Orderly Subspace Clustering

Jing Wang,[1] **Atsushi Suzuki,**[1] **Linchuan Xu,**[1,*] **Feng Tian,**[2] **Liang Yang,**[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]School of Artificial Intelligence, Hebei University of Technology, China
{jing_wang, atsushi_suzuki, linchuan_xu, yamanishi}@mist.i.u-tokyo.ac.jp, ftian@bournemouth.ac.uk, yangliang@vip.qq.com

Abstract

Semi-supervised representation-based subspace clustering is to partition data into their underlying subspaces by finding effective data representations with partial supervisions. Essentially, an effective and accurate representation should be able to uncover and preserve the true data structure. Meanwhile, a reliable and easy-to-obtain supervision is desirable for practical learning. To meet these two objectives, in this paper we make the first attempt towards utilizing the orderly relationship, such as the data a is closer to b than to c, as a novel supervision. We propose an orderly subspace clustering approach with a novel regularization term. OSC enforces the learned representations to simultaneously capture the intrinsic subspace structure and reveal orderly structure that is faithful to true data relationship. Experimental results with several benchmarks have demonstrated that aside from more accurate clustering against state-of-the-arts, OSC interprets orderly data structure which is beyond what current approaches can offer.

Introduction

Subspace clustering, which aims at partitioning high dimensional data into multiple low-dimensional subspaces with each subspace corresponding to one class or subject, has been widely used in machine learning, computer vision and pattern recognition area (Rao et al. 2010; Liu et al. 2013; Zhou et al. 2014; Zhang et al. 2018).

Recently, several subspace clustering methods have been proposed to explore the relationships among data with self-representation (Wu et al. 2015; Zhang et al. 2017). These representation-based methods can be roughly divided into two categories: the unsupervised and the semi-supervised. The former (Liu, Lin, and Yu 2010; Lu et al. 2012; Hu et al. 2014) mainly focuses on the intrinsic structure of data, while the later (Zhou et al. 2014; Fang et al. 2015; Wang et al. 2017a) utilizes a small amount of supervision information on top of data structure as a valuable guidance. The pairwise information, including must-link constraints and cannot-link constraints which specify whether the data must be or cannot be in the same cluster, has been demonstrated that it helps to improve performance (Wagstaff et al. 2001). However, the pairwise relationships only characterize the relationships between data and cluster, yet the relationships among data are far beyond what they can describe. Also, it is often difficult to identify whether two objects must or cannot belong to the same cluster if there is no cluster categories known in advance. To remedy this, we are inspired to seek for and utilize novel supervision information, so as to exploit data information more deeply and further achieve more accurate clustering.

In reality, one may easily observe that orderly relationships are ubiquitous among data. An orderly relationship represents that the data a is closer to b than to c. For example, an image of "apple" is often more related to that of "banana" than to that of "ball"; a frame of a video sequence is often more related to its neighbouring frames than to those far way. Such relative order naturally exists among data and it is reliable (Kendall 1948; Amid and Ukkonen 2015; Wang et al. 2018b). Preserving such relative relationships enables us to find a good representation that uncovers the true data relationship. With these advantages, the orderly relationship has been widely used in various applications, such as ordinal classification (Liu, Tsang, and Müller 2017), ordinal embedding (Terada and Luxburg 2014), hashing (Liu et al. 2016) and social networks (Song, Meyer, and Tao 2015; Wang et al. 2017b). Unfortunately, existing subspace clustering approaches have largely ignored the orderly relationship.

