A Simple Approach to Automated Spectral Clustering

Anonymous Author(s)
Affiliation
Address
email

Abstract

The performance of spectral clustering heavily relies on the quality of affinity matrix. A variety of affinity-matrix-construction (AMC) methods have been proposed but they have hyperparameters to determine beforehand, which requires strong experience and lead to difficulty in real applications especially when the inter-cluster similarity is high and/or the dataset is large. In addition, we often need to choose different AMC methods for different datasets, which still depends on experience. To solve these two challenging problems, in this paper, we present a simple yet effective method for automated spectral clustering. First, we propose to find the most reliable affinity matrix via grid search or Bayesian optimization among a set of candidates given by different AMC methods with different hyperparameters, where the reliability is quantified by the *relative-eigen-gap* of graph Laplacian introduced in this paper. Second, we propose a fast and accurate AMC method based on least squares representation and thresholding and prove its effectiveness theoretically. Finally, we extend the method to large-scale datasets such as MNIST, on which the time cost is less than 90s and the clustering accuracy is state-of-the-art. Extensive experiments of natural image clustering show that our method is more versatile, accurate, and efficient than baseline methods.

1 Introduction

Clustering is an important approach to data mining and knowledge discovery. Particularly, spectral clustering [Weiss, 1999; Shi and Malik, 2000; Ng *et al.*, 2002; Von Luxburg, 2007] has superior performance than k-means clustering [Steinhaus and others, 1956], hierarchical clustering [Johnson, 1967], DBSCAN [Ester *et al.*, 1996], and mixtures of probabilistic principal component analyzers [Tipping and Bishop, 1999] in many applications. Roughly speaking, spectral clustering consists of two steps: 1) construct an affinity matrix in which each element denotes the similarity between two data points; 2) perform normalized cut [Shi and Malik, 2000] on the graph corresponding to the affinity matrix. K-nearest neighbors (K-NN) and Gaussian kernel $k(\boldsymbol{x}, \boldsymbol{y}) = \exp(-\|\boldsymbol{x}-\boldsymbol{y}\|^2/(2\varsigma^2))$ are two popular methods to construct affinity matrices, where k and ς are hyperparameters.

As the performance of spectral clustering heavily relies on the quality of affinity matrix, in recent years, a variety of methods have been proposed to construct or learn affinity matrices for spectral clustering. Many of them are in the framework of self-expressive [Roweis and Saul, 2000; Elhamifar and Vidal, 2013] model, i.e., $\text{minimize}_{\boldsymbol{C}} \frac{1}{2}\|\boldsymbol{X} - \boldsymbol{X}\boldsymbol{C}\|_F^2 + \lambda\mathcal{R}(\boldsymbol{C})$. Here the columns of $\boldsymbol{X} \in \mathbb{R}^{m\times n}$ are the data points drawn from a union of subspaces. $\boldsymbol{C} \in \mathbb{R}^{n\times n}$ is a coefficient matrix. $\mathcal{R}(\boldsymbol{C})$ denotes a regularization operator on $\boldsymbol{C}$. λ is a hyperparameter to be determined in advance. Elhamifar and Vidal [2013] proposed to use $\mathcal{R}(\boldsymbol{C}) = \|\boldsymbol{C}\|_1 := \sum_{i=1}^n \sum_{j=1}^n |c_{ij}|$ under a constraint $\text{diag}(\boldsymbol{C}) = \boldsymbol{0}$. In [Elhamifar and Vidal, 2013], the affinity matrix for spectral clustering is given by $\boldsymbol{A} = |\boldsymbol{C}| + |\boldsymbol{C}|^\top$. The method is called Sparse Subspace Clustering (SSC). Some theoretical results of SSC can be found in [Wang and Xu, 2013; Soltanolkotabi *et al.*, 2014].

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

Following the self-expressive framework, Liu *et al.* [2013] let $\mathcal{R}(\boldsymbol{C}) = \|\boldsymbol{C}\|_*$ (nuclear norm of $\boldsymbol{C}$) and proposed a Low-Rank Representation (LRR) method for subspace clustering. Lu *et al.* [2012] and Pan Ji *et al.* [2014] used the least squares representation (LSR) model for subspace clustering. A few variants of LRR and SSC can be found in [Patel and Vidal, 2014; Li and Vidal, 2015; Patel *et al.*, 2015; Shen and Li, 2016; Li and Vidal, 2016; Lu *et al.*, 2018]. Recently, deep learning methods were also used to learn affinity matrices for spectral clustering [Ji *et al.*, 2017; Zhang *et al.*, 2019b,a] and have achieved state-of-the-art performance on many benchmark datasets.

One common limitation of these spectral or subspace clustering methods is that they have at least one hyperparameter to determine in advance. In the codes of SSC[1] and its variants provided by their authors, there is usually one more thresholding parameter for affinity matrix, which affects the clustering accuracy a lot. In the deep learning based methods, we need to determine the network structures and regularization parameters, which is much more difficult. Since clustering is an unsupervised learning problem, the hyperparameters cannot be tuned by cross validation that is widely used in supervised learning. Thus, we have to tune the hyperparameters in spectral clustering by experience, which is difficult when the given dataset is quite different from those in our experience and/or the inter-class similarity is high compared to the intra-cluster similarity. Note that SSC, LRR, and their kernel or deep learning extensions have quadratic or even cubic time complexity (per iteration of optimization), which further increases the difficulty of hyperparameter selection in clustering large datasets, though there have been a few work improving the computational efficiency [Peng *et al.*, 2013; Cai and Chen, 2014; Wang *et al.*, 2014; Peng *et al.*, 2015; You *et al.*, 2016a,b; Li and Zhao, 2017; You *et al.*, 2018; Matsushima and Brbic, 2019; Li *et al.*, 2020; Chen *et al.*, 2020; Kang *et al.*, 2021]. On the other hand, different datasets often require different methods for AMC, which is hard to tackle by experience.

This paper aims at model and hyperparameter selection for spectral clustering and wants to improve the convenience, accuracy, and efficiency of spectral clustering. Our contributions are as follows.

- We propose a *relative-eigen-gap* based automated spectral clustering (AutoSC) method. It finds the Laplacian matrix with largest *relative-eigen-gap* among a set of candidates constructed by different models with different hyperparameters.
- We also implement the AutoSC method via Bayesian optimization. The method can select the possibly best model and optimize the hyperparameters automatically. Note that any AMC methods (e.g. SSC) can be included into the framework of AutoSC.
- To improve the accuracy and efficiency of AutoSC, we propose a new AMC method based on least squares representation and thresholding and prove its effectiveness theoretically.
- We provide an extension for AutoSC to cluster large-scale datasets.

Experiments on seven benchmark image datasets demonstrate the effectiveness of our method. Particularly, our method is able to outperform state-of-the-art methods of large-scale clustering.

