2015 IEEE 15th International Conference on Data Mining Workshops

Matrix Factorization with Column L_0-norm Constraint for Robust Multi-subspace Analysis

Binghui Wang[1], Risheng Liu[1], Chuang Lin[2], and Xin Fan[1,⋆]
[1]School of Software, Dalian University of Technology, China
[2] Department of Neurorehabilitation Engineering, Bernstein Focus Neurotechnology Goettingen,
Bernstein Center for Computational Neuroscience, University Medical Center Goettingen, Georg-August University, Germany
⋆ Email:xin.fan@ieee.org

***Abstract*—We aim to study the subspace structure of data approximately generated from multiple categories and remove errors (e.g., noise, corruptions, and outliers) in the data as well. Most previous methods for subspace analysis learn only one subspace, failing to discover the intrinsic complex structure, while state-of-the-art methods use data itself as the basis (self-expressiveness property), showing degraded performance when data contain errors. To tackle the problem, we propose a novel method, called Matrix Factorization with Column L_0-norm constraint (MFC$_0$), from the matrix factorization perspective. MFC$_0$ *simultaneously* discovers the multi-subspace structure of either clean or contaminated data, and learns the basis for each subspace. Specifically, the learnt basis with the orthonormal constraint shows high robustness to errors by adding a regularization term. Owing to the column l_0-norm constraint, the generated representation matrix can be (approximate) block-diagonal after reordering its columns, with each block characterizing one subspace. We develop an efficient first-order optimization scheme to stably solve the nonconvex and nonsmooth objective function of MFC$_0$. Experimental results on synthetic data and real-world face datasets demonstrate the superiority over traditional and state-of-the-art methods on both representation learning, subspace recovery and clustering.**

I. Introduction

The observation data are extremely high dimensional in this 'big data' era. Typically, these data reside in a much lower-dimensional latent structure, instead of being uniformly distributed in the high-dimensional observation space. Thus, it is of great importance to reveal the underlying structure of the data as it helps to reduce the computational cost and enables a compact representation for learning.

Subspace methods have been widely used to analyze the data, and *linear subspace* is the most common choice for its simplicity and computational efficiency. Additionally, linear subspace has shown its effectiveness in modeling real-world problems such as motion segmentation [1], [2], face clustering [3], [4], and handwritten digits recognition [5]. Consequently, subspace analysis has been paid much attention in the past decade. For instance, principal component analysis (PCA) aims to learn a subspace while retaining maximal variances of the data. Nonnegative matrix factorization (NMF) [6] is designed to learn both nonnegative basis and nonnegative parts-based representation. Robust PCA (RPCA) [7] assumes that the data are approximately drawn from a low-rank subspace while perturbed by sparse noise. The basic assumption for these methods is the *single* subspace, which is not the case in many practical applications. A more reasonable way is to consider the data as lying near a union of linear subspaces. Unfortunately, the generalization to multiple subspaces is quite challenging.

Recently, multi-subspace analysis has attracted increasingly interests in visual data analysis [8], [9], [10]. Derived from recent advances in compressive sensing [11], [12], the methods including [13], [14], [15], [16], [17] have incorporated sparse and/or low-rank regularization into their formulations to model the mixture of linear structures for clean data[1] as well as dealing with errors in data, e.g., noise [18], missed entries [11], corruptions [7], and outliers [19]. Among them, sparse subspace clustering (SSC) [13] and low-rank representation (LRR) [14] are two pioneers, which formulate the discovery of multi-subspace structure by finding a sparse or low-rank representation of the data from a predefined self-expressive dictionary. It has been shown in literature that these methods can achieve more accurate subspace structure than traditional subspace analysis methods.

However, despite of their satisfactory results, these methods have two major drawbacks. Firstly, the original data are chosen to be the dictionary (basis), which cannot deliver the flexibility to the data. It is also questionable that the data containing errors are used as the basis for error correction. Secondly, neither sparse representation via an indirect l_1 penalty nor low-rank representation in a global constraint provides a direct description for individual data sample. That is, data samples cannot find the exact subspace where they lie, and thus their errors are unable to be removed.

In this paper, we focus on analyzing multi-subspace structure from the *matrix factorization* perspective, and propose a novel method, called Matrix Factorization with Column L_0-norm constraint (MFC$_0$), for robust multi-subspace analysis. Given a collection of data, we assume that they are generated from a union of independent subspaces, and all subspaces have an identical dimension. Our objective is to *simultaneously* learn the basis and representation matrix to discover the multi-subspace structure of the data. The basis is learnt with the orthonormal constraint, and has a strong ability to resist errors when adding a regularization term. By enforcing a column l_0-norm constraint, we are able to generate the (approximate) block-diagonal representation matrix, when the data samples are clean (contain errors). We also develop an efficient alternating direction type algorithm to handle the nonconvex and nonsmooth objective function. We highlight the contributions

[1]Data points are strictly sampled from the respective subspace

978-1-4673-8493-3/15 $31.00 © 2015 IEEE
DOI 10.1109/ICDMW.2015.80

of the proposed approach below:

- We propose an effective method to discover the structure of multi-subspace from the matrix factorization perspective. We also develop an efficient first-order optimization scheme to stably solve the nonconvex and nonsmooth objective function.
- MFC_0 learns an adaptive orthonormal basis from the data instead of using the data itself as basis. Moreover, the regularizer significantly enhances the robustness of MFC_0 to various errors, e.g., noise, random corruptions, and sample-specific outliers.
- Different from previous methods that use l_1-norm as an inexact surrogate to impose sparsity, MFC_0 directly takes advantage of a column l_0-norm constraint on the representation matrix, which has an exact description for each data sample and avoids tuning hyperparameters.

