Smooth Representation Clustering

Han Hu[1], Zhouchen Lin[2], Jianjiang Feng[1] and Jie Zhou[1]
[1]State Key Laboratory on Intelligent Technology and Systems, TNList
Department of Automation, Tsinghua University
[2]Key Lab. of Machine Perception, School of EECS, Peking University
huh04@mails.thu.edu.cn, zlin@pku.edu.cn, {jfeng,jzhou}@tsinghua.edu.cn

Abstract

Subspace clustering is a powerful technology for clustering data according to the underlying subspaces. Representation based methods are the most popular subspace clustering approach in recent years. In this paper, we analyze the grouping effect of representation based methods in depth. In particular, we introduce the enforced grouping effect conditions, which greatly facilitate the analysis of grouping effect. We further find that grouping effect is important for subspace clustering, which should be explicitly enforced in the data self-representation model, rather than implicitly implied by the model as in some prior work. Based on our analysis, we propose the SMooth Representation (SMR) model. We also propose a new affinity measure based on the grouping effect, which proves to be much more effective than the commonly used one. As a result, our SMR significantly outperforms the state-of-the-art ones on benchmark datasets.

1. Introduction

In many computer vision and machine learning problems, the data can be viewed as points drawn from multiple low-dimensional subspaces, with each subspace corresponding to one category or class, e.g., point trajectories of moving objects captured by an affine camera [24], images of several subjects under varying illumination or under different poses [6], and local patches or texture features of pixels/superpixels on an image [20]. A basic task for processing such kind of data is to cluster the points according to the underlying subspace. Such a task is called subspace clustering [26].

1.1. Related Work

Existing methods for subspace clustering can be roughly grouped into three categories: algebra based, statistics based, and spectral clustering based [26].

Most of the early studies on subspace clustering are algebra or statistics based. The two most well known algebraic methods are perhaps the shape interaction matrix (SIM) [2] and generalized principal component analysis (GPCA) [27]. Although with elegant formulations, in general the performance of these methods drops quickly in the presence of noise, degeneracy, or partially coupled subspaces. The statistics based methods treat subspace clustering as a mixed data inference problem and thus some popular methods from the more general statistical learning field can be used, such as random sample consensus (RANSAC) [5] and expectation maximization (EM) [12]. Although there have been several new techniques developed to improve the criterion (e.g., agglomerative lossy compression (ALC) [22]), model selection (e.g., Branch and Bound (BB) [10]), the performance of these methods is limited by their dependency on estimating the exact subspace models.

Many of the recent studies focus on the spectral clustering based methods [21, 30, 3, 16, 29, 19, 18, 17]. The major differences among these methods lie in the way they build the affinity matrix. A direct way is to compute affinity matrix from existing algebraic methods [21] or by defining a point-to-subspace or subspace-to-subspace distance metric [30]. More recently, many works apply the self-representation idea to compute affinities [3, 16, 29, 19, 18, 17], i.e., represent every sample by a linear combination of other samples, which result in state-of-the-art performance. These methods first compute a self-representation matrix Z^* by solving

$$\begin{aligned} \min_{Z} \quad & \alpha\|X - A(X)Z\|_l + \Omega(X, Z), \\ s.t. \quad & Z \in \mathcal{C}, \end{aligned} \tag{1}$$

where $X \in \mathbb{R}^{d\times n}$ is the data matrix with each column being a sample vector, $A(X)$ is a dictionary matrix which could be learnt or be simply set as $A(X) = X$, $\|\cdot\|_l$ is a proper norm, $\Omega(X, Z)$ and $\mathcal{C}$ are the regularizer and constraint set on Z, respectively, and $\alpha > 0$ is a trade-off parameter. Then Z^* is used to compute an affinity matrix, e.g.,

Table 1. The choices of $\Omega(X, Z)$, $\|\cdot\|_l$, and $\mathcal{C}$ of existing representation based methods.

	$\Omega(X,Z)$	$\Vert\cdot\Vert_l$	$\mathcal{C}$
SSC [3, 4]	$\Vert Z\Vert_1$	$\Vert\cdot\Vert_1$	$\{Z\mid Z_{ii}=0\}$
LRR [16, 15]	$\Vert Z\Vert_*$	$\Vert\cdot\Vert_{2,1}$	$\emptyset$
SSQP [29]	$\Vert Z^TZ\Vert_1$	$\Vert\cdot\Vert_F^2$	$\{Z\mid Z\geq 0,$ $Z_{ii}=0\}$
MSR [19]	$\Vert Z\Vert_1+\delta\Vert Z\Vert_*$	$\Vert\cdot\Vert_{2,1}$	$\{Z\mid Z_{ii}=0\}$
LSR [18]	$\Vert Z\Vert_F^2$	$\Vert\cdot\Vert_F^2$	$\emptyset$
LSR-Z [18]	$\Vert Z\Vert_F^2$	$\Vert\cdot\Vert_F^2$	$\{Z\mid Z_{ii}=0\}$
CASS [17]	$\sum_i \Vert X\text{diag}(\mathbf{z}_i)\Vert_*$	$\Vert\cdot\Vert_F^2$	$\emptyset$

$(|Z^*|+|Z^{*T}|)/2$, which is further input into a spectral clustering algorithm [23] to produce the final clustering result. The existing methods distinguish each other by employing different regularization terms $\Omega(Z)$ or constraint sets $\mathcal{C}$. Table 1 summarizes the choices of Ω, $\|\cdot\|_l$, and $\mathcal{C}$ of existing representation based methods, where $\|\cdot\|_1$ is the ℓ_l norm, i.e., sum of the absolute values of all entries, $\|\cdot\|_{2,1}$ is the $\ell_{2,1}$ norm, i.e., sum of the ℓ_2 norms of the column vectors, $\|\cdot\|_*$ is the nuclear norm, i.e., sum of singular values, and $\|\cdot\|_F$ is the Frobenious norm, i.e., square root of the sum of squared entries. Since the term $\|X - XZ\|_l$ concerns about representation error and it is not the main focus of our paper, we use $\|X - XZ\|_F^2$ in the sequel.

1.2. Contributions

Lu et al. [18, 17] discovered that Least Squares Regression (LSR) [18] and Correlation Adaptive Subspace Segmentation (CASS) [17] models both have the grouping effect defined as follows.

