Pattern Recognition 91 (2019) 147–161

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/patcog

Nesting-structured nuclear norm minimization for spatially correlated matrix variate

Lei Luo[a,c], Jian Yang[a,*], Yigong Zhang[a], Yong Xu[b], Heng Huang[c]

[a] *School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, PR China*
[b] *Bio-Computing Research Center, Shenzhen Graduate School, Harbin Institute of Technology, Shenzhen 518055, PR China*
[c] *Department of Electrical and Computer Engineering, University of Pittsburgh, Pittsburgh, PA, USA*

ARTICLE INFO

Article history:
Received 2 November 2016
Revised 30 January 2019
Accepted 18 February 2019
Available online 21 February 2019

Keywords:
Nesting-structured nuclear norm
Low rank
Structured sparsity
Matrix regression
Matrix completion
Sub-gradient method

ABSTRACT

Integrating the structure prior in modeling has achieved considerable attention in pattern recognition and computer vision. Most current state-of-the-art methods (such as low rank representation and structured sparsity) search for a structured metric to fit the structure of the estimated variate, which either bear high time complexity (e.g., compute singular value decomposition for large-scale matrices), or cannot effectively exploit structure information of a matrix variate. In this work, we introduce a nesting-structured nuclear norm to characterize the matrix variate with structure prior and provide a unified framework for solving nesting-structured nuclear norm minimization (NSNM) problem by resorting to an improved sub-gradient method. This not only takes local and global structures of the matrix variate into joint consideration, but also enjoys the lower time complexity than traditional nuclear norm minimization. The revealed statistical meaning explains the rationality of the proposed method. Moreover, we apply NSNM to matrix regression and completion problems, respectively. The extensive experiments for face recognition and large-scale matrix completion clearly demonstrate the superiority of NSNM over some existing methods.

1. Introduction

In the past few years, incorporating structured priors in statistical model has become a popular technique for coping with various estimation tasks since it can provide a natural characterization over the relationships between data. As the successful applications of this view, low rank minimization has shown great potential in numerous fields such as machine learning, signal processing and so on.

It is known that using low rank function to constrain a matrix variate can capture its global structure, which leads to the increasing interest in matrix completion and subspace segmentation problem since the intrinsic dimensionality of high-dimensional data is in fact much smaller, i.e., they often lie in low dimensional structures. The most representative models around this view are Low Rank Representation (LRR) [1] and Robust Principle Component Analysis (RPCA) [2]. Their common characteristic is regarding samples to be correlated, but the stretching of each sample inevitably destroys its spatial structure. In fact, there also exist correlations between pixels of a single sample in some applications. Taking advantage of this fact, nuclear norm based matrix regression (NMR) [3] and Nuclear Norm-Based 2-DPCA (N-2-DPCA) [4] integrated the relationships of pixels in the error images caused by the spatially contiguous variations into modeling for face recognition, which is carried out by virtue of low rank assumption of the error matrix. Chen et al. [5] and Luo et al. [6] considered the similar problem from the viewpoint of the dependent matrix distribution. Subsequently, Luo et al. [7] and Luo et al. [8] further extended the above models resorting to Schatten p-norm and tree structure, respectively. In addition, Xu et al. [9] exploited the low-rank structure of the multi-label predictor in multi-label learning. Yan et al. [10] used Maximum-Margin Matrix Factorization to accomplish collaborative prediction of rating data by emphasizing the low-rank structure of the desired data. For more related studies on rank function minimization, please see [11–14].

The promising results in the above work demonstrate that low rank can effectively characterize the global structure of a matrix variate, but the local structure for a matrix variate also exists and is very crucial for the face recognition problem. For instance, the face images of 120 individuals from AR database are resized to 45 × 30. For each individual, a clean face image and a face image with sunglasses are chosen (as shown on the left side of Fig. 1).

[*] Corresponding author.
E-mail address: csjyang@njust.edu.cn (J. Yang).

https://doi.org/10.1016/j.patcog.2019.02.011
0031-3203/© 2019 Elsevier Ltd. All rights reserved.

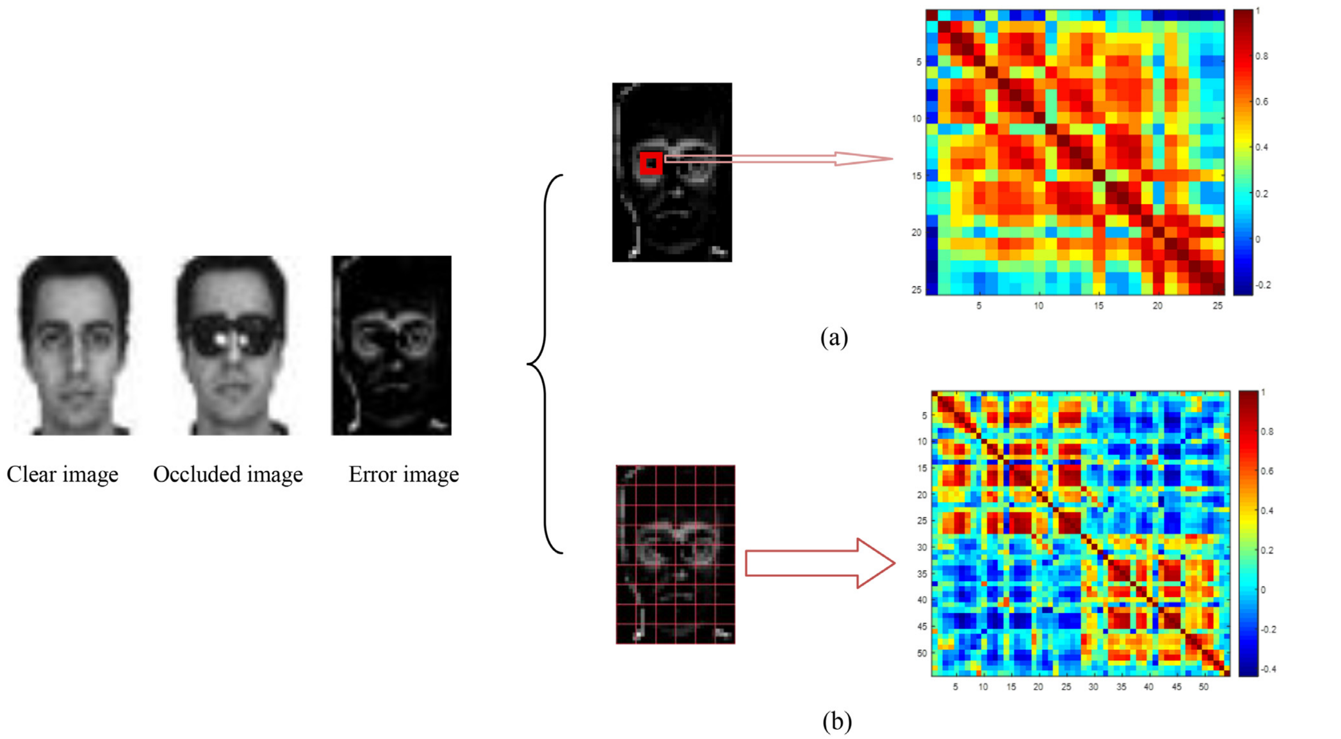


Fig. 1. (a) The illustration for showing the local correlation of the pixel-errors, where the right image is the correlation map of pixels in the red box area of the left error image; (b) The illustration for showing correlation of blocks, where the right image is the correlation map of nuclear norm of blocks (i.e., pixels in matrix B) in left error image.

Thus, 120 noise images caused by sunglasses are generated. Each noise image is partitioned into 9×6 blocks and the size of each block is 5×5. To describe the local correlation structure, we focus on the local block of each error image. Firstly, a certain block in the error image is chosen and marked as the red box. The correlation map of pixels in this block is presented in Fig. 1(a). It is clear that most pixels in red box are highly correlated. And this trend will become more evident with the narrowing of red box area. Next, we further investigate the relationships between blocks. Note that the structured attribute of nuclear norm, we attempt to use nuclear norm to measure each block, i.e., nuclear norm of each block is calculated to represent the corresponding block. Using the nuclear norm of each local block as an element and keeping their relative position, we can form a new matrix **B**, which to some extent reflects the global structure of the original error image. As a result, 120 random matrices of dimensions 9×6 are acquired. The correlation map of elements in **B** is shown in Fig. 1(b). It is found from Fig. 1(b) that these blocks are also correlated. By the above analysis, there is no doubt that *both local structure and structure among local blocks in an error image exist and the local structure factually plays a dominant role*. Accordingly, how to merge these useful structure information into a unified framework to realize the performance promotion turns into a valuable and challenging issue.

It is worth noting that some methods mentioned above, including LRR, RPCA and NMR, replace rank function with its tightest convex surrogate over the unit spectral norm ball, namely nuclear norm, to facilitate the design of algorithm. They suffer from the high computational cost to compute the singular value decomposition (SVD) in each iteration, especially for the large-scale matrix. Therefore, reducing the time complexity in SVD is extremely helpful for speeding up the algorithm. Toward this end, Lu et al. [15] presented a fast SVD method for multilevel block Hankel matrices. They used Lanczos process to reduce the MBH matrix into a bidiagonal or tridiagonal matrix and the SVD is implemented on the reduced matrix using the twisted factorization method. Majumdar et al. [16] decreased the complexity by computing a Cholesky decomposition instead of SVD. Cai et al. [17] computed the singular value thresholding (SVT) for a given matrix without SVD, which is carried out by two steps, namely, the polar decomposition step and the projection step done by Newton's method. Oh et al. [18] proposed a fast approximate SVT method by exploiting the property of iterative NNM procedures, which avoids the direct computation of SVD. The interested reader is referred to [19–21] for more techniques about the approximated SVD. These ways to some extend decrease the computation complexity in the traditional nuclear norm minimization. Nonetheless, they either focus on the specific cases, or are only the approximation of SVD, which may be far away from the essential attributes of nuclear norm characterizing the structure, leading to an impractical result.

On the basis of the above analysis, this paper will establish a nesting structure for a matrix and use it to induce a nesting-structured nuclear norm minimization model. This not only takes local and global structures of a matrix variate into joint consideration, but also affords the lower time complexity than traditional nuclear norm minimization. Specifically speaking, we partition an original matrix $\mathbf{A}^1$ into several blocks. Differing from the structured sparsity inducing norm [22–25], which stretches each block into a vector and uses L_2 or L_∞ norm to constrain it, here we keep the original matrix form of each block and use nuclear norm to directly characterize it. Due to the structure attribute of nuclear norm, this can exploit effectively the structure of each block which can be viewed as the local structure of the matrix $\mathbf{A}^1$. As the experiment in Fig. 1(b), we can obtain an external matrix $\mathbf{A}^2$, each element of which coincides with the nuclear norm of the corresponding block in $\mathbf{A}^1$. Such a process is repeated continually. We eventually obtain a matrix $\mathbf{A}^k$ which only includes an element. Obviously, each matrix $\mathbf{A}^l$, where $l = 1, \ldots, k$, possesses the smaller size than the original matrix $\mathbf{A}^l$. Then, the nuclear norm is acted

$$\mathbf{A}^1=\begin{pmatrix} a_{11}^1 & a_{12}^2 & \cdots & a_{1n}^1 \\ a_{21}^1 & a_{22}^1 & \cdots & a_{2n}^1 \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1}^1 & a_{m2}^1 & \cdots & a_{mn}^1 \end{pmatrix} \overset{\text{partition}}{\Rightarrow} \begin{pmatrix} \mathbf{A}_{11}^1 & \mathbf{A}_{12}^1 & \cdots & \mathbf{A}_{1q}^1 \\ \mathbf{A}_{21}^1 & \mathbf{A}_{22}^1 & \cdots & \mathbf{A}_{2q}^1 \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{A}_{p1}^1 & \mathbf{A}_{p2}^1 & \cdots & \mathbf{A}_{pq}^1 \end{pmatrix} \underset{w_{ij}^1\|\mathbf{A}_{ij}^1\|_{in}=a_{ij}^2}{\overset{\text{construct } \mathbf{A}^2}{\Rightarrow}} \mathbf{A}^2=\begin{pmatrix} a_{11}^2 & a_{12}^2 & \cdots & a_{1q}^2 \\ a_{21}^2 & a_{22}^2 & \cdots & a_{2q}^2 \\ \vdots & \vdots & \ddots & \vdots \\ a_{p1}^2 & a_{p2}^2 & \cdots & a_{pq}^2 \end{pmatrix} \overset{\text{partition}}{\Rightarrow} \cdots \underset{w_{11}^{k-1}\|\mathbf{A}_{11}^{k-1}\|_{in}=a_{11}^k}{\overset{\text{construct } \mathbf{A}^k}{\Rightarrow}} \mathbf{A}^k=\left(a_{11}^k\right)$$

Fig. 2. An example for explaining nesting structure.

on each external matrix $\mathbf{A}^l$ to further exploit the structure of the original matrix $\mathbf{A}^1$, which produces the *nesting-structured nuclear norm*. Using nuclear norm to characterize these external matrices factually captures the global structure of the original matrix variate since it incorporates all local structures into modeling. We establish a unified framework for nesting-structured nuclear norm minimization by developing an improved Sub-gradient method to solve the proposed model. This adds a popular accelerated scheme and a forcing descent technique into the traditional sub-gradient to stabilize and speed up the iterative process. Unlike the traditional nuclear norm minimization models, which compute the SVD on the original matrix, the proposed approach only needs to implement SVD on some generated small-scale matrices. Under a matrix variate distribution, the statistical meaning of the proposed framework is provided by seeking its maximum a posteriori probability estimation solution. Additionally, the essence of the proposed distribution for characterizing some structural variates is also revealed. The proposed method is applied on matrix regression and completion, respectively. A series of experiments on face recognition and large-scale matrix completion show the advantages of our method over some existing methods.

