of Theorem 3 can be found in the Supplementary Materials, available online. Theorem 3 shows that under mild conditions, any accumulation point (or limit point) of the sequence generated by Algorithm 1 is a critical point of the Lagrangian function $\mathcal{L}_{\mu}$, i.e., $(U_\star, V_\star, \widehat{U}_\star, \widehat{V}_\star, L_\star, S_\star, \{Y_i^\star\})$, which satisfies the first-order optimality conditions (i.e., the KKT conditions) of (17): $0 \in \frac{\lambda}{2}\partial\|\widehat{U}_\star\|_* + Y_1^\star$, $0 \in \frac{\lambda}{2}\partial\|\widehat{V}_\star\|_* + Y_2^\star$, $0 \in \partial\Phi(S_\star) + \mathcal{P}_{\Omega}(Y_4^\star)$, $\mathcal{P}_{\Omega^c}(Y_4^\star) = \mathbf{0}$, $L_\star = U_\star V_\star^T$, $\widehat{U}_\star = U_\star$, $\widehat{V}_\star = V_\star$, and $L_\star + S_\star = D$, where $\Phi(S) := \|\mathcal{P}_{\Omega}(S)\|_{\ell_{1/2}}^{1/2}$. Similarly, the convergence of Algorithm 2 can also be guaranteed. Theorem 3 is established for the proposed ADMM algorithm, which has only a single iteration in the inner loop. When the inner-loop iterations of Algorithm 1 iterate until convergence, it may lead to a simpler proof. We leave further theoretical analysis of convergence as future work.

Theorem 3 also shows that our ADMM algorithms have much weaker convergence conditions than the ones in [15], [45], e.g., the sequence of only one Lagrange multiplier is required to be bounded for our algorithms, while the ADMM algorithms in [15], [45] require the sequences of all Lagrange multipliers to be bounded.

5.2 Convergence Behavior

According to the KKT conditions of (17) mentioned above, we take the following conditions as the stopping criterion for our algorithms (see details in Supplementary Materials, available online),

$$\max\{\epsilon_1/\|D\|_F,\, \epsilon_2\} < \epsilon,$$

where $\epsilon_1 = \max\{\|U_k V_k^T - L_k\|_F, \|L_k + S_k - D\|_F, \|Y_1^k(\widehat{V}_k)^\dagger - (\widehat{U}_k^T)^\dagger (Y_2^k)^T\|_F\}$, $\epsilon_2 = \max\{\|\widehat{U}_k - U_k\|_F/\|U_k\|_F, \|\widehat{V}_k - V_k\|_F/\|V_k\|_F\}$, $(\widehat{V}_k)^\dagger$ is the pseudo-inverse of $\widehat{V}_k$, and ε is the stopping tolerance. In this paper, we set the stopping tolerance to $\epsilon = 10^{-5}$ for synthetic data and $\epsilon = 10^{-4}$ for real-world problems. As shown in the Supplementary Materials, available online, the stopping tolerance and relative squared error (RSE) of our methods decrease fast, and they converge within only a small number of iterations (usually within 50 iterations).

5.3 Computational Complexity

The per-iteration cost of existing Schatten quasi-norm minimization methods such as LpSq [31] is dominated by the computation of the thin SVD of an $m \times n$ matrix with $m \geq n$, and is $\mathcal{O}(mn^2)$. In contrast, the time complexity of computing SVD for (23) and (24) is $\mathcal{O}(md^2 + nd^2)$. The dominant cost of each iteration in Algorithm 1 corresponds to the matrix multiplications in the update of U, V and L, which take $\mathcal{O}(mnd + d^3)$. Given that $d \ll m, n$, the overall complexity of Algorithm 1, as well as Algorithm 2, is thus $\mathcal{O}(mnd)$, which is the same as the complexity of LMaFit [47], RegL1 [16], ROSL [64], Unifying [49], and factEN [15].

6 Experimental Results

In this section, we evaluate both the effectiveness and efficiency of our methods (i.e., $(\mathrm{S{+}L})_{1/2}$ and $(\mathrm{S{+}L})_{2/3}$) for solving extensive synthetic and real-world problems. We also compare our methods with several state-of-the-art methods, such as LMaFit[3] [47], RegL1[4] [16], Unifying [49], factEN[5] [15], RPCA[6] [44], PSVT[7] [45], WNNM[8] [46], and LpSq[9] [31].

6.1 Rank Estimation

As suggested in [6], [28], the regularization parameter λ of our two methods is generally set to $\sqrt{\max(m, n)}$. Analogous

3. http://lmafit.blogs.rice.edu/
4. https://sites.google.com/site/yinqiangzheng/
5. http://cpslab.snu.ac.kr/people/eunwoo-kim
6. http://www.cis.pku.edu.cn/faculty/vision/zlin/zlin.htm
7. http://thohkaistackr.wix.com/page
8. http://www4.comp.polyu.edu.hk/~cslzhang/
9. https://sites.google.com/site/feipingnie/

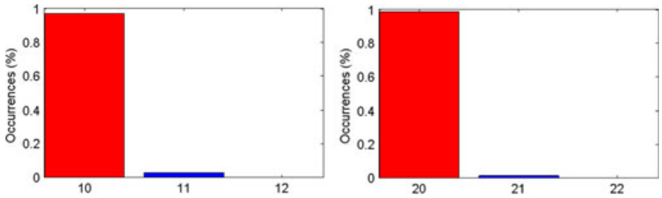


Fig. 3. Histograms with the percentage of estimated ranks for Gaussian noise and outlier corrupted matrices of size 500×500 (left) and $1{,}000 \times 1{,}000$ (right), whose true ranks are 10 and 20, respectively.

to other matrix factorization methods [15], [16], [47], [49], two proposed methods also have another important rank parameter, d. To estimate it, we design a simple rank estimation procedure. Since the observed data may be corrupted by noise/outlier and/or missing data, our rank estimation procedure combines two key techniques. First, we efficiently compute the k largest singular values of the input matrix (usually $k = 100$), and then use the basic spectral gap technique for determining the number of clusters [78]. Moreover, the rank estimator for incomplete matrices is exploited to look for an index for which the ratio between two consecutive singular values is minimized, as suggested in [79]. We conduct some experiments on corrupted matrices to test the performance of our rank estimation procedure, as shown in Fig. 3. Note that the input matrices are corrupted by both sparse outliers and Gaussian noise as shown below, where the fraction of sparse outliers varies from 0 to 25 percent, and the noise factor of Gaussian noise is changed from 0 to 0.5. It can be seen that our rank estimation procedure performs well in terms of robustness to noise and outliers.

