Robust Recovery of Corrupted Low-Rank Matrix by Implicit Regularizers

Ran He, *Member, IEEE*, Tieniu Tan, *Fellow, IEEE*, and Liang Wang, *Senior Member, IEEE*

Abstract—Low-rank matrix recovery algorithms aim to recover a corrupted low-rank matrix with sparse errors. However, corrupted errors may not be sparse in real-world problems and the relationship between ℓ_1 regularizer on noise and robust M-estimators is still unknown. This paper proposes a general robust framework for low-rank matrix recovery via implicit regularizers of robust M-estimators, which are derived from convex conjugacy and can be used to model arbitrarily corrupted errors. Based on the additive form of half-quadratic optimization, proximity operators of implicit regularizers are developed such that both low-rank structure and corrupted errors can be alternately recovered. In particular, the dual relationship between the absolute function in ℓ_1 regularizer and Huber M-estimator is studied, which establishes a connection between robust low-rank matrix recovery methods and M-estimators based robust principal component analysis methods. Extensive experiments on synthetic and real-world data sets corroborate our claims and verify the robustness of the proposed framework.

Index Terms—PCA, implicit regularizers, low-rank matrix recovery, correntropy, ℓ_1 regularization

1 Introduction

Principal component analysis (PCA) is a popular tool in signal processing and machine learning. It assumes that high-dimensional data reside in a low-dimensional linear subspace, and has been widely used for dimensionality reduction. Consider a data set of n samples $D = [d_1, \ldots, d_n]$ where d_i is a variable in m-dimensional euclidean space, $U = [u_1, \ldots, u_r] \in R^{m\times r}$ be a matrix whose columns constitute the bases of r-dimensional subspace, and $V = [v_1, \ldots, v_n] \in R^{r\times n}$ be principal components that are projection coordinates under U. From the viewpoint of mean square error (MSE), PCA assumes that data matrix D is generated by perturbing the matrix $A = UV \in R^{m\times n}$ whose columns reside in a subspace of dimension $r \ll min(m, n)$, i.e., $D = A + E$, where A is a rank-r matrix and E is a matrix whose entries are i.i.d. Gaussian random variables [1]. In this setting, PCA can be formulated as the following constrained optimization problem:

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

where $\|.\|_F$ is the Frobenius norm.

Although PCA can deal with small Gaussian noise in signal processing, it still has two on-going research issues. First, PCA is sensitive to outliers[1] because outliers dominate MSE such that they may significantly change principal subspaces [4], [5], [6]. Second, automatic selection of principal components [7], i.e., finding the intrinsic low-rank structure of data, remains challenging. To address the robustness problem, robust PCA methods have been developed. When matrix A is low-rank, robust PCA is also called the *low-rank matrix recovery* [8]. Traditionally, robust PCA methods [4], [5], [2], [9], [6] often treat some of samples (i.e., d_i) as outliers and replace the MSE in PCA with a robust M-estimator,[2] which can be summarized as the following general M-estimation problem,

$$\min_{U,V,\mu} \sum_{i=1}^{n} \phi(d_i - \mu - Uv_i), \tag{2}$$

where $\mu \in R^m$ is the robust center of D and ϕ is a robust estimator. Features (or samples) are iteratively reweighted and then uncorrupted features (or samples) are utilized to compute robust principal subspaces. To address the problem of automatic selection, Cai et al. [11] proposed a singular value thresholding (SVT) algorithm to find the rank r of a low-rank matrix by solving a convex optimization problem, namely, nuclear-norm minimization.

Recently, Wright et al. [12] showed that the robust PCA problem in (1) can be exactly solved by minimizing a combination of the nuclear norm and ℓ_1 norm if noise is known to be sparse. Some methods [13], [14], [15], [16] are accordingly developed to solve this robust PCA problem and recover the low-rank structure of corrupted matrices. In addition, recent theoretical analysis and experimental results in [14] show that, one can exactly recover a low-rank matrix by using the same optimization algorithm with an improved regularization parameter, even if corruptions are almost

[1] In robust statistics, outliers are those data points that deviate significantly from the rest of the data [2]. They can arise in practical applications due to either the process of data generation or mislabelled data [3].

[2] In statistics, one popular robust technique is the so-called M-estimators [10], which are obtained as the minima of sums of functions of the data. (See (27) in Appendix I, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TPAMI.2013.188.)

- *The authors are with the Center for Research on Intelligent Perception and Computing (CRIPAC) and the National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences, #95, Zhongguancun East Road, Haidian District, PO Box 2728, Beijing 100190, China. E-mail: {rhe, tnt, wangliang}@nlpr.ia.ac.cn.*

Manuscript received 22 Apr. 2012; revised 27 Apr. 2013; accepted 7 Sept. 2013; date of publication 30 Sept. 2014; date of current version 19 Mar. 2014.
Recommended for acceptance by A. Fitzgibbon.
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.2013.188

arbitrarily large. However, when the errors incurred by noise are dense, the solutions of ℓ_0-norm and ℓ_1-norm may not be equivalent, which makes the role of ℓ_1 regularizer still unclear. Although He et al. [17] showed that robust M-estimators and the iteratively reweighting strategy can be used to recover a corrupted low-rank matrix, the relationship between ℓ_1 regularization based robust PCAs and M-estimators based ones remains unknown.

This paper combines research outputs in low-rank matrix recovery [11], [12], proximity operator [18], [19], M-estimation [10], [20], half-quadratic (*HQ*) optimization [21] and information theoretic learning (*ITL*) [22]. The main contributions are summarized as follows:

1) Based on the additive form of HQ optimization, we derive implicit regularizers of robust M-estimators and their proximity operators to model arbitrarily corrupted errors, and propose a unified framework for robust matrix recovery. The introduction of implicit regularizers not only enriches the family of regularizers for robust learning but also provides a general scheme to develop new robust methods. When the implicit regularizer of Welsch M-estimator is used, the maximum correntropy criterion [22] gives a probabilistic foundation to support the proposed framework to recover arbitrarily corrupted low-rank matrices.

2) We study the properties of thresholding function of the ℓ_1 regularizer on errors under convex conjugacy, and point out a connection between the ℓ_1 regularizer solved by soft-thresholding operator and Huber loss function, which bridges the gap between ℓ_1 regularization based robust PCAs and M-estimators based ones.

3) For computational efficiency, we develop a generalized algorithm to solve the optimization problem with implicit regularizers based on the state-of-the-art accelerated proximal gradient (APG) algorithm [1]. Extensive experiments on simulated data set, background modeling, face reconstruction, and gait recognition corroborate our claims and demonstrate that when nonzero items of E are sparse, implicit regularizers of L1-L2 and Welsch M-estimators can also obtain a sparse representation of E. However, the two regularizers perform in a significantly different way in the estimation of noise (or outliers) compared with the ℓ_1 regularizer.

The remainder of this paper is structured as follows. We first review related theories and methods for robust recovery of low-rank matrices in Section 2. Then in Section 3, we propose a general framework for low-rank matrix recovery via implicit regularizers, and discuss the relationship between the ℓ_1 regularizer and Huber M-estimator. In Section 4, the proposed framework is validated by conducting a series of experiments on simulations and real-world applications. Finally, we conclude this paper and discuss future work in Section 5.

2 Related Work

In the last decades, robust PCA has been drawn much attention in the image processing, computer vision and machine learning communities.[3] Various robust PCA methods have been developed for different purposes. In this section, we first review proximity operators and half-quadratic optimization, which are commonly used to solve robust PCA problems. And then we briefly review the recent low-rank matrix recovery methods in robust PCA.

2.1 Proximity Operators

The ℓ_1 regularizer on errors in many robust sparse representation and low-rank matrix recovery methods are solved by soft-thresholding (also known as a shrinkage) operator, which belongs to proximity operator and is derived from a function minimization problem $\min_y h(x,y)$. The function $h(x,y)$ takes the following form,

$$h(x,y) \doteq \frac{1}{2}\|x-y\|_2^2 + \varphi(y), \tag{3}$$

where x and y are variables in real Hilbert space [18], and $\varphi(.)$ is a continuous function in separable regularizers.[4] $h(x,y)$ is often used as a denoising function in image restoration and signal recovery. When $\varphi(.)$ satisfies certain properties [19], the proximity operator of $h(x,y)$ w.r.t. y is unique. Based on Legendre transformation [18], [24], we can define the Moreau proximity operator (*MPO*) $\delta : \mathbb{R} \mapsto \mathbb{R}$ as $y \mapsto \delta(x)$ where

$$\delta(x) \doteq \arg\min_y \left\{ \frac{1}{2}\|x-y\|_2^2 + \varphi(y) \right\}. \tag{4}$$

Proximity operator was firstly introduced in [25]. It is a generalization of a convex projection operator [18], and has been used extensively in nonlinear signal recovery [18], image denoising [19] and sparse representation [26]. And it has recently been used to solve non-convex regularized optimization problems [27]. The soft-thresholding operator w.r.t the absolute function in ℓ_1 regularization is a special case of proximity operator (See [18, Example 2.15 and Equation (2.29)]). To better understand the concept of MPO, we give an example of proximity operator that is widely used in robust low-rank matrix recovery.

Example 1. Let $\varphi(y) = \lambda|y|$, the MPO $\delta(.)$ of (4) is the scalar soft-thresholding operator [18], i.e.,

$$\delta(x) = \begin{cases} 0 & |x| \leq \lambda, \\ x - \lambda \text{sign}(x) & |x| > \lambda, \end{cases} \tag{5}$$

where λ is a positive constant. And $\min_y h(x,y)$ is the Huber loss function in (8) w.r.t. x.

2.2 Half-Quadratic Optimization

Like MPO, half-quadratic optimization [21] is another commonly used optimization method based on Legendre transformation. Different from MPO based methods that mainly focus on regularizers, HQ tries to solve a nonlinear objective function by solving a number of least squares problems iteratively.

[3] More details on robust PCA can be found in our ICPR tutorial: http:// www.cripac.ia.ac.cn/People/rhe/ICPR2012.html.

[4] If a regularizer $\Phi(\mathbf{u})$ is separable, $\Phi(\mathbf{u}) = \sum_i \varphi(u_i)$ [23].

If a function $\phi(.)$ is differentiable and satisfies five conditions of the additive form of HQ optimization (i.e., [21, Equation (59)]), we have that for a fixed x, the following equation holds,[5]

$$\phi(x) = \min_y \frac{1}{2}\|x - y\|_2^2 + \varphi(y), \tag{6}$$

where $\varphi(.)$ is the dual potential function of $\phi(.)$ [21]. The y^* that minimizes (6) is determined by the minimizer function ($y^* = \delta(x)$),[6] which is only relative to a specific function $\phi(.)$. For every x, $\delta(x)$ is such that [21]

$$\frac{1}{2}\|x - \delta(x)\|_2^2 + \varphi(\delta(x)) \leq \frac{1}{2}\|x - y\|_2^2 + \varphi(y). \tag{7}$$

In HQ, one only focuses on $\phi(.)$ and its corresponding minimizer function. The exact formulation of dual potential function $\varphi(.)$ is often unknown. To better understand the concept of the minimizer function, we also give an example of half-quadratic optimization.