In this paper, we propose an orderly subspace clustering (OSC) approach by using the orderly relationship as a novel constraint to learn orderly representations. Compared with pairwise subspace clustering approaches, OSC works quite differently and is more advanced in three aspects. Firstly, the must-link/cannot-link is limited to paired data which belong to the same class or two different ones, but the orderly relationship is a triplet information among three data within one class or across multiple ones. Secondly, with must-link/cannot-link, the pairwise subspace clustering enforces representations to be close/far away, while OSC takes a step further by preserving relative order among data, even when they are from the same cluster. Thirdly, two pairwise relations may derive an orderly relationship, such as if the data a is must-link to b but cannot-link to c, we can derive that a is more related to b than to c. However, the reverse derivation may not hold. That is to say, the pairwise rela-

*Corresponding author

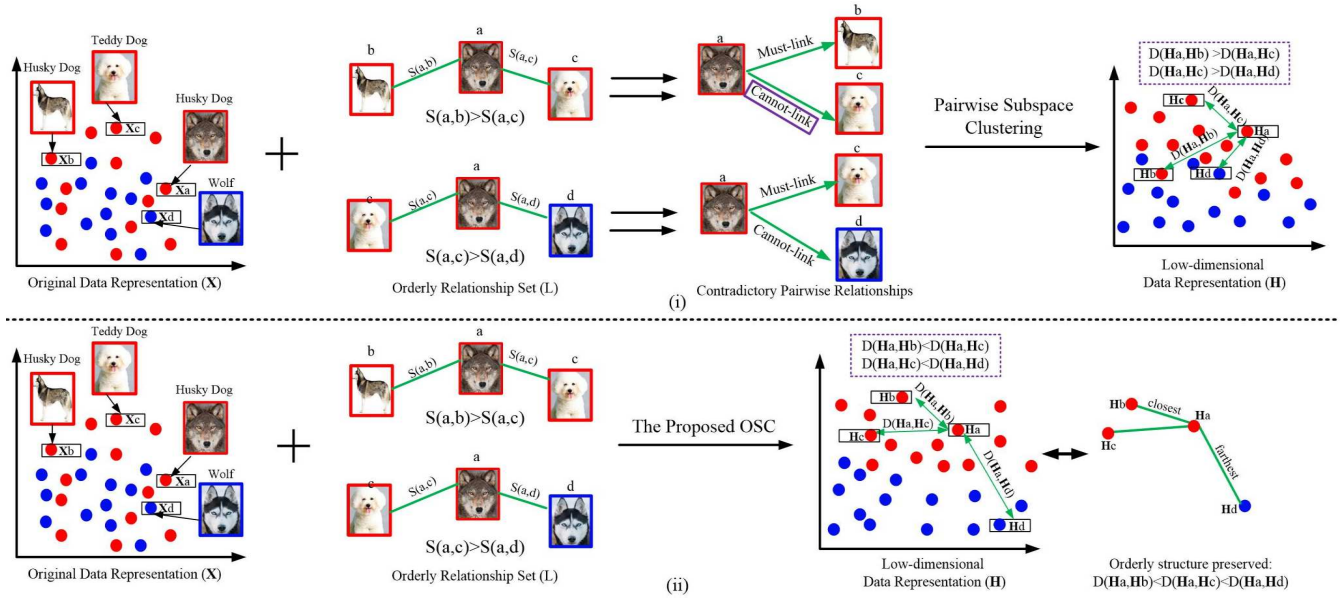


Figure 1: The comparison of existing pairwise subspace clustering methods and the proposed OSC for finding representations in low-dimensional spaces. The original dataset contains images of two classes, where the dog images (e.g., a, b and c) are in red and the wolves (e.g., d) are in blue. In particular, both a and b are Husky dogs and c is a Teddy dog. Existing pairwise approaches are infeasible to utilize orderly relationship as supervisions for semi-supervised learning, but OSC incorporates the orderly relationship effectively and enforces the learned representations to have the orderly structure which is faithful to the relationship.