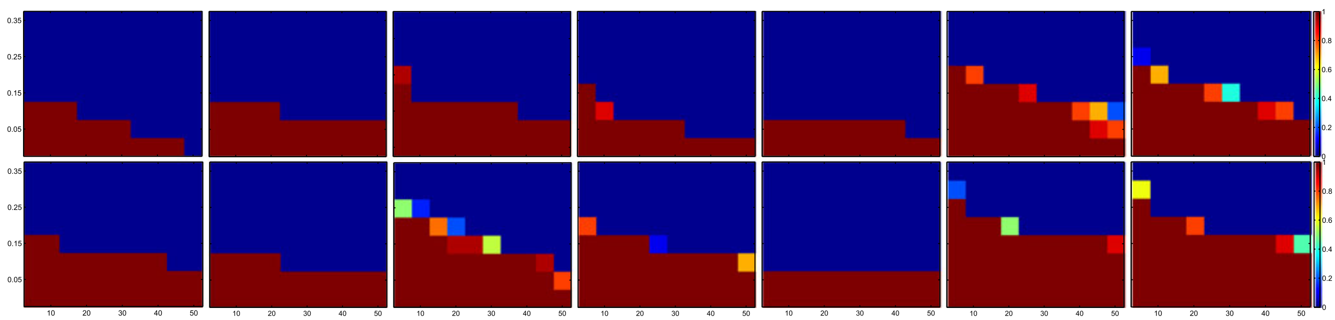
6.2 Synthetic Data

We generated the low-rank matrix $L^* \in \mathbb{R}^{m \times n}$ of rank r as the product PQ^T, where $P \in \mathbb{R}^{m \times r}$ and $Q \in \mathbb{R}^{n \times r}$ are independent matrices whose elements are independent and identically distributed (i.i.d.) random variables sampled from standard Gaussian distributions. The sparse matrix $S^* \in \mathbb{R}^{m \times n}$ is generated by the following procedure: its support is chosen uniformly at random and the non-zero entries are i.i.d. random variables sampled uniformly in the interval $[-5, 5]$. The input matrix is $D = L^* + S^* + N$, where the Gaussian noise is $N = nf \times \textbf{randn}$ and $nf \geq 0$ is the noise factor. For quantitative evaluation, we measured the performance of low-rank component recovery by the RSE, and evaluated the accuracy of outlier detection by the F-measure (abbreviated to F-M) as in [17]. The higher F-M or lower RSE, the better is the quality of the recovered results.

6.2.1 Model Comparison

We first compared our methods with RPCA (nuclear norm & ℓ_1-norm), PSVT (truncated nuclear norm & ℓ_1-norm), WNNM (weighted nuclear norm & ℓ_1-norm), LpSq (Schatten q-norm & ℓ_p-norm), and Unifying (bilinear spectral penalty & ℓ_1-norm), where p and q in LpSq are chosen from the range of $\{0.1, 0.2, \ldots, 1\}$. A phase transition plot uses magnitude to depict how likely a certain kind of low-rank matrices can be recovered by those methods for a range of different matrix ranks and corruption ratios. If the recovered matrix L has a RSE smaller than 10^{-2}, we consider the estimation of both L and S is regarded as successful. Fig. 4 shows the phase transition results of RPCA, PSVT, WNNM, LpSq, Unifying and both our methods on outlier corrupted matrices of size 200×200 and 500×500, where the corruption ratios varied from 0 to 0.35 with increment 0.05, and the true rank r from 5 to 50 with increment 5. Note that the rank parameter of PSVT, Unifying, and both our methods is set to $d = \lfloor 1.25r \rfloor$ as suggested in [34], [47]. The results show that both our methods perform significantly better than the other methods, which justifies the effectiveness of the proposed RPCA models (15) and (16).

To verify the robustness of our methods, the observed matrices are corrupted by both Gaussian noise and outliers, where the noise factor and outlier ratio are set to $nf = 0.5$ and 20 percent. The average results (including RSE, F-M, and running time) of 10 independent runs on corrupted matrices with different sizes are reported in Table 3. Note that the rank parameter d of PSVT, Unifying, and both our methods is computed by our rank estimation procedure. It is clear that both our methods significantly outperform all the other methods in terms of both RSE and F-M in all settings. Those non-convex methods including PSVT, WNNM, LpSq, Unifying, and both our methods consistently perform better than the convex method, RPCA, in terms of both RSE and F-M. Impressively, both our methods are much faster than the other methods, and at least 10 times faster than RPCA, PSVT, WNNM, and LpSq in the case when the size of matrices exceeds $5{,}000 \times 5{,}000$. This actually shows that our methods are more scalable,



(a) RPCA [44] (b) PSVT [45] (c) WNNM [46] (d) Unifying [49] (e) LpSq [31] (f) $(S+L)_{1/2}$ (g) $(S+L)_{2/3}$

Fig. 4. Phase transition plots for different algorithms on corrupted matrices of size 200×200 (top) and 500×500 (bottom). X-axis denotes the matrix rank, and Y-axis indicates the corruption ratio. The color magnitude indicates the success ratio $[0, 1]$, and a larger red area means a better performance of the algorithm (best viewed in colors).

TABLE 3
Comparison of Average RSE, F-M and Time (Seconds) on Corrupted Matrices

	RPCA [44]			PSVT [45]			WNNM [46]			Unifying [49]			LpSq [31]			$(S+L)_{1/2}$			$(S+L)_{2/3}$		
$m=n$	RSE	F-M	Time	RSE	F-M	Time	RSE	F-M	Time	RSE	F-M	Time	RSE	F-M	Time	RSE	F-M	Time	RSE	F-M	Time
500	0.1265	0.8145	6.87	0.1157	0.8298	2.46	0.0581	0.8419	30.59	0.0570	0.8427	1.96	0.1173	0.8213	245.81	*0.0469*	*0.8469*	*1.70*	**0.0453**	**0.8474**	**1.35**
1,000	0.1138	0.8240	35.29	0.1107	0.8325	16.91	0.0448	0.8461	203.52	0.0443	0.8462	6.93	0.1107	0.8305	985.69	*0.0335*	*0.8495*	*6.65*	**0.0318**	**0.8498**	**5.89**
5,000	0.1029	0.8324	1,772.55	0.0980	0.8349	1,425.25	0.0315	0.8483	24,370.85	0.0313	0.8488	171.28	—	—	—	*0.0152*	*0.8520*	*134.15*	**0.0145**	**0.8521**	**128.11**
10,000	0.1002	0.8340	20,321.36	0.0969	0.8355	1,5437.60	—	—	—	0.0302	0.8489	657.41	—	—	—	*0.0109*	*0.8524*	*528.80*	**0.0104**	**0.8525**	**487.46**

We highlight the best results in **bold** *and the second best in* italic *for each of three performance metrics. Note that WNNM and LpSq could not yield experimental results on the largest problem within 24 hours.*

and have even greater advantage over existing methods for handling large matrices.

We also report the RSE results of both our methods on corrupted matrices of size 500×500 and $1{,}000\times 1{,}000$ with outlier ratio 15 percent, as shown in Fig. 5, where the true ranks of those matrices are 10 and 20, and the noise factor ranges from 0.1 to 0.5. In addition, we provide the best results of two baseline methods, WNNM and Unifying. Note that the parameter d of Unifying and both our methods is computed via our rank estimation procedure. From the results, we can observe that both our methods are much more robust against Gaussian noise than WNNM and Unifying, and have much greater advantage over them in cases when the noise level is relatively large, e.g., 0.5.