2 Related work

2.1 Exploiting eigenvalue information for clustering

As the number of zero eigenvalues of a Laplacian matrix is equal to the number of connected components of the graph [Von Luxburg, 2007], a few researchers took advantage of eigenvalue information in spectral clustering [Meila *et al.*, 2005; Meila and Shortreed, 2006; Ji *et al.*, 2015; Hu *et al.*, 2017; Lu *et al.*, 2018]. For instance, Ji *et al.* [2015] utilized eigen-gap to determine the rank of the Shape Interaction Matrix. But the method requires to determine another hyperparameter γ beforehand and needs to perform spectral clustering for multiple times. The methods of [Meila *et al.*, 2005; Hu *et al.*, 2017; Lu *et al.*, 2018] are based on iterative optimization (need to perform eigenvalue decomposition at every iteration) and hence are not effective in handling large-scale datasets. In addition, the BDR method of [Lu *et al.*, 2018] has two hyperparameters (λ, γ) to determine by experience, although it outperformed SSC and LRR on some datasets. A comparison is shown in Figure 1.

[1] Wang and Xu [2013] and Soltanolkotabi *et al.* [2014] provided lower and upper bounds for the λ in SSC theoretically, which however depend on the unknown noise level.

2.2 Automated machine learning

Automated model and hyperparameter selection for supervised learning has been extensively studied Hutter *et al.* [2019]. In contrast, the study for unsupervised learning is very limited. The reason is that in unsupervised learning there is no ground truth or reliable metric to evaluate the performance of algorithms. Concurrently to our work, Poulakis [2020] also attempted to do automated clustering. Specifically, Poulakis [2020] proposed to use meta-learning to select clustering algorithm and use a heuristic combination of some clustering validity metrics such as Silhouette coefficient Liu *et al.* [2010] and S_Dbw [Halkidi and Vazirgiannis, 2001] as an objective to maximize via grid search or Bayesian optimization [Jones *et al.*, 1998]. One problem is that these metrics are mainly based on Euclidean distance or densities and hence may not be suitable to evaluate the clustering performance of non-distance or non-density based clustering algorithms. Another problem is that there is no unified metric to compare different clustering algorithms.

3 Automated Spectral Clustering (AutoSC)

3.1 Preliminary Knowledge

Let $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ be an affinity matrix constructed from a given a data matrix $\boldsymbol{X} \in \mathbb{R}^{m\times n}$. The corresponding graph is denoted by $G = (V, E)$, where $V = \{v_1, \ldots, v_n\}$ is the vertex set and $E = \{e_1, \ldots, e_l\}$ is the edge set. The degree matrix of a graph G is defined as $\boldsymbol{D} = \text{diag}(\boldsymbol{A}\mathbf{1})$, where $\mathbf{1} = [1, \ldots, 1]^\top$. Our goal is to partition the vertices into k disjoint nonempty subsets $C_1, \ldots, C_k$. Let $\mathcal{C} = \{C_1, \ldots, C_k\}$. It is expected to find a partition $\mathcal{C}$ that minimizes the following metric:

Definition 3.1 (MNCut)**.** The multiway normalized cut (MNCut) [Meila, 2001] is defined as

$$\text{MNCut}(\mathcal{C}) = \sum_{i=1}^{k} \sum_{j\neq i} \frac{Cut(C_i, C_j)}{Vol(C_i)}, \tag{1}$$

where $Cut(C_i, C_j) = \sum_{u\in C_i} \sum_{v\in C_j} A_{uv}$ and $Vol(C_i)$ denotes the sum of vertex degrees of C_i.

The normalized graph Laplacian matrix is defined as

$$\boldsymbol{L} = \boldsymbol{I} - \boldsymbol{D}^{-1/2}\boldsymbol{A}\boldsymbol{D}^{-1/2}, \tag{2}$$

where $\boldsymbol{I}$ is an identity matrix. The normalized graph Laplacian is often more effective than unnormalized one in spectral clustering (some theoretical justification was given by Von Luxburg Von Luxburg [2007]). Let $\sigma_i(\boldsymbol{L})$ be the i-th smallest eigenvalue of $\boldsymbol{L}$. The following claim shows the connection between MNCut($\mathcal{C}$) and $\boldsymbol{L}$.

Claim 3.2. *The sum of the k smallest singular values of $\boldsymbol{L}$ quantifies the potential connectivity among $C_1, \ldots, C_k$:*

$$\text{MNCut}(\mathcal{C}) \geq \textstyle\sum_{i=1}^{k} \sigma_i(\boldsymbol{L}).$$

The claim can be easily proved by using Lemma 4 of Meila [2001]. We defer all proof of this paper to the appendices. Because the multiplicity k of the eigenvalue 0 of $\boldsymbol{L}$ equals to the number of connected components in G Von Luxburg [2007], we expect to construct an affinity matrix $\boldsymbol{A}$ from $\boldsymbol{X}$ such that $\boldsymbol{L}$ has k zero eigenvalues. Thus the optimal partition means MNCut($\mathcal{C}$) $= \sum_{i=1}^{k} \sigma_i(\boldsymbol{L}) = 0$.

3.2 Relative Eigen-Gap Guided Search

In practice, we may construct an $\boldsymbol{A}$ such that $\sum_{i=1}^{k} \sigma_i(\boldsymbol{L})$ is as small as possible because guaranteeing zero eigenvalues is difficult. But this is not enough because $\boldsymbol{L}$ may have $k + 1$ or more very small or even zero eigenvalues. The second smallest eigenvalue of the Laplacian matrix of a graph G is called the algebraic connectivity of G (denoted by $ac(G)$) Fiedler [1973]. We have $ac(G) = 0$ if and only if G is not connected. When G has k disjointed components, there are k algebraic connectivities, denoted by $ac(C_1), \ldots, ac(C_k)$. Based on this, we have

Claim 3.3. *The k+1th smallest eigenvalue of $\boldsymbol{L}$ quantifies the least potential connectivity of partitions $C_1, \ldots, C_k$ of $\mathcal{C}$:*

$$\min_{1\leq i\leq k} \text{MNCut}(C_i) \geq \sigma_{k+1}(\boldsymbol{L}). \tag{3}$$

In other words, $\sigma_{k+1}(\boldsymbol{L})$ measures the difficulty in segmenting each of C_i into two subsets. Hence, when $\sigma_{k+1}(\boldsymbol{L})$ is large, the partitions $C_1, \ldots, C_k$ are stable. Based on Claim 3.2 and Claim 3.3, we may construct an $\boldsymbol{A}$ that has small $\sum_{i=1}^{k} \sigma_i(\boldsymbol{L})$ and large $\sigma_{k+1}(\boldsymbol{L})$ simultaneously, by solving

$$\begin{aligned} &\underset{\theta}{\text{maximize}} \;\; \sigma_{k+1}(\boldsymbol{L}) - \frac{1}{k}\sum_{i=1}^{k}\sigma_i(\boldsymbol{L}), \\ &\text{subject to} \;\; \boldsymbol{L} = \boldsymbol{I} - \boldsymbol{D}^{-1/2}\boldsymbol{A}\boldsymbol{D}^{-1/2}, \; \boldsymbol{A} = f_\theta(\boldsymbol{X}), \end{aligned} \tag{4}$$

where f_θ denotes a function with parameter θ. It is difficult to solve (4) because of the composition of f_θ, symmetric normalized Laplacian, and eigenvalue decomposition. On the other hand, in (4), we have to choose f in advance, which requires domain expert or strong experience because different dataset usually needs different f.

