Delving Deep into Least Square Regression Model for Subspace Clustering

Masataka Yamaguchi
masataka.yamaguchi.1016@gmail.com
NTT Corporation, Japan
Go Irie
go.irie.5@gmail.com
Takahito Kawanishi
kawanishi.takahito@lab.ntt.co.jp
Kunio Kashino
kashino.kunio@lab.ntt.co.jp

Abstract

Subspace clustering is the problem of clustering data drawn from a union of multiple subspaces. The most popular subspace clustering framework in recent years is the spectral clustering-based approach, which performs subspace clustering by first computing an affinity matrix and then applying spectral clustering to it. One of the representative methods for computing an affinity matrix is the least square regression (LSR) model, which is based on the idea of self-representation. Although its efficiency and effectiveness have been empirically validated, it lacks some theoritical analysis and practicality, *e.g.*: absense of interpretations, lack of theoretical analysis on its robustness, absense of guidelines for choosing the hyper-parameter, and the scalability. This paper aims at providing novel insights for better understanding on LSR, and also improving its practicality. For this purpose, we present four contributions: first, we present a novel interpretation of LSR, which is based on random sampling perspective. Second, we provide novel theoretical analysis on LSR's robustness toward outliers. Third, we theoretically and empirically demonstrate that selecting a larger value for the hyper-parameter tends to result in good clustering results. Finally, we derive another equivalent form of the LSR's solution, which can be computed with less time complexity than the original form regarding the data size.

1 Introduction

In many practical scenarios, high dimensional data often live in a union of low-dimensional linear subspaces. The problem of partitioning such data so that each cluster consists of all the data belonging to one subspace is called *Subspace Clustering*. Subspace clustering has greatly attracted attention as it has important and wide-ranging applications in various fields such as computer vision [23, 36], data mining [2, 27], network analysis [7, 11], system identification [3, 37], and biology [17, 24].

Prior Work: In the past few decades, many subspace clustering methods have been proposed, including algebraic methods [4, 8, 13, 15, 16, 25, 38], iterative methods [1, 5, 33, 42],

statistical methods [29, 30, 31, 41], and spectral clustering based methods [6, 9, 10, 14, 18, 20, 21, 22, 35, 39, 40]. Recent efforts have been made on spectral clustering based methods, as these often perform superiorly to the other methods in practical settings.

Most spectral clustering based methods are performed in two steps. The first step is to compute an affinity matrix such that a pair of data points belonging in the same subspace has higher affinity than those in different subspaces, and the second step is to partition data by applying spectral clustering to that affinity matrix. The first step is crucial to the success of subspace clustering, and several approaches to construct reasonable affinity matrices have been proposed [9, 10, 26, 40]. The most representative approach is self-representation based [9, 10].

The self-representation based approach first builds a self-representation matrix Z^* that is computed by representing each data points by a linear combination of the others and then computes an affinity matrix M using Z^* (e.g., $M_{ij} = |Z^*_{ij}| + |Z^{*\mathrm{T}}_{ij}|$). To compute the self-representation matrix Z^*, most self-representation based methods first solve the following problem:

$$\min_{Z \in \mathcal{C}} \mathrm{h}(E) + \lambda \mathrm{r}(Z),\ s.t.\ X = XZ + E, \tag{1}$$

where $X \in \mathbb{R}^{D \times N} = [\boldsymbol{x}_1, ..., \boldsymbol{x}_N]$ is a data matrix that consists of N D-dimensional data points, $\mathrm{h}(E)$ is the loss function for reconstruction residual E, $\mathrm{r}(Z)$ is a regularizer for a self-representation matrix Z, and $\mathcal{C}$ is a constraint set for Z. $\mathrm{h}(E)$ indirectly regularizes a self-representation matrix Z so that XZ is close to X and the most popular choice for $\mathrm{h}(E)$ is the squared Frobenius norm $\mathrm{h}(E) = ||E||_F^2$ [9, 10, 14, 18, 21, 22, 35, 39]. $\mathrm{r}(Z)$ regularizes a self-representation matrix Z so that each data point is reconstructed by using those in the same subspace and various types of regularizers, such as L1 norm $\mathrm{r}(Z) = ||Z||_1$ and nuclear norm $\mathrm{r}(Z) = ||Z||_*$, have been investigated so far [9, 10, 14, 18, 20, 21, 22, 35, 39].

Typically, Eq. (1) is solved via some iterative optimization algorithms, *e.g.*, alternating direction method of multipliers (ADMM). However, for certain pairs of $\mathrm{r}(Z)$ and $\mathrm{h}(E)$, Eq. (1) has a closed-form solution, and by using such a solution, we can efficiently solve Eq. (1) much faster than using iterative optimization algorithms. One of the most representative methods that have closed-form solusions is the least square regression (LSR) model, which we will explain in the following.

LSR: Lu *et al.* [21] theoretically showed that if Eq. (1) satisfies certain conditions, which they call the enforced block-diagonal (EBD) conditions, we are able to get a block diagonal solution under the independent subspaces assumption. As a special case of that condition, they presented LSR, which employs the Frobenius norm for both $\mathrm{r}(Z)$ and $\mathrm{h}(E)$, *i.e.*:

$$\min_{Z} ||E||_F^2 + \lambda ||Z||_F^2, s.t.\ X = XZ + E,\ Z \in \mathcal{C}. \tag{2}$$

In this paper, following other methods such as sparse subspace clustering (SSC), we specifically consider the case of $\mathcal{C} = \{C | C \in \mathrm{R}^{N \times N}, C_{ii} = 0\}$. When $\mathcal{C} = \{C | C \in \mathrm{R}^{N \times N}, C_{ii} = 0\}$, Eq. (2) has the following closed-form solution:

$$Z^* = I - D(\mathrm{Diag}(\mathrm{diag}(D)))^{-1},\ where\ D = (X^{\mathrm{T}}X + \lambda I)^{-1}, \tag{3}$$

where $\mathrm{Diag}(\boldsymbol{\lambda})$ converts a vector $\boldsymbol{\lambda}$ into a diagonal matrix the i^{th} diagonal entry of which is $(\boldsymbol{\lambda})_i$, and $\mathrm{diag}(\Lambda)$ converts Λ into a vector the i^{th} entry of which is the i^{th} diagonal entry of Λ.

