Maximum Correntropy Criterion-Based Sparse Subspace Learning for Unsupervised Feature Selection

Nan Zhou, Yangyang Xu, Hong Cheng, *Senior Member, IEEE*, Zejian Yuan, *Member, IEEE*, and Badong Chen, *Senior Member, IEEE*

***Abstract*— High-dimensional data contain not only redundancy but also noises produced by the sensors. These noises are usually non-Gaussian distributed. The metrics based on Euclidean distance are not suitable for these situations in general. In order to select the useful features and combat the adverse effects of the noises simultaneously, a robust sparse subspace learning method in unsupervised scenario is proposed in this paper based on the maximum correntropy criterion that shows strong robustness against outliers. Furthermore, an iterative strategy based on half quadratic and an accelerated block coordinate update is proposed. The convergence analysis of the proposed method is also carried out to ensure the convergence to a reliable solution. Extensive experiments are conducted on real-world data sets to show that the new method can filter out the outliers and outperform several state-of-the-art unsupervised feature selection methods.**

***Index Terms*— Machine learning, feature selection, maximum correntropy criterion (MCC), sparse subspace learning.**

I. Introduction

IN PATTERN recognition and computer vision, the dimensionality of the training data increases continuously. Although high dimensional data contain more useful information [1], they not only increase computation and memory requirements but also cause the performance degradation due to much redundancy and noises. Therefore, dimensionality reduction becomes a fundamental and important approach to preprocess data before performing certain learning tasks such as clustering and classification.

Generally speaking, dimensionality reduction approaches can be divided into two categories: feature selection and subspace learning. Feature selection aims to select a subset of relevant features and meanwhile to remove irrelevant and redundant ones out of data (e.g. [2]–[4]), while subspace learning aims to learn a transformation matrix to map the original high-dimensional data into a lower-dimensional representation (e.g. [5]–[7]). Although the two approaches are quite different, some efforts are directed to combine them into a unified framework, especially in unsupervised feature selection (e.g. [8]–[12]). The basic idea is to use a transformation matrix of linear subspace learning to guide the feature selection process. Cai *et al.* [8] regarded the unsupervised feature selection as a spectral sparse subspace learning problem and proposed the Multi-Cluster Feature Selection (MCFS) method. The MCFS method utilizes the ℓ_1-norm to constrain the transformation matrix, which may lead to poor interpretability of the model in selecting features. Gu *et al.* [9] utilize the ℓ_{21}-norm to replace the ℓ_1-norm to improve the interpretability of MCFS method. Because the ℓ_{21}-norm can enforce the row-sparsity of transformation matrix, the irrelevant features are corresponding to the zero-rows of the transformation matrix. Wang *et al.* [10] applied a global regression subspace learning with orthogonality constraint to deal with unsupervised feature selection and proposed the Matrix Factorization Feature Selection (MFFS) method. However, as mentioned in [13], the orthogonality constraint may limit its practicality, since in practice, feature weight vectors are not necessarily orthogonal to each other. For this reason, Zhou *et al.* [11] improved the MFFS method by replacing the orthogonality constraint with a row sparsity constraint and introducing a local structure in the model, and proposed the Global and Local Structure Preserving Sparse Subspace Learning (GLoSS) method for unsupervised feature selection . In both MFFS and GLoSS methods, the Frobenius norm is used to measure the distance between the original data and reconstructed ones, which is sensitive to heavy-tailed non-Gaussian noises or outliers.

Information Theoretic Learning (ITL) provides an elegant and unified way to improve the robustness of machine learning with respect to outliers [14], [15]. Due to its stability and robustness to outliers [16], [17], the Maximum Correntropy

Manuscript received July 20, 2016; revised July 9, 2017 and September 19, 2017; accepted December 8, 2017. Date of publication December 14, 2017; date of current version February 5, 2019. This work was supported in part by NSFC under Grant U1613223 and Grant 61673088, in part by 973 Program under Grant 2015CB351703, in part by NSF under Grant DMS-1719549, and in part by Fundamental Research Funds for Central Universities under Grant ZYGX2014Z009. This paper was recommended by Associate Editor V. Monga. *(Corresponding author: Hong Cheng.)*

N. Zhou is with the College of Control Engineering, Chengdu University of Information Technology, Chengdu 610225, China, and also with the Center for Robotics, University of Electronic Science and Technology of China, Chengdu 611731, China (e-mail: nzhouuestc@126.com).

Y. Xu is with the Department of Mathematical Sciences, Rensselaer Polytechnic Institute, Troy, NY 12180 USA.

H. Cheng is with the Center for Robotics, University of Electronic Science and Technology of China, Chengdu 611731, China (e-mail: hcheng@uestc.edu.cn).

Z. Yuan and B. Chen are with the Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University, Xi'an 710049, China.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCSVT.2017.2783364

Criterion (MCC) in ITL has recently been successfully applied to a wide range of applications, such as adaptive filtering [18]–[20], Kalman filtering [21] nonnegative matrix factorization [22], face recognition [23] and supervised feature selection [24]. However, to the best of our knowledge, there has been no study yet on MCC based unsupervised feature selection. In this paper, we will propose a MCC based robust unsupervised feature selection method called the Global and Local preserved Robust Sparse Subspace Learning (GLoRSS).

The main contributions of the present work are summarized as follows:

1) A novel robust sparse subspace learning model is proposed for unsupervised feature selection, in which the MCC is used for robust data fitting while preserving the local structure of the data.
2) A half-quadratic accelerated block coordinate update (HQ-BCU) method is proposed to solve this model which is shown to be very efficient with low computational complexity.
3) Convergence analysis has been carried out to ensure that the subsequence convergence can be obtained by our method. With nondegeneracy assumption, a stronger whole iterate sequence convergence can also be obtained.
4) Extensive experimental studies are presented in this paper. The effectiveness of the GLoRSS is verified by outlier detection and face reconstruction on the dataset contaminated by noises and outliers. In the unsupervised feature selection experiments, six real-world datasets from different domains are used, and the proposed GLoRSS is compared to eight state-of-the-art unsupervised feature selection algorithms. The sensitivity of the parameters in the GLoRSS is also studied. It is shown that the new method can perform well in a stable way within a large range of parameters.

The rest of the paper is organized as follows. Section II gives a brief review of subspace learning methods and robust methods for unsupervised feature selection. Section III revisits the Maximum Correntropy Criterion (MCC) and proposes a robust sparse subspace learning model based on MCC called GLoRSS. Section IV develops an iterative learning algorithm GLoRSS to search the solution of the model and analyzes its convergence. Section V presents the experimental studies. Finally, Section VI gives the conclusion.

To facilitate the presentation of the paper, the notations are listed in Table I

TABLE I
NOTATION

Notation	Description
n	Number of instances
d	Number of features
κ	Number of the selected features
$A_{i.}$	The i^{-th} row of the matrix A
K	Dimension of subspace
m	Number of nearest neighbors
$\|W\|_{2,1}$	$\sum_i \|W_{i.}\|_2$, the sum of the ℓ_2-norm of rows in W
$\|\mathbf{x}\|_0$	$\sharp\{x_i \neq 0\}$, the number of nonzero elements in vector $\mathbf{x}$
$\|\mathcal{I}\|$	Cardinality of set $\mathcal{I}$

II. Related Works

Consider n data samples $\{\mathbf{p}_i\}_{i=1}^n$ located in the d-dimensional space. $X = [\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_n]^\top \in \mathbb{R}^{n\times d}$ is the data set of samples. Based on reconstruction information of the data, the unsupervised feature selection problem can be formulated as follows [10], [11]:

$$\begin{aligned} \min_{W,H} \quad & \frac{1}{2}\|X - XWH\|_F^2 \\ \text{s.t. } & W \in \{0,1\}^{d\times K},\ W^\top \mathbf{1}_{d\times 1} = \mathbf{1}_{K\times 1}, \\ & \|W\mathbf{1}_{K\times 1}\|_0 = K, \end{aligned} \tag{1}$$

where the matrix W is the feature selection matrix with the entries "1" or "0" stating whether the feature is selected or not. For example, if the number of features is 5 and the indices of the selected features are $(2, 3, 5)$, then W comes in the following form:

$$W = \begin{pmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix} \tag{2}$$

The model (1) is of combinatorial nature which is extremely hard to solve. For this reason, Wang *et al.* [10] relaxed it to continuous version by orthogonality and nonnegativity constraints. However, this model only considers the global reconstruction information for subspace learning. In addition, the orthogonality constraint may limit its practicality. As mentioned in [13], feature weight vectors are not necessarily orthogonal to each other. Therefore, Zhou *et al.* [11] introduce the manifold embedding to preserve the local structure of the data. It is noted that all these methods use the matrix Frobenius norm to measure the distance between the original data space and the subspace, thus they are sensitive to outliers. Some methods are proposed for robust unsupervised feature selection. Qian and Zhai [13] used the ℓ_{21}-norm as the metric to measure the distance and proposed Robust Unsupervised Feature Selection (RUFS) methods. However, the ℓ_{21} norm induced lost function is hard to solve with extensively high computational complexity $O(d^3)$, thus it is not suitable for high dimensional data. Considering the global regression information, Zhu *et al.* [25] utilized the ℓ_{21}-norm to model the self-representation lost function with ℓ_{21}-norm constraint for robust unsupervised feature selection. All these robust unsupervised feature selection methods only use the ℓ_{21}-norm to improve the robustness of the model. Due to desirable properties of correntropy [14], in this paper, we utilize the Maximum Correntropy Criterion (MCC) to deal with robust unsupervised feature selection. Compared with the ℓ_{21}-norm approach, the MCC is much easier to solve with lower computational complexity.

III. Robust SSL Based on MCC

In this section, a robust sparse subspace learning model based on MCC is proposed. First, a brief review about the MCC is provided. Then, the original model (1) is extended to its robust version by introducing MCC measure. This

model is a combinatorial nature, thus it is not easy to find a reliable solution to the problem. Finally, in order to find a reliable solution, the combinatorial model is relaxed, in which sparsity and nonnegativity constraints are utilized to replace the original constraints in (1).

A. Maximum Correntropy Criterion

Correntropy is proposed in ITL to handle non-Gaussian noises and large outliers [14], and has been widely used in signal processing [16], [17] and machine learning [22]–[24]. It is directly related to Renyi's quadratic entropy and the probability of how similar two random variables are in a local range controlled by the kernel bandwidth. Adjusting the kernel bandwidth provides an effective way to eliminate the detrimental effect of outliers. The MCC can extract more information from the data for adaptation, and may yield, therefore, solutions that are more accurate than traditional MSE solution particularly in non-Gaussian signal processing [14]. In addition [14], correntropy induces a new metric which is equivalent to the ℓ_2 norm distance if points are close, behaves like the ℓ_1 norm distance as points get further apart and eventually approaches the ℓ_0 norm as points are far apart. This can explain why MCC is more robust to outliers than conventional ℓ_1-norm. Correntropy is defined as a similarity measure between two arbitrary random variables X and Y

$$V_\sigma(X,Y) = E[k_\sigma(X-Y)], \tag{3}$$

where $k_\sigma(\cdot)$ is the shift invariant mercer kernel with kernel bandwidth $\sigma > 0$ [26] which presents the correlation in kernel space and $E[\cdot]$ denotes the mathematical expectation.

