Attributed Subspace Clustering

Jing Wang[1], Linchuan Xu[1], Feng Tian[2], Atsushi Suzuki[1], Changqing Zhang[3,*] Kenji Yamanishi[1]
[1] Graduate School of Information Science and Technology, The University of Tokyo, Japan
[2] Faculty of Science and Technology, Bournemouth University, UK
[3] College of Intelligence and Computing, Tianjin University, Tianjin, China
jing_wang@mist.i.u-tokyo.ac.jp, linchuan_xu@mist.i.u-tokyo.ac.jp, ftian@bournemouth.ac.uk,
atsushi_suzuki@mist.i.u-tokyo.ac.jp, zhangchangqing@tju.edu.cn, yamanishi@mist.i.u-tokyo.ac.jp

Abstract

Existing methods on representation-based subspace clustering mainly treat all features of data as a whole to learn a single self-representation and get one clustering solution. Real data however are often complex and consist of multiple attributes or sub-features, such as a face image has expressions or genders. Each attribute is distinct and complementary on depicting the data. Failing to explore attributes and capture the complementary information among them may lead to an inaccurate representation. Moreover, a single clustering solution is rather limited to depict data, which can often be interpreted from different aspects and grouped into multiple clusters according to attributes. Therefore, we propose an innovative model called attributed subspace clustering (ASC). It simultaneously learns multiple self-representations on latent representations derived from original data. By utilizing Hilbert Schmidt Independence Criterion as a co-regularizing term, ASC enforces that each self-representation is independent and corresponds to a specific attribute. A more comprehensive self-representation is then established by adding these self-representations. Experiments on several benchmark image datasets have demonstrated the effectiveness of ASC not only in terms of clustering accuracy achieved by the integrated representation, but also the diverse interpretation of data, which is beyond what current approaches can offer.

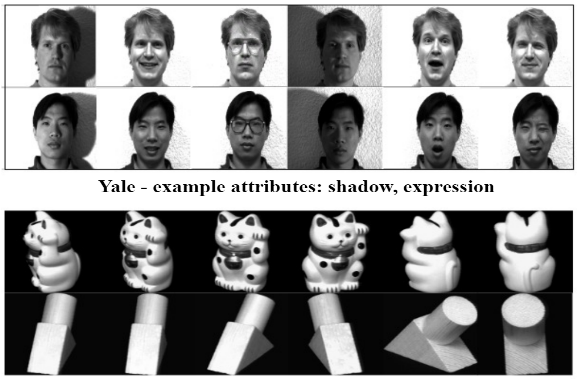
1 Introduction

Representation-based subspace clustering is to partition data points into their respective low-dimensional subspaces by finding effective self-representations [Vidal and Favaro, 2014]. It assumes that every data point in a union of subspaces can be represented as a linear combination of other data points. Due to the insensitivity to initialization and easy-to-solve with standard linear algebra [Liu *et al.*, 2014], the representation based subspace clustering has been widely used in computer vision and pattern recognition [Rao *et al.*, 2010; Liu *et al.*, 2013; Zhou *et al.*, 2014].

Recently, various self-representation based methods have been proposed [Vidal and Favaro, 2014; Wu *et al.*, 2015; Zhang *et al.*, 2017]. The sparse subspace clustering (SSC) [Elhamifar and Vidal, 2009] seeks the sparse solution of self-representation, which tends to be block diagonal. The low-rank representation (LRR) [Liu *et al.*, 2010] aims to preserve low-rank data structure. The least squares regression (LSR) [Lu *et al.*, 2012] uses grouping effect for modeling correlation structure of data and is more efficient than SSC and LRR. The correlation adaptive subspace segmentation (CASS) [Lu *et al.*, 2013] simultaneously performs automatic data selection and groups correlated data, which can adaptively balance SSC and LSR. Based on LSR, the smooth representation (SMR) [Hu *et al.*, 2014] incorporates a weight matrix that measures the spatial closeness of data. Given a dataset with multiple types of features, [Cao *et al.*, 2015] proposed a diverse multi-view subspace clustering (DiMSC) which learns a complementary representation shared by multiple features. [Zhang *et al.*, 2017] then proposed latent multi-view subspace clustering (LMSC) method, which clusters data points with latent representation and simultaneously explores underlying complementary information from multiple views. With the utilization of prior information, NNLRR [Fang *et al.*, 2015] encodes label information with graph-based regularization to seek low-rank and sparse representation simultaneously with nonnegativity constraints. Later, LRRADP [Wang *et al.*, 2018] was proposed by using adaptive distance penalty to construct an affinity graph, which enforces the representations of every two consecutive neighboring data to be similar.

These representation-based approaches, which incorporate regularization terms or prior information for more accurate learning, all tend to intuitively utilize the features of original data as a whole and learn a single self-representation. However, real data are complex and consist of multiple sub-features or attributes [Changpinyo *et al.*, 2013; Ou *et al.*, 2015]. As shown in Figure 1, the face images consist of multiple attributes including facial expressions, ethnicity and shadows, and the cat images have different rotations and shapes. Since each attribute represents one aspect of data and contains specific information, exploring diverse and complementary information among multiple attributes is of vital importance to learn a more accurate self-representation.

[*] Corresponding author



Yale - example attributes: shadow, expression

COIL20 - example attributes: rotation, shape

Figure 1: Sample images from the Yale and COIL20 datasets. Each row represents one subject with different attributes.

Moreover, the exploration of attributes may lead to multiple clustering solutions that provide a more comprehensive understanding of data. Taking the Yale in Figure 1 as an example, current subspace clustering methods can only obtain one clustering solution based on the single self-representation, i.e., grouping all face images of a person into one group, which is often rather limited to depict and understand the dataset adequately. However, with multiple self-representations corresponding to attributes (e.g. shadow and expression), the face images could be clustered as happy/dull expression, with/without glass or left/right-lit. In fact, the effectiveness of attribute exploration has also been demonstrated in some other research fields, including network embedding, ordinal embedding, information retrieval and so on [Niu *et al.*, 2014; Liao *et al.*, 2018; Mazaheri *et al.*, 2018].