tions could be utilized by OSC, but the pairwise approaches cannot incorporate the orderly relations effectively. We outline these differences and OSC's advantages in Figure 1. As shown in 1(i), given a set of orderly relationships such as a is closer to b than to c, i.e., $S(a,b) > S(a,c)$, an intuitive way to incorporate the relationship by existing pairwise approaches is to split the orderly relationship into two pairwise relations: (1) a and b are must-link and (2) a and c are cannot-link. However, this implies that a and c belong to different classes, which is incorrect. And from the relationship $S(a,c) > S(a,d)$, it can be derived that a is must-link to c and cannot-link to d, which is correct. However, while enforcing the representation of data a, i.e., $\mathbf{H}_a$ to be close to $\mathbf{H}_c$ and far away from $\mathbf{H}_d$), the pairwise approaches are not able to quantify the distance among their corresponding representations. As a result, it may end up with the relation $D(\mathbf{H}_a, \mathbf{H}_c) > D(\mathbf{H}_a, \mathbf{H}_d)$, which is apparently inconsistent with the true data relationships. This error may likely lead to a wrong clustering, e.g. it is not *b* but *d* which is clustered into the "*Dog*". In contrast, as shown in Figure 1(ii), OSC enforces representations to be faithful to the true data relationship, such as it learns $D(\mathbf{H}_a, \mathbf{H}_b) < D(\mathbf{H}_a, \mathbf{H}_c)$ with $S(a,b) > S(a,c)$. Moreover, from Figure 1(ii), we can easily deduce $S(a,b) > S(a,c) > S(a,d)$. Accordingly, $\mathbf{H}_a$ is enforced to be the closest to $\mathbf{H}_b$ and the furthest from $\mathbf{H}_d$, i.e., $D(\mathbf{H}_a, \mathbf{H}_b) < D(\mathbf{H}_a, \mathbf{H}_c) < D(\mathbf{H}_a, \mathbf{H}_d)$. As a result, a more discriminative representation matrix $\mathbf{H}$ is achieved with orderly structure preserved.

The main contribution of our work is that, to our best knowledge, OSC is the first approach to incorporate the orderly relationship as a novel supervision into subspace clustering for semi-supervised learning. Due to the natural availability of the orderly relationship, OSC is practical for real applications. With the orderly relationship, OSC ranks the learned representations effectively according to the true structure of data so that more discriminative representations are obtained. In fact, this strategy is not limited to one specific approach only but can also be easily adopted by several existing representation-based subspace clustering approaches such as SSC and LRR. Extensive experiments on several benchmark datasets have demonstrated the effectiveness of OSC against state-of-the-arts, not only in terms of accuracy, but also the interpretation of orderly data structure.

Related Work

Several representation-based subspace clustering methods have been proposed. To a large extent, unsupervised approaches all tend to exploit intrinsic data structure by adding regularization terms on the representations under different assumptions. For example, the sparse subspace clustering (SSC) (Elhamifar and Vidal 2009) seeks the sparse solution of data representation, which tends to be block diagonal. The low-rank representation (LRR has two versions: LRR_1 and $\text{LRR}_{2,1}$) (Liu, Lin, and Yu 2010) aims to take the correlation structure of data into account for finding a low-rank representation instead of a sparse representation. The least squares regression (LSR) (Lu et al. 2012) is more effective than than SSC and LRR for subspace clustering. It is

also efficient due to its closed form solution. 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.

To seek for more effective representations, several semi-supervised approaches have been proposed by incorporating pairwise information or labels as supervisions with graph-based regularization. In particular, a graph consists of "nodes" (data) and "edges" that indicate the relationships of data. If two data points are of must-link, a large positive weight is assigned to the edge. Otherwise, a non-positive weight is assigned. The graph is then incorporated into the objective function as a regularizer. CS-VFC (Zhou et al. 2014) incorporates such a graph into SSC to explore the unknown relationships among data, followed by adding the constraints directly to the affinity matrix. Since the block-diagonal structure is heavily desired for accurate sample clustering, (Feng et al. 2014) proposed a graph Laplacian constraint based SSC (BD-SSC) to construct exactly block-diagonal affinity matrices. NNLRR (Fang et al. 2015) employs the graph to encode label information to seek low-rank and sparse representation simultaneously with nonnegativity constraints on the matrices. Later, LRRADP (Wang et al. 2018a) was then proposed by using adaptive distance penalty to construct an affinity graph, which enforces the representations of every two consecutive neighboring data to be similar. However, none of existing approaches can capture the true data relationships, though the reflection of the true data structure is in essence for an effective representation.