Notations. Throughout this paper, the bold capital and bold lowercase symbols are used to represent matrices and vectors, respectively. $tr(\mathbf{X})$ denotes trace of a matrix $\mathbf{X}$ and $\exp(\cdot)$ represents the exponential function. If a square matrix $\mathbf{X}$ is positive semi-definite, we denote it by $\mathbf{X}\geq\mathbf{0}$. $\|\mathbf{X}\|_1$ denotes L_1-norm of the matrix $\mathbf{X}$. $\|\mathbf{X}\|_F$ denotes Frobenius norm of the matrix $\mathbf{X}$, which is equal to the L_2-norm of $\mathrm{Vec}(\mathbf{X})$ (i.e., $\|\mathrm{Vec}(\mathbf{X})\|_2$), where $\mathrm{Vec}(\cdot)$ is an operator converting a matrix into a vector. $\frac{\partial \mathrm{h}(\mathbf{B})}{\partial x_i}$ denotes the partial derivative of matrix function $\mathrm{h}(\mathbf{B})$ associated with x_i.

The remainder of this paper is organized as follows: in the next section we first give the definition of nesting-structured nuclear norm and present a nesting-structured nuclear norm minimization framework, which is solved by an improved sub-gradient method. Then, the statistical meaning and rationality of the proposed method are investigated in Section 3. The proposed framework is applied to matrix regression and matrix completion in Section 4. The convergence and complexity analysis for the proposed algorithm are presented in Section 5, and some experimental results are reported in Section 6. Section 7 contains the conclusions.

2. Nesting-structured nuclear norm minimization

In this section, we first introduce a nesting matrix structure and make use of it to induce a nesting-structured nuclear norm. Then, a nesting-structured matrix minimization model is presented and an improved sub-gradient is developed to solve it.

2.1. Nesting-structured nuclear norm

Partition an original matrix $\mathbf{A}=\mathbf{A}^1\in R^{n\times m}$ into $p\times q$ sub-matrices as shown in Fig. 2. Here each sub-matrix $\mathbf{A}_{ij}^1\in R^{p_i\times q_j}$, where $i=1,\dots,p$, $j=1,\dots,q$, $\sum_{i=1}^p p_i=n$, and $\sum_{i=1}^q q_i=m$. Given a matrix norm $\|\cdot\|_{in}$ and calculate $a_{ij}^2=w_{ij}^1\|\mathbf{A}_{ij}^1\|_{in}$, where $w_{ij}^1>0$ is the weight of $\|\mathbf{A}_{ij}^1\|_{in}$. Thereby we can obtain a new matrix $\mathbf{A}^2$. Such a process is repeated sequentially. Ultimately, the matrix $\mathbf{A}^k$ is generated. We call the procedure from $\mathbf{A}^1$ to $\mathbf{A}^k$ as *Nesting Structure* with depth $\boldsymbol{k}$. As the illustration in Fig. 1(a), we use nuclear norm to constrain each $\mathbf{A}_{ij}^l$ (i.e., the sum of singular values of $\mathbf{A}_{ij}^l$, which is denoted by $\|\mathbf{A}_{ij}^l\|_*$), where $l=1,\dots,k$, to exploit fully the structure of the block $\mathbf{A}_{ij}^l$, that is to say, $\|\cdot\|_{in}=\|\cdot\|_*$. Meanwhile, according to the observation in Fig. 1(b), we employ nuclear norm to act on each $\mathbf{A}^l$ to exploit the global structure of $\mathbf{A}$. Denote $\Omega(\mathbf{A})=\sum_{l=1}^k \sigma_l\|\mathbf{A}^l\|_*=\|\mathbf{A}\|_{*,w_{ij},*,\sigma_l}$, where $\sigma_l>0$, then $\Omega(\mathbf{A})$ is called as *nesting-structured nuclear norm*[1] with regard to $\mathbf{A}$.

Table 1
Various types of norms induced by nesting structure.

σ_l	$p,\ q$	$\|\cdot\|_{ex}$	$\|\cdot\|_{in}$	w_{ij}	$\Omega(\mathbf{A})$
$\sigma_1=1$, $\sigma_l=0(l\neq 1)$	/	$\|\cdot\|_*$ $\|\cdot\|_1$ $\|\cdot\|_F$	/	/	$\|\mathbf{A}\|_*$ $\|\mathbf{A}\|_1$ $\|\mathbf{A}\|_F$
$\sigma_2\neq 0$, $\sigma_l=0(l\neq 2)$	$p\geq 1$,	$\|\cdot\|_1$	$\|\cdot\|_2$	$w_{ij}>0$	group sparsity norm
$\sigma_l\geq 0$	$q\geq 1$	$\|\cdot\|_*$	$\|\cdot\|_*$		$\|\mathbf{A}\|_{*,w_{ij},*,\sigma_l}$

It is easy to see that the *nesting-structured nuclear norm not only captures the pixel-level structure of a matrix variate, but also considers the block-level structure*, while other norms such as nuclear norm or structured sparsity inducing norm cannot do these. *Throughout this paper, we only consider the second layer in this paper for the convenience*, i.e., $\sigma_2\neq 0$ but $\sigma_l=0(l\neq 2)$.

Remark 1. In the framework of nesting structure, some other choices can be acted on $\|\cdot\|_{ex}$ and $\|\cdot\|_{in}$, which induces various types of norms as summarized in Table 1.

2.2. Nesting-structured nuclear norm minimization

In this subsection, we propose a nesting-structured nuclear norm minimization (NSNM for short) model as follows:

$$\min_{\mathbf{X}} \|\mathrm{f}(\mathbf{X})\|_{*,w_{ij},*,\sigma_l}+\rho\mathrm{g}(\mathbf{X}). \tag{1}$$

where $\mathrm{f}(\mathbf{X}):R^{n\times m}\to R^{n\times m}$ is an affine function with regard to $\mathbf{X}$, $\mathrm{g}(\mathbf{X}):R^{n\times m}\to R$ is a general matrix function (note that $\mathrm{g}(\mathbf{X})$ may be non-smooth), and the parameter $\rho>0$ is a tradeoff between the two items (i.e., $\|\mathrm{f}(\mathbf{X})\|_{*,w_{ij},*,\sigma_l}$ and $\mathrm{g}(\mathbf{X})$). Meanwhile, it is also assumed that $\mathrm{f}(\mathbf{X})$ or $\mathrm{g}(\mathbf{X})$ includes the observation $\mathbf{D}$. Since nuclear norm is a special case of nesting-structured nuclear norm, model (1) factually generalizes the unconstrained versions of the general nuclear norm minimization problems [1–14]. Owing to the non-smoothness of $\mathrm{g}(\mathbf{X})$, some gradient based methods (e.g., SDM [26], APG [27]) cannot be used to solve (1) (since the convergence cannot be guaranteed). As for ADMM, we need to convert (1) into the following constrained version:

$$\min_{\mathbf{X},\mathbf{E}} \|\mathbf{E}\|_{*,w_{ij},*,\sigma_l}+\rho\mathrm{g}(\mathbf{X}),\ \mathrm{s.t.}\mathbf{E}=\mathrm{f}(\mathbf{X}). \tag{2}$$

[1] When k=1, $\Omega(\mathbf{A})$ strictly defines a norm that satisfies the three norm conditions, while it defines a quasi-norm when k>1 (the detailed proof is seen in supplemental materials). Because the mathematical formulations and derivations in this paper equally apply to both norm and quasi-norm, we do not differentiate these two concepts for notation brevity.

However, it is difficult to obtain a closed-form solution for the key sub-problem with regard to nesting-structured nuclear norm. Compared with the above two methods, sub-gradient based methods are easier to implement due to its relaxed conditions.

2.3. The proposed algorithm

Sub-gradient method [28] was originally developed by Shor in the 1970s, the core iteration of which can be considered as a generalization of gradient method for non-differentiable function. There have been many studies focusing on this method. Specifically, Boyd et al. [29] analyzed the convergence of sub-gradient method for different types of step size rules. Nedic and Ozdaglar [30] investigated sub-gradient method for computing the saddle points of a convex-concave function. Nesterov and Shikhman [31] developed a quasi-monotone sub-gradient method, which guarantees the best possible rate of convergence for the whole sequence of test points.

As we know, the concept of proximal operator plays the central role in proximal gradient methods. Analogously, Rockafellar et al. introduced proximal map for a non-convex (even non-smooth) function in [32]. Along this line, Bolte et al. [33] established a proximal alternating linearized minimization (PALM) algorithm for non-convex and non-smooth minimization problems and derived a new and simple globally convergent algorithm for solving the sparse nonnegative matrix factorization problem. Afterward, Cruz et al. [34] presented a variant of the proximal forward-backward splitting iteration for solving the non-smooth problem.

The main advantage of sub-gradient based methods is the wider application than some existing methods. In this paper, we will apply the above proposed methods to solving the model (1). And yet, sub-gradient based methods give the slow convergence, that is, it only achieves an optimal convergence rate: $O(1/t^{1/2})$ under the certain conditions. In addition, it is not a descent method. Thus, it is desired to integrate some accelerated schemes or descent strategy in design of algorithms to speed up the convergence and stabilize the iterative trend. To this end, a stochastic sub-gradient mirror-descent method with weighted iterate-averaging [35] is investigated and its per-iterate convergence rate is also analyzed. Surprisingly, by suitably choosing the step size values, one can obtain the rate of the order $1/t$ for strongly convex functions. Coincidentally, Neumaier [36] proposed a fast sub-gradient algorithm with optimal complexity both for the general non-smooth case and for the strongly convex case.

Zhang et al. [37] recently presented Rapidly Accelerated Proximal Gradient (RAPID) method for convex minimization. They introduced a simple line search step after each proximal gradient step in Accelerated Proximal Gradient (APG). A series of experiments showed the advantages of RAPID over APG. But for nuclear norm minimization, it still faces two SVD for the large-scale matrix due to the computation of auxiliary parameter. As mentioned before, RAPID or APG cannot be applied on problem (1) since both $\|\mathrm{f}(\mathbf{X})\|_{*,w_{ij},*,\sigma_l}$ and $\mathrm{g}(\mathbf{X})$ are non-differentiable. And yet, we are able to merge their accelerated schemes into the original sub-gradient method, which induces an improved gradient method. Denoting $\mathrm{J}(\mathbf{X}) = \|\mathrm{f}(\mathbf{X})\|_{*,w_{ij},*,\sigma_l} + \rho \mathrm{g}(\mathbf{X})$, the detailed iteration procedure of the improved gradient method is summarized in Algorithm 1.

3. The statistical meaning of the proposed model

In this section, we first analyze the statistical meaning of model (1) from the viewpoint of maximum likelihood estimation (MLE) using a matrix distribution, then account for the essence of the proposed distribution for characterizing the structural matrix variate.

Algorithm 1 NSNM by the improved sub-gradient method.

```
Input: data matrix $\mathbf{D}$ (or other known matrices), and parameter $\rho$.
Initialize: $\mathbf{V}^0 = \mathbf{0}, \mathbf{X}^0 = \mathbf{0}, \mathbf{Y}^0 = \mathbf{0}, \theta^0 = 1$
While not convergence do
Step 1. $\mathbf{Y}^{t+1} = \mathbf{V}^t - \mu^t \mathbf{W}^t$, where $\mathbf{W}^t \in \partial \|\mathbf{J}(\mathbf{V}^t)\|_{*,w_{ij},*,\sigma_l}$,
Step 2. $\mathbf{X}^{t+1} = \{ \begin{matrix} \mathbf{Y}^{t+1}, \mathrm{J}(\mathbf{Y}^{t+1}) \le \mathrm{J}(\mathbf{X}^t), \\ \mathbf{X}^t, \text{otherwise}, \end{matrix}$
Step 3. $\theta^{t+1} = \frac{1+\sqrt{1+4(\theta^t)^2}}{2}$,
Step 4. $\mathbf{V}^{t+1} = \mathbf{X}^{t+1} + (\frac{\theta^t - 1}{\theta^{t+1}})(\mathbf{X}^{t+1} - \mathbf{X}^t)$.
End while
Output: Optimal regression coefficient vector $\mathbf{X}^{t+1}$.
```

3.1. The derivation of model (1) by MLE

For the convenience of investigation, let $\mathrm{A}^1 = \mathrm{f}(\mathbf{X})$ in model (1). According to the definition of nesting-structured nuclear norm in Section 2.1, we can obtain a new matrix $\mathbf{A}^2$ denoted by $\mathbf{B}$ as shown Fig. 2. In addition, $\mathrm{f}(\mathbf{X})$ is assumed to include the observation $\mathbf{D}$ (e. g, $\mathrm{f}(\mathbf{X}) = \mathbf{D} - \mathrm{r}(\mathbf{X})$, where $\mathrm{r}(\mathbf{X})$: $R^{l\times m} \to R^{l\times m}$ is a matrix mapping). According to the results in [6], the matrix variate $\mathbf{B}$ can be assumed to has a distribution of

$$P(\mathbf{B}|\mathbf{M}, \mathbf{\Sigma}, \mathbf{\Delta}, \mathbf{C}) = C \exp\left(-\tfrac{1}{2} tr\left((\mathbf{B}-\mathbf{M})^T \mathbf{\Sigma}(\mathbf{B}-\mathbf{M})\mathbf{\Delta}\right)^{1/2}\right), \quad (3)$$

where C the positive proportionality constant. Notice that each element of $\mathbf{B}$ corresponds to each block of $\mathbf{A}^1$ and $\mathrm{A}^1 = \mathrm{f}(\mathbf{X})$, then $\mathbf{B}$ can be regarded as the matrix function associated with $\mathbf{X}$. Then, (3) can be rewritten as

$$P(\mathbf{D}|\mathbf{X}, \mathbf{M}, \mathbf{\Sigma}, \mathbf{\Delta}, \mathbf{C}) = C \exp\left(-\tfrac{1}{2} tr\left((\mathbf{B}-\mathbf{M})^T \mathbf{\Sigma}(\mathbf{B}-\mathbf{M})\mathbf{\Delta}\right)^{1/2}\right), \quad (4)$$

where $\mathbf{M} \in R^{n\times m}$, $\mathbf{\Sigma} \in R^{n\times n}$, $\mathbf{\Delta} \in R^{m\times m}$ and $\mathbf{\Sigma}, \mathbf{\Delta} \ge \mathbf{0}$.