6.2.2 Comparisons with Other Methods

Figs. 6a and 6b show the average F-measure and RSE results of different matrix factorization based methods on $1{,}000\times 1{,}000$ matrices with different outliers ratios, where the noise factor is set to 0.2. For fair comparison, the rank parameter of all these methods is set to $d=\lfloor 1.25r\rfloor$ as in [34], [47]. In all cases, RegL1 [16], Unifying [49], factEN [15], and both our methods have significantly better performance than LMaFit [47], where the latter has no regularizers. This empirically verifies the importance of low-rank regularizers including our defined bilinear factor matrix norm penalties. Moreover, we report the average RSE results of these matrix factorization based methods with outlier ratio 5 percent and various missing ratios in Figs. 6c and 6d, in which we also present the results of LpSq. One can see that only with a very limited number of observations (e.g., 80 percent missing ratio), both our methods yield much more accurate solutions than the other methods including LpSq, while more observations are available, both LpSq and our methods significantly outperform the other methods in terms of RSE.

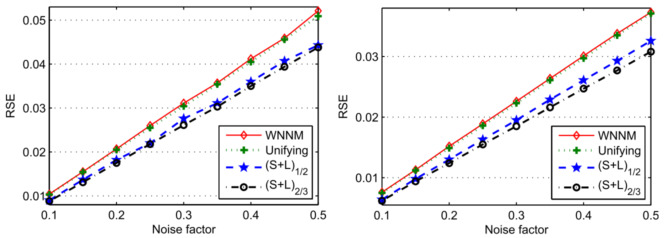
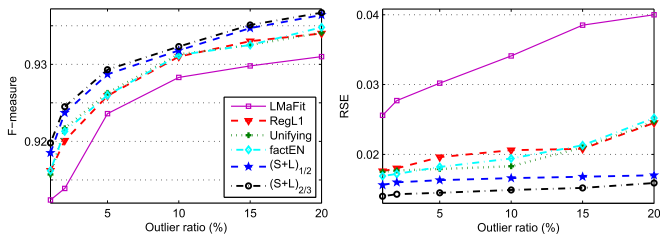


Fig. 5. Comparison of RSE results on corrupted matrices of size 500×500 (left) and $1{,}000\times 1{,}000$ (right) with different noise factors.

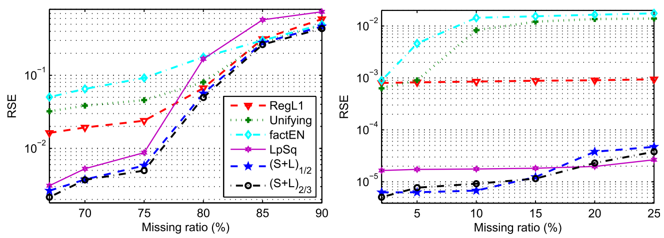
Finally, we report the performance of all those methods mentioned above on corrupted matrices of size $1{,}000\times 1{,}000$ as running time goes by, as shown in Fig. 7. It is clear that both our methods obtain significantly more accurate solutions than the other methods with much shorter running time. Different from all other methods, the performance of LMaFit becomes even worse over time, which may be caused by the intrinsic model without a regularizer. This also empirically verifies the importance of all low-rank regularizers, including our defined bilinear factor matrix norm penalties, for recovering low-rank matrices.

6.3 Real-World Applications

In this section, we apply our methods to solve various low-level vision problems, e.g., text removal, moving object detection, and image alignment and inpainting.



(a) F-measure vs. outlier ratio (b) RSE vs. outlier ratio



(c) RSE vs. high missing ratio (d) RSE vs. low missing ratio

Fig. 6. Comparison of different methods on corrupted matrices under varying outlier and missing ratios in terms of RSE and F-measure.

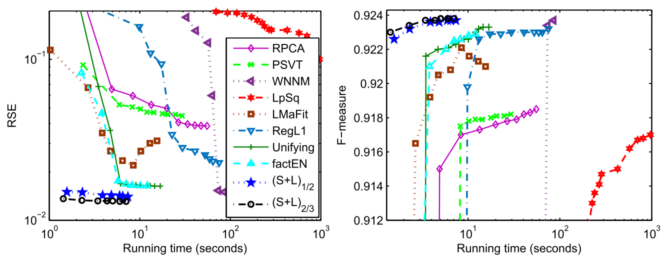
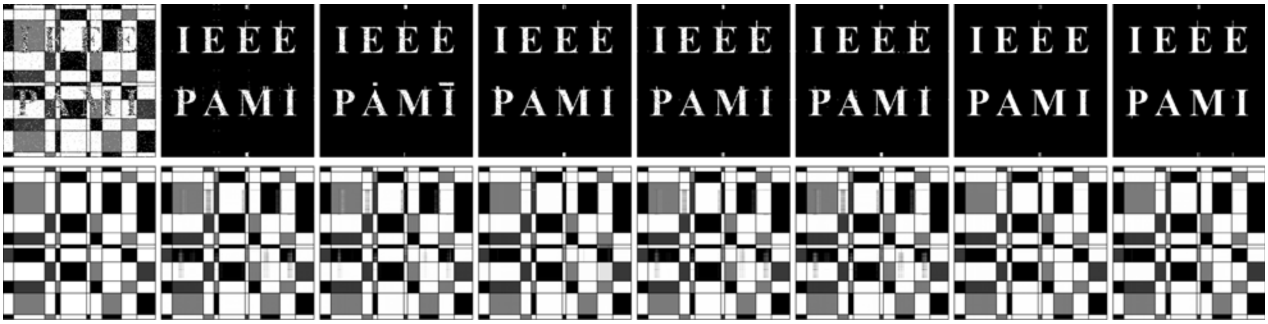
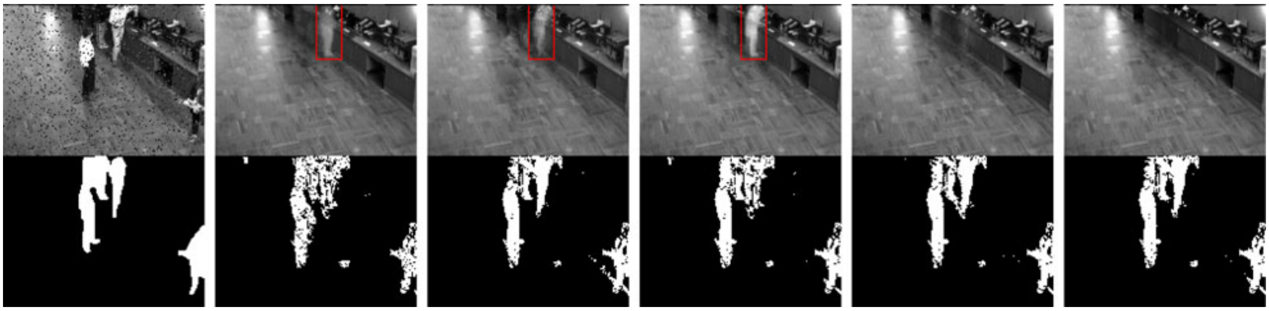


Fig. 7. Comparison of RSE (left) and F-measure (right) of different methods on corrupted matrices of size $1{,}000\times 1{,}000$ versus running time.