In this paper, we propose an innovative representation based subspace clustering approach, called Attributed Subspace Clustering (ASC), which interprets data from different perspectives and obtains multiple clustering solutions by leveraging the data's multiple attributes. Specifically, ASC learns multiple self-representations based on latent representations which are drawn from the original data. Utilizing Hilbert-Schmidt Independence Criterion (HSIC) [Gretton *et al.*, 2005] to co-regularize the latent representations, each self-representation is enforced to be independent and correspond to a particular attribute. Adding all these self-representations together results in an integrated representation which leads to a more accuracy clustering by capturing comprehensive information.

2 Attributed Subspace Clustering

2.1 Preliminary

Given n data matrix $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n] \in \mathbb{R}^{m \times n}$, where each data vector $\mathbf{x}_i$ is m-dimensional, we need to find effective self-representations for constructing an affinity matrix and applying spectral clustering methods [Shi and Malik, 2000] to cluster the data into their respective subspaces. To achieve so, representation-based approaches assume that every data point in a union of subspaces can be represented as a linear combination of other data points, i.e., $\mathbf{X} = \mathbf{XZ} + \mathbf{E}$, where $\mathbf{Z} \in \mathbb{R}^{n \times n}$ is the learned self-representation matrix and $\mathbf{E} \in \mathbb{R}^{n \times n}$ is the error term. It can be formulated as the following optimization problem to compute the optimal $\mathbf{Z}$:

$$\begin{aligned} &\min_{\mathbf{Z}} \Theta(\mathbf{E}) + \alpha\Omega(\mathbf{X}, \mathbf{Z}), \\ &s.t. \mathbf{X} = \mathbf{XZ} + \mathbf{E}, \mathbf{Z} \in \mathcal{T}, \end{aligned} \tag{1}$$

where α is the trade-off parameter, $\Theta(\mathbf{E})$ is the noise term. $\Omega(\mathbf{X}, \mathbf{Z})$ and $\mathcal{T}$ are the regularizer and constraint set on $\mathbf{Z}$, respectively.

Taking LRR, one of the most representative subspace clustering methods as an example, it seeks the low rank representation by solving the objective function:

$$\min_{\mathbf{Z}} \|\mathbf{X} - \mathbf{XZ}\|_F^2 + \alpha\|\mathbf{Z}\|_*. \tag{2}$$

Obviously, the self-representation of LRR is learned straightforwardly from the features of original data without exploring sub-features. Hence, it cannot represent multiple attributes and capture the diverse information among them. To understand and represent data thoroughly and in-depth, we propose our attributed subspace clustering (ASC) below.

2.2 The Proposed ASC

As seen in (2), the self-representation $\mathbf{Z}$ is learned from the original features of $\mathbf{X}$ by using $\mathbf{X}$ itself as a basis matrix. Hence, if $\mathbf{X}$ comes with V attributes, we can modify (2) to learn multiple self-representations $\mathbf{Z}^{(v)}$ simultaneously, with each one corresponding to one attribute:

$$\sum_{v=1}^{V} \|\mathbf{H}^{(v)} - \mathbf{H}^{(v)}\mathbf{Z}^{(v)}\|_F^2 + \alpha \sum_{v=1}^{V} \|\mathbf{Z}^{(v)}\|_*, \tag{3}$$

where $\mathbf{H}^{(v)}$ contains sub-features that can represent the v-th attribute of $\mathbf{X}$ and so called latent representation. Considering that every $\mathbf{H}^{(v)}$ describes one aspect of $\mathbf{X}$, given a basis matrix $\mathbf{W}^{(v)} \in \mathbb{R}^{m \times k^{(v)}}$ with a reduced dimensionality $k^{(v)}$, the product of each $\mathbf{W}^{(v)}$ and the corresponding $\mathbf{H}^{(v)}$ could well approximate $\mathbf{X}$, hence we can easily factorize $\mathbf{X}$ as

$$\min_{\mathbf{W}^{(v)}, \mathbf{H}^{(v)}} \sum_{v=1}^{V} \|\mathbf{X} - \mathbf{W}^{(v)}\mathbf{H}^{(v)}\|_F^2. \tag{4}$$

However, since there is no constraint for learning $\mathbf{H}^{(v)}$, the diverse information of multiple attributes may not be explored effectively, as every $\mathbf{H}^{(v)}$ could be very close to or even equal to each other. Consequently, if we simply combine (4) and (3), $\mathbf{Z}^{(v)}$ will also be similar to each other, so that the distinct information of each attribute cannot be fully explored.

Given a case of data $\mathbf{x}_i$, with two attributes, v and w, the latent distinct information of each attribute cannot be fully explored unless the representations of two attributes, i.e., $\mathbf{h}_i^{(v)}$ and $\mathbf{h}_i^{(w)}$, are enforced to be independent to each other. For n data vectors, if we assume that each vth attribute is drawn from $\mathcal{X}$ space and the wth attribute from $\mathcal{Y}$ space, essentially, we aim to learn a mapping function G of their representations from $S := \{(\mathbf{h}_1^{(v)}, \mathbf{h}_1^{(w)}), (\mathbf{h}_2^{(v)}, \mathbf{h}_2^{(w)}), \dots, (\mathbf{h}_n^{(v)}, \mathbf{h}_n^{(w)})\} \subseteq \mathcal{X} \times \mathcal{Y}$, i.e., G: $\mathcal{X} \rightarrow \mathcal{Y}$, to minimize the dependence between the data representations in the $\mathcal{X}$ and $\mathcal{Y}$. To do so, we employ the Hilbert-Schmidt Independence Criterion (HSIC) due to its simplicity and neat theoretical properties such as exponential convergence. HSIC computes the square of the norm of the cross-covariance operator over the domain $\mathcal{X} \times \mathcal{Y}$ in Hilbert Space. As an effective measure of dependence, HSIC has been applied to several machine learning tasks recently [Song *et al.*, 2007; Zhang and Zhou, 2010; Niu *et al.*, 2010]. Mathmatically, an empirical estimate of HSIC [Gretton *et al.*, 2005] is defined as

