# Half-quadratic based Iterative Minimization for Robust Sparse Representation

Ran He, *Member, IEEE*, Wei-Shi Zheng, *Member, IEEE*, Tieniu Tan, *Fellow, IEEE*, Zhenan Sun, *Member, IEEE*,

***Abstract*— Robust sparse representation has shown significant potential in solving challenging problems in computer vision such as biometrics and visual surveillance. Although several robust sparse models have been proposed and promising results have been obtained, they are either for error correction or for error detection, and learning a general framework that systematically unifies these two aspects and explore their relation is still an open problem. In this paper, we develop a half-quadratic (HQ) framework to solve the robust sparse representation problem. By defining different kinds of half-quadratic functions, the proposed HQ framework is applicable to performing both error correction and error detection. More specifically, by using the additive form of HQ, we propose an $\ell_1$-regularized error correction method by iteratively recovering corrupted data from errors incurred by noises and outliers; by using the multiplicative form of HQ, we propose an $\ell_1$-regularized error detection method by learning from uncorrupted data iteratively. We also show that the $\ell_1$-regularization solved by soft-thresholding function has a dual relationship to Huber M-estimator, which theoretically guarantees the performance of robust sparse representation in terms of M-estimation. Experiments on robust face recognition under severe occlusion and corruption validate our framework and findings.**

***Index Terms*— $\ell_1$-Minimization, Half-quadratic Optimization, Sparse Representation, M-estimator, Correntropy.**

## I. Introduction

Sparse signal representation arises in application of compressed sensing and has been considered as a significant technique in computer vision and machine learning [1][2][3]. Based on the $\ell_0$-$\ell_1$ equivalence theory [4][5], the solution of an $\ell_0$ minimization problem is equal to that of an $\ell_1$ minimization problem under certain conditions. Sparse representation has been widely applied in image analysis [6][7], compressive imaging [8][9], multi-sensor networks [10], and subspace segmentation [11]. Recent theoretical analysis [12] and experimental results [13] show that even if corruptions are high, one can almost recover corrupted data using $\ell_1$-based techniques. So far, all the sparse representation algorithms can be basically categorized into two major categories: error correction [12][13][14] and error detection [3][15][16]. The former aims to reconstruct the original data during robust learning, while the latter detects errors and learns from uncorrupted data. However, the theoretical support that $\ell_1$ regularization tends to achieve robustness still needs to be further studied [17][18]. More importantly, it is still an open issue to unify these two approaches in a general framework and study their intrinsic relation.

### A. Related Work

*1) Sparse Representation:* Given a predefined sample set $X \doteq [x_1, x_2, \ldots, x_n] \in \mathbb{R}^{d\times n}$ and an input sample $y \in \mathbb{R}^{d\times 1}$, where $n$ is the number of elements in $X$ and $d$ is the feature dimension, the sparse representation problem can be formulated as the following minimization problem [14],

$$\min_{\beta} ||\beta||_1 \quad s.t. \quad X\beta = y, \tag{1}$$

where $||\beta||_1 = \sum_{i=1}^{n} |\beta_i|$. Considering that a white noise $z$ satisfies $||z||_2 \le \varepsilon$, we can relax the equality constraint $X\beta = y$ in the following form,

$$y = X\beta + z. \tag{2}$$

Then the sparse representation $\beta$ can be computed via basis pursuit denoising (BPDN) method [19][20],

$$\min_{\beta} ||\beta||_1 \quad s.t. \quad ||y - X\beta||_2 \le \varepsilon. \tag{3}$$

By using the Lagrangian method, one can rewrite (3) as an unconstrained optimization problem [13],

$$\min_{\beta} \tfrac{1}{2}||y - X\beta||_2^2 + \lambda||\beta||_1, \tag{4}$$

where $\lambda$ is a positive regularization parameter. Iteratively reweighted methods, such as adaptive lasso [21], reweighted $\ell_1$ minimization [22] and multi-stage convex relaxation [23], are further developed to enhance sparsity for high-dimensional data. And various numerical methods [24][25] have been developed to minimize (3) or (4), where the iterative regularization method based on soft-shrinkage operator [26] is often used.

*2) Error Correction:* In robust statistics [27] and computer vision [1], the errors incurred by corruptions or occlusions may be arbitrarily large. Hence, one often addresses the following robust model [27],

$$y = X\beta + e + z, \tag{5}$$

where $e$ is a variable describing outliers that are intrinsically different from uncorrupted data. A number of algorithms have been developed to deal with outliers in (5) [1][3][13]. They are actually either for error correction or for error detection. The algorithms for error correction are mainly for recovering the groundtruth from corrupted data. One representative algorithm in the context of robust face recognition was proposed by Wright et al. [14], which assumes that the error $e$ has a sparse representation and seeks the sparsest solution via solving the following problem:

$$\min_{\beta,e} ||X\beta + e - y||_2^2 + \lambda(||\beta||_1 + ||e||_1), \tag{6}$$

where $e \in \mathbb{R}^{d\times 1}$ is an unknown error vector whose nonzero entries correspond to outliers. In [12][14], (6) is often solved by,

$$\min_{\omega} ||B\omega - y||_2^2 + \lambda||\omega||_1, \tag{7}$$

Manuscript received April 22, 2011

Ran He, Tieniu Tan, and Zhenan Sun are with Center for Research on Intelligent Perception and Computing (CRIPAC) and National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences, Beijing, China 100190. E-mail: {rhe, tnt, znsun}@nlpr.ia.ac.cn.

W.-S. Zheng is with the School of Information Science and Technology, Sun Yat-sen University, Guangzhou, China 510275. E-mail: wszheng@ieee.org.

Digital Object Indentifier 10.1109/TPAMI.2013.102

where $\omega = [\beta^T, e^T]^T \in \mathbb{R}^{(d+n)\times 1}$ and $B = [X, I] \in \mathbb{R}^{d\times(n+d)}$ ($I$ is the identity matrix). The algorithms to solve (7) estimate error as a vector variable and correct it in each iteration. And the $\ell_1$ norm is an approximation of $\ell_0$ norm in order to obtain a sparse solution. Recent analysis and experimental investigations in [12][13] show that the same $\ell_1$-minimization algorithm can be used to recover corrupted data even if the level of corruption is almost arbitrarily large.

*3) Error Detection:* The algorithms for error detection are mainly for selecting the most significant pixels (or uncorrupted pixels) in an image for better recognition performance [28][29][30]. These methods are often based on robust M-estimators or assume the occlusion masks are provided, which are widely used in subspace learning and Eigen tracking. Recently, based on maximum correntropy criterion (MCC) [31] and half-quadratic (HQ) optimization [32][33], He et al. [3] extended (4) by substituting mean square error with correntropy and iteratively computed a non-negative sparse solution for robust face recognition. He et al. [15] further studied an $\ell_1$ regularized correntropy problem for robust pattern recognition, where a robust sparse representation is computed by iteratively solving a weighted $\ell_1$-minimization problem. Furthermore, Yang et al. [16] modeled robust sparse representation as the robust regression problem with a sparse constraint and proposed an iteratively reweighted least squares algorithm. Li et al. [34] developed a structured sparse error coding method for continuous occlusion based on the error detection strategy.

To the best of our knowledge, currently there is not a general framework to unify these two kinds of sparse representation approaches aforementioned and study their relationship. For example, although the sparse representation method in [1] indeed improves robustness under tough conditions (e.g. 80% corruption), the reason why it can work under such a dense error still needs to be further investigated. In addition, robust sparse analysis is still an open and hot issue in information theory [18][17][35].

## B. Contribution

In order to unify robust sparse representation methods for error correction and error detection into one framework, we first address a general robust sparse representation problem, i.e.,

$$\min_{\beta} \sum\nolimits_{j=1}^{d} \phi((X\beta - y)_j) + \lambda ||\beta||_1, \tag{8}$$

where $\phi(.)$ is a robust M-estimator and can be optimized by half-quadratic (HQ) optimization[1], and $(.)_j$ denotes the $j$-th dimension of an input vector. We will investigate a general half-quadratic framework to minimize (8). Under this framework, a robust sparse representation problem is reduced to an iterative regularization problem, which can be optimized by solving a number of unconstrained quadratic problems. Then, we show that iteratively reweighted least squares and shrink operator based $\ell_1$ iterative regularization method are its two special cases.

Second, by utilizing the additive form of HQ, an $\ell_1$-regularized error correction method is developed to iteratively recover corrupted data through estimating errors incurred by noise and outliers; by harnessing the multiplicative form of HQ, an $\ell_1$-regularized error detection method is developed through iteratively using uncorrupted data to perform learning.

Third, we investigate possible M-estimators ($\phi(.)$) to show that the shrink operator on errors can be explained as the additive form of Huber M-estimator in iterative regularization, and the variable $e$ in (6) can be viewed as an auxiliary variable of Huber M-estimator in half-quadratic minimization. When the M-estimator in (8) is Welsch M-estimator, the undetermined linear system (i.e., $X\beta = y$) becomes a correntropy [31] adaptive system, which has a probabilistic support for large level of corruptions [31]. Numerical results on robust face recognition are run to validate our claims and demonstrate that our framework is sufficient in most cases.

In summary, the novelties of our work are as follows:

1. A unified framework is proposed to investigate both error correction and error detection. The connection and difference between error correction and detection are extensively investigated.
2. A deep investigation into Huber M-estimator is presented, which shows that the absolute function (in $\ell_1$ regularizer) solved by shrink operator can be viewed as the dual function of Huber M-estimator. To the best of our knowledge, it is the first time to present such a theoretical guarantee of the robustness of the $\ell_1$ based sparse representation methods from the viewpoint of M-estimation.
3. Robust and efficient $\ell_1$-regularized methods are developed using correntropy (i.e., Welsch M-estimator), which performs better than other M-estimators in terms of robustness with lower computational cost.

The rest of the paper is organized as follows. In Section II, we revisit the HQ minimization for convex or nonconvex functions. In Section III, we develop error correction and detection methods by harnessing the additive and the multiplicative form respectively. In Section IV, we investigate various robust M-estimators and discuss their robustness from the viewpoint of M-estimation. In Section V, the proposed approaches are validated by conducting robust face recognition experiments along with the comparison with related methods. Finally, we draw the conclusion and discuss future work in Section VI.

# II. Half-Quadratic Minimization

Since our investigation is relying on the half-quadratic theory, we first review some theoretical background and half-quadratic modeling based on conjugate function theory [36][37] for convex or non-convex minimization.

## A. Conjugate Function

Given a differentiable function $f(v): \mathbb{R}^n \to \mathbb{R}$, the conjugate $f^*(p): \mathbb{R}^n \to \mathbb{R}$ of the function $f$ is defined as [38]:

$$f^*(p) = \max_{v \in dom f} (p^T v - f(v)). \tag{9}$$

The domain of $f^*(p)$ is bounded above on **dom**$f$ [38]. Since $f^*(p)$ is the pointwise supremum of a family of convex functions of $p$, it is a convex function [38]. If $f(v)$ is convex and closed, the conjugate of its conjugate function is itself, i.e., $f^{**} = f$ [38].