In practice, the joint probability density function is often unknown, and only a finite number of data $\{(x_i, y_i)\}_{i=1}^n$ are available, which lead to the sample estimator of correntropy as follows:

$$\hat{V}_\sigma(X,Y) = \frac{1}{n}\sum_{i=1}^{n} k_\sigma(x_i - y_i). \tag{4}$$

The maximum of correntropy of error in (4) is called the maximum correntropy criterion (MCC). In this work, without mentioned otherwise, the kernel function is the Gaussian kernel $k_\sigma(X) = g(X,\sigma) \triangleq \exp(-X^2/2\sigma^2)$, thus the correntropy can be rewritten as follows:

$$\hat{V}_\sigma(X,Y) = \frac{1}{n}\sum_{i=1}^{n} g(x_i - y_i, \sigma), \tag{5}$$

which has a close relationship with M-estimation [27]. The term in (5) can be seen as the robust formulation of Welsch M-estimator, if we define $\rho(x) \triangleq 1 - g(x,\sigma)$. A main merit of MCC is that the kernel size decides the main properties of correntropy. Therefore, correntropy establishes a close relation between the M-estimation and ITL [14].

B. Sparse Subspace Learning Based on Maximum Correntropy Criterion

To introduce the correntropy measure into the model (1), we extend the MCC from vector space $\mathbb{R}^d$ to matrix space $\mathbb{R}^{n\times d}$ by replacing $|x_i - y_i|$ with $\|\mathbf{p}_i^\top - \mathbf{p}_i^\top WH\|_2$, the ℓ_2 norm distance between the sample $\mathbf{p}_i^\top$ and its reconstruction $\mathbf{p}_i^\top WH$. Using the correntropy measure instead of the Euclidean distance, we obtain a robust version of the model (1) as follows:

$$\begin{aligned} \max_{W,H}\ & \frac{1}{2}\sum_{i=1}^{n}\exp\left(\frac{-\|\mathbf{p}_i^\top - \mathbf{p}_i^\top WH\|_2^2}{2\sigma^2}\right) \\ \text{s.t. } & W\in\{0,1\}^{d\times K},\quad W^\top \mathbf{1}_{d\times 1} = \mathbf{1}_{K\times 1}, \\ & \|W\mathbf{1}_{K\times 1}\|_0 = K. \end{aligned} \tag{6}$$

Compared to $\ell_{2,1}$-norm that extends ℓ_1-norm and can also improve model's robustness to outliers, MCC is a local measure whose value mainly depends on the probability along $x = y$. It can truncate the effects by large error and thus better eliminate the negative affects of outliers than $\ell_{2,1}$-norm [14]. In addition, (6) is different from conventional kernel methods. The latters aim at building a nonlinear model using the kernel trick, while MCC employs a new cost function for training a model (not necessarily a kernel model). Their connection lies in that the MCC cost can be expressed as a distance in the kernel space.

To obtain a reliable solution, we follow [11] and relax the 0-1 variables in (6) to continuous ones by replacing the constraints in (6) with the sparsity and nonnegativity constraints. The local geometrical structure is a valuable discriminative information and plays important role in spectral clustering [28], [29]. In order to extract more information of the data, we add a local geometrical structure preserving term and obtain the following regularized model with continuous variables:

$$\begin{aligned} \max_{W,H}\ & \frac{1}{2}\sum_{i=1}^{n}\exp\left(\frac{-\|\mathbf{p}_i^\top - \mathbf{p}_i^\top WH\|_2^2}{2\sigma^2}\right) \\ & -\frac{\mu}{2}Tr(W^\top X^\top LXW) \\ \text{s.t. } & W\in\mathbb{R}_+^{d\times K},\ r(W)\le\kappa. \end{aligned} \tag{7}$$

Here L is the graph Laplacian matrix, which is used to describe the local geometrical structure of the data and can be formulated by many ways [30], $r(W)$ measures the row-sparsity of W, $\mathbb{R}_+^{d\times K}$ denotes the $d\times K$ nonnegative matrix. To alleviate the difficulty for solving the model (7), it is transformed to its equivalent formulation by penalizing the sparsity constraint into the objective as follows:

$$\begin{aligned} \max_{W,H}\ & \frac{1}{2}\sum_{i=1}^{n}\exp\left(\frac{-\|\mathbf{p}_i^\top - \mathbf{p}_i^\top WH\|_2^2}{2\sigma^2}\right) \\ & -\frac{\mu}{2}Tr(W^\top X^\top LXW) - \beta r(W) \\ \text{s.t. } & W\in\mathbb{R}_+^{d\times K}, \end{aligned} \tag{8}$$

where β is the parameter corresponding to κ.

IV. Solving the Robust Sparse Subspace Learning

In this section, we derive an iterative method to find a solution of (8). We first reformulate (8) by the half-quadratic (HQ) technique [31] and then apply the accelerated block coordinate

update (BCU) [32] to the reformulated problem. Although (8) involves fewer variables than the reformulated model, it lacks block-concave structure, and that can cause difficulty for a numerical approach to efficiently and stably find a solution to (8). On the contrary, the reformulated model has the nice block-concave structure, which enables acceleration of a numerical method. We choose BCU instead of the traditional gradient ascent due to the nice block structure of the reformulated problem that is amenable to BCU method as discussed in [33] and [34]. While the objective in (8) is highly nonconcave[1], the reformulated problem becomes concave with respect to each block variable. As shown in [32], for solving multi-block concave problems, BCU has desirable convergence behavior and also very nice practical performance.

A. Reformulation via Half Quadratic Technique

Let $\varphi(z) = z - z\ln(-z)$. It holds that

$$\exp(-x) = \sup_z \left\{ zx - \varphi(z) \right\} \tag{9}$$

by setting $\frac{\partial}{\partial z}(zx - \varphi(z)) = x + \ln(-z) = 0$, and one can find that the maximum value of (9) is reached at $z = -\exp(-x)$. Using this fact, we have

$$\exp\left(\frac{-\|\mathbf{p}_i^\top - \mathbf{p}_i^\top WH\|_2^2}{2\sigma^2}\right) = \sup_{y_i}\left\{ y_i \frac{\|\mathbf{p}_i^\top - \mathbf{p}_i^\top WH\|_2^2}{2\sigma^2} - \varphi(y_i) \right\},\ i = 1, \ldots, n. \tag{10}$$

Let

$$O^{MCC}(W, H, \mathbf{y}) = \frac{1}{2}\sum_{i=1}^{n}\left(y_i \frac{\|\mathbf{p}_i^\top - \mathbf{p}_i^\top WH\|_2^2}{2\sigma^2} - \varphi(y_i) \right),$$

$$f(W, H, \mathbf{y}) = O^{MCC}(W, H, \mathbf{y}) - \frac{\mu}{2} Tr(W^\top X^\top LXW).$$

Then according to (10), the problem (8) is equivalent to

$$\max_{W,H,\mathbf{y}} f(W, H, \mathbf{y}) - r_\beta(W),\ \text{s.t. } W \in \mathbb{R}_+^{d\times K}, \tag{11}$$

where $\mathbf{y} = (y_1, \ldots, y_n)$, and

$$r_\beta(W) = \beta \sum_{i=1}^{d} \|W_{i.}\|_2.$$

B. Iterative Method by the Accelerated BCU

Our algorithm is derived by applying the accelerated BCU method in [32] to (11). At each iteration, it updates the variables W, H and $\mathbf{y}$ sequentially, one at a time with the other two fixed to their most recent values, by maximizing the objective or a surrogate of its lower bound. Specifically, it iteratively performs the following updates:

$$W^{k+1} = \operatorname*{argmax}_{W\in\mathbb{R}_+^{K\times d}} \langle \nabla_W f(\hat{W}^k, H^k, \mathbf{y}^k), W - \hat{W}^k \rangle - \frac{L_w^k}{2}\|W - \hat{W}^k\|_F^2 - r_\beta(W), \tag{12a}$$

$$H^{k+1} = \operatorname*{argmax}_{H} O^{MCC}(W^{k+1}, H, \mathbf{y}^k), \tag{12b}$$

$$\mathbf{y}^{k+1} = \operatorname*{argmax}_{\mathbf{y}} O^{MCC}(W^{k+1}, H^{k+1}, \mathbf{y}), \tag{12c}$$

where L_w^k is the Lipschitz constant of $\nabla_W f(W, H^k, \mathbf{y}^k)$ with respect to W, and

$$\hat{W}^k = W^k + \omega_k(W^k - W^{k-1}) \tag{13}$$

is an extrapolated point with weight $\omega_k \in [0, 1)$. It is demonstrated in [35] and [36] that appropriate ω_k can significantly accelerate the BCU method for solving certain multi-block concave optimization problems.

Notice that we treat W and $H, \mathbf{y}$ in two different ways. The variable W is updated by a block proximal gradient method while the other two are updated by simple block maximization, because directly maximizing the objective of (11) with respect to W can be very difficult. In the way as done in (12), one can achieve a closed-form solution to each of the three subproblems. In the following, we discuss how to solve them explicitly.

1) Solution to W-Subproblem: Note that (12a) can be rewritten as

$$\min_{W\in\mathbb{R}_+^{K\times d}} \frac{1}{2}\|W - A\|_F^2 + \lambda\|W\|_{2,1}, \tag{14}$$

where $A = \hat{W}^k + \frac{1}{L_w^k}\nabla_W f(\hat{W}^k, H^k, \mathbf{y}^k)$ and $\lambda = \frac{\beta}{L_\omega^k}$. The problem (14) can be further decomposed into d smaller independent problems, each one involving one row of W and A and in the form of

$$\min_{\mathbf{w}\geq 0} \frac{1}{2}\|\mathbf{w} - \mathbf{a}\|_2^2 + \lambda\|\mathbf{w}\|_2. \tag{15}$$

As shown in [11], the problem (15) has a closed-form solution and thus (12a) can be solved explicitly. For readers' convenience, we give in Algorithm 1, the steps toward a solution to (14).

Algorithm 1 Proximal Operator for Nonnegative Group Lasso: $W = \text{Prox-NGL}(A, \lambda)$

```
for i = 1, ..., d do
  Let a be the i^th row of A and I the index set of positive components of a
  Set w to a zero vector
  if ||a_I||_2 > λ then
    Let w_I = (||a_I||_2 − λ) a_I / ||a_I||_2
  end if
  Set the i^th row of W to w
end for
```

[1]The work [32] is about block convex problems. But note that if a function $f(x)$ is convex, then $-f(x)$ would be concave.

2) Solution to H-Subproblem: Let

$$X^k = \text{Diag}\left(\sqrt{-\frac{\mathbf{y}^k}{2\sigma^2}}\right) X = \text{Diag}\left(\sqrt{-\frac{\mathbf{y}^k}{2\sigma^2}}\right) [\mathbf{p}_1, \ldots, \mathbf{p}_n]^\top. \tag{16}$$

Then it is easy to see that (12b) is equivalent to

$$H^{k+1} = \arg\min_H \frac{1}{2}\|X^k - X^k W^{k+1} H\|_F^2,$$

which can be solved by setting its first-order optimality condition $(X^k W^{k+1})^\top (X^k W^{k+1} H - X^k) = 0$. Hence, the update in (12b) can be explicitly written as

$$H^{k+1} = \left[(X^k W^{k+1})^\top X^k W^{k+1}\right]^\dagger (X^k W^{k+1})^\top X^k, \tag{17}$$

where "$A^\dagger$" denotes the Moore-Penrose pseudo-inverse of a matrix A.