A. Main Notations

In this paper, vectors and matrices are written as bold lowercase and uppercase symbols. The i-th entry of vector $\mathbf{u}$ is u_i. For matrix $\mathbf{W}$, its (i,j)-th entry, i-th column, and j-th row are denoted as $w_{i,j}$, $\mathbf{w}_i$, and $\mathbf{w}^j$ respectively. The superscript T stands for the transpose. For vector $\mathbf{u} \in \Re^m$, the l_p-norm is $\|\mathbf{u}\|_p = (\sum_{i=1}^m |u_i|^p)^{\frac{1}{p}}$, and the pseudo l_0-norm is the number of nonzero entries. For matrix $\mathbf{W} \in \Re^{m\times n}$, the Frobenius norm is defined as $\|\mathbf{W}\|_F = \sqrt{\sum_{j=1}^m \|\mathbf{w}^j\|_2^2}$. The $l_{2,1}$-norm is $\|\mathbf{W}\|_{2,1} = \sum_{i=1}^n \|\mathbf{w}_i\|_2$, and l_1-norm is $\|\mathbf{W}\|_1 = \sum_{j=1}^m \sum_{i=1}^n |w_{ij}|$. The inner product between two matrix is $\langle \mathbf{W}, \mathbf{U}\rangle = \mathbf{tr}(\mathbf{W}^T\mathbf{U})$, where $\mathbf{tr}$ is the trace operator.

II. The Proposed Method

In this section, we first review the basic concept of matrix factorization. Then we introduce our method for analyzing a collection of data generated from multiple subspaces.

A. Matrix Factorization

Given the observed data $\mathbf{Z} = \{\mathbf{z}_i\}_{i=1}^n \in \Re^{m\times n}$, the goal of matrix factorization is to decompose the data matrix into a product of matrices with tolerated errors. In the paper, we mainly consider two matrices $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^d \in \Re^{m\times d}$ and $\mathbf{Y} = \{\mathbf{y}_i\}_{i=1}^n \in \Re^{d\times n}$ such that

$$\min_{\mathbf{X},\mathbf{Y}} \|\mathbf{Z} - \mathbf{X}\mathbf{Y}\|_F^2, \tag{1}$$

where $\mathbf{X}$ and $\mathbf{Y}$ are the basis and representation matrix, respectively. Note that traditional matrix factorization methods, e.g., PCA and NMF [6], only focus on single subspace analysis. The generalization to deal with data generated from multiple subspaces motivates our proposed method.

B. Problem Formulation

Suppose the data $\mathbf{Z} = [\mathbf{Z}_1, \cdots, \mathbf{Z}_K]$ are drawn from a union of K independent subspaces $\{\mathcal{S}_k\}_{k=1}^K$ with an equal subspace dimension d_0, and $\mathbf{Z}_k \in \Re^{m\times n_k}$, $\sum_{k=1}^K n_k = n$. To begin with, we first write the matrix decomposition in the following form:

$$[\mathbf{Z}_1, \cdots, \mathbf{Z}_K] = [\mathbf{X}_1, \cdots, \mathbf{X}_K] \begin{bmatrix} \mathbf{Y}_1 \\ \vdots \\ \mathbf{Y}_K \end{bmatrix}, \tag{2}$$

where $\mathbf{X}$ is written as $\mathbf{X} = [\mathbf{X}_1, \cdots, \mathbf{X}_K]$ with $\mathbf{X}_k \in \Re^{m\times d_0}$, $\mathbf{Y} = [\mathbf{Y}_1; \cdots; \mathbf{Y}_K]$ with $\mathbf{Y}_k \in \Re^{d_0\times n}$, and $d = K \cdot d_0$.

Seeing Eq.(2), to analyze the structure of multiple subspaces, we expect to satisfy the following requirements: First, we impose an orthonormal constraint on the basis matrix to ensure that they are indeed the "basis". Moreover, orthogonality provides a way in dealing with insufficient data [20]; Second, we characterize each data as a convex combination of only the basis that span its underlying subspace. To achieve this, we apply a l_0-norm constraint to each column of the representation matrix to set that the number of nonzero coefficients equals to the subspace dimension. Finally, we use a regularization term to resist different types of errors contained in the data.

Mathematically, we define the objective function of our method that handles noise or corruptions as

$$\min_{\mathbf{X}\subset\mathcal{B}^m,\mathbf{Y}} \|\mathbf{Z} - \mathbf{X}\mathbf{Y}\|_F^2, \quad s.t. \quad \mathbf{Y} \geq 0,\ \|\mathbf{y}_i\|_0 = d_0, \forall i, \tag{3}$$

where $\mathcal{B}^m = \{\mathbf{x}_i \in \Re^m \big| \mathbf{X}^T\mathbf{X} = \mathbf{I}_d\}$ is the orthonormal constraint for the basis, where $\mathbf{I}_d$ is a $d \times d$ identity matrix. $\mathbf{Y} \geq 0$ indicates a convex combination, and $\|\mathbf{y}_i\|_0 = d_0$ is the requirement for the number of nonzero coefficients. For writing simplicity, we omit the "$\forall i$" symbol in the following associated equations.