Note that different f can result in very different distribution of eigenvalues and the small eigenvalues are sensitive to f, θ, and noise. Hence the objective in (4) is not effective to compared different f and θ. In this paper, we define a new metric *relative-eigen-gap* as follows

$$\text{reg}(\boldsymbol{L}) := \frac{\sigma_{k+1}(\boldsymbol{L}) - \frac{1}{k}\sum_{i=1}^{k}\sigma_i(\boldsymbol{L})}{\frac{1}{k}\sum_{i=1}^{k}\sigma_i(\boldsymbol{L}) + \varepsilon}, \tag{5}$$

where ε is a small constant (e.g. 10^{-6}) to avoid zero denominator. $\text{reg}(\boldsymbol{L})$ is not sensitive to the scale of the small eigenvalues. Therefore, instead of (4), we propose to solve

$$\begin{aligned} &\underset{(f,\theta)\in\mathcal{F}\times\Theta}{\text{maximize}} \;\; \text{reg}(\boldsymbol{L}), \\ &\text{subject to} \;\; \boldsymbol{L} = \boldsymbol{I} - \boldsymbol{D}^{-1/2}\boldsymbol{A}\boldsymbol{D}^{-1/2}, \; \boldsymbol{A} = f_\theta(\boldsymbol{X}), \end{aligned} \tag{6}$$

where $\mathcal{F}$ is a set of pre-defined functions and Θ is a set of hyperparameters. In fact, (6) is equivalent to choosing one $\boldsymbol{A}$ (or $\boldsymbol{L}$) from a set of candidates constructed by different f with different θ, of which the *relative-eigen-gap* is largest. The best θ can be found using grid search (or even random search). For convenience, we call the method AutoSC-GD. Table 1 shows a few examples of f and its parameters. One may use a weighted sum of affinity matrices given by different f, like Huang *et al.* [2012], which however will introduce more hyperparameters.

Table 1: A few examples of f and its θ for AMC (AASC: Huang *et al.* [2012])

f	K-NN	ϵ-neighborhood	Gaussian kernel	SSC	LRR	LSR	KSSC	AASC
θ	K	ϵ	σ	λ	λ	λ	λ, σ	$\sigma_1, \sigma_2, \ldots$

The following theorem[2] shows the connection between $\text{reg}(\boldsymbol{L})$ and the stability of the clustering $\mathcal{C}$.

Theorem 3.4. *Let $\mathcal{C}$ and $\mathcal{C}'$ be two partitions of the vertices of G, where $|\mathcal{C}| = |\mathcal{C}'| = k$. Define the distance between $\mathcal{C}$ and $\mathcal{C}'$ as*

$$\text{dist}(\mathcal{C}, \mathcal{C}') = 1 - \frac{1}{k}\sum_{C_i\in\mathcal{C}}\sum_{C'_j\in\mathcal{C}'}\frac{(Vol(C_i\cap C'_j))^2}{Vol(C_i)Vol(C'_j)}. \tag{7}$$

Suppose $\eta k\varepsilon \geq \sum_{i=1}^{k}\sigma_i(\boldsymbol{L}) \geq k\varepsilon$ and $\text{reg}(\boldsymbol{L}) > (k-1)\eta/2$. Let $\delta = \max\big(\text{MNCut}(\mathcal{C}) - \sum_{i=1}^{k}\sigma_i(\boldsymbol{L}), \text{MNCut}(\mathcal{C}') - \sum_{i=1}^{k}\sigma_i(\boldsymbol{L})\big)$. Then

$$\text{dist}(\mathcal{C}, \mathcal{C}') < \frac{1.5\delta\varepsilon^{-1}}{\text{reg}(\boldsymbol{L}) + (1-k)\eta/2}. \tag{8}$$

It indicates that when $\text{reg}(\boldsymbol{L})$ is large and δ is small, the partitions $\mathcal{C}$ and $\mathcal{C}'$ are close to each other. Thus, the clustering has high stability. When $\sum_{i=1}^{k}\sigma_i(\boldsymbol{L}) = k\varepsilon$, we have

$$\text{dist}(\mathcal{C}, \mathcal{C}') < \frac{6\delta}{\sigma_{k+1}(\boldsymbol{L}) - k\varepsilon},$$

[2]This theorem is a modified version of Theorem 1 in Meila *et al.* [2005], which is for the eigen-gap $\sigma_{k+1} - \sigma_k$ of $\boldsymbol{L}$. Here we consider $\text{reg}(\boldsymbol{L})$ instead.

which means the larger $\sigma_{k+1}(\boldsymbol{L})$ the more stable clustering.

Compare reg$(\boldsymbol{L})$ with [Meila and Shortreed, 2006] It is worth noting that Meila and Shortreed [2006] proposed to minimize $\mathrm{J}(\boldsymbol{L}) := \frac{1}{k}\sum_{i=1}^{k}\sigma_i(\boldsymbol{L}) + \alpha\left(\sigma_k(\boldsymbol{L}) - \sigma_{k+1}(\boldsymbol{L})\right)^2$ to find a good affinity matrix for spectral clustering. Although reg$(\boldsymbol{L})$ and $\mathrm{J}(\boldsymbol{L})$ seem similar, they are essentially different. As mentioned before, different AMC methods may lead to different scales for the small eigenvalues, which makes it difficult to compare different AMC methods using $\mathrm{J}(\boldsymbol{L})$. In addition, $\mathrm{J}(\boldsymbol{L})$ has a hyperparameter α to determine beforehand, which violates our goal of searching models and hyperparameters. A comparative study ($\alpha = 0$) is in Appendix E.5 of supplementary material.

3.3 AutoSC via Bayesian Optimization

Bayesian optimization (BO) Jones *et al.* [1998] has become a promising tool for hyperparameter optimization of supervised machine learning algorithms [Snoek *et al.*, 2012; Klein *et al.*, 2017]. Given a black-box function $g : \mathcal{X} \to \mathbb{R}$, BO aims to find an $\boldsymbol{x}^* \in \mathcal{X}$ that globally minimizes g and usually has three steps. The first step is finding the most promising point $\boldsymbol{x}_{t+1} \in \mathrm{argmax}_{\boldsymbol{x}}\, a_{p(g)}(\boldsymbol{x})$ by numerical optimization, where $a_{p(g)} : \mathcal{X} \to \mathbb{R}$ is an acquisition function (e.g. Expected Improvement) relying on an prior $p(g)$ (e.g. Gaussian processes [Williams and Rasmussen, 2006]). The second step is evaluating the expensive and possibly noisy function $y_{t+1} \sim g(\boldsymbol{x}) + \mathcal{N}(0, \sigma^2)$ and add the new sample $(\boldsymbol{x}_{t+1}, y_{t+1})$ to observation set $\mathcal{D}_t = \{(\boldsymbol{x}_{t+1}, y_{t+1}), \ldots, (\boldsymbol{x}_t, y_t)\}$. The last step is updating $p(g)$ and $a_{p(g)}$ using $\mathcal{D}_{t+1}$.