3) Solution to $\mathbf{y}$-Subproblem: As discussed at the beginning of section IV-A, the maximum value of (9) is reached at $z = -\exp(-x)$. Hence, the solution to (12c) can be explicitly written as

$$\mathbf{y}^{k+1} = \mathcal{T}(W^{k+1}, H^{k+1}), \tag{18}$$

where the mapping $\mathbf{y} = \mathcal{T}(W, H)$ is defined as

$$y_i = -\exp\left(\frac{-\|\mathbf{p}_i^\top - \mathbf{p}_i^\top W H\|_2^2}{2\sigma^2}\right), \quad i = 1, \ldots, n. \tag{19}$$

4) Parameter Settings: From (17) and (18), we see that the updates to H and $\mathbf{y}$ are parameter-free. However, to fully determine the W-update in (12a) requires the values of L_w^k and ω_k. In our algorithm, we set L_w^k to

$$L_w^k = \|(X^k)^\top X^k\|_2 \|H^k (H^k)^\top\|_2 + \mu \|X^\top L X\|_2, \tag{20}$$

which is a Lipschitz constant of $\nabla_W f(W, H^k, \mathbf{y}^k)$ with respect to W according to (28) in the appendix. Note that we can replace the matrix 2-norm in (20) to Frobenius norm. For large-size matrix, the latter one can be much cheaper but is usually larger. In our experiments, we have either small n or d, and since $\|(X^k)^\top X^k\|_2 = \|X^k (X^k)^\top\|_2$, we can choose to evaluate the small-sized one. For the extrapolation weight, we follow [36] and set it to

$$\omega_k = \min\left(\hat{\omega}_k, \delta_\omega \sqrt{\frac{L_\omega^{k-1}}{L_\omega^k}}\right), \tag{21}$$

where $\delta_\omega < 1$ is predetermined and $\hat{\omega}_k = (t_{k-1} - 1)/t_k$ with

$$t_0 = 1, t_k = \frac{1}{2}\left(1 + \sqrt{1 + 4t_{k-1}^2}\right). \tag{22}$$

Putting all discussions in this subsection together, we reach a complete iterative method toward finding a solution to (8), and its pseudo-code is shown in Algorithm 2, dubbed as GLoRSS.

Algorithm 2 **G**lobal and **Lo**cal Structure Preserving **R**obust **S**parse **S**ubspace Learning (GLoRSS)

1: **Input**: Data matrix $X \in \mathbb{R}^{n\times d}$, the number of selected features κ and parameter β, μ.
2: **Output**: Index set of selected features $\mathcal{I} \subseteq \{1, \ldots, d\}$ with $|\mathcal{I}| = \kappa$
3: **Initialize** $W^0 \in \mathbb{R}_+^{d\times K}$, $H^0 \in \mathbb{R}^{K\times d}$, set $\mathbf{y}^0$ by (18), choose a positive number $\delta_\omega < 1$, and set $k = 0$.
4: **while** Not convergent **do**
5: Compute X^k according to (16).
6: Compute L_w^k and ω_k according to (20) and (21) respectively.
7: Let $\hat{W}^k = W^k + \omega_k(W^k - W^{k-1})$.
8: Update $W^{k+1} \leftarrow$ by Algorithm 1.
9: **if** $f(W^{k+1}, H^k, \mathbf{y}^k) - r_\beta(W^{k+1}) \le f(W^k, H^k, \mathbf{y}^k) - r_\beta(W^k)$ **then**
10: Set $\hat{W}^k = W^k$ and back to **Step** 8.
11: **end if**
12: Update $H^{k+1} \leftarrow$ (17).
13: Update $\mathbf{y}^{k+1} \leftarrow$ (18).
14: Let $k \leftarrow k + 1$.
15: **end while**
16: Normalize each column of $W = W^{k+1}$.
17: Sort $\|W_{i\cdot}\|_2$, $i = 1, \ldots, d$ and select features corresponding to the κ largest ones.

C. Computational Complexity

Now, we evaluate the computational complexity per iteration of Algorithm 2. The analysis is for general case and no special structure is assumed to the data matrix X. If X has certain structure, e.g., sparsity, the computational complexity can be lower. We assume the dimension of subspace $K < \min(d, n)$. With X^k computed at the cost of $O(nd)$, the cost of updating W and H takes $O(ndK + nK^2 + dK^2)$ flops according to the complexity analysis of GLoSS in [11]. The update of $\mathbf{y}$ by (18) costs $O(ndK)$, and the evaluation of the objective value in line 9 also costs $O(ndK)$. The other computation in Algorithm 2 is much cheaper. Therefore, the per-iteration cost of the algorithm is $O(ndK + nK^2 + dK^2)$, which indicates scalability to the data size if $K = O(1)$.

D. Convergence Analysis

Next, we analyze the convergence of Algorithm 2. The algorithm enforces the monotonicity of the objective in line 9, and thus we can simply get the objective sequence convergence because the objective function in (11) is upper bounded by $\frac{n}{2}$. However, the objective convergence does not guarantee any optimality property of the iterate sequence. We next establish iterate subsequence convergence without any assumption on the algorithm and also whole iterate sequence convergence by assuming a nondegeneracy condition.

Let

$$F(W, H) = \tfrac{1}{2}\sum_{i=1}^n \exp\left(\frac{-\|\mathbf{p}_i^\top - \mathbf{p}_i^\top W H\|_2^2}{2\sigma^2}\right) - \tfrac{\mu}{2} Tr(W^\top X^\top L X W),$$

$$\Psi(W, H) = F(W, H) - r_\beta(W) - \iota_{\mathbb{R}_+^{d\times K}},$$

where $\iota_{\mathcal{X}}$ denotes the indicator function on a set $\mathcal{X}$. Then the problem (8) is equivalent to $\min_{W,H} \Psi(W, H)$, and its first-order optimality condition is $0 \in \partial\Psi(W, H)$. Any point (W, H) satisfying the condition is called a critical point of (8). The following two theorems summarize our iterate sequence convergence results.

Theorem 1 (Iterate Subsequence Convergence): Let $\{(W^k, H^k, \mathbf{y}^k)\}$ *be the sequence generated from Algorithm 2. Then any finite limit point* $(\bar{W}, \bar{H})$ *of* $\{(W^k, H^k)\}$ *is a critical point of* (8)*, namely, satisfying the first-order optimality condition* $0 \in \partial\Psi(\bar{W}, \bar{H})$.

The proof of the theorem is given in the Appendix. Here, we provide some insights to it. First, all the three updates in (12) result in sufficient increase of the objective value of (11). While the objective function is upper bounded by $\frac{n}{2}$, the total increase must be finite, and thus the difference of two consecutive iterates will eventually diminish. Hence, one can show the criticality of $(\bar{W}, \bar{H}, \bar{\mathbf{y}})$ to (11) where $\bar{\mathbf{y}} = \mathcal{T}(\bar{W}, \bar{H})$. Finally, we notice

$$\nabla_{(W,H)} f(\bar{W}, \bar{H}, \bar{\mathbf{y}}) = \nabla F(\bar{W}, \bar{H}) \tag{23}$$

to obtain the criticality of $(\bar{W}, \bar{H})$ to (8).

Assuming a nondegeneracy condition on $\bar{W}$, we show that there can be only one limit point, and thus the whole sequence $\{(W^k, H^k)\}$ converges. Without the condition, the whole sequence convergence cannot hold because in that case, the solution to (12b) is not uniquely determined in the limit. The nondegeneracy assumption is similar to that used in [37] for establishing the convergence of the higher-order orthogonality iteration, which is shown to possibly diverge even starting from a degenerate critical point.

Theorem 2 (Whole Iterate Sequence Convergence): Let $\{(W^k, H^k, \mathbf{y}^k)\}$ *be the sequence generated from Algorithm 2. If there is a finite limit point* $(\bar{W}, \bar{H})$ *of* $\{(W^k, H^k)\}$ *such that* $\bar{X}\bar{W}$ *is full-rank, then the whole sequence* $\{(W^k, H^k)\}$ *converges to* $(\bar{W}, \bar{H})$*, where* $\bar{X} = \text{Diag}(\sqrt{-\frac{\bar{\mathbf{y}}}{2\sigma^2}})X$ *and* $\bar{\mathbf{y}} = \mathcal{T}(\bar{W}, \bar{H})$.

The proof of this theorem is also provided in the Appendix.

V. Experimental Studies

Experimental results are illustrated in this section to show the performance of the proposed GLoRSS method. In particular to demonstrate the robust performance, the outlier image detection and noisy face image reconstruction experiments are conducted on Yale64 dataset. Then, the unsupervised feature selection performance of the new method is tested on six benchmark datasets and compared to eight state-of-the-art methods. Following the method in [22], the bandwidth of the Gaussian kernel is updated by

$$\sigma^k = \sqrt{\frac{\theta}{2n}\|X - XW^k H^k\|_F^2}, \tag{24}$$

where θ is a constant.

A. Datasets

In the experiments, six benchmark datasets coming from different domains are used, whose characteristics are shown in Table II. Yale64, WarpPIE, Orl64 and Orlraws are face images, and each instance of the datasets presents a single face image. USPS is a handwritten digit dataset, and each instance of the dataset presents a handwrite digit image. Isolet is a speech signal dataset containing 30 speakers' speech signal of alphabet twice.

TABLE II
The Datasets Detail

Dataset	♯ Instances	♯ Features	♯ Classes	Type of Data
Yale64	165	4096	15	Face image
WarpPIE	210	2420	10	Face image
Orl64	400	4096	50	Face image
Orlraws	100	10304	10	Face image
Usps	9298	256	10	Digit image
Isolet	1560	617	26	Speech signal

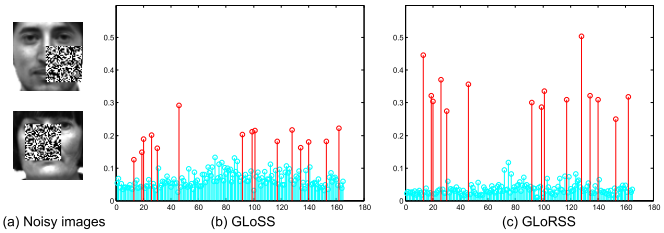


Fig. 1. Reconstruction error with artificial rectangle noises: (a) Two examples of noisy images. (b) Reconstruction errors obtained by GLoSS method. (c) Reconstruction errors obtained by GLoRSS method.

B. Outlier Detection

In outlier detection experiments, in order to show the robustness property of GLoRSS, we compare it to GLoSS [11] that uses the matrix Frobenius norm as the distance measure. Yale64 dataset is used in this problem. Two types of outlier are simulated. For the first type, 15 image are randomly selected, and each of them is occluded by rectangular area, each pixel in which is randomly set to 0 (black) or 255 (white). The size of the rectangular is set as $1/4$ size of the original image, and the location is randomly determined. Two examples of the occluded faces is shown in (a) of Fig. 1. For the second type of outliers, we add 15 dummy images (outlier) with the same size to the Yale64 dataset. The final 15 samples are outliers, and two of them are shown in (a) of Fig. 2. The reconstruction errors with both types of outliers have been presented in Fig. 1 and Fig. 2, where the red sticks are the positions of outliers. From the figures, we see that the GLoRSS method that is based on MCC measure can better detect the outliers and alleviated the impact of outliers.