LSR is a very pracical subspace clustering method due to its simplicity, effectiveness and efficiency. However, it lacks some theoritical analysis and practicality, *e.g.*:

- Absence of interpretations. Although Lu *et al*. presented LSR as a special case of the EBD condition, there are no specific reason to choose the Frobenius norm among various regularizers for $\mathrm{r}(Z)$ (except the existence of a closed-form solution).
- Lack of theoretical analysis on its robustness.
- Absence of guidelines for choosing the hyper-parameter λ. In the practical settings, one has to choose the hyper-parameter λ. However, to the best of our knowledge, no strategy are provided for choosing it. This is problematic especially when one cannot tune λ (*e.g*., when a validation dataset is not available).
- Scalability. Although the solution of LSR can be efficiently computed due to Eq. (3), it still takes $O(N^3)$ time complexity with reference to the data size N.

Contributions: This paper aims at providing novel insights for better understanding on LSR, and also improving its practicality. For this purpose, we present four novel theoretical and empirical analysis on LSR:

- We present a novel interpretation of LSR based on random sampling, a core technique in the field of robust subspace recovery. More specifically, we show that the Frobenius norm is derived as a regularizer for Z by introducing random sampling into Eq. (1) when data points are normalized and $\mathrm{h}(E) = ||E||_F^2$. Moreover, we also show that the hyper-paramter λ can be interpreted as $\lambda = \frac{1-\alpha}{\alpha}$, where α is a sampling rate of ramdom sampling.
- We provide novel theoretical analysis on LSR's robustness toward outliers.
- We prove that λZ^*, the solution of Eq. (3) multiplied by λ, approaches to a specific matrix as the hyper-parameter λ increases. Moreover, we empirically demonstrate that increasing λ does not cause a big drop in the accuracy, and hence suggest that selecting a larger value for the hyper-parameter λ would be better when one cannot tune it.
- We derive another equivalent form of Eq. (3), which can be computed with less time complexity than the original form regarding the data size.

2 Deriving LSR via Random Sampling Perspective

We first introduce a novel interpretation of the objective of LSR, *i.e*., Eq. (2). Our interpretation is derived by introducing random sampling into Eq. (1).

Why random sampling?: One of the major factors that degrade subspace clustering accuracy is the existence of outliers in a given data matrix X. When using self-representation-based methods, outliers may be connected each other. As a result, they may constitute incorrect clusters in the final clustering results.

For mitigating this problem, we consider utilizing random sampling, a core technique used in random sample consensus (RANSAC) [12]. RANSAC first repeats (1) sampling a few data from a given data matrix and (2) conducting subspace recovery using sampled data, and then outputs the most suitable subspace from all the computed subspaces. By using only a few sampled data for the subspace recovery step, RANSAC can reduce the number of

outliers to consider at the same time, which leads to reducing the negative impact of outliers. Although Yang *et al*. [41] proposed solving subspace clustering by applying RANSAC, it has some drawbacks, one of which is that its performance deteriorates quickly as the number of subspaces increases [34].

Inspired by RANSAC, we introduce random sampling into the problem (1). By doing so, we can also reduce the number of outliers to consider at the same time, which leads to preventing outliers from constituting incorrect clusters in the final clustering results.

Derivation: In the following, we introduce random sampling into the problem (1). Let $\boldsymbol{\theta} \in \{0,1\}^N$ be an indicator vector such that $(\boldsymbol{\theta})_i = 1$ if the i^{th} data of X are sampled. We assume $\boldsymbol{\theta}$ is sampled from a distribution $\mathrm{P}(\boldsymbol{\theta}|\alpha)$ that samples each element of $\boldsymbol{\theta}$ from a Bernoulli distribution with probability $\alpha \in (0,1)$ independently. In addition, we adopt the squared Frobenius norm $\mathrm{h}(E) = ||E||_F^2$ for a reconstruction error $\mathrm{h}(E)$. Using data randomly sampled from X by $\mathrm{P}(\boldsymbol{\theta}|\alpha)$, we redefine the problem (1) as follows:

$$\min_{Z\in\mathcal{C}} ||E_{\boldsymbol{\theta}}||_F^2 + \lambda \mathrm{r}(Z),\ s.t.\ XP_{\boldsymbol{\theta}} = XP_{\boldsymbol{\theta}}P_{\boldsymbol{\theta}}^{\mathrm{T}}ZP_{\boldsymbol{\theta}} + E_{\boldsymbol{\theta}},\ \boldsymbol{\theta} \sim \mathrm{P}(\boldsymbol{\theta}|\alpha), \tag{4}$$

where $P_{\boldsymbol{\theta}}$ is a selection matrix that keeps only data selected by $\boldsymbol{\theta}$, *i.e*., a matrix excluding the i^{th} column from an identity matrix if $(\boldsymbol{\theta})_i = 0$.