As an alternative to the grid search for (6), we can maximize reg$(\boldsymbol{L}_s)$ via BO. Suppose we have a set of different AMC models, i.e., $\mathcal{F} = \{f_1, f_2, \ldots, f_M\}$. For $i = 1, 2, \ldots, M$, let

$$g_i(\theta^{(i)}) := -\mathrm{reg}(\boldsymbol{L}(f_i(\theta^{(i)}|\mathbb{X}))),$$

where $\theta^{(i)}$ denotes the hyperparameters in f_i and $\mathbb{X}$ denotes the dataset. Then we use BO to find

$$\theta_*^{(i)} = \underset{\theta^{(i)} \in S^{(i)}}{\mathrm{argmin}}\; g_i(\theta^{(i)}), \tag{9}$$

where $S^{(i)}$ denotes the constraints on f_i's hyperparameters. Now we get the best model with its best hyperparameters

$$f_\star(\theta_*^{(\star)}|\mathbb{X}), \quad \text{where } \star = \underset{1 \le i \le M}{\mathrm{argmin}}\, g_i(\theta_*^{(i)}). \tag{10}$$

For convenience, we denote the method by AutoSC-BO. More details about the implementation of AutoSC-BO are in Appendix D. Note that $\mathcal{F}$ can include any AMC methods such as those in Table 1 and even DSC [Ji *et al.*, 2017] (see Appendix E.6).

4 AMC using Least Squares Representation with Thresholding

4.1 Motivation and Method

In AutoSC, the size of searching space is $|\mathcal{F}| \times \prod_j |\theta_j|$, which should be large enough to include effective models and their hyperparameters. A large number of works have shown that the self-expressive models [Elhamifar and Vidal, 2013; Liu *et al.*, 2013; Lu *et al.*, 2018; Ji *et al.*, 2017] often outperform other AMC models such as Gaussian kernel. However, the self-expressive model based AMC methods often require iterative optimization and has at least quadratic time complexity per iteration, which leads to huge time cost in AutoSC. Although LSR [Lu *et al.*, 2012] has closed-form solution, the clustering accuracy is not satisfactory [Lu *et al.*, 2018]. In this work, we show that LSR with a simple post-processing can be a good AMC method and is able to outperform SSC, LRR, and BDR [Lu *et al.*, 2018].

Specifically, the LSR model is given as

$$\underset{\boldsymbol{C}}{\mathrm{minimize}} \;\; \tfrac{1}{2}\|\boldsymbol{X} - \boldsymbol{X}\boldsymbol{C}\|_F^2 + \tfrac{\lambda}{2}\|\boldsymbol{C}\|_F^2, \tag{11}$$

of which the closed-form solution is

$$\boldsymbol{C} = (\boldsymbol{X}^\top\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{X}^\top\boldsymbol{X}. \tag{12}$$

Let $\mathrm{diag}(\boldsymbol{C}) = \boldsymbol{0}$ and $\boldsymbol{C} \leftarrow |\boldsymbol{C}|$, the affinity matrix can be constructed as $\boldsymbol{A} = (\boldsymbol{C} + \boldsymbol{C}^\top)/2$. One problem is that the off-diagonal elements of $\boldsymbol{A}$ are dense (leading to a connected graph), which can result in low clustering accuracy. Therefore, we propose to truncate $\boldsymbol{C}$ by keeping only the largest τ elements of each column of $\boldsymbol{C}$. Nevertheless, it is not easy to determine τ beforehand. When τ is too small, the corresponding graph will have $k+1$ or more connected components. When τ is too large, the corresponding graph will have $k-1$ or less connected components. However, τ can be automatically determined by our AutoSC-GD and AutoSC-BO.

In the case that the data have some low-dimensional nonlinear structures, the similarity between pair-wise columns of $\boldsymbol{X}$ cannot be well recognized by the linear regression (11). Therefore, we also consider the following nonlinear regression model

$$\underset{\boldsymbol{C}}{\text{minimize}} \quad \tfrac{1}{2}\|\phi(\boldsymbol{X}) - \phi(\boldsymbol{X})\boldsymbol{C}\|_F^2 + \tfrac{\lambda}{2}\|\boldsymbol{C}\|_F^2, \tag{13}$$

where ϕ denotes a nonlinear feature map performed on each column of the matrix, i.e. $\phi(\boldsymbol{X}) = [\phi(\boldsymbol{x}_1), \ldots, \phi(\boldsymbol{x}_n)]$. In (13), letting ϕ be some feature map induced by a kernel function $k(\cdot,\cdot)$, we get the kernel least square representation (KLSR):

$$\underset{\boldsymbol{C}}{\text{minimize}} \quad \tfrac{1}{2}\mathrm{Tr}\left(\boldsymbol{K} - 2\boldsymbol{K}\boldsymbol{C} + \boldsymbol{C}^\top\boldsymbol{K}\boldsymbol{C}\right) + \tfrac{\lambda}{2}\|\boldsymbol{C}\|_F^2, \tag{14}$$

where $\boldsymbol{K} = \phi(\boldsymbol{X})^\top\phi(\boldsymbol{X})$ and $[\boldsymbol{K}]_{ij} = k(\boldsymbol{x}_i, \boldsymbol{x}_j)$. The closed-form solution is

$$\boldsymbol{C} = (\boldsymbol{K} + \lambda\boldsymbol{I})^{-1}\boldsymbol{K}. \tag{15}$$

The post-processing is the same as that for (12). Polynomial kernel $k(\boldsymbol{x}_i, \boldsymbol{x}_j) = (\boldsymbol{x}_i^\top\boldsymbol{x}_j + b)^q$ and Gaussian kernel are two popular kernel functions, though there are many other choices.

Note that if $n \ll m$, using the *push-through identity* [Henderson and Searle, 1981], we reformulate (12) as $\boldsymbol{C} = \boldsymbol{X}^\top(\lambda\boldsymbol{I} + \boldsymbol{X}\boldsymbol{X}^\top)^{-1}\boldsymbol{X}$ to reduce the computational cost from $O(n^3)$ to $O(mn^2)$. In (15), when n is large (e.g.>5000), we perform randomized SVD [Halko *et al.*, 2011] on $\boldsymbol{K}$: $\boldsymbol{K} \approx \boldsymbol{V}_r\boldsymbol{\Sigma}_r\boldsymbol{V}_r^\top$. Then $\boldsymbol{C} \approx \boldsymbol{V}_r\boldsymbol{\Sigma}_r^{1/2}(\lambda\boldsymbol{I} + \boldsymbol{\Sigma})^{-1}\boldsymbol{\Sigma}_r^{1/2}\boldsymbol{V}_r^\top$, where $r = 20k$ works well in practical applications. The time complexity of computing $\boldsymbol{C}$ is $O(r\tau n + rn^2)$. The computation of the smallest $k+1$ eigenvalues of $\boldsymbol{L}$ is equivalent to compute the largest $k+1$ eigenvalues and eigenvectors of $\boldsymbol{D}^{-1/2}\boldsymbol{A}\boldsymbol{D}^{-1/2}$, which is sparse. The time complexity is $O(k\tau n)$. We have the follows.

Proposition 4.1. *Let $\hat{\boldsymbol{c}}$ be the optimal solution of* $\mathrm{minimize}_{\boldsymbol{c}}\, \frac{1}{2}\|\phi(\boldsymbol{y}) - \phi(\boldsymbol{X})\boldsymbol{c}\|^2 + \frac{\lambda}{2}\|\boldsymbol{c}\|^2$, *where ϕ is induced by Gaussian kernel and $\boldsymbol{y}$ is arbitrary. Then* $\|\hat{c}_i - \hat{c}_j\| \leq \sqrt{2 - 2\exp\left(-\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2/(2\varsigma^2)\right)}$.