$$\mathrm{HSIC}(\mathbf{H}^{(v)}, \mathbf{H}^{(w)}) = (n-1)^{-2} tr(\mathbf{R}\mathbf{K}^{(v)}\mathbf{R}\mathbf{K}^{(w)}), \quad (5)$$

where $\mathbf{K}^{(v)}$ and $\mathbf{K}^{(w)}$ are the centered Gram matrices of kernel functions defined over $\mathbf{H}^{(v)}$ and $\mathbf{H}^{(w)}$, respectively. $\mathbf{R} = \mathbf{I} - \frac{1}{n}\mathbf{e}\mathbf{e}^T$, where $\mathbf{I}$ is an identity matrix and $\mathbf{e}$ is an all-one column vector.

It is worth noticing that by utilizing the inner product kernel for HSIC, namely, $\mathbf{K}^{(v)} = {\mathbf{H}^{(v)}}^T\mathbf{H}^{(v)}$ and ignoring the scaling factor $(n-1)^{-2}$ for notational convenience, we can derive $\mathrm{HSIC}(\mathbf{H}^{(v)}, \mathbf{H}^{(w)})$ into a very simple F-norm formulation as

$$\begin{aligned} &\mathrm{HSIC}(\mathbf{H}^{(v)}, \mathbf{H}^{(w)}) = tr(\mathbf{R}\mathbf{K}^{(v)}\mathbf{R}\mathbf{K}^{(w)}) \\ &= tr(\mathbf{R}{\mathbf{H}^{(v)}}^T\mathbf{H}^{(v)}\mathbf{R}{\mathbf{H}^{(w)}}^T\mathbf{H}^{(w)}) = \|\mathbf{H}^{(v)}\mathbf{R}{\mathbf{H}^{(w)}}^T\|_F^2. \end{aligned} \quad (6)$$

Combining (3), (4) and (6) gives us the final objective function:

$$\begin{aligned} \min_{\mathbf{W}^{(v)},\mathbf{H}^{(v)},\mathbf{Z}^{(v)}} &\sum_{v=1}^{V} \|\mathbf{X} - \mathbf{W}^{(v)}\mathbf{H}^{(v)}\|_F^2 + \lambda_1 \sum_{v \neq w} \|\mathbf{H}^{(v)}\mathbf{R}{\mathbf{H}^{(w)}}^T\|_F^2 \\ &+ \lambda_2 \sum_{v=1}^{V} \|\mathbf{H}^{(v)} - \mathbf{H}^{(v)}\mathbf{Z}^{(v)}\|_F^2 + \lambda_3 \sum_{v=1}^{V} \|\mathbf{Z}^{(v)}\|_*. \end{aligned} \quad (7)$$

Here $\lambda_1 > 0$, $\lambda_2 > 0$ and $\lambda_3 > 0$ are the trade-off parameters to balance the diversity among attribute representations, errors and intrinsic data structures for all V attributes, respectively. Specifically, the first term uncovers the underlying representations $\{\mathbf{H}^{(v)}\}_{v=1}^V$ from the original data matrix $\mathbf{X}$ by using the second term as a regularizer. The first two terms guide the learning process of self-representation $\{\mathbf{Z}^{(v)}\}_{v=1}^V$ in the third term. The last term regularizes $\{\mathbf{Z}^{(v)}\}_{v=1}^V$ to capture the low rank structure of data of each attribute.

Remarks ASC is not limited to one specific subspace clustering method such as LRR. In fact, it can also be flexibly adapted and applied to most if not all current representation-based subspace clustering approaches such as SSC, LSR and SMR which differ mainly on regularization on $\mathbf{Z}$. For example, SSC can be advanced as

$$\begin{aligned} \min_{\mathbf{W}^{(v)},\mathbf{H}^{(v)},\mathbf{Z}^{(v)}} &\sum_{v=1}^{V} \|\mathbf{X} - \mathbf{W}^{(v)}\mathbf{H}^{(v)}\|_F^2 + \lambda_1 \sum_{v \neq w} \|\mathbf{H}^{(v)}\mathbf{R}{\mathbf{H}^{(w)}}^T\|_F^2 \\ &+ \lambda_2 \sum_{v=1}^{V} \|\mathbf{H}^{(v)} - \mathbf{H}^{(v)}\mathbf{Z}^{(v)}\|_F^2 + \lambda_3 \sum_{v=1}^{V} \|\mathbf{Z}^{(v)}\|_1, \\ s.t. \quad & diag(\mathbf{Z}^{(v)}) = 0. \end{aligned} \quad (8)$$

Thus, SSC could be more powerful and practical by simultaneously learning spare self-representation and exploring multiple attributed self-representations, which we will investigate in our future work.

2.3 Optimization

Since the optimization problem (7) is not convex for each variable $\mathbf{W}^{(v)}$, $\mathbf{H}^{(v)}$ and $\mathbf{Z}^{(v)}$, it is infeasible to find the global minimum. Hence we divide (7) into three subproblems and alternately update each subproblem with the other two fixed.

