Robust and Efficient Subspace Segmentation via Least Squares Regression

Can-Yi Lu[1,2], Hai Min[1], Zhong-Qiu Zhao[3], Lin Zhu[1],
De-Shuang Huang[4], and Shuicheng Yan[2]

[1] Department of Automation, University of Science and Technology of China, Hefei, China
[2] Department of Electrical and Computer Engineering, National University of Singapore
[3] School of Computer and Information, Hefei University of Technology, Hefei, China
[4] School of Electronics and Information Engineering, Tongji University, Shanghai, China
{canyilu,minhai361,zhongqiuzhao}@gmail.com,
zhomlynn@yahoo.com.cn, dshuang@tongji.edu.cn, eleyans@nus.edu.sg

Abstract. This paper studies the subspace segmentation problem which aims to segment data drawn from a union of multiple linear subspaces. Recent works by using sparse representation, low rank representation and their extensions attract much attention. If the subspaces from which the data drawn are independent or orthogonal, they are able to obtain a block diagonal affinity matrix, which usually leads to a correct segmentation. The main differences among them are their objective functions. We theoretically show that if the objective function satisfies some conditions, and the data are sufficiently drawn from independent subspaces, the obtained affinity matrix is always block diagonal. Furthermore, the data sampling can be insufficient if the subspaces are orthogonal. Some existing methods are all special cases. Then we present the Least Squares Regression (LSR) method for subspace segmentation. It takes advantage of data correlation, which is common in real data. LSR encourages a grouping effect which tends to group highly correlated data together. Experimental results on the Hopkins 155 database and Extended Yale Database B show that our method significantly outperforms state-of-the-art methods. Beyond segmentation accuracy, all experiments demonstrate that LSR is much more efficient.

1 Introduction

Subspace segmentation is an important clustering problem which attracts much attention in recent years. It arises in numerous applications in machine learning and computer vision literature, e.g. image representation [1], clustering [2] and motion segmentation [3] [4]. Given a set of data drawn from a union of subspaces, the goal of subspace segmentation is to segment (cluster or group) data into clusters with each cluster corresponding to a subspace. This problem is formally defined as follow:

Definition 1. *(Subspace Segmentation) Given a set of data vectors* $X = [X_1, \cdots, X_k] = [x_1, \cdots, x_n] \in \mathbb{R}^{d\times n}$ *drawn from a union of* k *subspaces* $\{\mathcal{S}_i\}_{i=1}^k$. *Let* X_i *ba a collection of* n_i *data vectors drawn from the subspace* $\mathcal{S}_i$, $n = \sum_{i=1}^k n_i$. *The task is to segment the data according to the underlying subspaces they are drawn from.*

A. Fitzgibbon et al. (Eds.): ECCV 2012, Part VII, LNCS 7578, pp. 347–360, 2012.
© Springer-Verlag Berlin Heidelberg 2012

1.1 Prior Works

During the past two decades, a number of subspace segmentation methods have been proposed. According to their mechanisms of representing the subspaces, existing works can be roughly divided into four categories: algebraic methods [5] [6], iterative methods [7] [8], statistical methods [9] [10] and spectral clustering-based methods [11] [12] [13] [14]. A review of these methods can be found in [15]. In this work, we review four most recent and related methods: Sparse Subspace Clustering (SSC) [11] [16], Low-Rank Representation (LRR) [12] [17] and their extensions [13] [14].

SSC and LRR are two spectral clustering-based methods. A main challenge in applying spectral clustering to subspace segmentation is to define a "good" affinity matrix (or graph) $Z \in \mathbb{R}^{n\times n}$. Each entry Z_{ij} measures the similarity between data points x_i and x_j. Ideally, the affinity matrix should be block diagonal, the between-cluster affinities are all zeros. Then it is easy to segment data on such well defined graph by spectral clustering. Typical choice for the measure of similarity is $Z_{ij} = \exp(-||x_i - x_j||/\sigma)$, $\sigma > 0$. However, this method is not able to characterize the structure of data from subspaces, and the affinity matrix is also not block diagonal. SSC and LRR provide a new way to construct the affinity matrix. They express each data point x_i as a linear combination of all other data $x_i = \sum_{j\neq i} Z_{ij}x_j$, and use the representational coefficient $(|Z_{ij}| + |Z_{ji}|)/2$ to measure the similarity between x_i and x_j. The difference between SSC and LRR lies in the regularization on Z: SSC enforces Z to be sparse while LRR encourages Z to be of low-rank.