Definition 1 (Grouping Effect): *Given a set of d-dimensional data points $X = [\mathbf{x}_1, \cdots, \mathbf{x}_n] \in \mathbb{R}^{d\times n}$, a self-representation matrix $Z = [\mathbf{z}_1, \cdots, \mathbf{z}_n] \in \mathbb{R}^{n\times n}$ has grouping effect if $\|\mathbf{x}_i - \mathbf{x}_j\|_2 \to 0 \Rightarrow \|\mathbf{z}_i - \mathbf{z}_j\|_2 \to 0$, $\forall i \neq j$.*

Inspired by [18, 17], we analyze the grouping effect of representation based method in depth. In particular, we introduce the enforced grouping effect (EGE) conditions, which can greatly facilitate the analysis of grouping effect of a representation based method. By the EGE conditions, we easily find new models that also have the grouping effect. In contrast, Lu et al. [18, 17] proved the grouping effect of LSR and CASS in a case-by-case way. Their proofs are specific and hence cannot be applied to other models.

We further find that grouping effect is actually important for subspace clustering. So we propose to explicitly enforce the grouping effect in the representation model. In contract, prior work [18, 17] only passively discovered that LSR and CASS have the grouping effect.

Finally, based on our analysis we propose the SMooth Representation (SMR) model. We also propose a novel affinity measure based on the grouping effect, which proves to be much more effective than the commonly used measure $(|Z^*| + |Z^{*T}|)/2$. Our experiments on benchmark datasets show that our SMR significantly outperforms the state-of-the-art approaches.

2. Grouping Effect

The grouping effect was first explicitly stated by Lu et al. [18, 17], who showed that in LSR and CASS when the samples are close to each other their representation coefficients are also close to each other. Their proofs are specific for LSR and CASS and cannot be applied to other models. In this section, we analyze the grouping effect of reconstruction based models in depth. We first introduce the Enforced Grouping Effect (EGE) conditions, which can help us identify the grouping effect easily. Then we investigate why grouping effect helps subspace clustering.

2.1. Enforced Grouping Effect Conditions

We introduce general sufficient conditions for the grouping effect as follows.

Definition 2 (Enforced Grouping Effect Conditions): The EGE conditions on problem (1) are:

(1) $A(X)$ is continuous with respect to X and $\Omega(X, Z)$ is continuous with respect to X and $Z \in \mathcal{C}$;

(2) Problem (1) has a unique solution Z^* and Z^* is not an isolated point of $\mathcal{C}$.

(3) $Z \in \mathcal{C}$ if and only if $ZP \in \mathcal{C}$, and $\Omega(X, Z) = \Omega(XP, ZP)$, for all permutation matrix P.

(4) $A(XP) = A(X)P$, $Z \in \mathcal{C}$ if and only if $P^TZP \in \mathcal{C}$, and $\Omega(X, Z) = \Omega(XP, P^TZP)$, for all permutation matrix P.

Then we have the following lemma.

Lemma 1: *If Problem (1) satisfies the EGE conditions (1) and (2), then its optimal solution Z^* is a continuous function of X.*

Proof: It is obvious that Z^* can be regarded as a function of X according to EGE condition (2). In the following, we prove the continuity of Z^* w.r.t X.

Suppose Z^* is discontinuous with respect to X and $X = X_1$ is a discontinuity point. We have: $\exists \varepsilon_1 > 0, \forall \delta_1 > 0$, there exist $\|X_2 - X_1\|_F < \delta_1$ that $\|Z_2^* - Z_1^*\|_F > \varepsilon_1$.

Denote $f(X, Z) = \|X - A(X)Z\|_l + \alpha\Omega(X, Z)$. Since Problem (1) has a unique solution, we have $\exists \varepsilon_2 > 0, f(X_2, Z_2^*) < f(X_2, Z_1^*) - \varepsilon_2$. According to EGE condition (1), $f(X, Z)$ is continuous with respect to X: for any $\varepsilon_3 > 0$, there exists $\delta_2 > 0$ such that for all $\|X - X_2\|_F < \delta_2 \Rightarrow |f(X, Z_2^*) - f(X_2, Z_2^*)| < \varepsilon_3$ and there exists some number $\delta_3 > 0$ such that for all $\|X - X_2\|_F < \delta_3 \Rightarrow |f(X, Z_1^*) - f(X_2, Z_1^*)| < \varepsilon_3$.

Suppose $2\varepsilon_3 < \varepsilon_2$, $\delta_1 \leq \delta_2$, and $\delta_1 \leq \delta_3$. We have

$$\begin{aligned} f(X_1, Z_2^*) & < f(X_2, Z_2^*) + \varepsilon_3 \\ & < f(X_2, Z_1^*) + \varepsilon_3 - \varepsilon_2 \\ & < f(X_1, Z_1^*) + 2\varepsilon_3 - \varepsilon_2 \\ & < f(X_1, Z_1^*). \end{aligned} \tag{2}$$

Eq. (2) indicates that Z_2^* is a better solution of Problem (1) when $X = X_1$, which is a contradiction. Hence the continuity of Z^* w.r.t. X is proved. □

Then we have the following proposition.

Proposition 1: *The optimal solution Z^* to problem (1) has grouping effect if EGE conditions (1), (2), and (3) are satisfied.*

Proof: We first instantiate X by X_1. Consider two sufficiently close points $\mathbf{x}_i$ and $\mathbf{x}_j$ in X_1. For simplicity we informally write $\|\mathbf{a} - \mathbf{b}\|_F \to 0$ to denote that $\mathbf{a}$ and $\mathbf{b}$ are close to each other. Exchanging the two columns $\mathbf{x}_i$ and $\mathbf{x}_j$, we get a new data matrix $X_2 = X_1 P_{ij}$, where P_{ij} is the permutation matrix by exchanging the i^{th} and j^{th} columns of the identity matrix. It is obvious that $\|X_2 - X_1\|_F \to 0$ and $A(X_2) - A(X_1)\|_F \to 0$.

Given EGE condition (3), it is easy to check that $Z_2^* = Z_1^* P$ is the unique optimal solution of problem (1) when $X = X_2$. By Lemma 1, we have that $\|X_2 - X_1\|_F \to 0 \Rightarrow \|Z_2^* - Z_1^*\|_F \to 0$. Therefore, $\|\mathbf{z}_i - \mathbf{z}_j\|_2 \to 0$ as Z_2^* and Z_1^* only differ in the i^{th} and j^{th} columns. □

We now check the grouping effect of existing representation based methods listed in Table 1 by the above EGE conditions. SSC [3, 4] does not satisfy the conditions. Indeed, it does not have the grouping effect. For example, considering $X = [\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3]$ with $\mathbf{x}_1 = \mathbf{x}_2 = \mathbf{x}_3$, any permutation matrix satisfying $\text{diag}(Z) = 0$ would be the optimal solution to SSC. For LSR, all EGE conditions are satisfied. So it has the grouping effect. Figures 1(a)-(c) also exemplify our above observations. For LRR [16, 15], it is obvious that EGE conditions (1) and (3) are satisfied. The uniqueness of the optimal solution to LRR can also be proven, as stated in Proposition 2. Hence, LRR has the grouping effect. The optimal solution of CASS [17] is unique according to [8], and it also has the grouping effect.