Further, we consider the general case dealing with sample-specific outliers $\|\mathbf{E}\|_{2,1}$ [2]. The objective function becomes

$$\begin{aligned} \min_{\mathbf{X}\subset\mathcal{B}^m,\mathbf{Y},\mathbf{E}} & \|\mathbf{Z} - \mathbf{X}\mathbf{Y} - \mathbf{E}\|_F^2 + \lambda \|\mathbf{E}\|_{2,1}, \\ s.t. \quad & \mathbf{Y} \geq 0,\ \|\mathbf{y}_i\|_0 = d_0. \end{aligned} \tag{4}$$

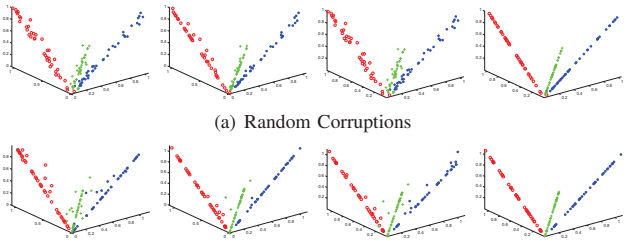
With above constraints, we point out that, when the data are clean, i.e., perfectly generated from K independent subspaces, $\mathbf{Y}$ is block-diagonal after reordering its columns, and $\mathbf{X}_k$ is the basis that characterizes data only from the k-th subspace.

When the data are contaminated with errors, e.g., random corruptions (a fraction of random entries are grossly corrupted) and sample-specific outliers (a fraction of data are far away from the subspaces), our experimental results show that $\mathbf{X}$ is still able to discover the multi-subspace structure and $\mathbf{Y}$ is approximate block-diagonal.

It is of great importance to note that state-of-the-art SSC and LRR can also be boiled down to the matrix factorization framework. They share some similarities with the proposed model, but there exist two major differences: On one hand, they take advantage of what they refer to as "self-expressiveness" property and use the data itself to form the basis. However, this pregiven basis cannot deliver the flexibility to the data. What's worse, they fail to perform data reconstruction when the data contain errors to some extent (See Figure 1). On the other hand, they respectively utilize a l_1-norm and nuclear norm to obtain

[2]For sparse outliers, one can replace $\|\mathbf{E}\|_{2,1}$ by $\|\mathbf{E}\|_1$ [16].

the sparse and low-rank representation of the data. However, they do not provide the exact basis for representing each data, and thus their ability for robust subspace clustering is limited (See Figure 5).



(a) Random Corruptions

(b) Sample-Specific Outliers

Fig. 1. Reconstruction results of data generated from 3 independent subspaces and contaminated with (a) random corruptions and (b) sample-specific outliers. From left to right are contaminated data and results of SSC, LRR, and MFC$_0$.

C. Problem Optimization

In this part, without loss of generality, we just solve for the general case in Eq.(4). Notice that we need to handle the nonconvex objective function, as well as the nonsmooth l_0-norm constraint. For this purpose, we propose an efficient first-order alternating direction type algorithm. The overall procedure of the algorithm is first introducing auxiliary variables and quadratic penalties into the objective function, then iteratively minimizing the augmented Lagrangian function with respect to each primal variable, and finally updating the multipliers.

To begin with, we introduce an auxiliary variable $\mathbf{V}$ to distinguish the same variable existing in both the object function and constraint. I.e., we reformulate Eq.(4) as

$$\begin{aligned} &\min_{\mathbf{X}\subset\mathcal{B}^m,\mathbf{Y},\mathbf{E},\mathbf{V}} \|\mathbf{Z}-\mathbf{XY}-\mathbf{E}\|_F^2+\lambda\|\mathbf{E}\|_{2,1}, \\ &s.t. \quad \mathbf{Y}=\mathbf{V},\mathbf{V}\geq 0, \|\mathbf{v}_i\|_0=d_0. \end{aligned} \tag{5}$$

Then, the augmented Lagrangian function of Eq.(5) is

$$\begin{aligned} \mathcal{L}_{\mathcal{A}}(\mathbf{X},\mathbf{Y},\mathbf{V},\mathbf{E},\mathbf{P}) &= \|\mathbf{Z}-\mathbf{XY}-\mathbf{E}\|_F^2+\lambda\|\mathbf{E}\|_{2,1} \\ &+\langle\mathbf{P},\mathbf{Y}-\mathbf{V}\rangle+\frac{\beta}{2}\|\mathbf{Y}-\mathbf{V}\|_F^2, \\ s.t. \quad \mathbf{V}\geq 0,\ &\|\mathbf{v}_i\|_0=d_0, \end{aligned} \tag{6}$$

where $\mathbf{P}$ is the Lagrangian multiplier, $\beta>0$ is the quadratic penalty parameter. Note that adding the penalty term does not change the optimal solution, since any feasible solution satisfying the constraint in Eq.(6) vanishes the penalty term.

Finally, our alternating direction algorithm consists of iteratively minimizing Eq.(6) with respect to one of $\mathbf{X},\mathbf{Y},\mathbf{E},\mathbf{V}$ while fixing the others, and updating the multiplier $\mathbf{P}$.

Update X: By discarding terms that are irrelevant to $\mathbf{X}$, we rewrite the subproblem with respect to $\mathbf{X}$ as

$$\min_{\mathbf{X}} \|\mathbf{Z}-\mathbf{XY}-\mathbf{E}\|_F^2, \quad s.t. \quad \mathbf{X}^\mathbf{T}\mathbf{X}=\mathbf{I}_d. \tag{7}$$

If the singular value decomposition (SVD) of $(\mathbf{Z}-\mathbf{E})\mathbf{Y}^\mathbf{T}=\mathbf{U}\mathbf{\Sigma}\mathbf{V}^\mathbf{T}$ [3], then from **Theorem 1** given below, we can achieve the closed-form solution of $\mathbf{X}$. That is,

$$\mathbf{X}=\mathbf{U}\mathbf{V}^\mathbf{T}. \tag{8}$$

Theorem 1. *Let $\mathbf{A}\in\Re^{m\times n}$ and $\mathbf{B}\in\Re^{d\times n}$ be any two matrices. Consider the orthonormal constrained minimization problem*

$$\min_{\mathbf{D}} \|\mathbf{A}-\mathbf{DB}\|_F^2 \quad s.t. \quad \mathbf{D}^\mathbf{T}\mathbf{D}=\mathbf{I}_d.$$

If the SVD of $\mathbf{A}\mathbf{B}^\mathbf{T}=\mathbf{U}\mathbf{\Sigma}\mathbf{V}^\mathbf{T}$, then we have $\mathbf{D}=\mathbf{U}\mathbf{V}^\mathbf{T}$.