Motivated by the fact that an ideal affinity matrix is block diagonal or sparse, SSC solves the following sparse representation problem:

$$\min ||Z||_0 \ \text{ s.t. } X = XZ, \ \text{diag}(Z) = 0, \tag{1}$$

where $||Z||_0$ is the ℓ^0-norm of Z, *i.e.* the number of nonzero elements. But such optimization problem is non-convex and NP-hard. Under some condition [18], it is equal to the following ℓ^1-minimization problem:

$$\min ||Z||_1 \ \text{ s.t. } X = XZ, \ \text{diag}(Z) = 0, \tag{2}$$

where $||Z||_1$ denotes the ℓ^1-norm of Z, $||Z||_1 = \sum_{i=1}^{n}\sum_{j=1}^{n} |Z_{ij}|$. It has been shown that when the subspaces are independent [1], the solution to problem (2) is block diagonal. But this solution does not guarantee to obtain a correct segmentation, since it may be "too sparse", which divides the within-cluster data into different groups. If there is a group of data points among which the pairwise correlations are very high, then sparse representation tends to select only one, at random [19]. Thus, SSC is not able to capture the correlation structure of data from the same subspace. Another drawback of SSC is that it is not efficient for solving the ℓ^1-minimization problem (2) for each data point.

LRR aims to take the correlation structure of data into account, it finds a low rank representation instead of a sparse representation. In the case of noise free data drawn from linear subspaces, the original LRR solves the following rank minimization problem:

$$\min \text{rank}(Z) \ \text{ s.t. } X = XZ, \tag{3}$$

[1] A collection of k subspaces $\{\mathcal{S}_i\}_{i=1}^{k}$ are independent if and only if $\sum_{i=1}^{k} \mathcal{S}_i = \oplus_{i=1}^{k}\mathcal{S}_i$.

where rank(Z) denotes the rank of Z. The rank minimization problem is also NP-hard, a common surrogate of rank function is the nuclear norm:

$$\min ||Z||_* \text{ s.t. } X = XZ, \tag{4}$$

where $||Z||_*$ is the nuclear norm of Z, *i.e.* the sum of all the singular values of Z. It can be shown that when the subspaces are independent, the solution to problem (4) is unique and block diagonal. The extended LRR model for data with noise is efficient and effective for subspace segmentation. However, is it really the low rank property of Z that makes LRR powerful for subspace segmentation? The intuition is not clear. It is not necessary to require a block diagonal matrix to be of low rank. The following is a special example based on the problem (3):

Example 1: Let $X_1 = \begin{bmatrix} 1 & 2 \\ 0 & 0 \end{bmatrix}$ and $X_2 = \begin{bmatrix} 0 & 0 \\ 1 & 2 \end{bmatrix}$ are some data points drawn from subspaces $\mathcal{S}_1$ and $\mathcal{S}_2$, respectively. We assume $\dim(\mathcal{S}_1) = \dim(\mathcal{S}_2) = 1$, thus $\mathcal{S}_1$ and $\mathcal{S}_2$ are orthogonal [2], and $\text{rank}(X_1) = \text{rank}(X_2) = 1$. Let $X = [X_1, X_2]$. The solutions to the problem (3) are not unique and one of which is

$$Z^* = \begin{bmatrix} 0.5 & 1 & 1 & 2 \\ 0.25 & 0.5 & -0.5 & -1 \\ 1 & 2 & 0.5 & 1 \\ -0.5 & -1 & 0.25 & 0.5 \end{bmatrix} \tag{5}$$

The above example shows that even when the subspaces are orthogonal, the solution to the original LRR model (3) does not guarantee to be block diagonal, and we are not able to get the true segmentation from (5). But we will theoretically show that it is easy to get a block diagonal solution for independent subspaces segmentation and even the data can be insufficient if the subspaces are orthogonal. Though LRR performs well by solving the nuclear norm minimization problem (4) which also leads to a solution with low rank, its superiority may have nothing to do with the rank of the solution. We may regard the nuclear norm (4) as a new criterion but not a simple surrogate of the rank function, at least for the subspace segmentation problem. The motivation of LRR comes from RPCA [20], but one should notice that their physical meanings are very different. RPCA aims to find a low-rank recovery of the observation X while LRR focuses on the matrix of representational coefficient.

Luo *et al.* proposes the Multi-Subspace Representation (MSR) [13] which combines both criteria of SSC and LRR by solving the following problem:

$$\min ||Z||_* + \delta||Z||_1 \text{ s.t. } X = XZ, \text{diag}(Z) = 0. \tag{6}$$

Another method, namely Subspace Segmentation via Quadratic Programming (SSQP) [14] is proposed by solving the following problem:

$$\min ||XZ - X||_F^2 + \lambda||Z^T Z||_1 \text{ s.t. } Z \geq 0, \text{diag}(Z) = 0. \tag{7}$$

The main differences among the above four methods are their objective functions. All these methods have a similar property: the solutions obtained by SSC, LRR and MSR

[2] The orthogonal subspaces must be independent, but not vice versa.

are block diagonal when the subspaces are independent, and SSQP requires an orthogonal subspaces assumption. Now a question is raised: what kind of objective function and data assumption guarantee a block diagonal solution? Furthermore, which will be a better choice? We try to answer these questions which are summarized in the contributions of this paper.