Based on conjugate function, a loss function in image restoration and signal recovery can be defined as [33][39][40],

$$f(v) = \min_{p} \{Q(v, p) + \varphi(p)\}, \tag{10}$$

[1] Note that $\phi(.)$ can be a convex function or a non-convex function. And not all M-estimators can be optimized by HQ. For example, absolute function cannot be optimized by HQ [33] and $|.|^{\alpha}(\alpha \in (1, 2])$ in $\ell_p$ M-estimator is not applicable in the additive form of HQ.

where $f(.)$ is a potential loss function (such as M-estimators in Table I), $v$ is a set of adjustable parameters of a linear system, $p$ is an auxiliary variable in HQ optimization, $Q(v,p)$ is a quadratic function ($Q(v,p) \doteq \sum_i p_i v_i^2$ for $p \in \mathbb{R}_+^d$ and $v \in \mathbb{R}^d$, or $Q(v,p) \doteq ||v-p||_2^2$ for $p \in \mathbb{R}^d$ and $v \in \mathbb{R}^d$), and $\varphi(.)$ is the dual potential function of $f(.)$ [2]. An example of (10) is given in (61) (Appendix III) for compressed signal recovery.

In the two-step iterative shrinkage/thresholding algorithms [13], the minimization function of (10) is also known as proximal mapping [39][40]; in half quadratic methods, the function $Q(v,p)+\varphi(p)$ is called the resultant (augmented) cost-function of $f(v)$, and can be optimized by a two-step alternating minimization way [33]. In [36][37], half-quadratic regularization is developed to minimize non-convex $f(v)$. In the following, we briefly review the HQ minimization and discuss its relationship to other methods.

### B. Half-quadratic Optimization

Let $\phi_v(.)$ be a function on a vector $v \in \mathbb{R}^d$ that is defined as

$$\phi_v(v) \doteq \sum\nolimits_{j=1}^{d} \phi(v_j), \tag{11}$$

where $\phi(.)$ is a potential loss function in HQ [33][41] and $v_j$ is the $j$-th entry of $v$. In machine learning and compressed sensing, one often aims to compute the following minimization problem:

$$\min_v \phi_v(v) + J(v), \tag{12}$$

where $J(v)$ is a convex penalty function on $v$. According to half-quadratic minimization [36][37], we know that for a fixed $v_j$, the following equation holds,

$$\phi(v_j) = \min_{p_j} Q(v_j,p_j) + \varphi(p_j), \tag{13}$$

where $\varphi(.)$ is the dual potential function of $\phi(.)$, and $Q(v_j,p_j)$ is the half quadratic function which can be modeled in the additive or the multiplicative form as shown later. Let $Q_v(v,p) \doteq \sum_{j=1}^{d} Q(v_j,p_j)$, we have the vector form of (13),

$$\phi_v(v) = \min_p Q_v(v,p) + \sum\nolimits_{j=1}^{d} \varphi(p_j). \tag{14}$$

By substituting (14) into (12), we obtain that

$$\min_v \{\phi_v(v) + J(v)\} = \min_{v,p} \{Q_v(v,p) + \sum_{j=1}^{d} \varphi(p_j) + J(v)\}, \tag{15}$$

where $p_j$ is determined by a minimization function $\delta(.)$ that is only related to $\phi(.)$ (See Table I for specific forms). In HQ optimization, $\delta(.)$ is derived from conjugate function and satisfies that $\{Q(v_j,\delta(v_j))+\varphi(\delta(v_j))\} \leq \{Q(v_j,p_j)+\varphi(p_j)\}$. Let $\delta_v(v) \doteq [\delta(v_1),\ldots,\delta(v_d)]$, and then one can alternately minimize (15) as follows,

$$p^{t+1} = \delta_v(v), \tag{16}$$

$$v^{t+1} = \arg\min_v Q_v(v,p^{t+1}) + J(v), \tag{17}$$

where $t$ indicates the $t$-th iteration. Algorithm 1 summarizes the optimization procedure. At each step, the objective function in (15) is reduced alternatingly until it converges.

**Algorithm 1**: Half-quadratic based Algorithms

```
Input: data matrix $X$, test sample $y$, and $v = \vec{0}$.
Output: $v$
while "not converged" do
    $p^{t+1} = \delta_v(v)$
    $v^{t+1} = \arg\min_v Q_v(v, p^{t+1}) + J(v)$
    $t = t + 1$
end while
```

### C. The Additive and Multiplicative Forms

In HQ minimization, the half-quadratic reformulation $Q(v_j,p_j)$ of an original cost-function has two forms [36][37]: the additive form denoted by $Q_A(v_j,p_j)$ and the multiplicative form denoted by $Q_M(v_j,p_j)$. Specifically, $Q_A(v_j,p_j)$ is formulated as [37],

$$Q_A(v_j,p_j) = (v_j\sqrt{c} - p_j/\sqrt{c})^2, \tag{18}$$

where $c$ is a constant and $c > 0$. The additive form indicates that we can expand a function $\phi(.)$ to a combination of quadratic terms and the auxiliary variable $p_j$ is related to $v_j$. During iterative minimization, the value of $v_j$ is updated and refined by $p_j$.

The multiplicative form $Q_M(v_j,p_j)$ is formulated in the form [36],

$$Q_M(v_j,p_j) = \frac{1}{2} p_j v_j^2. \tag{19}$$

It indicates that we can expand a non-convex (or convex) function $\phi(.)$ to quadratic terms of the multiplicative form. The auxiliary variable $p_j$ is introduced as a data-fidelity term. For $v_j$, $p_j$ indicates the contribution of $v_j$ to the whole data $v$.

Experimental results in [33] show that when the additive form is applicable, the number of iterations required to converge for the additive form seems to be larger than that for the multiplicative one. However, the computational cost of the additive form is much lower than that of the multiplicative one. Although the computational cost of the additive form is cheap, the additive form has not received much attention in the past decades [33].

By using different forms of quadratic functions, we are able to specifically obtain error correction or detection models as shown in Section III. Before that, we in the next section first discuss the relationship between the half-quadratic minimization with existing works.

### D. Connection to Sparse Representation Modeling

By combining the additive form or the multiplicative one, the half-quadratic framework is well connected with some existing popular sparse representation models. We specify $J(v)$ to be two convex functions (i.e., $J(.) = \lambda||.||_1$ and $J(.) = \lambda||.||_2^2$), and discuss its relationship with other algorithms.

First, by substituting $J(v) = \lambda||v||_1$ into (12), we obtain the following $\ell_1$-minimization problem:

$$\min_v \lambda||v||_1 + \frac{1}{2}\phi_v(v). \tag{20}$$

By combining (20), (18) and (14), we have

$$v^* = \arg\min_v \lambda||v||_1 + \frac{1}{2}\sum_{j=1}^{d} (v_j\sqrt{c} - p_j^{t+1}/\sqrt{c})^2. \tag{21}$$

According to the additive form of HQ [33], we have that the minimization function $\delta(v) = cv - \phi'(v)$. Then we have,

[2] Note that for different types of $Q(v,p)$, the dual potential functions $\varphi(.)$ may be different.

$$p_j^{t+1} = cv_j^t - \phi'(v_j^t). \tag{22}$$

By substituting (22) into (21), we have the iterative scheme,

$$v^{t+1} = \arg\min_v \lambda||v||_1 + \frac{1}{2}||\sqrt{c}v - (\sqrt{c}v^t - (\nabla\phi_v(v^t)/\sqrt{c}))||_2^2, \tag{23}$$

where $\nabla\phi_v(v^t) = [\phi'(v_1^t), \ldots, \phi'(v_d^t)]^T$. Let $c^t = \frac{1}{c}$, we can rewrite the above equation as follows,

$$v^{t+1} = \arg\min_v \lambda||v||_1 + \frac{1}{2c^t}||v - (v^t - c^t\nabla\phi_v(v^t))||_2^2. \tag{24}$$

The iterative scheme in (24) is a basic scheme in $\ell_1$ minimization [24]. Many algorithms [26][42][43][44], have been proposed based on (24). Each of its components $v_i$ can be independently obtained by soft shrinkage operator [24].

In addition, by substituting $v = X\beta - y$ and $J(\beta) = ||\beta||_2^2$ into (12), (12) then takes the form:

$$\min_\beta \phi_v(X\beta - y) + \lambda||\beta||_2^2. \tag{25}$$

If we make use of the multiplicative form of HQ in Algorithm 1 to solve (25), Algorithm 1 becomes the iterative reweighted least squares (IRLS) of robust statistics [28][31]. It has been shown in [32] that IRLS is a special case of half-quadratic minimization.

## III. Half-quadratic based Robust Sparse Representation Algorithms

Based on the HQ framework, this section addresses the robust problem in (8). We can rewrite (8) as follows:

$$J(\beta) \doteq \min_\beta \phi_v(X\beta - y) + \lambda||\beta||_1, \tag{26}$$

where $\phi(.)$ in $\phi_v(.)$ is a non-convex (or convex) M-estimator and can be optimized by HQ. (26) can be viewed as a robust formulation of (4) by substituting $\ell_2$-norm with $\phi_v(.)$. When $\phi(.)$ is Welsch M-estimator, (26) is actually based on maximum correntropy criterion (MCC) (see Appendix II in the supplementary file). Since $\phi(.)$ is a robust M-estimator, the minimization of (8) is actually a M-estimation of $\beta$. We adopt the two forms of HQ to optimize (26). When the additive form is used, we gain an $\ell_1$ regularized error correction algorithm. The auxiliary variable of HQ actually models errors incurred by noise. When the multiplicative form is used, we obtain an $\ell_1$ regularized error detection algorithm. The auxiliary variable can be viewed as a weight to detect noise. The following will detail the technique respectively.

### A. *Proposed Error Correction*

In this subsection, we utilize the additive form of HQ to optimize (26). Let $Q_v(X\beta - y, p) = ||X\beta - y - p||_2^2$, we have the following augmented objective function of (26) [32][33],

$$J_A(\beta, p) \doteq \min_{\beta,p} ||X\beta - y - p||_2^2 + \sum_{j=1}^d \varphi(p_j) + \lambda||\beta||_1, \tag{27}$$

where auxiliary variable $p$ is uniquely determined by the minimization function w.r.t. $\phi(.)$.

Let $f^{t+1} \doteq p^{t+1} + y$, we can alternatingly minimize (27) as follows,

$$f^{t+1} = y + \delta_v(X\beta^t - y), \tag{28}$$

$$\beta^{t+1} = \arg\min_\beta ||X\beta - f^{t+1}||_2^2 + \lambda||\beta||_1. \tag{29}$$

Note that, to save computational costs,it is efficient to find a solution in (29) that satisfies $J_A(\beta^{t+1}, p^{t+1}) \leq \hat{J}_A(\beta^t, p^{t+1})$.

Algorithm 2 summarizes the optimization procedure. As in Remark 2 in [33] (page 3), Algorithm 2 alternately minimizes the augmented objective function $J_A(\beta, p)$ until it converges (Proposition 2). In each iteration, it tries to re-estimate the value of an input sample $y$ ($f^{t+1}$). Since $\phi(.)$ is a robust M-estimator, corrupted entries in $y$ will be corrected step by step. Hence we denote Algorithm 2 as *error correction*. The minimization subproblem in (29) can be solved and expressed in a closed form as a shrinkage [24]. To easily tune the parameters, we give an active algorithm in Appendix I (See the supplementary file) to implement Algorithm 2.

**Algorithm 2**: $\ell_1$-regularized Error Correction

```
Input: data matrix $X$, test sample $y$, and $\beta = X^T y$.
Output: $\beta$
while "not converged" do
    $f^{t+1} = y + \delta_v(X\beta^t - y)$
    $\beta^{t+1} = \arg\min_\beta ||X\beta - f^{t+1}||_2^2 + \lambda||\beta||_1$
    $t = t + 1$
end while
```

*Proposition 1:* The sequence $\{J_A(\beta^t, p^t),\ t = 1, 2, \ldots\}$ generated by Algorithm 2 converges.