Proof: We first unfold the objective function:

$$\|\mathbf{A}-\mathbf{DB}\|_F^2=\mathbf{tr}(\mathbf{A}^\mathbf{T}\mathbf{A})-2\mathbf{tr}(\mathbf{A}\mathbf{B}^\mathbf{T}\mathbf{D}^\mathbf{T})+\mathbf{tr}(\mathbf{B}^\mathbf{T}\mathbf{D}^\mathbf{T}\mathbf{D}\mathbf{B}). \tag{9}$$

Using the constraint $\mathbf{D}^\mathbf{T}\mathbf{D}=\mathbf{I}$, and the SVD of $\mathbf{A}\mathbf{B}^\mathbf{T}=\mathbf{U}\mathbf{\Sigma}\mathbf{V}^\mathbf{T}$, then the minimization of Eq.(9) is transformed to the maximization problem

$$\mathbf{tr}(\mathbf{A}\mathbf{B}^\mathbf{T}\mathbf{D}^\mathbf{T})=\mathbf{tr}(\mathbf{U}\mathbf{\Sigma}\mathbf{V}^\mathbf{T}\mathbf{D}^\mathbf{T})=\mathbf{tr}(\mathbf{\Sigma}(\mathbf{D}\mathbf{V})^\mathbf{T}\mathbf{U}). \tag{10}$$

Note that $\|(\mathbf{DV})^\mathbf{T}\mathbf{U}\|_F^2=\mathbf{tr}((\mathbf{DV})^\mathbf{T}\mathbf{U}\mathbf{U}^\mathbf{T}(\mathbf{DV}))=\mathbf{tr}(\mathbf{I}_d)$ is a constant. Since $\mathbf{\Sigma}$ is diagonal, so the maximization of Eq.(10) is acquired when $(\mathbf{DV})^\mathbf{T}\mathbf{U}$ is a diagonal matrix, with positive diagonal elements. Therefore, we have $\mathbf{DV}=\mathbf{U}$, and thus $\mathbf{D}=\mathbf{U}\mathbf{V}^\mathbf{T}$.

Update Y: The subproblem of $\mathbf{Y}$ becomes

$$\min_{\mathbf{Y}} \|\mathbf{Z}-\mathbf{XY}-\mathbf{E}\|_F^2+\langle\mathbf{P},\mathbf{Y}-\mathbf{V}\rangle+\frac{\beta}{2}\|\mathbf{Y}-\mathbf{V}\|_F^2. \tag{11}$$

Taking derivative with respect to $\mathbf{Y}$ and setting it to zero, we have the closed-form solution

$$\mathbf{Y}=\frac{1}{1+\beta}\big(\mathbf{X}^T(\mathbf{Z}-\mathbf{E})+\beta\mathbf{V}-\mathbf{P}\big), \tag{12}$$

Update E: The subproblem of $\mathbf{E}$ is

$$\min_{\mathbf{E}} \|\mathbf{E}-(\mathbf{Z}-\mathbf{XY})\|_F^2+\lambda\|\mathbf{E}\|_{2,1}. \tag{13}$$

Eq.(13) can be seen as the proximal operator for $l_{2,1}$-norm. So, the update of $\mathbf{E}$ is easily obtained in the closed-form [14]. Denote $\mathbf{G}=\mathbf{Z}-\mathbf{XY}$, then

$$\mathbf{e}_i=\begin{cases}(1-\frac{\lambda/2}{\|\mathbf{g}_i\|_2})\mathbf{g}_i & \|\mathbf{g}_i\|_2\geq\frac{\lambda}{2},\\ 0 & \|\mathbf{g}_i\|_2<\frac{\lambda}{2}.\end{cases} \tag{14}$$

Update V: The objective function regrading $\mathbf{V}$ is

$$\min \big\|\mathbf{V}-(\mathbf{Y}+\beta^{-1}\mathbf{P})\big\|_F^2, \quad s.t \quad \|\mathbf{v}_i\|_0=d_0, \mathbf{v}_i\geq 0. \tag{15}$$

where "$\geq$" is taken component-wise.

Due to the discrete constraint of l_0-norm, we cannot achieve the solution using any (sub)gradient methods. Here, we develop a very simple yet efficient method and solve it column by column. We denote $\mathbf{U}=(\mathbf{Y}+\frac{\mathbf{1}}{\beta}\mathbf{P})$. For any $\mathbf{v}_i$, we define the operator $\mathrm{P}_{d_0}:\Re^d\to\Re^d$ by

$$\mathrm{P}_{d_0}(\mathbf{u}_i)=\arg\min_{\mathbf{v}_i}\{||\mathbf{u}_i-\mathbf{v}_i||_2^2:||\mathbf{v}_i||_0=d_0\}. \tag{16}$$

[3] Here, SVD is in the form that Σ is a diagonal matrix.