Meanwhile, the prior for the variate $\mathbf{X}$ is assumed to be:

$$P(\mathbf{X}|\nu) = c \exp\left(-\mathrm{g}(\mathbf{X})/\nu\right). \quad (5)$$

where $\nu > 0$ denotes the hyper-parameters associated with the prior of $\mathbf{X}$ and c is the positive proportionality constant. Thus, the posterior distribution for $\mathbf{X}$ can be written as

$$P(\mathbf{X}|\mathbf{M}, \mathbf{\Sigma}, \mathbf{\Delta}, C, \nu) \propto P(\mathbf{D}|\mathbf{X}, \mathbf{M}, \mathbf{\Sigma}, \mathbf{\Delta}, C) P(\mathbf{X}|\nu). \quad (6)$$

Taking the negative logarithm of Eq. (6) and omitting some constant terms, we can achieve the maximum posterior estimation of $\mathbf{X}$ by

$$\mathbf{X} = \arg\min_{\mathbf{X}} tr\left((\mathbf{B}-\mathbf{M})^T \mathbf{\Sigma}(\mathbf{B}-\mathbf{M})\mathbf{\Delta}\right)^{1/2} + \rho \mathrm{g}(\mathbf{X}). \quad (7)$$

where $\rho = 2\nu$.

Here, for the convenience, $\mathbf{\Sigma}$ and Δ are set as identity matrices, and $\mathbf{M}$ is chosen as a zero matrix, then (7) will be simplified as

$$\mathbf{X} = \arg\min_{\mathbf{X}} tr\left(\mathbf{B}^T\mathbf{B}\right)^{1/2} + \rho \mathrm{g}(\mathbf{X}). \quad (8)$$

Considering $tr(\mathbf{B}^T\mathbf{B})^{1/2} = \|\mathbf{B}\|_*$ and $\mathbf{B} = \mathbf{A}^2$ then (8) ultimately becomes (1).

3.2. Why use the distribution (3)?

In the following, the rationality of the proposed distribution (3) will be analyzed. Considering that $\mathbf{B}$ is the matrix function associated with $\mathbf{X}$ and connecting (3), the distribution of $\mathbf{X}$ can be concisely express as

$$P(\mathbf{X}) = C \exp\left(-\tfrac{1}{2} tr\left(\mathbf{B}^T\mathbf{B}\right)^{1/2}\right). \quad (9)$$

Let $\mathbf{L} = (\mathbf{B}^T\mathbf{B})^{-1/2}$, then (9) can be equivalently written as

$$P(\mathbf{X}) = C \exp\left(-\tfrac{1}{2} tr\left(\left(\mathbf{B}\mathbf{L}^{1/2}\right)^T \left(\mathbf{B}\mathbf{L}^{1/2}\right)\right)\right), \quad (10)$$

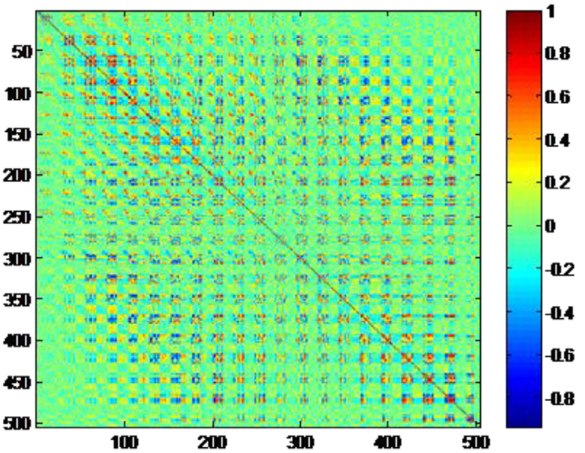


Fig. 3. Correlation map of elements in $\mathbf{BL}^{1/2}$ which corresponds to Fig. 1(b).

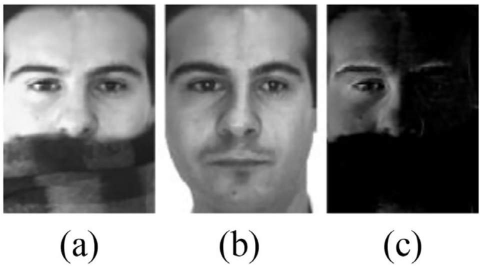


Fig. 4. (a) Original image, (b) recovered image, (c) noise image $\mathbf{A}^1$.

Denote $\mathbf{Z}=\mathbf{BL}^{1/2}$, then we have

$$P(\mathbf{Z})=C\exp\left(-\tfrac{1}{2}tr\left(\mathbf{Z}^T\mathbf{Z}\right)\right), \tag{11}$$

Thus, using the distribution (9) to characterize matrix variate **X** is equivalent to assuming that the induced matrix variate $\mathbf{Z}=\mathbf{BL}^{1/2}$ follows independent Gaussian distribution (11). In the following, it is verified that matrix variate **Z** actually follows independent Gaussian distribution.

Firstly, it is shown that $\mathbf{L}^{1/2}$ can alleviate the correlation between pixels of random matrix variate **B**. Based on the experimental setting in Fig. 1(b), 100 new matrices with dimensions of 24 × 21 (each image is denoted by **B**) are generated. Now multiplying **B** by $\mathbf{L}^{1/2}$, one obtains 100 random matrices $\mathbf{Z}=\mathbf{BL}^{1/2}$. The correlation map of elements in **Z** is shown in Fig. 3, from which we can see that *the correlations of pixels in Z are weaker than B and the elements in Z become independent approximately.*

Next, we show that $\mathbf{Z}=\mathbf{BL}^{1/2}$ approximately follows matrix variate Gaussian distribution. Fig. 4 shows an original image (size is 450 × 300) with scarf and slight illumination, and one can decompose (a) into the recovered term (b) and noise term (c) which is denoted by $\mathbf{A}^1$. Using the similar strategy in the previous experiment, $\mathbf{A}^1$ is partition into 30 × 20 blocks, where each block owns a size of 15 × 15. Thereby a new matrix **B** is obtained, each element of which consists with the nuclear norm of the corresponding block. Fig. 5(a) delineates **B** fitted by different distributions. One can see that Gaussian and Laplacian distributions are far away from empirical distribution. Fig. 5(b) shows the fitted distributions with regard to $\mathbf{BL}^{1/2}$ by different models. *Compared with* B*, the empirical distribution of* $\mathbf{BL}^{1/2}$ *more approaches Gaussian distribution.*

The above analysis reveals that the induced matrix variate $\mathbf{Z}=\mathbf{BL}^{1/2}$ approximately follows matrix variate Gaussian distribution and pixels in **Z** can be considered to be approximately independent. Thus, the effect of matrix $\mathbf{L}^{1/2}$ is indeed to alleviate the correlations between pixels in matrix **B** and make the matrix variate **B** approximately Gaussian. It leads to that the distribution (11) is reasonable for depicting the matrix variate **Z** since it provides an optimal characterization for Gaussian data. Back to the distribution (9), it is evident that using it to constrain **B** is closer to the real distribution of **B** than other distributions. *This explains the rationality of model (1) from the statistical viewpoint.* It is also noteworthy that the above analysis focuses on the matrix **B**, which represents the global structure of the original matrix $\mathbf{A}^1$. As for the local structures of $\mathbf{A}^1$, the similar results can be obtained.

4. Two applications of NSNM

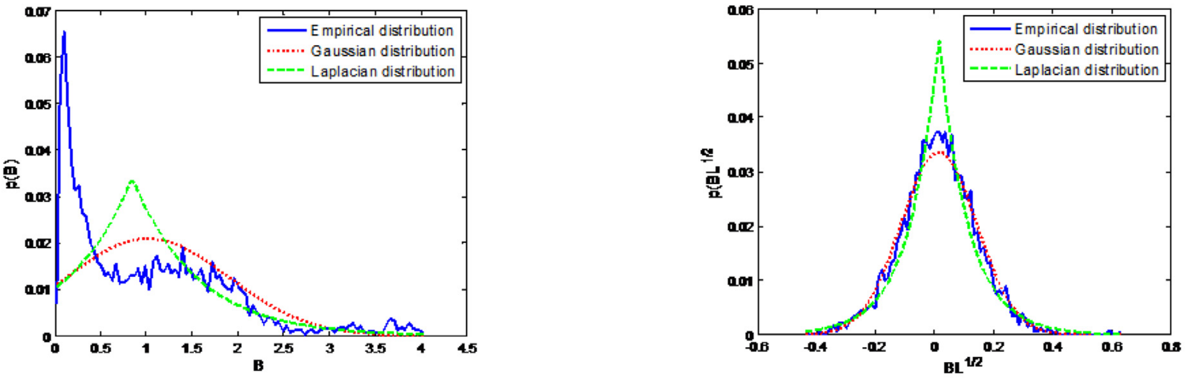
In this Section, we apply the proposed method to matrix regression and completion, respectively.

4.1. Nesting-structured nuclear norm based matrix regression

For face recognition, given a dictionary $\mathbf{M}=\{\mathbf{M}_1,\mathbf{M}_2,\ldots,\mathbf{M}_n\}$, where $\mathbf{M}_i\in R^{n\times m}$ $(i=1,\ldots,n)$ is a 2D image matrix. Then, we can use the dictionary **M** to represent a test image **D** $(\in R^{n\times m})$ linearly as follows:

$$\mathbf{D}=x_1\mathbf{M}_1+x_2\mathbf{M}_2+,\ldots,+x_n\mathbf{M}_n+\mathbf{E}, \tag{12}$$

where $\{x_1,x_2,\ldots,x_n\}$ is a set of representation coefficients, $x_1\mathbf{M}_1+x_2\mathbf{M}_2+,\ldots,+x_n\mathbf{M}_n$ is the reconstructed image and **E** is the representation residual. By defining the linear mapping



(a) The empirical distribution and the fitted distributions of **B**

(b) The empirical distribution and the fitted distributions of $\mathbf{BL}^{1/2}$

Fig. 5. The empirical distributions and the fitted distributions of the induced image **B** and $\mathbf{BL}^{1/2}$.

Table 2
The Schemes for Characterizing Residual Term in Some Existing Methods.

SRC	CRC	LRC	RLRC	RSC	CESR	SSRC [25]	NMR	NL$_1$R [6]
L$_1$ or L$_2$ norm	L$_2$ norm		Robust M-estimator			Structures sparsity induced norm	Nuclear norm	Nuclear-L$_1$ norm

from R^n to $R^{n\times m}$: $\mathrm{M}(\mathbf{x}) = x_1\mathbf{M}_1 + x_2\mathbf{M}_2 +, \dots, +x_n\mathbf{M}_n$, where $\mathbf{x} = [x_1, x_2, \dots, x_n]^T$. Then, the formula (12) can be expressed as

$$\mathbf{D} = \mathrm{M}(\mathbf{x}) + \mathbf{E}. \tag{13}$$

Eq. (13) can be viewed as the general linear regression model.

Let $\mathrm{f}(\mathbf{x}) = \mathbf{D} - \mathrm{M}(\mathbf{x})$ and $\mathrm{g}(\mathbf{x}) = \|\mathbf{x}\|_1$ in problem (1), then we can obtain a nesting-structured nuclear norm based matrix regression (Nesting-NMR) with L$_1$ regularization as follows:

$$\min_{\mathbf{x}} \|\mathbf{D} - \mathrm{M}(\mathbf{x})\|_{*,w_{ij},*,\sigma_l} + \rho\|\mathbf{x}\|_1. \tag{14}$$

Compared with some recent methods, such as CRC [38], SRC [39], LRC [40] and the scheme proposed in [41], there exist three significant advantages for model (14). Firstly, CRC, SRC and LRC need to stretch the error matrix into a vector in advance. It is unreasonable evidently for some structural noise caused by illumination, occlusion or real disguises due to the correlations between pixels. Our method directly considers the matrix form of error term: $\mathbf{D} - \mathrm{M}(\mathbf{x})$ and does not change the location of each element in error matrix, thus, *the spatial structure in error image is preserved*. Secondly, these existing methods generally use vector-level norm, such as L$_1$ or L$_2$-norm, to constrain residual term. From the statistical meaning, L$_1$ or L$_2$-norm provides an optimal characterization for some data following independent Laplace or Gaussian distribution. As shown in Fig. 5(a), the distributions of some practical noise will be extremely complicated, thus, independent Laplace or Gaussian distribution cannot characterize perfectly them. Compare with these known distributions, *the proposed distribution (3) is closer to the real distribution of some structural noise*. Finally, NMR also emphasizes the global structural information of noise, but it overlooks the local structure. *Our method takes local and global structures of residual term into joint consideration*, thus, it exploits the spatial structure more effectively than NMR.