*Proof:* According to the properties of the minimizer function $\delta(.)$ ($\{Q(v_j, \delta(v_j)) + \varphi(\delta(v_j))\} \leq \{Q(v_j, p_j) + \varphi(p_j)\}$), for a fixed $\beta^t$, we have $J_A(\beta^t, p^{t+1}) \leq J_A(\beta^t, p^t)$. And according to (29), for a fixed $p^{t+1}$, we have that $J_A(\beta^{t+1}, p^{t+1}) \leq \hat{J}_A(\beta^t, p^{t+1})$ such that

$$J_A(\beta^{t+1}, p^{t+1}) \leq \hat{J}_A(\beta^t, p^{t+1}) \leq \hat{J}_A(\beta^t, p^t).$$

Since $J_A$ is bounded below, the sequence

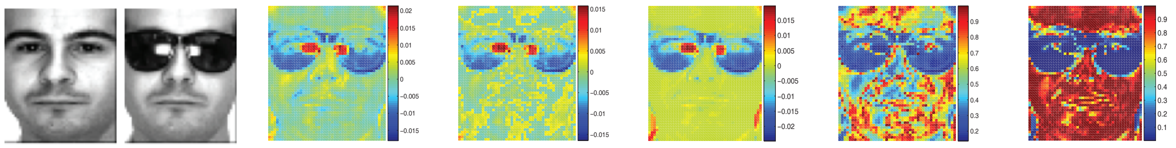
$$\{\ldots, J_A(\beta^t, p^t), J_A(\beta^t, p^{t+1}), J_A(\beta^{t+1}, p^{t+1}), \ldots\}$$

converges as $t \rightarrow \infty$. In particular, $J_A(\beta^{t+1}) \leq J_A(\beta^t)$, for all $t$, and the sequence $J_A(\beta^t)$ is convergent. ■

Similar to S1-$\ell_1$-MAGIC, Algorithm 2 also estimates noise at each iteration[3]. However, different from S1-$\ell_1$-MAGIC which assumes that noise has a sparse representation as well, Algorithm 2 has no such a specific assumption. If noise is indeed sparse in some applications, Algorithm 2 will naturally obtain a sparse solution of $p$ due to the fact that outliers are significantly different from uncorrupted entries.

Fig. 1 (d) and (e) show two examples of the auxiliary variables when Algorithm 2 converges. From Fig. 1 (b), we see that there are two occluded regions. One is highlight occlusion, and the other is sun-glasses occlusion. We see that Algorithm 2 can accurately estimate these two occlusions in this case. This is because M-estimators can efficiently deal with outliers (occlusions) that are significantly different from uncorrupted face pixels. As shown in Table I in Section IV, the minimizer functions of M-estimators in the additive form can estimate outliers and meanwhile keep the variations of uncorrupted data. More experimental validations are given in Section V. In Fig. 1 (d) and (e), the red regions with large positive values correspond to the highlight conclusion, and the blue regions with small negative values correspond to the sun-glasses conclusion.

[3] We denote the $\ell_1$-MAGIC toolbox used to solve (47) as S1-$\ell_1$-MAGIC.



(a) (b) (c) S1-$\ell_1$-MAGIC (d) Fair (e) Welsch (f) Fair (g) Welsch

Fig. 1. Error and weight images learned by different methods. An error (or weight) image is obtained by reshaping an error vector $e$ or auxiliary variable $p$. (a) An uncorrupted face image in the AR database. (b) An input face image $y$ with sun-glasses occlusion. (c) The error image of S1-$\ell_1$-MAGIC. (d) The error image of Fair M-estimator based error correction. (e) The error image of Welsch M-estimator based error correction. (f) The weight image of Fair M-estimator based error detection. (g) The weight image of Welsch M-estimator based error detection.

### B. Proposed Error Detection

In this subsection, we make use of the multiplicative form to optimize (26). Let $Q_v(X\beta - y, p) = \sum_j (p_j(y_j - \sum_i x_{ij}\beta_i)^2)$, we have the following augmented objective function of (26) [32][33],

$$J_M(\beta, p) \doteq \min_{\beta,p} \sum_{j=1}^{d} (p_j(y_j - \sum_{i=1}^{n} x_{ij}\beta_i)^2 + \varphi(p_j)) + \lambda||\beta||_1. \tag{30}$$

According to HQ optimization, a local minimizer $(\beta, p)$ of (30) can be alternately calculated by

$$p_j^{t+1} = \delta(y_j - \sum_{i=1}^{n} x_{ij}\beta_i^t), \tag{31}$$

$$\beta^{t+1} = \arg\min_{\beta} (y - X\beta)^T P(y - X\beta) + \lambda||\beta||_1, \tag{32}$$

where $P$ is a diagonal matrix whose diagonal element $P_{jj} = p_j^{t+1}$. The optimization problem in (32) can be rewritten as the following $\ell_1$-regularized quadratic problem:

$$\min_{\beta} ||\hat{X}\beta - f^{t+1}||_2^2 + \lambda||\beta||_1, \tag{33}$$

where $\hat{X} = \sqrt{P}X$ and $f^{t+1} = \sqrt{P}y$. Note that, to save computational cost, it is unnecessary to find the global solution of (33). It may be more efficient to find a sparse solution that satisfies $J_M(\beta^{t+1}, p^{t+1}) \leq \hat{J}_M(\beta^t, p^{t+1})$.

Algorithm 3 summarizes the optimization procedure. It alternatingly minimizes the augmented objective function $J_M(\beta, p)$ until it converges (Proposition 2). Since outliers are far away from the portion of uncorrupted data, their contributions to the optimization of the objective function will be smaller, as they always gain small values in matrix $P^{t+1}$. Therefore, outliers will have weaker influence on the estimation of $\beta$ such that Algorithm 3 can compute a sparse representation based on uncorrupted entries in $y$. And hence, we denote Algorithm 3 as *error detection*.

**Algorithm 3**: $\ell_1$-regularized Error Detection

```
Input: data matrix $X$, test sample $y$, and $\beta = X^T y$.
Output: $\beta$, $p$
while "not converged" do
    $P_{jj}^{t+1} = \delta(y_j - \sum_{i=1}^{n} x_{ij}\beta_i^t)$
    $f^{t+1} = \sqrt{P^{t+1}}y$ and $\hat{X} = \sqrt{P^{t+1}}X$
    $\beta^{t+1} = \arg\min_{\beta} ||\hat{X}\beta - f^{t+1}||_2^2 + \lambda||\beta||_1$
    $t = t + 1$
end while
```

*Proposition 2:* The sequence $\{\hat{J}_M(\beta^t, p^t),\ t = 1, 2, \ldots\}$ generated by Algorithm 3 converges.

*Proof:* According to the properties of the minimizer function $\delta(.)$ ($\{Q(v_j, \delta(v_j)) + \varphi(\delta(v_j))\} \leq \{Q(v_j, p_j) + \varphi(p_j)\}$), we have the following form for a fixed $\beta^t$, $J_M(\beta^t, p^{t+1}) \leq J_M(\beta^t, p^t)$. And for a fixed $p^{t+1}$, we have $J_M(\beta^{t+1}, p^{t+1}) \leq J_M(\beta^t, p^{t+1})$ such that

$$J_M(\beta^{t+1}, p^{t+1}) \leq \hat{J}_M(\beta^t, p^{t+1}) \leq \hat{J}_M(\beta^t, p^t).$$

Since $J_A$ is bounded below, the sequence

$$\{\ldots, J_M(\beta^t, p^t), J_M(\beta^t, p^{t+1}), J_M(\beta^{t+1}, p^{t+1}), \ldots\}$$

converges as $t \rightarrow \infty$. In particular, $J_M(\beta^{t+1}) \leq J_M(\beta^t)$, for all $t$, and the sequence $J_M(\beta^t)$ is convergent. ■

Fig. 1 (f) and (g) show two examples of auxiliary variables when Algorithm 3 converges. We see that Algorithm 3 treats the two occlusions in the same way. It assigns the two occluded regions small values (weights) due to the robustness of M-estimators. The auxiliary variable in Algorithm 3 actually plays a role of weighting function in each iteration.

## IV. M-Estimation for Large Corruptions

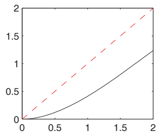
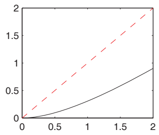
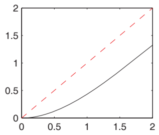
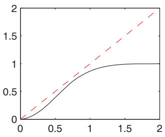
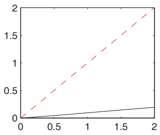
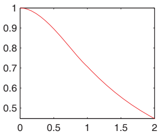
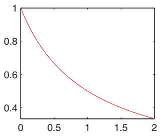
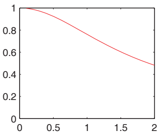
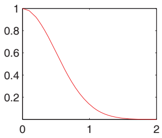
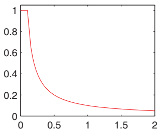
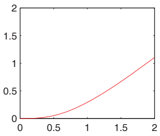
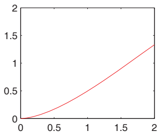
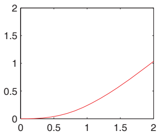
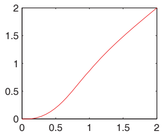
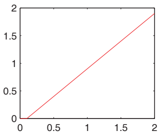
We leave the discussion of $\phi$ in the previous sections. In this section, we focus on it and connect it with the M-estimation. In robust statistics, one popular robust technique is M-estimation [45] (See Appendix II), which is defined as the minima of summation of functions of data and has been used with a history of more than 30 years. Table I tabulates the corresponding minimization functions $\delta$ related to the multiplicative and the additive forms of HQ respectively for different potential M-estimators $\phi$. The first row tabulates potential M-estimators and their curves [45]. The dashlines in these figures correspond to $\ell_1$ M-estimator[4]. The second row tabulates their corresponding minimization functions of the multiplicative form. The third row tabulates their corresponding minimization functions of the additive form. From Table I, we see that all M-estimators achieve the minima (zero) at origin.

In information theoretic learning (ITL) [46], it has been proved that the robustness of correntropy [31] and Renyi's quadratic entropy [41] based algorithms is actually related to Welsch M-estimator. If we substitute $\phi(.)$ with Welsch M-estimator, the adaptive linear system in (8) is a correntropy based adaptive system (See Appendix II). In ITL, correntropy has a probabilistic meaning of maximizing the error probability density at the origin [31] and its adaptation is applicable in any noisy environment

[4] In this section, we discuss the $\ell_1$ M-estimator in robust statistics [45].

TABLE I

MINIMIZATION FUNCTIONS $\delta$ RELEVANT TO THE MULTIPLICATIVE AND THE ADDITIVE FORM OF HQ FOR DIFFERENT POTENTIAL M-ESTIMATORS $\phi$. $\alpha$ IN M-ESTIMATOR IS A CONSTANT.

| estimators | $\ell_1$-$\ell_2$ | Fair | log-cosh | Welsch($\sigma^2 = 0.5$) | Huber ($\lambda = 0.1$) |
|---|---|---|---|---|---|
| Potential function $\phi(t)$ | $\sqrt{\alpha + t^2} - 1$<br> | $\frac{\lvert t\rvert}{\alpha} - \log(1 + \frac{\lvert t\rvert}{\alpha})$<br> | $\log(\cosh(\alpha t))$<br> | $1 - \exp(-\frac{t^2}{\sigma^2})$<br> | $\begin{cases} t^2/2 & \lvert t\rvert \le \lambda \\ \lambda\lvert t\rvert - \frac{\lambda^2}{2} & \lvert t\rvert > \lambda \end{cases}$<br> |
| Multiplica-tive form $\delta(t)$ | $1/\sqrt{\alpha + t^2}$<br> | $1/\alpha(\alpha + \lvert t\rvert)$<br> | $\alpha \frac{\tanh(\alpha t)}{t}$<br> | $\exp(-\frac{t^2}{\sigma^2})$<br> | $\begin{cases} 1 & \lvert t\rvert \le \lambda \\ \frac{\lambda}{\lvert t\rvert} & \lvert t\rvert > \lambda \end{cases}$<br> |
| Additive form $\delta(t)$ | $t - \frac{t}{\sqrt{\alpha + t^2}}$<br> | $t - \frac{t}{\alpha(\alpha + \lvert t\rvert)}$<br> | $t - \alpha\tanh(\alpha t)$<br> | $t - t\exp(-\frac{t^2}{\sigma^2})$<br> | $\begin{cases} 0 & \lvert t\rvert \le \lambda \\ t - \lambda sign(t) & \lvert t\rvert > \lambda \end{cases}$<br> |

when its distribution has the maximum at the origin [31]. This probabilistic property enables the theoretical support that correntropy adaptation can deal with large corruption. Hence learning algorithms will obtain high accuracy if the uncorrupted pixels are discriminative enough. In the following, we mainly discuss two commonly used M-estimators: Huber and Welsch.