Proposition 2: *LRR has a unique optimal solution.*

Proof: Please find it in the supplementary material. □

Proposition 1 not only provides a way to determine the grouping effect of existing methods, it may also help us to design new methods with grouping effect. For example, the following families of methods have grouping effects.

Proposition 3: *Problems (1) with the following $\Omega(Z)$ and $\mathcal{C}$ have grouping effect:*

(1) $\Omega(Z) = \sum_{j=1}^{n} \left(\sum_{i=1}^{n} |Z_{ij}|^p \right)^q, p > 1, q \geq \frac{1}{p}, \mathcal{C} = \emptyset.$

(2) $\Omega(Z) = tr((ZHZ^T)^p), H \succ 0, p \geq 1/2, \mathcal{C} = \emptyset.$

(3) $\Omega(Z) = tr((Z^T HZ)^p), H \succ 0, p \geq 1/2, \mathcal{C} = \emptyset.$

Proof: We put it in the supplementary material. □

It should be noted that the constraint set $\mathcal{C} = \{Z_{ii} = 0, \forall i\}$, as used by some existing methods, such as SSC [4], SSQP [29], MSR [19] and LSR-Z [18], does not satisfy the EGE condition (3). Accordingly, these methods do not have the grouping effect in a strict sense. However, these methods also perform well. So we generalize the concept of grouping effect as follows in order to comply with such an observation.

Definition 3 (Permutated Grouping Effect): *Given a set of data points $X = [\mathbf{x}_1, \ldots, \mathbf{x}_n] \in \mathbb{R}^{d\times n}$, a self-representation matrix $Z = [\mathbf{z}_1, \ldots, \mathbf{z}_n] \in \mathbb{R}^{n\times n}$ has permuted grouping effect if $\|\mathbf{x}_i - \mathbf{x}_j\|_2 \to 0 \Rightarrow \|P_{ij}^T \mathbf{z}_i - \mathbf{z}_j\|_2 \to 0$, where P_{ij}^T is the permutation matrix by exchanging the i^{th} and the j^{th} rows of the identity matrix.*

Then we have the following proposition.

Proposition 4: *The optimal solution Z^* to problem (1) has permutated grouping effect if EGE conditions (1), (2), and (4) are satisfied.*

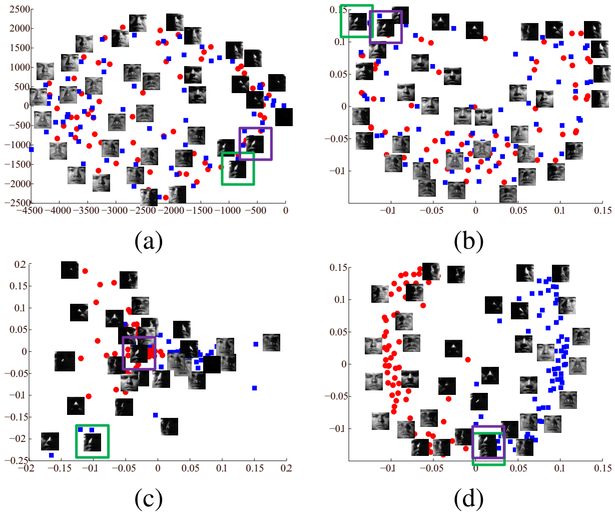
Proof: Similar to the proof of Proposition 1, we form a new data matrix $X_2 = X_1 P_{ij}$. By EGE conditions (2) and (4), $Z_2^* = P_{ij}^T Z_1^* P_{ij}$ is the unique optimal solution to problem (1) when $X = X_2$. If $\|\mathbf{x}_i - \mathbf{x}_j\|_F \to 0$, then by Lemma 1 we have $\|Z_2^* - Z_1^*\|_F \to 0$, implying $\|P_{ij}^T \mathbf{z}_i - \mathbf{z}_j\|_2 \to 0$. □

One may check that the constraint set $\mathcal{C} = \{Z_{ii} = 0, \forall i\}$ satisfy the EGE condition (4). So by Proposition 4, SSQP [29], MSR [19] and LSR-Z [18] have the permutated grouping effect.

2.2. Why Grouping Effect?

It was claimed by Lu et al. [18] that the effectiveness of LSR comes from the grouping effect. However, in [18] there is no convincing evidence to support this claim. In this section, we provide two viewpoints to advocate this property for representation based methods.

We first analyze the grouping effect from the viewpoint of optimization. The first term in Problem (1) penalizes the reconstruction error, which can be regarded as a first-order energy encoding the whole subspace structure of the data. The grouping effect of a representation matrix indicates that $\Omega(Z)$ and $\mathcal{C}$ in Problem (1) must include a second-order energy to penalize the discontinuities in the representation coefficients, either implicitly or explicitly. With this second-order energy, the representation will be stable. On the other hand, spatially close data points may help each other to prevent over-fitting in reconstruction the samples. For example, in Figure 1(a), we consider the faces marked by the green and purple squares. They are very close in appearance but with a large part shadowed, hence violating the subspace



(a) (b)

(c) (d)

Figure 1. Grouping effect of representation based methods. (a)-(d) Face images in the Extended Yale Face B dataset, displayed after reducing the dimensionalities of their features to two by PCA, where the features are: (a) the original image, (b)-(d) representation matrices computed by LSR [18], SSC [3], and our SMR, respectively. We can see that LSR and SMR map the spatially close samples in (a) (marked by green and purple boxes) to spatially close ones (green and purple boxes in (b) and (d)), while SSC does not (green and purple boxes in (c)). Also note that SMR maps the two samples closer than LSR and on the whole the face images are better separated by SMR.

constraints [6]. LSR and our proposed SMooth Representation (SMR, see Section 3.3) have the grouping effect. They represent the two faces closely in the new space and are also clustered correctly (Figures 1(b) and (d)). However, SSC does not include the discontinuity penalties, making the representations of the two faces far away from each other and finally being wrongly clustered (Figure 1(c)).

From the viewpoint of affinity measure, the most commonly used affinity measure is $(|Z^*| + |Z^{*T}|)/2$. In general, the grouping effect implies that spatially close points have similar affinities with other points. We describe this formally in Proposition 5.

