Robust Nuclear Norm-based Matrix Regression with Applications to Robust Face Recognition

Jianchun Xie, Jian Yang, Member, IEEE, Jianjun Qian, Ying Tai, and Hengmin Zhang

Abstract—Face recognition (FR) via regression analysis based classification has been widely studied in the past several years. Most existing regression analysis methods characterize the pixel-wise representation error via $l_1$-norm or $l_2$-norm, which overlook the two-dimensional structure of the error image. Recently, the nuclear norm based matrix regression (NMR) model is proposed to characterize low-rank structure of the error image. However, the nuclear norm cannot accurately describe the low-rank structural noise when the incoherence assumptions on the singular values does not hold, since it over-penalizes several much larger singular values. To address this problem, this paper presents the robust nuclear norm to characterize the structural error image and then extends it to deal with the mixed noise. The majorization-minimization (MM) method is applied to derive a iterative scheme for minimization of the robust nuclear norm optimization problem. Then, an efficiently alternating direction method of multipliers (ADMM) method is used to solve the proposed models. We use weighted nuclear norm as classification criterion to obtain the final recognition results. Experiments on several public face databases demonstrate the effectiveness of our models in handling with variations of structural noise (occlusion, illumination, etc.) and mixed noise.

Index Terms—robust nuclear norm, nuclear norm, robust matrix regression, ADMM, face recognition.

I. INTRODUCTION

FACE recognition can be considered as one of the most visible and challenging problems in computer vision and pattern recognition. In the past two decades, numerous face classification methods have been proposed by many scholars [1–10]. Among these methods, regression analysis based methods have achieved promising results.

Naseem et al. [12] presented the linear regression classifier (LRC) for FR, which aims to find a suitable representation of the testing face image, and classify it by checking which class can get a better representation than other classes. In order to avoid over-fitting, a regularization term is imposed upon the linear regression model. With the $l_1$-norm regularization on the coding coefficients, Wright et al. [1] presented a sparse representation based classification (SRC) scheme for FR. Zhang et al. [2] showed that it is not necessary to impose the $l_1$-norm on the coding coefficients, and proposed the collaborative representation classifier (CRC) based on ridge regression. CRC can achieve similar results as SRC but significantly speed up the algorithm. In these regression analysis methods, the representation residual is usually measured by the $l_2$-norm or $l_1$-norm, which assumes that the pixels of error follow Gaussian or Laplacian distribution independently [13]. However, the assumption might be unreasonable in some real situation because of occlusion, disguise, or illumination variation.

To be better character the error image, several improved regression methods are proposed by some researchers. For example, Yang et al. [3] proposed a robust sparse coding (RSC) method for FR, by seeking for an maximum likelihood estimation solution of the sparse coding problem. Based on the maximum correntropy criterion, He et al. [14, 15] presented robust sparse representation for FR. More recently, He et al. [16] made an effort to use different half-quadratic functions to measure the error image and unify it into the SRC scheme. In addition, Naseem et al. [5] further extended their LRC to the robust linear regression classification (RLRC) by using the Huber estimator to deal with severe random pixel noise and illumination variations. Zhou et al. [17] incorporated the markov random field model into the sparse representation framework for spatial continuity of the occlusion. Motivated by recent advances in structured sparsity, Jia et al. [18] proposed a structured sparse representation classifier (SSRC) by introducing a class of structured sparsity-inducing norm into the SRC framework. All of the above mentioned regression based methods have achieved promising results in FR.

However, these regression methods all use the 1-D pixel-based error model to address face classification problem. Actually, characterizing the error pixel by pixel neglects the whole structural information and relationship of the error image, such as occlusion, extreme illumination [1–3]. To overcome this limitation, Yang et al. [19] and Xu et al. [20] found that occlusion and illumination changes generally lead to a low-rank or approximately low-rank error image. As shown in Fig.1, (a) is a clear image which can be viewed as the expected, reconstructed image, and (b) is the corresponding occluded image. The error image (c), the difference between (a) and (b), is low-rank or nearly low-rank. To make use of this low-rank structure of the error image, Yang et al. proposed the nuclear norm based matrix regression (NMR) model for face representation and classification. The initial
idea of NMR is to minimize the rank function of the error matrix. It is known that the rank minimization problem is NP hard. As a common practice in rank minimization problems, Yang et al. [19] replaced the rank function with the nuclear norm to characterize the low rank structure of the error image. Qian et al. [21] proposed the robust nuclear norm regularized regression based classification (RNR) method for FR. RNR took advantage of the structural characteristics of noise and provided a unified framework for integrating error detection and error support into one regression model. The experimental results demonstrated that the RNR is robust to real disguise and random block occlusion.

As a convex relaxation of the rank function, however, the nuclear norm cannot accurately approximate the rank of matrix. Theoretically, if the incoherence assumptions on the singular values does not hold, solving the convex nuclear norm problem will lead to a suboptimal low rank solution [22]. As we know, the rank of a matrix is $l_0$-norm of the singular value vector, i.e. $\text{rank}(E) = \sum_i I(d_i > 0)$. The nuclear norm of a matrix is $l_1$-norm of the singular value vector, i.e. $\|E\|_* = \sum_i d_i = \sum_i d_i \times I(d_i > 0)$, which can be considered as the weighted rank function of the matrix. Under the condition that the singular values all are one, the nuclear norm will be the rank function, otherwise the nuclear norm will result in a biased estimator. Therefore, there is still big gap between the nuclear norm minimization and the rank minimization.

Inspired by Lu et al. [23], which used a family of nonconvex surrogates of $l_0$ norm on the singular values of matrix to approximate the rank function. This paper substitutes the nuclear norm with non-convex function for characterizing the low rank structure of the error image. Based on this characterization, a robust nuclear norm based matrix regression model is built. Furthermore, for the mixed noise, we assume the mixed noise is an additive combination of two components: sparse noise and structural noise. Then, we adopt the $l_1$-norm and the robust nuclear norm for measuring the two components, respectively. The ADMM has been successfully applied to optimize our model.

In summary, the contributions of this paper include the following aspects:

1) We introduce the robust nuclear norm to describe the error image, which does not need the assumption that the representation errors of pixels are statistically independent. Therefore, the error image with occlusion and illumination variation can be characterized by its low-rank structure.