1.2 Paper Contributions

In this work, we first theoretically show that if the objective function satisfies certain conditions, we are able to get a block diagonal solution based on the independent subspaces assumption. The above four criteria all satisfy these conditions and they are all special cases. If the sampling data are not sufficient, we further require the subspaces to be orthogonal. Second, as a new special case, we present the Least Squares Regression (LSR) model for subspace segmentation. The grouping effect of LSR help it group highly correlated data together. We further show that LSR is robust to a bounded disturbance, and is very efficient due to a closed form solution. At last, we experimentally show that LSR is more effective and efficient than SSC and LRR on the Hopkins 155 database and Extended Yale Database B.

2 Theoretical Analysis

For the ease of exploration, we assume the data are noise free in this section. SSC, LRR, MSR and SSQP use different criteria to control the within-cluster and between-cluster affinities. Under certain data assumption (independent or orthogonal), a block diagonal solution is obtained for true segmentation. A natural question is raised: what kind of criterion owns such property? We first consider a simple situation by using a basis of subspaces as dictionary.

Theorem 1. *Assume the subspaces $\{\mathcal{S}_i\}_{i=1}^k$ are independent, B_i is a matrix whose columns consist of a basis of the subspace $\mathcal{S}_i$, $B = [B_1, \cdots, B_k]$, and X_i is a matrix whose columns consist of some vectors from $\mathcal{S}_i$, $X = [X_1, \cdots, X_k]$. The solution Z^* to the following system*

$$X = BZ \tag{8}$$

is unique and block diagonal.

Proof. We only need to prove that, for any data point $y \in \mathcal{S}_i$, $y \neq 0$, there exists a unique z, $y = Bz$, where $z = [z_1^T, \cdots, z_k^T]^T$, with $z_i \neq 0$ and $z_j = 0$ for all $j \neq i$. Since the subspaces are independent, there exists a unique decomposition of y:

$$\begin{aligned} y &= 0 + \cdots + y + \cdots + 0 \\ &= B_1 z_1 + \cdots + B_i z_i + \cdots + B_k z_k, \end{aligned}$$

where $B_i z_i \in \mathcal{S}_i$, $i = 1, \cdots, k$. Thus, $B_i z_i = y$ and $B_j z_j = 0$ for all $j \neq i$. Considering B_i is a basis of $\mathcal{S}_i$, we have $z_i \neq 0$, z_i is unique, and $z_j = 0$ for all $j \neq i$.

From Theorem 1, we can learn a basis of X and use it as the dictionary. If the subspaces are independent, it is very easy to get a true segmentation by solving problem

(8). But this model cannot be directly extended for handling the data with noise which destroys the subspace structures. We consider a similar but more general model as the existing methods as follow:

$$\min f(Z) \ \text{ s.t. } Z \in \Omega = \{Z|X = XZ\}, \tag{9}$$

where $f(Z)$ is a matrix function. We show that the solution to problem (9) is block diagonal if $f(Z)$ satisfies the following *Enforced Block Diagonal (EBD)* conditions:

Enforced Block Diagonal Conditions. A function f is defined on $\Omega(\neq \varnothing)$ which is a set of matrices. For any $Z = \begin{bmatrix} A & B \\ C & D \end{bmatrix} \in \Omega$, $Z \neq 0$, where A and D are square matrices, B and C are of compatible dimension, A, $D \in \Omega$. Let $Z^D = \begin{bmatrix} A & 0 \\ 0 & D \end{bmatrix} \in \Omega$. We require

(1) $f(Z) = f(ZP)$, for any permutation matrix P, $ZP \in \Omega$.
(2) $f(Z) \geq f(Z^D)$, where the equality holds if and only if $B = C = 0$ (or $Z = Z^D$).
(3) $f(Z^D) = f(A) + f(D)$.

Theorem 2. *Assume the data sampling is sufficient* [3]*, and the subspaces are independent. If f satisfies the EBD conditions (1)(2), the optimal solution(s) Z^* to problem (9) is block diagonal:*

$$Z^* = \begin{bmatrix} Z_1^* & 0 & \cdots & 0 \\ 0 & Z_2^* & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & Z_k^* \end{bmatrix}$$

with $Z_i^* \in \mathbb{R}^{n_i \times n_i}$ corresponding to X_i, for each i. Furthermore, if f satisfies the EBD conditions (1)(2)(3), for each i, Z_i^* is also the optimal solution to the following problem:

$$\min f(Y) \ \textit{ s.t. } X_i = X_i Y \tag{10}$$

Proof. First, we still assume the columns of X are in general position: $X = [X_1, \cdots, X_k]$, since $f(Z) = f(ZP)$, for any permutation P, the objective function is invariant to any permutation. Let Z^* be an optimal solution to problem (9), we decompose Z^* to two parts $Z^* = Z^D + Z^C$, where

$$Z^D = \begin{bmatrix} Z_1^* & 0 & \cdots & 0 \\ 0 & Z_2^* & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & Z_k^* \end{bmatrix}, \ Z^C = \begin{bmatrix} 0 & * & \cdots & * \\ * & 0 & \cdots & * \\ \vdots & \vdots & \ddots & \vdots \\ * & * & \cdots & 0 \end{bmatrix},$$

with $Z_i^* \in \mathbb{R}^{n_i \times n_i}$. Denote $[M]_j$ as the j-th column of matrix M. Assume $[X]_j = [XZ^*]_j \in \mathcal{S}_l$, thus $[XZ^D]_j \in \mathcal{S}_l$, $[XZ^C]_j \in \oplus_{i \neq l} \mathcal{S}_i$. But $[XZ^C]_j = [XZ^*]_j -$

[3] The data sampling is sufficient which makes the problem (9) have a nontrivial solution.