It shows that when two data points in $\boldsymbol{X}$, e.g. $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$, are close to each other, the corresponding two elements in $\hat{\boldsymbol{c}}$, e.g. $\hat{c}_i$ and $\hat{c}_j$, have small difference. Hence (15) with Gaussian kernel utilizes local information to enhance $\boldsymbol{C}$.

4.2 Theoretical Guarantee

We introduce the following property (a necessary condition of successful subspace clustering) which is similar to the one used in [Wang and Xu, 2013; Soltanolkotabi *et al.*, 2014].

Definition 4.2 (Subspace Detection Property)**.** A symmetric affinity matrix $\boldsymbol{A}$ obtained from $\boldsymbol{X}$ has subspace detection property if for all i, the nonzero elements of $\boldsymbol{a}_i$ correspond to the columns of $\boldsymbol{X}$ in the same subspace as $\boldsymbol{x}_i$.

For convenience, let $\pi(i)$ be the index of the subspace $\boldsymbol{x}_i$ belongs to and C_j be the index set of the columns of $\boldsymbol{X}$ in subspace j. We consider the following deterministic model.

Definition 4.3 (Deterministic Model)**.** The columns of $\boldsymbol{X} \in \mathbb{R}^{m\times n}$ are drawn from a union of k different subspaces and are further corrupted by noise, where $\dim(\mathcal{S}_1 \cup \cdots \cup \mathcal{S}_k) = d \leq m/2 \leq n/2$. Let $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$ be the SVD of $\boldsymbol{X}$, where $\boldsymbol{\Sigma} = \mathrm{diag}(\sigma_1, \ldots, \sigma_n)$ and $\sigma_1 \geq \sigma_2 \geq \cdots \sigma_n$. Let $\gamma = \sigma_{d+1}/\sigma_d$. Denote $\boldsymbol{v}_i = (v_{i1}, \ldots, v_{in})$ the i-th row of $\boldsymbol{V}$ and let $\bar{\boldsymbol{v}}_i = (v_{i1}, \ldots, v_{id})$. Suppose the following conditions hold[3]: 1) for every $i \in [n]$, the $\bar{\tau}$-th largest element of $\{|\bar{\boldsymbol{v}}_i^\top\bar{\boldsymbol{v}}_j| : j \in C_{\pi(i)}\}$ is greater than α; 2) $\max_{i\in[n]}\max_{j\in[n]\setminus C_{\pi(i)}} |\bar{\boldsymbol{v}}_i^\top\bar{\boldsymbol{v}}_j| \leq \beta$; 3) $\max_{i,j,l}|v_{il}v_{jl}| \leq \mu$; 4) $\alpha - \beta > \frac{2\mu m\gamma^2}{1+\gamma^2}$.

[3] γ measures the noise level, β is dominated by the difference between subspaces, and μ quantifies the incoherence in the singular vectors.

Then the following theorem verifies the effectiveness of (11) followed by the truncation (thresholding) operation in subspace detection.

Theorem 4.4. *Suppose $\boldsymbol{X}$ is given by Definition 4.3 and $\boldsymbol{C}$ is given by* (11) *with*

$$\frac{\left(\Delta(1+\gamma^2)-2\mu m\gamma^2\right)\left(1-\sqrt{1-\rho^{-1}}\right)\sigma_d^2}{4\mu d-2\Delta} < \lambda < \frac{\left(\Delta(1+\gamma^2)-2\mu m\gamma^2\right)\left(1+\sqrt{1-\rho^{-1}}\right)\sigma_d^2}{4\mu d-2\Delta} \tag{16}$$

where $\rho = \frac{\left(\Delta(1+\gamma^2)-2\mu m\gamma^2\right)^2}{4(2\mu d-\Delta)(2\mu m-2\mu d-\Delta)}$ *and* $\Delta = \alpha-\beta$. *Then the $\boldsymbol{C}$ truncated by $\tau \leq \bar{\tau}$ has the subspace detection property.*

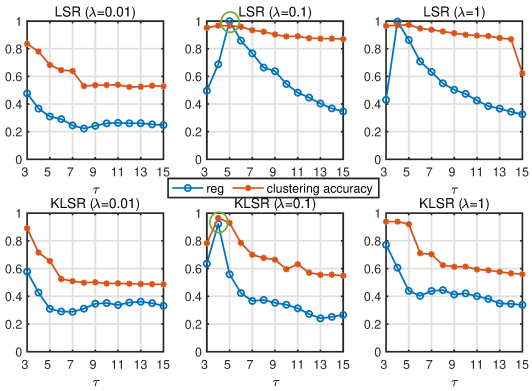
In Theorem 4.4, the width of the range of λ is $w = \frac{\left(\Delta(1+\gamma^2)-2\mu m\gamma^2\right)\sqrt{1-\rho^{-1}}\sigma_d^2}{2\mu d-\Delta}$. We see that larger σ_d, Δ, or smaller γ, d lead to wider range of λ, which corresponds to simper clustering problem. When $\rho < 1$ (too small Δ and too large γ), λ does not exist. Theorem 4.4 can be extended to the kernel case (14) without the restriction of $2d \leq m$ even when the columns of $\boldsymbol{X}$ are drawn from a union of nonlinear low-dimensional manifolds. See Definition C.1, Definition C.2, and Theorem C.3 in Appendix C of the supplementary material. Based on Theorem 4.4 and Theorem C.3, the follow proposition indicates that EGGS can cluster the data correctly.

Proposition 4.5. *Suppose the affinity matrix $\boldsymbol{A}$ given by AutoSC has the subspace or manifold detection property (defined in Appendix C of the supplementary material) and* $\text{reg}(\boldsymbol{L}) = \frac{\sigma_{k+1}}{\epsilon} > 0$. *Then each component of G consists of all columns of $\boldsymbol{X}$ in the same subspace or manifold.*

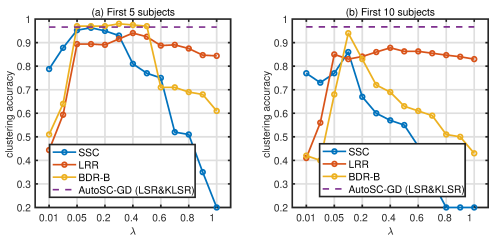
4.3 Examples of AutoSC-GD with Only LSR and KLSR

The algorithm of AutoSC-GD with Only LSR and KLSR is given by Algorithm 1 in the supplementary material. Figure 1(i) shows an intuitive example of the performance of LSR and KLSR in clustering a subset of the Extended Yale B database [Kuang-Chih *et al.*, 2005], where for (14) we use Gaussian kernel with $\varsigma = \frac{1}{n^2}\sum_{ij}\|\boldsymbol{x}_i - \boldsymbol{x}_j\|$. We see that: 1) in LSR and KLSR, for a fixed λ (or τ), the τ (or λ) with larger $\text{reg}(\boldsymbol{L})$ provides higher clustering accuracy; 2) for a fixed λ and a fixed τ, if LSR has a larger $\text{reg}(\boldsymbol{L})$, its clustering accuracy is higher than that of KLSR, and vice versa. We conclude that a larger $\text{reg}(\boldsymbol{L})$ indeed leads to a higher clustering accuracy, which is consistent with our theoretical analysis in Section 3.2.