The problem (4) excludes a part of data matrix X from the objective, hence the corresponding part of Z cannot be correctly learned. Therefore, instead of using $||E_{\boldsymbol{\theta}}||_F^2$ directly as in the problem (4), we use the expectation of $||E_{\boldsymbol{\theta}}||_F^2$ as follows:

$$\min_{Z\in\mathcal{C}} \mathrm{E}_{\boldsymbol{\theta}\sim\mathrm{P}(\boldsymbol{\theta}|\alpha)}[||E_{\boldsymbol{\theta}}||_F^2] + \lambda \mathrm{r}(Z),\ s.t.\ XP_{\boldsymbol{\theta}} = XP_{\boldsymbol{\theta}}P_{\boldsymbol{\theta}}^{\mathrm{T}}ZP_{\boldsymbol{\theta}} + E_{\boldsymbol{\theta}}. \tag{5}$$

As all the data in X are considered in the problem (5), a full part of Z is expected to be correctly learned.

We next consider how to compute the expectation $\mathrm{E}_{\boldsymbol{\theta}\sim\mathrm{P}(\boldsymbol{\theta}|\alpha)}[||E_{\boldsymbol{\theta}}||_F^2]$ in the problem (5). One way to compute it is to approximate $\mathrm{E}_{\boldsymbol{\theta}\sim\mathrm{P}(\boldsymbol{\theta}|\alpha)}[||E_{\boldsymbol{\theta}}||_F^2]$ by the Monte Carlo algorithm (*i.e*., computing $\frac{1}{N}\sum_i ||E_{\boldsymbol{\theta}_i}||_F^2$ with N stochastically sampled indicator vectors $\{\boldsymbol{\theta}_i\}_{i=1}^N$), which may degrade clustering results due to the approximation of the expectation $\mathrm{E}_{\boldsymbol{\theta}\sim\mathrm{P}(\boldsymbol{\theta}|\alpha)}[||E_{\boldsymbol{\theta}}||_F^2]$, and also requires high computational cost for computing $||E_{\boldsymbol{\theta}}||_F^2$ multiple times. However, fortunately, we find that the expectation $\mathrm{E}_{\boldsymbol{\theta}\sim\mathrm{P}(\boldsymbol{\theta}|\alpha)}[||E_{\boldsymbol{\theta}}||_F^2]$ can be explicitly represented as follows (see the supplementary for the derivation of this):

$$\begin{aligned}\mathrm{E}_{\boldsymbol{\theta}\sim\mathrm{P}(\boldsymbol{\theta}|\alpha)}[||E_{\boldsymbol{\theta}}||_F^2] =& \alpha||X - XZ'||_F^2 + (1-\alpha)||\mathrm{Diag}(\mathrm{diag}(X^{\mathrm{T}}X))^{\frac{1}{2}}Z'||_F^2 \\ &+ 2(1-\alpha)\mathrm{trace}(X^{\mathrm{T}}X(Z'-I)\mathrm{Diag}(\mathrm{diag}(Z'))) \\ &+ \frac{(2\alpha-1)(\alpha-1)}{\alpha}\mathrm{trace}(X^{\mathrm{T}}X\mathrm{Diag}(\mathrm{diag}(Z'))^2),\end{aligned} \tag{6}$$

where $Z' = \alpha Z$. By using Eq. (6), the expectation $\mathrm{E}_{\boldsymbol{\theta}\sim\mathrm{P}(\boldsymbol{\theta}|\alpha)}[||E_{\boldsymbol{\theta}}||_F^2]$ can be exactly computed without repeatedly sampling $\boldsymbol{\theta}$ and computing $||E_{\boldsymbol{\theta}}||_F^2$, which enables the objective in the problem (5) to be efficiently computed.

Finally, by $\mathcal{C} = \{Z | Z \in \mathbb{R}^{N\times N}, Z_{ii} = 0\}$, Eq. (6) can be simplified as follows:

$$\mathrm{E}_{\boldsymbol{\theta}\sim\mathrm{P}(\boldsymbol{\theta}|\alpha)}[||E_{\boldsymbol{\theta}}||_F^2] = \alpha||X - XZ'||_F^2 + (1-\alpha)||\mathrm{Diag}(\mathrm{diag}(X^{\mathrm{T}}X))^{\frac{1}{2}}Z'||_F^2. \tag{7}$$

To see this, note that $\mathrm{Diag}(\mathrm{diag}(Z')) = 0$ holds when $\mathcal{C} = \{Z | Z \in \mathbb{R}^{N\times N}, Z_{ii} = 0\}$. By Eq. (7), the problem (5) can be finally represented as follows:

$$\min_{Z\in\{Z|Z\in\mathbb{R}^{N\times N}, Z_{ii}=0\}} ||X - XZ'||_F^2 + \frac{1-\alpha}{\alpha}||\mathrm{Diag}(\mathrm{diag}(X^{\mathrm{T}}X))^{\frac{1}{2}}Z'||_F^2 + \frac{\lambda}{\alpha}\mathrm{r}(\frac{Z'}{\alpha}), \quad s.t.\ Z = \alpha Z' \tag{8}$$

From the problem (8), we can see that by applying random sampling to the problem (1), the objective of the problem (1) can be represented in a very simple form. It is worth noting that a novel term $||\mathrm{Diag}(\mathrm{diag}(X^{\mathrm{T}}X))^{\frac{1}{2}}Z'||_F^2$ in Eq. (7) can be considered as a regularizer for Z, which reveals that random sampling itself has an effect of regularizing Z. Moreover, under the typical assumption that each data point is normalized, $||\mathrm{Diag}(\mathrm{diag}(X^{\mathrm{T}}X))^{\frac{1}{2}}Z'||_F^2$ is equivalent to the Frobenius norm $||Z'||_F^2$, and hence Eq. (8) is equivalent to Eq. (2) by setting $\mathrm{r}(Z) = 0$ and replacing λ in Eq. (2) with $\frac{1-\alpha}{\alpha}$.