Remark 2. It should be noted that RLRC [42], RSC [43] and CESR [44] use robust M-estimator to fit some practical noise, but they are still dependent of the independent identically distributed hypothesis, which does not consist with some real- world noise on account of correlations between pixels. Although SSRC [25] attempts to embed the tree structure into noise and use mixed (L$_1$, L$_2$) or (L$_1$, L$_\infty$) norm characterize the spatially contiguous noise, these two norms only exploit the sparsity attribute among groups and the spatial structure information (as shown in Fig. 1(b)) caused by different groups is still neglected. For clarity, we list some existing approaches for characterizing residual term in Table 2.

To use Algorithm 1 to solve model (14), we first need to calculate the sub-gradient of the first part $\|\mathbf{D} - \mathrm{M}(\mathbf{x})\|_{*,w_{ij},*,\sigma_l}$ with regard to $\mathbf{x}$ in the objective (14). Let $\mathbf{A} = \mathbf{D} - \mathrm{M}(\mathbf{x})$,

$$\mathbf{B} = \begin{pmatrix} b_{11} & b_{12} & \dots & b_{1q} \\ b_{21} & b_{22} & \dots & b_{2q} \\ \vdots & \vdots & \ddots & \vdots \\ b_{p1} & b_{p2} & \dots & b_{pq} \end{pmatrix},$$

and $\mathrm{h}(\mathbf{B}) = \|\mathbf{B}\|_*$, where each $b_{ij} = w_{ij}\|\mathbf{A}_{ij}\|_*$ (as shown in Fig. 2). According to the chain rule, we can compute the partial derivative of $\mathrm{h}(\mathbf{B})$ associated with x_i as follows:

$$\frac{\partial \mathrm{h}(\mathbf{B})}{\partial x_i} = Tr\left[\left(\frac{\partial \mathrm{h}(\mathbf{B})}{\partial \mathbf{B}}\right)^T \frac{\partial \mathbf{B}}{\partial x_i}\right], \tag{15}$$

Let $\mathbf{P\Sigma Q}^T$ is the singular value decomposition of the matrix $\mathbf{B}$, then we have $\mathbf{PQ}^T \in \partial\|\mathbf{B}\|_* = \partial \mathrm{h}(\mathbf{B})$. Meanwhile,

$$\frac{\partial \mathbf{B}}{\partial \mathbf{x}} = \begin{pmatrix} \frac{\partial b_{11}}{\partial \mathbf{x}} & \frac{\partial b_{12}}{\partial \mathbf{x}} & \dots & \frac{\partial b_{1q}}{\partial \mathbf{x}} \\ \frac{\partial b_{21}}{\partial \mathbf{x}} & \frac{\partial b_{22}}{\partial \mathbf{x}} & \dots & \frac{\partial b_{2q}}{\partial \mathbf{x}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial b_{p1}}{\partial \mathbf{x}} & \frac{\partial b_{p2}}{\partial \mathbf{x}} & \dots & \frac{\partial b_{pq}}{\partial \mathbf{x}} \end{pmatrix}.$$

For each $\mathbf{A}_{ij}$, we assume its SVD is $\mathbf{P}_{ij}\mathbf{\Sigma}_{ij}\mathbf{Q}_{ij}{}^T$, where $\mathbf{P}_{ij} \in \mathbf{R}^{p_{ij}\times r_{ij}}$, $\mathbf{\Sigma}_{ij} \in \mathbf{R}^{r_{ij}\times r_{ij}}$, $\mathbf{Q}_{ij} \in \mathbf{R}^{q_{ij}\times r_{ij}}$, and r_{ij} is the rank of $\mathbf{A}_{ij}$, then it is easy to see that $w_{ij}\mathcal{M}_{ij}(\mathbf{P}_{ij}\mathbf{Q}_{ij}{}^T)$ is a sub-gradient of b_{ij} with regard to $\mathbf{x}$, where $\mathcal{M}_{ij} : \mathbf{R}^{p_{ij}\times q_{ij}} \to \mathbf{R}^n$ is the adjoint mapping of $\mathrm{M}(\cdot)$, i.e., $\mathcal{M}_{ij}(\mathbf{X}_{ij}) = (\mathrm{tr}(\mathbf{M}_{1,ij}{}^T\mathbf{X}_{ij}),\ \mathrm{tr}(\mathbf{M}_{2,ij}{}^T\mathbf{X}_{ij}), \dots, \mathrm{tr}(\mathbf{M}_{n,ij}{}^T\mathbf{X}_{ij}))$. Let $\mathbf{y}_{ij} = w_{ij}\mathcal{M}_{ij}(\mathbf{P}_{ij}\mathbf{Q}_{ij}{}^T)$, and $y_{ij,k}$ denote the kth element of $\mathbf{y}_{ij}$. Then,

$$\mathbf{Y}(k) = \begin{pmatrix} y_{11,k} & y_{12,k} & \dots & y_{1q,k} \\ y_{21,k} & y_{22,k} & \dots & y_{2q,k} \\ \vdots & \vdots & \ddots & \vdots \\ y_{p1,k} & y_{p2,k} & \dots & y_{pq,k} \end{pmatrix} \in \frac{\partial \mathbf{U}}{\partial x_k}.$$

Thus, $Tr[(\mathbf{PQ}^T)^T\mathbf{Y}(k)] \in \frac{\partial \mathrm{g}(\mathbf{B})}{\partial x_k}$.

By the above analysis, we can obtain a sub-gradient of $\|\mathbf{D} - \mathrm{M}(\mathbf{x})\|_{*,w_{ij},*,\sigma_l}$ w. r. t. $\mathbf{x}$: $\mathbf{w}_1 = [w_1, w_2, \dots, w_n]$, where each $w_k = Tr[(\mathbf{PQ}^T)^T\mathbf{Y}(k)]$, $k= 1, 2, \dots, n$. Since the sub-differential of $\|\mathbf{x}\|_1$ w. r. t. $\mathbf{x}$ can be written as $\mathbf{w}_2 = \mathrm{sgn}(\mathbf{x})$, where $\mathrm{sgn}(\cdot)$ denotes the symbolic function, then a sub-gradient of $\mathrm{J}(\mathbf{x}) = \mathrm{f}(\mathbf{x}) + \rho\mathrm{g}(\mathbf{x})$ is ultimately expressed as

$$\mathbf{w} = \mathbf{w}_1 + \rho\mathbf{w}_2. \tag{16}$$

Since the estimated variate is a vector in model (14), then all uppercase letters in Algorithm 1 will be rewritten as lowercase letters in this subsection. Then, the core iteration in Algorithm 1, namely step 1, becomes

$$\mathbf{y}^{t+1} = \mathbf{v}^t - \mu^t\mathbf{w}^t, \tag{17}$$

4.2. Nesting-structured nuclear norm based matrix completion

The matrix completion aims at recovering a low-rank matrix from partial observations of its entries. Several important real-world problems can be cast as a matrix completion problem, including remote sensing, system identification and recommendation systems. Utilizing nuclear norm minimization, Candès and Recht [45] first investigated the noiseless setting for matrix completion. Subsequently, Candès and Plan [46] further showed that matrix completion is provably accurate by nuclear norm minimization when the few observed entries are corrupted with a small amount of noise. Along with these theoretical results, a large number of efforts have been recently concentrated to develop low-computational yet effective algorithms to cope with nuclear norm minimization for matrix completion, such as interior-point method, singular value thresholding, Alternating Direction Method of Multipliers, singular value projection, accelerated proximal gradient method and so on. To better approximate the rank function, Hu et al. [47] introduced truncated nuclear norm regularization (TNNR) to characterize the matrix variate. Overall experiments on synthetic data and real visual data show the advantages of truncated nuclear norm, but it still confronts SVD for a large-scale matrix, and the two-step strategy will consume a large amount of time. For clarity, we first present the detailed definition of matrix completion in the following.

Suppose that there exists a location set Ω of size $l \times k$ such that the $\mathbf{x}_{ij}$ ($\mathbf{x}_{ij}$ is an element of matrix $\mathbf{X} \in R^{n \times m}$) is observed if and only if $(i, j) \in \Omega$. It is also assumed that Ω is sampled uniformly at random. Define $\mathrm{P}_\Omega(\mathbf{X})$ as the orthogonal projection of $\mathbf{X}$ onto the subspace of matrices that vanish outside. We aim at recovering $\mathbf{X}$ from $\mathrm{P}_\Omega(\mathbf{X})$. Combining model (1), let $\mathrm{f}(\mathbf{X}) = \mathbf{X}$ and $\mathrm{g}(\mathbf{X}) = \|\mathrm{P}_\Omega(\mathbf{X}) - \mathrm{P}_\Omega(\mathbf{D})\|_F^2$, then we can obtain a nesting-structured nuclear norm minimization for matrix completion (NNC) as follows:

$$\min_{\mathbf{X}} \|\mathbf{X}\|_{*,w_{ij},*,\sigma_l} + \rho\|\mathrm{P}_\Omega(\mathbf{X}) - \mathrm{P}_\Omega(\mathbf{D})\|_F^2, \tag{18}$$

The statistical meaning of (18) can be obtained by using the similar technique as in Section 3. The difference is that we need to exchange the roles of $\mathrm{f}(\mathbf{X})$ and $\mathrm{g}(\mathbf{X})$. Compared to the popular nuclear norm minimization for matrix completion, *the model (18) shares the lower time completion since the large-scale matrix X is partitioned into several small blocks*. Observing that our viewpoint is originated from the spatial dependence between pixels of $\mathbf{X}$, and combining the statistical meaning of nesting-structured nuclear norm, *we can see that the proposed problem (18) captures the global structure of X as well as the local structure more fully than the general nuclear norm minimization.*

In order for achieving the optimal solution of (18) by Algorithm 1, we first need to acquire a sub-gradient of $\|\mathbf{X}\|_{*,w_{ij},*,\sigma_l}$ with regard to $\mathbf{X}$. As in Section 2.1, we view $\mathbf{X}$ as the matrix $\mathbf{A}$, then we have $\mathbf{PQ}^T \in \partial\|\mathbf{B}\|_*$. Let $\mathbf{s} = (\mathrm{Vec}(\mathbf{QP}^T))^T$, where $\mathrm{Vec}(\cdot)$ is an operator converting a matrix into a vector, then a sub-gradient of $\|\mathbf{X}\|_{*,w_{ij},*,\sigma_l}$ with regard with $\mathbf{X}$ can be expressed as:

$$\mathbf{W}_1 = \begin{pmatrix} \mathbf{W}_{11} & \mathbf{W}_{12} & \dots & \mathbf{W}_{1q} \\ \mathbf{W}_{21} & \mathbf{W}_{22} & \dots & \mathbf{W}_{2q} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{W}_{p1} & \mathbf{W}_{p2} & \dots & \mathbf{W}_{pq} \end{pmatrix}, \tag{19}$$

where $\mathbf{W}_{ij} = \mathbf{s}_{i+i*(j-1)}\mathbf{P}_{ij}\mathbf{Q}_{ij}^T$.

Meanwhile, the gradient of $\mathrm{g}(\mathbf{X}) = \|\mathrm{P}_\Omega(\mathbf{X}) - \mathrm{P}_\Omega(\mathbf{D})\|_F^2$ can be written as $\mathbf{W}_2 = 2(\mathrm{P}_\Omega(\mathbf{X}) - \mathrm{P}_\Omega(\mathbf{D}))$. Thus, we can get a sub-gradient of $\mathrm{J} = \mathrm{f}(\mathbf{X}) + \rho\mathrm{g}(\mathbf{X})$ as: $\mathbf{W} = \mathbf{W}_1 + \rho\mathbf{W}_2$. Then, the step 1 in Algorithm 1 can be written as:

$$\mathbf{Y}^{t+1} = \mathbf{V}^t - \mu^t\mathbf{W}^t, \tag{20}$$

Compared to TNNR, our method computes the SVD on some small-scale matrices and does not carry out the nesting loop.

Remark 3. By (17) and (20), we see that Algorithm 1 is very simple and can be applied to a far wider variety of problems than other methods such as ADMM, gradient based methods and Newton's method.

5. Convergence and complexity analysis

From the forcing descent step, namely step 2 in Algorithm 1, it holds that the sequence $\{\mathrm{J}(\mathbf{X}^t)\}$ is decreasing monotonously, i.e., $\mathrm{J}(\mathbf{X}^{t+1}) \le \mathrm{J}(\mathbf{X}^t)$ for all $t \ge 0$. Connecting $\mathrm{J}(\mathbf{x}^t) \ge 0$, then the known monotone bounded theorem implies that the objective function sequence $\{\mathrm{J}(\mathbf{x}^t)\}$ generated by Algorithm 1 can ultimately achieve at an accumulation point as $t \to \infty$. Therefore, the convergence of Algorithm 1 does not rely on the convexity of the problem (1). Then, given a parameter $\varepsilon > 0$, it is enough for Algorithm 1 that we only need to adopt the following termination criterion:

$$|\mathrm{J}(\mathbf{X}^{t+1}) - \mathrm{J}(\mathbf{X}^t)| \le \varepsilon, \tag{21}$$

where $|\cdot|$ denotes the absolute value function.

5.1. The choice of step size rule

By a series of experiments, we find that the nonsummable diminishing rule [29] can obtain the better performance both in speed and accuracy. Thus, throughout this paper, $\mu^t = \xi/\sqrt{t+1}$, where $\xi \in [0.01, 0.1]$.