Proposition 5: *If the EGE conditions (1), (2), and (4) are satisfied, for all* $\|\mathbf{x}_i - \mathbf{x}_j\| \to 0$*, we have: (1)* $|Z^*_{ii} - Z^*_{jj}| \to 0, |Z^*_{ij} - Z^*_{ji}| \to 0$*; (2)* $\forall k \neq i, j$, $|Z^*_{ik} - Z^*_{jk}| \to 0$ *and* $|Z^*_{ki} - Z^*_{kj}| \to 0$.

Proof: According to Proposition 4, we have $\|Z^*_2 - Z^*_1\|_F \to 0$, where $Z^*_2 = P^T_{ij} Z^*_1 P_{ij}$. Hence (1) and (2) result. □

Proposition 5 indicates that grouping effect usually leads to a well balanced affinity graph, which is regarded to be helpful for spectral clustering [28]. In addition, based on Proposition 5, grouping effect enables us to define a new affinity measure for subspace clustering. We will show that this affinity measure performs better than the commonly used one $(|Z^*| + |Z^{*T}|)/2$ when the self-representation model have grouping effect.

3. Smooth Representation Clustering

In this section, based on the detailed analysis on grouping effect, we propose a novel subspace clustering method, called Smooth Representation (SMR) clustering. We first introduce how to explicitly enforce grouping effect in the representation model, then present the SMR model.

3.1. Enforcing Grouping Effect

As stated in Section 2.1, LRR and LSR utilize the grouping effect implicitly. The grouping effect can be understood as the smooth dependence of feature on the sample. We may write the regularization term of LSR as follows:

$$\begin{aligned}\Omega(Z) &= \mathrm{tr}(ZZ^T) \\ &= \tfrac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\|\mathbf{z}_i - \mathbf{z}_j\|_2^2 + \tfrac{1}{n}\|Z^T\mathbf{e}\|_2^2,\end{aligned} \tag{3}$$

where $\mathbf{e}$ is the all ones vector. It can be viewed as assigning equal weights to all pairs of representations, regardless of whether the representations are close to each other or not. By the analysis in Section 2, we should enforce the grouping effect explicitly by the affinity of samples. One possibility is adopting the following regularization term:

$$\begin{aligned}\Omega(Z) &= \tfrac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}\,\|\mathbf{z}_i - \mathbf{z}_j\|_2^2 \\ &= \mathrm{tr}(ZLZ^T),\end{aligned} \tag{4}$$

where $W = (w_{ij})$ is the weight matrix measuring the spatial closeness of points and $L = D - W$ is the Laplacian matrix, in which D is the diagonal degree matrix with $D_{ii} = \sum_{j=1}^{n} w_{ij}$. A common way to construct W is to use the k nearest neighbor (k-nn) graph with a heat kernel or 0-1 weights [9]. We find that the simple 0-1 weighted k-nn graph performs well enough in our experiments, as exemplified in Figure 1(d). So we use the 0-1 weighted k-nn graph (a default value of k is 4) in all our experiments. There are also more complex graphs, e.g., affinity graphs produced by other subspace clustering algorithms. However, usually we did not observe evident improvements over the k-nn graph.

3.2. Smooth Representation

To avoid numerical instability issue, we enforce L to be strictly positive definite by adding a ϵI and use $\tilde{L} = L + \epsilon I$ instead, where I is the identity matrix and $0 < \epsilon \ll 1$. A default value of ϵ is 0.01. Then we get our smooth representation model:

$$\min_{Z} \quad f(Z) = \alpha\|X - XZ\|_F^2 + \mathrm{tr}(Z\tilde{L}Z^T). \tag{5}$$

Problem (5) is a smooth convex program. Differentiating the objective function $f(Z)$ with respect to Z and setting it to zero, we get the optimal solution Z^* satisfies

$$\alpha X^T X Z^* + Z^* \tilde{L} = \alpha X^T X. \tag{6}$$

The above equation is a standard Sylvester equation [1]. It has a unique solution.

Proposition 6: *The Sylvester equation (6) has a unique solution.*

Proof: $X^T X$ is positive semi-definite. So all of its eigenvalues are nonnegative: $\lambda_i \geq 0, \forall i$. $\tilde{L}$ is positive definite. So all of its eigenvalues are positive $\mu_j > 0, \forall j$. Hence, for any eigenvalues of $X^T X$ and $\tilde{L}$, $\lambda_i + \mu_j > 0$. According to [14], the Sylvester equation (6) has a unique solution. □

A classical algorithm for the Sylvester equation is the Bartels-Stewart algorithm [1], which consists of transforming the coefficient matrices into Schur forms by QR decomposition, and then solving the resulting triangular system via back-substitution. The algorithm has a computational complexity of $\mathcal{O}(n^3)$.

The solution to Problem (5) also has several nice properties according to Proposition 7.

Proposition 7: *The solution to Problem (5) has the following properties: (1) it has grouping effect; (2) it is block diagonal when the subspaces are independent and the data is noise free.*

Proof: (1) We can easily check that Problem (5) satisfies EGE conditions (1), (2), and (3). According to Proposition 1, its solution has the grouping effect.

(2) When the columns of the data matrix X is permutated by any permutation matrix P, we have $\tilde{L}(XP) = P^T \tilde{L}(X) P$. Hence, $\Omega(ZP) = \text{tr}(ZP\tilde{L}(XP)P^T Z^T) = \text{tr}(Z\tilde{L}(X)Z^T) = \Omega(Z)$. Denote $Z^D = \begin{bmatrix} A & 0 \\ 0 & D \end{bmatrix}$, where A and D are from $Z = \begin{bmatrix} A & B \\ C & D \end{bmatrix}$. Substituting Z and Z^D into Equation (4), we have $\Omega(Z) \geq \Omega(Z^D)$, where the equality holds if and only if $B = C = 0$, and $\Omega(Z^D) = \Omega(A) + \Omega(D)$. So $\Omega(Z)$ satisfies the Enforced Block Diagonal Condition [18]. According to Theorem 2 in [18], its optimal solution is block diagonal when the subspaces are independent and the data is noise free. □

Furthermore, benefiting from the strengthening of grouping effect, the representations derived by SMR usually have much larger gap between the within-class and the between-class distances than those by LRR and LSR, as illustrated in qualitatively Figure 1(d) and quantitatively Figure 2. This property implies the SMR can derive more salient within-class affinities with regard to the between-class ones than the methods when using the measures described in Section 3.3.