C. Face Reconstruction

In this section, the robustness of GLoRSS is verified by face reconstruction on contaminated Yale64 dataset. The reconstructions by both GLoSS and GLoRSS methods are shown in three contaminated situations of noises and outliers that are described as follows:

1) The dataset has 15 outliers, which is the same as Fig. 2. The outliers are added to the dataset as last samples.

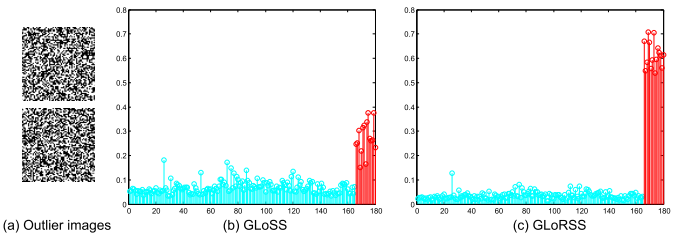


Fig. 2. Reconstruction error with outliers: (a) Two examples of outlier images. (b) Reconstruction errors obtained by GLoSS method. (c) Reconstruction errors obtained by GLoRSS method.

Fig. 3. Face reconstruction when dataset is contaminated by outliers: The first column shows the original images, the second column shows the reconstructed images by GLoSS method, the third column shows the reconstruction by GLoRSS method.

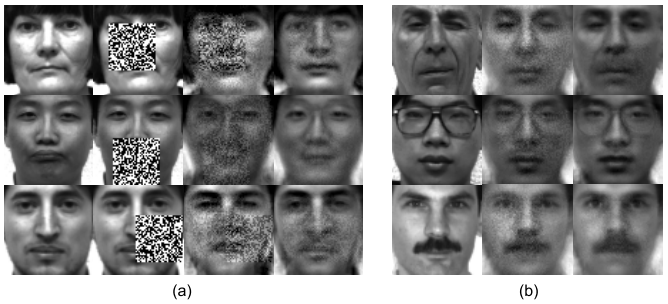


Fig. 4. Face reconstruction with rectangular noises. (a) Reconstruction for noisy images: The first column shows the original images. The second column shows the images with the noises. The third column shows the reconstruction by GLoSS method. The forth column shows the reconstruction by GLoRSS method. (b) Reconstruction for clean images: The first column shows the original images. The second column shows the reconstruction by GLoSS method. The third column shows the reconstruction by GLoRSS method.

2) 15 face images of the dataset are randomly chosen to add the rectangular noises which is the same as (a) of Fig. 1.
3) 15 face images of the datasets are randomly chosen to add the 5% pepper & salt noises.

Figs. 3-5 show the reconstructions when the dataset is contaminated by the three kinds of noise.

Fig. 3 shows that outliers affect the reconstructions by both methods, and the GLoRSS method can recover more details of the face than the GLoSS method. From Fig. 4, we see that the GLoSS method almost always fails to reconstruct the face that is contaminated by the rectangular noises, but GLoRSS can fix the face and preserve some of the facial characteristics outside and even inside the rectangular noise area. For clean faces, the reconstruction by GLoRSS looks much better than that by GLoSS. The results by the latter apparently contain lots

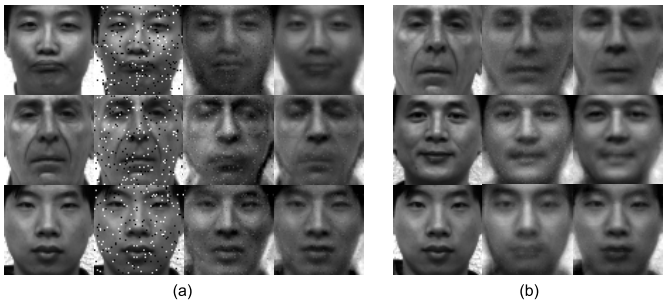


Fig. 5. Face reconstruction with pepper and salt noises. (a) Reconstruction for noisy images: The first column shows the original images. The second column shows the images with the noises. The third column shows the reconstruction by GLoSS method. The forth column shows the reconstruction by GLoRSS method. (b) Reconstruction for clean images: The first column shows the original images. The second column shows the reconstruction by GLoSS method. The third column shows the reconstruction by GLoRSS method.

of noise. Fig. 5 shows the results for the pepper and salt noise case. We notice that for noisy face reconstruction, the GLoSS method can hardly remove noises while the proposed GLoRSS method can filter out lots of noises on the face. For the clean images, the reconstruction by GLoRSS keeps key features better than that by GLoSS.

D. Unsupervised Feature Selection

In this section, the unsupervised feature selection performance of the proposed GLoRSS method is studied on six real-world benchmark datasets and compared to eight state-of-the-art unsupervised feature selection methods, listed below:

1) *LS*: Laplacian score (LS) method [38] selects the features individually that retain the samples' local similarity specified by a similarity matrix.
2) *MCFS*: Multi-cluster feature selection (MCFS) [8] selects the features by spectral sparse subspace learning method with ℓ_1-norm constraint.
3) *UDFS*: Unsupervised discriminative feature selection (UDFS) method [39] selects the features by their local discriminative property with the $\ell_{2,1}$-norm constraint.
4) *RSR*: Regularized self-representation (RSR) feature selection method [25] is a robust unsupervised feature selection method which uses the $\ell_{2,1}$-norm as the fitting measure and $\ell_{2,1}$-norm to promote row sparsity.
5) *NDFS*: Nonnegative Discriminative Feature Selection (NDFS) method [40] selects the features by nonnegative spectral sparse clustering with $\ell_{2,1}$-norm regularization term.
6) *GLSPFS*: Global and local structure preservation for feature selection (GLSPFS) method [30] uses both global and local similarity structure to model the feature selection problem.
7) *MFFS*: Matrix factorization feature selection (MFFS) method [10] selects the features by global regression subspace learning with orthogonal constraint.
8) *GLoSS*: Global and local structure preserving sparse subspace learning method [11] performs the sparse subspace learning by global data regression and locality

TABLE III

CLUSTERING RESULTS (ACC% ± STD%) OF DIFFERENT FEATURE SELECTION ALGORITHMS ON DIFFERENT DATASETS. THE BEST RESULTS ARE HIGHLIGHTED IN **BOLD** AND THE SECOND BEST RESULTS ARE UNDERLINED. (HIGHER ACC IS BETTER)

Dateset	Isolet	Yale64	Orl64	WarpPIE	Usps	Orlraw
LS	55.14 ± 3.15	41.25 ± 3.28	41.75 ± 1.71	32.33 ± 1.03	59.79 ± 2.72	66.12 ± 6.82
MCFS	54.95 ± 3.28	44.88 ± 3.72	50.75 ± 1.25	50.38 ± 2.25	66.55 ± 3.11	77.43 ± 7.15
UDFS	29.60 ± 2.73	38.21 ± 3.02	40.78 ± 1.03	55.57 ± 2.92	50.59 ± 1.97	65.32 ± 6.18
RSR	49.88 ± 3.75	45.48 ± 3.34	53.24 ± 1.83	37.52 ± 2.23	62.54 ± 2.34	72.54 ± 6.52
NDFS	54.33 ± 3.73	45.79 ± 3.81	49.85 ± 1.69	34.10 ± 3.81	63.32 ± 3.35	67.80 ± 6.48
GLSPFS	54.09 ± 3.22	50.84 ± 5.34	53.63 ± 2.62	45.94 ± 2.38	64.65 ± 3.69	78.00 ± 7.47
MFFS	55.39 ± 3.32	49.09 ± 3.64	50.19 ± 1.64	36.57 ± 2.32	63.30 ± 3.36	73.55 ± 7.68
GLoSS	62.45 ± 3.58	53.45 ± 3.88	54.27 ± 1.87	52.76 ± 2.12	67.24 ± 3.27	79.37 ± 7.34
GLoRSS	**63.29 ± 2.86**	**55.75 ± 3.62**	**54.75 ± 2.51**	**56.26 ± 3.76**	**70.93 ± 3.26**	**79.95 ± 7.56**

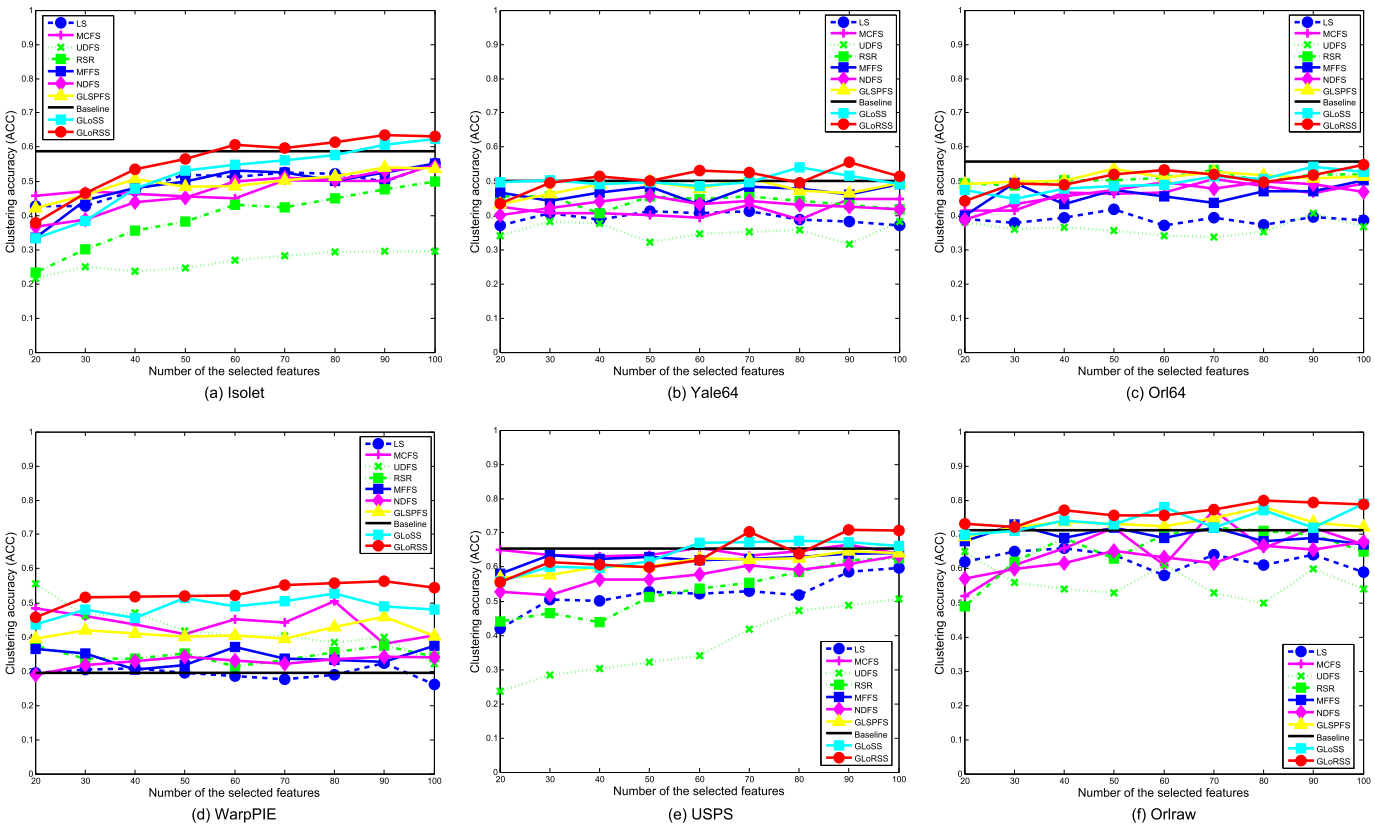


Fig. 6. Clustering accuracy (ACC) of using all features and selected features by different feature selection algorithms.

preserving with ℓ_{21}-norm constraint. In subspace learning process, the Frobenius norm is used to measure the distance between the data and reconstruction.

