Correntropy Induced L2 Graph for Robust Subspace Clustering

Canyi Lu[1], Jinhui Tang[2], Min Lin[1], Liang Lin[3], Shuicheng Yan[1], Zhouchen Lin[4,*]
[1] Department of Electrical and Computer Engineering, National University of Singapore
[2] School of Computer Science, Nanjing University of Science and Technology
[3] School of Software, Sun Yat-Sen University
[4] Key Laboratory of Machine Perception (MOE), School of EECS, Peking University
canyilu@gmail.com, jinhuitang@mail.njust.edu.cn, mavenlin@gmail.com
linliang@ieee.org, eleyans@nus.edu.sg, zlin@pku.edu.cn

Abstract

In this paper, we study the robust subspace clustering problem, which aims to cluster the given possibly noisy data points into their underlying subspaces. A large pool of previous subspace clustering methods focus on the graph construction by different regularization of the representation coefficient. We instead focus on the robustness of the model to non-Gaussian noises. We propose a new robust clustering method by using the correntropy induced metric, which is robust for handling the non-Gaussian and impulsive noises. Also we further extend the method for handling the data with outlier rows/features. The multiplicative form of half-quadratic optimization is used to optimize the non-convex correntropy objective function of the proposed models. Extensive experiments on face datasets well demonstrate that the proposed methods are more robust to corruptions and occlusions.

1. Introduction

In pattern recognition and computer vision community, the data usually follow certain type of simple structure that enables intelligent representation. The subspaces are possibly the most widely used data model, since many real-world data, such as face images and motions, can be well characterized by subspaces. Given a set of data points, assuming that they are drawn from multiple subspaces, the goal of subspace clustering is to (1) cluster these data points into clusters with each cluster corresponding to a subspace, and (2) predict the memberships of the subspaces, including the number of subspaces and the basis of each subspace. Subspace clustering is a fundamental problem and has numerous applications in the machine learning and computer vision literature, e.g. motion segmentation [21] and image clustering [13]. The challenge in these applications lies in that the only known information is the data points, and they are usually contaminated by various noises. Figure 1 illustrates some face images from three subjects. The face images with pixel corruption, sunglasses and/or scarf, deviate from their underlying subspaces. In this case, the subspace clustering is challenging. This paper aims to address the robust subspace clustering problem with various noises, such as the non-Gaussian noises.

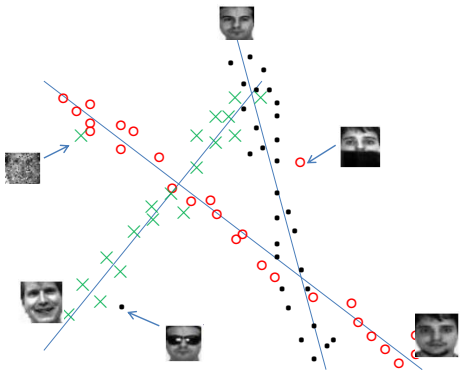

Figure 1. Face images belonging to different subjects lie in different subspaces. Noises by and corruptions deviate the data from the underlying subspaces.

1.1. Summary of Main Notations

In this work, matrices are represented with capital symbols. In particular, I denotes the identity matrix. For a matrix M, M_{ij} and $(M)_{ij}$ denote its (i,j)-th entry. M^i is its i-th row, and M_j is its j-th column. Diag(v) converts the vector v into a diagonal matrix in which the i-th diagonal entry is v_i. $\mathbb{R}_+$ denotes the set of non-negative real values and $\mathbb{S}_+^{d\times n}$ denote the set of positive semi-definite matrices. $M \succ 0$ denotes that M is symmetric and positive definite. C^1 denotes the set of continuous first derivative functions.

*Corresponding author.

$||v||_2$ and $||v||_\infty$ denote the L2 norm and infinity norm of vector v, respectively. L1 norm, L21 norm and nuclear norm of matrix M are defined as $||M||_1 = \sum_{ij} |M_{ij}|$, $||M||_{21} = \sum_j ||M_j||_2$, and $||M||_* = \sum_i \sigma_i$ (σ_i is the i-th singular value of M), respectively.

1.2. Related Work

Many subspace clustering methods have been proposed [21, 11, 6, 7]. In this work, we focus on the recent graph based subspace clustering methods [3, 4, 11, 10, 13]. These methods are based on the spectral clustering, and its first step aims to construct an affinity (or graph) matrix which is close to be block diagonal, with zero elements corresponding to data pair from different subspaces. After the affinity matrix is learned, the Normalized Cut [20] is employed to segment the data into multiple clusters. For a given data matrix $X \in \mathbb{R}^{d\times n}$, where d denotes the feature dimension and n is the number of data points, the most recent methods, including L1-graph [3] or Sparse Subspace Clustering (SSC) [4], Low-Rank Representation (LRR) [11, 10], Multi-Subspace Representation (MSR) [14] and Least Squares Representation (LSR) [13] learn the affinity matrix $Z \in \mathbb{R}^{n\times n}$ by solving the following common problem

$$\min_Z \mathcal{L}(X - XZ) + \lambda \mathcal{R}(Z). \tag{1}$$

For L1-graph or SSC, $\mathcal{L}(X - XZ) = ||X - XZ||_F^2$ and $\mathcal{R}(Z) = ||Z||_1$. The motivation of using SSC is that the L1-minimization will lead to a sparse solution tending to be block diagonal. As pointed out in [13], the L1-minimization does not exhibit the grouping effect, and thus is weak to group correlated data points together.