(a) Images (b) RPCA [44] (c) PSVT [45] (d) WNNM [46] (e) Unifying [49] (f) LpSq [31] (g) $(\mathrm{S+L})_{1/2}$ (h) $(\mathrm{S+L})_{2/3}$

Fig. 8. Text removal results of different methods: detected text masks (top) and recovered background images (bottom). (a) Input image (top) and original image (bottom); (b) RPCA (F-M: 0.9543, RSE: 0.1051); (c) PSVT (F-M: 0.9560, RSE: 0.1007); (d) WNNM (F-M: 0.9536, RSE: 0.0943); (e) Unifying (F-M: 0.9584, RSE: 0.0976); (f) LpSq (F-M: 0.9665, RSE: 0.1097); (g) $(\mathrm{S+L})_{1/2}$ (F-M: **0.9905**, RSE: **0.0396**); and (h) $(\mathrm{S+L})_{2/3}$ (F-M: *0.9872*, RSE: *0.0463*).



(a) Input,segmentation (b) RegL1 [16] (c) Unifying [49] (d) factEN [15] (e) $(\mathrm{S+L})_{1/2}$ (f) $(\mathrm{S+L})_{2/3}$

Fig. 9. Background and foreground separation results of different algorithms on the Bootstrap data set. The one frame with missing data of each sequence (top) and its manual segmentation (bottom) are shown in (a). The results of different algorithms are presented from (b) to (f), respectively. The top panel is the recovered background, and the bottom panel is the segmentation.

6.3.1 Text Removal

We first apply our methods to detect some text and remove them from the image used in [70]. The ground-truth image is of size 256×256 with rank equal to 10, as shown in Fig. 8a. The input data are generated by setting 5 percent of the randomly selected pixels as missing entries. For fairness, we set the rank parameter of PSVT [45], Unifying [49] and our methods to 20, and the stopping tolerance $\epsilon = 10^{-4}$ for all these algorithms. The text detection and removal results are shown in Fig. 8, where the text detection accuracy (F-M) and the RSE of recovered low-rank component are also reported. The results show that both our methods significantly outperform the other methods not only visually but also quantitatively. The running time of our methods and the Schatten quasi-norm minimization method, LpSq [31], is 1.36, 1.21 and 77.65 sec, respectively, which show that both our methods are more than 50 times faster than LpSq.

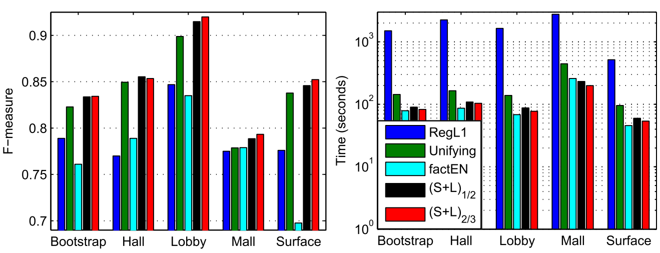


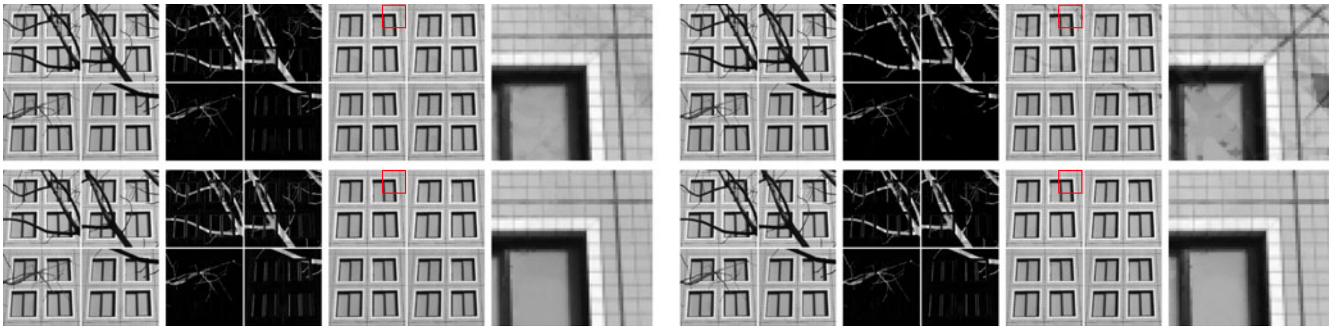
Fig. 10. Quantitative comparison for different methods in terms of F-measure (left) and running time (right, in seconds and in logarithmic scale) on five subsequences of surveillance videos.

6.3.2 Moving Object Detection

We test our methods on real surveillance videos for moving object detection and background subtraction as a RPCA plus matrix completion problem. Background modeling is a crucial task for motion segmentation in surveillance videos. A video sequence satisfies the low-rank and sparse structures, because the background of all the frames is controlled by few factors and hence exhibits low-rank property, and the foreground is detected by identifying spatially localized sparse residuals [6], [17], [40]. We test our methods on real surveillance videos for object detection and background subtraction on five surveillance videos: Bootstrap, Hall, Lobby, Mall and WaterSurface databases.[10] The input data are generated by setting 10 percent of the randomly selected pixels of each frame as missing entries, as shown in Fig. 9a.

Figs. 9b, 9c, 9d, 9e, and 9f show the foreground and background separation results on the Bootstrap data set. We can see that the background can be effectively extracted by RegL1, Unifying, factEN and both our methods, where their rank parameter is computed via our rank estimation procedure. It is clear that the decomposition results of both our methods are significantly better than that of RegL1 and factEN visually in terms of both background components and foreground segmentation. In addition, we also provide the running time and F-measure of different algorithms on all the five data

10. http://perception.i2r.a-star.edu.sg/



(a) (b) (c) (d) (e) (f) (g) (h)

Fig. 11. Alignment results on the Windows data set used in [50]. The first row: the results generated by RASL [50] (left) and PSVT [45] (right); The second row: the results generated by our $(\mathrm{S+L})_{1/2}$ (left) and $(\mathrm{S+L})_{2/3}$ (right) methods, respectively. (a) and (e): Aligned images; (b) and (f): Sparse occlusions; (c) and (g): Low-rank components; (d) and (h): Close-up (see red box in the left).

sets, as shown in Fig. 10, from which we can observe that both our methods consistently outperform the other methods in terms of F-measure. Moreover, Unifying and our methods are much faster than RegL1. Although factEN is slightly faster than Unifying and our methods, it usually has poorer quality of the results.

6.3.3 Image Alignment

We also study the performance of our methods in the application of robust image alignment: Given n images $\{I_1, \ldots, I_n\}$ of an object of interest, the image alignment task aims to align them using a fixed geometric transformation model, such as affine [50]. For this problem, we search for a transformation $\tau = \{\tau_1, \ldots, \tau_n\}$ and write $D \circ \tau = [\mathrm{vec}(I_1 \circ \tau_1)|\cdots|\mathrm{vec}(I_n \circ \tau_n)] \in \mathbb{R}^{m\times n}$. In order to robustly align the set of linearly correlated images despite sparse outliers, we consider the following double nuclear norm regularized model

$$\min_{U,V,S,\tau} \frac{\lambda}{2}(\|U\|_* + \|V\|_*) + \|S\|_{\ell_{1/2}}^{1/2}, \text{ s.t.}, D \circ \tau = UV^T + S. \quad (34)$$

Alternatively, our Frobenius/nuclear norm penalty can also be used to address the image alignment problem above.