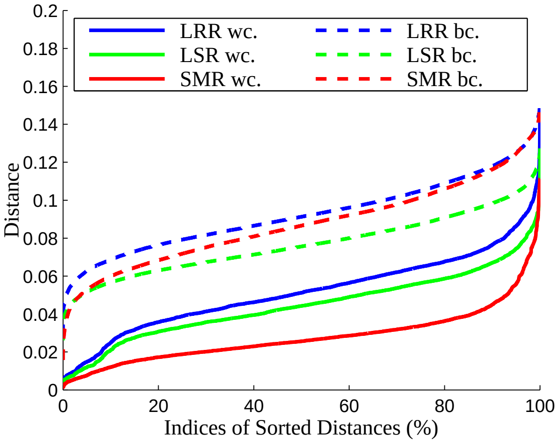


Figure 2. Comparing gaps between the within-class (wc.) and between-classes (bc.) representation distances of LRR, LSR and SMR on the USPS handwritten digit dataset. For each digit image, the 5% closest images in the wc. and bc. sets respectively are selected for illustration. We can see that SMR has much larger gaps than LRR and LSR.

3.3. Subspace Clustering by SMR

After obtaining the self-representation matrix Z^*, a common way for subspace clustering is to define an affinity matrix as

$$J_1 = (|Z^*| + |Z^{*T}|)/2 \tag{7}$$

and use the spectral clustering algorithm [23] to produce the final clustering results, as has been used by SSC, LRR and LSR.

The effectiveness of J_1 mainly comes from the block diagonal property of Z^*, leaving the nice property of grouping effect unexploited. To exploit the merit of grouping effect, we define a new affinity matrix as

$$J_2 = \left(\left| \frac{\mathbf{z}_i^{*T} \mathbf{z}_j^*}{\|\mathbf{x}_i\|_2 \|\mathbf{x}_j\|_2} \right|^{\gamma} \right), \tag{8}$$

where $\gamma > 0$ is used to control the affinity variances. The new affinity measure can be regarded as the inner product of the new representation vectors normalized by the norms of their original features. The normalization prevents the affinity measure from biased by the original feature amplitudes, which is very common in the motion segmentation problem whose trajectories usually vary a lot in amplitude. Figure 3 shows the affinity matrices of the two measures based on Z^* derived by SMR. It can be seen that J_2 strengthens the affinities within each cluster and weakens them across clusters.

The whole procedure of subapace clustering by SMR is summarized in Algorithm 1.

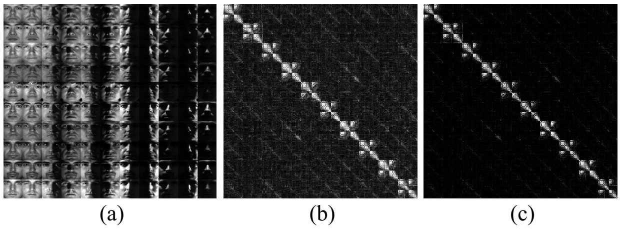


Figure 3. Comparing of J_1 and J_2. (a) Sample images from Extended Yale Face B datasets; (b) affinity matrix J_1; (c) affinity matrix J_2 ($\gamma = 2$). The block diagonal structure of J_2 is more salient than that of J_1. In particular, the magnitudes of off-block-diagonal entries are much smaller.

Algorithm 1 Subspace Clustering by SMooth Representation (SMR)

```
Require: Data points $X = [\mathbf{x}_1, \cdots, \mathbf{x}_n] \in \mathbb{R}^{d \times n}$, the number of subspaces $m$
  Build a $k$-nn graph $W$ and compute the corresponding Laplacian matrix $\tilde{L}$.
  Solve the Sylvester equation (6) by the Bartels-Stewart algorithm to get a representation matrix $Z^*$.
  Compute affinity matrix by either (7) or (8).
  Use spectral clustering algorithm to obtain $m$ clusters.
```

4. Experiments

In this section, we apply our SMR[1] to three applications of subspace clustering: motion segmentation, face clustering, and handwritten digit clustering. We also compare SMR with representative reconstruction based methods, such as SSC, LRR and LSR, whose performances are state-of-the-art.

4.1. Datasets and Evaluation Metric

We use three datasets for our experiments: Hopkins155 [25], Extended Yale Face B [7] and USPS [11], which are the most popular benchmark datasets used in the literature for evaluating subspace clustering algorithms. For all the algorithms, the best results are reported.

Hopkins155 [25] is a motion segmentation dataset, consisting of 155 video sequences with extracted feature points and their tracks across frames. See Figure 4 for some sample sequences. We use PCA to project the data into a 12-dimensional subspace. The same as in most literatures, for each algorithm, we use the same parameters for all sequences [26].

Extended Yale Face B [7] is a face clustering dataset, which consists of 192×168 pixel cropped face images, under varying poses and illuminations, from 38 human subjects. We use all the 64 frontal face images of the first 10 subjects, and resize the images to 32×32. We also use PCA to project the data into a 10×6-dimensional subspace.

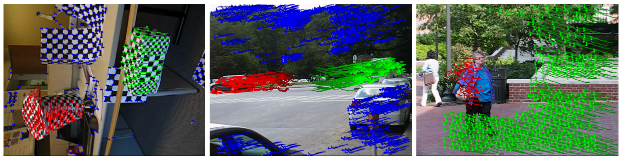


Figure 4. Some sample images from Hopkins155 datasets. The tracks marked in different color indicate different motions. From left to right, they are: *1R2RC*, *cars3* and *people2*.

Table 2. Clustering errors (*CE*) using affinity measure (7) and the computation times on Hopkins155 datasets. The computation time includes only the computation of Z^* in Problem (1).

method	SSC	LRR	LSR	SMR
CE (%)	3.90	4.11	3.01	**2.27**
time (s)	2.50	2.03	**0.12**	0.40

USPS [11] is a handwritten digit dataset of 9298 images, with each image having 16×16 pixels. We use the first 100 images of each digit for experiments.

The same as in most literatures, we use clustering error (CE) to measure the accuracy [25]. CE is the minimum error by matching the clustering result and the ground truth under the optimal permutation, formally defined as:

$$CE = 1 - \frac{1}{N}\sum_{i=1}^{N} \delta(p_i, \mathrm{map}(q_i)), \tag{9}$$

where q_i, p_i represent the output label and the ground truth one of the ith point; $\delta(x, y) = 1$ if $x = y$, and $\delta(x, y) = 0$ otherwise; $\mathrm{map}(q_i)$ is the best mapping function that permutes clustering labels to match the ground truth labels and it can be efficiently computed by the Kuhn-Munkres algorithm [13].