For LRR, $\mathcal{L}(X - XZ) = ||X - XZ||_{21}$ and $\mathcal{R}(Z) = ||Z||_*$. It aims to find a low rank affinity matrix. When the data are drawn from independent subspaces, LRR leads to a bock diagonal solution which can recover the true subspaces. For the noisy case, LRR uses the robust L21-norm to remove outlier samples.

MSR simply combines the criteria of SSC and LRR, $\mathcal{L}(X-XZ) = ||X-XZ||_{21}$ and $\mathcal{R}(Z) = ||Z||_1+\gamma||Z||_*$. Thus MSR can be regarded as a tradeoff between SSC and LRR, but it needs to tune one more parameter γ.

The LSR method uses the Frobenius norm to model both the reconstruction error and the representation matrix, $\mathcal{L}(X - XZ) = ||X - XZ||_F^2$ and $\mathcal{R}(Z) = ||Z||_F^2$. LSR has a closed form solution which makes it efficient, and the grouping effect makes it effective for subspace clustering.

The above methods share the common formulation as shown in (1). The Frobenius norm and L21 norm are used as the loss function while the L1 norm, nuclear norm and Frobenius are used to control the affinity matrix. Different formulations require different solvers for these problems. In this work, we show that the L1 norm, L21 norm and nuclear norm all satisfy certain conditions, and thus the previous subspace clustering methods, including SSC, LRR and MSR, can be unified within a general framework from the perspective of half-quadratic optimization [17]. The relationship between the general framework and the previous optimization methods for sparse and low rank minimization is also presented in this work.

Different from the previous methods which focus on a regularization term $\mathcal{R}(Z)$, this work focuses on the construction error term $\mathcal{L}(Z)$ for robust subspace learning. Previous works use the Frobenius norm to measure the quality of approximation, which is optimal for the case of independent and identically distributed (i.i.d.) Gaussian noise but not robust to outliers. LRR by using the L21 norm is able to remove the outlier samples, but it is sensitive to the outlier features. To overcome the weakness of mean squared error, we propose a new robust subspace clustering method which uses the correntropy induced metric as the loss function. The Frobenius norm is used to control the affinity matrix to preserve the grouping effect as in LSR. Then we minimize the non-convex correntropy objective of the proposed method by alternate minimization.

1.3. Contributions and Organization

We summarize the contributions of this work as follows:

- We propose a new robust subspace clustering method by Correntropy Induced L2 (CIL2) graph. It is able to handle data with non-Gaussian noises. We also extend CIL2 for handling data with outlier rows/features.

- We apply the correntropy induced L2 graph for face clustering under various types of corruptions and occlusions. Extensive experiments demonstrate the effectiveness of the proposed method by comparing it with the state-of-the-art methods.

The remainder of this paper is organized as follows. Section 2 gives a brief review of the half-quadratic analysis and presents a general half-quadratic framework for robust subspace clustering. Section 3 elaborates the proposed CIL2 graph for robust subspace clustering. Section 4 provides experimental results on face clustering under different settings. We conclude this paper in Section 5.

2. A General Half-Quadratic Framework for Robust Subspace Clustering

For a given data matrix $X \in \mathbb{R}^{d\times n}$, consider the following general problem:

$$\begin{aligned} &\min_Z \mathcal{J}(Z) = \mathcal{L}(E) + \lambda \mathcal{R}(Z) \\ &\text{s.t. } E = X - XZ, \end{aligned} \tag{2}$$

Table 1. The popular previous subspace clustering models can be solved by half-quadratic minimization.

Methods	Objective	Function	
	$\min_Z \mathcal{L}(X-XZ)+\lambda\mathcal{R}(Z)$	$\mathcal{L}(\cdot)$	$\mathcal{R}(\cdot)$
SSC [4]	$\min_Z \|\|X-XZ\|\|_F^2+\lambda\|\|Z\|\|_1$	$\|\|\cdot\|\|_F^2$	$\|\|\cdot\|\|_1$
LRR [11]	$\min_Z \|\|X-XZ\|\|_{21}+\lambda\|\|Z\|\|_*$	$\|\|\cdot\|\|_{21}$	$\|\|\cdot\|\|_*$
MSR [14]	$\min_Z \|\|X-XZ\|\|_{21}+\lambda\|\|Z\|\|_1+\lambda\gamma\|\|Z\|\|_*$	$\|\|\cdot\|\|_{21}$	$\|\|\cdot\|\|_1+\gamma\|\|\cdot\|\|_*$
LSR [13]	$\min_Z \|\|X-XZ\|\|_F^2+\lambda\|\|Z\|\|_F^2$	$\|\|\cdot\|\|_F^2$	$\|\|\cdot\|\|_F^2$

where $\mathcal{L}(E)$ is the loss function chosen to be robust to outliers or gross errors, and $\mathcal{R}(Z)$ is the regularization term. The loss function $\mathcal{L}(E)$ and regularization $\mathcal{R}(Z)$ may be non-quadratic. Thus it may be difficult to solve the problem (2). But if $\mathcal{L}(E)$ and $\mathcal{R}(Z)$ satisfy certain conditions, we can minimize $\mathcal{J}(Z)$ by half-quadratic analysis.