In the compressed sensing community, the fast iterative $\ell_1$-minimization algorithms [13][24] often rely on the soft-thresholding function. Since the $\ell_1$-norm $||p||_1 = \sum_j |p_j|$ in those algorithms is separable, each entry $p_j$ of $p$ is the minimization solution of the following problem,

$$\min_{p_j} \frac{1}{2}(x_j - p_j)^2 + \lambda|p_j|, \tag{34}$$

where $p_j \in R$ and $x_j \in R$ is the j-th entry of vector $x$. In compressed sensing, the optimal solution $p_j^*$ of (34) is popularly found by the following soft-thresholding function, i.e.,

$$p_j^* = soft(x_j, \lambda) = \begin{cases} 0 & |x_j| \le \lambda \\ x_j - \lambda sign(x_j) & |x_j| > \lambda \end{cases}. \tag{35}$$

By substituting $p_j^* = soft(x_j, \lambda)$ into (34), we have,

$$\min_{p_j} \frac{1}{2}(x_j - p_j)^2 + \lambda|p_j| = \begin{cases} x_j^2/2 & |x_j| \le \lambda \\ \lambda\,|x_j| - \frac{\lambda^2}{2} & |x_j| > \lambda \end{cases}. \tag{36}$$

The right part of equation (36) is often denoted as Huber loss function $\phi_H(.)$ in HQ (or Huber M-estimator in robust statistics). By conducting a summation on (36) over $j$, we have the following equation,

$$\min_p \frac{1}{2}||x - p||_2^2 + \lambda||p||_1 = \sum\nolimits_j \phi_H(x_j). \tag{37}$$

The formulations in (36) and (37) have also been studied for the HQ minimization in terms of the additive form where Huber loss function $\phi_H(.)$ has the following additive form (See Equations (3.26) to (3.28) in Section 3.3 of [33]),

$$\min_{x_j} \phi_H(x_j) = \min_{x_j, p_j} \frac{1}{2}(x_j - p_j)^2 + \lambda|p_j|, \tag{38}$$

where $p_j$ is an auxiliary variable and is uniquely determined by the minimization function of $\phi_H$, i.e., soft-thresholding function in (35). In HQ, the right-hand side of equation (38) is often called augmented objective function.

Note that no matter which side (the left- or the right-hand side of (36) and (38)) is used to model a problem, the dual relationship between Huber loss function and absolute function is uniquely determined by soft-thresholding function. And it always holds no matter vector $p$ is sparse or not. If $p$ is used to model dense noise, one can also correctly detect outliers due to the robustness of Huber M-estimator.

Based on this conjugate perspective, we learn that when the problem in (6) is solved using soft-thresholding methods, (6) is equivalent to the following problem by substituting $e_j$ with $-p_j^*$ in (35) and applying (38),

$$\min_\beta \sum\nolimits_{j=1}^d \phi_H((X\beta - y)_j) + \lambda||\beta||_1. \tag{39}$$

And according to the additive form of HQ, we can also have the augmented problem of (39), i.e.,

$$\min_{\beta, e} \sum\nolimits_{j=1}^d (((X\beta + e - y)_j)^2 + \lambda|e_j|) + \lambda||\beta||_1, \tag{40}$$

where the function $|.|$ is the dual potential function of Huber loss function [33]. When one resorts to iterative shrinkage thresholding for solving (6) and (40) and uses a descending $\lambda$ (i.e., $\lambda$ approaches 0), $\phi_H(x) = \min_e \frac{1}{2}(x - e)^2 + \lambda\,|e|$ will approach $\min_e \frac{1}{2}(x - e)^2$ (i.e., $e^* = \arg\min_e \frac{1}{2}(x - e)^2 = x$) such that the solutions of both (6) and (40) tend to be the solution of $X\beta - y = e$.

Different from the assumption that $e$ in (6) has a sparse representation [1][14], the $\ell_1$-norm in (40) means the dual potential function of Huber M-estimator (See (37)) rather than an approximation of $\ell_0$-norm. Hence both (39) and (40) are robust to outliers due to M-estimation. Recent experimental results on robust face recognition [13][14] also show that the model in (6) can achieve high recognition accuracy even when corruption is larger than 50%. From this dual viewpoint, we learn that this robustness is potentially because the soft-thresholding function used to solve (6) plays a role of robust Huber M-estimator.

Looking at the curve of Welsch M-estimator, we can observe that its segment between 0 and 1 is similar to that of $\ell_1$-norm. However, the segment of Welsch M-estimator between 1 and $\infty$ tends to be flattened out and is significantly different from $\ell_1$-norm. It imposes the same penalties to all outliers so that it can efficiently deal with those outliers with large magnitudes. Looking at the multiplicative minimization function of Welsch M-estimator, we can observe that the minimization function decreases dramatically when $t > 0.5$. That is to say, outliers will be given small weights during optimization. Fig. 1 (g) shows an example of the weight image of Welsch M-estimator by reshaping the auxiliary variable $p$ of our error detection algorithm. Looking at the additive minimization function of Welsch M-estimator, we can observe that the minimization function tends to be diagonal when $t > 0.5$. This means that an algorithm based on Welsch M-estimator can accurately estimate outliers that are intrinsically different from the uncorrupted ones. Fig. 1 (e) shows an example of the error image of Welsch M-estimator by reshaping auxiliary variable $p$ of our error correction algorithm.

Fig. 1 also shows visual results of error correction and detection algorithms for robust face recognition along with the comparison with S1-$\ell_1$-MAGIC. (More results will be reported in the experimental section.) Fig. 1 (a) shows an uncorrupted face image in the AR database [47]. And Fig. 1 (b) shows an input face image $y$ with sun-glasses occlusion. We can observe that there are two types of occlusions: one is incurred by sun-glasses and the other is incurred by the highlight in the sun-glasses. Fig. 1 (c) shows the error image of S1-$\ell_1$-MAGIC by reshaping error vector $e$ in (6). Fig. 1 (d) and (e) show the error images of Fair and Welsch M-estimators respectively by reshaping the auxiliary variable $p$ of our error correction algorithm; and Fig. 1 (f) and (g) show the weight images of Fair and Welsch M-estimator respectively by reshaping the auxiliary variable $p$ of our error detection algorithm.

From Fig. 1, we can observe that the methods based on Welsch M-estimator can accurately estimate or detect the occluded region in a face image. Compared with the results of Fair M-estimator, those of Welsch M-estimator are smoother, especially in non-occluded regions. This means that the method based on Welsch M-estimator can estimate occlusions more accurately than the one based on Fair M-estimator. Comparing Fig. 1 (c) with (e), we can observe that the error correction algorithm based on Welsch M-estimator likely computes a smoother result than S1-$\ell_1$-MAGIC.

## V. Experimental Results

In experiments, we focus on the robust face recognition problem [14][29] and demonstrate the effectiveness of the proposed methods for solving occlusion and corruption problems. Two public face recognition databases, namely the AR [47] and Extended Yale B [48] databases, were selected for experiments. Recognition rate and computational cost were used to evaluate the compared methods. All algorithms were implemented using MATLAB on an AMD Quad-Core 1.80GHz machine with 2GB memory.

### A. Experimental Setting and Face Databases

**Databases.** All grayscale images of the two public face databases were aligned by manually locating eyes of face images. The two selected databases are as follows.

1) *AR Database* [47]: it consists of over 4,000 face images from 126 subjects (70 men and 56 women). For each subject, 26 facial images were taken in two separate sessions. These images suffer different facial variations including various facial expressions (neutral, smile, anger, and scream), illumination variations (left light on, right light on and all side lights on), and occlusions by sun-glasses or scarf. This database is often used to compare robust face recognition methods. In our experiments, we used the same subset used in [3] that consists of 65 male subjects and 54 female subjects. Facial images were cropped and the resolution is $112 \times 92$.

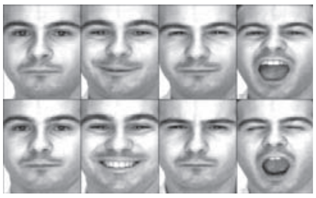


Fig. 2. Cropped facial images of the first subject in the AR database. Images in the first row are from the first session and images in the second row are from the second session.

2) *Extended Yale B Database*[48][49]: it is composed of 2,414 frontal face images from 38 subjects. Cropped and normalized $192 \times 168$ face images were captured under various controlled lighting conditions [49][14]. Fig. 3 shows some face images of the first subject in this database. For each subject, half of the images were randomly selected for training (i.e., about 32 images for each subject), and the rest were for testing.

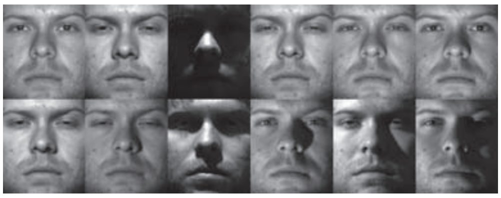


Fig. 3. Cropped facial images of one subject in the YALE B database.

**Methods.** We categorize the compared methods into two groups. The first group is not robust to outliers, and the second group is robust to outliers.

**First Group** In our experiments, the first group includes five sparse representation models, detailed as follows:

1) For the first sparse representation model formed by

$$\min_{\beta} ||\beta||_1 \quad s.t. \; ||X\beta - y||_2 \le \varepsilon, \tag{41}$$

we denote the $\ell_1$-MAGIC toolbox[5] used to solve (41) as S0-$\ell_1$-MAGIC.

[5] http://users.ece.gatech.edu/ justin/l1magic/

2) For the second sparse representation model formed by

$$\min_{\beta} \|X\beta - y\|_2^2 + \lambda||\beta||_1, \tag{42}$$

we denote the feature-sign search (FSS) algorithm [50][6], fast iterative shrinkage-thresholding algorithm (FISTA) [51] and Homotopy (HOMO) [52] algorithm used to solve (42) as S0-FSS, S0-FISTA and S0-HOMO respectively.

3) For the third sparse representation model formed by

$$\min_{\beta} ||\beta||_1 \quad s.t.\ X\beta = y, \tag{43}$$

we denote the polytope faces pursuit (PFP) [53] method used to solve (43) as S0-PFP.

4) For the fourth sparse representation model formed by

$$\min_{\beta} \|X\beta - y\|_2^2 + \lambda \sum\nolimits_i w_i\,|\beta_i|, \tag{44}$$

where $w = [w_1, \ldots, w_n]$ is a weight vector. We denote the method used to solve the above adaptive LASSO problem [21] as S0-ALASSO.