The running time of the proposed Algorithm 1 is mainly embodied in calculating the sub-gradient of the function $\|\mathrm{f}(\mathbf{X})\|_{*,w_{ij},*,\sigma_l}$ for the first step, which is closely related to the designed nesting structure in advance. For the sub-matrix $\mathbf{A}_{ij}$ of dimensions $p_{ij} \times q_{ij}$, where we assume $p_{ij} \ge q_{ij}$ and $p_{ij} = n/p, q_{ij} = m/q$ for all $1 \le i \le n$, $1 \le j \le m$, the time complexity of performing SVD is $O(p_{ij}{q_{ij}}^2)$. For the induced matrix $\mathbf{B}$ with size $p \times q$, the time complexity of performing SVD is $O(pq^2)$, where $p \ge q$ is assumed. Since we only consider the second layer in Fig. 2, the total time complexity for Algorithm 1 is $O(pqp_{11}{q_{11}}^2 + pq^2)$.

5.2. The comparison with nuclear norm minimization

For the nuclear norm minimization, the SVD is implemented on the original matrix $\mathbf{A}$ with dimensions of $n \times m$. Assuming $m \le n$, then the time complexity in the designed algorithm generally is $O(nm^2)$ [56,57]. Because p, q, p_{ij} and $q_{ij} <$ norm, we can obtain that $pqp_{11}{q_{11}}^2 + pq^2 < nm^2$ by the certain partition strategy. Particularly, suppose that $m = n = 100$, then the complexity of nuclear norm minimization is $O(10^6)$, while our method can achieve a complexity of $O(10^4)$. Therefore, compared with nuclear norm minimization, the proposed strategy greatly reduces the time complexity.

6. Experiment and analysis

In this Section, we implement face recognition on four standard face databases, namely the Extended Yale B database, the AR face database, Multi-PIE database and FERET database, to validate the robustness of model (14) for dealing with the structural noise. Several experiments on both synthetic data and real visual data are also conducted to show the effectiveness of model (14) for matrix completion. In this Section, all weights w_{ij} are set as 1.

In face recognition experiments, the ratio of nesting-structured nuclear norm of representation residual and L_2 norm of coefficients for each class is utilized to measure the distance between reconstruction image and classes, that is,

$$\mathrm{r}_i(\mathbf{D}) = \|\mathrm{M}(\mathbf{x}^*) - \mathrm{M}(\pi_i(\mathbf{x}^*))\|_{*,w_{ij},*,\sigma_l}/\|\pi_i(\mathbf{x}^*)\|_2,$$

for $i = 1, \dots, k$, where $\pi_i(\mathbf{x}^*)$ is a vector whose only nonzero entries are the entries in $\mathbf{x}^*$ that are associated with Class i, and $\mathbf{x}^*$ is the optimal representation coefficient obtained by Algorithm 1. Thus, the classification rule is defined as: if $\mathrm{r}_l(\mathbf{D}) = \min_i \mathrm{r}_i(\mathbf{D})$, then $\mathbf{D}$ belongs to Class l.

6.1. Databases

The AR face database [48] contains over 4000 color face images of 126 people (70 men and 56 women), including frontal views of faces with different facial expressions, lighting conditions and occlusions. The pictures of most persons were taken in two sessions (separated by two weeks). Each section contains 13 color images and 120 individuals (65 men and 55 women) participated in both sessions. The images of these 120 individuals were selected and used in our experiment. We manually cropped the face portion of the image and then normalized it to 50 × 40 pixels.

The extended Yale B face database [49] contains 38 human subjects under nine poses and 64 illumination conditions the light

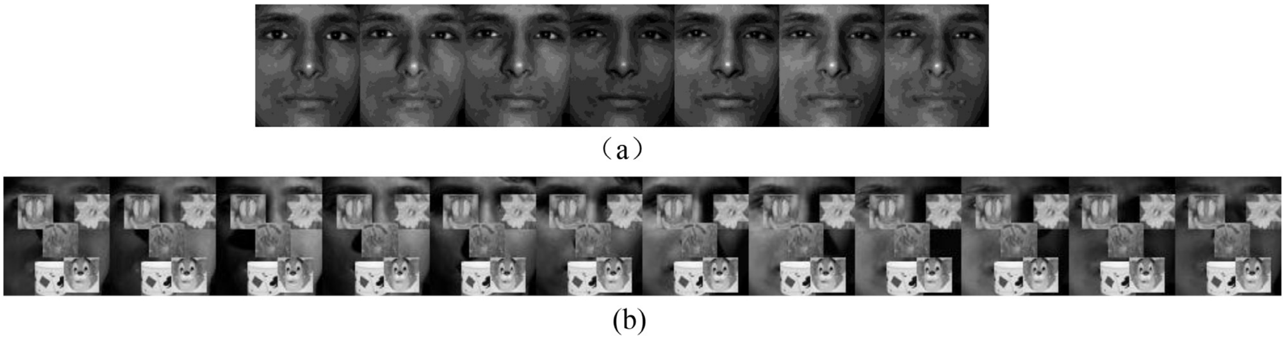


Fig. 6. (a) Seven training samples for one person from the extended Yale B database (b) Twelve test samples for one person from the extended Yale B database.

Table 3
The maximal recognition rates (%) of SRC, LRC, CRC, RSC, RLRC, CESR, SSEC, SSRC, NMR NL$_1$R and nesting-NMR for multiple random occlusions on the Extended Yale B face database.

SRC	CRC	LRC	RLRC	RSC	CESR	SSEC	SSRC	NMR	NL$_1$R	Nesting-NMR
41.7	14.7	18.6	45.9	46.7	42.5	18.6	44.1	47.4	47.8	**69.5**

source direction and the camera axis. The 64 images of a subject in a particular pose are acquired at camera frame rate of 30 frames/s, so there is only small change in head pose and facial expression for those 64 images. All frontal-face images marked with P00 are used, and each image is resized to 42 × 48 pixels and 96 × 84 pixels, respectively.

The CMU Multi-PIE database [50] contains images of 337 different subjects with variations in pose, expression and illumination. Individual attendance varies from 249, 203, 230 and 239 for Sessions 1–4. In our experiment, we use the frontal images with different illuminations and neutral expression. We manually cropped the face portion of the image and then normalized it to 60 × 45 pixels.

The FERET database [51] contains a total of 13,539 face images of 1565 subjects. The images vary in size, pose, illumination, facial expression, and age. We selected 1400 images of 200 individuals (each one has seven images). Each image was cropped to 40 × 40 pixels.

6.2. Experiments using the Extended Yale B database

We design two groups of experiments based on the Extended Yale B database. The first experiment is used to test the advantage of our algorithm for dealing with face recognition with artificial occlusion. To be fair, we use the same experiment setting as in [1]. Here Subsets 1 and Subsets 3 of Extended Yale B are utilized for training and testing, respectively. All the face images are resized to 96 × 84. We add five unrelated randomly block images into each test image in Subset 3. To be challenging, the locations of five block images are limited to some key parts in a face (but small-amplitude random changes), e.g., eyes, nose and mouth. The training samples and test samples for one person are shown in Fig. 6. The recognition rates of LRC [40], SRC [39], CRC [38], RSC [43], RLRC [42], CESR [44], SSEC [52], SSRC [25], NMR [3], NL$_1$R [6] and Nesting-NMR are summarized in Table 3. From Table 3, we can see that the advantage of the Nesting-NMR is quite evident, which achieves an improvement of 21.7% than the second best method: NL$_1$R (47.8%). Although both NMR and NL$_1$R consider the global structure of the error image, the ignoring of local structure leads to the undesired performance. Meanwhile, what is noteworthy is that NMR and NL$_1$R obtain the better results than other methods, which implies that it is indeed necessary to consider the structure of the error image in regression based models. Among these vector based methods, RSC (46.7%) and RLRC (45.9%) show some advantages as compared to SRC (41.7%) and LRC (18.6%), which demonstrates the effectiveness of M-estimator for characterizing noise image. The above results tell us that exploiting fully the structural information of noise image can effectively promote the performance of face recognition.

The second experiment is used to verify the robustness of the proposed algorithm to illumination. Under the settings of the previous experiment, Subset 1 is chosen as training images. Subset 4 and 5 are selected as testing images, respectively. Fig. 7 exhibits the samples for one person. Table 4 lists the results of some latest approaches and the proposed method. We can find that Nesting-NMR overall achieves much higher recognition rates than the other methods. Meanwhile, it is seen that some image-level methods, such as NMR, NL$_1$R and Nesting-NMR show the better results than vector-level methods such as SRC and CRC, which demonstrates that the face recognition performance benefits from the exploiting of the spatial structure for the face images. For Subset 4, at least 5.5% improvement is achieved by Nesting-NMR as compared to RLRC. For Subset 5, the advantage of Nesting-NMR is more apparent. The recognition rates of other methods are all less than 50%, while that of the proposed method reaches 63.5%. It seems that the effect of the structure information to recognition performance become more important with the increasing of illumination level. Therefore, considering the spatial structure of an error image is indispensable in face recognition with illumination changes. Meanwhile, connecting the global with local structures of the error image can further improve the performance of face recognition.

6.3. Experiments using the AR database

We evaluate the effectiveness of Nesting-NMR in coping with face recognition with real disguise on the AR database. Twenty-six face images of these 120 individuals are selected and used in our experiment. Eight images of them are used for training, which vary as follows: (a) neutral expression, (b) smiling, (c) angry, (d) screaming, (e)–(h) are taken under the same conditions. Eighteen images of them are used for testing, but we will set two different cases: (1) face images with glasses: Images from the testing set vary as follows: (i) wearing sunglasses (j) wearing sunglasses and left light on (k) wearing sunglasses and right light on, and (l)–(n) are taken under the same conditions as (i)–(k). (2) face images with scarf: Images from the testing set vary as follows: (i) wearing scarf (j) wearing scarf and left light on (k) wearing scarf and right light on, and (l)–(n) are taken under the same conditions as (i)–(k).

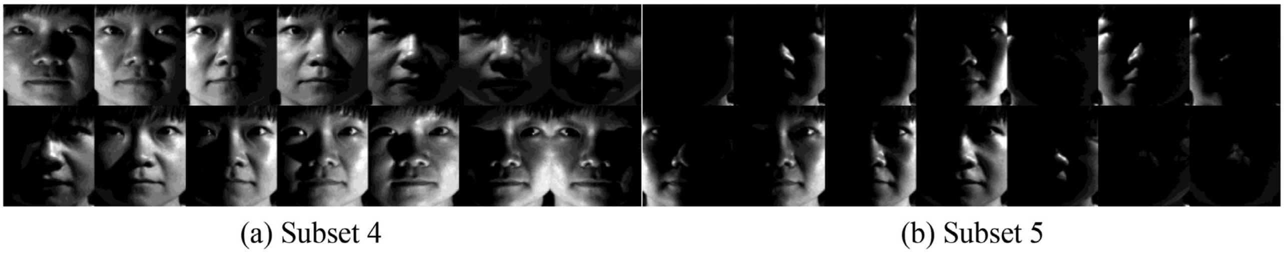


Fig. 7. Sample images with different illumination conditions from the Extended Yale B database.

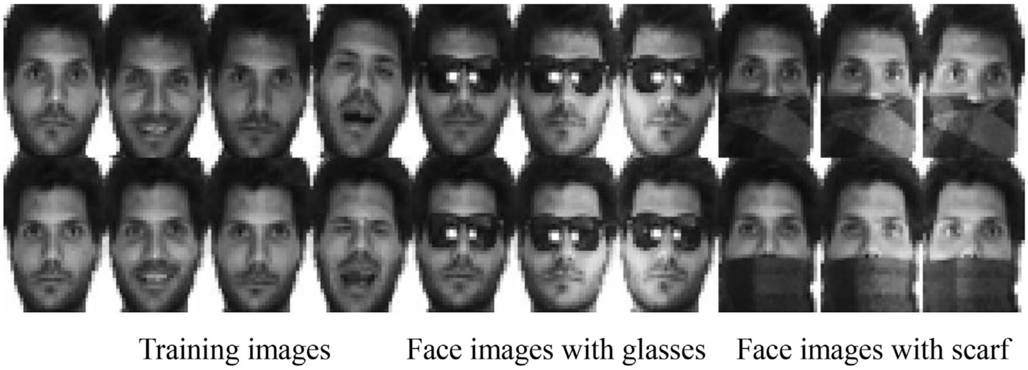


Fig. 8. The sample images for one person from AR face database.

Table 4
The maximal recognition rates (%) of SRC, LRC, CRC, RSC, RLRC, CESR, SSEC, SSRC, NMR NL_1R and nesting-NMR for illumination changes on the Extended Yale B face database.

Cases	SRC	CRC	LRC	RLRC	RSC	CESR	SSEC	SSRC	NMR	NL_1R	Nesting -NMR
Subset 4	78.4	88.0	87.6	89.7	80.5	36.8	20.6	79.3	90.2	93.9	**95.2**
Subset 5	28.8	35.7	42.2	43.2	36.7	22.2	12.5	32.4	47.9	48.1	**63.5**

Table 5
Recognition rates (%) of LRC, CRC, SRC, CESR, RSC, SSEC, NMR, NL_1R and nesting-NMR for real disguise on the AR Database.

Cases	SRC	CRC	LRC	RLRC	RSC	CESR	SSEC	SSRC	NMR	NL_1R	Nesting -NMR
Glasses	94.5	88.5	91.2	94.7	92.2	95.1	82.1	95.2	95.3	95.5	**96.5**
Scarf	54.6	61.0	27.6	52.9	60.2	34.9	42.9	61.3	65.0	64.8	**68.5**

All the face images are resized to 45 × 30. Thus, for each case, the total number of training samples is 840. Fig. 8 presents the sample images for one person from AR database.