In this work, we consider a general case of $\phi(x)$ that satisfies the following conditions [17]

$$\begin{aligned}
&\text{(a) } x\to\phi(x) \text{ is convex on } \mathbb{R},\\
&\text{(b) } x\to\phi(\sqrt{x}) \text{ is concave on } \mathbb{R}_+,\\
&\text{(c) } \phi(x)=\phi(-x), x\in\mathbb{R},\\
&\text{(d) } \phi(x) \text{ is } C^1 \text{ on } \mathbb{R},\\
&\text{(e) } \phi''(0^+)>0,\\
&\text{(f) } \lim_{x\to\infty}\phi(x)/x^2=0.
\end{aligned} \tag{3}$$

Or in the matrix form $\phi(M)$:

$$\begin{aligned}
&\text{(a) } M\to\phi(M) \text{ is convex on } \mathbb{R}^{N\times N},\\
&\text{(b) } M\to\phi(\sqrt{M}) \text{ is concave on } \mathbb{S}_+^{N\times N},\\
&\text{(c) } \phi(M)=\phi(-M),\ M\in\mathbb{R}^{N\times N},\\
&\text{(d) } \phi(M) \text{ is } C^1 \text{ on } \mathbb{R}^{N\times N},\\
&\text{(e) } \phi(M) \text{ is strictly convex on } 0,\\
&\text{(f) } \lim_{M\to\infty}\phi(M)/||M||_F^2=0.
\end{aligned} \tag{4}$$

If $\phi(\cdot)$ satisfies all the conditions in (3), there exists a dual function ψ [17] such that

$$\phi(x)=\inf_{s\in\mathbb{R}}\{\frac{1}{2}sx^2+\psi(s)\}, \tag{5}$$

where s is determined by the minimizer function $\delta(\cdot)$ with respect to $\phi(\cdot)$. $\delta(\cdot)$ admits an explicit form under certain restrictive assumptions:

$$s=\delta(t)=\begin{cases}\phi''(0^+), & \text{if } t=0,\\ \frac{\phi'(t)}{t}, & \text{if } t\neq 0.\end{cases} \tag{6}$$

If $\mathcal{L}(E)=\sum_{ij}\phi(E_{ij})$ (similar analysis can be performed on $\mathcal{R}(Z)$), problem (2) reads:

$$\begin{aligned}
&\min_Z \mathcal{J}(Z)=\sum_{ij}\phi(E_{ij})+\lambda\mathcal{R}(Z)\\
&\text{s.t. } E=X-XZ.
\end{aligned} \tag{7}$$

Using (7) on each E_{ij}, the augmented function of $\mathcal{J}$ of (7) is as follows

$$\mathcal{J}(Z,S)=\sum_{ij}(\frac{1}{2}S_{ij}E_{ij}^2+\psi(S_{ij}))+\lambda\mathcal{R}(Z). \tag{8}$$

Based on the half-quadratic optimization, $\mathcal{J}(Z,S)$ can be minimized by the following alternate procedure:

$$S_{ij}=\delta(E_{ij}), \tag{9}$$

$$Z=\arg\min_Z \sum_{ij}\frac{1}{2}S_{ij}E_{ij}^2+\lambda\mathcal{R}(Z). \tag{10}$$

The update sequence generated by the above scheme will converges. The objective function in (8) is nonincreasing under the update rules in (9)(10) [17].

For L1 norm, $\phi_1(x)=|x|=\sqrt{x^2}$ does not satisfy condition (d) in (3). We use $\phi_1(x)=\sqrt{x^2+\epsilon^2}$ as an approximation of $|x|$ with a small positive value ϵ. It can be easily seen that $\sqrt{x^2+\epsilon^2}$ satisfies all the conditions in (3). We roughly say the L1 norm satisfies all the conditions in (3) in this sense. Previous work [2] for solving the L1-minimization by iteratively reweighted least squares optimization can be interpreted as the half-quadratic optimization in (9) and (10). For L21 norm, $\phi_{21}(X)=||X||_{21}=\sum_i||X_i||_2\approx\sum_i(||X_i||_2^2+\epsilon)^{\frac{1}{2}}$, where ϵ is a small positive value. It is easy to check that $\phi_{21}(x)=(x^2+\epsilon)^{\frac{1}{2}}$ also satisfies all the conditions in (3). For nuclear norm, $\phi_*(X)=\text{Tr}(X^TX)^{\frac{1}{2}}\approx\text{Tr}(X^TX+\epsilon I)^{\frac{1}{2}}$, where ϵ is a small positive value. It is easy to check that $\text{Tr}(X^TX+\epsilon I)^{\frac{1}{2}}$ satisfies the conditions (a)-(e) in (4). For the condition (f), the i-th singular value σ_i of X converges to infinity when $X\to\infty$, and thus $\lim_{X\to\infty}\phi_*(X)/||X||_F^2=\lim_{\sigma_i\to\infty}\frac{\sum_i\sigma_i}{\sum_i\sigma_i^2}=0$. Therefore the nuclear norm also satisfies all the conditions in (4). The work [16] for solving low rank minimization by iteratively reweighted least squares minimization can be interpreted as the half-quadratic minimization.

If both two functions satisfy all the conditions in (3), the sum of them also satisfies these conditions. The optimization method in [14] for minimizing $||X||_1+\gamma||X||_*$ can be regarded as the half-quadratic optimization in (9)(10).