5) For the fifth sparse representation model formed by

$$\min_{\beta} \sum\nolimits_i w_i|\beta_i| \quad s.t.\ X\beta = y, \tag{45}$$

we denote the method used to solve the above reweighted $\ell_1$ minimization problem [22] as S0-$\ell_1$-W.

In addition, we also compare three linear representation methods. In the half-quadratic minimization for image processing, one considers the following model to deal with white Gaussian noise,

$$\min_{\beta} \|X\beta - y\|_2^2 + \lambda \sum\nolimits_i \phi(\beta_i - \beta_{i+1}), \tag{46}$$

where $\phi(x) = \sqrt{\varepsilon + x^2}$ is a half-quadratic loss function. And the regularization in (46) models the first order differences between neighboring elements in $\beta$. We denote the additive form and the multiplicative form to (46) as HQSA and HQSM respectively. Linear regression based classification (LRC) [54] and collaborative representation based classification (CRC) [55] are also compared.

**Second Group** The second group consists of three robust sparse representation models detailed as follows.

1) For the first sparse representation model formed by

$$\min_{\beta,e} ||\beta||_1 + \|e\|_1 \quad s.t.\ ||X\beta + e - y||_2 \le \varepsilon, \tag{47}$$

we denote the $\ell_1$-MAGIC toolbox used to solve (47) as S1-$\ell_1$-MAGIC.

2) For the second sparse representation model formed by

$$\min_{\beta,e} \|X\beta + e - y\|_2^2 + \lambda(||\beta||_1 + ||e||_1), \tag{48}$$

we denote the method FSS, FISTA and HOMO used to solve (48) as S1-FSS, S1-FISTA and S1-HOMO respectively.

3) For the third sparse representation model in the form

$$\min_{\beta,e} ||\beta||_1 + \|e\|_1 \quad s.t.\ X\beta + e = y, \tag{49}$$

we denote the polytope faces pursuit (PFP) [53] method used to solve (49) as S1-PFP.

PFP, FISTA, HOMO, and sparse reconstruction by separable approximation (SpaRSA) [56] methods were implemented by 'fast $\ell_1$ minimization' MATLAB package [13][7]. We tuned the parameters of all the compared methods to achieve the best performance on the training set, and then used these parameter settings on the testing set. Since these methods take different optimization strategies and a corrupted testing set may be different from a training one, different sparse representation methods may obtain different results.

**Classifier.** Wright et al. [1][14] proposed a linear classification method for sparse representation, and He et al. [3][15] developed a nonlinear one. In this section, to fairly evaluate different robust methods, we classify an input sample $y$ as suggested in [14]. For each class $c$, let $\psi_c : \mathbb{R}^n \to \mathbb{R}^{n_c}$ be a function which selects the coefficients belonging to class $c$, i.e. $\psi_c(\beta) \in \mathbb{R}^{n_c}$ is a vector whose entries are the entries in $\beta$ corresponding to class $c$. Utilizing only the coefficients associated to class $c$, a given sample $y$ is reconstructed as $\hat{y}_c = X_c\psi_c(\beta)$ where $X_c$ is a matrix whose samples all belong to the class $c$. Then $y$ can be classified by assigning it to the class corresponding to the minimal difference between $y$ and $\hat{y}_c$, i.e.,

$$\arg\min_{c} ||y - X_c\psi_c(\beta)||_2. \tag{50}$$

**Algorithm setting.** As suggested by [14], we normalized the columns of $X$ to have unit $\ell_2$-norm for all compared algorithms. We make use of a robust way to estimate the parameters of M-estimators. For Huber M-estimator, the threshold parameter is estimated as a function of median, i.e.,

$$\lambda = a \times \underset{j}{median}(|y_j - \sum\nolimits_{i=1}^{n} x_{ij}\beta_i^t|). \tag{51}$$

And the kernel size of other M-estimators is estimated as a function of mean [3],.i.e.,

$$\sigma^2 = a \times \underset{j}{mean}((y_j - \sum\nolimits_{i=1}^{n} x_{ij}\beta_i^t)^2). \tag{52}$$

The constant $a$ in (51) and (52) is empirically set to be 0.8 and 0.5 respectively. More experimental results on the parameter selection of M-estimators will be shown in Section V-D. There are various strategies for the implementation of Algorithm 2 and Algorithm 3. Here we implement them by the active set algorithm detailed in Appendix I.

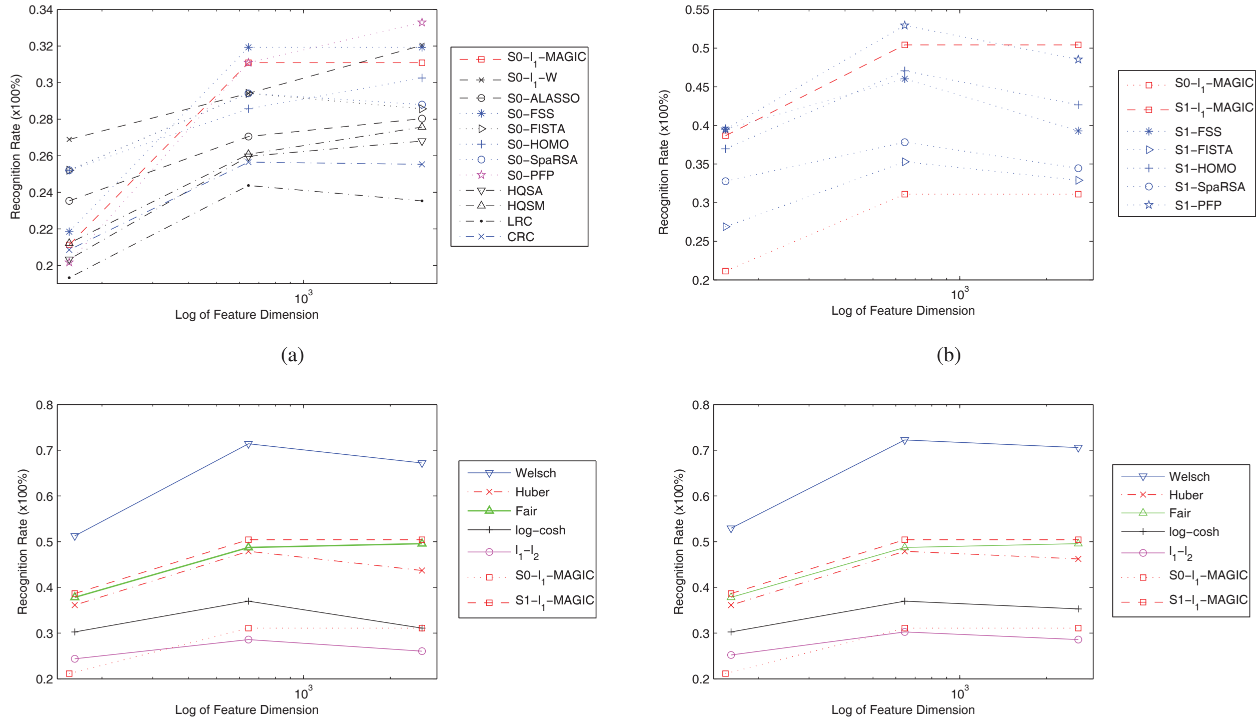
### *B. Sun-glasses Occlusion*

In this subsection, we investigate different methods against sun-glasses occlusion. For training, we used 952 non-occluded frontal view images (about 8 faces for each subject) with varying facial expressions in the AR database. Fig. 2 shows an example of 8 selected images of the first subject. For testing, we evaluated the methods on the images occluded by sun-glasses. Fig. 1 (b) shows a facial image from the testing set. Fig. 4 shows the recognition performance of different methods using different downsampled images of dimension 161, 644, and 2576 [14][54] corresponding to downsampling ratios of 1/8, 1/4, and 1/2, respectively.

Fig. 4 (a) shows experimental results of the methods that are not robust to outliers. Although these 'S0-' methods all aim to find a sparse solution, they obtain different recognition rates, as similarly shown in [13]. This is because they take different strategies for optimization and the corruption level in the testing set is unknown such that one can not tune parameters of each method to obtain the same result for each testing sample. We also see that sparse

[6] http://redwood.berkeley.edu/ bruno/sparsenet/

[7] http://www.eecs.berkeley.edu/ yang/software/l1benchmark/



(a)

(b)

(c)

(d)

Fig. 4. Recognition rates of different methods against sun-glasses occlusion in the AR database. (a) Recognition rates of the methods that are not robust to outliers. (b) Recognition rates of the sparse representation methods that are robust to outliers. (c) M-estimators are optimized by the additive form (Algorithm 2). (d) M-estimators are optimized by the multiplicative form (Algorithm 3).

representation methods outperform linear presentation methods (HQSA, HQSM, LRC and CRC). In addition, HQSA and HQSM perform slightly better than LRC and CRC.

Fig. 4 (b) plots the results of different sparse representation methods that are robust to outliers. Comparing Fig. 4 (b) with Fig. 4 (a), we see that recognition rates in Fig. 4 (b) are obviously higher than those in Fig. 4 (a). Although the same $\ell_1$ minimization methods are used, one 'S1-' method can deal with outliers better than its corresponding 'S0-' method. The recognition rate of S1-$\ell_1$-MAGIC is twice higher than that of S0-$\ell_1$-MAGIC. The results in Fig. 4 (a) and Fig. 4 (b) suggest that $\ell_1$ minimization methods 'S0-' fail to deal with outliers that are significantly different from the uncorrupted data. If outliers are not corrected, they will affect the estimation of sparsity largely. And if outliers are corrected as in the 'S1-' methods, the estimated sparse representation can be more accurate. We also illustrate the effect of outliers on sparse estimation in Appendix IV.

Fig. 4 (c) and Fig. 4 (d) show the results computed by the additive form (Algorithm 2) and the multiplicative form (Algorithm 3) respectively. We observe that recognition rates of the additive and multiplicative forms using the same M-estimator are very close. As shown by [3], S1-$\ell_1$-MAGIC is not robust enough to contiguous occlusion for face recognition. As the results shown in Fig. 4, we observe that algorithms based on Welsch M-estimator significantly outperforms S1-$\ell_1$-MAGIC and other methods. This is due to the fact that Welsch M-estimator places the same penalties to the outliers incurred by sun-glasses as shown in Table I. This is consistent with the results reported in correntropy [3][31] and M-estimation [45] where Welsch M-estimator (or non-convex M-estimators) has shown to be an efficient tool for big outliers and non-Gaussian noise. We can also observe that error correction algorithms (or error detection algorithms) based on Huber and Fair M-estimator achieve similar recognition accuracy as compared with S1-$\ell_1$-MAGIC. This is because they all make use of the absolute function in their objectives.

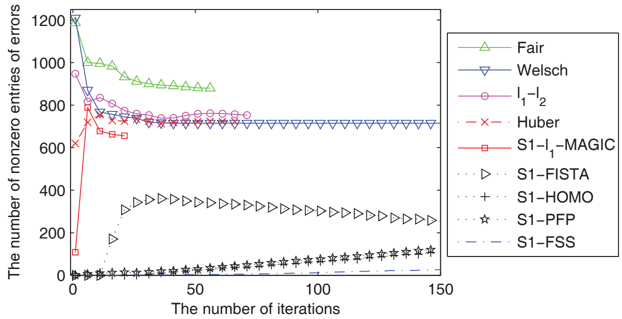
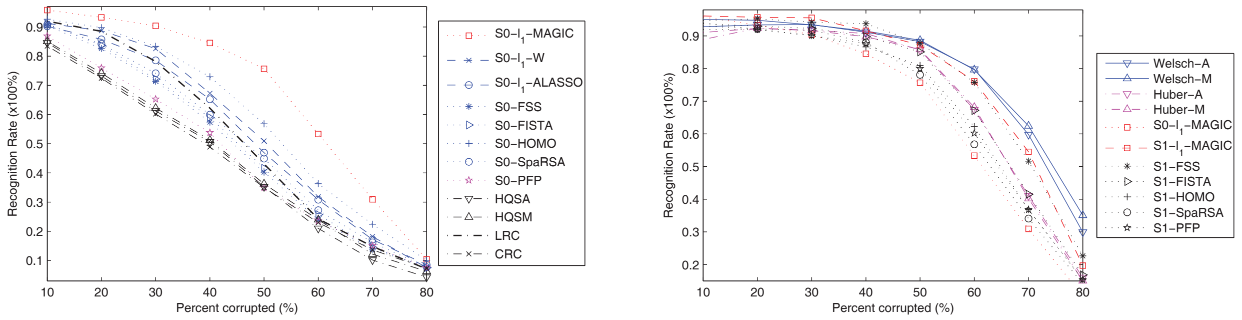


Fig. 5. Variation of nonzero entries of errors in error correction algorithms when the feature dimension is 2576.

