Robust Subspace Clustering via Thresholding Ridge Regression

Xi Peng[1], **Zhang Yi**[2,*], and **Huajin Tang**[1,2,*]
[1]Institute for Infocomm Research, Agency for Science, Technology and Research (A*STAR), Singapore 138632
[2]College of Computer Science, Sichuan University, Chengdu 610065, P.R. China.
pangsaai@gmail.com, zhangyi@scu.edu.cn, htang@i2r.a-star.edu.sg.

Abstract

Given a data set from a union of multiple linear subspaces, a robust subspace clustering algorithm fits each group of data points with a low-dimensional subspace and then clusters these data even though they are grossly corrupted or sampled from the union of dependent subspaces. Under the framework of spectral clustering, recent works using sparse representation, low rank representation and their extensions achieve robust clustering results by formulating the errors (e.g., corruptions) into their objective functions so that the errors can be removed from the inputs. However, these approaches have suffered from the limitation that the structure of the errors should be known as the prior knowledge. In this paper, we present a new method of robust subspace clustering by eliminating the effect of the errors from the projection space (representation) rather than from the input space. We firstly prove that ℓ_1-, ℓ_2-, and ℓ_∞-norm-based linear projection spaces share the property of intra-subspace projection dominance, i.e., the coefficients over intra-subspace data points are larger than those over inter-subspace data points. Based on this property, we propose a robust and efficient subspace clustering algorithm, called Thresholding Ridge Regression (TRR). TRR calculates the ℓ_2-norm-based coefficients of a given data set and performs a hard thresholding operator; and then the coefficients are used to build a similarity graph for clustering. Experimental studies show that TRR outperforms the state-of-the-art methods with respect to clustering quality, robustness, and time-saving.

Introduction

Subspace segmentation or subspace clustering (Vidal 2011) fits each group of data points using a low dimensional subspace and performs clustering in the projection space, which has attracted increasing interests from numerous areas such as image analysis (Cheng et al. 2010), motion segmentation (Gear 1998), and face clustering (Ho et al. 2003). When the data sets are clean and the subspaces are mutually independent, several existing approaches such as (Costeira and Kanade 1998) are able to exactly resolve the subspace clustering problem. However, the data sets probably contain various noises or lie on the intersection of multiple dependent subspaces. As a result, inter-cluster data points (i.e., the data points with different labels) may be wrongly grouped into the same cluster. Errors removing aims at eliminating the effect of these errors (i.e., noises, etc.), which has lain at the heart of subspace clustering. To achieve this so-called robust subspace clustering, various methods have been proposed, e.g., generalized principal component analysis (Vidal, Ma, and Sastry 2005), local subspace affinity (Yan and Pollefeys 2006), spectral curvature clustering (Chen and Lerman 2009), local best-fit flats (Zhang et al. 2012), fix rank representation (Liu et al. 2012), Sparse Subspace Clustering (SSC) (Elhamifar and Vidal 2013; Peng, Zhang, and Yi 2013), Low Rank Representation (LRR) (Lin, Liu, and Su 2011; Liu et al. 2013), and Least Squares Regression (LSR) (Lu et al. 2012).

In these approaches, representation-based spectral clustering methods have achieved state-of-the-art results in face clustering. The key of spectral clustering is to build an affinity matrix $\mathbf{W}$ of which each entry $\mathbf{W}_{ij}$ denotes the similarity between the connected data points. A 'good' affinity matrix is a block-diagonal matrix (sparse similarity graph), i.e., $\mathbf{W}_{ij} = 0$ unless the corresponding data points $\mathbf{x}_i$ and $\mathbf{x}_j$ belong to the same cluster. A frequently-used measurement of $\mathbf{W}_{ij}$ is Euclidean distance with Heat Kernel. However, this metric is sensitive to noise and cannot capture the structure of subspace. Recently, SSC and LRR provide a new way to construct the graph by using the sparse and low-rank representation, respectively. Moreover, they remove errors from the inputs by formulating the errors into their objective functions. Both theoretical analysis and experimental results have shown that SSC and LRR can handle some specific errors and have achieved impressive performance. Inspired by the success of SSC and LRR, numerous approaches have been proposed and the errors-removing method is widely adopted in this field (Liu et al. 2012; Lu et al. 2012; Liu and Yan 2011; Wang and Xu 2013; Deng et al. 2013). One major limitation of these approaches is that the structure of errors should be known as the priori. Clearly, this prior knowledge is difficult to achieve and these algorithms may fail unless the adopted assumption is consistent with the real structure of the errors.

Most existing methods solve the robust subspace clus-

*Corresponding Authors

Table 1: Notations.

Notation	Definition
n	data size
m	the dimensionality of samples
r	the rank of a given matrix
$\mathbf{x} \in \mathbb{R}^m$	a data point
$\mathbf{c} \in \mathbb{R}^n$	the representation of $\mathbf{x}$ over $\mathbf{D}$
$\mathbf{D} = [\mathbf{d}_1, \mathbf{d}_2, \ldots, \mathbf{d}_n]$	a given dictionary
$\mathbf{D}_x \in \mathbf{D}$	$\mathbf{x}$ and $\mathbf{D}_x$ belong to the same cluster
$\mathbf{D}_{-x}$	the data points of $\mathbf{D}$ except $\mathbf{D}_x$

tering problem by **removing the errors from the original data space** and obtaining a good affinity matrix based on a 'clean' data set. Differing from these approaches, we propose and prove that **the effect of errors can be eliminated from linear projection space because the coefficients with small values (trivial coefficients) always correspond to the projections over the errors**. This property, called intra-subspace projection dominance, is mathematically trackable. Based on our theoretical result, we further present an algorithm, Thresholding Ridge Regression (TRR), by considering ℓ_2-norm case. TRR has a closed-form solution and makes clustering data into multiple subspaces possible even though the structure of errors is unknown and the data are grossly corrupted.