Orderly Subspace Clustering (OSC)

Preliminary

Given a set of data points $\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n] \in \mathbb{R}^{m\times n}$, where each data vector $\mathbf{X}_i (1 \leq i \leq n)$ is a m-dimensional vector and n is the number of data. In order to cluster the data into their respective subspaces, we need to find effective representations for constructing an affinity matrix. 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., $\mathbf{X} = \mathbf{XH} + \mathbf{E}$, where $\mathbf{H} = [\mathbf{H}_1, \mathbf{H}_2, \ldots, \mathbf{H}_n] \in \mathbb{R}^{n\times n}$ is the data representation matrix and $\mathbf{E} \in \mathbb{R}^{m\times n}$ denotes errors. It can be formulated as the following optimization problem to compute the optimal data representation matrix $\mathbf{H}^*$:

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

where α is the tradeoff parameter, $\Theta(\mathbf{E})$ is the noise term. $\Omega(\mathbf{X}, \mathbf{H})$ and $\mathcal{T}$ are the regularizer and constraint set on $\mathbf{H}$, respectively. Under different assumptions, existing methods differ mainly in the choice of norms[1] for the noise term and regularization on $\mathbf{H}$.

As one of the most representative subspace clustering method, LSR (Lu et al. 2012) solves the following objective function:

$$\min_{\mathbf{H}} \|\mathbf{E}\|_F^2 + \alpha\|\mathbf{H}\|_F^2, \quad s.t. \mathbf{X} = \mathbf{XH} + \mathbf{E}. \tag{2}$$

Obviously, the representation learning process of LSR depends on the intrinsic data structure only. Since the orderly relationship reflects the true data relationships, which if incorporated, is of much help for more accurate learning. In the following we present our OSC which embeds orderly relationships to achieve more discriminative representations.

OSC-model

We use $S(a, b) > S(a, c)$ to denote an orderly relationship, i.e., the data a is closer/more similar to b than to c. The objective of OSC is to enforce the distances among representations of the data are in correspondence to the orderly relationship, i.e.,

$$S(a, b) > S(a, c) \Rightarrow D(\mathbf{H}_a, \mathbf{H}_b) < D(\mathbf{H}_a, \mathbf{H}_c), \tag{3}$$

where D denotes the distance measure. To achieve so, we minimize the following term:

$$I(S_+(a, b))I(D(\mathbf{H}_a, \mathbf{H}_b) \geq D(\mathbf{H}_a, \mathbf{H}_c))I(S_-(a, c)). \tag{4}$$

Here, $S_+(\cdot)/S_-(\cdot)$ indicates a closer/less close relationship between two data. $I(\cdot)$ is an indicator which equals to "1" if the condition in the parenthesis is satisfied and "0" otherwise. It is clear to see that given $S(a, b) > S(a, c)$, i.e., $I(S_+(a, b)) = I(S_-(a, c)) = 1$, minimizing (4) enforces $D(\mathbf{H}_a, \mathbf{H}_b) < D(\mathbf{H}_a, \mathbf{H}_c)$.

Use L to represent a set of orderly relationships, then (4) could be extended for $\forall(a, b, c) \in L$ as follows,

$$\sum_{(a,b,c)\in L} I(S_+(a, b))I(D(\mathbf{H}_a, \mathbf{H}_b) \geq D(\mathbf{H}_a, \mathbf{H}_c))I(S_-(a, c)). \tag{5}$$

Using Euclidean distance (Liu et al. 2018) for D, we can rewrite (5) as

$$\sum_{(a,b,c)\in L} I(S_+(a, b))I(\|\mathbf{H}_a - \mathbf{H}_b\|_2^2 \geq \|\mathbf{H}_a - \mathbf{H}_c\|_2^2)I(S_-(a, c)). \tag{6}$$