Based on the above analysis, previous subspace clustering methods by using the L1 norm, L21 norm and nuclear norm can be optimized by the half-quadratic analysis on (9)(10) by slightly relaxing the objective function. As

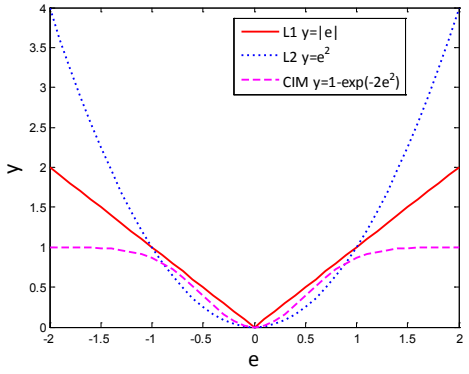

Figure 2. Comparison of different loss functions.

shown in Table 1, previous subspace clustering methods, including SSC, LRR, MSR and LSR, can be regarded as special cases of the problem (2) from the view of half-quadratic analysis. Note that the Frobenius norm $||\cdot||_F^2$ does not need half-quadratic representation because it is already quadratic. We also list it in Table 1 since it is widely used.

3. Correntropy Induced L2 Graph for Robust Subspace Clustering

3.1. Correntropy Induced Metric

The mean squared errors (MSE) are probably the most widely used methodologies for quantifying how similar two random variables are. Successful engineering solutions from this methodology rely heavily on the Gaussianity and linearity assumptions. The work in [5] extended the concept of mean squared error adaptation to information theoretic learning (ITL) to include the information theoretic criteria. Then they further proposed the concept of correntropy to process non-Gaussian and impulsive noises [12]. The correntropy is a generalized similarity measure between two arbitrary scalar random variables u and v defined by

$$V_\sigma(u,v) = \mathbf{E}[k_\sigma(e)], \tag{11}$$

where $e = u - v$, $\mathbf{E}[\cdot]$ is the expectation operator, and $k_\sigma(\cdot)$ is the kernel function. In this work we only consider the Gaussian kernel $k_\sigma(e) = \exp(-e^2/2\sigma^2)$. In practice, we usually have only a finite number of data $\{(u_i, v_i)\}_{i=1}^n$, which leads to the sample estimator of correntropy:

$$\hat{V}_\sigma(u,v) = \frac{1}{n}\sum_{i=1}^{n} k_\sigma(u_i - v_i). \tag{12}$$

Based on (12), Liu et al. [12] extended the concept of correntropy criterion for a general similarity measurement between any two vectors, which is called the Correntropy Induced Metric (CIM). It is formally defined as

$$\text{CIM}(u,v) = (k(0) - \frac{1}{n}\sum_{i=1}^{n} k_\sigma(e_i))^{1/2}, \tag{13}$$

where $e_i = u_i - v_i$, for each $i = 1, \cdots, n$.

Figure 2 shows a comparison of the absolute error, mean squared error and CIM. The mean squared error is a global metric which increases quadratically for large errors. CIM is a local metric which is close to the absolute error when the errors are relatively small. For large errors, the value of CIM is close to 1. Note that the large errors are usually caused by outliers, but their effect on CIM is limited. Therefore CIM will be more robust to the non-Gaussian noises. The effectiveness and robustness of correntropy have been verified in face recognition [9], feature selection [8] and signal processing [12]. This paper uses this concept for robust subspace clustering.

3.2. Correntropy Induced L2 Graph

For robust subspace clustering, we use the correntropy to replace the Frobenius norm in the LSR model to model the reconstruction error, leading to the Correntropy Induced L2 (CIL2) graph as follows:

$$\begin{aligned} &\min_Z \sum_{i,j}(1 - k_\sigma(E_{ij})) + \lambda||Z||_F^2 \\ &\text{s.t. } E = X - XZ. \end{aligned} \tag{14}$$

It is easy to check that $\phi_\sigma(x) = 1 - k_\sigma(x) = 1 - \exp(-x^2/2\sigma^2)$ satisfies all the conditions in (3). Therefore the above problem can be solved by the half-quadratic analysis. According to (8), problem (14) is equivalent to the following augmented objective function:

$$\begin{aligned} &\mathcal{J}(Z,S) = \sum_{ij}(\frac{1}{2}S_{ij}E_{ij}^2 + \psi(S_{ij})) + \lambda||Z||_F^2 \\ &\text{s.t. } E = X - XZ, \end{aligned} \tag{15}$$

where $\psi(\cdot)$ is the dual function corresponding to $\phi_\sigma(\cdot)$. We can minimize $\mathcal{J}(Z,S)$ in (15) by the following alternate procedure:

$$S_{ij} = \frac{1}{\sigma^2}\exp(-E_{ij}^2/2\sigma^2), \tag{16}$$

$$Z_i = \arg\min_{Z_i}(X - XZ)_i^T \text{Diag}(S_i)(X - XZ)_i + \lambda||Z_i||_2^2. \tag{17}$$

Let $\hat{X} = \text{Diag}(\sqrt{S_i})X$, then problem (17) is also a least square regression model:

$$\min_{Z_i} ||\hat{X} - \hat{X}Z_i||_2^2 + \lambda||Z_i||_2^2. \tag{18}$$

Since the kernel size σ may affect the performance of the proposed model. It is usually determined empirically. In this study, the kernel size is computed as the average reconstruction error,

$$\sigma^2 = \frac{1}{2dn}||X - XZ||_F^2. \tag{19}$$

From (15) or problem (18), we can see that the correntropy based LSR model can be regarded as a weighted LSR, where each weight S_{ij} corresponding to E_{ij} is used to control the effect of E_{ij}.