2) We provide two robust matrix regression models for dealing with the structural noise and mixed noise, by adopting the robust nuclear norm to describe the structural noise and $l_1$-norm of matrix to describe sparse noise, respectively.

3) Based on the MM algorithm, we minimize the robust nuclear norm of error matrix by solving the minimization of the weighted nuclear norm and employ ADMM to solve the models effectively. The convergence of the proposed algorithm is further analyzed in this paper.

This paper is an extended version of our conference paper [11]. The remainder of the paper is organized as follows: In Section II, we review the NMR model and point out the problems of the nuclear norm as error description. In Section III, we give the definition of the robust nuclear norm through the non-convex functions, and then propose two different robust matrix regression model for structural noise and mixed noise, respectively. The ADMM for solving the models and the RMR based classifier are also given in this section. In section IV, we conduct experiments to compare with the regression analysis based methods. Advices on parameters and the choice of the non-convex functions are given in this section. Finally, some conclusions are given in Section V.

Notations: Throughout this paper, $R^n$ denotes the space of $n$-dimensional real column vectors, and $R^{m_1 \times m_2}$ denotes the space of $m_1 \times m_2$ dimensional real matrices. For a matrix $E \in R^{m_1 \times m_2}$ (assuming $m_2 \leq m_1$), we write its singular value decomposition (SVD) as $E = UDV^T$ with $U \in R^{m_1 \times m_2}$, $V \in R^{m_2 \times m_2}$ and $D = \text{diag}(d_i, i = 1, 2, \ldots, m_2)$, where $d_i(E)$ is the $i$-th largest singular value of $E$. $d(E)$ is the rank function of $E$, i.e. the $l_0$-norm of the singular value vector. $\|E\|_* = \sum_{i=1}^{m_2} d_i$ is the nuclear norm of a matrix $E$, and $\|E\|_1 = \sum_{i,j} |E_{ij}|$ is the $l_1$-norm of matrix. $\text{Vec}()$ denotes an operator converting a matrix to a vector by column. The $l_1$-norm and $l_2$-norm of a vector are defined by $\|x\|_1 = \sum |x_i|$ and $\|x\|_2 = (\sum x_i^2)^{1/2}$. The indicator function is defined as $I_A(x) = \begin{cases} 1 & x \in A \\ 0 & x \notin A \end{cases}$.

II. NUCLEAR NORM FOR ERROR DESCRIPTION

In this section, we briefly review NMR method and then point out the problem of the nuclear norm based the error description.

A. Nuclear Norm based Matrix Regression

Given a set of image matrices $A_1, A_2, \ldots, A_n \in R^{m_1 \times m_2}$, an image matrix $Y \in R^{m_1 \times m_2}$ can be represented linearly as

$$Y = x_1 A_1 + x_2 A_2 + \cdots + x_n A_n + E,$$

where $x_1, x_2, \ldots, x_n$ is a set of coding coefficients and $E$ is the residual image. For convenience, one can define the linear mapping $A(x) = x_1 A_1 + x_2 A_2 + \cdots + x_n A_n$, then the formula (1) becomes $Y = A(x) + E$.

As shown in Fig. 1 (c), we can find that the residual image $E = Y - A(x)$ is nearly low rank. Therefore, the low rank constraint on the error matrix can be formulated as:

$$\min_x \text{rank}(Y - A(x)).$$

![Image of low-rank structures](image-url)
However, the rank minimization problem is usually NP-hard. Based on the fact that the nuclear norm is the best convex relaxation of the rank function [22], Yang et al. [19] presented the minimization of the nuclear norm of the error matrix as a criterion, which can be expressed as:

$$\min_{x} \| Y - A(x) \|_2,$$  \hspace{1cm} (3)

by posing an $l_2$-norm regularization term on coding coefficients, the NMR model can be expressed as follows:

$$\min_{x} \| Y - A(x) \|_2 + \frac{\lambda}{2} \| x \|_2^2. \hspace{1cm} (4)$$

### B. Problem of Nuclear Norm based Error Description

NMR adopts the nuclear norm as a substitute of the rank function to describe the low-rank structural error and demonstrates to be effectiveness in practice. It has been proved that under certain incoherence assumptions on the singular values, solving the convex nuclear norm problem leads to a near optimal low-rank solution [22]. However, such assumptions may be violated in real situation, and the obtained solution by using the nuclear norm may be suboptimal since the nuclear norm is not a perfect approximation of the rank function. A similar conclusion has been proved in the convex $l_1$-norm and non-convex $l_0$-norm for sparse vector recovery [24]. Fan and Li [25] proposed three criteria for good heuristic for non-convex $l_0$-norm: unbiasedness, sparsity and continuity. The $l_1$-norm satisfies both sparsity and continuity, but it is biased. Therefore, using the nuclear norm results in a biased estimator for rank function. In summary, using the nuclear norm as description for the low rank structure of the error matrix is not optimal.

Actually, the nuclear norm of a matrix is the sum of all singular values, so the nuclear norm generally over-emphasizes large singular value in practice. Using the nuclear norm as the error measurement for classifier maybe unsuitable. In the following, we will give an example to illustrate this phenomenon. Fig. 2 shows three face images from the Extended Yale B database with different illumination. The face images $A$, $A1$ come from the same class, and $B$ is from different class. By calculating, $\| B - A \|_2 = 51.01 < \| A1 - A \|_2 = 66.45$, adopting the nearest neighbor classifier, the image $B$ will be misclassified to the class of $A$. Fig. 2 (d) shows the singular values of the error images $B - A$ and $A1 - A$. one can find that the biggest singular value of the error matrix $A1 - A$ has an important influence on the nuclear norm, and plays an important role in the misclassification. So we adopt the weighted version of the nuclear norm for classifier measurement.