Example 2. Let $\phi(.)$ be the Huber function, i.e.,

$$\phi(x) = \begin{cases} \frac{1}{2}x^2 & |x| \leq \lambda, \\ \lambda|x| - \frac{1}{2}\lambda^2 & |x| > \lambda, \end{cases} \tag{8}$$

the absolute function $\lambda|.|$ is the dual potential function of the Huber function and the minimizer function $\delta(x)$ is the scalar soft-thresholding operator in (5) (See [21, Equations (83), (85), (86) and (87)]).

Although MPO and HQ are proposed from different motivations, they are both based on convex conjugacy. Comparing Example 1 with Example 2, we observe that they are quite similar in their objective functions, which motivates us to further analyze their relationship.

2.3 Robust Recovery of Low-Rank Matrix

Low-rank matrix recovery is a subfield of compressed sensing, and its research is started in the research results in [28] and [29]. Recently, there is growing interest in the robust recovery of a corrupted low-rank matrix, or so-called robust PCA [8], [12]. This problem occurs in a number of applications in machine learning and signal processing, and can be formulated as the following nuclear norm minimization problem [8],

$$\min_A \frac{1}{2}\|A - D\|_F^2 + \mu\|A\|_* \tag{9}$$

where $\|.\|_*$ denotes the nuclear norm of a matrix (i.e., the sum of its singular values) and μ is a constant. Cai et al., [11] derived a singular value thresholding algorithm to solve (9). The singular value thresholding operator is the Moreau proximity operator associated with nuclear norm [11].

By assuming that the error matrix E has a sparse representation, Wright et al [12] showed that the robust PCA problem can be exactly solved by minimizing a combination of nuclear norm and ℓ_1-norm. Then robust PCA can be formulated as [12]:

$$\min_{A,E} \|A\|_* + \lambda\|E\|_0 \quad s.t. \quad D = A + E, \tag{10}$$

where $\|.\|_0$ is the counting norm (i.e., the number of nonzero entries in the matrix), and λ is a positive constant. Since the problem in (10) is NP-hard and cannot be efficiently solved, one often considers its relaxation [12],

$$\min_{A,E} \|A\|_* + \lambda\|E\|_1 \quad s.t. \quad D = A + E, \tag{11}$$

where $\|.\|_1$ represents the matrix 1-norm (i.e., the sum of absolute values of all entries of a matrix). The nuclear norm and ℓ_1 norm in (11) are natural convex surrogates for the rank of A [30] and the sparsity of E [31] respectively. These two norms in (11) are generally intractable to optimize [32]. Hence one often uses a relaxed version of (11) [12], [1], [13], [32], i.e.,

$$\min_{A,E} \frac{1}{2}\|D - A - E\|_F^2 + \mu\|A\|_* + \mu\lambda\|E\|_1. \tag{12}$$

The solutions to (12) approach the solution set of (11) as μ decreases [12], [1]. In addition, the regularized formulation in (12) may be more suitable in certain applications and may have different recovery properties [32].

Various methods have been developed to optimize (9), (11) and (12). Wright et al. [12] adopted an iterative thresholding technique. To alleviate the slow convergence of the iterative thresholding method, fast algorithms [1], [33] are developed for recovering a corrupted low-rank matrix. In [13], augmented Lagrange multipliers are utilized to further reduce computational cost. In [16], random projection is introduced to deal with large-scale visual recovery problems. All of these methods are based on the soft-thresholding operator that belongs to Moreau proximity operator [25], [18]. Recently, iteratively reweighted least squares (IRLS) [34] and multiplicative form of HQ optimization [17] are used to solve (9) and (11) respectively. In [35], a convex optimization formulation is introduced such that the low-rank matrix recovery problem is reduced to a semidefinite programming.

An important issue in low-rank matrix recovery is how many sparse corruptions in (11) are allowed so that an accurate recovery can be obtained. Ganesh et al. [14] showed that one can exactly recover an arbitrarily corrupted low-rank matrix by using the same optimization algorithm with an improved regularization parameter. Hsu et al. [32] further gave stronger recovery guarantees to recover any low-rank matrix with high probability. Gross [36] presented new techniques in quantum mechanics for analyzing the problem of low-rank matrix recovery. Although many methods are developed to analyze the robustness of (11) and (12), the role of the ℓ_1 regularizer in (12) still needs to be further studied [32]. When the errors in E are dense, the solutions of ℓ_0-norm and ℓ_1-norm may be inequivalent.

[5] More details about (6) are given in [21, Appendix 7.2.]

[6] As analyzed in the following section, there is a close relationship between the minimizer function and proximity operator. Hence we make use of the same notation $\delta(.)$ to indicate them.

TABLE 1
Robust M-Estimators $\phi(.)$ and Their Minimizer Functions

	Huber	Fair	log-cosh	Welsch	L1-L2
$\phi(t)$	$\begin{cases} t^2/2 & \lvert t\rvert \leq \lambda \\ \lambda\lvert t\rvert - \frac{\lambda^2}{2} & \lvert t\rvert > \lambda \end{cases}$	$\frac{\lvert t\rvert}{\alpha} - \log(1 + \frac{\lvert t\rvert}{\alpha})$	$\log(\cosh(\alpha t))$	$1 - \exp(-\frac{t^2}{\sigma^2})$	$\sqrt{\alpha + \frac{t^2}{\sigma^2}} - 1$
$\delta(t)$	$\begin{cases} 0 & \lvert t\rvert \leq \lambda \\ t - \lambda sign(t) & \lvert t\rvert > \lambda \end{cases}$	$t - \frac{t}{\alpha(\alpha+\lvert t\rvert)}$	$t - \alpha\tanh(\alpha t)$	$t - t\exp(-\frac{t^2}{\sigma^2})$	$t - \frac{t}{\sqrt{\alpha+\frac{t^2}{\sigma^2}}}$

α is a positive constant.

3 Matrix Recovery by Implicit Regularizers

In this section, we first derive the definition of implicit regularizers of robust M-estimators and their proximity operators based on half-quadratic optimization. Then we propose a general framework for robust low-rank matrix recovery by implicit regularizers. Lastly, we study the function value of scalar soft-thresholding operator and discuss its relationship to Huber loss function.

3.1 Implicit Regularizers and Their Proximity Operators

In statistics and information theoretic learning, one common technique to solve the M-estimators based robust learning problem is the half-quadratic optimization [21]. Considering the similarity between HQ and MPO, we define the implicit regularizer of a potential function as follows.

Definition 1. Implicit regularizer. *An implicit regularizer $\varphi(y)$ is defined as the dual potential function of a robust loss function $\phi(x)$ and satisfies*

$$\phi(x) = \min_y \frac{1}{2}\|x - y\|_2^2 + \varphi(y), \tag{13}$$

where the proximity operator of $\varphi(.)$ exists.

According to Definition 1, we can study the properties of an implicit regularizer from both $\varphi(y)$ and its corresponding function $\phi(x)$. Furthermore, if $\phi(x)$ is given, the analytic form of function $\varphi(y)$ can be even unknown. During optimization, an implicit regularizer is iteratively solved by its proximity operator. Equation (13) is the same as to (6). That is, we make use of the additive form of HQ in (6) to define an implicit regularizer, which is similar to the idea in structured sparsity [37].[7] However, different from that HQ minimization focuses on $\phi(x)$, an implicit regularizer applies the right-hand side $\frac{1}{2}\|x - y\|_2^2 + \varphi(y)$ of (13) to model optimization problems.

Proposition 1. *If $\phi(.)$ and its dual potential function $\varphi(.)$ satisfy (13) and there exits a minimizer function $\delta(.)$ of $\phi(.)$, $\delta(.)$ is the proximity operator of regularizer $\varphi(.)$.*

Proof. According to the property of the minimizer function in (7) and the definition of MPO in (4), we have that the minimizer function $\delta(.)$ of $\phi(.)$ is the proximity operator of regularizer $\varphi(.)$. □

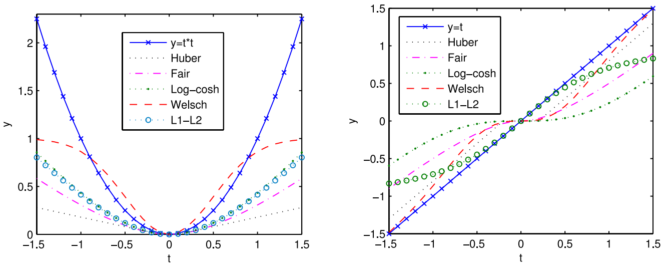
Proposition 1 shows that the minimizer function $\delta(.)$ of $\phi(.)$ in HQ is the proximity operator of regularizer $\varphi(.)$. This is due to the fact that both HQ and MPO are based on Legendre transformation. Table 1 tabulates five potential functions $\phi(.)$ [21], [38] and their corresponding minimizer functions, which are commonly used as the objectives in signal and image processing. In robust statistics [10], [20], these functions belong to M-estimators.

According to the definition of M-estimators (Appendix I, available in the online supplemental material), we see that all M-estimators achieve the minima (zero) at the origin. Fair M-estimator defines continuous derivatives of the first three orders, and yields a unique solution. Log-cosh M-estimator is a strictly convex function and is an approximation of Huber M-estimator. L1-L2 M-estimator takes both the advantage of L1 M-estimator ($|t|$) to reduce the influence of large errors and that of L2 estimator (t^2) to be continuous. Huber M-estimator [10] is a parabola in the vicinity of zero and increases linearly at a given level $|t| > \lambda$. Angst et al. [39] show that Huber M-estimator can efficiently handle outliers than ℓ_1 estimator for motion problems. Welsch M-estimator is widely used in information theoretic learning. It has been proved that the robustness of correntropy [22] based algorithms is actually related to the Welsch M-estimator. All properties of correntropy are controlled by its kernel size σ[8] (See Appendix I, available in the online supplemental material, for details).