Notations: Unless specified otherwise, **lower-case bold letters** represent column vectors and **upper-case bold ones** represent matrices. $\mathbf{A}^T$ and $\mathbf{A}^{-1}$ denote the transpose and pseudo-inverse of the matrix $\mathbf{A}$, respectively. $\mathbf{I}$ denotes the identity matrix. Table 1 summarizes some notations used throughout the paper.

Intra-subspace Projection Dominance

Let intra-subspace data points consist of the points belong to the same subspace and inter-subspace data points be the collection of points came from different subspaces. In this section, we show that the coefficients over intra-subspace data points are larger than those over inter-subspace data points in ℓ_1-, ℓ_2-, and ℓ_∞-norm-based projection space, namely, intra-subspace projection dominance. The proofs are presented in the supplementary material.

Let $\mathbf{x} \neq \mathbf{0}$ be a data point in the union of subspaces $\mathcal{S}_\mathbf{D}$ that is spanned by $\mathbf{D} = [\mathbf{D}_x \ \mathbf{D}_{-x}]$, where $\mathbf{D}_x$ and $\mathbf{D}_{-x}$ consist of the intra-cluster and inter-cluster data points, respectively. Note that, noise and outlier could be regarded as a kind of inter-cluster data point of $\mathbf{x}$. Without loss of generality, let $\mathcal{S}_{\mathbf{D}_x}$ and $\mathcal{S}_{\mathbf{D}_{-x}}$ be the subspace spanned by $\mathbf{D}_x$ and $\mathbf{D}_{-x}$, respectively. Hence, there are only two possibilities for the location of $\mathbf{x}$, i.e., in the intersection between $\mathcal{S}_{\mathbf{D}_x}$ and $\mathcal{S}_{\mathbf{D}_{-x}}$ (denoted by $\mathbf{x} \in \{\mathcal{S} | \mathcal{S} = \mathcal{S}_{\mathbf{D}_x} \cap \mathcal{S}_{\mathbf{D}_{-x}}\}$), or in $\mathcal{S}_{\mathbf{D}_x}$ except the intersection (denoted by $\mathbf{x} \in \{\mathcal{S} | \mathcal{S} = \mathcal{S}_{\mathbf{D}_x} \backslash \mathcal{S}_{\mathbf{D}_{-x}}\}$).

Let $\mathbf{c}_x^*$ and $\mathbf{c}_{-x}^*$ be the optimal solutions of

$$\min \ \|\mathbf{c}\|_p \quad \text{s.t. } \mathbf{x} = \mathbf{D}\mathbf{c}, \tag{1}$$

over $\mathbf{D}_x$ and $\mathbf{D}_{-x}$, respectively. $\|\cdot\|_p$ denotes the ℓ_p-norm and $p = \{1, 2, \infty\}$. We aim to investigate the conditions under which, for every nonzero data point $\mathbf{x} \in \mathcal{S}_{\mathbf{D}_x}$, if the ℓ_p-norm of $\mathbf{c}_x^*$ is smaller than that of $\mathbf{c}_{-x}^*$, then the coefficients over intra-subspace data points are larger than those over inter-subspace data points, i.e., $[\mathbf{c}_x^*]_{r_x,1} > [\mathbf{c}_{-x}^*]_{1,1}$ (intra-subspace projection dominance). Here, $[\mathbf{c}_x^*]_{r_x,1}$ denotes the r_x-th largest absolute value of the entries of $\mathbf{c}_x^*$ and r_x is the dimensionality of $\mathcal{S}_\mathbf{D}$.

In the following analysis, Lemma 1 and Lemma 3 show $[\mathbf{c}_x^*]_{r_x,1} > [\mathbf{c}_{-x}^*]_{1,1}$ when $\mathbf{x} \in \{\mathcal{S} | \mathcal{S} = \mathcal{S}_{\mathbf{D}_x} \backslash \mathcal{S}_{\mathbf{D}_{-x}}\}$ and $\mathbf{x} \in \{\mathcal{S} | \mathcal{S} = \mathcal{S}_{\mathbf{D}_x} \cap \mathcal{S}_{\mathbf{D}_{-x}}\}$, respectively. And Lemma 2 is a preliminary step toward Lemma 3.

Lemma 1 *For any nonzero data point $\mathbf{x}$ in the subspace $\mathcal{S}_{\mathbf{D}_x}$ except the intersection between $\mathcal{S}_{\mathbf{D}_x}$ and $\mathcal{S}_{\mathbf{D}_{-x}}$, i.e., $\mathbf{x} \in \{\mathcal{S} | \mathcal{S} = \mathcal{S}_{\mathbf{D}_x} \backslash \mathcal{S}_{\mathbf{D}_{-x}}\}$, the optimal solution of (1) over $\mathbf{D}$ is given by $\mathbf{c}^*$ which is partitioned according to the sets $\mathbf{D}_x$ and $\mathbf{D}_{-x}$, i.e., $\mathbf{c}^* = \begin{bmatrix} \mathbf{c}_x^* \\ \mathbf{c}_{-x}^* \end{bmatrix}$. Thus, we must have $[\mathbf{c}_x^*]_{r_x,1} > [\mathbf{c}_{-x}^*]_{1,1}$.*