### III. Robust Nuclear Norm for Matrix Regression

In this section, we will introduce the definition of the robust nuclear norm and adopt the robust nuclear norm as the metric of the error matrix, then we will present the robust matrix regression model to code the image and use the ADMM to solve the models. Subsequently, we also provide the complexity analysis of the proposed algorithm.
Since the function $\rho(\cdot)$ satisfies the properties of non-decrease and non-convex, we can find that the first-order Taylor expression of $\rho(d_i)$ in the proximity point $d_i^{(t)}$ satisfies
\[
\rho(d_i) \leq \rho(d_i^{(t)}) + \rho'(d_i^{(t)})(d_i - d_i^{(t)}),
\]
where $d_i^{(t)}$ is the $i$-th biggest singular value of $E$ at the $t$-th iteration. So
\[
F(d) \leq F(d^{(t)}) + F'(d^{(t)})^T(d - d^{(t)}),
\]
where $d = (d_1, d_2, \ldots, d_{m_2})^T$ is the singular vector and $F'(d^{(t)})$ is the first order derivative of $F(d)$ with respect to $d$. For ease of presentation, unless otherwise specified, we assume the right derivative of $\rho(\cdot)$ at 0 is finite. Let $J(d|d^{(t)}) = F(d^{(t)}) + F'(d^{(t)})^T(d - d^{(t)})$, the original non-convex problem can be processed by iteratively solving the minimization of the $J(d|d^{(t)})$. In detail, the MM algorithm proceeds by repeating two steps:

1. Construct a convex upper bound $J(d|d^{(t)})$ for non-convex prior term;
2. Minimize the upper bound until the sequence converges.

**Theorem 1:** [30] For a differential non-convex function $\rho(\cdot)$ on $[0, \infty]$, we have
\[
F(d) \leq J(d|d^{(t)}) \text{ and } F(d^{(t)}) = J(d^{(t)}|d^{(t)}),
\]
Furthermore, the LLA has descent property, that is,
\[
F(d^{(t+1)}) \leq F(d^{(t)}), \quad t = 1, 2, \ldots
\]
If the function $\rho(\cdot)$ is strictly concave then we always take “$<$” in Eq. (10).

By omitting constant terms from the upper bounding function $J(d|d^{(t)})$, the solution of $\min_x F(d)$ equals to $\min_x F'(d)^T d = \min_x \sum_{i=1}^{m_2} \rho'(d_i^{(t)}) d_i$. The $\rho'(d_i^{(t)})$ can be viewed as the weighted function of $d_i$, so the primordial objection function (6) can be solved by
\[
\min_{x} F'(d^{(t)})^T d = \min_x \sum_{i=1}^{m_2} u_i^{(t)} d_i = \|E\|_{w,*}^p,
\]
where $\rho(\cdot)$ is non-decreasing and non-convex function, then $u_i^{(t)} \geq 0$ and $u_i^{(t)}$ is decreasing with $d_i$.

**C. Robust Nuclear Norm based Matrix Regression**

According to the MM algorithm, the Robust Matrix Regression (RMR) can be expressed as follows:
\[
\min_{x} \|Y - A(x)\|_{w,*},
\]
in order to avoid over fitting, we pose regularization on the coding coefficients, and get the following regularized model
\[
\min_{x} \|Y - A(x)\|_{w,*} + \frac{\lambda}{2} \|x\|_p^p,
\]
where $\lambda > 0$ is a regularization parameter, if $p = 1$ (or $p = 2$), the last term in (13) becomes the $l_1$-norm (or the square of the $l_2$-norm) regularization pose on the coding coefficients. In this paper we name these two models as RMRL1 and RMRL2 respectively. Next, we develop the optimization algorithm to solve Eq. (13) by using ADMM.

**D. Optimization via ADMM**

In this section, we adopt ADMM to solve Eq. (13) efficiently. For more details of ADMM, one can refer to [31]. To deal with our problem, the model can be transformed as the following constrained problem:
\[
\min_{x} \|E\|_{w,*} + \frac{\lambda}{2} \|x\|_p^p \quad \text{s.t.} \quad Y - A(x) = E,
\]

The Augmented Lagrange function $L_{\mu}$ is expressed as
\[
L_{\mu}(E, x, Z) = \|E\|_{w,*} + \frac{\lambda}{2} \|x\|_p^p + \text{Tr}(Z^T(A(x) + E - Y)) + \frac{\mu}{2} \|A(x) + E - Y\|_F^2,
\]
where $\mu > 0$ is a penalty parameter, and $Z$ is the Lagrange multiplier. $\text{Tr}(\cdot)$ denotes the trace operator. Noting that Eq. (13) and Eq. (16) are the critical optimizations problems and Eq. (18) is a proximal minimization step of the Lagrange multiplies $Z$.

In detail, updating $x$: Fixing $E^{(t)}$ and $Z^{(t)}$, the solution of $x$ by
\[
\min_{x} \|A(x) - (Y - E^{(t)} - \frac{1}{\mu} Z^{(t)})\|_F^2 + \frac{\lambda}{\mu} \|x\|_p^p,
\]
denoting $H = [\text{Vec}(A_1), \ldots, \text{Vec}(A_{n})]$, $A(x)$ can be written as $Hx$ in the context of Frobenius norm. Letting $b^{(t)} = \text{Vec}(Y - E^{(t)} - \frac{1}{\mu} Z^{(t)})$, Eq. (19) is equivalent to
\[
x^{(t+1)} = \arg \min_{x} \|Hx - b^{(t)}\|_2^2 + \frac{\lambda}{\mu} \|x\|_p^p,
\]
different settings of $p$ lead to different instantiations of the representation model. When setting $p = 1$, the problem (20) can be solved by $l_1$ minimization solvers, and when setting $p = 2$, the problem (20) is a standard ridge regression model.

Updating $E$: First, we update the weighted vector $w^{(t)}$. By using the singular value decomposition, we get $E^{(t)} = U^{(t)} D^{(t)} V^{(t)^T}$, and update $w_i^{(t)} = \rho'(d_i^{(t)})$. Second, fixing $x^{(t+1)}$ and $Z^{(t)}$, the solution of $E$ by
\[
\min_{E} \frac{1}{2} \|E - (Y - A(x^{(t+1)} - \frac{1}{\mu} Z^{(t)})\|_F^2 + \frac{1}{\mu} \|E\|_{w,*},
\]
this optimal problem is normally solved via the singular value thresholding (SVT) operator with weighted vector $w^{(t)}$.

**Theorem 2:** [32] For each $0 \leq w_1 \leq \cdots \leq w_{m_2}$ and $Y \in R^{m_1 \times m_2}$, the weighted nuclear norm minimization problem
\[
\min_{x} \frac{1}{2} \|X - Y\|_F^2 + \tau \|X\|_{w,*}
\]
has the solution
\[
\hat{X} = USV^T,
\]
Algorithm 1 ADMM Algorithm for RMR

**Input:** An image matrix $Y$, a set of training images $A_1, A_2, \cdots, A_n$. The model parameters $\lambda, \mu$, the termination condition parameters $\varepsilon_1$ and $\varepsilon_2$.