Fig. 1 further depicts the five common M-estimators and their minimizer functions. In Fig. 1a, we observe that when $|t| > 1$, the value of MSE (i.e., $y = x^2$) increases rapidly whereas the values of the five M-estimators increase slowly and tend to be stable. This means the M-estimators can punish outliers such that they will generate small loss in M-estimation. If a descending parameter is used in an M-estimator and tends to be zero, the M-estimator will significantly penalize outliers (e.g., $|t| > \lambda$) such that outliers' contribution in the M-estimator based loss function tends to be zero. In information theoretic learning, if the kernel size of Welsch M-estimator is set to large values, correntropy defaults to MSE [22]. From Table 1 and Fig. 1a, it is interesting to observe that when the thresholding parameter λ in Huber M-estimator is set to large values, Huber M-estimator also defaults to MSE. In robust statistics, it seems that Huber M-estimator can efficiently deal with sparse corruptions whereas non-convex M-estimators

[7] The half-quadratic penalty function in structured sparsity [37] ([37, Equation (2.2)]) is relative to the multiplicative form of HQ optimization ([21, Equation (25)]).

[8] Since correntropy is derived from Parzen kernel estimator and satisfies Mercer kernel theorem [22], we denote the parameter σ in Welsch M-estimator by the kernel size as in [22].



(a) M-estimators (b) Proximity operators

Fig. 1. Graphic representations of five common estimators and their minimizer functions.

(such as Welsch) are more robust to larger errors and non-Gaussian noise [22].

In Fig. 1b, it is interesting to find that the five minimizer functions have similar character near the origin ($|t| < 0.5$). Like Huber M-estimator, the other four M-estimators have a smooth region near the origin. This character indicates that when proximity operators are used to estimate outliers, they will estimate ground-truth values of outliers and map variations of the uncorrupted data to zero because outliers are assumed to be far away from the origin. In addition, if the parameters of the five M-estimators tend to be zero, the five minimizer functions will be $\delta(t) = t$, which makes an M-estimator recover corrupted data from outliers (e.g., $E^*_{ij} = \delta(D_{ij} - A_{ij}) = D_{ij} - A_{ij}$). As discussed in correntropy [22] whose properties are controlled by the kernel size σ, the robustness of the five M-estimators is also affected by their parameters, which control the boundary of the smooth region around the origin. When outliers are sparse, the kernel size or threshold parameters in robust M-estimators are often determined by a robust and descending way [20], [22], such as mean and median.

3.2 A General Framework for Robust Matrix Recovery

Based on M-estimation and the proposed implicit regularizers, we propose a general framework for robust matrix recovery. By substituting the Frobenius norm in (9) with a robust M-estimator that has an implicit regularizer,[9] we have the following M-estimation problem,

$$\min_{A} \sum_{i=1}^{n} \sum_{j=1}^{m} \phi(D_{ij} - A_{ij}) + \mu \|A\|_* \tag{14}$$

where $\mu > 0$ is a regularization parameter. The optimal solution E^* of (14) depends on the optimal solution A^* (i.e., $E^* = D - A^*$). The optimization problem in (14) is a robust formulation of matrix completion in [11]. When $\phi(.)$ is a non-convex M-estimator, there are many local minima in (14). However, non-convex loss functions often can more effectively deal with non-Gaussian and large outliers in real world problems [20], [22]. And they also have drawn much attention in non-convex regularized sparse learning problems [27], [40].

Based on convex conjugacy and the additive form of HQ optimization, we have the augmented problem of (14),

$$\min_{A,E} \sum_{i=1}^{n} \sum_{j=1}^{m} \left\{ \frac{1}{2} (D_{ij} - A_{ij} - E_{ij})^2 + \varphi(E_{ij}) \right\} + \mu \|A\|_* \tag{15}$$

where $\varphi(.)$ is an implicit regularizer w.r.t. $\phi(.)$, E is an auxiliary matrix in HQ optimization, and E_{ij} is determined by the minimizer function of $\phi(.)$. If $\varphi(E) \doteq \sum_i \sum_j \varphi(E_{ij})$, we can reformulate (15) as the following regularization problem,

$$\min_{A,E} \frac{1}{2} \|D - A - E\|_F^2 + \varphi(E) + \mu \|A\|_*. \tag{16}$$

If the M-estimator $\phi(.)$ in (14) is Huber M-estimator, the implicit regularizer $\varphi(.)$ in (16) becomes $\mu\lambda\|.\|_1$ norm. When the M-estimator $\phi(.)$ in (14) is Welsch M-estimator, the minimization problem becomes the sample based maximum correntropy problem (See Appendix I, available in the online supplemental material). Compared with mean square error $\|.\|_F$ in (9), the model in (14) or (16) is more robust to outliers due to M-estimation. As illustrated in Fig. 1, it only minimizes the loss at the origin ($D - A = 0$) and treats nonzero loss with large values as outliers.

Comparing with the M-estimators based objective in (14) with those objectives in traditional robust PCAs [4], [2], [6] in (2), we find that although they are all based on robust estimators to deal with outliers, they are different. First, traditional robust PCAs often treat a sample (a column in D) as an outlier whereas the low-rank recovery model in (14) treats an item D_{ij} as an outlier. Second, the nuclear norm makes the model in (14) automatically determine the number of Eigenvectors.

In traditional robust PCAs, the problem in (14) can be solved by the multiplicative form of HQ optimization (or iteratively reweighted least squares) as in [17]. However, the iteratively reweighted methods involve matrix multiplication, which often results in a high computational cost [21]. Fortunately, we can solve (14) via solving (16), in which the optimization problem also belongs to the linear inverse problem with compound regularization [19], [41], [23] where two regularizers are nuclear norm and implicit regularizer.

In compressed sensing, proximal gradient methods have shown their advantage in solving linear inverse problems (or ℓ_1 minimization) [42], nuclear norm minimization [43], and low-rank matrix recovery [1]. This class of methods can be viewed as an extension of the classical gradient algorithm. Hence, we resort to the proximal gradient method [42] to solve the compound regularization problem in (16). Let the pair (A_k, E_k) be the value of the pair (A, E) in the kth iteration, we can solve (16) by iteratively solving the following subproblems [1], [12],

$$A_{k+1} = \arg\min_{A} \frac{1}{2} \|D - A - E_k\|_F^2 + \mu \|A\|_*, \tag{17}$$

$$E_{k+1} = \arg\min_{E} \frac{1}{2} \|D - A_{k+1} - E\|_F^2 + \varphi(E). \tag{18}$$

[9] Note that, both L1 M-estimator $|.|$ and Lp M-estimator $|.|^p$ are not applicable in the additive form of HQ minimization and have no implicit regularizers.

Here, we follow the approach of accelerated proximal gradient [1], [42] to solve the above two subproblems. Let $f(x) \doteq \frac{1}{2}\|x-b\|_2^2$ where b is a given constant and

$$Q(x,y) \doteq f(y) + < \nabla f(y), x-y > + \frac{L_f}{2}\|x-y\|_2^2 + \mu\varphi(x),$$

where the Lipschitz constant $L_f = 2$ [1]. Moreover, if we define $G \doteq y - \frac{1}{L_f}\nabla f(y)$, then

$$\arg\min_x Q(x,y) = \arg\min_x \frac{L_f}{2}\|x-G\|_2^2 + \mu\varphi(x). \tag{19}$$

Based on $Q(x,y)$ and (19), we can further solve subproblems in (17) and (18) by solving the following two subproblems [1],

$$A_{k+1} = \arg\min_A \frac{1}{2}\left\|A - G_k^A\right\|_F^2 + \mu\|A\|_*, \tag{20}$$

$$E_{k+1} = \arg\min_E \frac{1}{2}\left\|E - G_k^E\right\|_F^2 + \varphi(E). \tag{21}$$

The subproblem in (20) can be solved by the singular value thresholding operator [11] and the subproblem in (21) can be solved by the proximity operator [18], [19]. Let USV^T be the singular value decomposition (SVD) of G_k^A. Then the optimal solutions of (20) and (21) are

$$A_{k+1} = U\delta^*_{\frac{\mu_k}{2}}(S)V^T, \quad E_{k+1} = \delta\left(G_k^E\right), \tag{22}$$

where $\delta^*_{\frac{\mu_k}{2}}(S)$ is the singular value thresholding operator [11] on matrix S and $\delta(.)$ is the minimizer function listed in Table 1. As in compressed sensing and low-rank matrix recovery, the parameter in $\delta(.)$ (i.e., λ in Huber, α in Fair and log-cosh, σ in L1-L2 and Welsch) often has a descending sequence and approaches zero, or is set to a small value.

Algorithm 1: Generalized Accelerated Proximal Gradient (GAPG) Algorithm

```
Input: data matrix $D \in R^{m\times n}$, $\bar{\mu} = \theta\mu_0$, $\eta = 0.9$,
       $A_0 = A_{-1} = 0$, $E_0, E_{-1} = 0$, $t_0 = t_{-1} = 1$.
Output: $A_k$, $E_k$
 1: while 'not converged' do
 2:     $Y_k^A \leftarrow A_k + \frac{t_{k-1}-1}{t_k}(A_k - A_{k-1})$.
 3:     $Y_k^E \leftarrow E_k + \frac{t_{k-1}-1}{t_k}(E_k - E_{k-1})$.
 4:     $G_k^A \leftarrow Y_k^A - \frac{1}{2}(Y_k^A + Y_k^E - D)$.
 5:     $(U,S,V) \leftarrow svd(G_k^A)$, $A_{k+1} \leftarrow U\delta^*_{\frac{\mu_k}{2}}(S)V^T$.
 6:     $G_k^E \leftarrow Y_k^E - \frac{1}{2}(Y_k^A + Y_k^E - D)$.
 7:     $E_{k+1} \leftarrow \delta(G_k^E)$.
 8:     $t_{k+1} \leftarrow \frac{1+\sqrt{4t_k^2+1}}{2}$, $\mu_{k+1} \leftarrow max(\eta\mu_k, \bar{\mu})$ .
 9:     $k = k+1$.
10: end while
```

Algorithm 1 summarizes the procedure of our generalized accelerated proximal gradient (GAPG) algorithm. Since μ_k converges to $\overline{\mu} > 0$, the proof of convergence of Algorithm 1 is similar to the ones provided in [42] and [1]. As suggested in [1], μ varies from a large initial value μ_0 and decreases until it reaches the floor $\overline{\mu}$, and θ and η are set to 10^{-9} and 0.9 respectively. As in [1], the stopping criterion of Algorithm 1 is identical to the one in [43]. In step 7, Algorithm 1 obtains a robust and conjugated solution of gradient G_k^E via the MPO of an implicit regularizer, based on which it further finds a feasible solution Y_k^E to decrease its objective in step 3.