1) Experimental Setting: Thereare some parameters we need to set in advance. The dimension of subspace is fixed to $K = 100$ for both GLoSS and GLoRSS, and the number of selected features κ is taken from $\{20, 30, 40, 50, 60, 70, 80, 90, 100\}$ for all datasets. The Linear Preserve Projection (LPP) method is used to construct the Laplacian graph to preserve the local structure in GLSPFS, NDFS, GLoSS, MCFS, LS and GLoRSS because MCFS, NDFS and LS use the LPP method to construct Laplacian graph, and GLSPFS can obtain better performance with the LPP method. The number of nearest neighbors is set to $m = 5$ for LS, MCFS, UDFS, GLSPFS, NDFS, GLoSS and GLoRSS, because parameter m is required for UDFS to build the local total scatter and between-class scatter matrices and the other methods to build the similarity matrix.

For simplicity, the trade-off parameter of local structure preserving term μ is fixed to $\mu = 1$ in GLSPFS, GLoSS and GLoRSS for all tests. The sparsity controlling parameter β and the bandwidth control parameter θ are both tuned by the "grid-search" strategy with β from $\{0.001, 0.01, 0.1, 1, 10, 40, 70, 100\}$ and θ from {0.1, 1, 2, 3, 4, 5, 6, 7, 8, 9}. After completing the feature selection process, the selected features are used to cluster by K-means algorithm. Because the performance of K-means depends on the initial point, the algorithm is run 20 times with different random starting points, and the average results are reported.

The performance of the algorithms is evaluated based on their clustering results. For each dataset, the clustering number is set as its class number which is given by dataset. The clustering performance is measured by two criterions: the clustering

TABLE IV

CLUSTERING RESULTS (NMI% ± STD%) OF DIFFERENT FEATURE SELECTION ALGORITHMS ON DIFFERENT DATASETS. THE BEST RESULTS ARE HIGHLIGHTED IN **BOLD** AND THE SECOND BEST RESULTS ARE UNDERLINED. (HIGHER NMI IS BETTER)

Dateset	Isolet	Yale64	Orl64	WarpPIE	Usps	Orlraw
LS	69.73 ± 1.43	46.88 ± 2.07	62.61 ± 1.53	30.06 ± 2.89	56.62 ± 0.95	73.38 ± 3.12
MCFS	69.82 ± 1.37	53.70 ± 1.58	69.33 ± 1.62	54.37 ± 4.95	61.01 ± 0.92	83.91 ± 3.53
UDFS	44.98 ± 1.02	47.40 ± 1.64	62.38± 1.41	54.55 ± 4.38	41.31 ± 1.21	68.78 ± 3.45
RSR	63.47 ± 1.10	56.08 ± 1.43	72.33 ± 1.75	41.81 ± 3.75	55.32 ± 1.52	83.96 ± 4.35
NDFS	70.05 ± 2.00	54.67 ± 2.35	70.42 ± 1.14	28.16 ± 4.45	58.78 ± 0.99	78.81 ± 3.99
GLSPFS	68.80 ± 1.07	56.18 ± 3.40	73.05 ± 1.52	52.23 ± 4.42	60.33 ± 1.65	82.99 ± 4.73
MFFS	72.64 ± 1.73	56.17 ± 4.47	70.65 ± 1.25	40.95 ± 3.39	59.11 ± 0.76	81.09 ± 4.12
GLoSS	74.28 ± 1.25	58.87 ± 1.65	73.02 ± 2.02	55.76 ± 4.56	61.29 ± 1.25	85.65 ± 4.15
GLoRSS	**75.26 ± 1.39**	**61.66 ± 1.58**	**73.49 ± 1.46**	**59.26 ± 2.65**	**63.28± 1.11**	**86.46 ± 4.16**

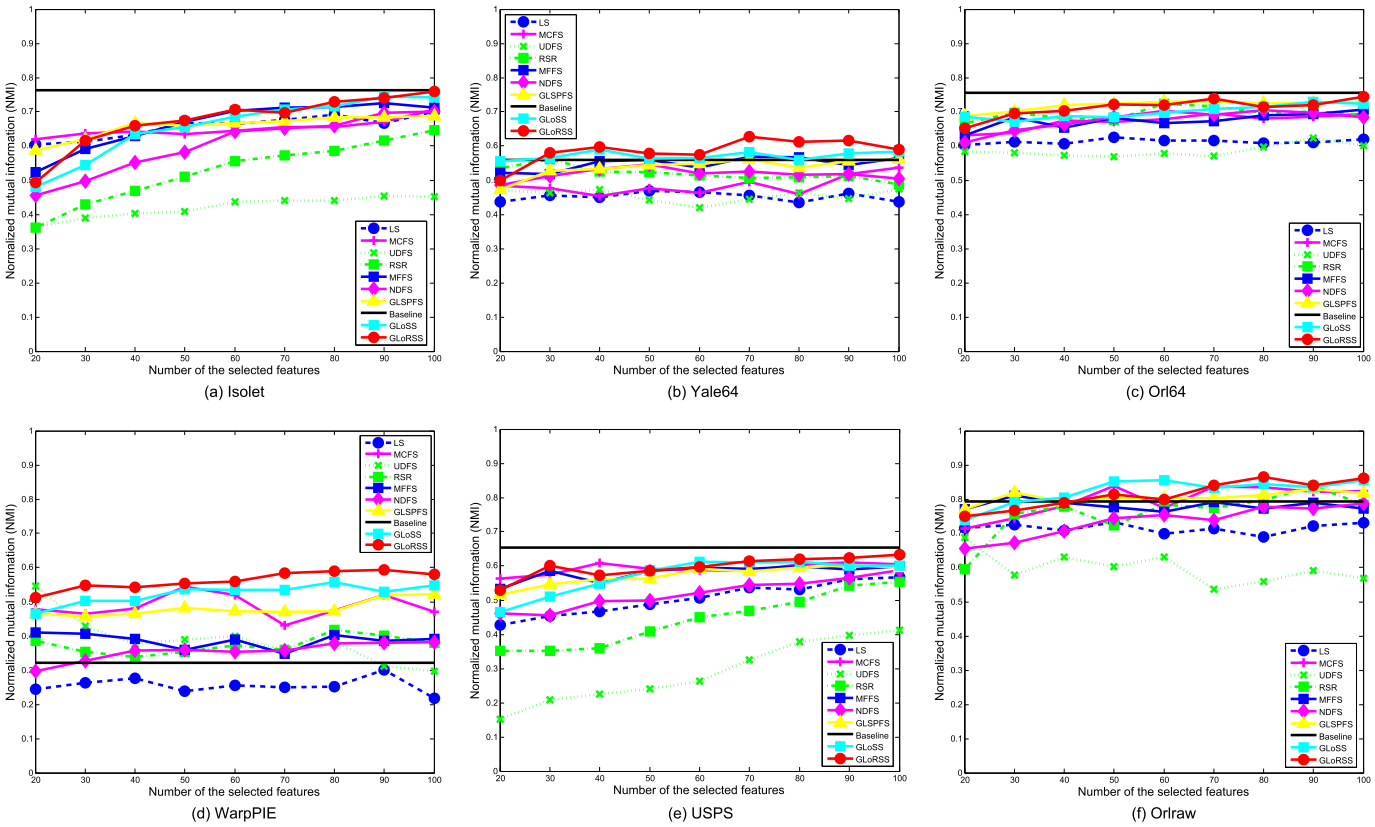


Fig. 7. Normalized mutual information (NMI) of using all features and selected features by different feature selection algorithms.

accuracy (ACC) and normalized mutual information (NMI), which are defined below. The ACC is computed as follows:

$$ACC = \frac{\sum_{i=1}^{n} \delta(q_i, map(p_i))}{n}, \tag{25}$$

where p_i and q_i are the predicted and true labels of the ith sample, and $\delta(\cdot\,,\,\cdot)$ is the indicator function where $\delta(a, b) = 1$ if $a = b$ and $\delta(a, b) = 0$ otherwise, and $map(\cdot)$ is a permutation mapping, which are realized by Kuhn-munkres algorithm [41]. The higher the value of ACC is, the better the clustering result is. The normalized mutual information (NMI) is another metric to measure the similarity of two clustering results. For two label vectors P and Q, it is defined as

$$NMI(P, Q) = \frac{I(P, Q)}{\sqrt{H(P)H(Q)}}, \tag{26}$$

where $I(P, Q)$ is the mutual information of P and Q, and $H(P)$ and $H(Q)$ are the entropies of P and Q [42]. In our experiments, P is obtained by the clustering labels using the selected features and Q is the true data labels given by the dataset. Higher value of NMI means better clustering result.

2) Performance Comparison: Tables III and IV present the ACC and NMI values produced by different methods. For each method, the number of selected features is varied among $\{20, 30, 40, \ldots, 100\}$ and the best result is reported. From the tables, we see that the proposed GLoRSS method performs the best among all the compared methods. It addition, note that the GLoRSS method outperforms GLoSS method for all datasets, and this is possiblly due to that the correntropy induced metric term is more robust than traditional Euclidean distance, thus it can alleviate the effects of noise existing in the data.

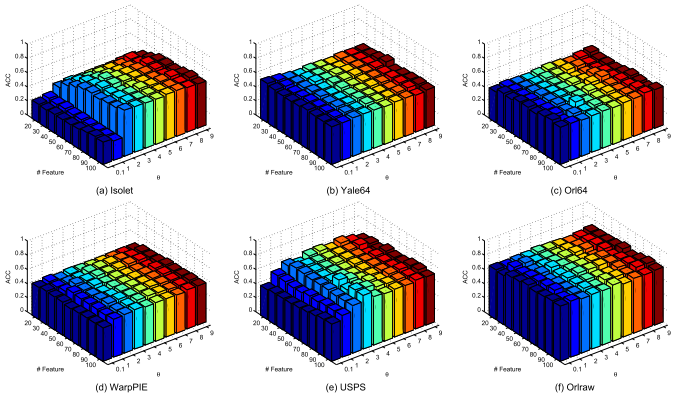


Fig. 8. The clustering accuracy (ACC) produced by GLoRSS with different κ and θ.

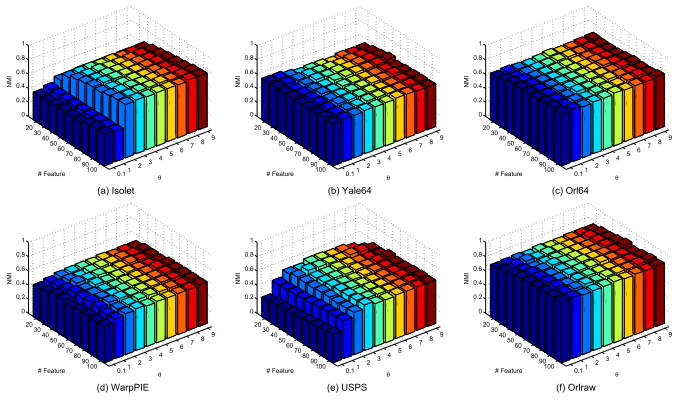


Fig. 9. The normalized mutual information (NMI) produced by GLoRSS with different κ and θ

To verify the effect of our proposed feature selection method, the results are compared between using all features and selected features given by different methods. Fig. 6 plots the ACC value and Fig. 7 the NMI value with respect to the number of selected features. The baseline corresponds to the results using all features for clustering. From the figures, we see that the proposed GLoRSS method exhibits the best in most cases, and it can produce comparable and even better clustering results than those using all features. Therefore, the high dimensional data do contain redundancy, and feature selection can eliminate the redundancy. Furthermore, it is noted that using fewer features can save the clustering time for the clustering problem, and thus feature selection can improve both clustering accuracy and efficiency.