Lemma 2 *Consider a nonzero data point $\mathbf{x}$ in the intersection between $\mathcal{S}_{\mathbf{D}_x}$ and $\mathcal{S}_{\mathbf{D}_{-x}}$, i.e., $\mathbf{x} \in \{\mathcal{S} | \mathcal{S} = \mathcal{S}_{\mathbf{D}_x} \cap \mathcal{S}_{\mathbf{D}_{-x}}\}$. Let $\mathbf{c}^*$, $\mathbf{z}_x$, and $\mathbf{z}_{-x}$ be the optimal solution of*

$$\min \ \|\mathbf{c}\|_p \quad \text{s.t. } \mathbf{x} = \mathbf{D}\mathbf{c} \tag{2}$$

over $\mathbf{D}$, $\mathbf{D}_x$, and $\mathbf{D}_{-x}$. $\mathbf{c}^* = \begin{bmatrix} \mathbf{c}_x^* \\ \mathbf{c}_{-x}^* \end{bmatrix}$ is partitioned according to the sets $\mathbf{D} = [\mathbf{D}_x \ \mathbf{D}_{-x}]$. If $\|\mathbf{z}_x\|_p < \|\mathbf{z}_{-x}\|_p$, then $[\mathbf{c}_x^*]_{r_x,1} > [\mathbf{c}_{-x}^*]_{1,1}$.

Lemma 3 *Consider the nonzero data point $\mathbf{x}$ in the intersection between $\mathcal{S}_{\mathbf{D}_x}$ and $\mathcal{S}_{\mathbf{D}_{-x}}$, i.e., $\mathbf{x} \in \{\mathcal{S} | \mathcal{S} = \mathcal{S}_{\mathbf{D}_x} \cap \mathcal{S}_{\mathbf{D}_{-x}}\}$, where $\mathcal{S}_{\mathbf{D}_x}$ and $\mathcal{S}_{\mathbf{D}_{-x}}$ denote the subspace spanned by $\mathbf{D}_x$ and $\mathbf{D}_{-x}$, respectively. The dimensionality of $\mathcal{S}_{\mathbf{D}_x}$ is r_x and that of $\mathcal{S}_{\mathbf{D}_{-x}}$ is r_{-x}. Let $\mathbf{c}^*$ be the optimal solution of*

$$\min \ \|\mathbf{c}\|_p \quad \text{s.t. } \mathbf{x} = \mathbf{D}\mathbf{c} \tag{3}$$

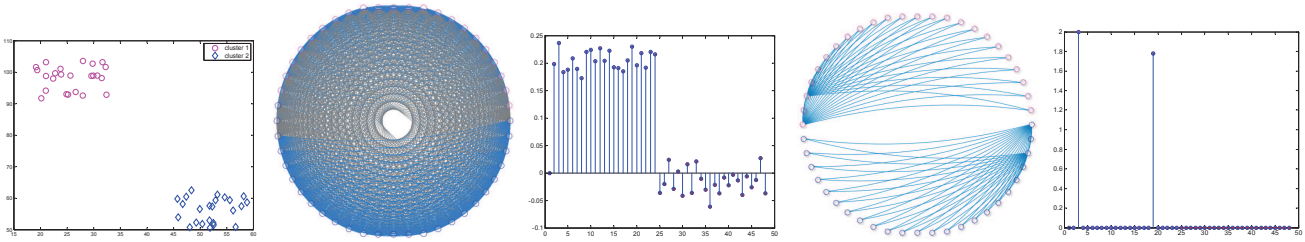
over $\mathbf{D} = [\mathbf{D}_x \ \mathbf{D}_{-x}]$ and $\mathbf{c}^* = \begin{bmatrix} \mathbf{c}_x^* \\ \mathbf{c}_{-x}^* \end{bmatrix}$ be partitioned according to the sets $\mathbf{D}_x$ and $\mathbf{D}_{-x}$. If

$$\sigma_{\min}(\mathbf{D}_x) \geq r_{-x} \cos\theta_{\min} \|\mathbf{D}_{-x}\|_{1,2}, \tag{4}$$

then $[\mathbf{c}_x^*]_{r_x,1} > [\mathbf{c}_{-x}^*]_{1,1}$. Here, $\sigma_{\min}(\mathbf{D}_x)$ is the smallest nonzero singular value of $\mathbf{D}_x$, $\theta_{\min}$ is the first principal angle between $\mathbf{D}_x$ and $\mathbf{D}_{-x}$, $\|\mathbf{D}_{-x}\|_{1,2}$ is the maximum ℓ_2-norm of the columns of $\mathbf{D}_{-x}$ and $[\mathbf{c}]_{r,1}$ denotes the r-th largest absolute value of the entries of $\mathbf{c}$.

According to the property of intra-subspace projection dominance, the coefficients over intra-subspace are always larger than those over the errors. Hence, we can eliminate the effect of the errors by keeping k largest entries and zeroing the other entries of the ℓ_p-norm-based representation, where k is the dimensionality of the corresponding subspace.

Figure 1 gives a toy example to illustrate the intra-subspace projection dominance in the ℓ_2-norm-based projection space, where the data points are sampled from two dependent subspaces corresponding to two clusters in $\mathbb{R}^2$. We plot the similarity graph (Figure 1(b) and Figure 1(d)



(a) (b) (c) (d) (e)

Figure 1: A toy example of the intra-subspace projection dominance in ℓ_2-norm-based projection space. (a) A given data sets come from two clusters, indicated by different shapes. Note that each cluster corresponds to a subspace, and the two subspaces are dependent. (b, c) The coefficients of a data point $\mathbf{x}$ and the similarity graph in ℓ_2-norm-based projection space. The first and the last 25 values in (c) correspond to the coefficients (similarity) over the intra-cluster and inter-cluster data points, respectively. (d, e) The coefficients of $\mathbf{x}$ and the similarity graph achieved by our method. For each data point, only the 2 largest coefficients are nonzero, corresponding to the projection over the base of $\mathbb{R}^2$. From (b) and (d), the inter-cluster data points connections are removed and the data are successfully separated into respective clusters.