Table 1. Criteria which satisfy the EBD conditions (1)(2)(3)

	$f(Z)$	Ω
SSC	$\|\|Z\|\|_0$ or $\|\|Z\|\|_1$	$\{Z\|X = XZ,\ \text{diag}(Z) = 0\}$
LRR	$\|\|Z\|\|_*$	$\{Z\|X = XZ\}$
SSQP	$\|\|Z^T Z\|\|_1$	$\{Z\|X = XZ, Z \geq 0, \text{diag}(Z) = 0\}$
MSR	$\|\|Z\|\|_1 + \delta\|\|Z\|\|_*$	$\{Z\|X = XZ, \text{diag}(Z) = 0\}$
Other choices	$(\sum_{i=1}^n \sum_{j=1}^n \lambda_{ij}\|Z_{ij}\|^{p_{ij}})^s$ $\lambda_{ij} > 0, p_{ij} > 0, s > 0$	$\{Z\|X = XZ, \text{diag}(Z) = 0\}$

$[XZ^D]_j \in \mathcal{S}_l$, since the subspaces are independent, $\mathcal{S}_l \cap \oplus_{i \neq l} \mathcal{S}_i = \{0\}$, so $[XZ^C]_j = 0$. Thus $XZ^C = 0$, $XZ^D = X$, Z^D is feasible for problem (9). By the EBD conditions (2), we have $f(Z^*) \geq f(Z^D)$. Notice Z^* is optimal, $f(Z^*) \leq f(Z^D)$. Therefore, $f(Z^*) = f(Z^D)$, the equality holds if and only if $Z^* = Z^D$. Hence, Z^* is block diagonal.

If the EBD conditions (3) is further satisfied, we have $f(Z^*) = \sum_{i=1}^k f(Z_i^*)$, $X = XZ^* = [X_1 Z_1^*, \cdots, X_k Z_k^*]$, $X_i = X_i Z_i^*$. Hence, Z_i^* is also the minimizer to problem (13).

From Theorem 2, it is easy to confirm SSC, LRR, MSR and SSQP are all special cases (see Table 1) by the following propositions:

Proposition 1. *If f satisfies the EBD conditions (1)(2)(3) on Ω, then also on $\Omega_1 \subset \Omega$, $\Omega_1 \neq \varnothing$.*

Proposition 2. *$\{f_i\}_{i=1}^m$ are a series of functions. For each i, if f_i satisfies the EBD conditions (1)(2)(3) on Ω_i, then also $\sum_{i=1}^m \lambda_i f_i$, ($\lambda_i > 0$) on $\cap_{i=1}^m \Omega_i (\neq \varnothing)$.*

The independence of subspaces guarantees the separability of data. Theorem 2 shows that many criteria can utilize such separability for correct segmentation. The minimization problem (13) helps us understand the within-cluster affinities which are not studied before. SSC not only enforces sparsity between-cluster but also within-cluster, which may lead to a too sparse solution. LRR encourages the within-cluster affinities to be of low nuclear norm (also low rank). But the physical meaning is unclear by a low rank graph.

For SSC, both $||Z||_0$ and $||Z||_1$ satisfy the EBD conditions (1)(2)(3). $||Z||_1$ is not only a good surrogate of $||Z||_0$ for optimization, but also an independent method for subspace segmentation. For the rank criterion, rank(Z) does not satisfy the EBD conditions (2), which is the key to enforce sparsity between-cluster. This is also the major difference between rank(Z) and $||Z||_*$. Thus, the original low rank representation (3) may be far from enough for subspace segmentation. $||Z||_*$ has its independent ability for modeling the data from different subspaces, it is not a simple surrogate of rank(Z) for subspace segmentation.

Notice SSQP requires the subspaces to be orthogonal for correct segmentation, but we show the independent subspaces assumption is enough. But the data sampling can be insufficient if the subspaces are orthogonal:

Theorem 3. *If the subspaces are orthogonal, and f satisfies the EBD conditions (1)(2), the optimal solution(s) to the following problem:*

$$\min ||X - XZ||_{2,p} + \lambda f(Z) \tag{11}$$

must be block diagonal, where $||\cdot||_{2,p}$ is defined as $||M||_{2,p} = (\sum_j (\sum_{i=1}^n M_{ij}^2)^{\frac{p}{2}})^{\frac{1}{p}}$, $p > 0$, and $\lambda > 0$ is a parameter which balances the effects of two terms.