3) Sensitivity of Parameters: To further study the proposed GLoRSS method, the sensitivity with regard to κ and θ in (24) is studied. The parameter β is fixed as their best value in GLoSS method. Figs. 8 and 9 plot the ACC and NMI values given by GLoRSS for different κ and θ's. From the figures, we see that except for Isolet and USPS, GLoRSS performs stably well for different combinations of κ and θ, and thus the users can choose the parameters within a large interval to have satisfactory feature selection performance.

VI. Conclusion

We have proposed a novel robust unsupervised feature selection model that is based on the MCC and sparse subspace learning. This model is derived from a combinatorial model by replacing the Euclidean distance measure with the correntropy measure and also relaxing the 0-1 variables to continuous ones. A half-quadratic accelerated block coordinate update iterative algorithm has been derived to solve the proposed model, and the convergence property has been proved. Experimental studies on noisy face images reconstruction have confirmed the robustness of the proposed model with respect to noises and outliers. In unsupervised feature selection experiments, the proposed method outperforms several state-of-the-art methods on six real-world data.

Appendix A
Lipschitz Constant Derivation

In this section, we derive the Lipschitz constant of $\nabla_W f(W, H^k, \mathbf{y}^k)$ with respect to W. By matrix calculus, it is not difficult to obtain that

$$\nabla_W f(W, H, \mathbf{y}^k) = -(X^k)^\top (X^k W H - X^k) H^\top - \mu X^\top L X W, \tag{27}$$

where X^k is defined in (16). In addition, for any $\tilde{W}$ and $\hat{W}$, we have

$$\begin{aligned}
&\|\nabla_W f(\tilde{W}, H, \mathbf{y}^k) - \nabla_W f(\hat{W}, H, \mathbf{y}^k)\|_F \\
&= \|(X^k)^\top X^k \hat{W} H H^\top + \mu X^\top L X \hat{W} - (X^k)^\top X^k \tilde{W} H H^\top \\
&\quad - \mu X^\top L X \tilde{W}\|_F \\
&\le \|(X^k)^\top X^k \hat{W} H H^\top - (X^k)^\top X^k \tilde{W} H H^\top\|_F \\
&\quad + \mu \|X^\top L X \hat{W} - X^\top L X \tilde{W}\|_F \\
&= \|(X^k)^\top X^k (\hat{W} - \tilde{W}) H H^\top\|_F + \mu \|X^\top L X (\hat{W} - \tilde{W})\|_F \\
&\le \|(X^k)^\top X^k\|_2 \|\hat{W} - \tilde{W}\|_F \|H H^\top\|_2 \\
&\quad + \mu \|X^\top L X\|_2 \|\hat{W} - \tilde{W}\|_F \\
&= \left(\|(X^k)^\top X^k\|_2 \|H H^\top\|_2 + \mu \|X^\top L X\|_2\right) \|\hat{W} - \tilde{W}\|_F,
\end{aligned}$$

where $\|A\|_2$ denotes the spectral norm and equals the largest singular value of A, the first inequality follows from the triangle inequality, and the last inequality uses the fact that $\|AB\|_F \le \|A\|_2 \|B\|_F$ for any matrices A and B of appropriate sizes. Hence, by definition, $\|(X^k)^\top X^k\|_2 \|(H^k H^k)^\top\|_2 + \mu \|X^\top L X\|_2$ is a Lipschitz constant of $\nabla_W f(W, H^k, \mathbf{y}^k)$ with respect to W.

Appendix B
Convergence Analysis

In the following, we prove Theorems 1 and 2. For simplicity, we assume $\omega_k = 0$, i.e., no extrapolation. The general case for positive ω_k can be shown similarly. Throughout this section, we define

$$\psi(W, H, y) = f(W, H, y) - r_\beta(W) - \iota_{\mathbb{R}_+^{d \times K}}. \tag{28}$$

A. Proof of Theorem 1

It is well-known that (c.f. [32, Lemma 2.1])

$$\begin{aligned}
&\psi(W^{k+1}, H^k, \mathbf{y}^k) - \psi(W^k, H^k, \mathbf{y}^k) \\
&\qquad \ge \frac{L_w^k}{2} \|W^{k+1} - W^k\|_F^2 \ge \frac{L_w}{2} \|W^{k+1} - W^k\|_F^2,
\end{aligned} \tag{29}$$

where $L_w = \mu\|X^\top LX\|_2$ is a lower bound of L_w^k. From [43, Lemma 3.1], we have

$$\frac{1}{2}\|X^k - X^kW^{k+1}H^k\|_F^2 - \frac{1}{2}\|X^k - X^kW^{k+1}H^{k+1}\|_F^2 = \frac{1}{2}\|X^kW^{k+1}H^k - X^kW^{k+1}H^{k+1}\|_F^2, \quad (30)$$

and

$$X^kW^{k+1}H^k - X^kW^{k+1}H^{k+1} = U^{k+1}(U^{k+1})^\top\left(X^kW^{k+1}H^k - X^k\right), \quad (31)$$

where U^{k+1} contains the left leading singular vectors of X^kW^{k+1} corresponding to its nonzero singular values. In addition, note that $y_i^k \geq -1$, $\forall i, \forall k$, and $f(W, H, \mathbf{y})$ is strongly concave with modulus $\frac{1}{2}$ with respect to $\mathbf{y}$ restricted on the set $[-1, 0)^n$. Hence,

$$\psi(W^{k+1}, H^{k+1}, \mathbf{y}^{k+1}) - \psi(W^{k+1}, H^{k+1}, \mathbf{y}^k) \geq \frac{1}{4}\|\mathbf{y}^{k+1} - \mathbf{y}^k\|_2^2. \quad (32)$$

Summing up (29), (30), and (32) gives

$$\psi(W^{k+1}, H^{k+1}, \mathbf{y}^{k+1}) - \psi(W^k, H^k, \mathbf{y}^k) \geq \frac{L_w}{2}\|W^{k+1} - W^k\|_F^2 + \frac{1}{2}\|X^kW^{k+1}H^k - X^kW^{k+1}H^{k+1}\|_F^2 + \frac{1}{4}\|\mathbf{y}^{k+1} - \mathbf{y}^k\|_2^2,$$

which together with the upper boundedness of ψ implies that

$$\lim_{k\to\infty} W^{k+1} - W^k = 0, \quad (33a)$$

$$\lim_{k\to\infty} X^kW^{k+1}H^k - X^kW^{k+1}H^{k+1} = 0, \quad (33b)$$

$$\lim_{k\to\infty} \mathbf{y}^{k+1} - \mathbf{y}^k = 0. \quad (33c)$$

We have from the definition of X^k in (16) and (33c) that $X^{k+1} - X^k \to 0$ as $k \to \infty$. Hence, together from (31) and (33b), it follows that

$$\lim_{k\to\infty} U^k(U^k)^\top\left(X^kW^kH^k - X^k\right) = 0.$$

Since $r_\beta(W)$ is coercive, $\{W^k\}$ and thus $\{X^kW^k\}$ must be bounded. From $(X^kW^k)^\top = (X^kW^k)^\top U^k(U^k)^\top$, left multiplying $(X^kW^k)^\top$ in the above equation gives

$$\lim_{k\to\infty} (X^kW^k)^\top(X^kW^kH^k - X^k) = 0. \quad (34)$$

Assume $(\bar{W}, \bar{H})$ is a finite limit point of $\{(W^k, H^k)\}_{k\geq 1}$. Due to the continuity of the mapping $\mathcal{T}$ defined in (19), there is a subsequence $\{(W^k, H^k, \mathbf{y}^k)\}_{k\in\mathcal{K}}$ convergent to $(\bar{W}, \bar{H}, \bar{\mathbf{y}})$. Letting $\mathcal{K} \ni k \to \infty$ in (34) yields

$$\nabla_H f(\bar{W}, \bar{H}, \bar{\mathbf{y}}) = (\bar{X}\bar{W})^\top(\bar{X}\bar{W}\bar{H} - \bar{X}) = 0.$$

If necessary, taking another subsequence, we can assume that L_w^k converges to some number $\bar{L}_w$ as $\mathcal{K} \ni k \to \infty$. Hence, letting $\mathcal{K} \ni k \to \infty$ in (12a), we have

$$\bar{W} = \operatorname*{argmax}_{W\in\mathbb{R}^{d\times K}} \langle\nabla_W f(\bar{W}, \bar{H}, \bar{\mathbf{y}}), W - \bar{W}\rangle - \frac{\bar{L}_w}{2}\|W - \bar{W}\|_F^2 - r_\beta(W),$$

and thus $0 \in \partial\psi(\bar{W}, \bar{H}, \bar{\mathbf{y}})$. From (23), we conclude that $(\bar{W}, \bar{H})$ is a critical point of Ψ, and thus complete the proof of Theorem 1.

B. Proof of Theorem 2

Denote $Z = (W, H)$. Let $\bar{\sigma}_{\min} = \sigma_{\min}(\bar{X}\bar{W}) > 0$ be the smallest singular value of $\bar{X}\bar{W}$. By the continuity of the singular value function and spectral norm of a matrix, there is $\delta_1 > 0$ such that

$$\sigma_{\min}(XW) \geq \frac{\sigma_{\min}(\bar{X}\bar{W})}{2}, \text{ and}$$

$$\|XW\|_2 \leq 2\|\bar{X}\bar{W}\|_2, \ \forall\ (X, W) \in \mathcal{B}_{\delta_1}(\bar{X}, \bar{W}), \quad (35)$$

$$\|HH^\top\|_2 \leq 2\|\bar{H}\bar{H}^\top\|_2, \ \forall H \in \mathcal{B}_{\delta_1}(\bar{H}), \quad (36)$$

$$\|X^\top X\|_2 \leq 2\|\bar{X}^\top\bar{X}\|_2, \ \forall X \in \mathcal{B}_{\delta_1}(\bar{X}), \quad (37)$$

where $\mathcal{B}_\delta(\bar{Z}) := \{Z : \|Z - \bar{Z}\|_F \leq \delta\}$ denotes the δ-ball centered at $\bar{Z}$. By the continuity of $\mathcal{T}$, there is $\delta_2 > 0$ such that

$$\mathrm{Diag}\left(\sqrt{\frac{-\mathcal{T}(Z)}{2\sigma^2}}\right) \in \mathcal{B}_{\frac{\delta_1}{4}}(\bar{X}), \ \forall Z \in \mathcal{B}_{\delta_2}(\bar{Z}). \quad (38)$$

Note that the function $\frac{1}{2}\sum_{i=1}^n \exp\left(\frac{-\|\mathbf{p}_i^\top - \mathbf{p}_i^\top WH\|_2^2}{2\sigma^2}\right)$ is real analytic, and the other terms in $\Psi(Z)$ are all semi-algebraic and continuous. Hence, according to [32], $\Psi(Z)$ satisfies the so-called Kurdyka-Łojasiewicz property (c.f. [44]): in a neighborhood $\mathcal{B}_\rho(\bar{Z})$, there exists a function $\phi(s) = c \cdot s^{1-\theta}$ for some $c > 0$ and $0 \leq \theta < 1$ such that

$$\phi'(|\Psi(Z) - \Psi(\bar{Z})|)\mathrm{dist}(0, \partial\Psi(Z)) \geq 1, \text{ for any } Z \in \mathcal{B}_\rho(\bar{Z}) \cap \mathrm{dom}(\Psi) \text{ and } \Psi(Z) \neq \Psi(\bar{Z}), \quad (39)$$

where $\mathrm{dist}(0, \partial\Psi(Z)) = \min_Y\{\|Y\|_F : Y \in \partial\Psi(Z)\}$. In the remaining analysis, we take

$$\rho < \min\left(\frac{\delta_1}{4}, \delta_2\right).$$

Denote $\Psi_k = \Psi(\bar{Z}) - \Psi(Z^k)$ and $\phi_k = \phi(\Psi_k)$. Then Ψ_k is nonnegative and nonincreasing with respect to k. In addition, let L_G be the uniform Lipschitz constant of $\mathcal{T}$ and ∇F within $\mathcal{B}_{\frac{\delta_1}{4}}(\bar{Z})$ and also that of $\nabla f(Z, \mathbf{y})$ within $\mathcal{B}_{\frac{\delta_1}{4}}(\bar{Z}) \times [-1, 0)^n$. Note that L_G must be finite since F, f and $\mathcal{T}$ are second-order differentiable.