Figure 1(ii) shows the clustering accuracy of SSC [Elhamifar and Vidal, 2013], LRR [Liu *et al.*, 2013], and BDR-B [Lu *et al.*, 2018] with different hyperparameters and our method Auto-GD with LSR and KLSR on the Extended Yale Face B subset. SSC and BDR-B are sensitive to the value of λ, especially for the relatively difficult task, say Figure 1(ii-b). LRR is not sensitive to the value of λ but its accuracy is low. LSR and KLSR are more accurate and efficient than other methods.



(i) Clustering accuracy and $\text{reg}(\boldsymbol{L})$ (rescaled between 0 and 1) and clustering accuracy of LSR and KLSR on the first 5 subjects of the Extended Yale Face B database.



(ii) Clustering accuracies of SSC, LRR, and BDR-B with different hyperparameter λ in comparison to AutoSC-GD with LSR and KLSR on the Extended Yale Face B database. The value of λ used in BDR-B has been divided by 10. The γ in BDR-B is chosen from $\{0.01, 0.1, 1\}$ and the best one is used for each λ. In Case (b), the time costs of SSC, LRR, BDR-B, and AutoSC-GD are 9.5s, 33.0s, 7.6s, and 1.6s respectively.

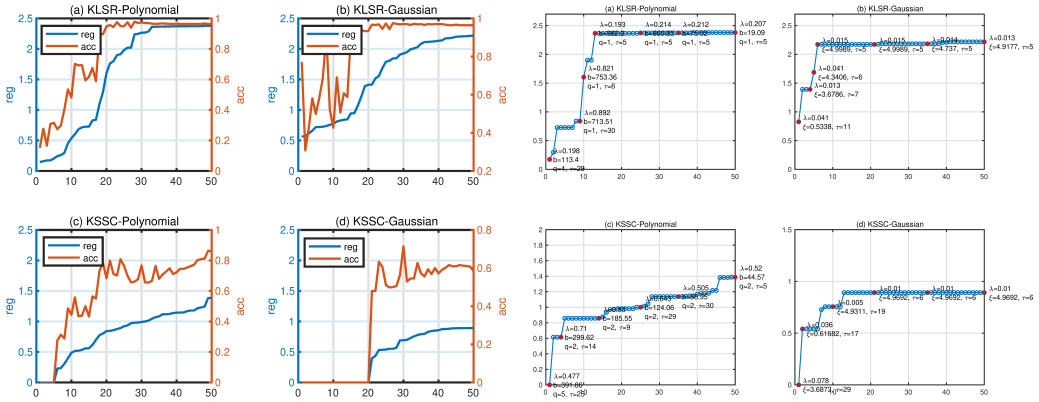
Figure 1: Examples about $\text{reg}(\boldsymbol{L})$, clustering accuracy, and hyperparameters on Extended Yale B.

4.4 Examples of AutoSC-BO with Many AMC Methods

We show the performance of AutoSC-BO with many AMC methods such as SSC and LSR. Taking the KLSR model (14) with polynomial kernel as an example, the parameters are $\theta = (\lambda, b, q, \tau)^\top$ and the constraints are given by $S = \{\lambda \in \mathbb{R} : \lambda_{min} \leq \lambda \leq \lambda_{max};\ b \in \mathbb{R} : b_{min} \leq b \leq b_{max};\ q \in \mathbb{Z}^+ : q_{min} \leq q \leq q_{max};\ \tau \in \mathbb{Z}^+ : \tau_{min} \leq \tau \leq \tau_{max}\}$. More details are in Appendix E.6 of the supplement. Shown in Table 2, larger reg corresponds to higher clustering accuracy and KLSR with polynomial kernel (the optimal q is 1) performs best. Figure 2 shows the performance of KSSC and KLSR in each iteration of AutoSC-BO.

Table 2: Clustering accuracy of AutoSC-BO with many methods on Yale Face B dataset (first 10 subjects). All hyperparameters of the kernel functions were optimized via Bayesian optimization.

AMC	ϵ-neigh -borhood	Polynomial kernel	Gaussian kernel	KSSC (Gauss)	KSSC (Poly)	KLSR (Gauss)	KLSR (Poly)
reg_{max}	0.776	1.294	1.307	0.892	1.388	2.217	2.379
Accuracy	0.325	0.389	0.393	0.584	0.859	0.963	0.966



(i) reg($\boldsymbol{L}$) and clustering accuracy of KLSR and KSSC in each iteration of BO.

(ii) reg($\boldsymbol{L}$) and hyperparameters of KLSR and KSSC in each iteration of BO.

Figure 2: AutoSC-BO with KSSC and KLSR on the first 10 subjects of YaleB Face dataset.

5 AutoSC+NSE for Large-Scale Data

See Appendix B of the supplementary material.

6 Experiments

We test our AutoSC on Extended Yale B Face [Kuang-Chih *et al.*, 2005], ORL Face [Samaria and Harter, 1994], COIL20 [Nene *et al.*, 1996], AR Face [Martínez and Kak, 2001], MNIST [LeCun *et al.*, 1998], Fashion-MNIST [Xiao *et al.*, 2017], GTSRB [Stallkamp *et al.*, 2012], subsets and extracted features of MNIST and Fashion-MNIST.

6.1 Performance on small datasets

We first compare AutoSC with SSC [Elhamifar and Vidal, 2013], LRR [Liu *et al.*, 2013], EDSC Pan Ji *et al.* [2014][4], KSSC [Patel and Vidal, 2014], SSC-OMP [You *et al.*, 2016b], BDR-Z Lu *et al.* [2018], and BDR-B [Lu *et al.*, 2018]. The clustering accuracy and time cost on four (more the results are in the supplementary material) small datasets are reported in Table 3. AutoSC-GD

[4] It's an improved version of LSR Lu *et al.* [2012]. The original LSR and its kernel version have low accuracy [Lu *et al.*, 2018] and hence are not compared in this paper.

and AutoSC-BO outperformed other methods significantly. SSC-OMP and AutoSC-GD are more efficient than other methods. The time cost of AutoSC-BO is much higher than that of AutoSC-GD because we optimize all hyperparameters.

Table 3: Clustering performance on the six small datasets. For the MNIST1k, we report the average results of 20 trials because the subset is formed randomly. AutoSC chose LSR for Yale B and AR and chose KLSR for others datasets The NMI results are in Table 1 of Appendix E.3.