3.3. Row Based Correntropy Induced L2 Graph

In some real-world applications, the data may be occluded with outlier rows/features. For example, some rows of the face images with sunglasses and scarf are outliers, which are not discriminative for classification and clustering. In this case, we should measure the quality of the reconstruction error based on the entire row. The effect of rows can be controlled by assigning different weights, and each element in the same row has the same weight. To this end, we have the row based Correntropy Induced L2 (rCIL2) graph by solving the following problem

$$\begin{aligned} &\min_{Z} \sum_{i} (1 - k_\sigma(||E^i||_2)) + \lambda ||Z||_F^2 \\ &\text{s.t. } E = X - XZ. \end{aligned} \tag{20}$$

According to the half-quadratic analysis, the above problem is equivalent to the following problem

$$\mathcal{J}_r(Z, w) = \sum_i (\frac{1}{2} w_i ||X^i - X^i Z||_2^2 + \psi(w_i)) + \lambda ||Z||_F^2. \tag{21}$$

Problem (21) can be solved by updating Z, w, and σ alternately as follows:

$$w_i = \frac{1}{\sigma^2} \exp(-(X^i - X^i Z)^2 / 2\sigma^2), \tag{22}$$

$$Z = \arg\min_{Z} \text{Tr}((X - XZ)^T \text{Diag}(w)(X - XZ)) + \lambda ||Z||_F^2, \tag{23}$$

$$\sigma^2 = \frac{1}{2d} \sum_i ||X^i - X^i Z||_2^2. \tag{24}$$

According to (6) and (10), it is easy to prove that the sequences $\{\hat{\mathcal{J}}(Z^t, S^t), t = 1, 2, \cdots\}$ in (15) and $\{\hat{\mathcal{J}}(Z^t, w^t), t = 1, 2, \cdots\}$ in (21) converge.

3.4. The Grouping Effect

The CIL2 and rCIL2 graphs also use the L2 regularization as in LSR [13]. It is expected that they also have the grouping effect, i.e. the coefficients of a group of correlated data are approximately equal. The obtained solutions by CIL2 in (17) and by rCIL2 in (23) are the weighted least square regression model which owns the grouping effect:

Proposition 1 *Given a data vector $y \in \mathbb{R}^d$, data points $X \in \mathbb{R}^{d \times n}$, the weight vector $w \in \mathbb{R}^d$ corresponding to each row of X, and a parameter λ. Assume that each data point of X is normalized. Let z^* be the optimal solution to the following weighted LSR (in vector form) problem:*

$$\min_{z} ||Diag(w)(y - Xz)||_2^2 + \lambda ||z||_2^2. \tag{25}$$

We have

$$\frac{||z_i^* - z_j^*||_2}{||w||_2 ||Diag(w)y||_2} \leq \frac{1}{\lambda} \sqrt{2(1 - r)}, \tag{26}$$

where $r = X_i^T X_j$ is the sample correlation.

We omit the proof of Proposition 1 here, it can be proved in the same way as the Theorem 7 in [13].

The mechanism of correntropy and the Proposition 1 ensure that both CIL2 and rCIL2 are not only robust to noises but also preserve the grouping effect.

3.5. Algorithm for Subspace Clustering

Similar to the previous subspace clustering method LSR, which uses the representation coefficient matrix to construct the graph for clustering, we apply the learned solution Z^* by CIL2 and rCIL2 to construct a graph with weights $W = (|Z^*| + |Z^{*T}|)/2$, and then Normalized Cut [20] is applied to cluster the data points into multiple clusters.

4. Experiments

4.1. Datasets and Settings

Our experiments are performed on three face datasets: Yale, ORL and AR. Descriptions of these data sets are given as follows.

The Yale face dataset [1] contains 165 grayscale images of 15 individuals. The images demonstrate variations in lighting condition and facial expression (normal, happy,

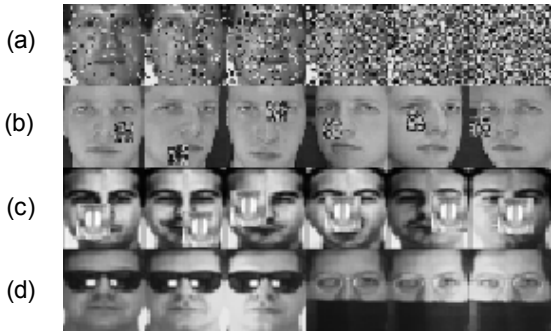

Figure 3. (a) Some corrupted face images from the Yale dataset, with 10%, 20% 30%, 50%, 70% and 90% of pixels corrupted, respectively; (b) Some face images with random block occlusion from the ORL dataset; (c) Some face images with 20% occlusion by monkey face from the AR dataset; (d) Some face images with contiguous occlusion by sunglasses and scarf from the AR dataset.

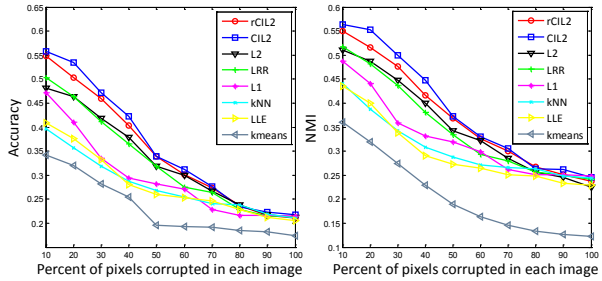

Figure 4. Clustering accuracy and NMI on the Yale dataset with pixel corruption for different algorithms.