$\mathbf{W}^{(v)}$**-subproblem**: It is a standard matrix factorization decomposition, so we obtain

$$\mathbf{W}^{(v)} = \mathbf{X}{\mathbf{H}^{(v)}}^T(\mathbf{H}^{(v)}{\mathbf{H}^{(v)}}^T)^{-1}. \quad (9)$$

$\mathbf{H}^{(v)}$**-subproblem**: Updating $\mathbf{H}^{(v)}$ with other subproblems fixed leads to the following objective function:

$$\begin{aligned} &\min_{\mathbf{H}^{(v)}} J(\mathbf{H}^{(v)}) = \|\mathbf{X} - \mathbf{W}^{(v)}\mathbf{H}^{(v)}\|_F^2 \\ &+ \lambda_1 \sum_{w=1, w\neq v}^{V} \|\mathbf{H}^{(v)}\mathbf{R}{\mathbf{H}^{(w)}}^T\|_F^2 + \lambda_2\|\mathbf{H} - \mathbf{H}^{(v)}\mathbf{Z}^{(v)}\|_F^2. \end{aligned} \quad (10)$$

The problem (10) is smooth convex. Differentiating (10) with respect to $\mathbf{H}^{(v)}$ and setting it to be 0, we get the following optimal solution $\mathbf{H}^{(v)}$ which satisfies

$$\begin{aligned} &\mathbf{A}^{(v)}\mathbf{H}^{(v)} + \mathbf{H}^{(v)}\mathbf{B}^{(v)} = \mathbf{C}^{(v)}. \\ s.t. \quad &\mathbf{A}^{(v)} = {\mathbf{W}^{(v)}}^T\mathbf{W}^{(v)}, \mathbf{C}^{(v)} = {\mathbf{W}^{(v)}}^T\mathbf{X}, \\ \mathbf{B}^{(v)} &= \lambda_1\mathbf{R}\sum_{w=1,w\neq v}^{V}{\mathbf{H}^{(w)}}^T\mathbf{H}^{(w)}\mathbf{R} + \lambda_2(\mathbf{I} - {\mathbf{Z}^{(v)}}^T + {\mathbf{Z}^{(v)}}^T\mathbf{Z}^{(v)}). \end{aligned} \quad (11)$$

The equation above is a standard Sylvester equation [Bartels and Stewart, 1972]. To avoid the instability issue, we ensure $\mathbf{A}^{(v)}$ to be strictly positive definite by $\mathbf{A}^{\hat{(v)}} = \mathbf{A}^{(v)} + \epsilon\mathbf{I}$, where $\mathbf{I}$ is an identity matrix and $\epsilon \in (0, 1]$. In the following, we prove it has a unique solution.

Proposition 1. *The Sylvester equation* (11) *has a unique solution.*

Proof. *The Sylvester equation* $\mathbf{A}^{(v)}\mathbf{H}^{(v)} + \mathbf{H}^{(v)}\mathbf{B}^{(v)} = \mathbf{C}^{(v)}$ *has a unique solution for* $\mathbf{H}^{(v)}$ *exactly when there are no common eigenvalues of* $\mathbf{A}^{(v)}$ *and* $-\mathbf{B}^{(v)}$ *[Bartels and Stewart, 1972]. Since* $\mathbf{A}^{(v)}$ *is a positive definite matrix, all of its eigenvalues are positive:* $\alpha_i^{(v)} > 0$. *For* $\mathbf{B}^{(v)}$ *is a positive semi-definite matrix, all of its eigenvalues are nonnegative:* $\beta_i^{(v)} \geq 0$. *Hence, for any eigenvalues of* $\mathbf{A}^{(v)}$ *and* $\mathbf{B}^{(v)}$, $\alpha_i^{(v)} + \beta_j^{(v)} > 0$. *Accordingly, the Sylvester equation* (11) *has a unique solution.*

$\mathbf{Z}^{(v)}$**-subproblem**: Updating $\mathbf{Z}^{(v)}$ with other subproblems fixed leads to

$$\min_{\mathbf{Z}^{(v)}} J(\mathbf{Z}^{(v)}) = \|\mathbf{H}^{(v)} - \mathbf{H}^{(v)}\mathbf{Z}^{(v)}\|_F^2 + \lambda_3\|\mathbf{Z}^{(v)}\|_*. \quad (12)$$

This can be solved via the following lemma.

Lemma 1 [Favaro *et al.*, 2011]. *Let* $\mathbf{A} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^T$ *be the SVD of a given matrix A. The optimal solution to* $\min_{\mathbf{C}} \|\mathbf{C}\|_* + \frac{\tau}{2}\|\mathbf{A} - \mathbf{AC}\|_F^2$ *is*

$$\mathbf{C} = \mathbf{V}_1(\mathbf{I} - \frac{1}{\tau}\mathbf{\Lambda}_1^{-2})\mathbf{V}_1^T, \tag{13}$$

where $\mathbf{U} = [\mathbf{U}_1\mathbf{U}_2]$, $\mathbf{\Lambda} = diag(\mathbf{\Lambda}_1, \mathbf{\Lambda}_2)$, *and* $\mathbf{V} = [\mathbf{V}_1\mathbf{V}_2]$ *are partitioned according to the sets* $\mathbf{I}_1 = \{i : \sigma_i > 1/\sqrt{\tau}\}$ *and* $\mathbf{I}_2 = \{i : \sigma_i \leq 1/\sqrt{\tau}\}$, *where* σ_i *is the ith entry of* $\mathbf{\Lambda}$. *Moreover, the optimal value is*

$$\psi(\mathbf{A}) = \sum_{i=\mathbf{I}_1}(1 - \frac{1}{2\tau}\sigma_i^{-2}) + \frac{\tau}{2}\sum_{i\in\mathbf{I}_2}\sigma_i^2. \tag{14}$$

The whole procedure of ASC is summarized in Algorithm 1.