We first test both our methods on the Windows data set (which contains 16 images of a building, taken from various viewpoints by a perspective camera, and with occlusions due to tree branches) used in [50] and report the aligned results of RASL[11] [50], PSVT [45] and our methods in Fig. 11, from which it is clear that, compared with RASL and PSVT, both our methods not only robustly align the images, correctly detect and remove the occlusion, but also achieve much better performance in terms of low-rank components, as shown in Figs. 11d and 11h, which give the close-up views of the red boxes in Figs. 11c and 11g, respectively.

6.3.4 Image Inpainting

Finally, we applied the defined D-N and F-N penalties to image inpainting. As shown by Hu et al. [38],[12] the images of natural scenes can be viewed as approximately low rank matrices. Naturally, we consider the following D-N penalty regularized least squares problem

$$\min_{U,V,L} \frac{\lambda}{2}(\|U\|_* + \|V\|_*) + \frac{1}{2}\|\mathcal{P}_\Omega(L - D)\|_F^2, \text{ s.t.}, L = UV^T. \quad (35)$$

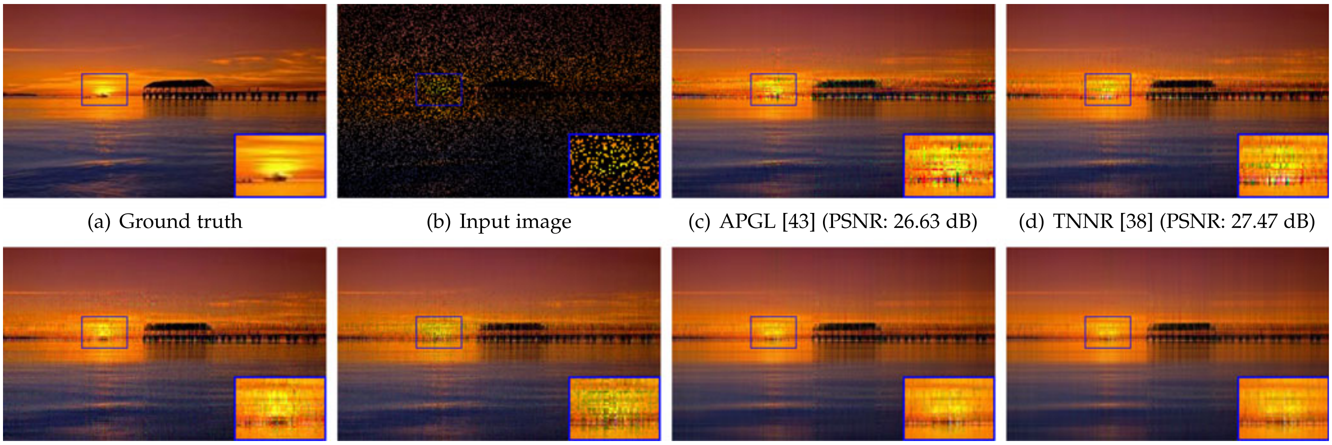
The F-N penalty regularized model and the corresponding ADMM algorithms for solving both models are provided in the Supplementary Materials, available online. We compared our methods with one nuclear norm solver [43], one weighted nuclear norm method [37], [46], two truncated nuclear norm methods [38], [45], and one Schatten-q quasi-norm method [19].[13] Since both of the ADMM algorithms in [38], [45] have very similar performance as shown in [45], we only report the results of [38]. For fair comparison, we set the same values to the parameters d, ρ and μ_0 for both our methods and [38], [45], e.g., $d = 9$ as in [38].

Fig. 12 shows the 8 test images and some quantitative results (including average PSNR and running time) of all those methods with 85 percent random missing pixels. We also show the inpainting results of different methods for random mask of 80 percent missing pixels in Fig. 13 (see the Supplementary Materials, available online, for more results with different missing ratios and rank parameters). The results show that both our methods consistently produce much better PSNR results than the other methods in all the settings. As analyzed in [34], [70], our D-N and F-N penalties not only lead to two scalable optimization problems, but also require

Index		1	2	3	4	5	6	7	8
PSNR (dB)	[43]	23.37	17.29	20.33	21.56	18.07	21.54	26.39	16.45
	[38]	24.29	17.93	19.96	21.87	18.49	21.76	25.87	16.28
	[46]	23.76	17.42	20.65	21.04	17.86	21.31	26.76	16.70
	[19]	23.90	17.99	20.48	21.63	18.56	22.01	26.98	17.05
	D-N	**24.83**	*18.95*	*21.02*	*22.26*	*19.33*	*22.65*	**27.60**	*17.29*
	F-N	*24.69*	**19.18**	**21.23**	**22.49**	**19.40**	**22.82**	*27.23*	**17.46**
Time (sec.)	[43]	13.18	13.39	12.90	16.03	14.93	18.20	16.13	14.57
	[38]	216.70	227.04	194.86	196.38	233.25	247.94	236.41	263.87
	[46]	97.12	87.89	84.68	91.55	98.40	95.13	64.67	87.61
	[19]	823.46	750.69	676.43	687.50	813.72	836.28	449.76	807.63
	D-N	*2.73*	*3.46*	*3.16*	*3.23*	*3.34*	*3.55*	*2.40*	*3.21*
	F-N	**1.91**	**2.57**	**2.29**	**2.46**	**2.68**	**2.73**	**1.87**	**2.63**

Fig. 12. The natural images used in image inpainting [38] (top), and quantitative comparison of inpainting results (bottom).

11. http://perception.csl.illinois.edu/matrix-rank/
12. https://sites.google.com/site/zjuyaohu/
13. As suggested in [19], we chose the ℓ_q-norm penalty, where q was chosen from the range of $\{0.1, 0.2, \ldots, 1\}$.



(a) Ground truth (b) Input image (c) APGL [43] (PSNR: 26.63 dB) (d) TNNR [38] (PSNR: 27.47 dB)

(e) WNNM [46] (PSNR: 27.95 dB) (f) IRNN [19] (PSNR: 26.86 dB) (g) D-N (PSNR: 28.79 dB) (h) F-N (PSNR: 28.64 dB)

Fig. 13. Image inpainting results of different methods on Image 7 with 80 percent random missing pixels (best viewed in colors).

significantly fewer observations than traditional nuclear norm solvers, e.g., [43]. Moreover, both our methods are much faster than the other methods, in particular, more than 25 times faster than the methods [19], [38], [46].

7 Conclusion and Discussions

In this paper we defined the double nuclear norm and Frobenius/nuclear hybrid norm penalties, which are in essence the Schatten-$1/2$ and $2/3$ quasi-norms, respectively. To take advantage of the hyper-Laplacian priors of sparse noise/outliers and singular values of low-rank components, we proposed two novel tractable bilinear factor matrix norm penalized methods for low-level vision problems. Our experimental results show that both our methods can yield more accurate solutions than original Schatten quasi-norm minimization when the number of observations is very limited, while the solutions obtained by the three methods are almost identical when a sufficient number of observations is observed. The effectiveness and generality of both our methods are demonstrated through extensive experiments on both synthetic data and real-world applications, whose results also show that both our methods perform more robust to outliers and missing ratios than existing methods.