$H = [\text{Vec}(A_1), \cdots, \text{Vec}(A_n)]$, $Z(0) = 0 \in R^{m_1 \times m_2}$, while not converged do

1. Updating $x(t+1)$.
   - if $p = 1$, $x^{(t+1)} = \arg\min \|Hx - b(t)\|^2 + \frac{\lambda}{2} \|x\|_1$, can be solved by $l_1$ magic [33];
   - if $p = 2$, $x$ can be solved with standard ridge regression, that is $x^{(t+1)} = (H^TH + \frac{\lambda}{\mu} I)^{-1} H^T b(t)$;
2. Enhance low rank using Algorithm 2 and get solution $E^* \rightarrow E^{(t+1)}$;
3. Updating $Z(t+1)$ by Eq. (18).

end while

**Output:** Optimal regression coefficient vector $x^*$.

Algorithm 2 The reweighted nuclear norm minimization algorithm for solving minimization problem (17)

**Input:** $E(0)$, decompose $E(0) = U(0) D(0) V_T^T(0)$, $D(0) = \text{diag}(d_{i0}, i = 1, \cdots, m_2)$.

while not converged do

1. Updating $w_i(k); w_i(k) = \rho'(d_i(k))$;
2. Updating $E_{(k+1)}; E_{(k+1)} = U(k) S_{w(k)} D(k) V_T^T(k)$.

end while

**Output:** Solution $E^*$.

where $Y = UDV^T$ is the SVD of $Y$, and $S_{w,r}(D)$ is the SVT operator with weighted vector $w$,

$$S_{w,r}(D) = \text{diag}(\max(d_i - w_i, 0), i = 1, \cdots, m_2).$$

As suggested in [32], we need to choose the proper termination parameters $\varepsilon_1$ and $\varepsilon_2$, and use the following termination conditions:

$$\|A(x^{(t+1)}) + E^{(t+1)} - Y\|^2_F / \|Y\|^2_F < \varepsilon_1,$$

$$\max(\|x^{(t+1)} - x^{(t)}\|^2_F, \|E^{(t+1)} - E^{(t)}\|^2_F) / \|Y\|^2_F < \varepsilon_2.$$  

Thus, we propose the iterative scheme for robust matrix regression by alternatively solving $E, x$, and update weighted function $w$. The whole algorithms for RMRL1 and RMRL2 are summarized in Algorithm 1, and as the sub-problem, the algorithm for updating $E$ is summarized in Algorithm 2.

E. Extended RMR for dealing with mixed noise

In real-world applications, some data could be corrupted by sparse noise and structural noise due to the acquisition error. So, the mixed noise can be modeled as an additive combination of two independent components, we can decompose the mixed noise and characterize them respectively. In this subsection, we will propose the sparse regularized RMR for dealing with mixed noise.

In the mixed noise case, we assume that the error image matrix $E$ can be decomposed into a low rank part and a sparse part. Therefore, we can build the following Sparse regularized Robust Matrix Regression model (S-RMR):

$$\min_{E, h, x} \|E\|_{w,b} + \alpha \|E_s\|_1 + \beta \|x\|_p$$

$$s.t. \ Y - A(x) = E + h,$$

where $\|E\|_{w,b}$ characterizes the structural noise and $\|E_s\|_1$ characterizes the sparse noise, respectively. If $p = 1$ (or $p = 2$), the last term in (26) becomes the $l_1$-norm (or the square of the $l_2$-norm). We abbreviate our model as S-RMRL1 and S-RMRL2.

Eq.(26) is a constrained optimization problem which can be solved by the extended ADMM method. By introducing auxiliary variable $u$ which satisfy $u = x$, the corresponding augmented Lagrangian function is defined as

$$L_{\mu} = \|E\|_{w,b} + \alpha \|E_s\|_1 + \beta \|u\|_p^p$$

$$+ \text{Tr}(Z^T(Y - A(x) - E_t - E_s)) + \frac{\mu}{2} \|Y - A(x) - E_t - E_s\|^2_F + \|x - u\|^2_2),$$

where $\alpha > 0$ and $\beta > 0$ are constants that determine the penalty for large representation error and coding coefficients, $Z_1, Z_2$ are Lagrange multipliers. The ADMM takes advantage of the separable forms with respect to $E_t, E_s$ and $x$, we propose iterative scheme for S-RMRL model, the whole algorithm is summarized in Algorithm 3.

In detail, according to Theorem 2, one can get the optimal solution of $E_t$ by Algorithm 2 with $D = Y - A(x(t)) - E_s^{(t)} + \frac{1}{\mu} Z^{(t)}$. The optimal solution of $E_s$ is given by

$$E_s^{(t+1)} = \mathbb{D}_{\mu}(Y - A(x(t)) - E_l^{(t+1)} + \frac{1}{\mu} Z^{(t)}),$$

where the function $\mathbb{D}_\mu$ is the shrinkage operator defined as

$$\mathbb{D}_\mu(v)_{ij} = \text{sign}(v_{ij}) \cdot \max(\|v_{ij}\| - \mu, 0).$$

The optimal solution of $x$ can be obtained by:

$$x^{(t+1)} = (H^T H + I)^{-1} H^T g^{(t+1)} + u^{(t)} - \frac{1}{\mu} z^{(t)},$$

where $g^{(t+1)} = \text{Vec}(Y - E_s^{(t+1)} - E_l^{(t+1)} + \frac{1}{\mu} z^{(t)}).$

For step 4 in Algorithm 3, the optimal solution can be obtained by calculating the stationary point for the corresponding objective function,

$$u^{(t+1)} = \begin{cases} \mathbb{D}_\mu \left(\frac{u^{(t)} + \frac{1}{\mu} z^{(t)}}{\mu} \right), & p = 1 \\
\left(1 + \frac{2\mu}{\mu} \right) \frac{1}{\mu} z^{(t)}, & p = 2. \end{cases}$$

Algorithm 1 and Algorithm 3 can be interpreted as using multi-step iterations strategy. As we can see, each step of the LLA procures results in a weighted nuclear norm minimization, and the ADMM algorithm repeatedly calls the SVT operator with weighted vector in the Algorithm 2. The computation costs can be more than the ADMM algorithm for solving NMR. To alleviate the computations, we adopt one-step LLA strategy, which is studied in [34]. The one-step LLA runs the Algorithm 2 only once instead of waiting to converge or reach the maximum number of iterations. Our experiments reveal that the one-step method is as efficient as the fully iterative method but saves much more time.
Algorithm 3: Solve S-RMR Algorithm via ADMM

Input: An image matrix $Y$, a set of training images $A_1, A_2, \cdots, A_n$. The model parameters $\alpha, \beta$, the termination condition parameters $\varepsilon_1, \varepsilon_2$.