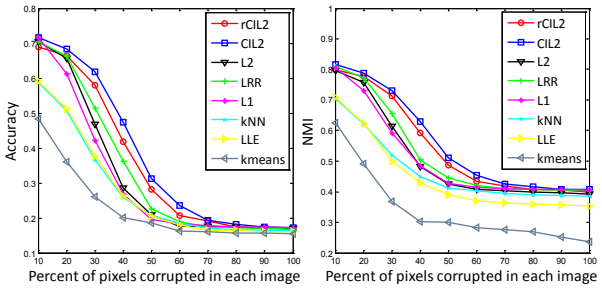

Figure 5. Clustering accuracy and NMI on the ORL dataset with pixel corruption for different algorithms.

sad, sleepy, surprised, and wink). The grayscale images are resized to a resolution of 32×32 pixels.

The ORL face dataset [19] contains 400 images of 40 individuals. Some images were captured at different times and have different variations including expression (open or closed eyes, smiling or non-smiling) and facial details (glasses or no glasses). The images were taken with a tolerance for some tilting and rotation of the face up to 20 degrees. Each image is resized to 32×32 pixels.

The AR database [15] consists of over 4,000 facial images from 126 subjects. For each subject, 26 facial images are taken in two separate sessions. These images suffer different facial variations, including various facial expressions (neutral, smile, anger, and scream), illumination variations (left light on, right light on, and all side lights on), and occlusion by sunglasses or scarf. We select a subset of the data set consisting of 50 male subjects and 50 female subjects. The grayscale images are resized to a resolution of 32×32 pixels.

4.2. Evaluation Metrics

The clustering result is evaluated by the accuracy and normalized mutual information (NMI) metric as in[22]. For each data point x_i, let p_i and y_i be the obtained cluster label and the label provided by the ground truth, respectively. The accuracy is defined as follows:

$$Accuracy = \frac{\sum_{i=1}^{n} \delta(y_i, map(p_i))}{n}, \tag{27}$$

where $\delta(a, b)$ is the delta function that equals one if $a = b$ and equals zero otherwise, and $map(p_i)$ is the permutation mapping function that maps each cluster label p_i to the equivalent label in y.

Let C denote the set of clusters obtained from the ground truth and C' obtained by the segmentation method. Their mutual information metric $MI(C, C)$ is defined as follows:

$$MI(C, C') = \sum_{c_i \in C, c'_j \in C'} p(c_i, c'_j) log_2 \frac{p(c_i, c'_j)}{p(c_i)p(cj')}, \tag{28}$$

where $p(c_i)$ and $p(c'_j)$ are the probabilities that a sample point arbitrarily selected from the data point belongs to the clusters c_i and c'_j, respectively, and $p(c_i, c'_j)$ is the joint probability that the arbitrarily selected data point belongs to the clusters c_i as well as c'_j at the same time. We use the normalized mutual information (NMI) as follows:

$$NMI(C, C') = \frac{MI(C, C')}{\max(H(C), H(C'))}, \tag{29}$$

where $H(C)$ and $H(C')$ are the entropies of C and C', respectively. It is easy to see that $NMI(C, C)$ ranges from 0 to 1. $NMI = 1$ if the two sets of clusters are identical, and $NMI = 0$ if the two sets are independent.

4.3. Algorithm Settings

We compare our rCIL2 and CIL2 graphs with several graph construction methods for subspace clustering, including the L1-graph [3] (or SSC [4]), L2-graph (LSR) [13], and LRR-graph [10]. kNN and LLE [18] are also applied to construct graphs for subspace clustering. Kmeans is used as the baseline for comparison. The model parameters of these methods are searched from the candidate value sets and the best results are reported.

4.4. Results under Random Pixel Corruption

In some practical scenarios, the face images may be partially corrupted. We evaluate the algorithmic robustness on the Yale and ORL face datasets. Each image is corrupted by replacing a percentage of randomly chosen pixels with i.i.d. samples from a uniform distribution (uniform on [0, 255]). The corrupted pixels are randomly chosen for each image, and the locations are unknown. We vary the percentage r of corrupted pixels from 10% to 100%. Figure 3 (a) shows some examples of those corruptions. To the human eyes, beyond 50% corruption, the corrupted images are barely recognizable as face images. Since the images are with random corruption, we repeat the experiments for 20 times for each r, and the means of accuracy and NMI are reported for evaluation.

Figures 4 and 5 show the means of clustering accuracy and NMI of different methods as functions of the corruption level. It can be found that both the accuracy and NMI decrease when more pixels of each image are randomly

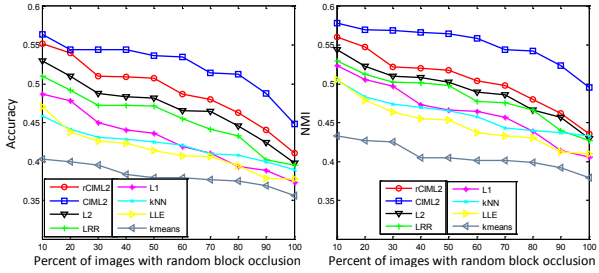

Figure 6. Clustering accuracy and NMI on the Yale dataset with block occlusion for different algorithms.