An interesting direction of future work is the theoretical analysis of the properties of both of our bilinear factor matrix norm penalties compared to the nuclear norm and the Schatten quasi-norm. For example, how many observations are sufficient for both our models to reliably recover low-rank matrices, although in our experiments we found that our methods perform much better than existing Schatten quasi-norm methods with a limited number of observations. In addition, we are interested in exploring ways to regularize our models with auxiliary information, such as graph Laplacian [27], [80], [81] and hyper-Laplacian matrix [82], or the elastic-net [15]. We can apply our bilinear factor matrix norm penalties to various structured sparse and low-rank problems [5], [17], [28], [33], e.g., corrupted columns [9] and Hankel matrix for image restoration [25].

Acknowledgments

Fanhua Shang, James Cheng and Yuanyuan Liu were supported in part by Grants (CUHK 14206715 & 14222816) from the Hong Kong RGC, ITF 6904079, Grant 3132821 funded by the Research Committee of CUHK. Zhi-Quan Luo was supported in part by the National Science Foundation (NSF) under Grant CCF-1526434, the National Natural Science Foundation of China under Grants 61571384 and 61731018, and the Leading Talents of Guang Dong Province program, Grant No. 00201510. Zhouchen Lin was supported by National Basic Research Program of China (973 Program) (grant no. 2015CB352502), National Natural Science Foundation (NSF) of China (grant nos. 61625301, 61731018, and 61231002), and Qualcomm. Fanhua Shang and Yuanyuan Liu are the corresponding authors.

References

[1] M. Aharon, M. Elad, and A. Bruckstein, "K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Trans. Signal Proces.*, vol. 54, no. 11, pp. 4311–4322, Nov. 2006.

[2] J. Wright, A. Yang, A. Ganesh, S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 2, pp. 210–227, Feb. 2009.

[3] F. De la Torre and M. J. Black, "A framework for robust subspace learning," *Int. J. Comput. Vis.*, vol. 54, no. 1, pp. 117–142, 2003.

[4] E. Elhamifar and R. Vidal, "Sparse subspace clustering: Algorithm, theory, and applications," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 11, pp. 2765–2781, Nov. 2013.

[5] G. Liu, Z. Lin, S. Yan, J. Sun, Y. Yu, and Y. Ma, "Robust recovery of subspace structures by low-rank representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 171–184, Jan. 2013.

[6] E. J. Candès, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis?" *J. ACM*, vol. 58, no. 3, pp. 1–37, 2011.

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

[8] T. Zhou and D. Tao, "GoDec: Randomized low-rank & sparse matrix decomposition in noisy case," in *Proc. 28th Int. Conf. Mach. Learn.*, 2011, pp. 33–40.

[9] Y. Chen, H. Xu, C. Caramanis, and S. Sanghavi, "Robust matrix completion with corrupted columns," in *Proc. 28th Int. Conf. Mach. Learn.*, 2011, pp. 873–880.

[10] D. Krishnan and R. Fergus, "Fast image deconvolution using hyper-Laplacian priors," in *Proc. Adv. Neural Inf. Process. Syst.*, 2009, pp. 1033–1041.

[11] S. Mallat, "A theory for multiresolution signal decomposition: The wavelet representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 11, no. 7, pp. 674–693, Jul. 1989.

[12] L. Luo, J. Yang, J. Qian, and Y. Tai, "Nuclear-L_1 norm joint regression for face reconstruction and recognition with mixed noise," *Pattern Recognit.*, vol. 48, no. 12, pp. 3811–3824, 2015.

[13] A. Eriksson and A. van den Hengel, "Efficient computation of robust low-rank matrix approximations in the presence of missing data using the L_1 norm," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2010, pp. 771–778.

[14] Q. Ke and T. Kanade, "Robust L_1 norm factorization in the presence of outliers and missing data by alternative convex programming," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2005, pp. 739–746.

[15] E. Kim, M. Lee, and S. Oh, "Elastic-net regularization of singular values for robust subspace learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 915–923.

[16] Y. Zheng, G. Liu, S. Sugimoto, S. Yan, and M. Okutomi, "Practical low-rank matrix approximation under robust L_1-norm," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 1410–1417.

[17] X. Zhou, C. Yang, and W. Yu, "Moving object detection by detecting contiguous outliers in the low-rank representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 3, pp. 597–610, Mar. 2013.

[18] J. Fan and R. Li, "Variable selection via nonconcave penalized likelihood and its Oracle properties," *J. Am. Statist. Assoc.*, vol. 96, no. 456, pp. 1348–1360, 2001.

[19] C. Lu, J. Tang, S. Yan, and Z. Lin, "Generalized nonconvex nonsmooth low-rank minimization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 4130–4137.

[20] C. Zhang, "Nearly unbiased variable selection under minimax concave penalty," *Ann. Statist.*, vol. 38, no. 2, pp. 894–942, 2010.

[21] Z. Xu, X. Chang, F. Xu, and H. Zhang, "$L_{1/2}$ regularization: A thresholding representation theory and a fast solver," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 23, no. 7, pp. 1013–1027, Jul. 2012.

[22] W. Zuo, D. Meng, L. Zhang, X. Feng, and D. Zhang, "A generalized iterated shrinkage algorithm for non-convex sparse coding," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2013, pp. 217–224.

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

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

[25] K. Jin and J. Ye, "Annihilating filter-based low-rank Hankel matrix approach for image inpainting," *IEEE Trans. Image Process.*, vol. 24, no. 11, pp. 3498–3511, Nov. 2015.

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

[27] N. Shahid, V. Kalofolias, X. Bresson, M. Bronstein, and P. Vandergheynst, "Robust principal component analysis on graphs," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 2812–2820.

[28] H. Zhang, Z. Lin, C. Zhang, and E. Chang, "Exact recoverability of robust PCA via outlier pursuit with tight recovery bounds," in *Proc. 29th AAAI Conf. Artif. Intell.*, 2015, pp. 3143–3149.

[29] Z. Zhang, A. Ganesh, X. Liang, and Y. Ma, "TILT: Transform invariant low-rank textures," *Int. J. Comput. Vis.*, vol. 99, no. 1, pp. 1–24, 2012.

[30] Z. Gu, M. Shao, L. Li, and Y. Fu, "Discriminative metric: Schatten norm versus vector norm," in *Proc. 21st Int. Conf. Pattern Recognit.*, 2012, pp. 1213–1216.

[31] F. Nie, H. Wang, X. Cai, H. Huang, and C. Ding, "Robust matrix completion via joint Schatten p-norm and L_p-norm minimization," in *Proc. IEEE Int. Conf. Data Mining*, 2012, pp. 566–574.

[32] M. Lai, Y. Xu, and W. Yin, "Improved iteratively rewighted least squares for unconstrained smoothed ℓ_p minimization," *SIAM J. Numer. Anal.*, vol. 51, no. 2, pp. 927–957, 2013.