From (35), we have that if $(X^k, W^{k+1}) \in \mathcal{B}_{\delta_1}(\bar{X}, \bar{W})$, then

$$\psi(W^{k+1}, H^{k+1}, \mathbf{y}^k) - \psi(W^{k+1}, H^k, \mathbf{y}^k) \geq \frac{\sigma_{\min}^2(X^kW^{k+1})}{2}\|H^{k+1} - H^k\|_F^2 \geq \frac{\bar{\sigma}_{\min}^2}{8}\|H^{k+1} - H^k\|_F^2, \quad (40)$$

which together with (29) and the fact $\Psi(Z^{k+1}) = \psi(Z^{k+1}, \mathbf{y}^{k+1})$ implies

$$\Psi_k - \Psi_{k+1} = \Psi(Z^{k+1}) - \Psi(Z^k) \geq \frac{L_w}{2}\|W^{k+1} - W^k\|_F^2 + \frac{\bar{\sigma}_{\min}^2}{8}\|H^{k+1} - H^k\|_F^2 \geq \min\left(\frac{L_w}{2}, \frac{\bar{\sigma}_{\min}^2}{8}\right)\|Z^{k+1} - Z^k\|_F^2. \quad (41)$$

Since $\bar{Z}$ is a limit point of $\{Z^k\}_{k\geq 1}$, there must exist k_0 such that Z^{k_0} is sufficiently close to $\bar{Z}$. In addition, note that $\{Z^k\}_{k\geq 1}$ converges *if and only if* $\{Z^k\}_{k\geq k_0}$ converges. Hence, without loss of generality, we can assume Z^0 is sufficiently close to $\bar{Z}$ such that

$$2\sqrt{\frac{\Psi_0}{C_2}} + \|Z^0 - \bar{Z}\|_F + \frac{C_1}{2C_2}\phi_0 < \rho, \tag{42a}$$

$$\sqrt{\frac{\Psi_0}{C_2}} < \frac{\delta_1}{4} - \rho. \tag{42b}$$

where $C_1 = L_G + L_G^2 + 4\|\bar{X}^\top \bar{X}\|_2 \|\bar{H}\bar{H}^\top\|_2 + \mu\|X^\top L X\|_2$ and $C_2 = \min\left(\frac{L_w}{2}, \frac{\bar{\sigma}_{\min}^2}{8}\right)$.

From (29), we have $\|W^1 - W^0\|_F \leq \sqrt{\frac{2}{L_w}\Psi_0}$, and thus

$$\begin{aligned}\|W^1 - \bar{W}\|_F &\leq \|W^1 - W^0\|_F + \|W^0 - \bar{W}\|_F \\ &\overset{(42a)}{\leq} \sqrt{\frac{2}{L_w}\Psi_0} + \rho \overset{(42b)}{\leq} \frac{\delta_1}{4}.\end{aligned}$$

In addition, since $\|Z^0 - \bar{Z}\|_F \leq \rho < \delta_2$, we have from (38) that $X^0 \in \mathcal{B}_{\frac{\delta_1}{4}}(\bar{X})$. Hence, $(X^0, W^1) \in \mathcal{B}_{\delta_1}(\bar{X}, \bar{W})$, and (40) holds for $k = 0$. From (41) and noting $\Psi_k \geq 0$, we have $\|Z^1 - Z^0\|_F \leq \sqrt{\frac{\Psi_0}{C_2}}$ and thus

$$\|Z^1 - \bar{Z}\|_F \leq \|Z^1 - Z^0\|_F + \|Z^0 - \bar{Z}\|_F \leq \rho,$$

which implies $Z^1 \in \mathcal{B}_\rho(\bar{Z})$.

Assume $Z^k \in \mathcal{B}_\rho(\bar{Z})$. By the same arguments as above, we have $(X^k, W^{k+1}) \in \mathcal{B}_{\delta_1}(\bar{X}, \bar{W})$, and thus (40) holds. Hence, $\|Z^{k+1} - Z^k\|_F \leq \sqrt{\frac{\Psi_0}{C_2}}$, and

$$\begin{aligned}\|Z^{k+1} - \bar{Z}\|_F &\leq \|Z^{k+1} - Z^k\|_F + \|Z^k - \bar{Z}\|_F \\ &\leq \sqrt{\frac{\Psi_0}{C_2}} + \rho < \frac{\delta_1}{4}.\end{aligned} \tag{43}$$

From the updates in (12a) and (12b), it follows that

$$\begin{aligned}0 \in\ & \nabla_W f(W^k, H^k, \mathbf{y}^k) - L_w^k(W^{k+1} - W^k) \\ & -\partial\left(r_\beta(W^{k+1}) + \iota_{\mathbb{R}_+^{d\times K}}(W^{k+1})\right), \\ 0 =\ & \nabla_H f(W^{k+1}, H^{k+1}, \mathbf{y}^k),\end{aligned}$$

which indicates that

$$\begin{aligned}&\nabla_W F(Z^{k+1}) - \nabla_W f(W^k, H^k, \mathbf{y}^k) + L_w^k(W^{k+1} - W^k) \\ &\in \nabla_W F(Z^{k+1}) - \partial\left(r_\beta(W^{k+1}) + \iota_{\mathbb{R}_+^{d\times K}}(W^{k+1})\right), \\ &\nabla_H F(Z^{k+1}) - \nabla_H f(W^{k+1}, H^{k+1}, \mathbf{y}^k) = \nabla_H F(Z^{k+1}).\end{aligned}$$

Hence,

$$\begin{aligned}&\mathrm{dist}(0, \partial\Psi(Z^{k+1})) \\ &\leq \|\nabla_W F(Z^{k+1}) - \nabla_W f(W^k, H^k, \mathbf{y}^k) + L_w^k(W^{k+1} - W^k)\|_F \\ &\quad + \|\nabla_H F(Z^{k+1}) - \nabla_H f(W^{k+1}, H^{k+1}, \mathbf{y}^k)\|_F \\ &= \|\nabla_W F(Z^{k+1}) - \nabla_W F(Z^k) + L_w^k(W^{k+1} - W^k)\|_F \\ &\quad + \|\nabla_H f(W^{k+1}, H^{k+1}, \mathbf{y}^{k+1}) - \nabla_H f(W^{k+1}, H^{k+1}, \mathbf{y}^k)\|_F \\ &\leq L_G\|Z^{k+1} - Z^k\|_F + \left(4\|\bar{X}^\top \bar{X}\|_2\|\bar{H}\bar{H}^\top\|_2 + \mu\|X^\top L X\|_2\right) \\ &\quad * \|W^{k+1} - W^k\|_F + L_G\|\mathbf{y}^{k+1} - \mathbf{y}^k\|_2 \\ &\leq C_1\|Z^{k+1} - Z^k\|_F,\end{aligned} \tag{44}$$

where we have used (23) to have the equality, and in the second inequality, we note that $4\|\bar{X}^\top \bar{X}\|_2\|\bar{H}\bar{H}^\top\|_2 + \mu\|X^\top L X\|_2$ is an upper bound of L_w^k from (36) and (37).

In addition, we have

$$\begin{aligned}\phi_k - \phi_{k+1} &\geq \phi'(\Psi_k)(\Psi_k - \Psi_{k+1}) \\ &\overset{(39)}{\geq} \frac{\Psi_k - \Psi_{k+1}}{\mathrm{dist}(0, \partial\Psi(Z^k))} \\ &\overset{(44)}{\geq} \frac{\Psi_k - \Psi_{k+1}}{C_1\|Z^k - Z^{k-1}\|} \\ &\overset{(41)}{\geq} \frac{C_2\|Z^{k+1} - Z^k\|_F^2}{C_1\|Z^k - Z^{k-1}\|_F}\end{aligned} \tag{45}$$

where the first inequality is from the concavity of ϕ. From (45), it follows that

$$\begin{aligned}&C_2\|Z^{k+1} - Z^k\|_F^2 \leq C_1\|Z^k - Z^{k-1}\|_F(\phi_k - \phi_{k+1}) \\ \Rightarrow\ &\sqrt{C_2}\|Z^{k+1} - Z^k\|_F \leq \sqrt{C_1\|Z^k - Z^{k-1}\|_F(\phi_k - \phi_{k+1})} \\ \Rightarrow\ &\sqrt{C_2}\|Z^{k+1} - Z^k\|_F \leq \frac{\sqrt{C_2}}{2}\|Z^k - Z^{k-1}\|_F \\ &\qquad + \frac{C_1}{2\sqrt{C_2}}(\phi_k - \phi_{k+1}).\end{aligned}$$

Summing the above inequality over k and arranging terms give

$$\sum_{k=1}^{K}\|Z^{k+1} - Z^k\|_F \leq \|Z^1 - Z^0\|_F + \frac{C_1}{2C_2}(\phi_1 - \phi_{K+1}). \tag{46}$$

Hence,

$$\begin{aligned}\|Z^{K+1} - \bar{Z}\|_F &\leq \sum_{k=1}^{K}\|Z^{k+1} - Z^k\|_F + \|Z^1 - \bar{Z}\|_F \\ &\leq \|Z^1 - \bar{Z}\|_F + \|Z^1 - Z^0\|_F + \frac{C_1}{2C_2}\phi_0 \\ &\leq 2\|Z^1 - Z^0\|_F + \|Z^0 - \bar{Z}\|_F + \frac{C_1}{2C_2}\phi_0 \\ &\leq 2\sqrt{\frac{\Psi_0}{C_2}} + \|Z^0 - \bar{Z}\|_F + \frac{C_1}{2C_2}\phi_0 \\ &\overset{(42a)}{\leq} \rho,\end{aligned}$$

which indicates $Z^{K+1} \in \mathcal{B}(\bar{Z}, \rho)$. By induction, we have $Z^k \in \mathcal{B}(\bar{Z}, \rho)$, $\forall k$, and thus (46) holds for all K. Therefore, $\{Z^k\}_{k=1}^\infty$ is a Cauchy sequence and converges. Since $\bar{Z}$ is a limit point, it must hold that $\lim_{k\to\infty} Z^k = \bar{Z}$. This completes the proof.