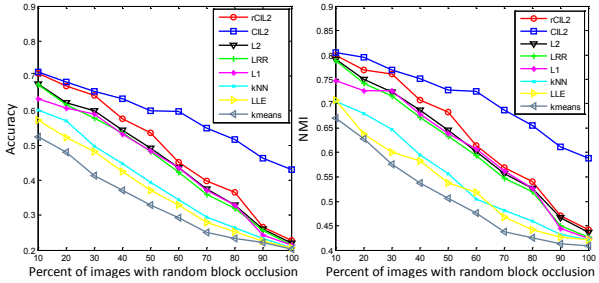

Figure 7. Clustering accuracy and NMI on the ORL dataset with block occlusion for different algorithms.

corrupted. Our proposed CIL2 and rCIL2 outperform the compared methods in most cases. In particular, the CIL2 usually performs better than rCIL2 when the percentage of the corrupted pixels is no more than 50% on the Yale dataset and 70% on the ORL dataset. This is because each row of images may not be regarded as outliers when the level of the random pixel corruption is low. LRR and L2-graph perform competitively on both datasets, which also verifies the effectiveness of the grouping effect of these two methods for subspace clustering. When the images are with high percentage of pixel corruptions, none of the compared methods perform well due to insufficient discriminative information.

4.5. Results under Contiguous Occlusion

In this subsection we simulate various types of contiguous occlusions by replacing a randomly selected local region in some randomly selected images with a black-white square and an unrelated monkey image.

The first experiment is conducted on the Yale and ORL datasets with random block occlusion. Figure 3 (b) shows some face images with such black-white occlusions, in size of 8×8 pixels. In each dataset, we select r percentage of the images for occlusion, with r varying from 10% to 100%. The experiments are repeated 20 times for each r, and the means of accuracy and NMI are reported for evaluation.

Figures 6 and 7 show the means of clustering accuracy and NMI of each method on different percentages of corrupted images. CIL2-graph achieves the best accuracy and NMI on both Yale and ORL datasets in all cases. Compared with previous subspace clustering methods, the improvement by rCIL2 is still limited. The phenomenon is similar to the random pixel corruption scenario, since the images with block occlusion will not lead to outlier rows. rCIL2-graph will not be very effective in this case. Notice that in this experiment, r percentage of the images in each dataset is selected to be occluded with a size of 8×8 block, and thus the decreasing curves of the clustering accuracy and NMI are flatter than those in Figures 4 and 5.

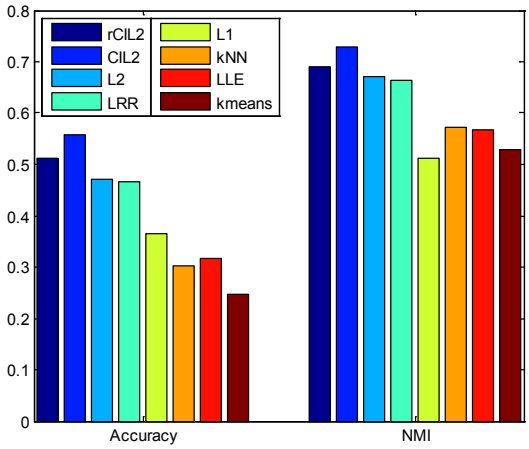

Figure 8. Clustering accuracy and NMI on the AR dataset with an unrelated image occlusion for different algorithms.

The second experiment is conducted on a subset of AR dataset. This subset consists of 1,400 images from 100 subjects, 50 males and 50 females. These images are of non-occluded frontal views with various facial expressions in Sessions 1 and 2. For each image, we randomly select a local region to be replaced by an unrelated monkey image. The size of monkey image is 14×14, i.e. about 20% pixels of each image are occluded. Figure 3 (c) shows some face images with such unrelated image occlusions.

Figure 8 shows the clustering accuracy and NMI of each method on the AR dataset with unrelated monkey image occlusion. The experimental results are similar to the above experiment. Still, CIL2 obtains the best results, and rCIL2, LRR and L2-graph are competitive on this experiment.

4.6. Results on Real-World Malicious Occlusion

In real-world face recognition systems, people may wear sunglasses or scarfs which make the classification or clustering more challenging. In this subsection, we evaluate the robustness of the proposed method on the AR dataset with sunglasses and scarf occlusions. The AR dataset contains two separate sessions. In each session, each subject has 7 face images with different facial variations, 3 face images with sunglasses occlusion and 3 face images with scarf occlusion. Figure 3 (d) shows some face images with such an occlusion. In each session, we conduct two experiments corresponding to the sunglasses and scarf occlusions. For sunglasses occlusion, we use the first 2 normal face images

Table 2. The clustering accuracy (%) and NMI (%) of different algorithms on the AR dataset.

Methods	Accuracy				NMI			
	Session 1		Session 2		Session 1		Session 2	
	Sunglasses	Scarf	Sunglasses	Scarf	Sunglasses	Scarf	Sunglasses	Scarf
rCIL2	**85.2**	**78.4**	**86.4**	**81.2**	**93.8**	**90.1**	**94.0**	**91.5**
CIL2	81.2	75.4	85.4	79.0	89.9	87.9	93.8	88.8
L2	78.2	71.6	80.0	72.6	86.3	83.8	90.7	83.8
LRR	77.2	72.2	79.6	74.6	86.6	84.7	90.7	84.7
L1	43.8	40.6	27.8	40.2	72.3	67.7	55.8	60.8
kNN	26.4	25.6	26.8	27.2	65.4	66.0	66.1	65.9
LLE	28.0	27.6	33.2	27.2	63.4	62.9	66.6	61.7
kmeans	30.0	29.4	30.8	29.8	65.4	63.8	65.3	65.5

and 3 face images with sunglasses of each subject. For scarf occlusion, we use the first 2 normal face images and 3 face images with scarf of each subject.