) using the visualization toolkit NodeXL. In this example, the errors (i.e., the intersection between two dependent subspaces) lead to the connections between the inter-cluster data points and the weights of these connections are smaller than the edge weights between the intra-cluster data points.

Thresholding Ridge Regression for Robust Subspace Clustering

The property of intra-subspace projection dominance holds for ℓ_1, ℓ_2, and ℓ_∞ cases. However, we only present an algorithm by considering ℓ_2-norm case because ℓ_2-norm-minimization problem has a closed form solution.

Let $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ be a collection of data points located on a union of dependent or disjoint or independent subspaces $\{S_1, S_2, \ldots, S_L\}$ and $\mathbf{X}_i = [\mathbf{x}_1, \ldots, \mathbf{x}_{i-1}, \mathbf{0}, \mathbf{x}_{i+1}, \ldots, \mathbf{x}_n], (i = 1, \cdots, n)$ be the dictionary for $\mathbf{x}_i$, we aim to solve the following problem:

$$\min_{\mathbf{c}_i} \frac{1}{2}\|\mathbf{x}_i - \mathbf{X}_i\mathbf{c}_i\|_2^2 + \lambda\|\mathbf{c}_i\|_2^2, \tag{5}$$

where λ is a positive real number.

(5) is actually the well known ridge regression (Hoerl and Kennard 1970), whose optimal solution is $(\mathbf{X}_i^T\mathbf{X}_i + \lambda\mathbf{I})^{-1}\mathbf{X}_i^T\mathbf{x}_i$. However, this solution requires $O(mn^4)$ for n data points with dimensionality of m. To solve (5) efficiently, we rewrite it as

$$\min_{\mathbf{c}_i} \frac{1}{2}\|\mathbf{x}_i - \mathbf{X}\mathbf{c}_i\|_2^2 + \lambda\|\mathbf{c}_i\|_2^2, \ \text{ s.t. } \mathbf{e}_i^T\mathbf{c}_i = 0. \tag{6}$$

Using Lagrangian method, we have

$$\mathbb{L}(\mathbf{c}_i) = \frac{1}{2}\|\mathbf{x}_i - \mathbf{X}\mathbf{c}_i\|_2^2 + \lambda\|\mathbf{c}_i\|_2^2 + \gamma\mathbf{e}_i^T\mathbf{c}_i, \tag{7}$$

where γ is the Lagrangian multiplier. Clearly,

$$\frac{\partial\mathbb{L}(\mathbf{c}_i)}{\partial\mathbf{c}_i} = \left(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}\right)\mathbf{c}_i - \mathbf{X}^T\mathbf{x}_i + \gamma\mathbf{e}_i. \tag{8}$$

Let $\frac{\partial\mathbb{L}(\mathbf{c}_i)}{\partial\mathbf{c}_i} = 0$, we obtain

$$\mathbf{c}_i = \left(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}\right)^{-1}\left(\mathbf{X}^T\mathbf{x}_i - \gamma\mathbf{e}_i\right). \tag{9}$$

Algorithm 1: Robust Subspace Clustering via Thresholding Ridge Regression

Input: A collection of data points $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^n$ sampled from a union of linear subspaces $\{S_i\}_{i=1}^L$, the balance parameter λ and thresholding parameter k;

1: Calculate $\mathbf{P} = \left(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}\right)^{-1}$ and $\mathbf{Q} = \mathbf{P}\mathbf{X}^T$ and store them.
2: For each point $\mathbf{x}_i$, obtain its representation $\mathbf{c}_i$ via (11).
3: For each $\mathbf{c}_i$, eliminate the effect of errors in the projection space via $\mathbf{c}_i = \mathcal{H}_k(\mathbf{c}_i)$, where the hard thresholding operator $\mathcal{H}_k(\mathbf{c}_i)$ keeps k largest entries in $\mathbf{c}_i$ and zeroes the others.
4: Construct an affinity matrix by $\mathbf{W}_{ij} = |\mathbf{c}_{ij}| + |\mathbf{c}_{ji}|$ and normalize each column of $\mathbf{W}$ to have a unit ℓ_2-norm, where $\mathbf{c}_{ij}$ is the jth entry of $\mathbf{c}_i$.
5: Construct a Laplacian matrix $\mathbf{L} = \mathbf{D}^{-1/2}\mathbf{W}\mathbf{D}^{-1/2}$, where $\mathbf{D} = \text{diag}\{d_i\}$ with $d_i = \sum_{j=1}^n \mathbf{W}_{ij}$.
6: Obtain the eigenvector matrix $\mathbf{V} \in \mathbb{R}^{n\times L}$ which consists of the first L normalized eigenvectors of $\mathbf{L}$ corresponding to its L smallest nonzero eigenvalues.
7: Perform k-means clustering algorithm on the rows of $\mathbf{V}$.

Output: The cluster assignment of $\mathbf{X}$.

Multiplying both sides of (9) by $\mathbf{e}_i^T$, and since $\mathbf{e}_i^T\mathbf{c}_i = 0$, it holds that

$$\gamma = \frac{\mathbf{e}_i^T\left(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}\right)^{-1}\mathbf{X}^T\mathbf{x}_i}{\mathbf{e}_i^T\left(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}\right)^{-1}\mathbf{e}_i}. \tag{10}$$

Substituting γ into (10), the optimal solution is given by

$$\mathbf{c}_i^* = \mathbf{P}\left[\mathbf{X}^T\mathbf{x}_i - \frac{\mathbf{e}_i^T\mathbf{Q}\mathbf{x}_i\mathbf{e}_i}{\mathbf{e}_i^T\mathbf{P}\mathbf{e}_i}\right], \tag{11}$$

where $\mathbf{Q} = \mathbf{P}\mathbf{X}^T$, $\mathbf{P} = \left(\mathbf{D}^T\mathbf{D} + \lambda\mathbf{I}\right)^{-1}$, and the union of $\mathbf{e}_i$ $(i = 1, \cdots, n)$ is the standard orthogonal basis of $\mathbb{R}^n$, i.e., all entries in $\mathbf{e}_i$ are zeroes except the i-th entry is one.