4.2. Experimental Results

For fair comparison, we adopt the Frobenius norm for the reconstruction error term for all the algorithms. Table 2 lists the motion segmentation errors of the four methods on the Hopkins155 datasets using the typical affinity measure (7). SMR achieves a clustering error of 2.27%, while the best result of other algorithms is 3.01% by LSR. *It should be noted that the numbers in Table 2 are different from those listed in [18] because they used an approximate computation of CE, which is also observed by [4].* Noting that most sequences are easy to be segmented and hence all the algorithms get zero errors on them, the performance improvement by SMR over others is significant. The computational costs of all the algorithms are also listed in Table 2. SMR is a bit slower than LSR but much faster than SSC and LRR.

[1] Codes available at *https://sites.google.com/site/hanhushomepage/*

Table 3. Clustering errors (*CE*) using affinity measure (8) with $\gamma = 1$ on Hopkins155 datasets.

method	SSC	LRR	LSR	SMR
CE (%)	6.27	2.83	1.90	**1.13**

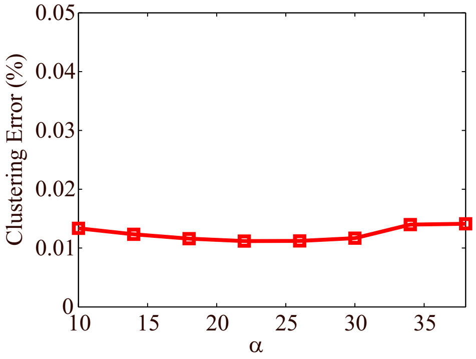


Figure 5. The performance of SMR with varying α on Hopkins155 datasets (4-nn graph is used).

As stated in Section 3.3, the affinity measure (8) is better in exploiting the grouping effect. So we also use affinity measure (8) for experiments and report the clustering errors in Table 3. It can be seen that the performances of LRR, LSR and SMR are significantly improved, and SMR get the minimum segmentation error with 1.13%. The SSC with the new affinity measure (8) performs worse than using the traditional affinity (7). These results support the use of affinity measure (8) rather than (7) for subspace clustering when the self-representation model has grouping effect.

We also test the performance of SMR with varying parameters α (in the objective function of (1)) and k (for constructing the k-nn graph). The results are shown in Figure 5 and Figure 6. SMR performs very stably with varying α and the number k of neighborhood. Since other algorithms rely on only one parameter α, to be fair we use a 4-nn graph for all our experiments without tuning k on different datasets.

Tables 4 and 5 show the clustering errors on Extended Yale Face B and USPS datasets, respectively. To make fair comparison, we use the traditional affinity measure (7) in all the algorithms. The number of our algorithm using the new affinity (8) is also shown in the last column for reference. It can be seen that SMR outperforms the others significantly, especially on the USPS datasets. We also illustrate the affinity matrices using (7) in Figure 7, where we can clearly observe better grouping effect from SMR than from the others. When the affinity (8) J_2 is used, the performances are further improved. For example, we achieve 3.75% on Extended Yale Face B.

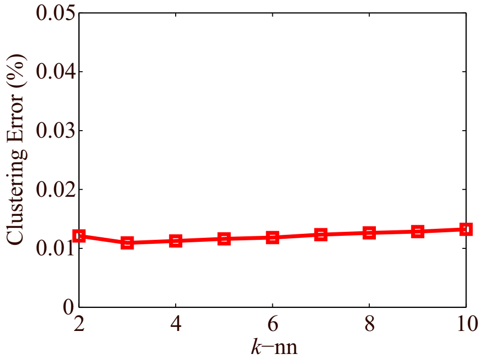


Figure 6. The performance of SMR with varying k in k-nn graph construction, where for each k, the optimal α is reported.

Table 4. Clustering errors (*CE*) on Extended Yale Face B datasets.

method	SSC	LRR	LSR	SMR	SMR(J_2)
CE (%)	48.81	35.00	27.50	**26.56**	**3.75**

Table 5. Clustering errors (*CE*) on USPS datasets.

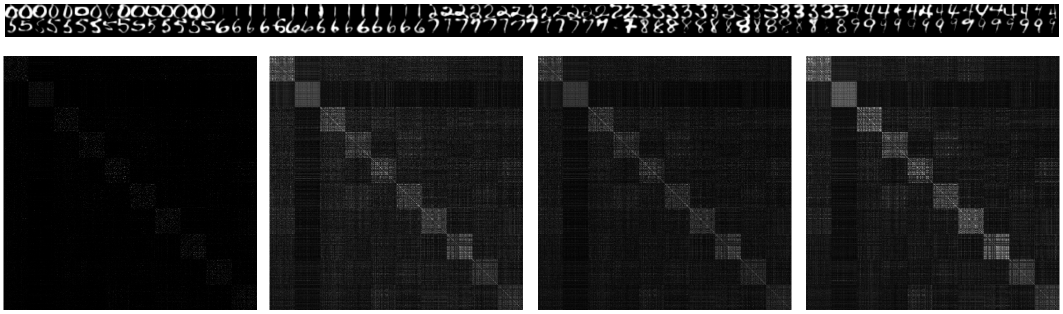
method	SSC	LRR	LSR	SMR	SMR(J_2)
CE (%)	43.10	22.60	26.10	**12.70**	**11.20**

5. Conclusions and Future Work

In this paper, we analyze the grouping effect of representation based methods in depth. We introduce Enforced Grouping Effect conditions to verify whether a representation based model has the grouping effect in a systematic manner. We also provide insights to the importance of grouping effect for subspace clustering. Based on our detailed analysis, we propose a novel subspace clustering model, Smooth Representation, to explicitly enforce the grouping effect in the model. We further propose a novel affinity measure that better utilizes the grouping effect among representation coefficients. Extensive experiments on benchmark datasets testify to the great advantage of SMR over the state-of-the-art subspace clustering methods. In the future, we plan to utilize the grouping effect in a wider scope, e.g., semi-supervised learning.

Acknowledgement

This work is supported by the National Natural Science Foundation of China under Grants 61225008, 61020106004 and 61373074, the National Basic Research Program of China under Grant 2014CB349304, the Ministry of Education of China under Grant 20120002110033, and the Tsinghua University Initiative Scientific Research Program. Zhouchen Lin is supported by NSF China (grant nos. 61272341, 61231002, 61121002) and MSRA.