Fig. 5 further shows the variation of the $\ell_0$ norm (i.e., the number of nonzero entries of error $e$) of error $e$ in error correction methods as a function of the number of iterations. We see that robust M-estimator with kernel size parameter in (51) or (52) performs significantly different as compared with Huber M-estimator and $\ell_1$ minimization methods. Since S1-FSS, S1-HOMO and S1-PFP all use the active set method, they estimate only one entry as the error in each iteration. As a result, when



(a)

(b)

Fig. 6. Recognition rates of various methods under random corruption. Capital letters 'A' and 'M' indicate the additive and the multiplicative forms respectively. (a) Recognition rates of the methods that are not robust to outliers. (b) Recognition rates of the methods that are robust to outliers.

the level of corruptions is large, these three methods will count on more iterations.

For the Huber M-estimator based method and $\ell_1$ minimization methods, the $\ell_0$-norm of error $e$ increases until they converge. For other robust methods with the kernel size parameter, the $\ell_0$ norm of error $e$ decreases as the number of iterations increases. This may be because the median operator in Huber M-estimator and the soft-threshold operator in $\ell_1$ minimization adaptively decrease the value of their threshold parameters. A large value of the parameter can lead to the scenario that only a small number of data entries are estimated as outliers at the beginning of iterations. In contrast, the kernel size parameter does not play a role of truncation function. As shown in Table I, there is a nonzero segment around the kernel size parameter so that its corresponding methods estimate a large number of data entries as noise at the beginning of iterations. Fig. 5 also shows that different M-estimators will cause different strategies to estimate errors incurred by noise although they are used in the same error correction algorithm.

### C. Random Pixel Corruption

We tested the robustness of different methods on the Extended Yale B Face Database. For each subject, half of the images were randomly selected for training, and the rest half were for testing. The training and testing set contained 1205 and 1209 images respectively. Since the images in the testing set incurred large variations due to different lighting conditions or facial expressions, it is a difficult recognition task. Each image was resized to $24 \times 21$ [8] and stacked it into a 504-$D$ vector. Each test image was corrupted by replacing a set of randomly selected pixels with a random pixel value which follows a uniform distribution over [0, 255]. We vary the percentage of image pixels that suffer corruptions from 10% to 80% [14].

Fig. 6 shows the recognition accuracy of different methods, as a function of the level of corruption. Here we focus on Welsch and Huber M-estimator based methods. We see that the recognition rates of all compared methods are close when the level of corruption is 10%. But the recognition rates of those methods in Fig. 6 (a) decrease rapidly as the level of corruption increases. In Fig. 6 (b), we can observe that S1-FSS, S1-PFP, S1-SpaSRA and S1-HOMO perform worse than the other methods. This may be due to the different modeling on sparsity under different parameters and optimization strategies. The two methods based on Welsch M-estimator can perform slightly better than $\ell_1$ minimization methods when the level of corruption is larger than 50%. And the two methods based on Huber M-estimator perform slightly worse than S1-$\ell_1$-MAGIC and almost obtain similar results as S1-FISTA. This is due to that the Huber M-estimator based methods and S1-FISTA all resort to soft-thresholding function. When the level of corruption is larger than 30%, the sparsity assumption on the noise vector $e$ cannot be made. However, methods based on (26) can still achieve high recognition rates due to the use of M-estimator in their objective functions. Even if there are large corruptions, the error correction algorithm (or the error detection algorithm) can still utilize uncorrupted pixels to correct (or detect) errors incurred by the noise. As a result, they achieve high recognition accuracy.

### D. Robustness, Sparsity and Computational Cost

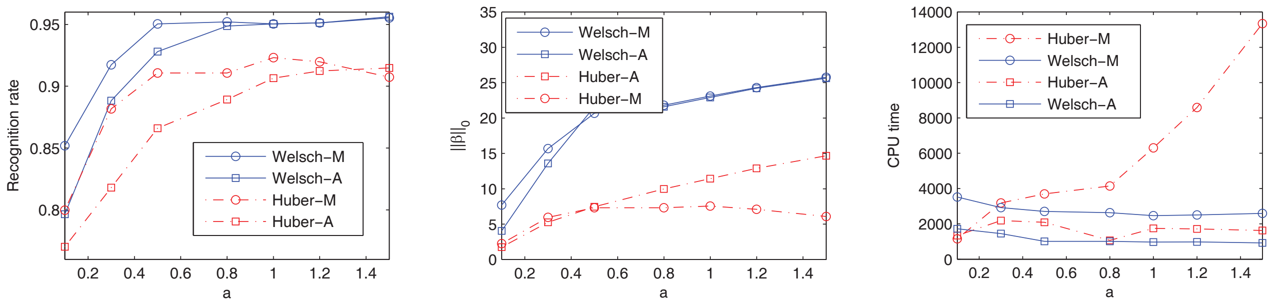
In real-world applications, occlusion and corruption are often unknown. Hence, a parameter obtained from cross-validation on uncorrupted training set may not be realistic for corrupted testing data. A common way for parameter selection of M-estimators is robust parameter selection method, such as median [28] and Silverman's rule [31]. However, those robust parameter selection methods are only developed for small level of corruptions. When the corruption and occlusion are larger than 50%, those methods will fail. To the best of our knowledge, there seems no existing work to discuss the parameter selection of M-estimators for large corruptions. In order to overcome this difficulty, we follow the approach in [31] to investigate parameter selection and discuss its effect on recognition accuracy, computational cost and sparsity of coefficient $\beta$.

We conducted two experiments in this subsection. The setting of the first experiment is the same as that of Section V-C. In the second experiment, we study our proposed methods via phase transition diagram [27][57]. As in [27], we address the following problem,

$$y = X\beta_0 + e$$

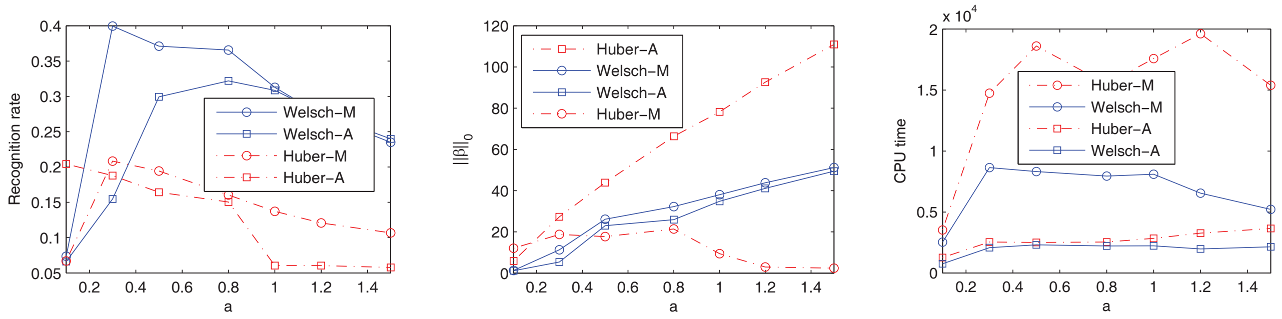
where $\beta_0$ is zero except for $s$ entries drawn from $N(0,1)$, each $X_{ij} \sim N(0,1)$ with column normalized to unit length, and $e$ is zero except for $(0.1 \times d)$ entries (i.e., the level of corruption is 10%.). To simulate outliers, the nonzero entries of $e$ are drawn

[8] The matlab 'imresize' function was used to resize image.



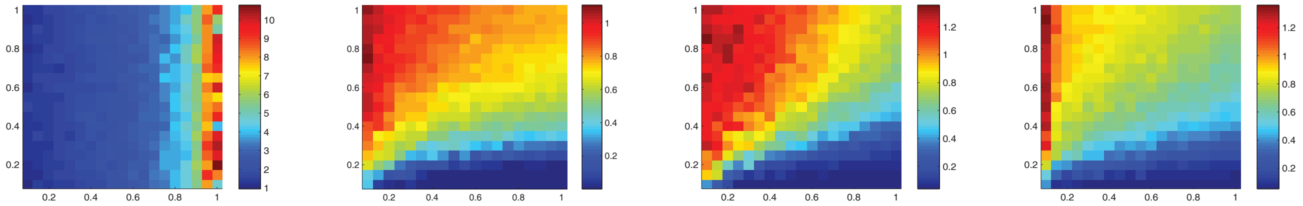
(a) Recognition rate (b) Sparsity (c) Computation cost

Fig. 7. Recognition accuracy, sparsity, and total CPU time of Huber and Welsch M-estimator based methods under 10% corruption.



(a) Recognition rate (b) Sparsity (c) Computation cost

Fig. 8. Recognition accuracy, sparsity, and total CPU time of Huber and Welsch M-estimator based methods under 80% corruption.



(a) S0-$\ell_1$-MAGIC (b) S1-$\ell_1$-MAGIC (c) Huber-A (d) Huber-M

Fig. 9. Phase transition diagrams [57][27] of different methods, where the level of corruption is fixed at 10% and the number of samples $n$ is fixed at 200. Horizontal axis: $\delta = d/n$ (the number of feature dimension / the number of samples). Vertical axis: $\rho = s/d$ (the number of nonzero elements / the number of feature dimension). Each color indicates a different median of normalized $l_2$ error of $||\hat{\beta} - \beta_0||_2 / \|\beta_0\|_2$ over 30 runs.

from $\{2 \times \max_j |(X\beta_0)_j|\} \times N(0,1)$ [9]. As in [27], white noise is not used. Since Huber loss function has a dual relationship to absolute function $|.|$, we only report phase transition diagrams of Huber M-estimator based methods.

The recognition accuracy, sparsity, and CPU time of Huber and Welsch M-estimator based methods are plotted in Fig. 7 and Fig. 8 as a function of the value of $a$ (in (51) and (52)) respectively. And the phase transition diagrams of the four compared methods are shown in Fig. 9. Capital letters 'A' and 'M' indicate the additive and the multiplicative form respectively. The main observations from the experiments are summarized below.

***Parameter selection***: As discussed in correntropy, the kernel size of Welsch M-estimator controls all properties of robustness [31]. We can see that for the two selected M-estimators, their parameters control recognition rates, sparsity of coefficient $\beta$, and computational cost. Moreover, the effect of their parameters seems to be different under different levels of corruptions. In the case of 10% corruption, the best accuracy is achieved when $a$ is around 1.2, suggesting the use of a large $a$; in the case of 80% corruption, the best accuracy is achieved when $a$ is around 0.3, suggesting the use of a small $a$. When the percentage of corruption or occlusion is smaller than 50%, the mean (or median) is mainly dominated by uncorrupted pixels, and therefore $a$ can be set to a larger value to adapt to uncorrupted pixels; when the level of corruption is larger than 50%, the mean (or median) is mainly dominated by corrupted pixels, and therefore $a$ can be set to a smaller value to punish outliers seriously.