Algorithm 1 Solving ASC

```
Input: Original data matrix $\mathbf{X}$, Number of attributes $V$,
    Number of subspaces $p$, Parameters $\lambda_1$, $\lambda_2$, $\lambda_3$
Output: Clustering result.
  for $v = 1 : V$ do
    Initialize $\mathbf{W}^{(v)}$, $\mathbf{H}^{(v)}$ and $\mathbf{Z}^{(v)}$ with random values.
  end for
  while not converge do
    for $v = 1 : V$ do
      Fixing $\mathbf{H}^{(v)}$ and $\mathbf{Z}^{(v)}$, update $\mathbf{W}^{(v)}$ by (9).
      Fixing $\mathbf{W}^{(v)}$ and $\mathbf{Z}^{(v)}$, update $\mathbf{H}^{(v)}$ by solving (11).
      Fixing $\mathbf{W}^{(v)}$ and $\mathbf{H}^{(v)}$, update $\mathbf{Z}^{(v)}$ by solving (12).
    end for
  end while
  Construct an integrated self-representation matrix $\mathbf{Z} = \sum_{v=1}^{V}\mathbf{Z}^{(v)}$ and a similarity matrix $\mathbf{S} = |\mathbf{Z}| + |\mathbf{Z}^T|$.
  Segment the data into $p$ groups by Normalized Cuts.
```

2.4 Complexity and Convergence Analysis

The complexity of updating $\mathbf{W}^{(v)}$ in (9) is $O(mnk^{(v)}+{k^{(v)}}^3)$ and $\mathbf{H}^{(v)}$ in (11) is $O({k^{(v)}}^3)$ by using Bartels Stewart algorithm for solving the Sylvester equation. Updating $\mathbf{Z}^{(v)}$ has $O(n^3)$ complexity as it involves SVD decomposition. Since usually $k^{(v)} \ll n$, the overall cost is $\sum_{v=1}^{V} O(mnk^{(v)} + n^3)$ which is the same as that of LRR. Hence, we can conclude that ASC does not increase the complexity with respect to n as the result of exploring attributes, while in the meantime, multiple representations are learned to depict data comprehensively. Since the optimizations of all subproblems are convex and we can obtain each optimal solution, the value of (7) is monotonically decreasing. Therefore, the objective function (7) converges as it has trivial lower-bound 0.

3 Experiments

3.1 Dataset

To demonstrate the effectiveness of ASC, we carried experiments on several benchmark datasets. The Yale[1] contains 11 face images for each of 15 subjects. The face images of each subject are either in different facial expressions (such as happy or sad) or configurations (such as with or without glasses). The COIL20[2] is composed of 1440 images for 20 objects. The 72 images of each object were captured by a fixed camera at a pose intervals of 5 degree. Some sample images of the two objects are shown in Figure 1. The Extended YaleB (YaleB) [Hu *et al.*, 2014] contains 38 individuals and around 64 near frontal images under different illuminations for each individual. We use the first 10 classes of this dataset, which consists of 640 frontal face images. The ORL[3] dataset consists of 40 subjects with each containing 10 face images with various lighting and facial expressions. The Notting-Hill [Zhou *et al.*, 2014] is a video face dataset, which is derived from the movie "Notting Hill". We used the 4660 faces from 5 main casts. The USPS [Hu *et al.*, 2014] contains 9298 images of handwritten digits belonging to ten digits, 0-9. We use the first 100 images of each digit for our experiments. The images vary in each class but have common stroke attributes (such as 3 and 8) in different classes.

3.2 Experiment Setup

We compared ASC against the state-of-the-arts SSC [Elhamifar and Vidal, 2009], LRR [Liu *et al.*, 2013], LSR [Lu *et al.*, 2012], SMR [Hu *et al.*, 2014] and L2-Graph [Peng *et al.*, 2017]. The parameters of these methods were tuned to achieve the best performance for a fair comparison. For ASC, we empirically fixed $(\lambda_1, \lambda_2) = (0.2\lambda_3, 0.1\lambda_3)$ and tuned λ_3 from [0.1, 0.2, 0.3, 0.4, 0.5]. We also fixed the number of attributes $V = 3$ and each reduced dimension $k^{(v)} = 50$ for all experiments. For all approaches, the similarity matrices were conducted on the typical similarity measures [Georghiades *et al.*, 2001] and Normalized Cuts [Shi and Malik, 2000] was employed to produce the final clustering results. Four widely used evaluation metrics, including accuracy, normalized mutual information (NMI), Purity and F-score are used to assess the quality of the results with a comprehensive evaluation, and the best results are highlighted in **boldface**.

3.3 Performance Analysis

Clustering Results. Table 1 summarizes the average clustering results along with standard deviations. We can see that ASC performs the best on all datasets, which proves the effectiveness of exploring diverse information among attributes. Especially for YaleB dataset, ASC outperforms the second best result achieved by LRR with a large margin, i.e., 11.68%. This could be due to two reasons. One is that YaleB comes with heavy noises and outliers. LRR learns self-representation based on the original data $\mathbf{X}$ but ASC is based on latent representations $\{\mathbf{H}^{(v)}\}_{v=1}^3$, the features of which are extracted from the original data with noises alleviated by the first term in (7). Hence, the learned self-representations of ASC are more accurate and lead to a higher accuracy. The other reason is that ASC learns more comprehensive information of data by exploring diverse informa-