From this derivation, we can see that introducing random sampling into the problem (1) leads a regularization effect. In particular, the regularization effect can be represented to the Frobenius norm under some typical settings, *i.e.*, $\mathcal{C} = \{Z | Z \in \mathbb{R}^{N\times N}, Z_{ii} = 0\}$, $\mathrm{h}(E) = ||E||_F^2$ and $||\boldsymbol{x}_i||_2 = 1$ for all i. This gives a novel interpretation of the objective of LSR. Moreover, we can give a novel interpretation to the hyper-parameter λ that, under those conditions, the hyper-parameter λ in the problem (1) is equivalent to $\frac{1-\alpha}{\alpha}$, where α is a sampling rate.

3 Robustness toward Outliers

In the previous section, we derived the objective of LSR by introducing random sampling to reduce the outliers' influence, but it has been unclear how effective it is at reducing outliers' influence. In the following, we analyze the robustness of LSR toward robustness.

To conduct the analysis, we consider two types of normalized data: (1) N^m inliers $X^c = [\boldsymbol{x}_1^m, ..., \boldsymbol{x}_{N^m}^m]$ belonging to a main low-dimensional subspace $\mathbb{S}^m$ ($\boldsymbol{x}_i^m \in \mathbb{S}^m$ for $i = 1, ..., N^m$) and (2) N^m outliers $X^o = [\boldsymbol{x}_1^o, ..., \boldsymbol{x}_{N^o}^o]$ belonging to an outlier subspace $\mathbb{S}^o$ ($\mathbb{S}^m \subset \mathbb{S}^o$ and $\boldsymbol{x}_i^o \in \mathbb{S}^o$ for $i = 1, ..., N^o$). We assume that all data in X^m and X^o are sampled from distributions $\mathrm{P}^m(\boldsymbol{x})$ and $\mathrm{P}^o(\boldsymbol{x})$ that satisfy $\mathrm{P}^m(\boldsymbol{x}) = 0$ if $\boldsymbol{x} \notin \mathbb{S}^m$ or $||\boldsymbol{x}||_2 \neq 1$ and $\mathrm{P}^o(\boldsymbol{x}) = 0$ if $\boldsymbol{x} \notin \mathbb{S}^o$ or $||\boldsymbol{x}||_2 \neq 1$, respectively. Also, let $Z^* = \left[Z^{c^*}, Z^{o^*}\right] = \left[\boldsymbol{z}_1^{m^*}, ..., \boldsymbol{z}_{N^m}^{m^*}, \boldsymbol{z}_1^{o^*}, ..., \boldsymbol{z}_{N^o}^{o^*}\right]$ be the self-representation matrix computed by the problem (3).

To analyze the robustness of LSR, we have to define the metric for evaluating robustness. Among some metrics that can be used for evaluating robustness, due to its tractability, we consider how much all the outliers X^o contribute to reconstructing each data $\boldsymbol{x}_i^m$ belonging to the main subspace $\mathbb{S}^m$. More specifically, we define inliers' contribution and outliers' contribution as follows:

Definition 1 (Inliers' Contribution and Outliers' Contribution)**.** *Let O^m and O^o be the zero matrices that are the same sizes as X^m and X^o, respectively. We define the inliers' contribution to a reconstructed vector $X\boldsymbol{z}_i^m$ for x_i^m as* $\mathrm{Cont}_i^m = ([X^m, O^o]\,\boldsymbol{z}_i^m)^{\mathrm{T}}\boldsymbol{x}_i^m$. *Similarly, we define the outliers' contribution to a reconstructed vector $X\boldsymbol{z}_i^m$ for $\boldsymbol{x}_i^m$ as* $\mathrm{Cont}_i^o = ([O^m, X^o]\,\boldsymbol{z}_i^m)^{\mathrm{T}}\boldsymbol{x}_i^m$.

$[X^m, O^o]\,\boldsymbol{z}_i^m$ is a vector constructed by extracting the part composed by data belonging to $\mathbb{S}^m$ from a linear combination $X\boldsymbol{z}_i^m$, and Cont_i^m is the length of a vector computed by projecting $[X^m, O^o]\,\boldsymbol{z}_i^m$ onto $\boldsymbol{x}_i^m$, since $\boldsymbol{x}_i^m$ is normalized. Cont_i^o can be interpreted similarly.

We use Cont_i^m and Cont_i^o as proxies for how much X_m and X_o contribute to reconstructing data $\boldsymbol{x}_i^m$, respectively.

We find that the ratio of Cont_i^o to Cont_i^m approaches a specific value as the amount of sampled data approaches to infinity, as shown in the following theorem:

Theorem 1. *Suppose* $\mathrm{P}^m(\boldsymbol{x})$ *and* $\mathrm{P}^o(\boldsymbol{x})$ *are independent of the direction of* $\boldsymbol{x}$ *as long as* $\boldsymbol{x}$ *is in each subspace* $\mathbb{S}^m$ *and* $\mathbb{S}^o$*, respectively (i.e.,*$\mathrm{P}^m(\boldsymbol{x}) = \mathrm{P}^m(\boldsymbol{x}')$ *if* $\boldsymbol{x} \in \mathbb{S}^m$*,* $\boldsymbol{x}' \in \mathbb{S}^m$ *and* $||\boldsymbol{x}||_2 = ||\boldsymbol{x}'||_2 = 1$*, and* $\mathrm{P}^o(\boldsymbol{x}) = \mathrm{P}^o(\boldsymbol{x}')$ *if* $\boldsymbol{x} \in \mathbb{S}^o$*,* $\boldsymbol{x}' \in \mathbb{S}^o$ *and* $||\boldsymbol{x}||_2 = ||\boldsymbol{x}'||_2 = 1$*). If the ratio of* N^m *to* N^o *is fixed, i.e.,* N^m *and* N^o *can be represented as* $N^m = M^m k$ *and* $N^o = M^o k$ *by using a natural number* k *and positive numbers* M^m *and* M^o*, we have:*

$$\operatorname*{plim}_{k\to\infty} \frac{\mathrm{Cont}_i^o}{\mathrm{Cont}_i^m} = \frac{M^o}{M^m} \frac{\dim(\mathbb{S}^m)}{\dim(\mathbb{S}^o)}, \tag{9}$$

where plim *is the probability limit operator.*

Based on Theorem 1, the LSR's solution approximately satisfies the following two properties when N^m and N^o are sufficiently large:

- The larger N^m is than N^o, the smaller the outliers' influence on a reconstruct vector $X\boldsymbol{z}_i^m$ is.
- The smaller $\dim(\mathbb{S}^m)$ is than $\dim(\mathbb{S}^o)$, the smaller the outliers' influence on a reconstruct vector $X\boldsymbol{z}_i^m$ is.

From the first property, the influence of outliers on reconstructed vectors can be guaranteed to be small if a given data matrix X contains only a few outliers, which is a convincing and also desirable result. From the second property, the influence of outliers on reconstructed vectors can be guaranteed to be small if the dimension of the outliers' influence $\dim(\mathbb{S}^o)$ is sufficiently large compared to the dimension of a main subspace $\dim(\mathbb{S}^m)$. Given that the dimension of a subspace built by outliers contained in real-world data (e.g., images with missing entries) is typically large, the influence of outliers on reconstructed vectors is expected to be small in many real-world settings.

4 Hyper-parameter Choice

In this section, we discuss how to choose the hyper-parameter λ. If we have no labeled data or cannot annotate labels to data, we cannot adjust λ based on data. For such a case, we should have some guidelines for choosing the hyper-parameter λ.

To decide how to choose the hyper-parameter λ, we conducted preliminary experiments for investigating how the hyper-parameter λ affects clustering results. More specifically, we conducted experiments with various hyper-parameters λ on three datasets: the Extended Yale Face Database B (EYaleB) [19], the MNIST dataset and the Hopkins 155 motion segmentation database (Hopkins155) [32]. For details of experimental settings, see the supplementary.

We show the experimental results in Fig. 2. It can be seen that if the hyper-parameter λ is too small, the accuracy is much worse than its peak. On the other hand, interestingly, it can be seen that increasing the hyper-parameter λ does not cause a big drop in the accuracy.

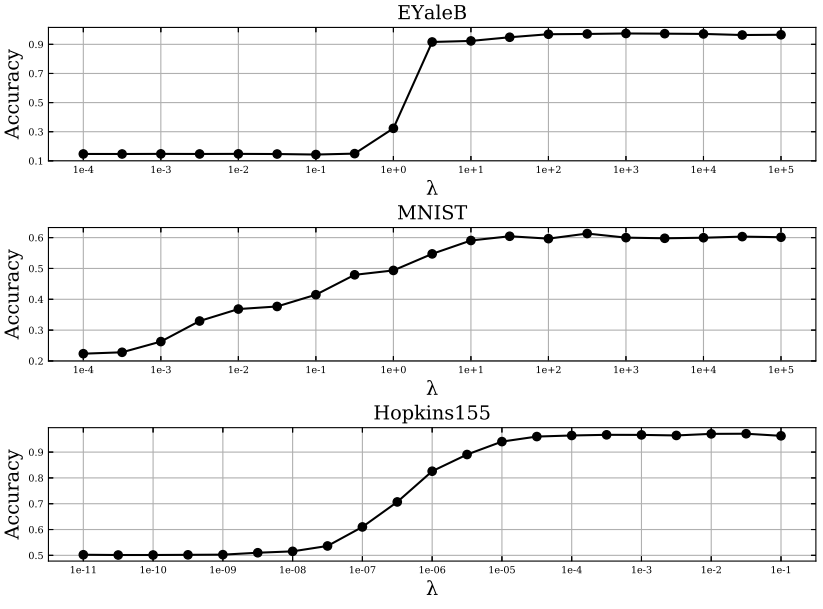

Figure 1: Mean clustering accuracy on three datasets achieved by LSR with different hyper-parameters.

Surprizingly, we find that λZ^* converges to a non-zero matrix when λ approaches ∞ as follows:

$$\lim_{\lambda\to\infty} \lambda Z^* = X^{\mathrm{T}}X - \mathrm{Diag}(\mathrm{diag}(X^{\mathrm{T}}X)) \tag{10}$$

Derivation. First, we replace λ with $\frac{1}{\tau}$ as follows:

$$Z^* = I - D(\mathrm{Diag}(\mathrm{diag}(D)))^{-1}, \textit{ where } D = (X^{\mathrm{T}}X + \frac{1}{\tau}I)^{-1} \tag{11}$$

By L'Hospital's rule, we have:

$$\begin{aligned}\lim_{\lambda\to\infty} \lambda Z^* &= \lim_{\tau\to+0} \frac{1}{\tau} Z^* \\ &= \lim_{\tau\to+0} \frac{\frac{\partial Z^*}{\partial \tau}}{\frac{\partial \tau}{\partial \tau}} \\ &= \lim_{\tau\to+0} \frac{\partial Z^*}{\partial \tau}\end{aligned} \tag{12}$$