In particular, when Huber M-estimator is used in (14) and a descending parameter $\mu\lambda$ is set as in [1], $\delta(.)$ and $\varphi(.)$ in (22) and (21) become soft-thresholding function and $\mu\lambda\|E\|_1$ respectively. From this viewpoint, the algorithm in [1] is a special case of Algorithm 1. In addition, according to convex conjugacy [21], the following equation of (21) for a fixed G_k^E always exists,

$$\begin{aligned}&\min_E \frac{1}{2}\left\|E - G_k^E\right\|_F^2 + \mu\lambda\|E\|_1\\ &= \frac{1}{2}\left\|\delta\left(G_k^E\right) - G_k^E\right\|_F^2 + \mu\lambda\left\|\delta\left(G_k^E\right)\right\|_1 = \phi_H^{\mu\lambda}\left(G_k^E\right),\end{aligned} \tag{23}$$

where $\phi_H^{\mu\lambda}(.)$ denotes the Huber M-estimator in (8) and $\mu\lambda$ is its thresholding parameter. That is, the optimum solution E^*_{ij} in (21) is a dual variable of Huber M-estimator at point G_k^E.

3.3 ℓ_1 Regularizer and Huber Loss Function

In this subsection, we further study the problems in (14) and (15) when $\phi(.)$ is the Huber M-estimator denoted by $\phi_H^{\mu\lambda}(.)$. According to the definitions of $\|.\|_F$ and $\|.\|_1$, we can rewrite (12) as follows,

$$\min_{A,E} \sum_{i=1}^{n}\sum_{j=1}^{m}\left(\frac{1}{2}(D_{ij} - A_{ij} - E_{ij})^2 + \mu\lambda\left|E_{ij}\right|\right) + \mu\|A\|_*. \tag{24}$$

By substituting soft-thresholding operator back into (24), (24) takes the following form,

$$\min_A \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{m}\phi_H^{\mu\lambda}(D_{ij} - A_{ij}) + \mu\|A\|_*. \tag{25}$$

Since $\phi_H^{\mu\lambda}(.)$ belongs to M-estimators, we learn that the model in (25) can treat errors incurred by occlusion or corruption whether the errors are sparse or not. Comparing (11), (12), (24) and (25), we have the following observations:

1) If we resort to soft-thresholding operator (denoted by $\delta_H^{\mu\lambda}(x)$) to solve (12) and (25), the optimal solution of (12) and (25) tends to be that of (11) when μ in soft-thresholding operator tends to be zero. When $\mu\lambda|x|$ tends to be zero, soft-thresholding operator $\delta_H^{\mu\lambda}(x) \to x$ such that $E^*_{ij} = \delta_H^{\mu\lambda}(D_{ij} - A^*_{ij}) = D_{ij} - A^*_{ij}$. As a result, the optimal solution (A^*,E^*) of (12) and (25) satisfies $D = A^* + E^*$ and approaches the solution set of (11). Whether the errors modeled by E are sparse or dense, the ℓ_1 regularizer solved by soft-thresholding function is always related to the dual potential function of Huber M-estimator. If outliers are significantly different from uncorrupted data, the model in (12) can efficiently deal with dense or sparse outliers.

2) By substituting the equality constraint $E = D - A$ into the objective of (11), we directly have the following minimization problem,

$$\min_A \lambda\|D - A\|_1 + \|A\|_*, \tag{26}$$

where the ℓ_1-norm is a natural convex surrogate for sparsity, and is generally intractable to optimize [31], [32]. It has been

TABLE 2
Correct Recovery for Randomly Corrupted Matrices of Varying Size

m	$rank(A_0)$	$\|\|E_0\|\|_0$	$\|\|\hat{A}-A_0\|\|_F/\|\|A_0\|\|_F$			$rank(\hat{A})$			$\|\|\hat{E}\|\|_0/\|\|E_0\|\|_0$		
			Huber	Welsch	L1-L2	Huber	Welsch	L1-L2	Huber	Welsch	L1-L2
200	10	2,000	1.1×10^{-7}	1.9×10^{-7}	1.6×10^{-7}	10.0	10.0	10.0	1.00	1.00	1.00
200	20	4,000	1.7×10^{-8}	1.5×10^{-8}	1.7×10^{-8}	20.0	20.0	20.0	1.00	1.00	1.00
500	25	12,500	1.1×10^{-7}	1.5×10^{-7}	1.0×10^{-7}	25.0	25.0	25.0	1.00	1.00	1.00
500	50	25,000	9.0×10^{-8}	3.1×10^{-8}	2.5×10^{-8}	50.0	50.0	50.0	1.00	1.00	1.00
1000	50	50,000	5.9×10^{-8}	1.9×10^{-7}	7.0×10^{-8}	50.0	50.0	50.0	1.00	1.00	1.00
1000	100	100,000	8.2×10^{-8}	2.5×10^{-8}	2.5×10^{-8}	100.0	100.0	100.0	1.00	1.00	1.00

shown that the sparse solution of (12) approaches that of (11) and (10) when μ tends to be zero and E (or $D-A$) is sparse [12], [1]. If we treat the ℓ_1-norm in (26) as L1 M-estimator[10] in robust statistics rather than sparsity surrogate, (26) can be solved by iteratively reweighted least squares [20] when $D_{ij} \neq A_{ij}$ for $\forall i,j$.[11] However, since there will be many zero elements in $E = D - A$ (i.e., E is sparse), we cannot simply treat (26) as an L1 M-estimation problem and use IRLS to solve (26).

3) From (12) and (25), we learn that ℓ_1 regularized low-rank matrix recovery methods for solving (12) have a potential relationship with traditional M-estimators based PCA methods in (2). They are all related to M-estimation. In HQ optimization, there are multiplicative and additive forms [21]. Traditional M-estimator based PCA methods often apply the multiplicative form whereas the ℓ_1 regularized low-rank matrix recovery methods harness the additive form of Huber M-estimator (i.e., soft-thresholding operator). The merit of using the additive form is to avoid weighting samples, which results in a low computational cost. Another advantage of low-rank matrix recovery is to automatically determine a low-rank eigenspace.

4 Experiments

In this section, numerical simulations are first run to evaluate the recovery ability of different M-estimators. Then three computer vision applications involving background modeling, face reconstruction and gait recognition are further used to verify the robustness of the proposed framework. All algorithms were implemented in MATLAB based on the code available in Yi Ma's website.[12] The singular value decomposition is implemented by PROPACK,[13] which uses the iterative Lanczos algorithm to compute the SVD directly. The parameters μ_0 and t_0 in Algorithm 1 are set as the same as those of accelerated proximal gradient algorithm in [1]. Note that the goal of our experiments is to compare different regularizers based methods for robust learning rather than to just achieve the highest recognition accuracy on these data sets.

Since the minimizer function $\delta(.)$ in Algorithm 1 corresponds to a special M-estimator, we mainly study Huber M-estimator, L1-L2 M-estimator, and Welsch M-estimator for Algorithm 1. When Huber M-estimator is used, the thresholding parameter in $\delta(.)$ is set to $\mu\lambda$. Then Algorithm 1 becomes the accelerated proximal gradient algorithm in [1]. For vision applications, the parameter σ^2 in Welsch and L1-L2 M-estimator is estimated by the robust mean value [38], i.e., $\sigma^2 = mean_{i,j}((G_k^E)^2)$. For succinct notation, we denote GAPG+Huber, GAPG+L1-L2, and GAPG+Welsch by Huber, L1-L2, and Welsch respectively.

4.1 Simulation Results

4.1.1 Simulation Conditions

As suggested in [12], [13], [43], random matrices are generated to quantitatively evaluate different M-estimators. Without loss of generality and for simplicity, we assume that the unknown matrix $A \in R^{m\times m}$ is square [12]. The ordered pair $(A_0, E_0) \in R^{m\times m} \times R^{m\times m}$ denotes the true solution. And the observation matrix $D = A_0 + E_0$ is the input to all algorithms, and the ordered pair $(\hat{A}, \hat{E})$ denotes the output. The matrix A_0 is generated as a product UV^T according to the random orthogonal model of rank r [12]. The matrices U and V are independent $m \times r$ matrices whose elements are i.i.d. Gaussian random variables with zero mean and unit variance. To simulate outliers, we generate error matrix E_0 as a matrix whose zero elements are chosen uniformly at random and non-zero elements are i.i.d. uniformly in the interval $[-500, 500]$. The distributions of A_0 and E_0 are identical to those used in [12], [13], [43]. The maximum iteration number of our GAPG algorithm is set to 500. All of these simulations are averaged over 20 runs. To achieve the best recovery ability, we tune parameters of the Huber M-estimator based algorithm for each experimental setting, i.e., λ, $\overline{\mu}$, θ, μ_0, and η. And for the remaining two M-estimators, we set $\sigma^2 = 200 \times median_{i,j}((G_k^E)^2)$.

We use the reconstruction error to quantitatively evaluate different methods under different levels of corruptions. The reconstruction error is computed by $\rho_c \doteq \|\hat{A} - A_0\|_F / \|A_0\|_F$ and the level of corruptions is defined as $\rho_e \doteq \|E_0\|_0 / m^2$.

4.1.2 Recovery of Low-Rank Matrix

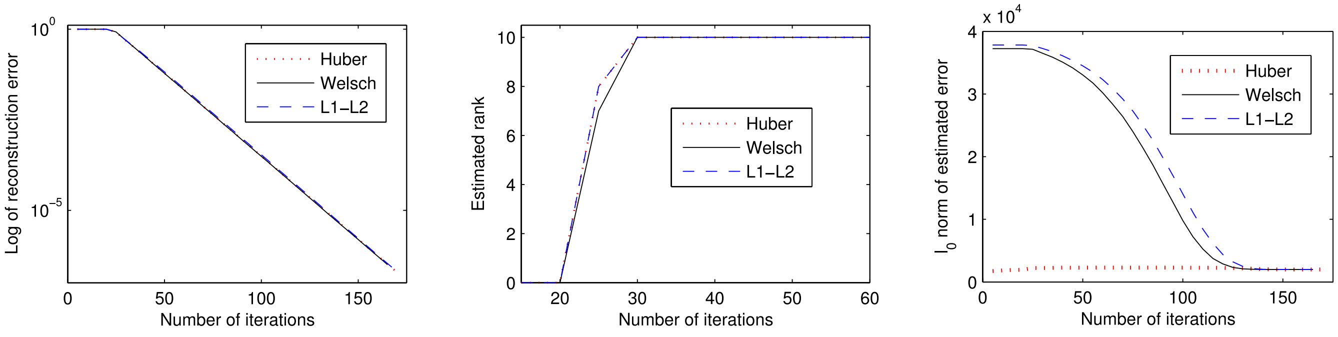
Table 2 shows experimental results of the three M-estimators based algorithms with respect to different dimensions (m), different ranks ($rank(A_0)$) and different levels of corruptions ($\rho_e \doteq \|E_0\|_0/m^2$). We see that the three methods can all obtain small reconstruction errors and find the ground truth low-rank. These results corroborate that the

[10] L1 M-estimator is not stable because $|x|$ is not strictly convex in x [20]. Although L1 M-estimator reduces the influence of large errors, these errors still have an influence because L1 M-estimator has no cut off point[10], [20].