(a) SSC (b) LRR (c) LSR (d) SMR

Figure 7. Affinity matrices derived by SSC, LRR, LSR, and SMR on USPS datasets using (7). The affinities are normalized by $0.6 \cdot \max(Z^*)$ to have better view. The grouping effect of SMR is much more salient than those of others.

References

[1] R. Bartels and G. Stewart. Solution of the matrix equation $AX + XB = C$. *Communications of the ACM*, 15(9):820–826, Sept. 1972.

[2] J. Costeira and T. Kanade. A multibody factorization method for independently moving objects. *IJCV*, 29(3):108–121, 1998.

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

[4] E. Elhamifar and R. Vidal. Sparse subspace clustering: Algorithm, theory, and applications. *IEEE TPAMI*, 2013.

[5] M. A. Fischler and R. C. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM*, 24(6):381–395, 1981.

[6] A. Georghiades, P. Belhumeur, and D. Kriegman. From few to many: Illumination cone models for face recognition under variable lighting and pose. *IEEE TPAMI*, 23(6):643–660, 2001.

[7] A. S. Georghiades, P. N. Belhumeur, and D. J. Kriegman. From few to many: Illumination cone models for face recognition under variable lighting and pose. *IEEE TPAMI*, 23(6):643–660, 2001.

[8] E. Grave, G. Obozinski, and F. Bach. Trace lasso: a trace norm regularization for correlated designs. In *NIPS*, pages 2187–2195, 2011.

[9] X. He and P. Niyogi. Locality preserving projections. In *NIPS*, 2003.

[10] H. Hu, J. Feng, and J. Zhou. Exploiting unsupervised and supervised constraints for subspace clustering. *CoRR*, 2014.

[11] J. J. Hull. A database for handwritten text recognition research. *IEEE TPAMI*, 16(5):550–554, 1994.

[12] K. Kanatani and Y. Sugaya. Multi-state optimization for multi-body motion segmentation. In *Australia-Japan Advanced Workshop on Computer Vision*, 2003.

[13] H. Kuhn. The Hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955.

[14] P. Lancaster. Explicit solutions of linear matrix equations. *SIAM Review*, 12(4):pp. 544–566, 1970.

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

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

[17] C. Lu, Z. Lin, and S. Yan. Correlation adaptive subspace segmentation by trace lasso. In *ICCV*, 2013.

[18] C. Y. Lu, H. Min, Z.-Q. Zhao, L. Zhu, D.-S. Huang, and S. Yan. Robust and effficient subspace segmentation via least squares regression. In *ECCV*, pages 347–360, 2012.

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

[20] Y. Ma, H. Derksen, W. Hong, and J. Wright. Segmentation of multivariate mixed data via lossy data coding and compression. *IEEE TPAMI*, 29(9):1546–1562, 2007.

[21] J. H. Park, H. Zha, and R. Kasturi. Spectral clustering for robust motion segmentation. In *ECCV (4)*, pages 390–401, 2004.

[22] S. Rao, R. Tron, R. Vidal, and Y. Ma. Motion segmentation in the presence of outlying, incomplete, or corrupted trajectories. *IEEE TPAMI*, 32(10):1832–1845, 2010.

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

[24] C. Tomasi and T. Kanade. Shape and motion from image streams under orthography: a factorization method. *IJCV*, 9(2):137–154, 1992.

[25] R. Tron and R. Vidal. A benchmark for the comparison of 3-d motion segmentation algorithm. In *Proc. CVPR*, 2007.

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

[27] R. Vidal, Y. Ma, and S. Sastry. Generalized principal component analysis (GPCA). *IEEE TPAMI*, 27(12):1945–1959, 2005.

[28] U. von Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416, 2007.

[29] S. Wang, X. Yuan, T. Yao, S. Yan, and J. Shen. Efficient subspace segmentation via quadratic programming. In *AAAI*, pages 519–524, 2011.

[30] J. Yan and M. Pollefeys. A general framework for motion segmentation: Independent, articulated, rigid, non-rigid, degenerate and non-degenerate. In *Proc. ECCV*, 2006.

Supplementary Material of
Smooth Representation Clustering

Anonymous CVPR submission

Paper ID 1242

In this document, we prove Proposition 2 and Proposition 3 in detail.

To prove Proposition 2, we first provide two lemmas:

Lemma S.1 [1]: *Given a subspace S spanned by a set of orthogonal basis $[\mathbf{u}_1, \dots, \mathbf{u}_r]$ ($\mathbf{u}_i \in \mathbb{R}^{n \times 1}$) and its orthogonal complement $S_\perp$, for any matrix $M \in \mathbb{R}^{n \times k}, \forall k$, there exist a unique pair $M_1 \in S$ and $M_2 \in S_\perp$ such that*

$$M = M_1 + M_2. \tag{1}$$

Lemma S.2: *Let A and B be matrices of the same size. If $AB^T = 0$ and $A^T B = 0$, then $\|A + B\|_* = \|A\|_* + \|B\|_*$.*

Proof: Note the singular value decompositions (SVDs) of A and B as:

$$A = U_A \Sigma_A V_A^T, \ B = U_B \Sigma_B V_B^T, \tag{2}$$

where U_A and U_B are left-invertible; and V_A and V_B are right-invertible. From the condition $AB^T = 0$, we get $V_A^T V_B = 0$. Similarly, $A^T B = 0$ implies $U_A^T U_B = 0$. Hence,

$$A + B = \left[\begin{array}{cc} U_A & U_B \end{array} \right] \left[\begin{array}{cc} \Sigma_A & \\ & \Sigma_B \end{array} \right] \left[\begin{array}{cc} V_A & V_B \end{array} \right]^T \tag{3}$$

is a valid SVD of $A + B$. It is easy to check that $\|A + B\|_* = \|A\|_* + \|B\|_*$. □

Proposition 2: *The LRR problem (4) [2] has a unique optimal solution.*

$$\min_Z \ f(Z) = \alpha \|X - XZ\|_F^2 + \|Z\|_*. \tag{4}$$

Proof: Note the SVD of X as $X = U \Sigma V^T$ with $U \in \mathbb{R}^{d \times r}$, $\Sigma = \text{diag}(\mathbf{s})$ ($\mathbf{s}_i > 0, \forall 1 \leq i \leq r$) and $V \in \mathbb{R}^{n \times r}$. Note S as the subspace spanned by columns of V, and $S_\perp$ as the orthogonal complement of S.

Suppose Z^* is an optimal solution of problem (4). According to Lemma S.1, there exist a unique pair $Z_1^* \in S$ and $Z_2^* \in S_\perp$ that $Z^* = Z_1^* + Z_2^*$. Next we prove that Z_2^* must equal 0.