After projecting the data set into the linear space spanned by itself via (11), the proposed algorithm, named Thresholding Ridge Regression (TRR), handles the errors by performing a hard thresholding operator $\mathcal{H}_k(\cdot)$ over $\mathbf{c}_i$, where $\mathcal{H}_k(\cdot)$ keeps k largest entries in $\mathbf{c}_i$ and zeroing the others. Generally, the optimal k equals to the dimensionality of corresponding subspace. Algorithm 1 summarizes our approach and steps 5–7 are normalized spectral clustering (Ng, Jordan, and Weiss 2002).

Related Works

Our work is related to several existing representation-based subspace clustering methods mainly including Sparse Subspace Clustering (SSC) (Elhamifar and Vidal 2013) and Low Rank Representation (LRR) (Liu et al. 2013).

SSC constructs a similarity graph using the sparse representation of a given data set. To handle the errors that probably exist in the data set, SSC formulates the errors into its objective function as follows:

$$\begin{aligned} &\min_{\mathbf{C},\mathbf{E},\mathbf{Z}} \|\mathbf{C}\|_1 + \lambda_E\|\mathbf{E}\|_1 + \lambda_Z\|\mathbf{Z}\|_F \\ &\text{s.t. } \mathbf{X} = \mathbf{X}\mathbf{C} + \mathbf{E} + \mathbf{Z}, \mathbf{C}\mathbf{1}^T = \mathbf{1}, \text{diag}(\mathbf{C}) = 0, \end{aligned} \quad (12)$$

where $\mathbf{C} \in \mathbb{R}^{n\times n}$ is the sparse representation of the data set $\mathbf{X} \in \mathbb{R}^{m\times n}$, $\mathbf{E}$ corresponds to the sparse outlying entries, $\mathbf{Z}$ denotes the reconstruction errors owing to the limited representational capability, and the parameters λ_E and λ_Z balance the terms of the objective function. If the data located into the linear subspace, then the affine constraint $\mathbf{C}\mathbf{1}^T = \mathbf{1}$ could be removed.

Different from SSC, LRR uses the lowest-rank representation instead of the sparest one to build the graph via

$$\min \ \|\mathbf{C}\|_* + \lambda\|\mathbf{E}\|_p \quad \text{s.t. } \mathbf{x} = \mathbf{X}\mathbf{C} + \mathbf{E}, \quad (13)$$

where $\|\mathbf{C}\|_* = \sum_i \sigma_i(\mathbf{C})$, $\sigma_i(\mathbf{C})$ is the i-th singular value of $\mathbf{C}$, and $\|\cdot\|_p$ could be chosen as $\ell_{2,1}$-, ℓ_1-, or Frobenius-norm. The choice of the norm only depends on which kind of error is assumed in the data set. Specifically, $\ell_{2,1}$-norm is usually adopted to depict sample-specific corruption and outliers, ℓ_1-norm is used to characterize random corruption, and Frobenius norm is used to describe the Gaussian noise.

From (12) and (13), it is easy to find that SSC and LRR remove the pre-specified errors from the input space. This strategy of errors removing has been adopted by numerous works such as (Liu et al. 2012; Liu and Yan 2011; Wang and Xu 2013; Deng et al. 2013). In contrast, our approach eliminates the effect of errors from the projection space. The proposed method takes a different way to handle the errors and does not suffer from the limitation of estimating the structure of the errors as SSC and LRR did.

Experimental Verification and Analysis

In this section, we investigate the performance of TRR for robust face clustering with respect to clustering quality, robustness, and computational efficiency.

Experimental Configurations

We compared TRR[1] with several recently-proposed subspace clustering algorithms, i.e., SSC (Elhamifar and Vidal 2013), LRR (Liu et al. 2013), and two variants of LSR (LSR1 and LSR2) (Lu et al. 2012). Moreover, we used the coefficients of Locally Linear Embedding (LLE) (Roweis and Saul 2000) to build the similarity graph for subspace clustering as (Cheng et al. 2010) did, denoted as LLE-graph.

For fair comparisons, we performed the same spectral clustering algorithm (Ng, Jordan, and Weiss 2002) on the graphs built by the tested algorithms and reported their best results with the tuned parameters. For the SSC algorithm, we experimentally found an optimal α from 1 to 50 with an interval of 1. For LRR, the optimal λ was found from 10^{-6} to 10 as suggested in (Liu et al. 2013). For LSR and TRR, the optimal λ was chosen from 10^{-7} to 1. Moreover, a good k was found from 3 to 14 for TRR and from 1 to 100 for LLE-graph.

Evaluation metrics: Two popular metrics, *Accuracy* (or called *Purity*) and Normalized Mutual Information (*NMI*) (Cai, He, and Han 2005), are used to evaluate the clustering quality. The value of *Accuracy* or *NMI* is 1 indicates perfect matching with the ground truth, whereas 0 indicates perfect mismatch.

Data sets: We used two popular facial databases, i.e., Extended Yale Database B (Georghiades, Belhumeur, and Kriegman 2001) (ExYaleB) and AR database (Martinez and Benavente 1998). ExYaleB contains 2414 frontal-face images with size 192×168 of 38 subjects (about 64 images per subject), while the first 58 samples per subject were used and each image was downsized to 54×48. Moreover, we tested a subset of AR which consists of 1400 clean faces distributed over 50 male subjects and 50 female subjects. All the AR images were downsized and normalized from 165×120 to 55×40. For computational efficiency, we performed Principle Component Analysis (PCA) to reduce the dimensionality of the data by reserving 98% energy.