[11] The weighing function of L1 M-estimator $|x|$ is $1/|x|$ [20].

[12] http://perception.csl.uiuc.edu/matrix-rank/sample_code.html.

[13] http://sun.stanford.edu/~rmunk/PROPACK/.



(a) Reconstruction error ρ_c

(b) $rank(\hat{A})$

(c) $||\hat{E}||_0$

Fig. 2. Comparison of the three M-estimators based algorithms when $m = 200$, $\frac{rank(A_0)}{m} = 0.05$, and $\rho_e = 0.05$. (a) Reconstruction error as a function of the number of iterations. (b) Estimated rank $rank(\hat{A})$ as a function of the number of iterations. (c) Estimated error $\|\hat{E}\|_0$ as a function of the number of iterations.

GAPG algorithm based on different M-estimators can accurately recover a corrupted low-rank matrix.

Fig. 2 further shows the variations of reconstruction errors, ranks and corrupted errors estimated by the three M-estimators based algorithms when $m = 200$, $\frac{rank(A_0)}{m} = 0.05$, $\rho_e = 0.05$. From Figs. 2a and 2b, we see that all the three methods reduce reconstruction errors and increase estimated ranks step by step. When the number of iterations is larger than 30, they all find the ground-truth rank. However, we see from Fig. 2c that they estimate error matrix E in a different way. Both L1-L2 and Welsch M-estimator based algorithms make use of the median operator to estimate outliers. In L1-L2 and Welsch M-estimator based algorithms, all entries of E will have large values due to the poor reconstruction of low-rank matrix on the first several iterations. Since M-estimators are robust to outliers, the two algorithms finally estimate real outliers when they converge.

4.1.3 Different Levels of Corruptions

This subsection is to evaluate the performance of different M-estimators based algorithms under different levels of corruptions in terms of ρ_c, $rank(\hat{A})$, $\|\hat{E}\|_0$, and the number of iterations. Fig. 3 shows experimental results of the three algorithms.

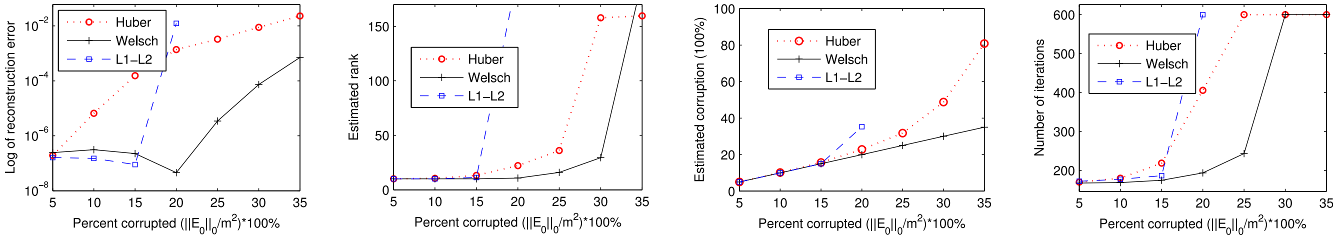
Figs. 3a and 3b show reconstruction errors and estimated ranks respectively. We see that the reconstruction errors of all methods increase as the level of corruptions increases. When the level of corruptions is larger than 20 percent, all methods fail to estimate the ground-truth rank, i.e., $rank(A_0)$. It is interesting to observe that the Welsch M-estimator based method achieves the lowest reconstruction error when the level of corruption is 20 percent. This is because the kernel size in Welsch M-estimator controls its robustness [22] and is directly related to the level of corruptions [38], [6]. Hence this phenomenon shows that this parameter setting of Welsch M-estimator is the best for 20 percent corruption.

Fig. 3c shows estimated corruption $((\|\hat{E}\|_0/m^2) \times 100\%)$ as a function of the level of corruptions. We see that the Huber M-estimator based methods estimate 80 percent entries of matrix D as outliers when the level of corruptions is 35 percent, whereas the Welsch M-estimator based method can almost accurately estimate corrupted entries. Welsch M-estimator seems to be more effective to control large outliers than the other two M-estimators.

Fig. 3d shows the variation of the number of iterations as a function of the level of corruptions. We see that the number of iterations increases significantly as the level of corruptions becomes larger. The reason for this increase may be that the three methods need more iterations to estimate the errors incurred by corruptions.

4.2 Background Modeling from Video

One natural application of low-rank matrix recovery is video analysis due to the correlation between video frames [12]. Background modeling and foreground detection [44] are two of the most basic algorithmic tasks in video



(a) $\frac{||\hat{A}-A_0||_F}{||A_0||_F}$

(b) $rank(\hat{A})$

(c) $(||\hat{E}||_0/m^2) \times 100\%$

(d) Iterations

Fig. 3. Comparison of the three M-estimators based algorithms under different levels of corruptions when $m = 200$, $\frac{rank(A_0)}{m} = 0.05$. Since the method based on L1-L2 M-estimator fails to find the ground-truth outliers when the level of corruptions is larger than 20 percent in this case, we only report the results of L1-L2 M-estimator when the level of corruptions is smaller than 20 percent.

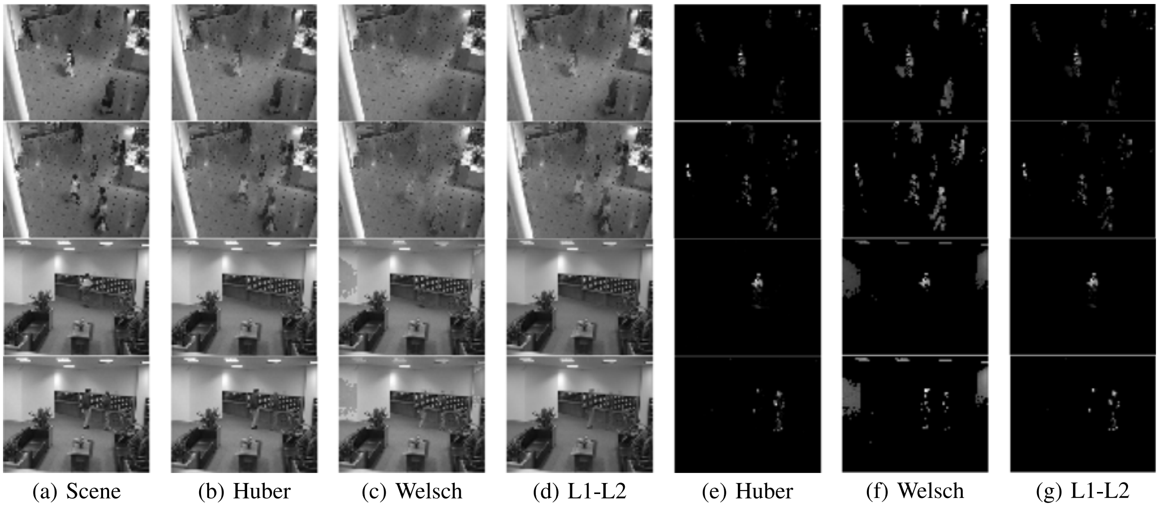


Fig. 4. Background modeling from video [8], [12]. (a) Two frames in first and second rows taken in an airport scene; two frames in third and fourth rows taken in a lobby scene with changing illumination [44]. (b)-(d) Low-rank matrix $\hat{A}$ obtained by Huber, Welsch and L1-L2 M-estimator based GAPGs respectively. (e)-(g) Sparse components obtained by Huber, Welsch and L1-L2 M-estimator based methods respectively.

analysis. The tests are performed on two example videos introduced in [44], [12], [8]. (1) In the first airport scene,[14] a total of 200 grayscale frames were used and the size of each frame is 64×80. And so D is a $5,120 \times 200$ matrix. This video sequence has a relatively static background, but significant foreground variations [8]. The first and second row in Fig. 4a show two frames from this video. (2) The second scene includes a sequence of 550 grayscale frames taken in the lobby scene with drastic illumination changes.[15] The size of each frame is 64×80, and so D is a $5,120 \times 550$ matrix. The third and fourth rows in Fig. 4a show two frames from this scene.

Figs. 4b, 4c, and 4d show the low-rank $\hat{A}$ obtained by Huber, Welsch and L1-L2 M-estimator based methods respectively. And Figs. 4e, 4f, and 4g show the foreground components obtained by Huber, Welsch and L1-L2 M-estimator based methods respectively. Since foreground variation is significantly different from static background, the pixels corresponding to the foreground can be treated as outliers. As a result, all methods can estimate low-rank components and separate foregrounds from the background. For the airport scene, Welsch M-estimator seems to outperform the other two methods. Since foreground variations in Fig. 4a are not sparse, the Huber M-estimator based method only estimates some parts of the foreground.

For the lobby scene, the low-rank components learned by Huber and L1-L2 M-estimator based methods correctly identify the main illuminations as background and the sparse part corresponding to the motion in this scene. On the other hand, the results produced by the Welsch M-estimator based method treat some of the illumination as foreground, which is the same as that produced by robust estimator based PCA in [4]. This inaccurate estimation is due to the unique setting of the kernel size σ^2 in Welsch M-estimator for fair comparison, though well-tuned parameter could further improve the accuracy of estimation [6], [22]. We see that there are two variations in this lobby scene. One is human motion and the other is illumination change. The current kernel size makes Welsch M-estimator treat both of the two variations as outliers. An appropriate selection of the kernel size is important for Welsch M-estimator to detect outliers. Fig. 5 further shows the variation of estimated errors (or foreground) in this scene. As illustrated in Fig. 2c, the L1-L2 and Welsch M-estimator based methods perform significantly different from the Huber M-estimator based method whereas they can all find the ground-truth foreground.

Fig. 5 also shows that all compared methods reach the maximum iteration number. We consider this phenomenon as an agreement with these experiments. This is because that the number of iterations for real scenes is typically higher than that in the simulation results in Section 4.1.2 [8]. In addition, the level of corruptions is often larger than 10 percent in real scenes. As shown in Fig. 3d, the number of iterations for low-rank matrix recovery often becomes larger as the level of corruption increases.

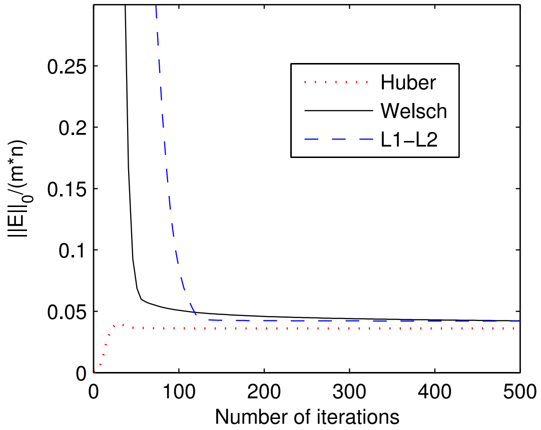


Fig. 5. Estimated error (or foreground) $\|E\|_0/m^2$ as a function of the number of iterations.