To ensure a distance between $\|\mathbf{H}_a - \mathbf{H}_b\|_2^2$ and $\|\mathbf{H}_a - \mathbf{H}_c\|_2^2$, a tunable threshold $\delta > 0$ is incorporated to regulate the distances in-between as

$$\sum_{(a,b,c)\in L} I(S_+(a, b))I(\|\mathbf{H}_a - \mathbf{H}_b\|_2^2 > (\|\mathbf{H}_a - \mathbf{H}_c\|_2^2 - \delta))I(S_-(a, c)). \tag{7}$$

Because (7) is non-continuous, we use a ReLU loss function $f(t) = \max(0, t)$ and obtain our orderly structured term

$$\sum_{(a,b,c)\in L} I(S_+(a, b)) \max(0, t)I(S_-(a, c)), \tag{8}$$

[1] In this paper, Frobenius norm, L_2 norm, $L_{2,1}$ norm and nuclear norm are represented by $\|\cdot\|_F$, $\|\cdot\|_2$, $\|\cdot\|_{2,1}$ and $\|\cdot\|_*$, respectively.

where $t = \|\mathbf{H}_a - \mathbf{H}_b\|_2^2 - (\|\mathbf{H}_a - \mathbf{H}_c\|_2^2 - \delta)$.

By incorporating (8) into (2), we propose a simple yet effective objective function as follows:

$$\begin{aligned}\mathcal{F} = \min_{\mathbf{H}} & \underbrace{\|\mathbf{X} - \mathbf{X}\mathbf{H}\|_F^2}_{\text{error}} + \underbrace{\alpha\|\mathbf{H}\|_F^2}_{\text{smoothness}} \\ & \underbrace{+ \beta \sum_{(a,b,c)\in L} I(S_+(a,b)) \max(0,t) I(S_-(a,c))}_{\text{orderly structure}},\end{aligned} \tag{9}$$

where β is the trade-off parameter of the orderly structured term. Note that, without orderly relationship, the objective function (9) becomes the same as that in LSR (Liu, Lin, and Yu 2010); with relationships, for each $S(a,b) > S(a,c)$, (9) penalizes t to encourage $\|\mathbf{H}_a - \mathbf{H}_b\|_2^2 - \|\mathbf{H}_a - \mathbf{H}_c\|_2^2 \to -\delta$ for $t > 0$ or maintains the structure for $t < 0$.

Remarks. It is worth pointing that the orderly structured term is not limited to LSR only. In this paper we base OSC on LSR simply because LSR not only achieves effective performance, but also has been proved efficient due to its closed form solution, which makes our approach more practical. In fact, any other representation-based subspace clustering approach which satisfies (1) can also be extended to the semi-supervised setting with the orderly structure term. For example, $\text{LRR}_{2,1}$ can be directly extended by incorporating the term as

$$\begin{aligned}\min_{\mathbf{H}} & \|\mathbf{X} - \mathbf{X}\mathbf{H}\|_{2,1} + \alpha\|\mathbf{H}\|_* \\ & + \beta \sum_{(a,b,c)\in L} I(S_+(a,b)) \max(0,t) I(S_-(a,c)).\end{aligned}$$

Such extension unifies the process of low rank representation learning with the orderly structure preserving so that more accurate performances are expected, which would be our future work.