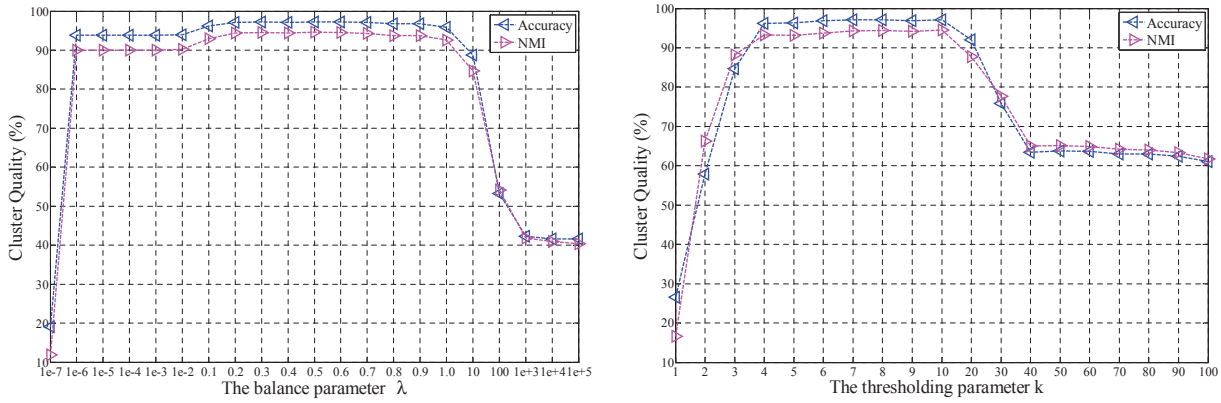
Model Selection

TRR has two parameters, the balance parameter λ and the thresholding parameter k. The values of these parameters depend on the data distribution. In general, a bigger λ is more suitable to characterize the corrupted images and k equals to the dimensionality of the corresponding subspace.

To examine the influence of these parameters, we carried out some experiments using a subset of ExYaleB which contains 580 images from the first 10 individuals. We randomly selected a half of samples to corrupt using white Gaussian noise via $\tilde{\mathbf{y}} = \mathbf{x} + \rho\mathbf{n}$, where $\tilde{\mathbf{y}} \in [0\ 255]$, $\mathbf{x}$ denotes the chosen sample, $\rho = 10\%$ is the corruption ratio, and $\mathbf{n}$ is the noise following the standard normal distribution.

Figure 2 shows that: 1) while λ increases from 0.1 to 1.0 and k ranges from 4 to 9, *Accuracy* and *NMI* almost remain unchanged; 2) the thresholding parameter k is helpful to improve the robustness of our model. This verifies the

[1]The codes can be downloaded at the authors' website http://www.machineilab.org/users/pengxi/.



(a) (b)

Figure 2: The influence the parameters of TRR. (a) The influence of λ, where $k = 7$. (b) The influence of k, where $\lambda = 0.7$.

correctness of our claim that the trivial coefficients correspond to the codes over the errors. 3) a larger k will impair the discrimination of the model, whereas a smaller k cannot provide enough representative ability. Indeed, the optimal value of k can be found around the intrinsic dimensionality of the corresponding subspace. According to (Costa and Hero 2004), the intrinsic dimensionality of the first subject of Extended Yale B is 6, which shows that the optimal k of TRR equals to the dimension of the corresponding subspace.

Clustering on Clean Images

In this section, we evaluate the performance of TRR using 1400 clean AR images (167 dimension). The experiments were carried out on the first L subjects of the data set, where L increases from 20 to 100. Figure 3 shows that: 1) TRR is more competitive than the other examined algorithms, e.g., with respect to $L = 100$, the *Accuracy* of TRR is at least, 1.8% higher than that of LSR1, 2.7% higher than that of LSR2, 24.5% higher than that of SSC , 8.8% higher than that of LRR and 42.5% higher than that of LLE-graph. 2) With increasing L, the NMI of TRR almost remain unchanged, slightly varying from 93.0% to 94.3%. The possible reason is that NMI is robust to the data distribution (increasing subject number).

Clustering on Corrupted Images

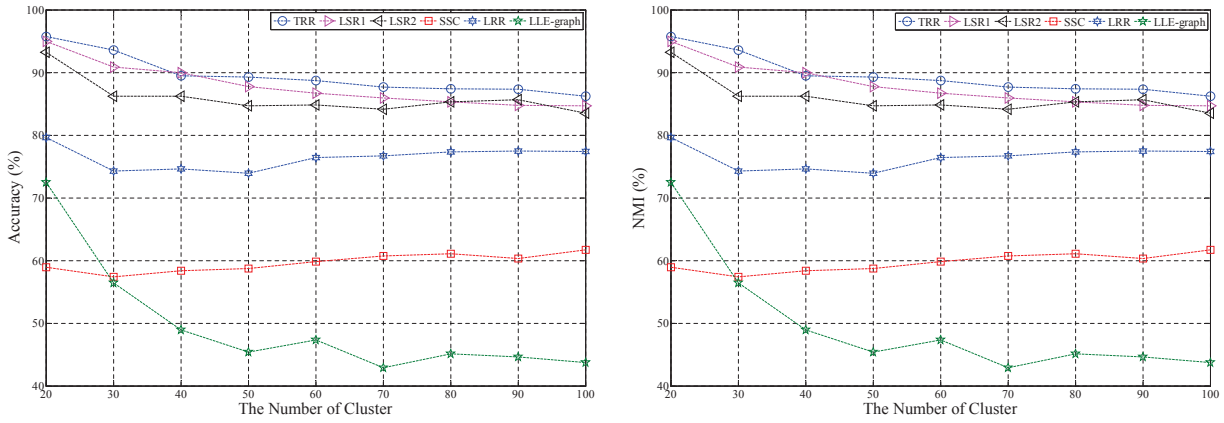
Our error removing strategy can improve the robustness of TRR without the prior knowledge of the errors. To verify this claim, we test the robustness of TRR using ExYaleB over 38 subjects. For each subject of the database, we randomly chose a half of images (29 images per subject) to corrupt by white Gaussian noise or random pixel corruption (see Figure 4), where the former is additive and the latter is non-additive. In details, for the image $\mathbf{x}$, we added white Gaussian noise and increased the corruption ratio ρ from 10% to 90%. For the random pixel corruption, we replaced the value of a percentage of pixels randomly selected from the image with the values following a uniform distribution over $[0,\ p_{max}]$, where p_{max} is the largest pixel value of $\mathbf{x}$. To avoid randomness, we produced ten data sets beforehand and then performed the evaluated algorithms over these data partitions.