Notice the error term $||X - XZ||_{2,p}$ is not caused by noise, but the limited representational capability with insufficient data. We omit the proof here since it is very similar to Theorem 2 in [14], one only needs to replace $||Z^T Z||_1$ as $f(Z)$ and complete the proof similarly. Theorem 3 guarantees the block diagonal structure of the solution to problem (11) from insufficient data. Though the orthogonal subspaces assumption is possibly violated in real data, Theorem 3 provides a theoretical low bound for correct segmentation from insufficient data.

3 Subspace Segmentation via LSR

3.1 Least Squares Regression

The theoretical analysis in Section 2 not only summarizes some existing works, but also helps us design a new criterion. Theorem 2 shows that many criteria guarantee the block diagonal property between-cluster. But the within-cluster affinities are also very important for subspace segmentation. Sparse representation encourages sparsity not only between-cluster, but also within-cluster, thus it misses the important correlation structure in the data. However, most data exhibit strong correlations. The subspace segmentation problem studied in this paper assumes data are drawn from a union of subspaces. If the sampling data are sufficient, they tend to be highly correlated. For modeling such data, we present the *Least Squares Regression (LSR)* method as follow:

$$\min ||Z||_F \ \text{ s.t. } X = XZ,\ \text{diag}(Z) = 0. \tag{12}$$

Here $||Z||_F$ denotes the Frobenius norm of Z, *i.e.* $||Z||_F = (\sum_{i=1}^n \sum_{j=1}^n Z_{ij}^2)^{\frac{1}{2}}$, which is a special case of $(\sum_{i=1}^n \sum_{j=1}^n \lambda_{ij} |Z_i j|^{p_{ij}})^s$. As corollaries to Theorem 2 and Theorem 3, we have

Theorem 4. *Assume the data sampling is sufficient, and the subspaces are independent. The optimal solution Z^* to problem (12) is block diagonal:*

$$Z^* = \begin{bmatrix} Z_1^* & 0 & \cdots & 0 \\ 0 & Z_2^* & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & Z_k^* \end{bmatrix},$$

where $Z_i^* \in \mathbb{R}^{n_i \times n_i}$ is also the optimal solution to the following problem:

$$\min ||Y||_F \ \textit{ s.t. } X_i = X_i Y,\ \textit{diag}(Y) = 0. \tag{13}$$

Theorem 5. *If the subspaces are orthogonal, the optimal solution to the following problem*

$$\min ||X - XZ||_{2,p} + \lambda ||Z||_F^2 \tag{14}$$

is block diagonal, where $p > 0$, $\lambda > 0$.

The above two theorems show LSR can reveal the true subspace membership under some data assumption.

3.2 LSR with Noise

The data from real applications are always contaminated with noise. Similar to SSC, we use the Frobenius norm to penalize the noise as follow:

$$\min ||X - XZ||_F^2 + \lambda ||Z||_F^2 \text{ s.t. } \mathrm{diag}(Z) = 0, \tag{15}$$

where $\lambda > 0$ is a parameter used to balance the effects of the two parts. Problem (15) can be efficiently solved by the following analytical solution:

Theorem 6. *The optimal solution to problem (15) is*

$$Z^* = -D(diag(D))^{-1} \text{ and } diag(Z^*) = 0, \tag{16}$$

where $D = (X^T X + \lambda I)^{-1}$.

The proof of Theorem 6 is presented in Appendix.

The constraint $\mathrm{diag}(Z) = 0$ can be removed from problem (15) which leads to another formulation of LSR as follow:

$$\min ||X - XZ||_F^2 + \lambda ||Z||_F^2. \tag{17}$$

It also has an analytical solution:

$$Z^* = (X^T X + \lambda I)^{-1} X^T X. \tag{18}$$

The problem (17) is actually the well known Tikhonov regularization [21] or ridge regression [22]. We will discuss why it is better than SSC and LRR for subspace segmentation in next subsection.

3.3 Why LSR?

The motivation by using LSR for subspace segmentation is that it tends to shrink coefficients of correlated data and groups them together. LSR exhibits the grouping effect that the coefficients of a group of correlated data are approximately equal. The grouping effect of LSR is stated in the following theorem:

Theorem 7. *Given a data vector* $y \in \mathbb{R}^d$, *data points* $X \in \mathbb{R}^{d \times n}$ *and a parameter* λ. *Assume each data point of* X *are normalized. Let* z^* *be the optimal solution to the following LSR (in vector form) problem:*

$$\min ||y - Xz||_2^2 + \lambda ||z||_2^2. \tag{19}$$

Algorithm 1. Subspace Segmentation via LSR

Input: data matrix X, number of subspaces k.

1. Solve the LSR problem by (16) or (18).
2. Define the affinity matrix by $(|Z^*| + |(Z^*)^T|)/2$.
3. Segment the data into k subspaces by Normalized Cuts.

We have

$$\frac{||z_i^* - z_j^*||_2}{||y||_2} \leq \frac{1}{\lambda}\sqrt{2(1-r)}, \tag{20}$$

where $r = x_i^T x_j$ is the sample correlation.

The proof of Theorem 7 is presented in Appendix.