***Sparsity***: From Fig. 7, Fig. 8, Fig. 9 and Fig. 13, we see that outliers will significantly affect the estimation of sparse representation. Since S0-$\ell_1$-MAGIC method does not concern outliers very well, it fails to find the ground truth solution $\beta_0$ in Fig. 9 (a) and Fig. 13 (a). And for all the compared methods,

[9] Since there will be several types of outliers in real world problems, we generate outliers in a different way from [27].

the estimation errors of $\beta$ increase as the level of corruption increases. Although M-estimators do not directly penalize the sparse coefficient $\beta$, they control uncorrupted pixels during learning. In each iteration, robust methods make use of uncorrupted pixels to estimate corrupted pixels. For the additive form, the corrupted pixels of a test sample are iteratively corrected; for the multiplicative form, the corrupted pixels are eliminated by reweighting. Since the sparsity of coefficient $\beta$ is related to the linear model which mainly depends on uncorrupted pixels, M-estimators will also affect sparsity.

***Computational cost***: the computational cost of the multiplicative form is much higher than that of the additive form. That is, the minimization using the additive form is faster than the one using the multiplicative one. One simple reason is that the method based on the multiplicative form often involves matrix multiplication in each iteration (e.g., weighting the data in the training set). These results are consistent with those reported in [33]. Therefore the use of the additive form is recommended for high dimensional data. In addition, since the proposed methods aim to learn a sparse representation and deal with outliers at the same time, their computational cost will be larger than those 'S0' methods. However, since our methods are based on M-estimation rather than sparse assumption on errors, our methods are potentially helpful for dense errors.

***Convex and non-convex M-estimators***: From the experimental results, we see that recognition rates of the multiplicative and the additive forms are close for each M-estimator. And different M-estimators will result in different recognition rates. In the case of 80% corruption, Welsch M-estimator (non-convex) seems to consistently outperform Huber M-estimator (convex) in terms of recognition rate. Although convex loss functions have a global solution, they do not handle outliers well. In real world applications, there are often several types of outliers, such as sunglasses and highlight occlusion in Fig. 1 (b). As plotted in Table I, a convex M-estimator gives different errors different loss such that it may give much attention on large errors. In contrast, non-convex loss functions often enhance sparsity for high dimensional problem [23], or improve robustness to outliers in [45]. Moreover, the study in information theoretic learning [31] shows that the performance sensitivity to kernel size is much lower than the selection of thresholds in M-estimators. Hence the selection of M-estimators is important for robust sparse representation and a non-convex M-estimator may be more applicable.

***Error correction and detection***: The performance of error correction and detection is different in the aspects of recognition rates, sparsity and computation cost, although they optimize the same objective function from the viewpoint of HQ. This difference of the two forms always exists in HQ methods [33]. Note that the augmented objective functions of the multiplicative form is convex only when its auxiliary variables are fixed, which makes local minimum solutions for the pair $(\beta, p)$. In addition, the recognition rate of error detection algorithms seems to be higher than that of error correction algorithms in a large range of $a$. And in Fig. 9 and Fig. 13, the green and blue regions of the multiplicative form are larger than those of the additive form, which indicates better recovery performance. Theoretically speaking, the two forms of HQ methods should have similar results if their parameters are well tuned for each testing sample. However, in practice, the multiplicative form is often more robust. This is because the parameter of the multiplicative form seems to be more adaptive and easily tuned for different corruption levels.

## VI. Conclusion and Future Work

We have presented a general half-quadratic framework for solving the problem of robust sparse representation. This framework unifies algorithms for error correction and detection by using the additive and the multiplicative forms respectively. Some effective M-estimators for the proposed half-quadratic framework have been investigated. We have shown that the absolute function in $\ell_1$ regularizer solved by soft-thresholding function can be viewed as the dual form of Huber M-estimator, which gives a theoretical guarantee of the robustness of robust sparse representation methods in terms of M-estimation. Experimental results on robust face recognition have shown that when applicable, Welsch M-estimator is potentially attractive and effective to handle large occlusion and corruption than other M-estimators, and error correction algorithms are suitable for high-dimensional data than error detection algorithms.

Our study of the HQ based robust sparse representation also shows that outliers significantly affect the estimation of sparse coding and different M-estimators will result in different robustness. Non-convex M-estimators seem to be more robust to real-world outliers that are often complex. The first avenue for future research is to introduce structure prior of errors, such as smooth constraint [33] and tree structure, into our robust framework to deal with a particular occlusion task. In addition, soft-thresholding function and iteratively reweighted methods have drawn much attention in subspace segmentation [11], robust alignment [58], nuclear norm minimization [59][60] and structure sparsity [61][62] where the used minimization functions are related to HQ optimization. The second avenue is to establish the relationship between different methods in these applications and to develop new methods by HQ optimization. Lastly, considering that linear representation (including sparse representation) methods in Appendix IV need a lot of samples to form a dictionary (or subspace), another avenue is to study the undersampling case for linear representation methods.

## Acknowledgment

The authors would like to greatly thank the associate editor and the reviewers for their valuable comments and advice, and thank Prof. Liang Wang, Dr. Xiao-Tong Yuan and Mr. Bo Dai for helping to revise the paper. This work is funded by the Research Foundation for the Doctoral Program of the Ministry of Education of China (Grant No. 20100041120009, 20110171120051), the National Basic Research Program of China (Grant No. 2012CB316300), National Natural Science Foundation of China (Grant No. 60736018, 61075024, 61103155, 61102111), and the Strategic Priority Research Program of the Chinese Academy of Sciences (Grant No. XDA06030300).

## References

[1] J. Wright, Y. Ma, J. Mairal, G. Sapiro, T. S. Huang, and S. Yan, "Sparse representation for computer vision and pattern recognition," *Proceedings of IEEE*, vol. 98, no. 6, pp. 1031–1044, 2010.

[2] B. Cheng, J. Yang, S. Yan, Y. Fu, and T. S. Huang, "Learning with $\ell_1$-graph for image analysis," *IEEE Transactions on Image Processing*, vol. 4, pp. 858–866, 2010.

[3] R. He, W.-S. Zheng, and B.-G. Hu, "Maximum correntropy criterion for robust face recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 8, pp. 1561–1576, 2011.

[4] E. J. Candés and T. Tao, "Near optimal signal recovery from random projections: universal encoding strategies," *IEEE Transactions on Information Theory*, vol. 52, no. 12, pp. 5406–5425, 2006.

[5] D. L. Donoho, "Compressed sensing," *IEEE Transactions on Information Theory*, vol. 52, no. 4, pp. 1289–1306, 2006.

[6] M. Aharon, M. Elad, and A. M. Bruckstein, "The k-svd: An algorithm for designing of overcomplete dictionaries for sparse representations," *IEEE Transactions on Signal Processing*, vol. 54, no. 11, pp. 4311–4322, 2006.

[7] J. Mairal, G. Sapiro, and M. Elad, "Learning multiscale sparse representations for image and video restoration," *SIAM Multiscale Modeling & Simulation*, vol. 7, no. 1, pp. 214–241, 2008.

[8] M. Wakin, J. Laska, M. Duarte, D. Baron, S. Sarvotham, D. Takhar, K. Kelly, and R. Baraniuk, "An architecture for compressive image," in *Proceedings of International Conference on Image Processing*, 2006, pp. 1273–1276.

[9] D. Takhar, J. Laska, M. Wakin, M. Duarte, D. Baron, S. Sarvotham, K. Kelly, , and R. Baraniuk, "A new compressive imaging camera architecture using optical-domain compression," in *Proceedings of Computational Imaging IV at SPIE Electronic Imaging*, 2006, pp. 43–52.

[10] W. Bajwa, J. Haupt, A. Sayeed, and R. Nowak, "Compressive wireless sensing," in *Proceedings of International Conference on Information Processing in Sensor Networks*, 2006.

[11] E. Elhamifar and R. Vidal, "Sparse subspace clustering: Algorithm, theory, and applications," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2013.

[12] J. Wright and Y. Ma, "Dense error correction via $\ell_1$-minimization," *IEEE Transactions on Information Theory*, vol. 56, no. 7, pp. 3540–3560, 2010.

[13] A. Y. Yang, S. S. Sastry, A. Ganesh, and Y. Ma, "Fast $\ell_1$-minimization algorithms and an application in robust face recognition: A reivew," in *Proceedings of International Conference on Image Processing*, 2010.

[14] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 2, pp. 210–227, 2009.

[15] R. He, W.-S. Zheng, B.-G. Hu, and X.-W. Kong, "A regularized correntropy framework for robust pattern recognition," *Neural Computation*, vol. 23, no. 8, pp. 2074–2100, 2011.

[16] M. Yang, L. Zhang, J. Yang, and D. Zhang, "Robust sparse coding for face recognition," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2011, pp. 625–632.

[17] N. H. Nguyen and T. D. Tran, "Exact recoverability from dense corrupted observations via $\ell_1$-minimization," *IEEE Transactions on Information Theory*, vol. 59, no. 4, pp. 2017–2035, 2013.

[18] S. Vaiter, G. Peyre, C. Dossal, and J. Fadili, "Robust sparse analysis regularization," *IEEE Transactions on Information Theory*, vol. 59, no. 4, pp. 2001–2016, 2013.

[19] S. Chen, D. L. Donoho, and M. Saunders, "Atomic decomposition by basis pursuit," *SIAM Review*, vol. 43, no. 1, pp. 129–159, 2001.

[20] E. J. Candés, J. Romberg, and T. Tao, "Stable signal recovery from incomplete and inaccurate measurements," *Communications on Pure and Applied Math*, vol. 59, no. 8, pp. 1207–1223, 2006.

[21] H. Zou, "The adaptive lasso and its oracle properties," *Journal of the American Statistical Association*, vol. 101, no. 476, pp. 1418–1429, 2006.

[22] E. J. Candés, M. Wakin, and S. Boyd, "Enhancing sparsity by reweighted $\ell_1$ minimization," *Journal of Fourier Analysis and Applications*, vol. 14, no. 5, pp. 877–905, 2008.

[23] T. Zhang, "Multi-stage convex relaxation for learning with sparse regularization," in *Proceedings of Neural Information Processing Systems*, 2008, pp. 16–21.

[24] W. Yin, S. Osher, D. Goldfarb, and J. Darbon, "Bregman iterative algorithms for $\ell_1$-minimization with applications to compressed sensing," *SIAM Journal on Imaging Sciences*, vol. 1, no. 1, pp. 143–168, 2008.

[25] P. L. Combettes and J.-C. Pesquet, "Proximal thresholding algorithm for minimization over orthonormal bases," *SIAM Journal on Optimization*, vol. 18, no. 4, pp. 1531–1376, 2007.

[26] R. Nowak and M. Figueiredo, "Fast wavelet-based image deconvolution using the EM algorithm," in *Proceedings of Asilomar Conference on Signals, Systems, and Computers*, vol. 1, 2001, pp. 371–375.

[27] D. L. Donoho and J. Tanner, "Observed universality of phase transitions in high-dimensional geometry, with implications for modern data analysis and signal processing," *Philosophical Transactions of The Royal Society A*, vol. 367, no. 1906, pp. 4273–4293, 2009.

[28] F. De la Torre and M. Black, "A framework for robust subspace learning," *International Journal of Computer Vision*, vol. 54, no. 1-3, pp. 117–142, 2003.

[29] S. Fidler, D. Skocaj, and A. Leonardis, "Combining reconstructive and discriminative subspace methods for robust classification and regression by subsampling," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 3, pp. 337–350, 2006.