		SSC	LRR	EDSC	KSSC	SSC-OMP	BDR-Z	BDR-B	AutoSC-GD	AutoSC-BO
Yale B	acc	0.723	0.643	0.806	0.649	0.768	0.596	0.719	**0.897**	**0.909**
	time	273.8	928.1	58.6	464.3	**8.9**	368.8	368.8	**19.1**	78.3
ORL	acc	0.711	0.762	0.712	0.707	0.665	0.739	0.735	**0.795**	**0.803**
	time	2.7	8.8	**2.0**	2.6	**0.4**	3.9	3.9	2.3	20.5
AR	acc	0.718	0.769	0.673	0.726	0.669	0.745	0.751	**0.786**	**0.826**
	time	317.5	1220.6	69.1	627.4	**57.6**	578.7	578.7	**43.4**	130.6
MNIST -1k	acc	0.596 (0.054)	0.513 (0.037)	0.536 (0.035)	0.577 (0.053)	0.542 (0.038)	0.576 (0.037)	0.578 (0.043)	**0.615** (0.041)	**0.619** (0.038)
	time	24.9	69.1	5.2	32.8	**1.3**	26.9	26.9	**2.5**	21.7

6.2 Performance on large datasets

We compare our AutoSC+NSE with LSC-K [Chen and Cai, 2011], SSC-OMP [You *et al.*, 2016b], and S^5C [Matsushima and Brbic, 2019], and S^3COMP-C [Chen *et al.*, 2020]. The parameter settings and more results are in Appendix E.7 and Appendix E.8 respectively. Table 4 shows the clustering accuracy and standard deviation of 10 repeated trials on the raw-pixel data of MNIST and Fashion-MNIST. Our methods have the highest clustering accuracy in every case. It is worth mentioning that if NSE is ablated, the accuracy of AutoSC-GD on MNIST-10k, -20k, and -30k are 0.9772, 0.9783 and 0.9864 respectively, higher than that of AutoSC-GD+NSE, though the time costs increased.

Table 4: Clustering accuracy and time cost (second) on MNIST and Fashion MNIST. "—" means the computation is out of memory.

		LSC-K	SSC-OMP	S^5C	S^3COMP-C	AutoSC-GD+NSE	AutoSC-BO+NSE
MNIST-10k	acc	0.652(0.037)	0.431(0.014)	0.646(0.045)	0.623(0.028)	**0.687**(0.035)	**0.679**(0.034)
	time	**18.9**	26.4	82.3	710.4/20	**16.2**	48.3
MNIST	acc	0.665(0.021)	0.453(0.017)	0.627(0.025)	—	**0.755**(0.022)	**0.750**(0.009)
	time	329.2	1178.3	961.5	—	**86.9**	**123.6**
Fashion-MNIST-10k	acc	0.571(0.025)	0.509(0.038)	0.565(0.021)	0.569(0.024)	**0.576**(0.011)	**0.572**(0.019)
	time	**18.6**	26.8	107.3	707.2/20	**17.3**	50.9
Fashion-MNIST	acc	0.561(0.015)	0.359(0.017)	0.559(0.013)	—	**0.586**(0.008)	**0.578**(0.012)
	time	335.1	1156.6	932.6	—	**88.7**	**122.8**

Table 5: Clustering accuracy (mean value and standard deviation) and time cost (second) on MNIST and Fashion-MNIST with feature extraction. "—" means the computation is out of memory and "/" means the algorithm was performed on a computational platform not comparable to ours. The underlined values are from Chen *et al.* [2020].

		LSC-K	SSC-OMP	S^5C	S^3COMP-C	AutoSC-GD+NSE	AutoSC-BO+NSE
MNIST	acc	0.8659(0.0215)	<u>0.8159</u>	0.7829(0.0283)	<u>0.9632</u>	**0.9775**(0.0034)	**0.9741**(0.0044)
	time	273.6	<u>280.6</u>	907.5	<u>416.8</u>	**59.2**	**115.3**
Fashion-MNIST	acc	0.6131(0.0298)	0.3796(0.0217)	0.6057(0.0227)	—	**0.6398**(0.0133)	**0.6461**(0.0104)
	time	251.8	1013.9	913.2	—	**61.9**	**112.6**
GTSRB	acc	0.8711(0.0510)	<u>0.8252</u>	0.9044(0.0267)	<u>0.9554</u>	**0.9873**(0.0126)	**0.9881**(0.0078)
	time	**31.2**	/	98.7	/	**16.8**	69.4

7 Conclusion

We have proposed an automated spectral clustering method. Extensive experiments showed the effectiveness and superiority of our methods over baseline methods. The efficiency improvement is from the closed-form solutions of the least squares regressions. The accuracy improvement is from the effectiveness of LSR and KLSR with thresholding.

References

Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Found. Trends Mach. Learn.*, 3(1):1–122, 2011.

Joan Bruna and Stéphane Mallat. Invariant scattering convolution networks. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1872–1886, 2013.

Deng Cai and Xinlei Chen. Large scale spectral clustering via landmark-based sparse representation. *IEEE transactions on cybernetics*, 45(8):1669–1680, 2014.

Xinlei Chen and Deng Cai. Large scale spectral clustering with landmark-based representation. In *Twenty-fifth AAAI conference on artificial intelligence*. Citeseer, 2011.

Ying Chen, Chun-Guang Li, and Chong You. Stochastic sparse subspace clustering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4155–4164, 2020.

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

Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd*, volume 96, pages 226–231, 1996.

Jicong Fan, Yuqian Zhang, and Madeleine Udell. Polynomial matrix completion for missing data imputation and transductive learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 3842–3849, 2020.

Miroslav Fiedler. Algebraic connectivity of graphs. *Czechoslovak mathematical journal*, 23(2):298–305, 1973.

M. Halkidi and M. Vazirgiannis. Clustering validity assessment: finding the optimal partitioning of a data set. In *Proceedings 2001 IEEE International Conference on Data Mining*, pages 187–194, 2001.

N. Halko, P. G. Martinsson, and J. A. Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Review*, 53(2):217–288, 2011.

Harold V Henderson and Shayle R Searle. On deriving the inverse of a sum of matrices. *Siam Review*, 23(1):53–60, 1981.

Juhua Hu, Qi Qian, Jian Pei, Rong Jin, and Shenghuo Zhu. Finding multiple stable clusterings. *Knowledge and Information Systems*, 51(3):991–1021, 2017.

Hsin-Chien Huang, Yung-Yu Chuang, and Chu-Song Chen. Affinity aggregation for spectral clustering. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 773–780. IEEE, 2012.

Frank Hutter, Lars Kotthoff, and Joaquin Vanschoren. *Automated machine learning: methods, systems, challenges*. Springer Nature, 2019.

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 (ICCV)*, December 2015.

Pan Ji, Tong Zhang, Hongdong Li, Mathieu Salzmann, and Ian Reid. Deep subspace clustering networks. In *Advances in Neural Information Processing Systems*, pages 24–33, 2017.

Stephen C Johnson. Hierarchical clustering schemes. *Psychometrika*, 32(3):241–254, 1967.

Donald R Jones, Matthias Schonlau, and William J Welch. Efficient global optimization of expensive black-box functions. *Journal of Global optimization*, 13(4):455–492, 1998.

Zhao Kang, Zhiping Lin, Xiaofeng Zhu, and Wenbo Xu. Structured graph learning for scalable subspace clustering: From single view to multiview. *IEEE Transactions on Cybernetics*, pages 1–11, 2021.

Mohsen Kheirandishfard, Fariba Zohrizadeh, and Farhad Kamangar. Deep low-rank subspace clustering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2020.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Aaron Klein, Stefan Falkner, Simon Bartels, Philipp Hennig, and Frank Hutter. Fast bayesian optimization of machine learning hyperparameters on large datasets. In *Artificial Intelligence and Statistics*, pages 528–536. PMLR, 2017.