In all cases mentioned above, SRC, LRC, RLRC, CRC, RSC, CESR, SSEC, SSRC, NMR, NL_1R and the proposed Nesting-NMR are, respectively, used for face classification. The maximal recognition rate of each method is compared in Table 5, from which we find that Nesting-NMR gets the better results than state-of-the-art methods. For face recognition with glasses, the performances of CESR and SRC is very competitive, which achieve the recognition rates of 95.1% and 94.5%, respectively. For scarf disguise, NMR works the second best, but lags behind our method by 3.5%. As the previous experiments, CRC (88.5%, 61.0%) and LRC (91.2%, 27.6%) are still not suitable for describing the structural noise, while some structural methods such as NL_1R (95.5%, 64.8%) and SSRC (95.2%, 61.3%) present the better results. This experiment implies that the proposed nesting-structured nuclear norm fits better to characterize real disguises than L_1, L_2 or even nuclear norm.

6.4. Experiments using the multi-PIE database

In this subsection, an experiment is conducted on the Multi-PIE database. There are 249 subjects in Session 1, and 166, 160, 175 subjects in Sessions 2, 3 and 4, respectively. The similar experimental setting as in [55] is adopted. We choose 7 frontal images with slight illuminations (Session 1) {05, 06, 07, 08, 15, 16, 17} from each subject as training images and another 7 frontal images with severe illuminations {00, 01, 02, 11, 12, 13, 19} per subject from Session 2 are used as test images. We add randomly three irregular block occlusions, including tiger, butterfly and lotus, into the test images to validate the robustness of our methods to illumination plus irregular occlusion (as shown in Fig. 9). Table 6 lists recognition rates of all methods. It is clear that Nesting-NMR obtains the highest recognition rate: 95.6%. NMR achieves the second highest recognition rate: 94.3%. That means that 1.3% recognition errors can be avoided by using Nesting-NMR instead of NMR. Meanwhile, we also see that LRC and CRC are sensitive to irregular occlusion. Thus, we can draw a conclusion that the proposed method is effective for face recognition with irregular occlusion, which implies the statistical meaning of Nesting-NMR is closer to the nature of the structural noise.

6.5. Experiments using the FERET database

In this part, a challenging experiment is designed to handle face recognition with pose, illumination and facial expression on FERET

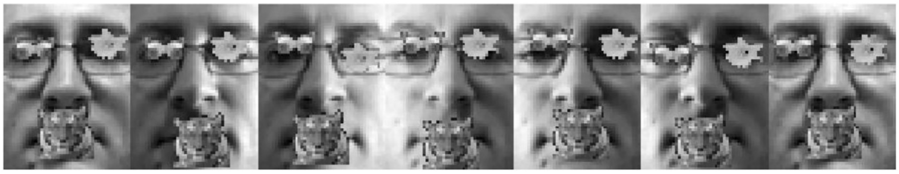


Fig. 9. The testing images for one person from Multi-PIE database.

Fig. 10. The samples for one person from FERET face database.

Table 6
Recognition rates (%) of LRC, RLRC, CRC, SRC, CESR, RSC, SSEC, SSRC, NMR, NL$_1$R and nesting-NMR for illumination variations on the Multi-PIE Database.

SRC	CRC	LRC	RLRC	RSC	CESR	SSEC	SSRC	NMR	NL$_1$R	Nesting -NMR
94.1	63.3	68.1	93.8	91.7	94.2	81.6	94.5	94.3	94.2	**95.6**

Table 7
Recognition rates (%) of LRC, RLRC, CRC, SRC, CESR, RSC, SSEC, SSRC, NMR, NL$_1$R and our methods on the FERET face database.

SRC	CRC	LRC	RLRC	RSC	CESR	SSEC	SSRC	NMR	NL$_1$R	Nesting-NMR
72.95 ± 13.06	51.29 ± 13.23	72.12 ± 12.65	73.08 ± 13.60	73.06 ± 12.78	74.87 ± 14.07	59.75 ± 13.72	74.12 ± 14.89	71.63 ± 11.91	74.29 ± 12.9	**76.67** ± 12.06

database. 7 images from FERET database for each subject are randomly selected and in total 200 subjects are used for our training and test, and each image is resized to 40 × 40 pixels. Fig. 10 shows the samples of a person from this dataset. We choose a random subset with 4 images per subject to form the training set and take the rest for test. This experiment is repeated over 10 random splits of the data set. The average accuracy and the standard deviation of each algorithm are exhibited in Table 7. Compared with other robust methods such as SRC, RSC and RLRC, CESR obtains better performance. But they are still inferior to the proposed method. Nesting-NMR integrates the spatial structure of the noise into modeling, while other methods do not take this information into account. Thus, our algorithm obtains the leading results. In addition, standard deviation of Nesting-NMR is relatively low as compared to CESR, that is, it is robust to different partitioning data. The above experiment further demonstrates that preserving the structural information of the noise image does contribute to face recognition with pose, illumination and facial expression.

6.6. Matrix completion on synthetic data

We generate the rank-r matrix $\mathbf{D}$ as a product $\mathbf{UV}^T$, where $\mathbf{U}$ and $\mathbf{V}$ are independent $m \times r$ matrices whose elements are independent and identically distributed (i.i.d) samples from standard Gaussian distribution $N(0, 1)$. Thus, entries of $\mathbf{D}$ have mean 0 and variance r. The locations of observed indices Ω are sampled uniformly at random. Let Λ be the percentage of observed entries over $m \times m$. We generate synthetic data $\mathbf{X}$ by $\mathbf{X} = \mathbf{D} + \lambda\mathbf{E}$, where $\mathbf{E}$ is Gaussian white noise with zero mean and standard deviation of one and λ is the noise level. Suppose that $\mathbf{X}_{sol}$ is the recovered solution by a certain algorithm. We define the total reconstruction error by $RE = \|\mathbf{X}_{sol} - \mathbf{D}\|_F / \|\mathbf{D}\|_F$, which is a widely used metric in matrix completion. Note that our programming environment is Matlab 2011, and all algorithms are implemented on a Core Duo 2.93 GHz with 4 G RAM desktop.

In this experiment, we first set $\Lambda = \lambda = 0.5$, and select $m = 1000$ and $m = 10000$, respectively. All the chosen algorithms are run 20 times with the underlying rank r lying between 10 and 100. To be fair, we try to choose the optimal parameters for all methods. The mean errors for TNNR, SVP [53], SVT [54] and our method are reported in Fig. 11(a) and (b), respectively. What is worth mentioning is that we consider TNNR-ADMMAP as TNNR since it gets the best results in [47]. Compared with other methods, our approach achieves higher accuracy. It is chiefly because our method combines the global and local structures in modeling, which exploits fully the structure information of $\mathbf{X}$. Meanwhile, the result of TNNR is considerably competitive. And the performance of SVT and SVP is almost the same.

Then, we fix the rank $r = 100$, $m = n$ and $\Lambda = \lambda = 0.5$. All the chosen algorithms are run 20 times with the underlying size m lying from 1000 to 10,000. The average time consuming for all methods is compared in Fig. 11(c). Since the proposed method carries out SVD on the smaller matrices and avoids the nesting loop, it consumes less time than TNNR. With the increasing of dimensions, the advantage of the speed for our method becomes more obvious. Additionally, it is found that the time consuming of SVT is relatively high as compared to other methods.

6.7. Matrix completion on real visual data

In order to verify the effectiveness of our algorithm on real data, as well as illustrate it in a visible approach, we apply the nesting-structured nuclear norm based matrix completion for image recovery. As a colored image is commonly represented as three matrices (containing red, green and blue components respectively), we independently deal with each of three matrices and combine them together to obtain the final results. We compare the proposed method with several representative matrix completion algorithms, including SVT, SVP and TNNR. Performances of different algorithms are evaluated by the well-known PSNR (Peak

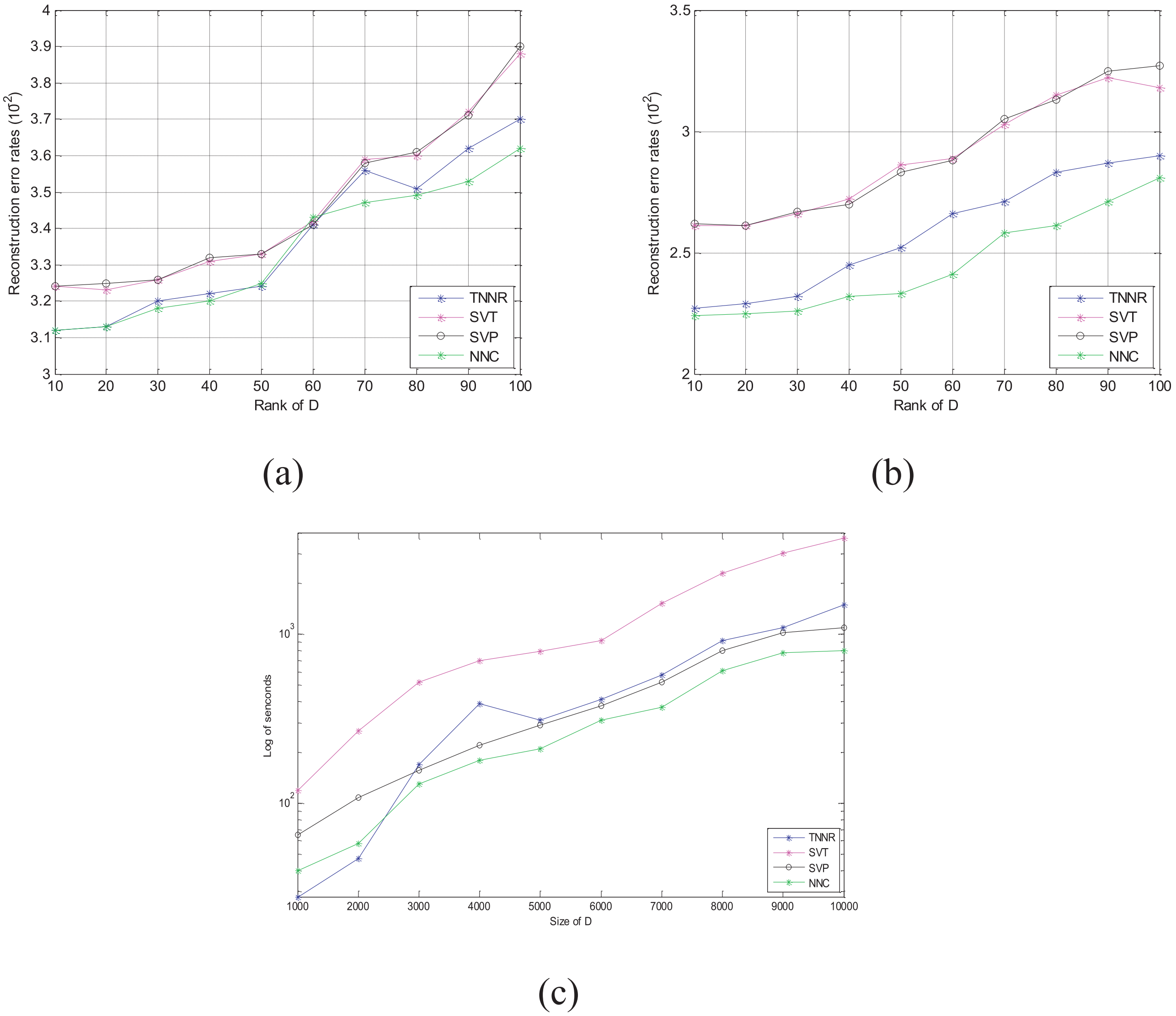


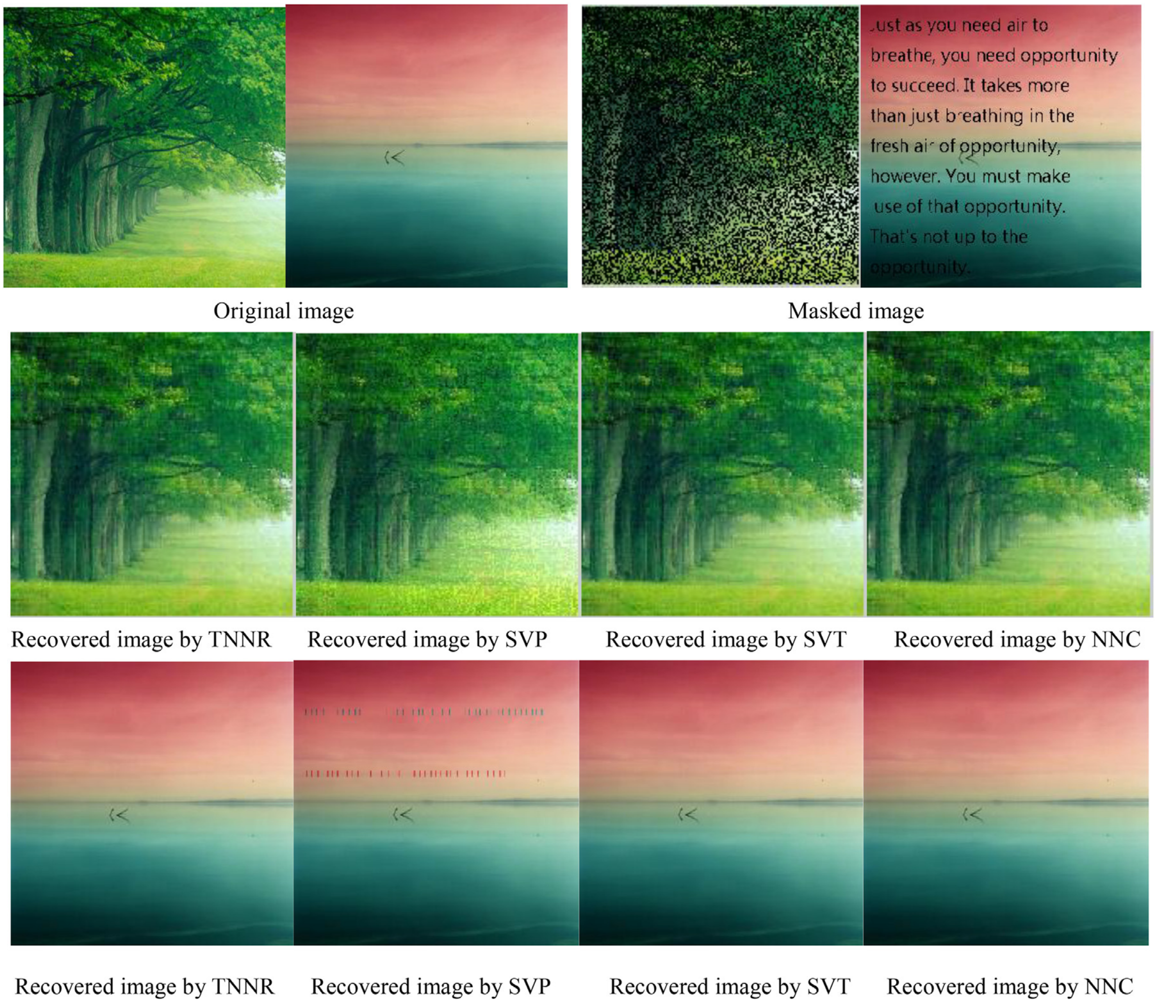
Fig. 11. (a) The reconstruction error versus the matrix rank for $m = 1000$; (b) The reconstruction error versus the matrix rank for $m = 10000$; (c) The time consuming versus the size of matrix from 1000 to 10,000.