Optimizations

We rewrite the optimization problem (9) as

$$\begin{aligned}\mathcal{F} = & \min_{\mathbf{H}_u,\mathbf{H}_a,\mathbf{H}_b,\mathbf{H}_c} \sum_{u\notin L} \|\mathbf{X}_u - \mathbf{X}\mathbf{H}_u\|_2^2 + \sum_{a\in L} \|\mathbf{X}_a - \mathbf{X}\mathbf{H}_a\|_2^2 \\ & + \sum_{b\in L} \|\mathbf{X}_b - \mathbf{X}\mathbf{H}_b\|_2^2 + \sum_{c\in L} \|\mathbf{X}_c - \mathbf{X}\mathbf{H}_c\|_2^2 + \alpha \sum_{u\notin L} \|\mathbf{H}_u\|_2^2 \\ & + \alpha \sum_{a\in L} \|\mathbf{H}_a\|_2^2 + \alpha \sum_{b\in L} \|\mathbf{H}_b\|_2^2 + \alpha \sum_{c\in L} \|\mathbf{H}_c\|_2^2 \\ & + \beta \sum_{(a,b,c)\in L} I(S_+(a,b)) \max(0,t) I(S_-(a,c)),\end{aligned} \tag{10}$$

where $\mathbf{X}_u$ indicates a data vector without prior information and $\mathbf{H}_u$ is the corresponding representation. Since (10) is non-convex, it is non-trivial to find the global minimum. Here we divide (10) into several subproblems for alternately updating each subproblem with the others fixed.

The optimization of **$\mathbf{H}_u$-subproblem** leads to the standard LSR formulation (Lu et al. 2012), so we obtain

$$\mathbf{H}_u = (\mathbf{X}^T\mathbf{X} + \alpha\mathbf{I})^{-1}\mathbf{X}^T\mathbf{X}_u, \tag{11}$$

where $\mathbf{I}$ is an identity matrix.

$\mathbf{H}_a$-subproblem: Updating $\mathbf{H}_a$ with other subproblems fixed leads to

$$\begin{aligned}\min_{\mathbf{H}_a} & \mathcal{F}(\mathbf{H}_a) = \|\mathbf{X}_a - \mathbf{X}\mathbf{H}_a\|_2^2 \\ & + \alpha\|\mathbf{H}_a\|_2^2 + \beta \sum_{(b,c)\in L} I(S_+(a,b)) \max(0,t) I(S_-(a,c)).\end{aligned} \tag{12}$$

Since $\max(0,t)$ is non-continuously differentiable with respect to $\mathbf{H}_a$, we relax the optimization by replacing $\max(0,t)$ by $I(t>0)t$ and solve the modified problem instead of directly optimizing (13). This modification leads to the minimization of the following function:

$$\begin{aligned}\min_{\mathbf{H}_a} & \mathcal{F}(\mathbf{H}_a) = \|\mathbf{X}_a - \mathbf{X}\mathbf{H}_a\|_2^2 \\ & + \alpha\|\mathbf{H}_a\|_2^2 + \beta \sum_{(b,c)\in L} I(t>0) I(S_+(a,b)) t I(S_-(a,c)).\end{aligned} \tag{13}$$

Differentiating $\mathcal{F}(\mathbf{H}_a)$ with respect to $\mathbf{H}_a$ and setting it to zero, we get

$$\begin{aligned}& \mathbf{X}^T\mathbf{X}_a - \mathbf{X}^T\mathbf{X}\mathbf{H}_a + \alpha\mathbf{H}_a \\ & + \beta \sum_{(b,c)\in L} I(t>0) I(S_+(a,b))(\mathbf{H}_c - \mathbf{H}_b) I(S_-(a,c) = 0.\end{aligned} \tag{14}$$

This is the fixed point equation, and we can get the closed form solution as

$$\begin{aligned}\mathbf{H}_a = & (\mathbf{X}^T\mathbf{X} + \alpha\mathbf{I})^{-1}(\mathbf{X}^T\mathbf{X}_a \\ & + \beta \sum_{(b,c)\in L} I(t>0) I(S_+(a,b))(\mathbf{H}_b - \mathbf{H}_c) I(S_-(a,c)).\end{aligned} \tag{15}$$