Figure 4: The samples with real possible corruptions. Top row: the images with white Gaussian noise; Bottom row: the images with random pixel corruption. From left to right, the corruption rate increases from 10% to 90% (with an interval of 20%).

From Table 2, we have the following conclusions: (1) all the investigated methods perform better in the case of white Gaussian noise. The result is consistent with a widely-accepted conclusion that non-additive corruptions are more challenging than additive ones in pattern recognition. (2) TRR is more robust than LSR1, LSR2, SSC, LRR and LLE-graph by a considerable performance margin. For example, with respect to white Gaussian noise, the performance gain in *Accuracy* between TRR and LSR2 varied from 14.0% to 22.8%; with respect to random pixel corruption, the performance gain varied from 5.0% to 13.2%.

Running Time

In this section, we report the time costs of these algorithms for clustering and building the similarity graph. Table 3 reports the time costs obtained by averaging the elapsed CPU time over 5 independent experiments for each algorithm. We carried out the experiments using 2204 images from Extended Yale Database B over 38 subjects and 1400 samples from AR database over 100 subjects. From the result, TRR is remarkably faster than the other methods to get the clustering results.



(a) (b)

Figure 3: The clustering quality (*Accuracy* and *NMI*) of different algorithms on the first L subjects of AR data set.

Table 2: The performance of TRR, LSR (Lu et al. 2012), SSC (Elhamifar and Vidal 2013), LRR (Liu et al. 2013), and LLE-graph (Roweis and Saul 2000) on the **ExYaleB (116 dimension)**. ρ denotes the corrupted ratio; The values in the parentheses denote the optimal parameters for the reported *Accuracy*, i.e., TRR (λ, k), LSR (λ), SSC(α), LRR (λ), and LLE-graph (k).

Corruption	ρ	TRR		LSR1		LSR2		SSC		LRR		LLE-graph	
		Accuracy	NMI	Accuracy	NMI	Accuracy	NMI	Accuracy	NMI	Accuracy	NMI	Accuracy	NMI
Clean Data	0	**86.78**(1.0,5)	**92.84**	76.50(1e-3)	80.59	74.59(1e-4)	79.05	68.60(8)	75.04	85.25(10)	91.19	51.82(3)	61.61
White Gaussian Noise	10	**89.25**(1e-4,6)	**92.71**	72.28(1e-2)	78.36	73.19(1e-4)	78.52	68.38(8)	74.25	87.79(0.7)	92.12	47.82(5)	69.40
	30	**88.70**(0.7,6)	**92.18**	71.14(1e-4)	75.93	74.55(1e-4)	78.30	66.02(10)	71.50	81.31(5.0)	86.05	46.51(6)	59.84
	50	**86.57**(0.7,4)	**90.43**	63.61(1e-2)	70.58	63.16(1e-4)	71.79	55.85(22)	61.99	84.96(0.4)	79.15	37.48(5)	52.10
	70	**74.32**(0.6,7)	**77.70**	52.72(1e-3)	63.08	51.54(1e-4)	63.02	49.00(30)	58.64	60.66(0.7)	69.57	32.76(5)	44.96
	90	**56.31**(0.6,7)	**63.43**	43.15(0.1)	55.73	42.33(1e-4)	55.64	44.10(36)	51.79	49.96(0.2)	57.90	29.81(5)	42.90
Random Pixels Corruption	10	**82.76**(1.0,4)	**88.64**	72.35(1e-3)	77.09	72.35(1e-4)	77.11	64.97(48)	68.40	78.68(0.3)	87.19	46.82(6)	59.26
	30	**68.97**(0.7,7)	**75.89**	56.48(1e-4)	63.19	56.48(1e-2)	63.28	56.13(49)	59.96	60.80(0.6)	67.47	33.26(5)	42.33
	50	**48.15**(1.0,6)	**56.67**	42.15(1e-4)	50.53	43.16(0.4)	53.09	45.60(39)	51.69	38.61(0.2)	49.93	19.51(5)	27.77
	70	**34.98**(1e-2,5)	**45.56**	27.86(1e-3)	35.88	27.50(1e-2)	35.73	34.71(48)	41.14	30.54(0.2)	38.13	13.39(6)	18.82
	90	**30.04**(1e-4,4)	**38.39**	19.78(1e-3)	28.00	19.19(0.1)	28.22	20.78(47)	30.03	19.01(0.2)	29.16	14.07(6)	23.04

Table 3: Average running time (seconds).