REFERENCES

[1] D. Chen, X. Cao, F. Wen, and J. Sun, "Blessing of dimensionality: High-dimensional feature and its efficient compression for face verification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 3025–3032.

[2] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *J. Mach. Learn. Res.*, vol. 3, pp. 1157–1182, Jan. 2003.

[3] G. Herman, B. Zhang, Y. Wang, G. Ye, and F. Chen, "Mutual information-based method for selecting informative feature sets," *Pattern Recognit.*, vol. 46, no. 12, pp. 3315–3327, 2013.

[4] Z. Zhao, L. Wang, H. Liu, and J. Ye, "On similarity preserving feature selection," *IEEE Trans. Knowl. Data Eng.*, vol. 25, no. 3, pp. 619–632, Mar. 2013.

[5] L. Wang, H. Cheng, Z. Liu, and C. Zhu, "A robust elastic net approach for feature learning," *J. Vis. Commun. Image Represent.*, vol. 25, no. 2, pp. 313–321, 2014.

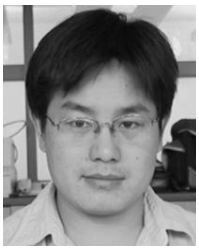
[6] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*. Hoboken, NJ, USA: Wiley, 2012.

[7] X. He, D. Cai, S. Yan, and H.-J. Zhang, "Neighborhood preserving embedding," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2005, pp. 1208–1213.
[8] D. Cai, C. Zhang, and X. He, "Unsupervised feature selection for multi-cluster data," in *Proc. 16th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2010, pp. 333–342.
[9] Q. Gu, Z. Li, and J. Han, "Joint feature selection and subspace learning," in *Proc. Int. Joint Conf. Artif. Intell. (IJCAI)*, 2011, vol. 22. no. 1, p. 1294.
[10] S. Wang, W. Pedrycz, Q. Zhu, and W. Zhu, "Subspace learning for unsupervised feature selection via matrix factorization," *Pattern Recognit.*, vol. 48, no. 1, pp. 10–19, 2015.
[11] N. Zhou, Y. Xu, H. Cheng, J. Fang, and W. Pedrycz, "Global and local structure preserving sparse subspace learning: An iterative approach to unsupervised feature selection," *Pattern Recognit.*, vol. 53, pp. 87–101, May 2016.
[12] N. Zhou, H. Cheng, W. Pedrycz, Y. Zhang, and H. Liu, "Discriminative sparse subspace learning and its application to unsupervised feature selection," *ISA Trans.*, vol. 61, pp. 104–118, Mar. 2016.
[13] M. Qian and C. Zhai, "Robust unsupervised feature selection," in *Proc. 23rd Int. Joint Conf. Artif. Intell.*, 2013, pp. 1621–1627.
[14] W. Liu, P. P. Pokharel, and J. C. Príncipe, "Correntropy: Properties and applications in non-Gaussian signal processing," *IEEE Trans. Signal Process.*, vol. 55, no. 11, pp. 5286–5298, Nov. 2007.
[15] J. C. Príncipe, *Information Theoretic Learning: Renyi's Entropy and Kernel Perspectives*. New York, NY, USA: Springer, 2010.
[16] B. Chen and J. C. Príncipe, "Maximum correntropy estimation is a smoothed map estimation," *IEEE Signal Process. Lett.*, vol. 19, no. 8, pp. 491–494, Aug. 2012.
[17] B. Chen, J. Wang, H. Zhao, N. Zheng, and J. C. Príncipe, "Convergence of a fixed-point algorithm under maximum correntropy criterion," *IEEE Signal Process. Lett.*, vol. 22, no. 10, pp. 1723–1727, Oct. 2015.
[18] B. Chen, L. Xing, H. Zhao, N. Zheng, and J. C. Príncipe, "Generalized correntropy for robust adaptive filtering," *IEEE Trans. Signal Process.*, vol. 64, no. 13, pp. 3376–3387, Jul. 2016.
[19] Z. Wu, S. Peng, B. Chen, and H. Zhao, "Robust Hammerstein adaptive filtering under maximum correntropy criterion," *Entropy*, vol. 17, no. 10, pp. 7149–7166, 2015.
[20] B. Chen, L. Xing, J. Liang, N. Zheng, and J. C. Príncipe, "Steady-state mean-square error analysis for adaptive filtering under the maximum correntropy criterion," *IEEE Signal Process. Lett.*, vol. 21, no. 7, pp. 880–884, Jul. 2014.
[21] B. Chen, X. Liu, H. Zhao, and J. C. Príncipe, "Maximum correntropy Kalman filter," *Automatica*, vol. 76, pp. 70–77, Feb. 2017.
[22] J. J.-Y. Wang, X. Wang, and X. Gao, "Non-negative matrix factorization by maximizing correntropy for cancer clustering," *BMC Bioinformat.*, vol. 14, no. 1, p. 107, 2013.
[23] R. He, W.-S. Zheng, and B.-G. Hu, "Maximum correntropy criterion for robust face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 8, pp. 1561–1576, Aug. 2011.
[24] R. He, T. Tan, L. Wang, and W.-S. Zheng, "$\ell_{2,1}$ Regularized correntropy for robust feature selection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 2504–2511.
[25] P. Zhu, W. Zuo, L. Zhang, Q. Hu, and S. C. K. Shiu, "Unsupervised feature selection by regularized self-representation," *Pattern Recognit.*, vol. 48, no. 2, pp. 438–446, 2015.
[26] V. Vapnik, *The Nature of Statistical Learning Theory*. New York, NY, USA: Springer, 2013.
[27] P. J. Huber, *Robust Statistics*. Berlin, Germany: Springer, 2011.
[28] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 8, pp. 888–905, Aug. 2000.
[29] A. Y. Ng, M. I. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," in *Proc. Adv. Neural Inf. Process. Syst.*, 2002, pp. 849–856.
[30] X. Liu, L. Wang, J. Zhang, J. Yin, and H. Liu, "Global and local structure preservation for feature selection," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 25, no. 6, pp. 1083–1095, Jun. 2013.
[31] D. Geman and G. Reynolds, "Constrained restoration and the recovery of discontinuities," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 14, no. 3, pp. 367–383, Mar. 1992.
[32] Y. Xu and W. Yin, "A block coordinate descent method for regularized multiconvex optimization with applications to nonnegative tensor factorization and completion," *SIAM J. Imag. Sci.*, vol. 6, no. 3, pp. 1758–1789, 2013.
[33] Z. Peng, T. Wu, Y. Xu, M. Yan, and W. Yin, "Coordinate friendly structures, algorithms and applications," *Ann. Math. Sci. Appl.*, vol. 1, no. 1, pp. 57–119, 2016.
[34] H.-J. M. Shi, S. Tu, Y. Xu, and W. Yin. (Sep. 2016). "A primer on coordinate descent algorithms." [Online]. Available: https://arxiv.org/abs/1610.00040
[35] Y. Xu and W. Yin. (2014). "A globally convergent algorithm for nonconvex optimization based on block coordinate update." [Online]. Available: https://arxiv.org/abs/1410.1386
[36] Y. Xu, "Alternating proximal gradient method for sparse nonnegative Tucker decomposition," *Math. Programm. Comput.*, vol. 7, no. 1, pp. 39–70, 2015.
[37] Y. Xu. (2015). "On the convergence of higher-order orthogonality iteration." [Online]. Available: https://arxiv.org/abs/1504.00538
[38] X. He, D. Cai, and P. Niyogi, "Laplacian score for feature selection," in *Proc. Adv. Neural Inf. Process. Syst.*, 2005, pp. 507–514.
[39] Y. Yang, H. T. Shen, Z. Ma, Z. Huang, and X. Zhou, "$\ell_{2,1}$-Norm regularized discriminative feature selection for unsupervised learning," in *Proc. Int. Joint Conf. Artif. Intell. (IJCAI)*, 2011, vol. 22. no. 1, p. 1589.
[40] Z. Li, Y. Yang, J. Liu, X. Zhou, and H. Lu, "Unsupervised feature selection using nonnegative spectral analysis," in *Proc. AAAI*, 2012, pp. 1026–1032.
[41] L. Lovász and M. D. Plummer, *Matching Theory*, vol. 367. Providence, RI, USA: AMS, 2009.
[42] R. M. Gray, *Entropy and Information Theory*. New York, NY, USA: Springer, 2011.
[43] Y. Xu, "Fast algorithms for higher-order singular value decomposition from incomplete data," *J. Comput. Math.*, vol. 35, no. 4, pp. 395–420, 2017.
[44] J. Bolte, A. Daniilidis, and A. Lewis, "The Łojasiewicz inequality for nonsmooth subanalytic functions with applications to subgradient dynamical systems," *SIAM J. Optim.*, vol. 17, no. 4, pp. 1205–1223, 2007.

Nan Zhou received the Ph.D. degree from the Center for Robotics, University of Electronic Science and Technology of China, Chengdu. He is currently a Lecturer with the College of Control Engineering, Chengdu University of Information Technology, Chengdu. His research interests include machine learning, feature learning, computer vision, and sparse modeling.

Yangyang Xu received the Ph.D. degree from the Department of Computational and Applied Mathematics, Rice University, in 2014. He is currently an Assistant Professor with the Department of Mathematical Sciences, Rensselaer Polytechnic Institute. He has published over 20 papers in top-ranking journals. His research interests are optimization theory and methods and their applications in areas, such as machine learning, statistics, and signal processing. He has developed optimization algorithms for compressed sensing, matrix completion, and tensor factorization and learning. His current research focuses on first-order methods, operator splitting, stochastic optimization, and high performance parallel computing.

Hong Cheng (SM'14) is currently a Professor with the School of Automation, University of Electronic Science and Technology of China (UESTC), Chengdu. He is also the Founding Director of Pattern Recognition and Machine Intelligence Laboratory, UESTC. He has over 50 academic publications including two books. He has been a Senior Member of ACM, and an Associate Editor of the IEEE *Computational Intelligence Magazine*. He is a Reviewer for many important journals and conferences, such as the IEEE TITS, MAV, CVPR, ICCV, ITSC, IVS, and ACCV.



Zejian Yuan received the M.S. degree in electronic engineering from the Xi'an University of Technology in 1999, and the Ph.D. degree in pattern recognition and intelligent system from Xi'an Jiaotong University, China, in 2003. He was a Visiting Scholar at the Advanced Robotics Laboratory, Chinese University of Hong Kong from 2008 to 2009. He is currently an Associate Professor with the Department of Automatic Engineering, Xi'an Jiaotong University, and a member of the Chinese Association of Robotics. His research interests include image processing, pattern recognition, and machine learning methods in computer vision.

Badong Chen (SM'13) received the Ph.D. degree in computer science and technology from Tsinghua University in 2008. He is currently a Professor with the Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University. He has published two books, four chapters, and over 100 papers in various journals and conference proceedings. His research interests are in signal processing, information theory, machine learning, and their applications in cognitive science and engineering. He is an Associate Editor of the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS and the *Journal of the Franklin Institute*, and has been on the Editorial Board of Entropy.