Similarly, we obtain the closed solutions for **$\mathbf{H}_b$-subproblem** and **$\mathbf{H}_c$-subproblem** as

$$\begin{aligned}\mathbf{H}_b = & (\mathbf{X}^T\mathbf{X} + \alpha\mathbf{I} + \beta \sum_{(a,c)\in L} I(t>0) I(S_+(a,b)\mathbf{I} I(S_-(a,c)))^{-1} \\ & \cdot (\mathbf{X}^T\mathbf{X}_b + \beta \sum_{(a,c)\in L} I(t>0) I(S_+(a,b)\mathbf{H}_a I(S_-(a,c))),\end{aligned} \tag{16}$$

and

$$\begin{aligned}\mathbf{H}_c = & (\mathbf{X}^T\mathbf{X} + \alpha\mathbf{I} - \beta \sum_{(a,b)\in L} I(t>0) I(S_+(a,b)\mathbf{I} I(S_-(a,c)))^{-1} \\ & \cdot (\mathbf{X}^T\mathbf{X}_c + \beta \sum_{(a,b)\in L} I(t>0) I(S_+(a,b)\mathbf{H}_a I(S_-(a,c))).\end{aligned} \tag{17}$$

Since the alternating minimization relies on the comparison between the initial values of t and 0, it is important to have a sensible initialization. Here we initialize the whole representations matrix $\mathbf{H}$ using LSR, which is a special case of our method (when $\beta = 0$ in (9)).

After obtaining representations of each vector, the representation matrix $\mathbf{H}^*$ is formed. The affinity matrix is then constructed and NCuts (Shi and Malik 2000) is applied onto the affinity matrix to produce the final clustering results. The whole procedure of OSC is summarized in Algorithm 1.

Experiments

To demonstrate the effectiveness of OSC, we conducted experiments on a variety of datasets, including image datasets

Table 1: Clustering accuracies (%) on Extended Yale B

Methods	k-NN	SSC	LRR_1	$\text{LRR}_{2,1}$	LSR	CASS	SMR	TSC	CS-VFC	SemiLSR	OSC
5 subjects	71.56	92.19	81.88	86.56	92.19	94.03	85.31	89.69	88.81	92.31	**94.06**
10 subjects	49.59	63.97	60.56	65.00	73.59	81.88	72.50	62.19	74.34	75.24	**87.19**
30 subjects	52.59	50.19	58.23	61.24	58.38	N/A	56.94	56.20	63.57	62.65	**66.67**
38 subjects	47.53	45.89	55.11	57.39	57.73	N/A	56.88	54.18	54.35	58.21	**62.80**

Table 2: Clustering accuracies (%) on USPS

Methods	k-NN	SSC	LRR_1	$\text{LRR}_{2,1}$	LSR_1	LSR	CASS	SMR	TSC	CS-VFC	SemiLSR	OSC
10 subjects	77.70	73.72	74.40	74.40	72.40	72.20	82.40	84.20	73.10	84.57	85.52	**88.00**

Algorithm 1 Solving OSC

Input: data matrix $\mathbf{X}$, orderly relationship set L, number of subspaces p, parameters λ, β and δ.
Output: Clustering result.
Initialize $\mathbf{H}$ by solving (2).
repeat
  for $i = 1 : n$ **do**
    Obtain representations for each $\mathbf{X}_i$ by solving (11), (15), (16) and (17), accordingly.
  end for
until max iteration times or convergence
Construct the affinity matrix by $(|\mathbf{H}^*| + |\mathbf{H}^{*T}|)/2$.
Segment the data into p groups by NCuts.

(Extended-Yale B, USPS), a document dataset (20newsgroups) and a motion sequence (Hopkins 155).

Extended Yale B is a face clustering dataset which contains 2414 frontal face images of 38 subjects, with approximately 64 images per subject taken under different illumination conditions. We chose different numbers of subjects in the experiment, ranging from the first 5, 10, 30, to 38 (i.e., all of the subjects in the dataset).

USPS contains 9298 handwritten digit images (16×16 each). It consists of 10 classes corresponding to the 10 digits, $0 \sim 9$. We used the first 100 examples of each digit for this experiment.