Algorithms	Total costs		Time for building graph	
	AR	ExYaleB	AR	ExYaleB
TRR	**307.69**	**230.78**	10.69	28.76
LSR1	1190.38	653.91	1.39	1.25
LSR2	1255.98	641.49	1.21	**0.50**
SSC	1299.41	584.14	67.49	242.71
LRR	1295.6	849.66	44.74	118.45
LLE-graph	2030.51	527.71	**1.10**	1.41

Conclusions

Under the framework of graph-oriented learning (Yan et al. 2007), most of the recent approaches achieve the robust clustering result by removing the errors from the original space and then build the neighboring relation based on a 'clean' data set. In contrast, we propose and prove that it is possible to eliminate the effect of the errors from the linear projection space (representation). Based on this mathematically tractable property, we present a simple but effective method for robust subspace clustering. Extensive experimental results validate the good performance of our approach.

The work might be extended or improved from the following aspects. Except subspace clustering, similarity graph is also a fundamental problem in subspace learning. Therefore, the proposed method can be extended for feature extraction. Moreover, it is interesting to develop supervised or semi-supervised method based on our framework.

Acknowledgments

The authors would like to thank the anonymous reviewers for their valuable comments and suggestions to improve the quality of this paper. This work was supported by Agency for Science, Technology, and Research (A*STAR), Singapore under SERC Grant 12251 00002 and National Nature Science Foundation of China under grant No.61432012.

References

Cai, D.; He, X. F.; and Han, J. W. 2005. Document clustering using locality preserving indexing. *IEEE Transactions on Knowledge and Data Engineering* 17(12):1624–1637.

Chen, G. L., and Lerman, G. 2009. Spectral curvature clustering (SCC). *International Journal of Computer Vision* 81(3):317–330.

Cheng, B.; Yang, J.; Yan, S.; Fu, Y.; and Huang, T. 2010. Learning with L1-graph for image analysis. *IEEE Transactions on Image Processing* 19(4):858–866.

Costa, J., and Hero, A. 2004. Geodesic entropic graphs for dimension and entropy estimation in manifold learning. *IEEE Transactions on Signal Processing* 52(8):2210–2221.

Costeira, J. P., and Kanade, T. 1998. A multibody factorization method for independently moving objects. *International Journal of Computer Vision* 29(3):159–179.

Deng, Y.; Dai, Q.; Liu, R.; Zhang, Z.; and Hu, S. 2013. Low-rank structure learning via nonconvex heuristic recovery. *IEEE Transactions on Neural Networks and Learning Systems* 24(3):383–396.

Elhamifar, E., and Vidal, R. 2013. Sparse subspace clustering: Algorithm, theory, and applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35(11):2765–2781.

Gear, C. W. 1998. Multibody grouping from motion images. *International Journal of Computer Vision* 29(2):133–150.

Georghiades, A. S.; Belhumeur, P. N.; and Kriegman, D. J. 2001. From few to many: Illumination cone models for face recognition under variable lighting and pose. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 23(6):643–660.

Ho, J.; Yang, M.-H.; Lim, J.; Lee, K.-C.; and Kriegman, D. 2003. Clustering appearances of objects under varying illumination conditions. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, 11–18. IEEE.

Hoerl, A. E., and Kennard, R. W. 1970. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* 12(1):55–67.

Lin, Z.; Liu, R.; and Su, Z. 2011. Linearized alternating direction method with adaptive penalty for low-rank representation. In *Proc. of Neural Information Processing Systems*, volume 2, 6.

Liu, G. C., and Yan, S. C. 2011. Latent low-rank representation for subspace segmentation and feature extraction. *IEEE International Conference on Computer Vision* 1615–1622.

Liu, R.; Lin, Z.; la Torre, F. D.; and Su, Z. 2012. Fixed-rank representation for unsupervised visual learning. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, 598–605. IEEE.

Liu, G.; Lin, Z.; Yan, S.; Sun, J.; Yu, Y.; and Ma, Y. 2013. Robust recovery of subspace structures by low-rank representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35(1):171–184.

Lu, C.-Y.; Min, H.; Zhao, Z.-Q.; Zhu, L.; Huang, D.-S.; and Yan, S. 2012. Robust and efficient subspace segmentation via least squares regression. In *Proc. of European Conference on Computer Vision*, 347–360. Springer.

Martinez, A., and Benavente, R. 1998. The AR face database.

Ng, A.; Jordan, M.; and Weiss, Y. 2002. On spectral clustering: Analysis and an algorithm. In *Proc. of Advances in Neural Information Processing Systems*, volume 14, 849–856.

Peng, X.; Zhang, L.; and Yi, Z. 2013. Scalable sparse subspace clustering. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, 430–437. IEEE.

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

Vidal, R.; Ma, Y.; and Sastry, S. 2005. Generalized principal component analysis (GPCA). *IEEE Transactions on Pattern Analysis and Machine Intelligence* 27(12):1945–1959.

Vidal, R. 2011. Subspace clustering. *IEEE Signal Processing Magazine* 28(2):52–68.

Wang, Y., and Xu, H. 2013. Noisy sparse subspace clustering. In *Proc. of the International Conference on Machine Learning*, volume 28, 89–97.

Yan, J., and Pollefeys, M. 2006. A general framework for motion segmentation: Independent, articulated, rigid, non-rigid, degenerate and non-degenerate. In *Proc. of European Conference on Computer Vision*, 94–106. Springer.

Yan, S. C.; Xu, D.; Zhang, B. Y.; Zhang, H. J.; Yang, Q.; and Lin, S. 2007. Graph embedding and extensions: A general framework for dimensionality reduction. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 29(1):40–51.

Zhang, T.; Szlam, A.; Wang, Y.; and Lerman, G. 2012. Hybrid linear modeling via local best-fit flats. *International Journal of Computer Vision* 100(3):217–240.