$H = [\text{Vec}(A_1), \cdots, \text{Vec}(A_n)], \quad Z(0) = E(s(0) = 0 \in R_{m_1 \times n_2}^n$.

while not converged do

1. Updating $E_t$: $E_t^{(t+1)} = \arg\min_{E_t} \frac{1}{\mu} \|E_t\|_{\alpha(t), \ast} + \frac{1}{2} \|E_t - (Y - A(x(t)) - E_{s(t)} + \frac{1}{\mu} Z_{s(t)}^{(t)})\|_F^2$.

2. Updating $E_s$: $E_s^{(t+1)} = \arg\min_{E_s} \frac{1}{\mu} \|E_s\|_1 + \frac{1}{2} \|E_s - (Y - A(x(t)) - E_{s(t+1)} + \frac{1}{\mu} Z_{s(t)}^{(t)})\|_F^2$.

3. Updating $x$: $x(t+1) = \arg\min_x \|Y - A(x) - E_{s(t+1)} - E_{s(t)} + \frac{1}{\mu} Z_{s(t)}^{(t)}\|_F^2 + \|x - u(t) + \frac{1}{\mu} Z_{s(t)}^{(t)}\|_2^2$.

4. Updating $u$: $u^{(t+1)} = \arg\min_u \frac{1}{\mu} \|u\|_p^p + \frac{1}{2} \|u - (x(t+1) + E_{s(t+1)} - E_{s(t)})\|_2^2$.

5. Updating $Z_{1}$ and $Z_{2}$: $Z_{1(t)}^{(t+1)} = Z_{2(t)}^{(t)} + \mu(Y - A(x(t+1)) - E_{s(t+1)} - E_{s(t)})$, $Z_{2(t)}^{(t+1)} = Z_{2(t)}^{(t)} + \mu(x(t+1) - u(t+1))$.

end while

Output: Optimal regression coefficient vector $x^*$.

F. Complexity and convergence analysis

We now discuss the computational complexity of Algorithms using the one-step method. The computational complexity is mainly determined by the singular value decomposition and the matrix multiplications. Suppose that the training number is $n$ and the image size is $m_1 \times m_2$. For convenience, we assume that $m_2 \leq m_1$, then the computational complexity for performing SVD on $m_1 \times m_2$ matrix is $O(m_1 m_2^2)$, and the complexity of matrix multiplications is $O(n m_1 m_2 + n^2)$. Thus the total time complexity of the Algorithm 1 for RMRL2 and Algorithm 3 for S-RMRL1, S-RMRL2 is $O((m_1 m_2^2 + nm_1 m_2 + n^2))$, where $i$ is the number of iteration.

For RMRL1, the coding coefficients are solved by an $l_1$-regularized problem. It is reported that the $l_1$ magic for solving $l_1$ minimization have an empirical complexity of $O(m_1^2 m_2 n^{1.3})$ [35], so the computational complexity of RMRL1 is $O(t(m_1 m_2^2 + nm_1 m_2 + n^2 + m_1^2 m_2 n^{1.3}))$.

When $p = 1$ or $p = 2$, the solution of the problem (15) is a saddle point of the following Lagrangian function: $L = \|E_{w(t)}\|_p + \lambda \|x_{p(t)}\|_p + \text{Tr}(Z_p^{(t)}(A(x) - E - Y))$. According to the optimization theory, finding the optimal solutions of original and dual problem is equivalent to find a saddle point of the function $L$. The solution satisfies the KKT conditions of (14),

\[
\begin{align*}
Z^* &\in \partial\|E^*\|_{\alpha, \ast}, \\
\lambda x^* + H\text{Vec}(Z^*) &= 0, \\
A(x^*) - Y &= E^*.
\end{align*}
\]

Theorem 3: If $\mu > 0$, the sequence $(E^{(t)}, x^{(t)}, Z^{(t)})$ generated by Algorithm 1 converges to $(E^*, x^*, Z^*)$, where $(E^*, x^*)$ is a solution of (13).

The proof and an convergence example are given in supplemental materials.

G. RMR based classification

Similar to the strategy of NMR [19], given the training samples $A_1, A_2, \cdots, A_n$ from different classes, a new testing sample $Y$ can be approximated by the linear span $Y = x_1 A_1 + \cdots + x_n A_n$. Based on the optimal solution $x^*$, one can get the reconstructed image of $Y$ as $\hat{Y} = A(x^*)$. Let $\delta_l : R^n \rightarrow R^n$ be the characteristic function that selects the coefficients associated with the $i$-th class. One can get the reconstruction of $Y$ in class $i$ as $\hat{Y}_i = A(\delta_l(x^*))$. The corresponding class reconstruction error is defined by

\[
e_i(Y) = \|A(x^*) - A(\delta_l(x^*))\|_{w_{i,final}}^s, \\
\text{where} \|E\|_{w_{i,final}}^s = \sum_{i=1}^{m_2} w_{i,final} d_i, \text{ w_{final} is the final weighted vector.}
\]

The decision rule is

\[
\text{identity}(Y) = \arg\min_i e_i(Y).
\]

IV. EXPERIMENTS

In this section, standard face databases are selected to evaluate the effectiveness and robustness of our algorithms. Several competitive face recognition methods including LRC[12], CRC[2], SRC[1], SSRC[18], RLRC[5], CESR[14], RSC[3], half-quadratic with the additive form HQ-A[16], half-quadratic with the multiplicative form HQ-M[16], NMR[19] and RNR[21] are tested as comparisons. In our experiments, we use the two RMR algorithms for structural noise caused by occlusion, illumination, and real disguise, and use the two S-RMR algorithms for mixed noise. The parameter settings of all the methods follow the author’s suggestions. Because the log-sum function need less parameters and the recognition rates can nearly achieve the same value comparing with other non-convex function, for convenience, we choose the log-sum as the non-convex function in our experiments. The databases are described in supplemental materials.