[1] http://www.cad.zju.edu.cn/home/dengcai/Data/FaceData.html

[2] http://www1.cs.columbia.edu/CAVE/software/softlib/coil-20.php

[3] http://www.cad.zju.edu.cn/home/dengcai/Data/FaceData.html

Methods	Metrics	Yale	YaleB	ORL	USPS	COIL20	NHill
SSC	Accuracy	58.84±0.79	62.72±0.21	75.05±1.29	75.64±0.00	73.19±0.00	68.35±0.46
	NMI	60.20±0.64	63.26±0.16	87.91±0.26	75.54±0.10	89.03±0.80	82.64±1.22
	Purity	60.00±0.74	62.72±0.21	78.90±1.05	79.24±0.00	80.42±0.00	76.46±0.56
	F-score	36.60±1.14	38.43±0.30	63.32±1.46	65.66±0.16	68.84±0.00	60.66±0.94
LRR	Accuracy	57.27±0.74	62.38±0.28	78.80±0.93	78.40±0.00	82.79±0.56	67.51±0.13
	NMI	54.23±0.52	63.62±0.21	88.15±0.25	81.22±0.00	91.16±0.32	80.31±1.05
	Purity	50.42±0.51	62.47±0.20	83.10±0.88	82.80±0.00	87.87±0.00	74.36±1.02
	F-score	33.17±1.39	44.92±0.81	64.29±1.18	72.27±0.00	80.21±0.70	58.25±0.65
LSR	Accuracy	57.09±0.66	66.56±0.00	74.60±1.28	69.13±0.33	69.44±0.76	72.16±0.45
	NMI	58.64±0.39	62.54±0.00	86.81±0.69	65.74±0.24	78.83±0.39	83.12±0.16
	Purity	57.82±0.54	67.34± 0.00	78.40±1.17	73.19±0.18	69.96±0.88	75.67±0.32
	F-score	35.10±0.54	42.09±0.00	66.26±1.71	53.73±0.33	62.08±1.06	60.03±0.47
SMR	Accuracy	55.39±0.33	63.81±0.00	76.50±0.61	79.72±0.00	73.33±0.00	69.93±0.33
	NMI	56.53±0.55	64.52±0.00	85.27±0.18	67.69±0.11	86.05±0.22	84.39±0.32
	Purity	56.00±0.33	61.81±0.00	77.10±0.38	74.72±0.00	76.99±0.00	73.88±0.48
	F-score	32.06±0.57	40.96±0.15	60.79±0.33	59.23±0.00	62.87±0.00	57.76±1.24
L2-Graph	Accuracy	54.42±1.27	63.53±0.00	74.15±1.24	69.52±0.11	67.46±0.16	66.65±1.13
	NMI	57.48±1.43	63.91±0.00	85.39±1.17	66.73±0.31	76.38±0.29	80.67±0.24
	Purity	55.03±1.27	62.53±0.00	76.15±1.46	73.36±0.13	68.38±0.17	74.64±1.09
	F-score	33.27±0.93	40.77±0.00	62.42±1.41	54.26±0.36	59.09±0.28	60.95±0.81
ASC	Accuracy	**60.12±1.57**	**74.06±3.85**	**81.20±1.07**	**79.04±0.81**	**84.25±0.92**	**72.64±1.46**
	NMI	**60.72±0.44**	**71.01±2.67**	**88.97±0.63**	**82.19±0.72**	**93.14±0.55**	**84.68±0.56**
	Purity	**60.12±1.57**	**74.06±3.85**	**84.10±0.84**	**83.44±0.92**	**88.03±0.22**	**76.11±0.65**
	F-score	**37.32±1.43**	**56.26±4.03**	**68.84±1.55**	**73.75±1.14**	**80.69±0.99**	**61.48±1.12**

Table 1: Clustering results (mean ± standard deviation).

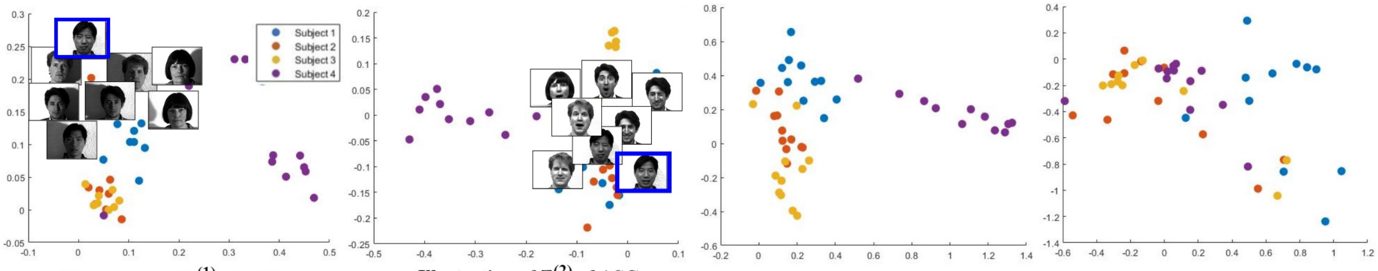


Illustration of $Z^{(1)}$ of ASC (Face images with shadow are close)

Illustration of $Z^{(2)}$ of ASC (Happy/surprised face images are close)

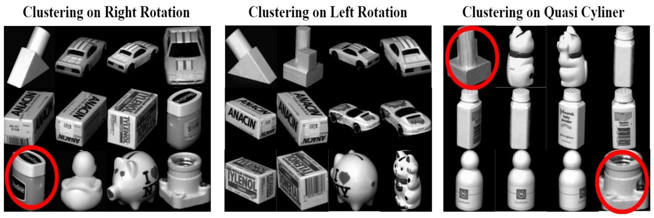
Illustration of Z of ASC

Illustration of Z of LRR

Figure 2: Illustrations of self-representations of ASC against LRR on the Yale, with their dimensionalities of features being reduced to 2-D by PCA. The solid circles in different colors indicate the face images of different subjects and the blue rectangle indicates the same image. The attributed self-reresentations $\mathbf{Z}^{(1)}$ and $\mathbf{Z}^{(2)}$ of ASC uncover multiple attributes, demonstrating that the images sharing the common attribute are close together. The integrated $\mathbf{Z}$ of ASC reveals a clearer structure than that of LRR - the distribution of a subject's images is more coherent.

tion among multiple independent self-representations and integrating them together. However, LRR learns a single self-representation based on the features of original data only. To validate our reasoning, below we take a close look at the self-representations.