Moreover, we define the nonnegative orthant mapping by

$$\mathrm{P}_+(\mathbf{u}_i) = \arg\min_{\mathbf{v}_i} \{||\mathbf{u}_i - \mathbf{v}_i||_2^2 : \mathbf{v}_i \geq 0\} = \max\{0, \mathbf{u}_i\}. \tag{17}$$

Denote the indicator function $I_{\mathcal{V}}(\mathbf{v}_i)$, where the set $\mathcal{V} = I(\mathbf{v} \in \Re^d : \mathbf{v} \geq 0, ||\mathbf{v}||_0 = d_0)$, then we have the following proximal operator

$$\begin{aligned} \mathrm{prox}_{\frac{1}{2}}^{I_{\mathcal{V}}}(\mathbf{u}_i) &= \arg\min_{\mathbf{v}_i} \{||\mathbf{u}_i - \mathbf{v}_i||_2^2 : \mathbf{v}_i \geq 0, ||\mathbf{v}_i||_0 = d_0\} \\ &= \mathrm{P}_{d_0}(\mathrm{P}_+(\mathbf{u}_i)). \end{aligned} \tag{18}$$

Above proximal operator can be solved by selecting d_0 largest nonnegative entries of $\mathbf{u}_i$ with corresponding indexes $\mathbf{q}_j = [q_{1,j}, \cdots, q_{d_0,j}]$. To be specific, for $i = 1, \cdots, d$, we set

$$v_{i,j} = \begin{cases} u_{i,j} & i \in \mathbf{q}_j, \\ 0 & i \notin \mathbf{q}_j. \end{cases} \tag{19}$$

Update P: Finally, for the multiplier $\mathbf{P}$, based on the dual optimal condition [21], its updating rule is

$$\mathbf{P} := \mathbf{P} + \mu(\mathbf{Y} - \mathbf{V}), \tag{20}$$

where $\mu = \min(\rho\mu, \mu_{max})$, with pregiven ρ and μ_{max}.

Algorithm 1 shows the algorithmic procedure for our alternating direction scheme of the optimization program Eq.(6). For subspace clustering, we apply Normalized Cut [22] to $\mathbf{Y}^T\mathbf{Y}$ after obtaining the representation $\mathbf{Y}$ by solving Eq.(6).

Algorithm 1 Solving the problem Eq.(6)

Input:

Data matrix $\mathbf{Z}$, subspace dimension d_0.

Initialize:

Randomize $\mathbf{X}^{(0)}, \mathbf{E}^{(0)} = 0$, $\mathbf{V}^{(0)} = 0$, $\mathbf{P}^{(0)} = 0$, $\mu^{(0)} = 10^{-3}$; $\rho = 1.2$, $\mu_{max} = 10^3$, $\epsilon = 10^6$.

Process:

1: **while** not converged **do**

2: Update $\mathbf{X}, \mathbf{Y}, \mathbf{E}, \mathbf{V}, \mathbf{P}$ using corresponding equations.

3: Check the convergence conditions:

$$\|\mathbf{Z} - \mathbf{X}\mathbf{Y}\|_\infty \leq \epsilon \text{ and } \|\mathbf{Y} - \mathbf{V}\|_\infty \leq \epsilon.$$

4: **end while**

Output: $\mathbf{X}, \mathbf{Y}$

III. Experimental Results

In this section, we carry out different types of experiments on both synthetic data and real-world datasets to test the performance of the proposed MFC_0. On synthetic data, we test MFC_0's ability to discover the block-diagonal structure of the representation matrix and its stability to obtain the local solution. On real-world face datasets, we pay close attention to recovering face subspace structure of multiple subjects and dealing with face clustering. Experimental configurations are described in the corresponding parts. For MFC_0, we randomly generate a matrix with size $m \times Kd_0$ to initialize the basis matrix. The hyperparameters of all comparable methods are selected via cross-validation. The stopping criterion is that either the method reaches the maximal iteration 1000 or the difference between neighboring iterations is less than 10^{-6}. All experiments are conducted in Matlab R2010b with the platform Intel(R) Core(TM) i3-2100, CPU 3.10 GHz, and RAM 16.0 GB.

A. Synthetic Data

We first visualize data reconstruction results in the 3-dim ambient space in Figure 1. The data samples are drawn from three 1-dim independent subspaces, and disturbed by random corruptions ($||\mathbf{E}||_1$) and sample-specific outliers ($||\mathbf{E}||_{2,1}$). We compare MFC_0 with SSC [13] and LRR [14]. It can be seen from Figure.3 that SSC and LRR are unable to remove the errors, while MFC_0 can fully eliminate them. The reason is that both SSC and LRR use the original noisy data as the basis, which is indeed problematic for reconstruction. In contrast, for MFC_0, it tries to learn the basis that span the underlying subspace—the clean data lies in, the noisy absorbed in the error regularization term.

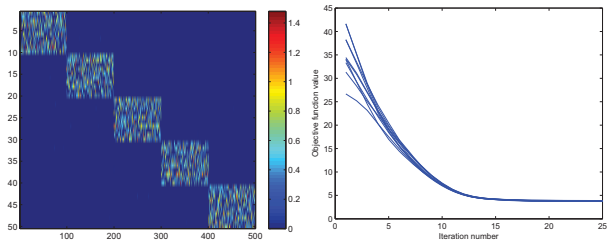
For high-dimensional data analysis, the synthetic data are generated from $K = 5$ independent subspaces with each containing $n_k = 100$ samples. All subspaces have the same dimension $d_0 = 10$ embedded in a $D = 100$ dimensional ambient space. The procedure of generating above subspaces is similar to that of [5]: the basis $\mathbf{U}_k$ of each subspace are calculated by $\mathbf{U}_{k+1} = \mathbf{T}\mathbf{U}_k, 1 \leq k \leq 4$, where $\mathbf{T} = \mathrm{orth}(\mathrm{rand}(\mathrm{D})) \in \Re^{D\times D}$ is a random orthonormal matrix and $\mathbf{U}_1 \in \Re^{D\times d_0}$ is a random column orthogonal matrix. The data samples from each subspace are sampled by $\mathbf{X}_k = \mathbf{U}_k\mathbf{R}_k$, with $\mathbf{R}_k \in \Re^{d_0\times n_k}$ with random distribution.