Suppose $Z_2^* \neq 0$. We have $\|Z_2^*\|_* > 0$. The condition $\mathbf{Z}_2^* \in S_\perp$ implies $XZ_2^* = U \Sigma V^T Z_2^* = 0$. Then

$$\begin{aligned} f(Z^*) &= \alpha ||X - XZ^*||_F^2 + ||Z||_* \\ &= \alpha ||X - X(Z_1^* + Z_2^*)||_F^2 + ||Z_1^* + Z_2^*||_* \\ &= \alpha ||X - XZ_1^*||_F^2 + ||Z_1^*||_* + ||Z_2^*||_* \\ &> f(Z_1^*) \end{aligned} \tag{5}$$

Equation (5) indicates Z_1^* is a better solution of problem (4) than Z^*, which is a contradiction. Hence $Z_2^* = 0$ is proved. As a result, we have $Z^* = Z_1^*$.

The condition $Z_1^* \in S$ indicates that there exists a unique matrix $W \in \mathbb{R}^{r \times n}$ that

$$Z_1^* = VW. \tag{6}$$

Substituting equation (6) into problem (4), we get a new optimization about W as

$$\min_{W}\; g(W) = \alpha\|X - XVW\|_F^2 + \|VW\|_* = \alpha\|X - U\Sigma W\|_F^2 + \|W\|_*. \tag{7}$$

It is easy to verify that the Hessian matrix of the first term

$$H_1 = I \otimes \Sigma U^T U \Sigma = I \otimes \Sigma^2 \succ 0, \tag{8}$$

where $I \in \mathbb{R}^{n\times n}$ is the identity matrix, and $\otimes$ is the Kronecker product operator. According to equation (8), problem (7) is strictly convex and it has a unique solution W^*. This implies that the solution of problem (4), Z^*, is also unique, and $Z^* = VW^*$. □

Next we prove Proposition 3. Recall the optimization problem for self-representation based methods as (9).

$$\begin{aligned} \min_{Z} \quad & f(Z) = \alpha\|X - A(X)Z\|_l + \Omega(X, Z), \\ s.t. \quad & Z \in \mathcal{C}, \end{aligned} \tag{9}$$

Proposition 3: *Problems (9) with the following $\Omega(Z)$ and $\mathcal{C}$ have grouping effect:*

(1) $\Omega(Z) = \sum\limits_{j=1}^{n}\left(\sum\limits_{i=1}^{n}|Z_{ij}|^p\right)^q, p > 1, q \geq 0, \mathcal{C} = \emptyset.$

(2) $\Omega(Z) = tr((ZHZ^T)^p), H \succ 0, p \geq 1/2, \mathcal{C} = \emptyset.$

(3) $\Omega(Z) = tr((Z^T HZ)^p), H \succ 0, p \geq 1/2, \mathcal{C} = \emptyset.$

Proof: (1) It is easy to verify that EGE conditions (1) and (3) are satisfied.

Noting that the regularity term $\Omega(Z) = \sum\limits_{j=1}^{n}\left(\sum\limits_{i=1}^{n}|Z_{ij}|^p\right)^q = \sum\limits_{j=1}^{n}\|Z_j\|_p^{pq}$, where $\|Z_j\|_p = \left(\sum\limits_{i=1}^{n}|Z_{ij}|^p\right)^{1/p}$ is the ℓ_p vector-norm, we have $\Omega(Z)$ is strictly convex w.r.t Z. As a result, problem (9) has a unique solution. According to Proposition 1 in the paper, the grouping effect of this solution is also guaranteed.

(2) Regarding H defined by $X = [\mathbf{x}_1, \ldots, \mathbf{x}_n]$ with $H(XP) = P^T H(X)P$, we can verify that $\Omega(Z) = \mathrm{tr}((ZHZ^T)^p)$ satisfy EGE conditions (1) and (3).

When $p > 1/2$, $\Omega(Z) = \mathrm{tr}((ZHZ^T)^p)$ is strictly convex w.r.t Z, and thus problem (9) has a unique solution. In the following, we will prove that when $p = 1/2$, problem (9) also has a unique solution

Since $H \succ 0$, we can find an invertible matrix $L \in \mathbb{R}^{n\times n}$ such that $H = LL^T$. Substituting $Z = YL^{-1}$ into $f(Z)$, we have

$$f(Z) = h(Y) = \alpha\|X - XYL^{-1}\|_F^2 + \mathrm{tr}((YY^T)^{1/2}). \tag{10}$$

Noting that $\|Y\|_* = \mathrm{tr}((YY^T)^{1/2})$, similar as the proof of Proposition 2, we conclude that $Y^* = VW \in S$ and thus an optimization problem w.r.t W is obtained as

$$\min_{W}\; g(W) = \alpha\|X - XVWL^{-1}\|_F^2 + \|VW\|_* = \alpha\|X - U\Sigma WL^{-1}\|_F^2 + \|W\|_*. \tag{11}$$

The Hessian matrix of the first term of $g(W)$ is

$$H_1 = (LL^T)^{-T} \otimes \Sigma^2 = H^{-T} \otimes \Sigma^2. \tag{12}$$

Since $H \succ 0$ and $\Sigma^2 \succ 0$, we get $H_1 \succ 0$, which indicates the uniqueness of the solution of problem (11). Hence, Problem (9) with $\Omega(Z) = \mathrm{tr}((ZHZ^T)^{1/2}), H \succ 0, \mathcal{C} = \emptyset$ also has a unique solution.

According to Proposition 1, the grouping effect is proved.

(3) When $p > 1/2$, the uniqueness and grouping effect of the solution can be easily proved.

In the following, we prove the proposition with $p = 1/2$. There exists a decomposition $H = LL^T, L \in \mathbb{R}^{n\times n}$. Substituting $Z = L^{-T}Y$ into $f(Z)$, we get

$$f(Z) = h(Y) = \alpha\|X - XL^{-T}Y\|_F^2 + \|Y\|_*. \tag{13}$$

Note $U\Sigma V^T$ as the SVD of XL^{-T} and S as the subspace spanned by the columns of XL^{-T}. Similarly as the proof of Proposition 2, we have $Y^* \in S$ and it is unique, which also implies the uniqueness of Z^*. As a result, Z^* has grouping effect. □

References

[1] R. A. Horn and C. R. Johnson. *Matrix Analysis*. Cambridge University Press, 2007. 1

[2] G. Liu, Z. Lin, and Y. Yu. Robust subspace segmentation by low-rank representation. In *ICML*, pages 663–670, 2010. 1