Lee Kuang-Chih, J. Ho, and D. J. Kriegman. Acquiring linear subspaces for face recognition under variable lighting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(5):684–698, 2005.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Chun-Guang Li and Rene Vidal. Structured sparse subspace clustering: A unified optimization framework. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 277–286, 2015.

Chun-Guang Li and Rene Vidal. A structured sparse plus structured low-rank framework for subspace clustering and completion. *IEEE Transactions on Signal Processing*, 64(24):6557–6570, 2016.

Jun Li and Handong Zhao. Large-scale subspace clustering by fast regression coding. In *IJCAI*, 2017.

Jun Li, Hongfu Liu, Zhiqiang Tao, Handong Zhao, and Yun Fu. Learnable subspace clustering. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–15, 2020.

Yanchi Liu, Zhongmou Li, Hui Xiong, Xuedong Gao, and Junjie Wu. Understanding of internal clustering validation measures. In *2010 IEEE international conference on data mining*, pages 911–916. IEEE, 2010.

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

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.

Canyi Lu, Jiashi Feng, Zhouchen Lin, Tao Mei, and Shuicheng Yan. Subspace clustering by block diagonal representation. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):487–501, 2018.

Aleix M Martínez and Avinash C Kak. PCA versus LDA. *IEEE transactions on pattern analysis and machine intelligence*, 23(2):228–233, 2001.

Bertil Matérn. *Spatial variation*, volume 36. Springer Science & Business Media, 2013.

Shin Matsushima and Maria Brbic. Selective sampling-based scalable sparse subspace clustering. In *Advances in Neural Information Processing Systems*, pages 12416–12425, 2019.

Marian Meila and Susan Shortreed. Regularized spectral learning. *Journal of machine learning research*, 1(1):1–20, 2006.

Marian Meila, Susan Shortreed, and Liang Xu. Regularized spectral learning. In *AISTATS*. PMLR, 2005.

Marina Meila. The multicut lemma. *UW Statistics Technical Report*, 417, 2001.

S. A. Nene, S. K. Nayar, and H. Murase. Columbia object image library (coil-20). Report, Columbia University, 1996.

Andrew Y Ng, Michael I Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. In *Advances in neural information processing systems*, pages 849–856, 2002.

Pan Ji, M. Salzmann, and Hongdong Li. Efficient dense subspace clustering. In *IEEE Winter Conference on Applications of Computer Vision*, pages 461–468, 2014.

V. M. Patel and R. Vidal. Kernel sparse subspace clustering. In *2014 IEEE International Conference on Image Processing (ICIP)*, pages 2849–2853, Oct 2014.

V. M. Patel, Nguyen Hien Van, and R. Vidal. Latent space sparse and low-rank subspace clustering. *Selected Topics in Signal Processing, IEEE Journal of*, 9(4):691–701, 2015.

Xi Peng, Lei Zhang, and Zhang Yi. Scalable sparse subspace clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 430–437, 2013.

Xi Peng, Huajin Tang, Lei Zhang, Zhang Yi, and Shijie Xiao. A unified framework for representation-based subspace clustering of out-of-sample and large-scale data. *IEEE transactions on neural networks and learning systems*, 27(12):2499–2512, 2015.

Giannis Poulakis. Unsupervised automl: a study on automated machine learning in the context of clustering. Master's thesis, $\Pi\alpha\nu\varepsilon\pi\iota\sigma\tau\acute{\eta}\mu\iota o$ $\Pi\varepsilon\iota\rho\alpha\iota\acute{\omega}\varsigma$, 2020.

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

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

Jie Shen and Ping Li. Learning structured low-rank representation via matrix factorization. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, volume 51 of *Proceedings of Machine Learning Research*, pages 500–509, Cadiz, Spain, 09–11 May 2016. PMLR.

Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *IEEE Transactions on pattern analysis and machine intelligence*, 22(8):888–905, 2000.

Jasper Snoek, Hugo Larochelle, and Ryan P Adams. Practical bayesian optimization of machine learning algorithms. *Advances in neural information processing systems*, 25, 2012.

Mahdi Soltanolkotabi, Ehsan Elhamifar, and Emmanuel J Candes. Robust subspace clustering. *The Annals of Statistics*, 42(2):669–699, 2014.

Sho Sonoda and Noboru Murata. Neural network with unbounded activation functions is universal approximator. *Applied and Computational Harmonic Analysis*, 43(2):233 – 268, 2017.

Johannes Stallkamp, Marc Schlipsing, Jan Salmen, and Christian Igel. Man vs. computer: Benchmarking machine learning algorithms for traffic sign recognition. *Neural networks*, 32:323–332, 2012.

Hugo Steinhaus et al. Sur la division des corps matériels en parties. *Bull. Acad. Polon. Sci*, 1(804):801, 1956.

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

Ulrike Von Luxburg. A tutorial on spectral clustering. *Statistics and computing*, 17(4):395–416, 2007.

Yu-Xiang Wang and Huan Xu. Noisy sparse subspace clustering. In *International Conference on Machine Learning*, pages 89–97. PMLR, 2013.

Shusen Wang, Bojun Tu, Congfu Xu, and Zhihua Zhang. Exact subspace clustering in linear time. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence*, pages 2113–2120, 2014.

Y. Weiss. Segmentation using eigenvectors: a unifying view. In *Proceedings of the Seventh IEEE International Conference on Computer Vision*, volume 2, pages 975–982 vol.2, 1999.

Christopher K Williams and Carl Edward Rasmussen. *Gaussian processes for machine learning*, volume 2. MIT press Cambridge, MA, 2006.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms, 2017.

Chong You, Chun-Guang Li, Daniel P Robinson, and René Vidal. Oracle based active set algorithm for scalable elastic net subspace clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3928–3937, 2016.

Chong You, Daniel Robinson, and René Vidal. Scalable sparse subspace clustering by orthogonal matching pursuit. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3918–3927, 2016.

Chong You, Chi Li, Daniel P Robinson, and René Vidal. Scalable exemplar-based subspace clustering on class-imbalanced data. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 67–83, 2018.

Junjian Zhang, Chun-Guang Li, Chong You, Xianbiao Qi, Honggang Zhang, Jun Guo, and Zhouchen Lin. Self-supervised convolutional subspace clustering network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5473–5482, 2019.

Tong Zhang, Pan Ji, Mehrtash Harandi, Wenbing Huang, and Hongdong Li. Neural collaborative subspace clustering. In *International Conference on Machine Learning*, pages 7384–7393, 2019.

Checklist

The checklist follows the references. Please read the checklist guidelines carefully for information on how to answer these questions. For each question, change the default **[TODO]** to [Yes] , [No] , or [N/A] . You are strongly encouraged to include a **justification to your answer**, either by referencing the appropriate section of your paper or providing a brief inline description. For example:

- Did you include the license to the code and datasets? [Yes] See Section.
- Did you include the license to the code and datasets? [No] The code and the data are proprietary.
- Did you include the license to the code and datasets? [N/A]

Please do not modify the questions and only use the provided macros for your answers. Note that the Checklist section does not count towards the page limit. In your paper, please delete this instructions block and only keep the Checklist section heading above along with the questions/answers below.

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [N/A]
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...

(a) Did you state the full set of assumptions of all theoretical results? [Yes]
(b) Did you include complete proofs of all theoretical results? [Yes]

3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]

5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]