Table 8
The Comparison of PSNR Values and average running time (second) by different matrix completion algorithms (corresponding to Fig. 12).

Cases	TNNR		SVP		SVT		NNC	
	PSNR	Time	PSNR	Time	PSNR	Time	PSNR	Time
Random mask	24.389	18.10	18.263	20.2	21.398	60.10	**25.183**	**10.3**
Text mask	37.638	23.1	10.695	21.9	31.532	59.1	**39.325**	**8.90**

Signal-to-Noise Ratio) metric [43]. We try to tune the parameters to be optimal of the chosen algorithms and report the best result.

Two experiments on the real images are implemented. The first is a relatively easy matrix completion problem with random mask, where the missing entries are randomly distributed on the 300 × 300 image (as shown in Figs. 12 or 13). Second experiment uses text mask. It is generally agreed that image inpainting with text mask is more difficult since the observed pixels are not randomly sampled and text mask may result in loss of important image information. We report our results in Figs. 12 and 13. It is seen that the results of the compared methods are encouraging for random mask, but SVP is still sensitive to text mask. Notice that the higher PSNR value represents the better performance. The results in Tables 8 and 9 further demonstrate the advantage of our method over the other methods. Therefore, considering the structural information (including local and global structures) does play a pivotal role in matrix completion problem. Next, we compare the time consuming of all methods for the above two experiment. Each of the above experiments is run 20 times and we compute the average time consuming. The results are also shown in Tables 8 and 9, from which we find that our method is still fastest. Especially, the time consuming of NNC is less than half of TNNR. SVT is significantly more time-consuming than SVP. Thus, our method is reliable in both time consuming and performance.



ust as you need air to breathe, you need opportunity to succeed. It takes more than just b'eathing in the fresh ai' of opportunity, however. You must make use of that opportunity. That's not up to the opportunity.

Original image

Masked image

Recovered image by TNNR

Recovered image by SVP

Recovered image by SVT

Recovered image by NNC

Recovered image by TNNR

Recovered image by SVP

Recovered image by SVT

Recovered image by NNC

Fig. 12. Comparison of image recovery by using different matrix completion algorithms for random mask.

Table 9
The Comparison of PSNR values and average running time (second) by different matrix completion algorithms (corresponding to Fig. 13).

Cases	TNNR		SVP		SVT		NNC	
	PSNR	Time	PSNR	Time	PSNR	Time	PSNR	Time
Random mask	25.541	19.4	20.450	33.2	16.684	55.6	**25.973**	**6.78**
Text mask	17.389	12.7	15.367	28.9	16.231	59.1	**17.582**	**6.24**

Table 10
The influence of different factorization strategies to the recognition performance on the Extended Yale B database.

Factorization		1 × 1	3 × 3	4 × 4	6 × 6	8 × 7	12 × 12	16 × 14	24 × 21	48 × 42	96 × 94
Cases	Occlusion	43.2	59.3	65.2	67.2	**69.5**	68.5	62.0	61.2	52.7	47.4
	Subset 4	83.0	87.2	91.5	93.2	**95.2**	94.1	94.8	93.8	93.6	95.2
	Subset 4	45.1	53.9	60.7	62.7	**63.5**	62.1	61.2	59.2	50.8	47.9

Table 11
The influence of different factorization strategies to the recognition performance on the AR database.

Factorization		1 × 1	3 × 3	5 × 3	5 × 5	5 × 6	9 × 10	15 × 15	45 × 30
Cases	Sunglasses	91.2	93.7	95.2	96.2	**96.5**	96.1	95.1	95.3
	Scarf	59.5	63.2	66.3	68.1	**68.5**	67.8	65.2	65.0

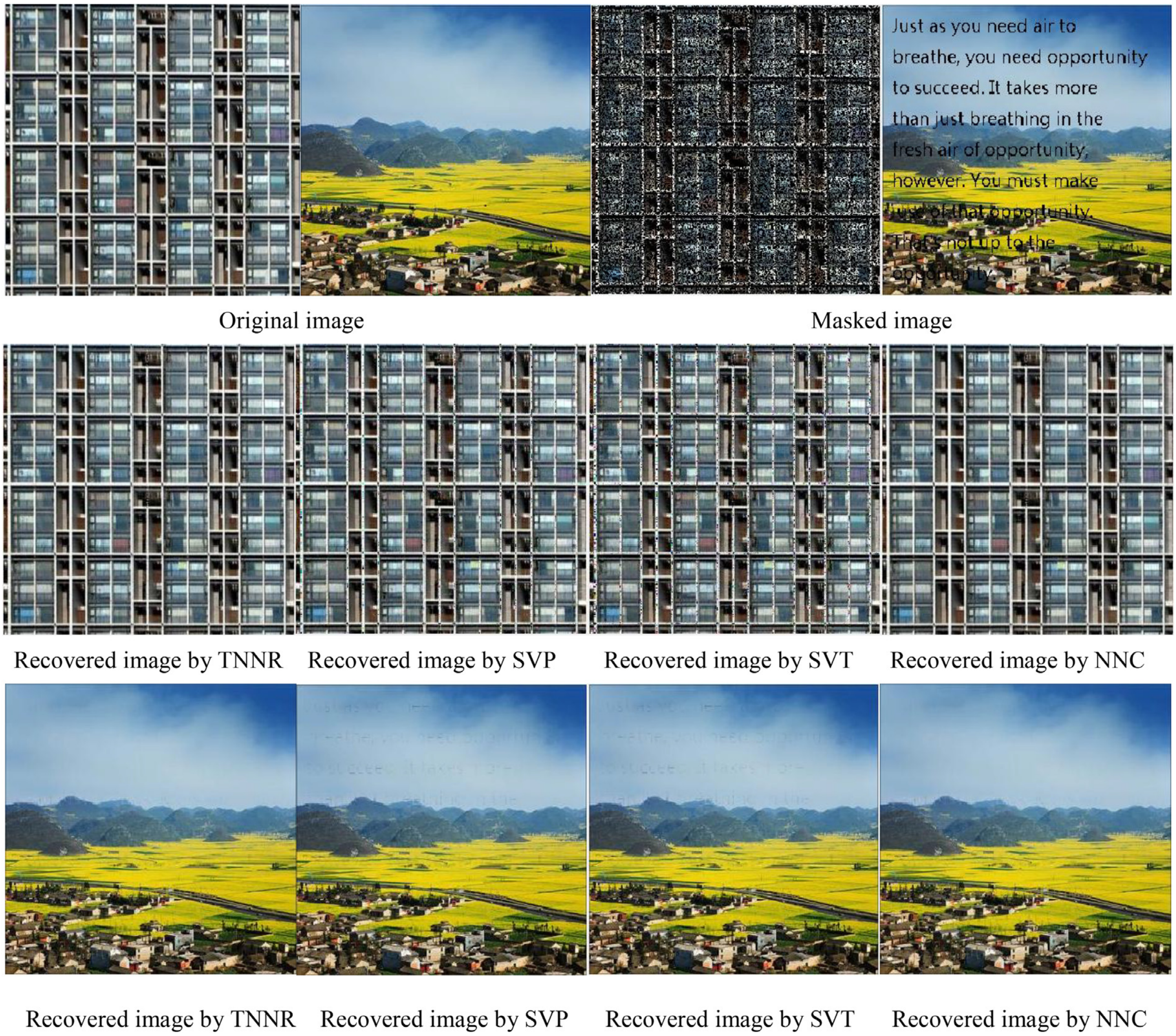


Fig. 13. Comparison of image recovery by using different matrix completion algorithms for random mask.

6.8. Performances of different factorization strategies

We take the extended Yale B and AR databases for the examples to investigate the optimal factorization approach in this subsection. Different factorization strategies are chosen to deal with face recognition with occlusion, illumination or real disguises (which corresponds to Tables 3–5). The experimental results are listed in Tables 10 and 11, respectively. It is seen that when $p,\ q \in [5,\ 10]$, the optimal recognition rates are achieved. In fact, each sub-matrix with dimensions of $p \times q(\text{where}\, p \times q \in [5,\ 10])$ can reflect the local structure of original error image. Thus, such a factorization strategy can capture the local structure information well, leading to the outstanding performance.

6.9. The choice of parameters

In this subsection, we discuss how the parameters in model (1) affect the performance of the proposed method. Throughout this paper, the face images are carried out the normalization processing. This makes our method robust to different parameters. For the convenience, all weights w_{ij}^1 are set as 1. Thus, there is only a parameter ρ in the proposed model. Fig. 14 shows the influence of different parameters ρ to the recognition performance on the extended Yale B (Section 6.2), AR (Section 6.3), Multi-PIE (Section 6.4) and FERET databases (Section 6.5). From Fig. 14, it is found that when ρ is around 10, the optimal recognition rates are obtained for different face recognition scenarios. In addition, the choice of parameters on matrix completion is also discussed. The similar

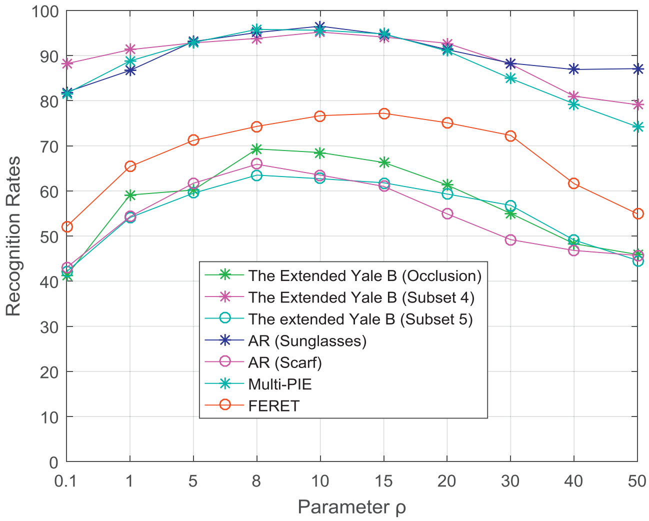


Fig. 14. The influence of different parameters ρ to the recognition performance on the Extended Yale B, AR, Multi-PIE and FERET databases.

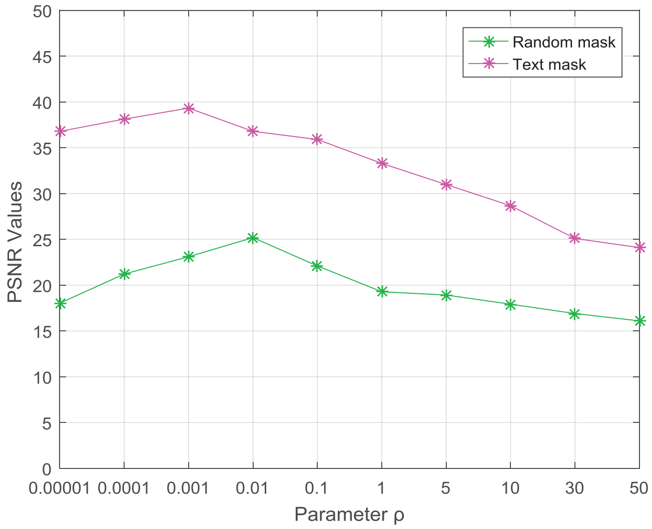


Fig. 15. The influence of different parameters ρ to the recognition performance for matrix completion.

experiment setting as in Section 6.7 is used. The influence of different parameters ρ to the matrix completion is exhibited in Fig. 15, from which it is seen that the best results are achieved in the range of [0.001, 0.01].