[33] C. Lu, Z. Lin, and S. Yan, "Smoothed low rank and sparse matrix recovery by iteratively reweighted least squares minimization," *IEEE Trans. Image Process.*, vol. 24, no. 2, pp. 646–654, Feb. 2015.

[34] F. Shang, Y. Liu, and J. Cheng, "Scalable algorithms for tractable Schatten quasi-norm minimization," in *Proc. 30th AAAI Conf. Artif. Intell.*, 2016, pp. 2016–2022.

[35] G. H. Golub and C. F. V. Loan, *Matrix Computations*, 4th ed. Baltimore, MD, USA: Johns Hopkins Univ. Press, 2013.

[36] S. Wang, L. Zhang, and Y. Liang, "Nonlocal spectral prior model for low-level vision," in *Proc. Asian Conf. Comput. Vis.*, 2013, pp. 231–244.

[37] S. Gu, L. Zhang, W. Zuo, and X. Feng, "Weighted nuclear norm minimization with application to image denoising," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 2862–2869.

[38] Y. Hu, D. Zhang, J. Ye, X. Li, and X. He, "Fast and accurate matrix completion via truncated nuclear norm regularization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 9, pp. 2117–2130, Sep. 2013.

[39] W. Bian, X. Chen, and Y. Ye, "Complexity analysis of interior point algorithms for non-Lipschitz and nonconvex minimization," *Math. Program.*, vol. 149, no. 1, pp. 301–327, 2015.

[40] J. Wright, A. Ganesh, S. Rao, Y. Peng, and Y. Ma, "Robust principal component analysis: Exact recovery of corrupted low-rank matrices by convex optimization," in *Proc. Adv. Neural Inf. Process. Syst.*, 2009, pp. 2080–2088.

[41] J. Wright, A. Ganesh, K. Min, and Y. Ma, "Compressive principal component pursuit," *Inform. Infer.*, vol. 2, no. 1, pp. 32–68, 2013.

[42] F. Shang, Y. Liu, H. Tong, J. Cheng, and H. Cheng, "Robust bilinear factorization with missing and grossly corrupted observations," *Inform. Sci.*, vol. 307, pp. 53–72, 2015.

[43] K.-C. Toh and S. Yun, "An accelerated proximal gradient algorithm for nuclear norm regularized least squares problems," *Pac. J. Optim.*, vol. 6, pp. 615–640, 2010.

[44] Z. Lin, M. Chen, and L. Wu, "The augmented Lagrange multiplier method for exact recovery of corrupted low-rank matrices," Coordinated Sci. Lab., Univ. Illinois, Urbana-Champaign, Il, USA, Tech. Rep. UILU-ENG-09–2215 DC-247, Mar. 2009.

[45] T. Oh, Y. Tai, J. Bazin, H. Kim, and I. Kweon, "Partial sum minimization of singular values in robust PCA: Algorithm and applications," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 4, pp. 744–758, Apr. 2016.

[46] S. Gu, Q. Xie, D. Meng, W. Zuo, X. Feng, and L. Zhang, "Weighted nuclear norm minimization and its applications to low level vision," *Int. J. Comput. Vis.*, vol. 121, no. 2, pp. 183–208, 2017.

[47] Y. Shen, Z. Wen, and Y. Zhang, "Augmented Lagrangian alternating direction method for matrix separation based on low-rank factorization," *Optim. Method Softw.*, vol. 29, no. 2, pp. 239–263, 2014.

[48] Y. Liu, L. Jiao, and F. Shang, "A fast tri-factorization method for low-rank matrix recovery and completion," *Pattern Recognit.*, vol. 46, no. 1, pp. 163–173, 2013.

[49] R. Cabral, F. De la Torre, J. Costeira, and A. Bernardino, "Unifying nuclear norm and bilinear factorization approaches for low-rank matrix decomposition," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2013, pp. 2488–2495.

[50] Y. Peng, A. Ganesh, J. Wright, W. Xu, and Y. Ma, "RASL: Robust alignment by sparse and low-rank decomposition for linearly correlated images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 11, pp. 2233–2246, Nov. 2012.

[51] H. Ji, C. Liu, Z. Shen, and Y. Xu, "Robust video denoising using low rank matrix completion," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2010, pp. 1791–1798.

[52] D. L. Donoho, "De-noising by soft-thresholding," *IEEE Trans. Inform. Theory*, vol. 41, no. 3, pp. 613–627, May 1995.

[53] H. Xu, C. Caramanis, and S. Sanghavi, "Robust PCA via outlier pursuit," in *Proc. Adv. Neural Inf. Process. Syst.*, 2010, pp. 2496–2504.

[54] Y. Yang, Z. Ma, A. G. Hauptmann, and N. Sebe, "Feature selection for multimedia analysis by sharing information among multiple tasks," *IEEE Trans. Multimedia*, vol. 15, no. 3, pp. 661–669, Apr. 2013.

[55] Z. Ma, Y. Yang, N. Sebe, and A. G. Hauptmann, "Knowledge adaptation with partially shared features for event detection using few exemplars," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 9, pp. 1789–1802, Sep. 2014.

[56] M. Shao, D. Kit, and Y. Fu, "Generalized transfer subspace learning through low-rank constraint," *Int. J. Comput. Vis.*, vol. 109, no. 1, pp. 74–93, 2014.

[57] Z. Ding, M. Shao, and S. Yan, "Missing modality transfer learning via latent low-rank constraint," *IEEE Trans. Image Process.*, vol. 24, no. 11, pp. 4322–4334, Nov. 2015.

[58] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Found. Trends Mach. Learn.*, vol. 3, no. 1, pp. 1–122, 2011.

[59] T. Oh, Y. Matsushita, Y. Tai, and I. Kweon, "Fast randomized singular value thresholding for nuclear norm minimization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 4484–4493.

[60] R. Larsen, "PROPACK-software for large and sparse SVD calculations," 2005.

[61] J. Cai and S. Osher, "Fast singular value thresholding without singular value decomposition," *Methods Anal. Appl.*, vol. 20, no. 4, pp. 335–352, 2013.

[62] F. Jiang, M. Oskarsson, and K. Åstrom, "On the minimal problems of low-rank matrix factorization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 2549–2557.

[63] S. Xiao, W. Li, D. Xu, and D. Tao, "FaLRR: A fast low rank representation solver," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 4612–4620.

[64] X. Shu, F. Porikli, and N. Ahuja, "Robust orthonormal subspace learning: Efficient recovery of corrupted low-rank matrices," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 3874–3881.

[65] F. Shang, Y. Liu, J. Cheng, and H. Cheng, "Robust principal component analysis with missing data," in *Proc. 23rd ACM Int. Conf. Inf. Knowl. Manag.*, 2014, pp. 1149–1158.

[66] G. Liu and S. Yan, "Active subspace: Toward scalable low-rank learning," *Neural Comput.*, vol. 24, no. 12, pp. 3371–3394, 2012.

[67] N. Srebro, J. Rennie, and T. Jaakkola, "Maximum-margin matrix factorization," in *Proc. Adv. Neural Inf. Process. Syst.*, 2004, pp. 1329–1336.

[68] J. Feng, H. Xu, and S. Yan, "Online robust PCA via stochastic optimization," in *Proc. Adv. Neural Inf. Process. Syst.*, 2013, pp. 404–412.