We plot the representation matrix for clean data in Figure $2(\mathrm{a})$. Consider that the objective function of MFC_0 in Eq.(5) is nonconvex with discrete l_0-norm constraint, we check whether our first-order optimization algorithm is stable. The objective function values versus iteration number for 10 times is shown in Figure $2(\mathrm{b})$. We can see from Figure 2 that, MFC_0 can precisely discover the block-diagonal structure of the representation matrix when the data are clean. Moreover, it can converge to a very stable value although using different random initializations.

We further consider the cases that the data are contaminated with errors. For random corruptions and sample-specific outliers with ratio $=$ 0.1, 0.3, and 0.5, the corresponding representation matrices are plotted in Figure 3 and Figure 4, respectively. We observe that even the data are contaminated with different types of errors to a large extent, the representation matrix learnt via MFC_0 is still close to be block-diagonal.

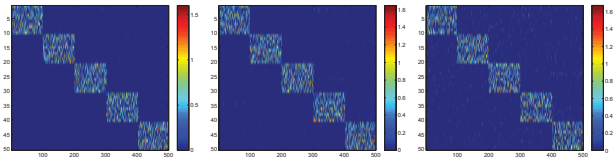
Finally, we compare our MFC_0 with SSC and LRR for subspace clustering. The performance versus different ratios, ranging from 0 to 0.8, of random corruptions and sample-specific outliers in the data is depicted in Figure 5. It can be seen that the clustering accuracy of MFC_0 is higher than that of LRR and SSC in most cases. Surprisingly, even when the ratio is 0.6, MFC_0 can still obtain $> 95\%$ accuracy in both cases. Whereas, for LRR and SSC, the highest accuracies are nearly 70% and 85%, respectively.

To sum up, from above results, we can safely come to the conclusion that our MFC_0 method has a strong ability to resist errors in the data and is more robust than LRR and SSC.



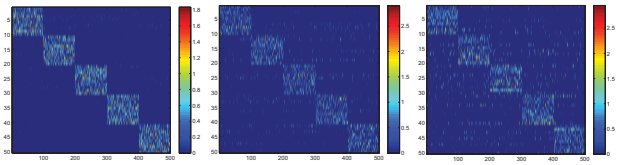
(a) (b)

Fig. 2. (a) Representation matrix learnt by clean data. (b) Objective function values vs. iteration number.



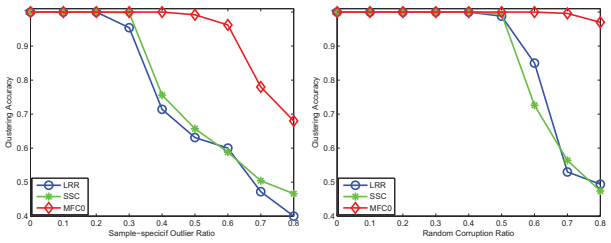
(a) ratio=0.1 (b) ratio=0.3 (c) ratio=0.5

Fig. 3. Representation matrix learnt by different ratios of random corruptions in the data.



(a) ratio=0.1 (b) ratio=0.3 (c) ratio=0.5

Fig. 4. Representation matrix learnt by different ratios of sample-specific outliers in the data.



(a) (b)

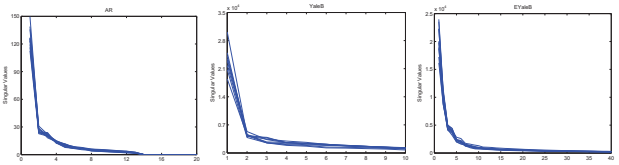
Fig. 5. Clustering accuracy vs. ratio of (a) random corruptions and (b) sample-specific outliers in the data.

B. Real-World Data

In this part, we evaluate the performance of our MFC_0 method in dealing with two tasks: face subspace recovery and face clustering. Experiments are carried out on the AR [4], Yale [5], and Extended Yale B (EYaleB) [6] [23] face datasets.

AR contains over 3,000 images of 126 individuals. They are taken at two different occasions with different facial expressions, illumination conditions, and occlusions (sun glasses and scarf). In the experiment, images are resized to 48x48 pixels. Yale contains 165 grayscale images of 15 individuals. These images are taken under different lighting conditions, facial expressions, and with/without glasses. We simply use the cropped images and resize them to 32x32 pixels. EYaleB consists of 2414 images of 38 subjects under 9 poses and 64 illumination conditions. They are captured under various lighting conditions and cropped to 32x32 pixels.

[4]http://www2.ece.ohio-state.edu/ aleix/ARdatabase.html
[5]http://vision.ucsd.edu/content/yale-face-database
[6]http://vision.ucsd.edu/ leekc/ExtYaleDatabase/ExtYaleB.html



(a) (b) (c)

Fig. 6. Singular values of several random chosen subjects in (a) AR, (b) Yale, and (c) EYaleB.

For MFC_0, we need to know the subspace dimension of each subject in advance. A simple yet efficient way is to calculate the number of nonzero singular values of each subject. Figure 6 plots the singular values of several subjects in the three datasets. The subspace dimensions d_0 for AR, Yale, and EYaleB are approximate 12, 10, and 10, respectively. Note that for real-world datasets, the singular values often gradually close to zero, validating that face images are contaminated with errors. The higher dimension of AR reflects its more complicated face distribution than Yale and EYaleB.

1) Face Basis Learning and Subspace Recovery: Subspace learning (SL) aims to learn the representative basis that provides a compact presentation for the data. For face subspace, traditional SL methods assume that data samples of all subjects can be represented by a single lower-dimensional subspace. However, previous results [4] have proven that, under the Lambertian assumption, face images of a subject with a fixed pose and varying illumination lie close to a linear subspace. Thus for multiple subjects, we need to learn each subject its underlying face subspace with the corresponding basis. We testify that our method is capable of correcting errors and learning the multiple face subspaces simultaneously (Note that SSC and LRR are unable to learn the basis). For comparison, we choose single subspace based PCA. For PCA, we preserve 95 percent of total variance.