7. Conclusions

Considering the structural information of the matrix variate has become the hot topic in pattern recognition and computer vision. This paper establishes a nesting matrix structure and uses it to induce a nesting-structured nuclear norm. Based on this norm, we present a nesting-structured nuclear norm minimization (NSNM) model and develop an improved sub-gradient method to solve it. NSNM captures effectively the global and local structures of a matrix variate, and owns the lower time complexity than traditional nuclear norm minimization due to the effect of partition. The analyzed statistical meaning shows fully the reasonability of our method. What is more, the proposed framework is applied to matrix regression and completion problems. In our opinion, the main limit of our method is in dealing with sparse noise. Since our method integrates the structural information of noise into modeling, its advantage is not obvious for sparse noises which are independently generated. Meanwhile, our method may not handle well non-aligned face recognition problem, which is the common limitation of regression based methods. Therefore, in the future, we need to develop an effective framework to overcome the above limitations. We will analyze the essential attribute of the noise in real data from statistical and physics viewpoints. Leveraging these properties, we can automatically fit noise to improve the performance of models. In addition, it is an interesting research direction to provide the theoretical analysis for the NNC model.

Acknowledgments

The authors would like to thank the editor and the anonymous reviewers for their critical and constructive comments and suggestions. This work was supported by the National Science Fund of China under Grant No. U1713208, Program for Changjiang Scholars.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.patcog.2019.02.011.

References

[1] G.C. Liu, Z.C. Lin, S.C. Yan, J. Sun, Y. Yu, Y. Ma, Robust recovery of subspace structures by low-rank representation, IEEE Trans. Pattern Anal. Mach. Intell. 35 (1) (2013) 171–184.
[2] E.J. Candès, X.D. Li, Y. Ma, J. Wright, Robust principal component analysis? J. ACM 58 (May (3)) (2011) Art. ID 11.
[3] J. Yang, L. Luo, J.J. Qian, Y. Tai, F.L. Zhang, Y. Xu, Nuclear norm based matrix regression with applications to face recognition with occlusion and illumination changes, IEEE Trans. Pattern Anal. Mach. Intell. 39 (1) (2017) 156–171.
[4] F. Zhang, J. Yang, J.J. Qian, Y. Xu, Nuclear norm-based 2-DPCA for extracting features from images, IEEE Trans. Neural Netw. Learn. Syst. 26 (10) (2015) 2247–2260.
[5] J. Chen, J. Yang, L. Luo, J. Qian, W. Xu, Matrix variate distribution-induced sparse representation for robust image classification, IEEE Trans. Neural Netw. Learn. Syst. 26 (10) (2015) 2291–2300.
[6] L. Luo, J. Yang, J.J. Qian, Y. Tai, Nuclear-L_1 norm joint regression for face reconstruction and recognition with mixed noise, Pattern Recognit. 48 (12) (2015) 3811–3824.
[7] L. Luo, L. Chen, J. Yang, J. Qian, B. Zhang, Tree-structured nuclear norm approximation with applications to robust face recognition, IEEE Trans. Image Process. 25 (12) (2016) 5757–5767.
[8] L. Luo, J. Yang, J. Qian, Y. Tai, G.F. Lu, Robust image regression based on the extended matrix variate power exponential distribution of dependent noise, IEEE Trans. Neural Netw. Learn. Syst. (2016) 1–15.
[9] C. Xu, T. Liu, D. Tao, C. Xu, Local rademacher complexity for multi-label learning, IEEE Trans. Image Process. 25 (3) (2016) 1495–1507.
[10] Y. Yan, M. Tan, I. Tsang, Y. Yang, C. Zhang, Q. Shi, Scalable maximum margin matrix factorization by active riemannian subspace search, IJCAI, 2015.
[11] J.J. Qian, L. Luo, J. Yang, F.L. Zhang, Z.C. Lin, Robust nuclear norm regularized regression for face recognition with occlusion, Pattern Recognit. 48 (10) (2015) 3145–3159.
[12] X. Zhang, F. Sun, G. Liu, Y. Ma, Fast low-rank subspace segmentation, IEEE Trans. Knowl. Data Eng. 26 (5) (2014) 1293–1297.
[13] T. Bouwmans, A. Sobral, S. Javed, S.K. Jung, and E.H. Zahzah, Decomposition into low-rank plus additive matrices for background/ foreground separation: a review for a comparative evaluation with a large-scale dataset, 2015, arXiv:1511.01245.
[14] J.X. Li, L. Luo, F.L. Zhang, J. Yang, D. Rajan, Double low rank matrix recovery for saliency fusion, IEEE Trans. Image Process. 25 (9) (2016) 4421–4432.
[15] L. Lu, W. Xu, S.Z. Qiao, A fast SVD for multilevel block Hankel matrices with minimal memory storage, Numer. Algo. (2014) 1–17.
[16] A. Majumdar, R. Ward, Fast SVD free low-rank matrix recovery: application to dynamic MRI reconstruction, MedCom, 2014.
[17] J.F. Cai, O. Stanley, Fast singular Value Thresholding Without Singular Value Decomposition, 2010 UCLA CAM Report 5.
[18] T.H. Oh, Y. Matsushita, Y.W. Tai, I.S. Kweon, Fast randomized singular value thresholding for nuclear norm minimization, in: CVPR, 2015, pp. 4484–4493.
[19] Z. Lin, M. Chen, L. Wu, Y. Ma, The Augmented Lagrange Multiplier Method for Exact Recovery of Corrupted Low-Rank Matrices, 2009 *UIUC Technical Report*, November.
[20] X. Liu, Z. Wen, and Y. Zhang, Limited memory block Krylov subspace optimization for computing dominant singular value decomposition, Preprint, 2012.
[21] Z. Lin and S. Wei, A block Lanczos with warm start technique for accelerating nuclear norm minimization algorithms, Preprint, 2010.
[22] X. Shi, Y. Yang, Z. Guo, Z. Lai, Face recognition by sparse discriminant analysis via joint $L_{2,1}$-norm minimization, Pattern Recognit. 47 (7) (2014) 2447–2453.
[23] J. Wang, M. Wang, P.P. Li, L.Q. Liu, Z.Q. Zhao, X.G. Hu, X.D. Wu, Online feature selection with group structure analysis, IEEE Trans. Knowl. Data Eng. 27 (11) (2015) 3029–3041.
[24] N. Rao, R. Nowak, C. Cox, and T. Rogers, Classification with sparse overlapping groups, 2014, arXiv:1402.4512.
[25] K. Jia, T.H. Chan, Y. Ma, Robust and practical face recognition via structured sparsity, ECCV, 2012.
[26] E. Polak, Computational Methods in Optimization, Academic press, New York, 1971.
[27] S. Ji, J.P. Ye, An accelerated gradient method for trace norm minimization, ICML, 2009.
[28] N. Shor, Minimization Methods for Non-differentiable Functions, Springer Series in Computational Mathematics, 1985.
[29] S. Boyd, L. Xiao, A. Mutapcic, Subgradient methods, Lecture Notes of EE392o, Stanford University, 2004 Autumn Quarter.
[30] A. Nedić, A. Ozdaglar, Subgradient methods for saddle-point problems, J. Optim. Theory Appl. 142 (1) (2009) 205–228.
[31] Y. Nesterov, V. Shikhman, Quasi-monotone subgradient methods for nonsmooth convex minimization, J. Optim. Theory Appl. 165 (3) (2015) 917–940.
[32] R.T. Rockafellar, R. Wets, Variational analysis, Volume 317 of Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences], 1998.
[33] J. Bolte, S. Sabach, M. Teboulle, Proximal alternating linearized minimization for nonconvex and nonsmooth problems, Math. Program. 146 (1) (2014) 459–494.
[34] J.Y.B. Cruz, On proximal subgradient splitting method for minimizing the sum of two nonsmooth convex functions, 2014, arXiv:1410.5477.

[35] O. Shamir, and T. Zhang, Stochastic gradient descent for non-smooth optimization: convergence results and optimal averaging schemes, 2012, arXiv:1212.1824.

[36] A. Neumaier, OSGA: a fast subgradient algorithm with optimal complexity, Math. Program. (2014) 1–21.

[37] Z. Zhang, and V. Saligrama, RAPID: Rapidly Accelerated Proximal Gradient Algorithms for Convex Minimization, 2014, arXiv:1406.4445.

[38] L. Zhang, M. Yang, X.C. Feng, Sparse representation or collaborative representation which helps face recognition? ICCV, 2011.

[39] J. Wright, A.Y. Yang, A. Ganesh, S.S. Sastry, Y. Ma, Robust face recognition via sparse representation, IEEE PAMI 31 (2) (2009) 210–227.

[40] I. Naseem, R. Togneri, M. Bennamoun, Linear regression for Face Recognition, IEEE PAMI 32 (11) (2010) 2106–2112.

[41] M. Yang, Z. Feng, S.C. Shiu, L. Zhang, Fast and robust face recognition via coding residual map learning based adaptive masking, Pattern Recognit. 47 (2) (2014) 535–543.

[42] I. Naseem, R. Togneri, M. Bennamoun, Robust regression for face recognition, Pattern Recognit. 45 (1) (2012) 104–118.

[43] M. Yang, L. Zhang, J. Yang, D. Zhang, Robust sparse coding for face recognition, CVPR, 2011.

[44] R. He, W.S. Zheng, B.G. Hu, Maximum correntropy criterion for robust face recognition, IEEE PAMI 33 (8) (2011) 1561–1576.

[45] E.J. Candès, B. Recht, Exact matrix completion via convex optimization, Found. Comput. Math. 9 (6) (2009) 717–772.

[46] E.J. Candès, Y. Plan, Matrix completion with noise, IEEE Proc. 98 (2010) 925–936 2010.

[47] Y. Hu, D. Zhang, J. Ye, X. Li, X. He, Fast and accurate matrix completion via truncated nuclear norm regularization, IEEE PAMI 35 (9) (2014) 2117–2130.

[48] A.M. Martinez, The ar face database, 1998 Technical Report 24, CVC.

[49] K. Lee, J. Ho, D. Kriegman, Acquiring linear subspaces for face recognition under variate lighting, IEEE PAMI 27 (5) (2005) 684–698.

[50] R. Gross, I. Matthews, J. Cohn, T. Kanade, S. Baker, Multi-PIE, Image Vis. Comput. 28 (5) (2010) 807–813.

[51] P.J. Phillips, H. Wechsler, J. Huang, P.J. Raussa, The FERET database and evaluation procedure for face-recognition algorithms, Image Vis. Comput. 16 (5) (1998) 295–306.

[52] X.X. Li, D.Q. Dai, X.F. Zhang, C.-X. Ren, Structured sparse error coding for face recognition with occlusion, *IEEE Trans. Image Process.* 22 (5) (2013) 1889–1999.

[53] R. Meka, P. Jain, I.S. Dhillon, Guaranteed rank minimization via singular value projection, NIPS, 2010.

[54] J.F. Cai, E.J. Cande's, Z. Shen, A singular value thresholding algorithm for matrix completion, SIAM J. Optim. 20 (2010) 1956–1982.

[55] L. Luo, J. Yang, B. Zhang, J.L. Jiang, H. Huang, Nonparametric Bayesian correlated group regression with applications to image classification, IEEE Trans. Neural Netw. Learn. Syst. 99 (2018) 1–15.

[56] L. Zhang, Q. Zhang, L. Zhang, D. Tao, X. Huang, B. Du, Ensemble manifold regularized sparse low-rank approximation for multiview feature embedding, Pattern Recognit. 48 (10) (2015) 3102–3112.

[57] J. Wen, B. Zhang, Y. Xu, J. Yang, N. Han, Adaptive weighted nonnegative low-rank representation, Pattern Recognit. 81 (2018) 326–340.

Lei Luo received the B.S. degree from Xinyang Normal University, Xinyang, China in 2008, the M.S. degree from Nanchang University, Nanchang, China in 2011. 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 pattern recognition and optimization algorithm.

Jian Yang received the Ph.D. degree from Nanjing University of Science and Technology (NUST), on the subject of pattern recognition and intelligence systems in 2002. In 2003, he was a Postdoctoral researcher at the University of Zaragoza. From 2004 to 2006, he was a Postdoctoral Fellow at Biometrics Centre of Hong Kong Polytechnic University. From 2006 to 2007, he was a Postdoctoral Fellow at Department of Computer Science of New Jersey Institute of Technology. Now, he is a Chang-Jiang professor in the School of Computer Science and Technology of NUST. He is the author of more than 200 scientific papers in pattern recognition and computer vision. His papers have been cited more than 6000 times in the Web of Science, and 15,000 times in the Scholar Google. His research interests include pattern recognition, computer vision and machine learning. Currently, he is/was an associate editor of Pattern Recognition, Pattern Recognition Letters, IEEE Trans. Neural Networks and Learning Systems, and Neurocomputing. He is a Fellow of IAPR.

Yigong Zhang 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 SLAM.

Yong Xu (M'06) was born in Sichuan, China, in 1972. He received the B.S. and M.S. degrees in 1994 and 1997, respectively, and the Ph.D. degree in pattern recognition and intelligence system from the Nanjing University of Science and Technology, Nanjing, China, in 2005. Currently, he is with the Bio-Computing Research Center, Shenzhen Graduate School, Harbin Institute of Technology, Shenzhen, China. His current research interests include pattern recognition, biometrics, machine learning, image processing, and video analysis.

Heng Huang received both B.S. and M.S. degrees from Shanghai Jiao Tong University, Shanghai, China, in 1997 and 2001, respectively. He received the Ph.D. degree in Computer Science from Dartmouth College in 2006. Since 2017, he joined University of Pittsburgh as John A. Jurenko Endowed Full Professor in Computer Engineering. His research interests include machine learning, data mining, bioinformatics, neuroinformatics, health informatics.