[69] R. Sun and Z.-Q. Luo, "Guaranteed matrix completion via non-convex factorization," *IEEE Trans. Inform. Theory*, vol. 62, no. 11, pp. 6535–6579, Nov. 2016.

[70] F. Shang, Y. Liu, and J. Cheng, "Tractable and scalable Schatten quasi-norm approximations for rank minimization," in *Proc. 19th Int. Conf. Artif. Intell. Statist.*, 2016, pp. 620–629.

[71] L. Mirsky, "A trace inequality of John von Neumann," *Monatsh. Math.*, vol. 79, pp. 303–306, 1975.

[72] J. von Neumann, "Some matrix-inequalities and metrization of matric-space," *Tomsk Univ. Rev.*, vol. 1, pp. 286–300, 1937.

[73] R. Mazumder, T. Hastie, and R. Tibshirani, "Spectral regularization algorithms for learning large incomplete matrices," *J. Mach. Learn. Res.*, vol. 11, pp. 2287–2322, 2010.

[74] W. Cao, J. Sun, and Z. Xu, "Fast image deconvolution using closed-form thresholding formulas of l_q $(q = \frac{1}{2}, \frac{2}{3})$ regularization," *J. Vis. Commun. Image R.*, vol. 24, no. 1, pp. 31–41, 2013.

[75] J. Bolte, S. Sabach, and M. Teboulle, "Proximal alternating linearized minimization for nonconvex and nonsmooth problems," *Math. Program.*, vol. 146, no. 1, pp. 459–494, 2014.

[76] M. Hong, Z.-Q. Luo, and M. Razaviyayn, "Convergence analysis of alternating direction method of multipliers for a family of nonconvex problems," *SIAM J. Optim.*, vol. 26, no. 1, pp. 337–364, 2016.

[77] C. Chen, B. He, Y. Ye, and X. Yuan, "The direct extension of ADMM for multi-block convex minimization problems is not necessarily convergent," *Math. Program.*, vol. 155, no. 1–2, pp. 57–79, 2016.

[78] U. von Luxburg, "A tutorial on spectral clustering," *Stat. Comput.*, vol. 17, no. 4, pp. 395–416, 2007.

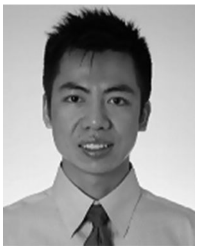
[79] R. Keshavan, A. Montanari, and S. Oh, "Low-rank matrix completion with noisy observations: A quantitative comparison," in *Proc. 47th Conf. Commun. Control Comput.*, 2009, pp. 1216–1222.

[80] Y. Yang, D. Xu, F. Nie, S. Yan, and Y. Zhuang, "Image clustering using local discriminant models and global integration," *IEEE Trans. Image Process.*, vol. 19, no. 10, pp. 2761–2773, Oct. 2010.

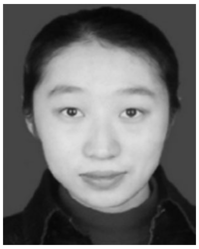
[81] Y. Yang, F. Nie, D. Xu, J. Luo, Y. Zhuang, and Y. Pan, "A multimedia retrieval framework based on semi-supervised ranking and relevance feedback," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 4, pp. 723–742, Apr. 2012.

[82] M. Yin, J. Gao, and Z. Lin, "Laplacian regularized low-rank representation and its applications," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 3, pp. 504–517, Mar. 2016.

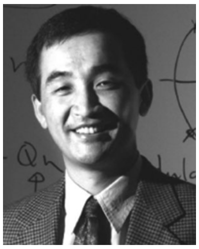
Fanhua Shang (M'14) received the PhD degree in circuits and systems from Xidian University, Xi'an, China, in 2012. He is currently a research associate in the Department of Computer Science and Engineering, The Chinese University of Hong Kong. From 2012 to 2013, he was a post-doctoral research associate in the Department of Electrical and Computer Engineering, Duke University, Durham, North Carolina. From 2013 to 2015, he was a post-doctoral research fellow in the Department of Computer Science and Engineering, The Chinese University of Hong Kong. His current research interests include machine learning, data mining, pattern recognition, and computer vision. He is a member of the IEEE.



James Cheng is an assistant professor in the Department of Computer Science and Engineering, The Chinese University of Hong Kong, Hong Kong. His current research interests include distributed computing systems, large-scale network analysis, temporal networks, and big data.



Yuanyuan Liu received the PhD degree in pattern recognition and intelligent system from Xidian University, Xi'an, China, in 2013. She is currently a post-doctoral research fellow in the Department of Computer Science and Engineering, The Chinese University of Hong Kong. Prior to that, she was a post-doctoral research fellow in the Department of Systems Engineering and Engineering Management, CUHK. Her current research interests include image processing, pattern recognition, and machine learning.



Zhi-Quan (Tom) Luo (F'07) received the BSc degree in applied mathematics from Peking University, Beijing, China, in 1984. He received the PhD degree in operations research from the Operations Research Center and the Department of Electrical Engineering and Computer Science, MIT, Cambridge, Massachusetts, in 1989. Subsequently, he was selected by a joint committee of the American Mathematical Society and the Society of Industrial and Applied Mathematics to pursue PhD degree in the USA. From 1989 to 2003, He held a faculty position in the Department of Electrical and Computer Engineering, McMaster University, Hamilton, Canada, where he eventually became the department head and held a Canada Research chair in information processing. Since April of 2003, he has been in the Department of Electrical and Computer Engineering, University of Minnesota (Twin Cities) as a full professor. His research interests lies in the union of optimization algorithms, signal processing and digital communication. Currently, he is with the Chinese University of Hong Kong, Shenzhen, where he is a professor and serves as the vice president academic. He received the 2004, 2009 and 2011 IEEE Signal Processing Society Best Paper Awards, the 2011 EURASIP Best Paper Award and the 2011 ICC Best Paper Award. He was awarded the 2010 Farkas Prize from the INFRMS Optimization Society. He chaired the IEEE Signal Processing Society Technical Committee on the Signal Processing for Communications (SPCOM) from 2011-2012. He has held editorial positions for several international journals including the *Journal of Optimization Theory and Applications*, the *SIAM Journal on Optimization*, *Mathematics of Computation*, and the *IEEE TRANSACTIONS ON SIGNAL PROCESSING*. In 2014 he is elected to the Royal Society of Canada. He is a fellow of the IEEE and ACM.

Zhouchen Lin (M'00-SM'08) received the PhD degree in applied mathematics from Peking University, in 2000. He is currently a professor in the Key Laboratory of Machine Perception, School of Electronics Engineering and Computer Science, Peking University. His research interests include computer vision, image processing, machine learning, pattern recognition, and numerical optimization. He is an area chair of CVPR 2014/2016, ICCV 2015, and NIPS 2015, and a senior program committee member of AAAI 2016/2017/2018 and IJCAI 2016. He is an associate editor of the *IEEE Transactions on Pattern Analysis and Machine Intelligence* and the *International Journal of Computer Vision*. He is a senior member of the IEEE. He is a fellow of the IAPR.