Table 2 shows the clustering results on the AR dataset for the images with sunglasses and scarf occlusions. Different from the above experiments, rCIL2 achieves the best clustering accuracy and NMI in all cases. That is because the face images with sunglasses and scarf occlusions contain many outlier rows/features, and rCIL2 is designed for such a task. Both LRR and L2 graphs perform better than L1 graph, which is consistent with the result in [10, 13].

5. Conclusions

In this paper, we study the robust subspace clustering problem, and present a general framework from the viewpoint of half-quadratic optimization to unify the L1 norm, Frobenius norm, L21 norm and nuclear norm based subspace clustering methods. Previous iteratively reweighted least squares optimization methods for the sparse and low rank minimization can be regarded as the half-quadratic optimization. As a new special case, we use the correntropy as the loss function for robust subspace clustering to handle the non-Gaussian and impulsive noises. An alternate minimization algorithm is used to optimize the non-convex correntropy objective. Extensive experiments on the face clustering with various types of corruptions and occlusions well demonstrate the effectiveness and robustness of the proposed methods by comparing with the state-of-the-art subspace clustering methods.

Acknowledgements

This research is supported by the Singapore National Research Foundation under its International Research Centre @Singapore Funding Initiative and administered by the IDM Programme Office. Z. Lin is supported by National Nature Science Foundation of China (Grant nos. 61272341, 61231002, and 61121002).

References

[1] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman. Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection. *TPAMI*, 19(7):711–720, 1997.

[2] R. Chartrand and W. Yin. Iteratively reweighted algorithms for compressive sensing. In *ICASSP*, pages 3869–3872, 2008.

[3] B. Cheng, J. Yang, S. Yan, Y. Fu, and T. S. Huang. Learning with ℓ^1-graph for image analysis. *TIP*, 19:858–866, 2010.

[4] E. Elhamifar and R. Vidal. Sparse subspace clustering. In *CVPR*, pages 2790–2797, 2009.

[5] D. Erdogmus and J. C. Principe. An error-entropy minimization algorithm for supervised training of nonlinear adaptive systems. *TSP*, 50(7):1780–1786, 2002.

[6] Y. Fang, R. Wang, and B. Dai. Graph-oriented learning via automatic group sparsity for data analysis. In *ICDM*, pages 251–259. IEEE, 2012.

[7] J. Gui, Z. Sun, W. Jia, R. Hu, Y. Lei, and S. Ji. Discriminant sparse neighborhood preserving embedding for face recognition. *Pattern Recognition*, 45(8):2884–2893, 2012.

[8] R. He, T. Tan, L. Wang, and W.-S. Zheng. L21 regularized correntropy for robust feature selection. In *CVPR*, pages 2504–2511, 2012.

[9] R. He, W.-S. Zheng, and B.-G. Hu. Maximum correntropy criterion for robust face recognition. *TPAMI*, 33(8):1561–1576, 2011.

[10] G. Liu, Z. Lin, S. Yan, J. Sun, Y. Yu, and Y. Ma. Robust recovery of subspace structures by low-rank representation. *TPAMI*, 35(1):171–184, 2013.

[11] G. Liu, Z. Lin, and Y. Yu. Robust subspace segmentation by low-rank representation. In *ICML*, pages 663–670, 2010.

[12] W. Liu, P. P. Pokharel, and J. C. Príncipe. Correntropy: properties and applications in non-Gaussian signal processing. *TSP*, 55(11):5286–5298, 2007.

[13] C.-Y. Lu, H. Min, Z.-Q. Zhao, L. Zhu, D.-S. Huang, and S. Yan. Robust and efficient subspace segmentation via least squares regression. In *ECCV*, 2012.

[14] D. Luo, F. Nie, C. Ding, and H. Huang. Multi-subspace representation and discovery. In *ECML PKDD*, pages 405–420, 2011.

[15] A. M. Martinez. The AR face database. *CVC Technical Report*, 24, 1998.

[16] K. Mohan and M. Fazel. Iterative reweighted algorithms for matrix rank minimization. In *JMLR*, volume 13, pages 3441–3473, 2012.

[17] M. Nikolova and M. K. Ng. Analysis of half-quadratic minimization methods for signal and image recovery. *SIAM Journal on Scientific computing*, 27(3):937–966, 2005.

[18] S. T. Roweis and L. K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, 2000.

[19] F. S. Samaria and A. C. Harter. Parameterisation of a stochastic model for human face identification. In *IEEE Workshop on Applications of Computer Vision*, pages 138–142, 1994.

[20] J. B. Shi and J. Malik. Normalized cuts and image segmentation. *TPAMI*, 22(8):888–905, 2000.

[21] R. Tron and R. Vidal. A benchmark for the comparison of 3-D motion segmentation algorithms. In *CVPR*, pages 1–8. IEEE, 2007.

[22] W. Xu, X. Liu, and Y. Gong. Document clustering based on non-negative matrix factorization. In *International ACM SIGIR conference on Research and development in informaion retrieval*, pages 267–273, 2003.