Study of Attributes. Figure 2 illustrates the learned self-representation matrices of ASC and LRR for the Yale dataset. We only demonstrate the images of four subjects for clearer visualization, as well as two (out of three) attributed self-representations $\mathbf{Z}^{(1)}$ and $\mathbf{Z}^{(2)}$ of ASC due to page limitation. We can see that the closeness among images are different in $\mathbf{Z}^{(1)}$ and $\mathbf{Z}^{(2)}$. For example, the face images with shadow are very close to each other in $\mathbf{Z}^{(1)}$ while the faces with happy/surprised expressions are close in $\mathbf{Z}^{(2)}$. Apparently, these representations enable the understanding of the data from multiple perspectives, which could be hardly achievable by current subspace clustering approaches. It is worthy noticing that the image that enclosed in blue rectangle in $\mathbf{Z}^{(1)}$ is also represented in $\mathbf{Z}^{(2)}$, which means that the same image could be grouped differently based on different attributes.



Clustering on Right Rotation

Clustering on Left Rotation

Clustering on Quasi Cyliner

Figure 3: Diverse clustering solutions of the COIL20 dataset. Each clustering is based on an attribute: left rotation, right rotation and quasi cyliner shape. Images circled in red are outliers.

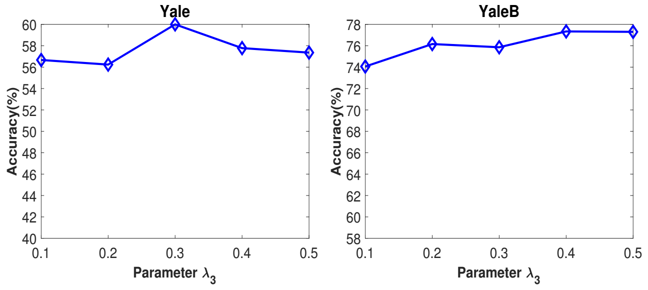


Figure 4: The effect of the parameter λ_3.

Moreover, in terms of the whole features of data (i.e., subject), the integrated $\mathbf{Z}$ of ASC shown in Figure 2 has a much clearer data structure than that of LRR. This will undoubtedly lead to better clustering results which are in line with the results given in the Table 1. We further tested ASC on COIL20, a larger dataset, and demonstrated some example clustering results in Figure 3. Again, the results are quite impressive and promising, with multiple clusters being obtained through different attributes (rotation, shape, etc), which is beyond what existing subspace clustering approaches can offer.

Parameter Analysis. Here we tested the sensitivity of parameters λ_1, λ_2 and λ_3 on clustering results. We varied λ_3 from 0.1 to 0.5 with an increment of 0.1 and we experimentally fixed $(\lambda_1, \lambda_2) = (0.2\lambda_3, 0.1\lambda_3)$. Since the performance on each dataset has a similar tendency, we show the parameters' effects on the accuracy for Yale and YaleB only. As shown in Figure 4, ASC shows a relatively stable performance on the two datasets, which demonstrates the robustness to parameter tuning. Also, the performance of ASC is consistently better than LRR when λ_3 is tuned in a suitable range. Taking the Yale as an example, ASC performs better when $\lambda_3 \in [0.3, 0.5]$ than 0.5727 that is achieved by LRR (Table 1). Worth to mention that, the better accuracies could be expected by grid search λ_1, λ_2 and λ_3, though it will inevitably incur a higher cost of the parameter tuning. We also evaluated the effect of the number of attributes V for both Yale and YaleB. Here we fixed $\lambda_3 = 0.5$ and gradually increased V from 1 to 5. As illustrated it in Figure 5, the accuracy for both datasets increases when V is tuned from 1 to 3, which signifies the effectiveness of the exploration of multiple attributes. However, the accuracy decreases when V increases from 3 to 5. The fluctuation could be due to a compromise between the amount of available features for each $\mathbf{H}^{(v)}$ and the diverse information among them. When V increases, more diverse information can be utilized. However, when V exceeds the number of prominent attributes, the amount of discriminative features for each $\mathbf{H}^{(v)}$ could be insufficient for representing an attribute.

Convergence Analysis. Having proven the convergence of ASC in the Section 2.4, here we experimentally demonstrate its convergence in Figure 6, where the horizontal axis is the number of iterations and the vertical axis is the value of objective function. It can be seen that the values of the objective function are non-increasing and drop sharply around 10 iterations on the Yale.

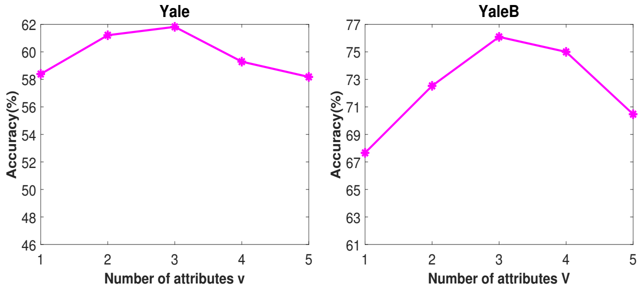


Figure 5: The effect of the number of attributes V.

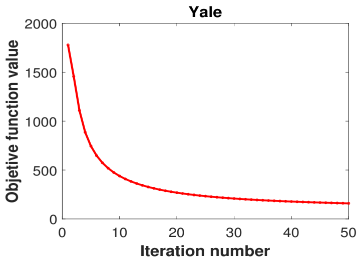


Figure 6: Convergence curves.

4 Conclusion

In this paper, we have proposed an attributed subspace clustering (ASC) approach which explores multiple attributes of data to understand data from various aspects. Different from existing subspace clustering approaches that seek for a single self-representation based on original data, ASC simultaneously learns multiple self-representations with each one corresponding to one attribute and obtains an aggregated self-representation by adding them together. Extensive experiments on six image benchmarks have clearly shown that ASC not only achieves multiple clustering solutions with each one reflecting one property of data, but also improves the clustering accuracy based on an aggregated self-representation. Though ASC has shown the effectiveness for each dataset with the same number of attributes being fixed in our experiments, nevertheless, investigating the optimal number of attributes for each dataset with model selection which could potentially enable ASC to be more desirable and practical. This could be our future work.