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

TABLE 3
Comparison of the Three M-Estimators Based Algorithms for Face Recognition

Method	Error rate	Std	Min	Max
PCA	13.8%	1.7%	11.2%	15.8%
Huber+PCA	12.5%	1.9%	9.2%	15.1%
Welsch+PCA	12.2%	1.8%	9.2%	15.1%
L1-L2+PCA	12.3%	1.6%	9.2%	14.5%

A lower average error rate means a better reconstruction ability.

4.3 Face Reconstruction: Removing Shadows and Specularities

Another application of low-rank matrix recovery is face reconstruction [8], [12]. It has been demonstrated that variation of many face images under variable lighting can be effectively modeled by low dimensional linear spaces [45]. Under certain ideal circumstances, aligned facial images of the same subject approximately exist in a nine dimensional linear subspace [12]. However, in face recognition, face images often suffer from self-shadowing or saturations in brightness under directional illumination [12]. As illustrated in Table 3, although PCA can deal with small Gaussian noise, it fails to deal with the errors incurred by shadows and specularities.

In this subsection, we evaluate the reconstruction ability of different M-estimators on the Extended Yale B database [45]. A total of 31 different illuminations are used for each person. All facial images are cropped and aligned to 96×84 pixels according to two eyes' positions. The matrix D contains well-aligned training images of a person's face under various illumination conditions. Fig. 6 shows the results of all methods on face images. We see that all methods can detect specularities in the eyes and shadows around the nose region. It seems that the Huber method obtains better reconstruction results in Fig. 6b and the Welsch method detects corrupted regions more accurately in Fig. 6f.

Facial image restoration is an important step in a face recognition system. Here we make use of GAPGs as a preprocessing step and then perform classification to quantitatively evaluate the reconstruction ability of different GAPGs. We randomly divided 31 different face images of one subject into two subsets. One for training contains 27 face images per subject, and the other for testing contains four face images per subject. We performed GAPGs for each individual on the training set and projected all recovered data into the subspace learned by PCA. Then the nearest-neighbor classifier is used, and average error rates of different methods are reported. A lower average error rate means a better reconstruction ability.

Table 3 shows the statistical results of the compared four methods. Although PCA can deal with small Gaussian noise in face images, it fails to deal with the errors incurred by shadows and specularities. Hence it obtains the highest average error rate. The error rates of the three M-estimators based methods are 90, 89, and 88 percent of that of the PCA method respectively, which suggests that all the three methods can efficiently deal with shadows and specularities. In addition, error rates of the three M-estimators based methods are close to those of HQ-SVTs in [17]. This is because their objective functions are based on the same M-estimators albeit different optimization methods are adopted. The experimental results suggest that low-rank matrix recovery methods are effective as a preprocessing tool for face recognition.

4.4 Gait Recognition

Gait is a commonly used behavioral biometric to recognize persons at a distance [46], [47]. The representation of human gait in real-world surveillance often gets out of control due to variations of viewing angles or carrying conditions. Recently, Zheng et al. [48] showed that there is a low-rank subspace of view transformation model computed to transform the gait features in probe viewing angle to those in gallery viewing angles.

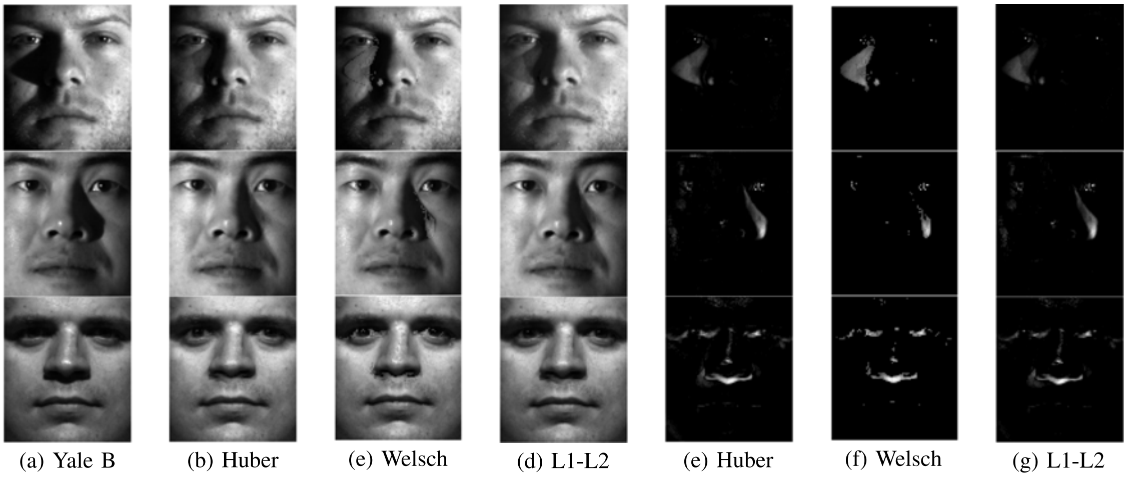
In this experiment, low-rank matrix recovery methods are used to deal with the variations of viewing angles and carrying conditions in gait recognition. The CASIA Gait database[16] is used to evaluate gait recognition accuracy as it is one of the most widely used data set in the recent literature. This database is composed of human walking videos for 124 subjects, each under 11 viewpoints. For one subject under each view, there are six normal walking sequences, two bag-carrying sequences and two coat-wearing sequences. Gait energy image (GEI) [49] was constructed as gait feature descriptor. The nearest neighbor classifier is used.

To systematically evaluate the robustness of different methods, we perform tests on two subsets of the CASIA Gait database. (1) The first subset includes a total number of 4,092 GEIs of 124 individuals from the first two normal walking sequences and the first bag-carrying sequence. (2) In the second subset, the training set is the same as the subset in (1). The first testing set contains a total number of 1,364 GEIs of 124 individuals from the third normal walking sequence, and the second testing set constants a total number of 1,364 GEIs of 124 individuals from the second bag-carrying sequence. Figs. 7a and 8a show GEIs of normal walking sequence and bag-carrying sequence respectively.

In the first test, we use GAPGs to recover a low-rank matrix of a GEI and its sparse components for each individual. Fig. 7 shows the recovered low-rank matrices and sparse components on gait energy images of a normal walking sequence. Fig. 8 shows the recovered low-rank matrices and sparse components on gait energy images of a bag-carrying sequence. We observe that all three methods can find the sparse components corresponding to the bag on GEIs. In the sparse components in Fig. 7b, the Huber M-estimator based method seems to estimate more gait variations as sparse components (denoted by light-gray pixels) in error matrix E than the other two methods. For normal walking sequence, the smaller the light-gray area of sparse components is, the better a robust method is.

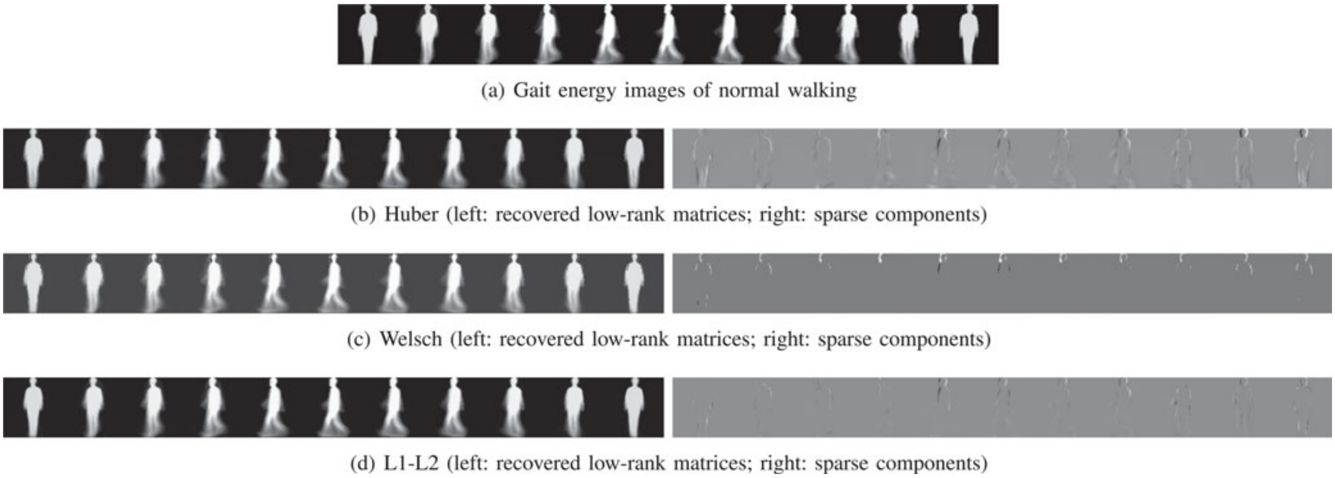
To quantitatively evaluate the reconstruction accuracy, we divide the recovered low-rank matrices into probe set and gallery set [48]. The probe set contains a total number of 1,364 recovered GEIs of 124 individuals from the bag-carrying sequence in the first test. And the gallery set includes the recovered GEIs of one view of each individual from the two normal working sequences in the first test. Then we obtain 10 gallery sets that corresponds to

[16] http://www.cbsr.ia.ac.cn/english/Gait Databases.asp.



(a) Yale B (b) Huber (e) Welsch (d) L1-L2 (e) Huber (f) Welsch (g) L1-L2

Fig. 6. Removing shadows from face images. (a) Original images of a face under different illuminations from the Extended Yale B database. (b)-(d) Low-rank matrix $\hat{A}$ obtained by Huber, Welsch and L1-L2 M-estimator based methods respectively. (e)-(g) Sparse components corresponding to specularities in the eyes and shadows around the nose region.