The grouping effect of LSR presented in the above theorem shows that the solution is correlation dependent. If x_i and x_j are highly correlated, *i.e.* $r = 1$ (if $r = -1$ then consider $-x_j$), Theorem 7 says that the difference between the coefficient paths of x_i and x_j is almost 0. Thus x_i and x_j will be grouped in the same cluster.

Sparse representation does not have the grouping effect and even it is unstable. The grouping effect of LRR is still unclear [4], but note that (20) is a tight bound and so we can expect LSR to possess greater grouping capabilities. It is interesting that the solution (18) of LSR is also of low rank, $\text{rank}(Z^*) = \text{rank}(X)$. But the effectiveness of LSR for subspace segmentation comes from its grouping effect, it has nothing to do with the property of its low rankness.

3.4 Algorithm of Subspace Segmentation by LSR

Similar to SSC and LRR, our method is also a spectral clustering-based method. We first solve the LSR problem by (16) or (18) (the solution is denoted by Z^*), then define the affinity matrix as $(|Z^*| + |(Z^*)^T|)/2$. The spectral clustering algorithm such as Normalized Cuts [23] is employed to produce the ultimate segmentation results. In summary, we have the entire Least Squares Regression (LSR) algorithm for segmenting data drawn from multiple subspaces in Algorithm 1.

4 Experimental Verification

In this section, we evaluate LSR [5] on the Hopkins 155 motion database and Extended Yale Database B, in comparison with SSC [6] and LRR [7]. We implement two versions of

[4] The original LRR by rank minimization (3) does not have the grouping effect.

[5] The Matlab code: `http://home.ustc.edu.cn/~canyilu/`

[6] The Matlab code: `http://www.vision.jhu.edu/code/`, it uses the CVX package for solving the sparse representation problem, we instead use the SPAMS package (`http://www.di.ens.fr/willow/SPAMS/index.html`), which is very efficient due to its implementation by C++. It is not fair, but SSC is still slower than LRR and LSR.

[7] The Matlab code: `http://sites.google.com/site/guangcanliu/`

Table 2. The parameter, segmentation errors (%), and running time (s) of each method on the Hopkins 155 Database

		SSC	LRR	LSR1	LSR2
Parameter λ		2×10^{-3}	2.4	4.8×10^{-3}	4.6×10^{-3}
Error	Max	39.53	36.36	36.36	36.36
	Mean	4.02	3.23	**2.50**	2.84
	Median	0.90	0.50	**0.31**	0.34
	STD	10.04	6.06	**5.62**	6.16
Running times		149.70	129.30	24.33	**21.35**

LSR, LSR1 denotes the LSR algorithm by (16), and LSR2 denotes the LSR based on (18). All experiments are carried out by by using Matlab on a PC with 2.4GHz CPU and 2GB RAM.

4.1 Motion Segmentation

We apply LSR on the Hopkins 155 database [8]. It consists of 156 sequences of two or three motions (a motion corresponding to a subspace). Each sequence is a sole segmentation task and so there are 156 subspace segmentation tasks totally. We first use PCA to project the data onto a 12-dimensional subspace. Then SSC, LRR and LSR are employed for comparison. The parameter is manually tuned for each method and we report the best result. Table 2 shows the parameter, segmentation errors, and running time of each method on the Hopkins 155 database. Notice the running time show in the Table 2 is the cost that we compute the affinity matrix by different methods.

4.2 Face Clustering

We evaluate LSR on the Extended Yale Database B [25]. In this experiment, we use the first 5 and 10 classes data, each class contains 64 images. The images are resized into 32×32. Then the data are projected onto a 5×6-dimensional subspace for 5 classes clustering problem by PCA, and a 10×6-dimensional subspace for 10 classes clustering problem. The three methods all perform well on the above setting. Table 3 lists the parameter, segmentation accuracies, and running time of each method on the Extended Yale Database B.

4.3 Experimental Results and Discussions

Based on the experimental results shown in Table 2 and Table 3, we have the following observations and discussions:

(1) LSR outperforms SSC and LRR on the Hopkins 155 database and Extended Yale Database B. The advantage of LSR is mainly due to its grouping effect for modeling the correlation structure of data. Notice the motion data exhibit strong correlations, the dimension of each affine subspace is at most three [11]. The correlation structure of face images has been widely used for face recognition [24] [25].

[8] `http://www.vision.jhu.edu/data/hopkins155/`

Table 3. The parameter, segmentation accuracies (%), and running time (s) of each method on the Extended Yale Database B

		SSC	LRR	LSR1	LSR2
Parameter λ	5	0.05	0.15	0.4	0.4
	10	0.05	1.5	0.004	0.004
Accuracy	5	76.88	81.88	88.13	**91.56**
	10	47.81	65.00	70.16	**72.34**
Running time	5	0.657	0.602	0.018	**0.009**
	10	4.760	2.261	0.101	**0.045**

(2) LSR is robust. Theorem 1.9 in [26] shows that problem (17) is equivalent to a robust optimization problem, when the data X is subjected to a bounded matrix disturbance with a Frobenius norm.
(3) Beyond the performance and robustness, the experimental results show that LSR is more efficient than SSC and LRR. SSC solves the ℓ^1-minimization problem, which is non-smooth and requires much computational cost. The optimization of LRR by inexact ALM which demands hundreds of singular value decomposition and the convergence is not strictly proved in theory.
(4) The most important is that LSR is simpler and better. The previous models for subspace segmentation, SSC, LRR, MSR and SSQP, are unnecessary sophistication.