Acknowledgements

This work was partially supported by JSPS KAKENHI (No. 19H01114), JST AIP and JSPS KAKENHI (No. 18J12201).

References

[Bartels and Stewart, 1972] Richard H. Bartels and GW Stewart. Solution of the matrix equation ax+ xb= c [f4]. *Communications of the ACM*, 15(9):820–826, 1972.

[Cao *et al.*, 2015] Xiaochun Cao, Changqing Zhang, Huazhu Fu, Si Liu, and Hua Zhang. Diversity-induced multi-view subspace clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–594, 2015.

[Changpinyo *et al.*, 2013] Soravit Changpinyo, Kuan Liu, and Fei Sha. Similarity component analysis. In *Advances in Neural Information Processing Systems*, pages 1511–1519, 2013.

[Elhamifar and Vidal, 2009] 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.

[Fang *et al.*, 2015] Xiaozhao Fang, Yong Xu, Xuelong Li, Zhihui Lai, and Wai Keung Wong. Robust semi-supervised subspace clustering via non-negative low-rank representation. *IEEE transactions on cybernetics*, 2015.

[Favaro *et al.*, 2011] Paolo Favaro, René Vidal, and Avinash Ravichandran. A closed form solution to robust subspace estimation and clustering. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 1801–1807. IEEE, 2011.

[Georghiades *et al.*, 2001] Athinodoros S Georghiades, Peter N Belhumeur, and David J Kriegman. From few to many: Illumination cone models for face recognition under variable lighting and pose. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 23(6):643–660, 2001.

[Gretton *et al.*, 2005] Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf. Measuring statistical dependence with hilbert-schmidt norms. In *International conference on algorithmic learning theory*, pages 63–77. Springer, 2005.

[Hu *et al.*, 2014] Han Hu, Zhouchen Lin, Jianjiang Feng, and Jie Zhou. Smooth representation clustering. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 3834–3841. IEEE, 2014.

[Liao *et al.*, 2018] Lizi Liao, Xiangnan He, Hanwang Zhang, and Tat-Seng Chua. Attributed social network embedding. *IEEE Transactions on Knowledge and Data Engineering*, 2018.

[Liu *et al.*, 2010] 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.

[Liu *et al.*, 2013] Guangcan Liu, Zhouchen Lin, Shuicheng Yan, Ju Sun, Yong Yu, and Yi Ma. Robust recovery of subspace structures by low-rank representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(1):171–184, 2013.

[Liu *et al.*, 2014] Junmin Liu, Yijun Chen, Jiangshe Zhang, and Zongben Xu. Enhancing low-rank subspace clustering by manifold regularization. *Image Processing, IEEE Transactions on*, 23(9):4022–4030, 2014.

[Lu *et al.*, 2012] 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 *Computer Vision–ECCV 2012*, pages 347–360. Springer, 2012.

[Lu *et al.*, 2013] 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.

[Mazaheri *et al.*, 2018] Amir Mazaheri, Boqing Gong, and Mubarak Shah. Learning a multi-concept video retrieval model with multiple latent variables. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 14(2):46, 2018.

[Niu *et al.*, 2010] Donglin Niu, Jennifer G Dy, and Michael I Jordan. Multiple non-redundant spectral clustering views. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 831–838, 2010.

[Niu *et al.*, 2014] Donglin Niu, Jennifer G Dy, and Michael I Jordan. Iterative discovery of multiple alternativeclustering views. *IEEE transactions on pattern analysis and machine intelligence*, 36(7):1340–1353, 2014.

[Ou *et al.*, 2015] Mingdong Ou, Peng Cui, Fei Wang, Jun Wang, and Wenwu Zhu. Non-transitive hashing with latent similarity components. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 895–904. ACM, 2015.

[Peng *et al.*, 2017] Xi Peng, Zhiding Yu, Zhang Yi, and Huajin Tang. Constructing the l2-graph for robust subspace learning and subspace clustering. *IEEE transactions on cybernetics*, 47(4):1053–1066, 2017.

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

[Shi and Malik, 2000] Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22(8):888–905, 2000.

[Song *et al.*, 2007] Le Song, Alex Smola, Arthur Gretton, Karsten M Borgwardt, and Justin Bedo. Supervised feature selection via dependence estimation. In *Proceedings of the 24th international conference on Machine learning*, pages 823–830. ACM, 2007.

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

[Wang *et al.*, 2018] Chang-Peng Wang, Jiang-She Zhang, Fang Du, and Guang Shi. Symmetric low-rank representation with adaptive distance penalty for semi-supervised learning. *Neurocomputing*, 316:376–385, 2018.

[Wu *et al.*, 2015] Fei Wu, Yongli Hu, Junbin Gao, Yanfeng Sun, and Baocai Yin. Ordered subspace clustering with block-diagonal priors. *Cybernetics, IEEE Transactions on*, 2015.

[Zhang and Zhou, 2010] Yin Zhang and Zhi-Hua Zhou. Multi-label dimensionality reduction via dependence maximization. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 4(3):14, 2010.

[Zhang *et al.*, 2017] Changqing Zhang, Qinghua Hu, Huazhu Fu, Pengfei Zhu, and Xiaochun Cao. Latent multi-view subspace clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4279–4287, 2017.

[Zhou *et al.*, 2014] Chengju Zhou, Changqing Zhang, Xuewei Li, Gaotao Shi, and Xiaochun Cao. Video face clustering via constrained sparse representation. In *Multimedia and Expo (ICME), 2014 IEEE International Conference on*, pages 1–6. IEEE, 2014.