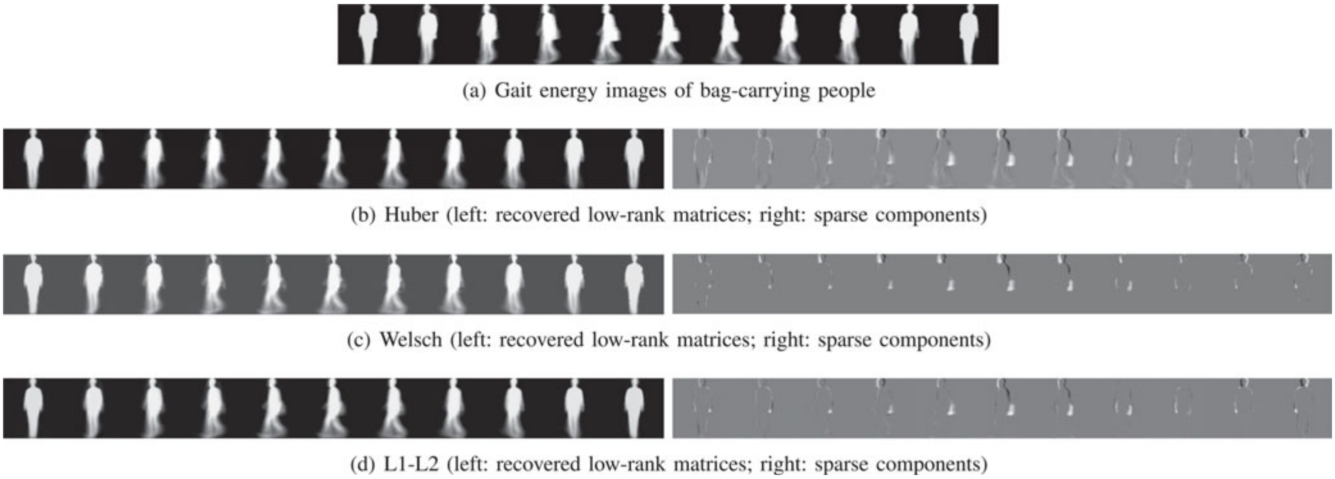
(a) Gait energy images of normal walking

(b) Huber (left: recovered low-rank matrices; right: sparse components)

(c) Welsch (left: recovered low-rank matrices; right: sparse components)

(d) L1-L2 (left: recovered low-rank matrices; right: sparse components)

Fig. 7. Recovered low-rank matrices and sparse components (light-gray pixels) on gait energy images of normal walking under 11 different viewpoints.



(a) Gait energy images of bag-carrying people

(b) Huber (left: recovered low-rank matrices; right: sparse components)

(c) Welsch (left: recovered low-rank matrices; right: sparse components)

(d) L1-L2 (left: recovered low-rank matrices; right: sparse components)

Fig. 8. Recovered low-rank matrices and sparse components (light-gray pixels) corresponding to the bag on gait energy images of bag-carrying people.

10 different views. Each gallery set contains 248 recovered GEIs. Fig. 9 shows the recognition rates of different methods under different views. We observe that the GEI around 90 degree in Fig. 7a is significantly different from that in Fig. 8a in the case of carrying bag. It seems that carrying bag leads to more difference on the GEIs around

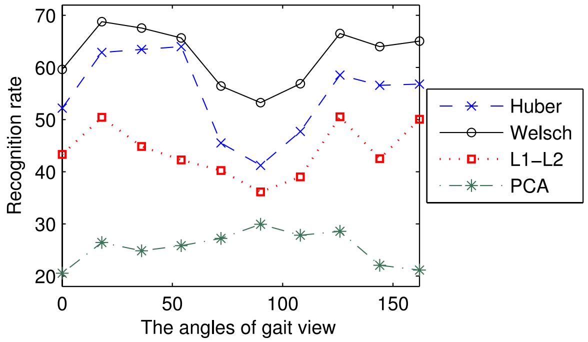


Fig. 9. Recognition rates of different methods under different views. The 10 angles (from 0 degree to 162 degree at interval 18 degree) correspond to the first 10 GEIs in Fig. 7.

90 degree than those around other degrees such that the GEIs around 90 degree in the probe set are significantly different from those around 90 degree in the gallery set. As a result, all methods obtain relatively lower recognition rates around 90 degree. We also observe that the methods can be ordered in the ascending recognition rates as PCA, L1-L2, Huber, and Welsch. The Welsch M-estimator based method achieves the highest recognition rate. This is because it can accurately reconstruct GEIs in both the gallery and probe sets. As shown in Figs. 7 and 8, Welsch M-estimator can keep useful information in A and estimate bags in error matrix E. These results demonstrate that low-rank matrix recovery methods can accurately estimate the errors incurred by bag-carrying.

In the second test, we perform PCA on the whole recovered low-rank matrices to learn a subspace. Then we project the training data and the testing data into this subspace. Table 4 shows the error rates of different methods. As expected, three low-rank methods all perform better than PCA. Since there are bag-carrying sequences in both training and testing sets in this test, the recognition rates of all methods are very close and higher than those in Fig. 9.

TABLE 4
The Error Rates of Different Methods for Normal Walking Sequence and Bag-Carrying Sequence

	Huber+PCA	Welsch+PCA	L1-L2+PCA	PCA
norm	3.1%	2.9%	3.1%	3.9%
bag	11.1%	10.8%	10.8%	11.3%

4.5 Discussion

ℓ_1 regularizer and M-estimators. Experimental results on the simulated and real-world data sets demonstrate that all the compared M-estimators can be used to recover corrupted low-rank data, and further verify the relationship between the absolute function in ℓ_1 regularizer and Huber M-estimator in Section 3.3. Since M-estimators have similar properties as shown in Fig. 1, all the methods can detect and correct corrupted errors if they are significantly different from uncorrupted data. In some applications, L1-L2 and Welsch M-estimators based methods even achieve higher accuracy than ℓ_1 regularizer based one (Huber M-estimator). This may be because the two M-estimators make use of an adaptive and robust way to estimate their parameters. As shown in Fig. 1, if parameters of M-estimators are well tuned, all M-estimators can have similar robustness.

Kernel size and threshold parameters. Experimental results also show that without sparsity assumption, both L1-L2 and Welsch M-estimators based methods can estimate sparse errors. However, they perform in a different way to estimate errors compared with Huber M-estimator. As discussed in correntropy, the kernel size based M-estimator runs in a different way against the threshold based one [22]. Since outliers are those data points that are significantly different from other data points, both the kernel size and threshold based methods can find the ground truth outliers due to the robustness of M-estimators. The analysis between kernel size and threshold based methods suggests that the soft-threshold based methods are more preferable for sparse errors and the kernel size based methods are more preferable for dense errors.

5 Conclusion and Future Work

This paper has studied the low-rank matrix recovery problem from the viewpoint of implicit regularizers which are derived from conjugated functions. It provides a unified view to analyze recent ℓ_1 regularization based and traditional M-estimators based robust PCAs. It also gives an M-estimation explanation of the robustness of robust low-rank matrix recovery methods. Moreover, our study enriches the family of regularizers for robust learning. Based on proximity operators of implicit regularizers, a robust framework is developed for robust low-rank matrix recovery, which is solved by a generalized accelerated proximal gradient algorithm. A series of experiments on simulations and real-world applications have verified the robustness of the proposed framework.

Recently, iteratively reweighted least squares methods have been developed to solve nuclear norm minimization [34] and trace norm minimization [50] where the used weighting functions ([34, p. 6]) are just the same multiplicative minimizer functions of half-quadratic minimization [21]. It has been shown in [51] that IRLS for some certain functions is a special case of HQ. Hence, future work is to study the nuclear norm (or trace norm) minimization problem from the two forms of HQ.

Acknowledgments

The authors would like to greatly thank the associate editor and the reviewers for their valuable comments and suggestions, and thank Dr. Xiao-tong Yuan, Dr. Wei-shi Zheng, Dr. Jing Dong, Shuai Zheng and Haiqing Li for designing the experiments and helping revising the paper. This work is funded by National Basic Research Program of China (Grant No. 2012CB316300), National Natural Science Foundation of China (Grant No. 61103155,61175003), and the Strategic Priority Research Program of the Chinese Academy of Sciences (Grant No. XDA06030300).

References

[1] Z. Lin, A. Ganesh, J. Wright, L. Wu, M. Chen, and Y. Ma, "Fast Convex Optimization Algorithms for Exact Recovery of a Corrupted Low-Rank Matrix," UIUC Technical Report UILU-ENG-09-2214, 2009.

[2] C. Ding, D. Zhou, X. He, and H. Zha, "R1-PCA: Rotational Invariant l1-Norm Principal Component Analysis for Robust Subspace Factorization," *Proc. Int'l Conf. Machine Learning*, pp. 281-288, 2006.

[3] C.M. Bishop, *Patten Recognition and Machine Learning*. Springer, 2006.

[4] F. De La Torre and M. Black, "A Framework for Robust Subspace Learning," *Int'l J. Computer Vision*, vol. 54, no. 1-3, pp. 117-142, 2003.

[5] R.A. Maronna, "Principal Components and Orthogonal Regression Based on Robust Scales," *Technometrics*, vol. 47, pp. 264-273, 2005.

[6] R. He, B.-G. Hu, W.-S. Zheng, and X.W. Kong, "Robust Principal Component Analysis Based on Maximum Correntropy Criterion," *IEEE Trans. Image Processing*, vol. 20, no. 6, pp. 1485-1494, June 2011.

[7] X. Wang and X. Tang, "A Unified Framework for Subspace Face Recognition," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 26, no. 9, pp. 1222-1228, Sept. 2004.

[8] E.J. Candés, X. Li, Y. Ma, and J. Wright, "Robust Principal Component Analysis?" *J. ACM*, vol. 58, no. 3, article 11, 2011.

[9] J. Iglesias, M. de Bruijne, M. Loog, F. Lauze, and M. Nielsen, "A Family of Principal Component Analyses for Dealing with Outliers," *Proc. 10th Int'l Conf. Medical Image Computing & Computer-Assisted Intervention*, pp. 178-185, 2007.

[10] P. Huber, *Robust Statistics*. John Wiley & Sons, 1981.

[11] J. Cai, E.J. Candés, and Z. Shen, "A Singular Value Thresholding Algorithm for Matrix Completion," *SIAM J. Optimization*, vol. 20, no. 4, pp. 1956-1982, 2010.

[12] J. Wright, A. Ganesh, S. Rao, and Y. Ma, "Robust Principal Component Analysis: Exact Recovery of Corrupted Low-Rank Matrices via Convex Optimization," *J. ACM*, vol. 4, pp. 1-44, 2009.

[13] Z. Lin, M. Chen, L. Wu, and Y. Ma, "The Augmented Lagrange Multiplier Method for Exact Recovery of Corrupted Low-Rank Matrices," UIUC Technical Report UILU-ENG-09-2215, 2009.

[14] A. Ganesh, J. Wright, X. Li, E.J. Candés, and Y. Ma, "Dense Error Correction for Low-Rank Matrices via Principal Component Pursuit," *The Computing Research Repository*, 2010.

[15] G. Liu, Z. Lin, S. Yan, J. Sun, Y. Yu, and Y. Ma, "Robust Recovery of Subspace Structures by Low-Rank Representation," *The Computing Research Repository*, 2010.

[16] Y. Mu, J. Dong, X. Yuan, and S. Yan, "Accelerated Low-Rank Visual Recovery by Random Projection," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pp. 2609-2616, 2011.

[17] R. He, Z. Sun, T. Tan, and W.-S. Zheng, "Recovery of Corrupted Low-Rank Matrices via Half-Quadratic Based Nonconvex Minimization," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2011.