For deriving $\lim_{\tau\to+0}\frac{\partial Z^*}{\partial\tau}$, we first derive $\frac{\partial D}{\partial\tau}$ and $\frac{\partial \mathrm{Diag}(\mathrm{diag}(D))^{-1}}{\partial\tau}$ as follows:

$$\begin{aligned}\frac{\partial D}{\partial\tau}=&\frac{1}{\tau^2}D^2\\=&\frac{1}{\tau}D(X^{\mathrm{T}}X+\frac{1}{\tau}I-X^{\mathrm{T}}X)D\\=&\frac{1}{\tau}(D-DX^{\mathrm{T}}XD)\end{aligned}\tag{13}$$

$$\begin{aligned}\frac{\partial \mathrm{Diag}(\mathrm{diag}(D))^{-1}}{\partial\tau}=&-\frac{1}{\tau}\mathrm{Diag}(\mathrm{diag}(D-DX^{\mathrm{T}}XD))\mathrm{Diag}(\mathrm{diag}(D))^{-2}\\=&-\frac{1}{\tau}\mathrm{Diag}(\mathrm{diag}(D)-\mathrm{diag}(DX^{\mathrm{T}}XD))\mathrm{Diag}(\mathrm{diag}(D))^{-2}\\=&-\frac{1}{\tau}\mathrm{Diag}(\mathrm{diag}(D))^{-1}+\frac{1}{\tau}\mathrm{Diag}(\mathrm{diag}(DX^{\mathrm{T}}XD))\mathrm{Diag}(\mathrm{diag}(D))^{-2}\end{aligned}\tag{14}$$

By Eq. (13) and Eq. (14), $\frac{\partial Z^*}{\partial\tau}$ can be represented as follows:

$$\begin{aligned}\frac{\partial Z^*}{\partial\tau}=&\frac{\partial I}{\partial\tau}-\frac{\partial D(\mathrm{Diag}(\mathrm{diag}(D)))^{-1}}{\partial\tau}\\=&-\frac{\partial D}{\partial\tau}(\mathrm{Diag}(\mathrm{diag}(D)))^{-1}-D\frac{\partial(\mathrm{Diag}(\mathrm{diag}(D)))^{-1}}{\partial\tau}\\=&-\frac{1}{\tau}(D-DX^{\mathrm{T}}XD)(\mathrm{Diag}(\mathrm{diag}(D)))^{-1}+\frac{1}{\tau}D\mathrm{Diag}(\mathrm{diag}(D))^{-1}\\&-\frac{1}{\tau}D\mathrm{Diag}(\mathrm{diag}(DX^{\mathrm{T}}XD))\mathrm{Diag}(\mathrm{diag}(D))^{-2}\\=&\frac{1}{\tau}DX^{\mathrm{T}}XD(\mathrm{Diag}(\mathrm{diag}(D)))^{-1}-\frac{1}{\tau}D\mathrm{Diag}(\mathrm{diag}(DX^{\mathrm{T}}XD))\mathrm{Diag}(\mathrm{diag}(D))^{-2}\\=&D_\tau X^{\mathrm{T}}XD_\tau(\mathrm{Diag}(\mathrm{diag}(D_\tau)))^{-1}-D_\tau\mathrm{Diag}(\mathrm{diag}(D_\tau X^{\mathrm{T}}XD_\tau))\mathrm{Diag}(\mathrm{diag}(D_\tau))^{-2}\end{aligned}\tag{15}$$

where $D_\tau=\frac{D}{\tau}=(\tau X^{\mathrm{T}}X+I)^{-1}$. By Eq. (15) and $\lim_{\tau\to+0}D_\tau=I$, $\lim_{\tau\to+0}\frac{\partial Y}{\partial\tau}$ can be represented as follows:

$$\begin{aligned}&\lim_{\tau\to+0}\frac{\partial Z^*}{\partial\tau}\\=&\lim_{\tau\to+0}D_\tau X^{\mathrm{T}}XD_\tau(\mathrm{Diag}(\mathrm{diag}(D_\tau)))^{-1}-D_\tau\mathrm{Diag}(\mathrm{diag}(D_\tau X^{\mathrm{T}}XD_\tau))\mathrm{Diag}(\mathrm{diag}(D_\tau))^{-2}\\=&X^{\mathrm{T}}X-\mathrm{Diag}(\mathrm{diag}(X^{\mathrm{T}}X)).\end{aligned}\tag{16}$$

By Eq. (12) and Eq. (16), we have $\lim_{\lambda\to\infty}\lambda Z^*=X^{\mathrm{T}}X-\mathrm{Diag}(\mathrm{diag}(X^{\mathrm{T}}X))$.

□

It is worth noting that typical graph clustering methods produce same results even after multiplying λ to Z^*. Moreover, when subspaces are orthogonal, the right side of Eq. (10) still satisfies the self-representation property (*i.e.*, $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ belong to a same subspace if $Z^*_{ij}\neq 0$).