5 Conclusions

This paper explores the subspace segmentation problem. We first theoretically provide an Enforced Block Diagonal (EBD) conditions, and show that if the subspaces are independent, many criteria which satisfy the EBD conditions always produce a block diagonal solution. The EBD conditions are general and easy to be satisfied, the existing methods, SSC, LRR, MSR and SSQP are all special cases. Furthermore, the data sampling can be insufficient when the subspaces are orthogonal. Second, considering that sufficient data which can be characterized by linear subspace are usually highly correlated, we further propose the Least Squares Regression (LSR) method which takes advantage of the correlation of data. We theoretically show that the grouping effect of LSR makes it group the highly correlated data together, and also it is robust to noise. Experimental results on real data demonstrate that LSR is efficient and effective, with comparison to the state-of-the-art subspace segmentation methods SSC and LRR. LSR is simple while the existing methods are sophisticated.

Acknowledgment. This research is partially supported by the Singapore National Research Foundation under its International Research Centre @ Singapore Funding Initiative and administered by the IDM Programme Office, the grants of the NSFC, Nos. 61005010, 61005007, 31071168, 61133010 and 61100161.

Appendix

Proof of Theorem 6

Proof. Assume $X = [x_1, x_2, \cdots, x_n] \in \mathbb{R}^{d\times n}$. Remove the i-th column of X, we get $Y_i = [x_1, \cdots, x_{i-1}, x_{i+1}, \cdots, x_n]$. The i-th column of Z^* is $[Z^*]_i = E_i Y_i^T x_i$, where $E_i = (Y_i^T Y_i + \lambda I)^{-1}$. But it is not efficient to compute an inverse of matrix for each x_i. We will show how to obtain $[Z^*]_i$ from $D = (X^T X + \lambda I)^{-1}$ which can be pre-calculated. First, we arrange X as $XP = [Y_i\ x_i]$, where P is a permutation matrix, $PP^T = P^T P = I$. Thus we have

$$[P^T(X^T X + \lambda I)P]^{-1} = P^T D P. \tag{21}$$

On the other hand, we can compute $[P(X^T X + \lambda I)P]^{-1}$ by using the Woodbury formula [27] as follow:

$$\begin{aligned}
&[P^T(X^T X + \lambda I)P]^{-1} \\
&= \begin{bmatrix} Y_i^T Y_i + \lambda I & Y_i^T x_i \\ x_i^T Y_i & x_i^T x_i + \lambda \end{bmatrix}^{-1} \\
&= \begin{bmatrix} E_i & 0 \\ 0 & 0 \end{bmatrix} + \beta_i \begin{bmatrix} b_i b_i^T & b_i \\ b_i^T & 1 \end{bmatrix}
\end{aligned}$$

where

$$b_i = -[Z^*]_i,$$
$$\beta_i = x_i^T x_i + \lambda - x_i^T Y_i E_i Y_i^T x_i.$$

Thus $[Z^*]_i = -b_i$ is what we need, and we can get b_i from (21) by considering the property of P, we get

$$Z^*_{ji} = \begin{cases} -\frac{D_{ji}}{D_{ii}}, & j \neq i, \\ 0, & j = i. \end{cases}$$

The solution can be rewritten as $Z^* = -D(\text{diag}(D))^{-1}$, $\text{diag}(Z^*) = 0$.

Proof of Theorem 7

Proof. Let $L(z) = ||y - Xz||_2^2 + \lambda||z||_2^2$. Since z^* is the optimal solution to problem (19), it satisfies

$$\left.\frac{\partial L(z)}{\partial z_k}\right|_{z=z^*} = 0. \tag{22}$$

Thus we have

$$-2x_i^T(y - Xz^*) + 2\lambda z_i^* = 0, \tag{23}$$

$$-2x_j^T(y - Xz^*) + 2\lambda z_j^* = 0. \tag{24}$$

Equation (23) and (24) gives

$$z_i^* - z_j^* = \frac{1}{\lambda}(x_i^T - x_j^T)(y - Xz^*). \tag{25}$$

Since each column of X is normalized, $||x_i - x_j||_2 = \sqrt{2(1-r)}$ where $r = x_i^T x_j$. Notice z^* is optimal to problem (19), we get

$$||y - Xz^*||_2^2 + \lambda||z^*||_2^2 = L(z^*) \leq L(0) = ||y||_2^2 \tag{26}$$