[30] H. Jia and A. M. Martinez, "Support vector machines in face recognition with occlusions," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 136–141.

[31] W. F. Liu, P. P. Pokharel, and J. C. Principe, "Correntropy: Properties and applications in non-gaussian signal processing," *IEEE Transactions on Signal Processing*, vol. 55, no. 11, pp. 5286–5298, 2007.

[32] J. Idier, "Convex half-quadratic criteria and interacting auxiliary variables for image restoration," *IEEE Transactions on Image Processing*, vol. 10, no. 7, pp. 1001–1009, 2001.

[33] M. Nikolova and M. K. NG, "Analysis of half-quadratic minimization methods for signal and image recovery," *SIAM Journal on Scientific Computing*, vol. 27, no. 3, pp. 937–966, 2005.

[34] X.-X. Li, D.-Q. Dai, X.-F. Zhang, and C.-X. Ren, "Structured sparse error coding for face recognition with occlusion," *IEEE Transactions on Image Processing*, vol. 22, no. 5, pp. 1889–1900, 2013.

[35] N. H. Nguyen and T. D. Tran, "Robust lasso with missing and grossly corrupted observations," *IEEE Transactions on Information Theory*, vol. 59, no. 4, pp. 2036–2058, 2013.

[36] D. Geman and G. Reynolds, "Constrained restoration and recovery of discontinuities," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 14, pp. 367–383, 1992.

[37] D. Geman and C. Yang, "Nonlinear image recovery with half-quadratic regularization," *IEEE Transactions on Image Processing*, vol. 4, no. 7, pp. 932–946, 1995.

[38] S. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge University Press, 2004.

[39] P. L. Combettes and V. R. Wajs, "Signal recovery by proximal forwardbackward splitting," *SIAM Journal on Multiscale Modeling & Simulation*, vol. 4, no. 5, pp. 1168–1200, 2005.

[40] J. Bioucas-Dias and M. Figueiredo, "A new twist: Two-step iterative shrinkage/thresholding algorithms for image restoration," *IEEE Transactions on Image Processing*, vol. 16, no. 12, pp. 2992–3004, 2007.

[41] X. T. Yuan and B. G. Hu, "Robust feature extraction via information theoretic learning," in *Proceedings of International Conference on Machine Learning*, 2009, pp. 1193–1200.

[42] M. J. Fadili and J. L. Starck, "Sparse representation-based image deconvolution by iterative thresholding," in *Astronomical Data Analysis ADA*, 2006.

[43] M. Figueiredo, R. Nowak, and S. J. Wright, "Gradient projection for sparse reconstruction: application to compressed sensing and other inverse problems," *IEEE Journal of Selected Topics in Signal Processing: Special Issue on Convex Optimization Methods for Signal Processing*, vol. 1, no. 4, pp. 586–597, 2007.

[44] T. Ragheb, S. Kirolos, J. Laska, A. Gilbert, M. Strauss, R. Baraniuk, and Y. Massoud, "Implementation models for analog-to-information conversion via random sampling," in *Proceeding of Midwest Symposium on Circuits and Systems*, 2007, pp. 325–328.

[45] Z. Zhang, "Parameter estimation techniques: A tutorial with application to conic fitting," *Image and Vision Computing*, vol. 15, no. 1, pp. 59–76, 1997.

[46] J. C. Principe, D. Xu, and J. W. Fisher, "Information-theoretic learning," in *S. Haykin, editor, Unsupervised Adaptive Filtering, Volume 1: Blind-Souurce Separation. Wiley*, 2000.

[47] A. M. Martinez and R. Benavente, "The AR face database," Computer Vision Center, Tech. Rep., 1998.

[48] A. S. Georghiades, P. N. Belhumeur, and D. J. Kriegman, "From few to many: Illumination cone models for face recognition under variable lighting and pose," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 6, pp. 643–660, 2001.

[49] K.-C. Lee, J. Ho, and D. Kriegman, "Acquiring linear subspaces for face recognition under variable lighting," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 5, pp. 684–698, 2005.

[50] H. Lee, A. Battle, R. Raina, and A. Y. Ng, "Efficient sparse coding algorithms," in *Proceedings of Neural Information Processing Systems*, vol. 19, 2006, pp. 801–808.

[51] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM Journal on Imaging Sciences*, vol. 2, no. 1, pp. 183–202, 2009.

[52] D. L. Donoho and Y. Tsaig, "Fast solution of $l_1$-norm minimization problems when the solution may be sparse," *IEEE Transactions on Information Theory*, vol. 54, no. 11, pp. 4789–4812, 2008.

[53] M. Plumbley, "Recovery of sparse representations by polytope faces pursuit," in *Proceedings of International Conference on Independent Component Analysis and Blind Source Separation*, 2006, pp. 206–213.

[54] I. Naseem, R. Togneri, and M. Bennamoun, "Linear regression for face recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 11, pp. 2106–2112, 2010.

[55] L. Zhang, M. Yang, and X. Feng, "Sparse representation or collaborative representation: Which helps face recognition?" in *Proceedings of IEEE International Conference on Computer Vision*, 2011.

[56] S. Wright, R. Nowak, and M. Figueiredo, "Sparse reconstruction by separable approximation," in *Proceedings of IEEE Conference on Acoustics, Speech and Signal Processing*, 2008.

[57] D. L. Donoho and V. Stodden, "Breakdown point of model selection when the number of variables exceeds the number of observations," in *Proceedings of the International Joint Conference on Neural Networks*, 2006, pp. 16–21.

[58] A. Wagner, J. Wright, A. Ganesh, Z. Zhou, H. Mobahi, and Y. Ma, "Toward a practical face recognition system: Robust alignment and illumination by sparse representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 2, pp. 372–386, 2012.

[59] E. J. Candés, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis?" *Journal of the ACM*, vol. 8, no. 58, pp. 1–37, 2010.

[60] R. He, Z. Sun, T. Tan, and W.-S. Zheng, "Recovery of corrupted low-rank matrices via half-quadratic based nonconvex minimization," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2011, pp. 2889–2896.

[61] R. Jenatton, G. Obozinski, and F. Bach, "Structured sparse principal component analysis," in *Proceedings of Artificial Intelligence and Statistics*, 2010, pp. 366–373.

[62] R. He, T. Tan, L. Wang, and W.-S. Zheng, "$\ell_{2,1}$ regularized correntropy for robust feature selection," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 2504–2511.

[63] P. Hellier, C. Barillot, E. Memin, and P. Perez, "An energy-based framework for dense 3D registration of volumetric brain images," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2000.

[64] F. Champagnat and J. Idier, "A connection between half-quadratic criteria and EM algorithms," *IEEE Signal Processing Letters*, vol. 11, no. 9, pp. 709–712, 2004.

[65] M. Allain, J. Idier, and Y. Goussard, "On global and local convergence of half-quadratic algorithms," *IEEE Transactions on Image Processing*, vol. 15, no. 5, pp. 1030–1042, 2006.

[66] X. T. Yuan and S. Li, "Half quadratic analysis for mean shift: with extension to a sequential data mode-seeking method," in *Proceedings of IEEE International Conference on Computer Vision*, 2007.

[67] R. Chartrand, "Exact reconstruction of sparse signals via nonconvex minimization," *IEEE Signal Processing Letters*, vol. 14, no. 10, pp. 707–710, 2007.

[68] I. Daubechies, R. DeVore, M. Fornasier, and C. Gunturk, "Iteratively re-weighted least squares minimization for sparse recovery," *Communications on Pure and Applied Mathematics*, vol. 63, no. 1, pp. 1–38, 2010.

[69] S. Seth and J. C. Principe, "Compressed signal reconstruction using the correntropy induced metric," in *Proceedings of IEEE Conference on Acoustics, Speech and Signal Processing*, 2008, pp. 3845–3848.

[70] R. Chartrand and W. Yin, "Iteratively reweighted algorithms for compressive sensing," in *Proceedings of IEEE Conference on Acoustics, Speech and Signal Processing*, 2008, pp. 3869–3872.

[71] G. Davis, S. Mallat, and M. Avellaneda, "Adaptive greedy approximations," *Journal of Constructive Approximation*, vol. 13, pp. 57–98, 1997.

[72] I. Daubechies, M. Defrise, and C. De Mol, "An iterative thresholding algorithm for linear inverse problems with a sparsity constraint," *Communications in Pure and Applied Mathematics*, vol. 57, pp. 1413–1457, 2006.

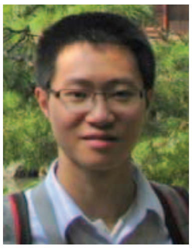
[73] M. Fornasier, "Theoretical foundations and numerical methods for sparse recovery," *Radon Series on Computational and Applied Mathematics*, vol. 9, pp. 1–121, 2010.

[74] S. Z. Li, "Face recognition based on nearest linear combinations," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 1998, pp. 839–844.

[75] Q. Shi, A. Eriksson, A. van den Hengel, and C. Shen, "Face recognition really a compressive sensing problem," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2011, pp. 553–560.

[76] X. Sun, *Matrix Perturbation Analysis*. Chinese Science Press, 2001.

**Ran He** received the B.E. degree in Computer Science from Dalian University of Technology, the M.S. degree in Computer Science from Dalian University of Technology, and Ph.D. degree in Pattern Recognition and Intelligent Systems from Institute of Automation, Chinese Academy of Sciences in 2001, 2004 and 2009, respectively. Since September 2010, Dr. He has joined NLPR where he is currently Associate Professor. He currently 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.



**Wei-Shi Zheng** received his Ph.D. degree in Applied Mathematics at Sun Yat-sen University, China, 2008. After that, he has been a Postdoctoral Researcher on the European SAMURAI Research Project at the Department of Computer Science, Queen Mary University of London, UK. He has now joined Sun Yat-sen University as an Associate Professor under the one-hundred-people program of Sun Yat-sen University. He has published widely in IEEE TPAMI, IEEE TNN, IEEE TIP, Pattern Recognition, IEEE TSMC-B, IEEE TKDE, ICCV, CVPR and AAAI. His current research interests are in object association and categorization for visual surveillance. He is also interested in discriminant/sparse feature extraction, dimension reduction, transfer learning, and face image analysis.

**Tieniu Tan** received his B.Sc. degree in electronic engineering from Xi'an Jiaotong University, China, in 1984, and his M.Sc. and Ph.D. degrees in electronic engineering from Imperial College London, U.K., in 1986 and 1989, respectively. He was the Director General of the CAS Institute of Automation from 2000-2007, and has been Professor and Director of the NLPR since 1998. He also serves as 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 9 books. He holds more than 30 patents. His current research interests include biometrics, image and video understanding, and information forensics and security.

Dr Tan is a Fellow of the IEEE and the IAPR (the International Association of Pattern Recognition). He has served as chair or program committee member for many major national and international conferences. He currently serves as Vice President of the IAPR, the Executive Vice President of the Chinese Society of Image and Graphics, Deputy President of the Chinese Association for Artificial Intelligence, and was 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.

**Zhenan Sun** received the BE degree in industrial automation from Dalian University of Technology in 1999, the MS degree in system engineering from Huazhong University of Science and Technology in 2002, and the PhD degree in pattern recognition and intelligent systems from CASIA in 2006. He is an associate professor in the Institute of Automation, Chinese Academy of Sciences (CASIA). In March 2006, he joined the Center of Biometrics and Security Research (CBSR) in the National Laboratory of Pattern Recognition (NLPR) of CASIA as a faculty member. He is a member of the IEEE and the IEEE Computer Society. His research focuses on biometrics, pattern recognition, and computer vision.