A. Experiments on the Extended Yale B Database

1) Face recognition with different Illumination: As we know, the extreme illumination variation is a challenging task for most FR methods. In this subsection, we design two training modes to test the performance of RMR method under different illumination conditions. The first is the “single training sample” protocol, and the second is the “multi training samples” protocol. It should also be noted that for the single training sample protocol, we choose the first image of each person from Subset 1 for training, and for multi training samples protocol, we choose Subset 1 for training. Subsets 4 and 5 shown in Fig. 3 with extreme lighting conditions are used for testing, respectively. The experimental results of all methods are shown in Fig. 4.

From Fig. 4, we find that RMR achieves better results than the other methods no matter the training sample is single or not. Especially, for Subset 5, the recognition rate of our method is 93.7% (RMRL1), which achieves improvement over
RNR. Some robust sparse representation methods work well for pixel-level noise, but they seem to be very sensitive to the extreme illumination changes.

2) Face recognition with random block occlusions: In this subsection, we design two experiments to investigate the robustness of the proposed methods in dealing with different level of contiguous noise. In every experiment, we use the first image of each person from Subset 1 for training as the single training sample protocol.

In the first experiment, we use the similar experiment setting as in [1]. Subset 1 and 2 are used for training and Subset 3 for testing, but with different kinds of occlusions: cup, dollar, cartoon mask, book, flower and puzzle in testing images. We compare our methods with several related methods. The recognition results of each method are displayed in Fig. 5.

From Fig. 5, we can find that the proposed RMR model achieve the best result among all methods. This experiment demonstrates that RMR is more robust than others for FR with different contiguous occlusions even only with single training sample for every subject.

Fig. 6. Recognition rates (%) of each classifier under different occlusion level. (a) a square black block, (b) a square random block.

3) Face recognition with mixed noise: Finally, we conduct two more changing experiments to evaluate the performance of our methods for FR with mixed noise. Some sparse noise is added to Subset 5 of Extended Yale B and pixel corruption plus unrelated random block occlusion (level from 10% to 50%) to Subset 3 of Extended Yale B. The test images for one person are shown in Fig. 7. The training samples are the same as the corresponding experiments which have been performed in the above section. Table 1 shows the recognition rates obtained using different methods. We can see that our methods are dominant compared to all the other methods. For Subset 5, the advantages of our methods are more evident, S-RMRL1 and S-RMRL2 achieve the highest recognition rates (89.1%, 88.6%), which implies that our methods are very robust to illumination and suitable for dealing with mixed noise. For Subset 3 with pixel corruption plus block occlusion, the recognition rates drop down dramatically with the corruption level increases. However, our methods still achieve the relative robust performance. For example, S-RMRL1 gains 61.1% recognition rate, compared to NMR at least 33.5% improvement.
Table I

<table>
<thead>
<tr>
<th>Case</th>
<th>Subset 5</th>
<th>10%</th>
<th>20%</th>
<th>30%</th>
<th>40%</th>
<th>50%</th>
</tr>
</thead>
<tbody>
<tr>
<td>LRC</td>
<td>9.1</td>
<td>99.6</td>
<td>88.8</td>
<td>63.8</td>
<td>41.9</td>
<td>17.8</td>
</tr>
<tr>
<td>CRC</td>
<td>5.3</td>
<td>100</td>
<td>97.1</td>
<td>75.0</td>
<td>50.7</td>
<td>24.1</td>
</tr>
<tr>
<td>SRC</td>
<td>21.9</td>
<td>99.8</td>
<td>93.6</td>
<td>67.5</td>
<td>46.1</td>
<td>28.1</td>
</tr>
<tr>
<td>SSRC</td>
<td>23.6</td>
<td>98.8</td>
<td>94.2</td>
<td>68.5</td>
<td>52.1</td>
<td>30.3</td>
</tr>
<tr>
<td>RLRC</td>
<td>12.3</td>
<td>99.6</td>
<td>89.2</td>
<td>70.0</td>
<td>51.3</td>
<td>19.3</td>
</tr>
<tr>
<td>CESR</td>
<td>25.2</td>
<td>94.7</td>
<td>91.7</td>
<td>76.8</td>
<td>53.1</td>
<td>22.1</td>
</tr>
<tr>
<td>RSC</td>
<td>28.3</td>
<td>100</td>
<td>99.3</td>
<td>89.9</td>
<td>58.1</td>
<td>29.7</td>
</tr>
<tr>
<td>HQ_A</td>
<td>10.6</td>
<td>97.6</td>
<td>91.3</td>
<td>85.6</td>
<td>51.1</td>
<td>21.9</td>
</tr>
<tr>
<td>HQ_M</td>
<td>15.8</td>
<td>98.3</td>
<td>92.2</td>
<td>87.9</td>
<td>56.3</td>
<td>22.2</td>
</tr>
<tr>
<td>NMR</td>
<td>7.8</td>
<td>99.8</td>
<td>94.3</td>
<td>79.4</td>
<td>51.8</td>
<td>27.6</td>
</tr>
<tr>
<td>RNR</td>
<td>10.2</td>
<td>100</td>
<td>99.2</td>
<td>90.2</td>
<td>58.8</td>
<td>30.2</td>
</tr>
<tr>
<td>S-RMRL1</td>
<td>89.1</td>
<td>100</td>
<td>100</td>
<td>97.1</td>
<td>84.8</td>
<td>61.1</td>
</tr>
<tr>
<td>S-RMRL2</td>
<td>88.6</td>
<td>100</td>
<td>100</td>
<td>97.6</td>
<td>82.9</td>
<td>52.2</td>
</tr>
</tbody>
</table>