Thus $||y - Xz^*||_2 \leq ||y||_2$. Then equation (25) implies

$$\frac{||z_i^* - z_j^*||_2}{||y||_2} \leq \frac{1}{\lambda}\sqrt{2(1-r)}. \tag{27}$$

References

1. Hong, W., Wright, J., Huang, K., Ma, Y.: Multiscale hybrid linear models for lossy image representation. IEEE Transactions on Image Processing 15(12), 3655–3671 (2006)
2. Ho, J., Yang, M.-H., Lim, J., Lee, K.-C., Kriegman, D.: Clustering appearances of objects under varying illumination conditions. In: IEEE Conference on Computer Vision and Pattern Recognition, vol. 1 (2003)
3. Kanatani, K.: Motion segmentation by subspace separation and model selection. In: International Conference on Computer Vision, vol. 2, pp. 586–591 (2001)
4. Yan, J., Pollefeys, M.: A General Framework for Motion Segmentation: Independent, Articulated, Rigid, Non-rigid, Degenerate and Non-degenerate. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006. LNCS, vol. 3954, pp. 94–106. Springer, Heidelberg (2006)
5. Costeira, J.P., Kanade, T.: A multibody factorization method for independently moving objects. International Journal of Computer Vision 29(3), 159–179 (1998)
6. Vidal, R., Ma, Y., Sastry, S.: Generalized principal component analysis (gpca). IEEE Transactions on Pattern Analysis and Machine Intelligence 27(12), 1945–1959 (2005)
7. Bradley, P.S., Mangasarian, O.L.: k-plane clustering. Journal of Global Optimization 16(1), 23–32 (2000)
8. Zhang, T., Szlam, A., Lerman, G.: Median k-flats for hybrid linear modeling with many outliers. In: International Conference on Computer Vision Workshops, pp. 234–241 (2009)
9. Tipping, M.E., Bishop, C.M.: Mixtures of probabilistic principal component analyzers. Neural Computation 11(2), 443 (1999)
10. Ma, Y., Derksen, H., Hong, W., Wright, J.: Segmentation of multivariate mixed data via lossy data coding and compression. IEEE Transactions on Pattern Analysis and Machine Intelligence 29(9), 1546–1562 (2007)
11. Elhamifar, E., Vidal, R.: Sparse subspace clustering. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 2790–2797 (2009)
12. Liu, G., Lin, Z., Yu, Y.: Robust subspace segmentation by low-rank representation. In: International Conference on Machine Learning, pp. 663–670 (2010)
13. Luo, D., Nie, F., Ding, C., Huang, H.: Multi-Subspace Representation and Discovery. In: Gunopulos, D., Hofmann, T., Malerba, D., Vazirgiannis, M. (eds.) ECML PKDD 2011, Part II. LNCS(LNAI), vol. 6912, pp. 405–420. Springer, Heidelberg (2011)
14. Wang, S., Yuan, X., Yao, T., Yan, S., Shen, J.: Efficient subspace segmentation via quadratic programming. In: Innovative Applications of Artificial Intelligence Conference, vol. 1, pp. 519–524 (2011)
15. Vidal, R.: A tutorial on subspace clustering. IEEE Signal Processing Magazine (2010)
16. Elhamifar, E., Vidal, R.: Clustering disjoint subspaces via sparse representation. In: 2010 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2010, March 14-19, pp. 1926–1929 (2010)

17. Liu, G., Lin, Z., Yan, S., Sun, J., Yu, Y., Ma, Y.: Robust recovery of subspace structures by low-rank representation. Arxiv preprint arXiv:1010.2955 (2010)
18. Donoho, D.L.: For most large underdetermined systems of linear equations the minimal ℓ^1-norm solution is also the sparsest solution. Communications on Pure and Applied Mathematics 59(6), 797–829 (2006)
19. Zou, H., Hastie, T.: Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society Series B-Statistical Methodology 67, 301–320 (2005)
20. Candes, E.J., Li, X.D., Ma, Y., Wright, J.: Robust principal component analysis? Journal of the ACM 58(3) (2011)
21. Tikhonov, A.N.: Solution of incorrectly formulated problems and the regularization method (1963)
22. Hoerl, A.E., Kennard, R.W.: Ridge regression: Biased estimation for nonorthogonal problems, 12 (1970)
23. Shi, J.B., Malik, J.: Normalized cuts and image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence 22(8), 888–905 (2000)
24. Lee, K.C., Ho, J., Kriegman, D.J.: Acquiring linear subspaces for face recognition under variable lighting. IEEE Transactions on Pattern Analysis and Machine Intelligence 27(5), 684–698 (2005)
25. Wright, J., Yang, A.Y., Ganesh, A., Sastry, S.S., Ma, Y.: Robust face recognition via sparse representation. IEEE Transactions on Pattern Analysis and Machine Intelligence 31(2), 210–227 (2009)
26. Sra, S., Nowozin, S., Wright, S.J. (eds.): Optimization for Machine Learning. The MIT Press (2010)
27. Golub, G.H., Loan, C.F.V.: Matrix computations, vol. 3. Johns Hopkins Univ. Pr. (1996)