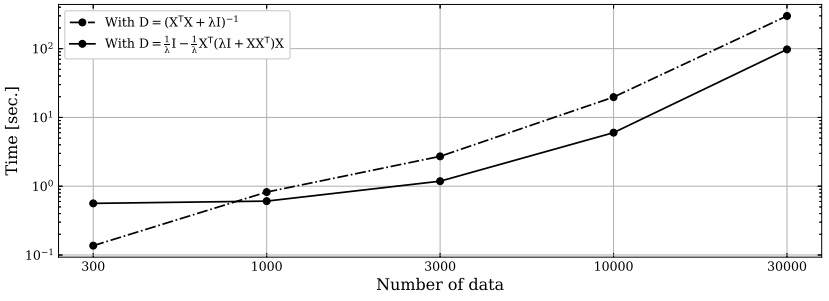

Figure 2: The time versus the number of data N. We ran experiments with Intel Xenon CPU (2.30GHz) on a single thread.

Given these facts, we can expect that LSR may still cluster data with high accuracy even when Z^* is approximately equal to the right side of Eq. (10). Based on these results, we suggest selecting a large value for the hyper-parameter λ when one cannot tune it (e.g., when a validation dataset cannot be used).

5 Improving Scalability

Although LSR has a closed-form solution, *i.e.*, Eq. (3), computing $D = (X^{\mathrm{T}}X + \lambda I)^{-1}$ requires computing an inverse matrix of $N \times N$, therefore it takes $O(N^3)$ times with reference to the data size N. In this section, we show a technique that can reduce the time complexity from $O(N^3)$ to $O(N^2)$.

To reduce the time complexity, we apply the Woodbury identity [28] to $D = (X^{\mathrm{T}}X + \lambda I)^{-1}$ as follows:

$$\begin{aligned} D &= (X^{\mathrm{T}}X + \lambda I)^{-1} \\ &= \frac{1}{\lambda} I - \frac{1}{\lambda} X^{\mathrm{T}} (\lambda I + XX^{\mathrm{T}})^{-1} X. \end{aligned} \tag{17}$$

Note that the computational complexity of Eq. (17) is $O(N^2)$, which is less than the computational complexity of directly computing $D = (X^{\mathrm{T}}X + \lambda I)^{-1}$[1].

In Fig. 2, we also show that the mean time required to computing Eq. (3) with $D = (X^{\mathrm{T}}X + \lambda I)^{-1}$ and Eq. (17) on the MNIST dataset. We randomly chose $N = \{300, 1000, 3000, 10000, 30000\}$ data points from the dataset, and measured the time for computing Eq. (3) for each N. From this figure, we can see that, by using Eq. (17), we can compute Eq. (3) faster than using $D = (X^{\mathrm{T}}X + \lambda I)^{-1}$ when the number of data is large. Therefore, we suggest that using Eq. (17) for computing Eq. (3) when N is much larger than D.

Based on these, we suggest that using Eq. (17), rather than using $D = (X^{\mathrm{T}}X + \lambda I)^{-1}$, would be better for computing Eq. (3) when N is large.

[1]Note that the time complexity of Eq. (17) is $O(D^3)$ with reference to D, whereas that of $D = (X^{\mathrm{T}}X + \lambda I)^{-1}$ is $O(D^2)$. From this, we can expect that using Eq. (17) is more effective especially when D is small.

6 Conclusion

To improve understanding on LSR and its practicality, we provided four novel theoretical and empirical analysis on it: first, we presented a novel interpretation of LSR based on random sampling, a core technique in the field of robust subspace recovery. Second, we provide novel theoretical analysis on LSR's robustness toward outliers. Third, we theoretically and empirically showed that increasing the hyper-parameter λ does not cause a big drop in the clustering accuracy. Finally, we derive another equivalent form of its closed-form solution, which can be computed with $O(N^2)$ time complexity, which is less than the time complexity of the original form, *i.e.*, $O(N^3)$.

References

[1] Pankaj K Agarwal and Nabil H Mustafa. K-means projective clustering. In *Proceedings of the twenty-third ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 155–165. ACM, 2004.

[2] Rakesh Agrawal, Johannes Gehrke, Dimitrios Gunopulos, and Prabhakar Raghavan. *Automatic subspace clustering of high dimensional data for data mining applications*, volume 27. ACM, 1998.

[3] Laurent Bako. Identification of switched linear systems via sparse optimization. *Automatica*, 47(4):668–677, 2011.

[4] Terrance E Boult and L Gottesfeld Brown. Factorization-based segmentation of motions. In *Visual Motion, 1991., Proceedings of the IEEE Workshop on*, pages 179–186. IEEE, 1991.

[5] Paul S Bradley and Olvi L Mangasarian. K-plane clustering. *Journal of Global Optimization*, 16(1):23–32, 2000.

[6] Guangliang Chen and Gilad Lerman. Spectral curvature clustering (scc). *International Journal of Computer Vision*, 81(3):317–330, 2009.

[7] Yudong Chen, Ali Jalali, Sujay Sanghavi, and Huan Xu. Clustering partially observed graphs via convex optimization. *Journal of Machine Learning Research*, 15(1):2213–2238, 2014.

[8] João Paulo Costeira and Takeo Kanade. A multibody factorization method for independently moving objects. *International Journal of Computer Vision*, 29(3):159–179, 1998.

[9] Ehsan Elhamifar and René Vidal. Sparse subspace clustering. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 2790–2797. IEEE, 2009.

[10] Ehsan Elhamifar and Rene Vidal. Sparse subspace clustering: Algorithm, theory, and applications. *IEEE transactions on pattern analysis and machine intelligence*, 35(11): 2765–2781, 2013.

[11] Brian Eriksson, Laura Balzano, and Robert Nowak. High-rank matrix completion. In *Artificial Intelligence and Statistics*, pages 373–381, 2012.

[12] Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. In *Readings in computer vision*, pages 726–740. Elsevier, 1987.

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

[14] Han Hu, Zhouchen Lin, Jianjiang Feng, and Jie Zhou. Smooth representation clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3834–3841, 2014.