Table II

<table>
<thead>
<tr>
<th>Case</th>
<th>baboon occlusion</th>
<th>Sunglasses</th>
<th>Scarf</th>
</tr>
</thead>
<tbody>
<tr>
<td>LRC</td>
<td>52.3</td>
<td>92.8</td>
<td>30.7</td>
</tr>
<tr>
<td>CRC</td>
<td>68.6</td>
<td>93.5</td>
<td>63.6</td>
</tr>
<tr>
<td>SRC</td>
<td>70.0</td>
<td>94.4</td>
<td>57.6</td>
</tr>
<tr>
<td>SSRC</td>
<td>94.0</td>
<td>95.4</td>
<td>66.7</td>
</tr>
<tr>
<td>RLRC</td>
<td>95.1</td>
<td>94.6</td>
<td>53.3</td>
</tr>
<tr>
<td>CESR</td>
<td>88.9</td>
<td>95.0</td>
<td>33.5</td>
</tr>
<tr>
<td>RSC</td>
<td>94.8</td>
<td>94.2</td>
<td>66.8</td>
</tr>
<tr>
<td>HQ_A</td>
<td>94.2</td>
<td>94.7</td>
<td>48.7</td>
</tr>
<tr>
<td>HQ_M</td>
<td>94.6</td>
<td>95.0</td>
<td>50.1</td>
</tr>
<tr>
<td>NMR</td>
<td>95.1</td>
<td>96.9</td>
<td>73.5</td>
</tr>
<tr>
<td>RNR</td>
<td>95.2</td>
<td>97.2</td>
<td>76.8</td>
</tr>
<tr>
<td>RMRL1</td>
<td>97.3</td>
<td>97.5</td>
<td>79.1</td>
</tr>
<tr>
<td>RMRL2</td>
<td>96.9</td>
<td>97.1</td>
<td>78.1</td>
</tr>
</tbody>
</table>

B. Experiments on the AR Database

1) Face recognition with continuous occlusion: In this experiment, we evaluate the robustness of our methods in dealing with real disguise on the AR database. The images of 120 individuals are selected and used in our experiments. We manually crop the face portion of the image and then normalize it to 50 × 40 pixels, the images of one person are shown in Fig.8. We divide the images into four sets: (1) the first four images (with various facial expressions) from sessions 1 and 2, (2) the first 4 images of sessions 1 and 2 corrupted by a square block of “baboon” with 60% occlusion level; (3) 6 images with sunglasses from both sessions, and (4) 6 images with scarves from both sessions. In our experiments, we use the set (1) for training and the others for testing to verify the effectiveness of our method in dealing with real disguise and high level occlusion.

The results of FR are listed in Table II. As it is shown in Table II, compared with other algorithms, our methods achieve the highest recognition rate for each test, when the faces are disguised with sunglasses, our methods show some improvement, the recognition rate is 97.5% (RMRL1). When the faces are disguised with scarves, the advantage of RMR becomes evident, it achieves 79.1% (RMRL1) recognition rate which improves 5.6% compared with NMR. When the testing images were corrupted by a very high level artificial occlusion, RMRL2 can still achieve 96.9% recognition rate. We think the inherent difference between the natural and artificial occlusion that plays a dominant role in the recognition task, and the natural occlusion is more complicated and more difficult than the artificial one.

2) Face recognition with mixed noise: In order to further evaluate the robustness of S-RMR (including S-RMRL1, S-RMRL2) in dealing with mixed noise. We use the eight images of the first set for training, and add the random pixel corruption to real disguise images (as shown in Fig. 9) as testing samples, respectively. Table III gives the recognition rates of the methods. From Table III, we can find that our methods get the highest recognition rates in all classifiers. Our experimental results imply that CRC, and LRC are not suitable for characterizing the proposed mixed noise.

C. Experiment on the LFW Database

For the LFW Database, we use the image LFWa and simply crop the face image to remove the background, the images are resized to 64 × 64 pixels. In our experiment, we choose 6 face images from each subject and in total 100 subjects are chosen. We utilize 4 images for training, and 2 images for testing, respectively. The samples for a person are given in Fig. 10, and Table IV lists the results of all methods. Compared with other algorithms, It is evident that RMRL1 and RMRL2 achieve the highest recognition rates (52.3%, 51.5%), RNR achieves the second highest recognition rate (46.6%). This result demonstrates that our methods are relatively more robust...
to handle the unconstrained setting than the other method based regression analysis.

![Example images from LFW face database. (a) The training images. (b) The testing images.](image)

**D. Experiments on the PubFig Database**

For the PubFig Database, we select 20 images of 200 persons for our training and testing, and each image is resized to 80 × 70 pixels. 10 images for each person are used as testing images and the rest of the images are used as training. Thus, there are 2000 testing images and 2000 training images in total. The samples for a person are given in Fig.11, and Table IV lists the results of some approaches based regression analysis. Although, we can find that the performances of other methods obtain competitive recognition rates, it slightly lags behind our methods. This means that it is a real challenge for regression analysis in facing inaccurate alignment changes.

![The samples for one person from PubFig face database.](image)

**E. The choice of parameters**

We now discuss how to choose the value of the regularization parameters in our models. k-fold cross validation is a popular method for comparing different parameters, and we use it in our paper. Note that there is only one parameter in the RMR model, we pick the values for λ, say (0.01, 0.1, 1, 10, 50, 100). Then for each λ, the algorithm of RMRL1 and RMRL2 produces the entire solution, we choose the λ by the highest recognition rate.

For S-RMR model, there are two tuning parameters. As suggested in RPCA, we fix the parameter α = 1/√m2, and the method for choosing the other parameter β in S-RMRL1 or S-RMRL2 model are similar with RMR model.

**F. The choice of non-convex functions**

In this subsection, we discuss the choice of non-convex functions. As the discussion in section III, we mainly focus on the four non-convex functions (e.g. l_q, log-sum, atan, log-exp).

First, we study the weighted function how to influence the solution of weighted nuclear norm minimization problem. The theorem 2 gives the SVT operator with weighted vector, the function \( s_{w,\tau}(d_i) = \max(0, d_i - \tau w_i) \) is the core of SVT operator. With the different non-convex functions, the relationship of \( d_i \) and \( s_{w,\tau}(d_i) \) is illustrated in Fig. 12 (a). From the Fig. 12 (a), one can find that the \( s_{w,\tau}(d_i) \) can tend to be consistent only need to choose the appropriate parameters in the \( \rho(\cdot) \) function, so with the different weighted function we can get similar solution of \( E \) via Eq. (23).