[18] P.L. Combettes and V.R. Wajs, "Signal Recovery by Proximal Forwardbackward Splitting," *SIAM J. Multiscale Modeling & Simulation*, vol. 4, no. 5, pp. 1168-1200, 2005.

[19] J.M. Bioucas-Dias and M.A.T. Figueiredo, "A New Twist: Two-Step Iterative Shrinkage/Thresholding Algorithms for Image Restoration," *IEEE Trans. Image Processing*, vol. 16, no. 12, pp. 2992-3004, Dec. 2007.

[20] Z. Zhang, "Parameter Estimation Techniques: A Tutorial with Application to Conic Fitting," *Image and Vision Computing*, vol. 15, no. 1, pp. 59-76, 1997.

[21] M. Nikolova and M.K. NG, "Analysis of Half-Quadratic Minimization Methods for Signal and Image Recovery," *SIAM J. Scientific Computing*, vol. 27, no. 3, pp. 937-966, 2005.

[22] W.F. Liu, P.P. Pokharel, and J.C. Principe, "Correntropy: Properties and Applications in Non-Gaussian Signal Processing," *IEEE Trans. Signal Processing*, vol. 55, no. 11, pp. 5286-5298, Nov. 2007.

[23] M.V. Afonso, J.M. Bioucas-Dias, and M.A.T. Figueiredo, "An Augmented Lagrangian Approach to Linear Inverse Problems with Compound Regularization," *Proc. Int'l Conf. Image Processing*, pp. 4169-4172, 2010.

[24] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge Univ. Press, 2004.

[25] J.-J. Moreau, "Fonctions Convexes Duales et Points Proximaux dans un Espace Hilbertien," *Comptes Rendus de l'Académie des Sciences. Série I. Mathématique. Académie des Sciences*, vol. 255, pp. 2897-2899, 1962.

[26] J. Wright, Y. Ma, J. Mairal, G. Sapiro, T.S. Huang, and S. Yan, "Sparse Representation for Computer Vision and Pattern Recognition," *Proc. IEEE*, vol. 98, no. 6, pp. 1031-1044, June 2010.

[27] P. Gong, C. Zhang, Z. Lu, J. Huang, and J. Ye, "A General Iterative Shrinkage and Thresholding Algorithm for Non-Convex Regularized Optimization Problems," *Proc. Int'l Conf. Machine Learning*, 2013.

[28] E.J. Candés and T. Tao, "The Power of Convex Relaxation: Near-Optimal Matrix Completion," *IEEE Trans. Information Theory*, vol. 56, no. 5, pp. 2053-2080, May 2010.

[29] A. Singer and M. Cucuringu, "Uniqueness of Low-Rank Matrix Completion by Rigidity Theory," *SIAM J. Matrix Analysis and Applications*, vol. 31, pp. 1621-1641, 2010.

[30] M. Fazel, "Matrix Rank Minimization with Applications," PhD dissertation, Stanford Univ., 2002.

[31] R. Tibshirani, "Regression Shrinkage and Selection via the Lasso," *J. Royal Statistical Soc. B*, vol. 58, no. 1, pp. 267-288, 1996.

[32] D. Hsu, S.M. Kakade, and T. Zhang, "Robust Matrix Decomposition with Sparse Corruptions," *IEEE Trans. Information Theory*, vol. 57, no. 11, pp. 7221-7234, Nov. 2011.

[33] A. Ganesh, Z. Lin, J. Wright, L. Wu, M. Chen, and Y. Ma, "Fast Algorithms for Recovering a Corrupted Low-Rank Matrix," *Proc. Int'l Workshop Computational Advances in Multi-Sensor Adaptive Processing*, 2008.

[34] M. Fornasier, H. Rauhut, and R. Ward, "Low Rank Matrix Recovery via Iteratively Reweighted Least Squares Minimization," submitted to *SIAM J. Optimization*, pp. 1-22, 2011.

[35] V. Chandrasekaran, S. Sanghavi, P.A. Parrilo, and A.S. Willsky, "Sparse and Low-Rank Matrix Decompositions," *Proc. IEEE Conf. Comm., Control, and Computing*, pp. 962-967, 2009.

[36] D. Gross, "Recovering Low-Rank Matrices from Few Coefficients in Any Basis," *IEEE Trans. Information Theory*, vol. 57, no. 3, pp. 1548-1566, Mar. 2011.

[37] C.A. Micchelli, J.M. Morales, and M. Pontil, "A Family of Penalty Functions for Structured Sparsity," *Proc. Advances in Neural Information Processing Systems*, pp. 1-9, 2010.

[38] R. He, W.-S. Zheng, and B.-G. Hu, "Maximum Correntropy Criterion for Robust Face Recognition," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 33, no. 8, pp. 1561-1576, Aug. 2011.

[39] R. Angst, C. Zach, and M. Pollefeys, "The Generalized Trace Norm and Its Application to Structure from Motion Problems," *Proc. IEEE Int'l Conf. Computer Vision*, pp. 2502-2509, 2011.

[40] L. Yuan, J. Liu, and J. Ye, "Efficient Methods for Overlapping Group Lasso," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 35, no. 9, pp. 2104-2116, Sept. 2013.

[41] J.M. Bioucas-Dias and M.A.T. Figueiredo, "An Iterative Algorithm for Linear Inverse Problems with Compound Regularizers," *Proc. Int'l Conf. Image Processing*, pp. 685-688, 2008.

[42] A. Beck and M. Teboulle, "A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems," *SIAM J. Imaging Sciences*, vol. 2, no. 1, pp. 183-202, 2009.

[43] K.-C. Toh and S. Yun, "An Accelerated Proximal Gradient Algorithm for Nuclear Norm Regularized Least Squares Problems," *Pacific J. Optimization*, vol. 6, pp. 615-640, 2010.

[44] L. Li, W. Huang, I. Gu, and Q. Tian, "Statistical Modeling of Complex Backgrounds for Foreground Object Detection," *IEEE Trans. Image Processing*, vol. 13, no. 11, pp. 1459-1472, Nov. 2004.

[45] K.-C. Lee, J. Ho, and D. Kriegman, "Acquiring Linear Subspaces for Face Recognition under Variable Lighting," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 27, no. 5, pp. 684-698, May 2005.

[46] L. Wang, T. Tan, H. Ning, and W. Hu, "Silhoutte Analysis Based Gait Recognition for Human Identification," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 25, no. 12, pp. 1505-1518, Dec. 2003.

[47] S. Sarkar, P.J. Phillips, Z. Liu, I.R. Vega, P. Grother, and K.W. Bowyer, "The Humanid Gait Challenge Problem: Data Sets, Performance, and Analysis," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 27, no. 2, pp. 162-177, Feb. 2005.

[48] S. Zheng, K. Huang, T. Tan, R. He, and J. Zhang, "Robust View Transformation Model for Gait Recognition," *Proc. Int'l Conf. Image Processing*, 2011.

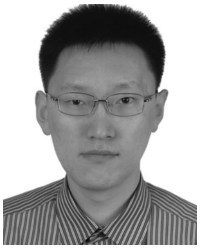
[49] J. Han and B. Bhanu, "Individual Recognition Using Gait Energy Image," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 28, no. 2, pp. 316-322, Feb. 2006.

[50] E. Grave, G. Obozinski, and F. Bach, "Trace Lasso: A Trace Norm Regularization for Correlated Designs," *Proc. Advances in Neural Information Processing Systems (NIPS)*, 2012.

[51] J. Idier, "Convex Half-Quadratic Criteria and Interacting Auxiliary Variables for Image Restoration," *IEEE Trans. Image Processing*, vol. 10, no. 7, pp. 1001-1009, July 2001.

[52] W.F. Liu, P.P. Pokharel, and J.C. Principe, "Error Entropy, Correntropy and M-Estimation," *Proc. IEEE Workshop Machine Learning for Signal Processing*, pp. 179-184, 2006.

[53] S. Seth and J. Principe, "Compressed Signal Reconstruction Using the Correntropy Induced Metric," *Proc. Int'l Conf. Acoustics, Speech, and Signal Processing*, 2008.



Ran He received the BE and ME degrees in computer science from Dalian University of Technology, and the PhD degree in pattern recognition and intelligent systems from the Institute of Automation, Chinese Academy of Sciences in 2001, 2004, and 2009, respectively. Since September 2010, he has joined NLPR where he is currently an associate professor. He is currently a member of the IEEE, serves as an associate editor of *Neurocomputing* (Elsevier), and serves on the program committee of several conferences. His research interests focus on information theoretic learning, pattern recognition, and computer vision.

Tieniu Tan received the BSc degree in electronic engineering from Xi'an Jiaotong University, China, in 1984, and the MSc and PhD degrees in electronic engineering from Imperial College London, United Kingdom, in 1986 and 1989, respectively. He was the director general of the CAS Institute of Automation from 2000 to 2007, and has been a professor and the director of the NLPR since 1998. He also serves as a deputy secretary-general (for cyber-infrastructure and international affairs) of the CAS. He has published more than 300 research papers in refereed journals and conferences in the areas of image processing, computer vision and pattern recognition, and has authored or edited nine books. He holds more than 30 patents. His current research interests include biometrics, image and video understanding, and information forensics and security. He has served as chair or program committee member for many major national and international conferences. He currently serves as a vice president of the IAPR, the executive vice president of the Chinese Society of Image and Graphics, the deputy president of the Chinese Association for Artificial Intelligence, and was the deputy president of the China Computer Federation and the Chinese Automation Association. He has given invited talks and keynotes at many universities and international conferences, and has received numerous national and international awards and recognitions. He is a fellow of the IEEE and the International Association of Pattern Recognition (IAPR).

Liang Wang received the BEng and MEng degrees from Anhui University in 1997 and 2000, respectively, and the PhD degree from the Institute of Automation, Chinese Academy of Sciences (CAS), Beijing, China, in 2004. From 2004 to 2010, he was a research assistant with Imperial College London, London, United Kingdom, and Monash University, Australia, a research fellow with the University of Melbourne, Australia, and a lecturer with the University of Bath, United Kingdom, respectively. Currently, he is a professor of Hundred Talents Program at the Center for Research on Intelligent Perception and Computing (CRIPAC) and the National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences, Beijing, China. His major research interests include machine learning, pattern recognition, and computer vision. He has widely published in highly ranked international journals and leading international conferences. He is an associate editor for the *International Journal of Image and Graphics, Signal Processing, Neurocomputing* and *IEEE Transactions on Systems, Man and Cybernetics-Part B*. He received the Special Prize of the Presidential Scholarship of Chinese Academy of Sciences. He has been a guest editor for seven special issues, a coeditor of five edited books, and a cochair of 10 international workshops. He is a member of BMVA and a senior member of the IEEE.