[15] Pan Ji, Mathieu Salzmann, and Hongdong Li. Shape interaction matrix revisited and robustified: Efficient subspace clustering with corrupted and incomplete data. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4687–4695, 2015.

[16] Ken-ichi Kanatani. Motion segmentation by subspace separation and model selection. In *Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference on*, volume 2, pages 586–591. IEEE, 2001.

[17] Hans-Peter Kriegel, Peer Kröger, and Arthur Zimek. Clustering high-dimensional data: A survey on subspace clustering, pattern-based clustering, and correlation clustering. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 3(1):1, 2009.

[18] Hanjiang Lai, Yan Pan, Canyi Lu, Yong Tang, and Shuicheng Yan. Efficient k-support matrix pursuit. In *European Conference on Computer Vision*, pages 617–631. Springer, 2014.

[19] Kuang-Chih Lee, Jeffrey Ho, and David J Kriegman. Acquiring linear subspaces for face recognition under variable lighting. *IEEE Transactions on pattern analysis and machine intelligence*, 27(5):684–698, 2005.

[20] Guangcan Liu, Zhouchen Lin, and Yong Yu. Robust subspace segmentation by low-rank representation. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 663–670, 2010.

[21] Can-Yi Lu, Hai Min, Zhong-Qiu Zhao, Lin Zhu, De-Shuang Huang, and Shuicheng Yan. Robust and efficient subspace segmentation via least squares regression. In *European conference on computer vision*, pages 347–360. Springer, 2012.

[22] Canyi Lu, Jiashi Feng, Zhouchen Lin, and Shuicheng Yan. Correlation adaptive subspace segmentation by trace lasso. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, pages 1345–1352. IEEE, 2013.

[23] Yi Ma, Harm Derksen, Wei Hong, and John Wright. Segmentation of multivariate mixed data via lossy data coding and compression. *IEEE transactions on pattern analysis and machine intelligence*, 29(9), 2007.

[24] Brian McWilliams and Giovanni Montana. Subspace clustering of high-dimensional data: a predictive approach. *Data Mining and Knowledge Discovery*, 28(3):736–772, 2014.

[25] Quanyi Mo and Bruce A Draper. Semi-nonnegative matrix factorization for motion segmentation with missing data. In *European Conference on Computer Vision*, pages 402–415. Springer, 2012.

[26] JinHyeong Park, Hongyuan Zha, and Rangachar Kasturi. Spectral clustering for robust motion segmentation. In *European Conference on Computer Vision*, pages 390–401. Springer, 2004.

[27] Lance Parsons, Ehtesham Haque, and Huan Liu. Subspace clustering for high dimensional data: a review. *Acm Sigkdd Explorations Newsletter*, 6(1):90–105, 2004.

[28] Kaare Brandt Petersen, Michael Syskind Pedersen, et al. The matrix cookbook. *Technical University of Denmark*, 7(15):510, 2008.

[29] Shankar Rao, Roberto Tron, Rene Vidal, and Yi Ma. Motion segmentation in the presence of outlying, incomplete, or corrupted trajectories. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(10):1832–1845, 2010.

[30] Yasuyuki Sugaya and Kenichi Kanatani. Geometric structure of degeneracy for multibody motion segmentation. In *International Workshop on Statistical Methods in Video Processing*, pages 13–25. Springer, 2004.

[31] Michael E Tipping and Christopher M Bishop. Mixtures of probabilistic principal component analyzers. *Neural computation*, 11(2):443–482, 1999.

[32] Roberto Tron and René Vidal. A benchmark for the comparison of 3-d motion segmentation algorithms. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pages 1–8. IEEE, 2007.

[33] Paul Tseng. Nearest q-flat to m points. *Journal of Optimization Theory and Applications*, 105(1):249–252, 2000.

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

[35] René Vidal and Paolo Favaro. Low rank subspace clustering (lrsc). *Pattern Recognition Letters*, 43:47–61, 2014.

[36] René Vidal and Richard Hartley. Motion segmentation with missing data using powerfactorization and gpca. In *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, volume 2, pages II–II. IEEE, 2004.

[37] René Vidal, Stefano Soatto, Yi Ma, and Shankar Sastry. An algebraic geometric approach to the identification of a class of linear hybrid systems. In *Decision and Control, 2003. Proceedings. 42nd IEEE Conference on*, volume 1, pages 167–172. IEEE, 2003.

[38] Rene Vidal, Yi Ma, and Shankar Sastry. Generalized principal component analysis (gpca). *IEEE transactions on pattern analysis and machine intelligence*, 27(12):1945–1959, 2005.

[39] Shusen Wang, Xiaotong Yuan, Tiansheng Yao, Shuicheng Yan, and Jialie Shen. Efficient subspace segmentation via quadratic programming. In *AAAI*, volume 1, pages 519–524, 2011.

[40] Jingyu Yan and Marc Pollefeys. A general framework for motion segmentation: Independent, articulated, rigid, non-rigid, degenerate and non-degenerate. In *European conference on computer vision*, pages 94–106. Springer, 2006.

[41] Allen Y Yang, Shankar R Rao, and Yi Ma. Robust statistical estimation and segmentation of multiple subspaces. In *Computer Vision and Pattern Recognition Workshop, 2006. CVPRW'06. Conference on*, pages 99–99. IEEE, 2006.

[42] Teng Zhang, Arthur Szlam, and Gilad Lerman. Median k-flats for hybrid linear modeling with many outliers. In *Computer Vision Workshops (ICCV Workshops), 2009 IEEE 12th International Conference on*, pages 234–241. IEEE, 2009.