Second, in order to further evaluate the influence of the different non-convex function on FR rates, we conduct four experiments with RMRL2 method. For the Extended Yale B database, we choose Subset 1 for training, Subsets 4 and 5 for testing, respectively. For the AR database, we chose the first set for training, and use the third set and fourth set (sunglasses and scarves) for testing, respectively. We show the best recognition rates in Fig.12 (b). From the Fig.12 (b) one can find that the recognition rates can nearly achieve the same value, so it is not important to choose non-convex functions.

![The SVT operators and the Recognition rates with different non-convex functions (a) \( s_{w,\tau}(d) \), (b) Recognition rate (%) on different database.](image)

**TABLE III**

**RECOGNITION RATES(%) OF EACH CLASSIFIER FOR FR ON THE AR DATABASE WITH RANDOM NOISE.**

<table>
<thead>
<tr>
<th>Case</th>
<th>LRC</th>
<th>CRC</th>
<th>SRC</th>
<th>SSRC</th>
<th>RLRC</th>
<th>CESR</th>
<th>RSC</th>
<th>HQ_A</th>
<th>HQ_M</th>
<th>NMR</th>
<th>RNR</th>
<th>S-RMRL1</th>
<th>S-RMRL2</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a)</td>
<td>40.6</td>
<td>31.7</td>
<td>89.5</td>
<td>93.3</td>
<td>70.6</td>
<td>92.7</td>
<td>68.5</td>
<td>48.1</td>
<td>70.2</td>
<td>79.6</td>
<td>81.2</td>
<td>97.3</td>
<td>96.1</td>
</tr>
<tr>
<td>(b)</td>
<td>6.9</td>
<td>12.9</td>
<td>45.7</td>
<td>46.2</td>
<td>32.1</td>
<td>32.5</td>
<td>47.0</td>
<td>19.6</td>
<td>36.0</td>
<td>42.2</td>
<td>54.3</td>
<td>61.1</td>
<td>60.9</td>
</tr>
</tbody>
</table>

**V. CONCLUSIONS**

In this work, we present the robust nuclear norm for measuring the error matrix and propose a novel robust matrix regression model including RMR and S-RMR. The main merit of RMR model is its robustness to occlusions, illumination variations in FR, the reason lies in the fact that the robust nuclear norm is suitable for characterizing the low rank structural information of image matrix. The main merit of S-RMR model is that it can characterize the sparse noise and image-wise noise, simultaneously. The proposed model can be solved by the ADMM. Extensive experimental results demonstrate that the proposed method is more robust than state-of-the-art regression based methods for FR when dealing with the variations of occlusion and illumination.

**ACKNOWLEDGMENT**

The authors would like to thank the editor and the anonymous reviewers for their critical and constructive comments and suggestions.
TABLE IV
RECOGNITION RATES (%) OF DIFFERENT CLASSIFIER METHOD ON THE LFW, PUBFIG AND CHUK FACE DATABASES.

<table>
<thead>
<tr>
<th>Database</th>
<th>LRC</th>
<th>CRC</th>
<th>SRC</th>
<th>SRCR</th>
<th>RLRC</th>
<th>CESR</th>
<th>RSC</th>
<th>HQ_A</th>
<th>HQ_M</th>
<th>NMR</th>
<th>NRR</th>
<th>RMR_L1</th>
<th>RMR_L2</th>
</tr>
</thead>
<tbody>
<tr>
<td>LFW</td>
<td>27.5</td>
<td>43.0</td>
<td>40.0</td>
<td>43.5</td>
<td>30.6</td>
<td>42.5</td>
<td>31.0</td>
<td>43.6</td>
<td>45.1</td>
<td>44.5</td>
<td>46.6</td>
<td>52.3</td>
<td>51.5</td>
</tr>
<tr>
<td>PubFig</td>
<td>27.4</td>
<td>43.1</td>
<td>44.5</td>
<td>45.8</td>
<td>33.2</td>
<td>36.9</td>
<td>47.2</td>
<td>46.1</td>
<td>46.5</td>
<td>46.3</td>
<td>46.1</td>
<td>50.2</td>
<td>49.3</td>
</tr>
</tbody>
</table>

REFERENCES


**Jianchun Xie** received the M.S. degree from Southeast University, Nanjing, China in 2006. He is currently pursuing the Ph.D. degree in pattern recognition and intelligence system from school of computer science and engineering, Nanjing University of Science and Technology (NUST), Nanjing, China. Now, he is an assistant professor in the school of science of NUST. His current research interests include pattern recognition and optimization algorithm.

**Jian Yang** received the BS degree in mathematics from the Xuzhou Normal University in 1995, the MS degree in applied mathematics from the Changsha Railway University in 1998, and the PhD degree from the Nanjing University of Science and Technology (NUST), Nanjing, China, on the subject of pattern recognition and intelligence systems in 2002. In 2003 to 2007, he was a postdoctoral fellow at the University of Zaragoza, Hong Kong Polytechnic University and New Jersey Institute of Technology, respectively. He is currently a professor in the School of Computer Science and Technology of NUST. He is the author of more than 100 scientific papers in pattern recognition and computer vision. His journal papers have been cited more than 4,000 times in the ISI Web of Science, and 8,000 times in the Web of Scholar Google. His research interests include pattern recognition, computer vision, and machine learning. He is currently an associate editor of Pattern Recognition Letters and IEEE Transactions Neural Networks and Learning Systems, respectively.

**Jianjun Qian** received the BS and MS degrees in 2007 and 2010, respectively, and the PhD degree in pattern recognition and intelligence systems from the Nanjing University of Science and Technology (NUST), in 2014. Now, he is an assistant professor in the school of computer science and engineering of NUST. His research interests include pattern recognition, computer vision, and face recognition in particular.

**Ying Tai** received the B.S. degree in the school of computer science and Engineering from Nanjing University of Science and Technology (NUST), Nanjing, China, in 2012. Currently, he is pursuing the Ph.D. degree in NUST. His current research interests include pattern recognition, computer vision, and especially face recognition.

**Hengmin Zhang** received the bachelor degree in mathematics from the Heze University in 2010, and the master degree in the college of science from China University of Mining and Technology University (CUMT), Xuzhou, China in 2013. He is currently pursuing the Ph.D. degree in pattern recognition and intelligence system from School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, China. His current research interests include computer vision, machine learning and nonconvex optimization.