We randomly chose three subjects (subject 1, 2, and 55, each 15 samples) in AR to learn the basis, which is then used for face reconstruction. Some learnt typical basis of three subjects in AR, as well as some examples of corrected faces and reconstruction errors are plotted in Figure 7. From it we can see that: (1) The learnt basis via MFC_0 can be separated into three parts, i.e., basis in each of the three rows are used for representing only the corresponding subject. In this sense, face images of each subject lie on its own underlying subspace. However, for PCA, the learnt basis are mixed together. Each basis contains information of face images across the three subjects; (2) MFC_0 shows its effectiveness in face correction with gross corruptions (sunglasses). The reconstructed face has a very clear contour. While for PCA,

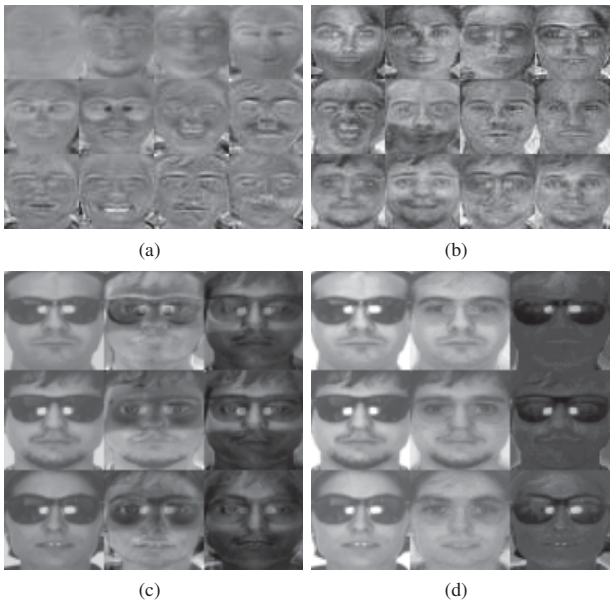


Fig. 7. Some examples of learnt face basis, corrected faces, and reconstruction errors of three subjects in AR. (a) and (b) are some representative face basis obtained by PCA and MFC_0. (c) and (d) are some corrected faces and reconstruction errors obtained by PCA and MFC_0.

it is blurry and untidy; (3) For MFC_0, the remaining error contains useful information, e.g., the sunglasses. Interestingly, these errors contain discriminant features and can be used for object recognition [3]. For instance, we can utilize errors to detect whether one wearing a sunglass. Nevertheless, the errors obtained by PCA cannot provide any useful information.

2) Face Clustering: In this part, we focus on the face clustering task: Given face images of multiple subjects, we group them according to their respective subject. We compare MFC_0 with K-Means, PCA, NMF, and state-of-the art SSC and LRR in terms of clustering performance. For PCA, we preserve 95 percent of total variance; For NMF, we set their basis number equalling to MFC_0's, i.e., $K \times d_0$. To study the effect of different number of subjects K, i.e., subspaces, we set K in an ascending order. For Yale, K is ranging from 2 to 11; For EYaleB, it is from 2 to 20 with an interval 2.

Table I and II show the clustering performance on the Yale and EYaleB, respectively. We have the following observations: (1) In both cases, our MFC_0 performs the best. The reason is that, by automatically learning each face subspace for the respective subject, MFC_0 is discriminative to separate faces from different subjects. Actually, this is the key point that makes it outperform single subspace methods. By adding the regularization term, MFC_0 is robust to errors contained in the faces. Moreover, using a column l_0-norm constraint to directly control the sparsity, MFC_0 shows its superiority to the indirect l_1 penalty and avoids the burden of hyperparameter tuning; (2) SSC and LRR are multi-subspace learning methods with error correction, so they can achieve promising results only next to MFC_0. However, the irrationality is that they use original faces as the basis, which is questionable since the faces are contaminated with errors. The accuracy gap between SSC, LRR, and MFC_0 validates this argument; (3) K-Means, PCA, and NMF can be recast into single subspace learning methods. When the sample size of each subject is small in Yale, the multi-subspace structure may be not so clear and the accuracy gap between MFC_0 with them is not remarkable. However, all of them perform poorly with modest sample size in EYaleB. Even worse, they fail to work when the number of subjects K increases to a certain value. While for MFC_0, it still performances very well.

TABLE I. CLUSTERING PERFORMANCE ON YALE

K	K-Means	PCA	NMF	SSC	LRR	MFC_0
2	0.76	0.76	0.82	0.87	**0.95**	**0.95**
3	0.61	0.61	0.70	0.64	0.68	**0.74**
4	0.59	0.63	0.64	0.61	0.64	**0.76**
5	0.55	0.53	0.64	0.67	0.67	**0.69**
6	0.53	0.56	0.53	0.65	0.58	**0.68**
7	0.55	0.53	0.55	0.62	0.57	**0.61**
8	0.48	0.55	0.49	0.60	**0.63**	0.60
9	0.46	0.43	0.38	0.55	0.57	**0.58**
10	0.47	0.45	0.47	**0.54**	**0.54**	**0.54**
11	0.47	0.54	0.49	0.57	0.57	**0.59**
Avg.	0.55	0.56	0.57	0.63	0.64	**0.66**

TABLE II. CLUSTERING PERFORMANCE ON EYALEB

K	K-Means	PCA	NMF	SSC	LRR	MFC_0
2	0.51	0.51	0.80	0.98	**1.00**	**1.00**
4	0.35	0.35	0.59	0.88	0.86	**0.89**
6	0.30	0.30	0.72	0.84	0.82	**0.86**
8	0.25	0.20	0.37	0.76	0.74	**0.76**
10	0.23	0.23	0.45	0.74	0.71	**0.75**
12	0.19	0.22	0.37	0.67	**0.70**	**0.70**
14	0.20	0.18	0.36	0.64	0.68	**0.72**
16	0.17	0.18	0.37	0.63	0.67	**0.73**
18	0.14	0.15	0.35	0.63	0.68	**0.74**
20	0.13	0.15	0.37	0.60	0.66	**0.72**
Avg.	0.25	0.25	0.47	0.74	0.75	**0.79**

IV. CONCLUSIONS

In this paper, we propose a novel multi-subspace learning method, called Matrix Factorization with Column l_0 constraint (MFC_0), for robustly analyzing the structure of data approximately generated from multiple categories. By automatically learning the basis matrix with an orthonormal constraint, MFC_0 is able to discover the mixture subspace structure and is robust to different types of errors. Moreover, MFC_0 directly imposes a column l_0-norm constraint on the representation matrix, which achieves a (or approximate) block-diagonal structure when the data samples are clean (or contain errors). We propose an alternating direction type algorithm to stably solve the nonconvex and nonsmooth optimization program of MFC_0. Experimental results verify that MFC_0 can achieve better clustering performance against traditional and state-of-the-art methods on both representation learning and multi-subspace recovery/clustering tasks.

Acknowledgements: This work is supported by the Natural Science Foundation of China under Grant Nos. 61272371, 61300086, 61432003, and U0935004, Fundamental Research Funds for the Central Universities, No.DUT14QY16, and the Open Project Program of the State Key Laboratory of CAD&CG, Zhejiang University, China.

REFERENCES

[1] S. Rao, R. Tron, R. Vidal, and Y. Ma, "Motion segmentation in the presence of outlying, incomplete, or corrupted trajectories," *IEEE TPAMI*, vol. 32, no. 10, pp. 1832–1845, 2010.

[2] J. Yan and M. Pollefeys, "A general framework for motion segmentation: Independent, articulated, rigid, non-rigid, degenerate and non-degenerate," in *ECCV*. Springer, 2006, pp. 94–106.

[3] G. Liu and S. Yan, "Latent low-rank representation for subspace segmentation and feature extraction," in *ICCV*, 2011, pp. 1615–1622.

[4] R. Basri and D. Jacobs, "Lambertian reflectance and linear subspaces," *IEEE TPAMI*, vol. 25, no. 2, pp. 218–233, 2003.

[5] K. Tang, R. Liu, Z. Su, and J. Zhang, "Structure-constrained low-rank representation," *IEEE TNNLS*, vol. 25, no. 12, pp. 2167–2179, 2014.

[6] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, 1999.

[7] E. J. Candès, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis?" *J. ACM*, vol. 58, no. 3, p. 11, 2011.

[8] R. Liu, Z. Lin, and Z. Su, "Learning markov random walks for robust subspace clustering and estimation," *Neural Networks*, vol. 59, pp. 1–15, 2014.

[9] G. Liu, Z. Lin, X. Tang, and Y. Yu, "Unsupervised object segmentation with a hybrid graph model (hgm)," *IEEE TPAMI*, vol. 32, no. 5, pp. 910–924, 2010.

[10] R. Vidal, Y. Ma, and S. Sastry, "Generalized principal component analysis (GPCA)," *IEEE TPAMI*, vol. 27, no. 12, pp. 1945–1959, 2005.

[11] E. J. Candès and B. Recht, "Exact matrix completion via convex optimization," *FOUND COMPUT MATH*, vol. 9, no. 6, pp. 717–772, 2009.

[12] D. L. Donoho, "For most large underdetermined systems of linear equations the minimal l1-norm solution is also the sparsest solution," *COMMUN PUR APPL MATH*, vol. 59, no. 6, pp. 797–829, 2006.

[13] E. Elhamifar and R. Vidal, "Sparse subspace clustering: Algorithm, theory, and applications," *IEEE TPAMI*, vol. 35, no. 11, pp. 2765–2781, 2013.

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

[15] R. Vidal and P. Favaro, "Low rank subspace clustering (LRSC)," *PRL*, vol. 43, pp. 47–61, 2014.

[16] R. Liu, Z. Lin, F. De la Torre, and Z. Su, "Fixed-rank representation for unsupervised visual learning," in *IEEE CVPR*, 2012, pp. 598–605.

[17] Y. Ni, J. Sun, X. Yuan, S. Yan, and L.-F. Cheong, "Robust low-rank subspace segmentation with semidefinite guarantees," in *ICDMW*. IEEE, 2010, pp. 1179–1188.

[18] E. J. Candes and Y. Plan, "Matrix completion with noise," *Proc IEEE*, vol. 98, no. 6, pp. 925–936, 2010.

[19] H. Xu, C. Caramanis, and S. Sanghavi, "Robust pca via outlier pursuit," in *NIPS*, 2010, pp. 2496–2504.

[20] S. Wang, X. Yuan, T. Yao, S. Yan, and J. Shen, "Efficient subspace segmentation via quadratic programming." in *AAAI*, 2011.

[21] Z. Lin, R. Liu, and H. Li, "Linearized alternating direction method with parallel splitting and adaptive penalty for separable convex programs in machine learning," *Machine Learning*, pp. 1–39, 2013.

[22] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE TPAMI*, vol. 22, no. 8, pp. 888–905, 2000.

[23] K.-C. Lee, J. Ho, and D. Kriegman, "Acquiring linear subspaces for face recognition under variable lighting," *IEEE TPAMI*, vol. 27, no. 5, pp. 684–698, 2005.