20 newsgroups is a collection of approximately 20,000 newsgroup documents, partitioned evenly across 20 different newsgroups. Following (Breitenbach and Grudic 2005), we used 3970 document vectors which belong to four topics in an 8014-dimensional space.

Hopkins 155 contains 155 motion sequences, each of which contains two or three motions (one motion corresponds to one subspace). As in (Hu et al. 2014), we used PCA to project the data into a 12-dimensional subspace.

Experimental settings

We compared OSC with several unsupervised baselines: k-NN using heat kernel distance, SSC (Elhamifar and Vidal 2009), LRR (LRR_1 and $\text{LRR}_{2,1}$) (Liu et al. 2013), LSR (Lu et al. 2012), CASS (Lu et al. 2013), SMR (Hu et al. 2014) and TSC (Heckel and Bölcskei 2015). Since existing semi-supervised approaches and OSC utilize different supervision information, it is practically infeasible to compare them. Nevertheless, to experimentally demonstrate that the orderly relationship is infeasible to be implemented by pairwise approaches (as in Figure 1), we split each orderly relation into a pair of must-link and cannot-link and applied the pair onto an existed approach CS-VFC and a constructed approach SemiLSR[2]. The parameters for each compared method were tuned to achieve the best performance for comparison. For OSC, we varied the regularization parameters α and β within [0.1, 0.2, 0.3, 0.4, 0.5] and [0.01, 0.02, 0.03, 0.04, 0.05], respectively, and δ was fixed to be 0.002 for all the experiments. To construct orderly relations for OSC, we first randomly selected 10% data for each dataset, and then constructed 30 orderly relations for each selected data as in (Chang et al. 2014). Without losing generality, the orderly relations were constructed in two days. For image and document datasets, we used the labels because they can be directly used for constructing orderly relationships, although the relationships can also be formed by observing similarities among images/documents. In particular, one half of the orderly relationships were constructed with labels for data with the same label are more related than those with different ones. The other half were constructed from the q-nearest neighbouring graph since data are usually more related to their nearest neighbours than those far away and q was set as 5 according to (Gong et al. 2017). For the motion dataset, we chose the first 15 relations with each containing two frames from the same scene and one from another. Each of the second 15 relations is formed by choosing from a scene two neighbouring frames and one farther away. To ensure a fair comparison, the affinity matrices of all approaches were conducted on the typical affinity measures (Georghiades, Belhumeur, and Kriegman 2001) and NCuts (Shi and Malik 2000) was employed to produce the final clustering results. The clustering performance is evaluated by accuracy/error (Hu et al. 2014) and the best results

[2]SemiLSR is modified by incorporating pairwise information onto LSR. Following (Liu et al. 2012), we have modified $\|\mathbf{H}\|_F^2$ to $tr(\mathbf{HIH}^T)$ by setting $\mathbf{I}_{ij} = 1$ if data are of must-link, or $\mathbf{I}_{ij} = 0$ for cannot-link.

Table 3: Clustering accuracies (%) on 20 Newsgroups

Methods	k-NN	SSC	LRR_1	$\text{LRR}_{2,1}$	LSR	CASS	SMR	TSC	CS-VFC	SemiLSR	OSC
4 subjects	91.41	83.63	91.59	92.19	84.81	N/A	86.32	89.09	89.43	89.36	**93.32**

Table 4: Clustering errors (%) on Hopkins 155

Methods	k-NN	SSC	LRR_1	$\text{LRR}_{2,1}$	LSR	CASS	SMR	TSC	CS-VFC	SemiLSR	OSC
MAX	45.59	39.53	36.36	32.50	36.36	32.85	35.83	45.21	37.25	36.15	36.39
MEAN	13.44	4.02	3.23	3.13	2.84	2.42	2.27	12.04	3.24	2.25	**1.31**
STD	12.90	7.21	6.60	5.90	6.16	5.84	5.41	11.24	4.64	5.36	4.01

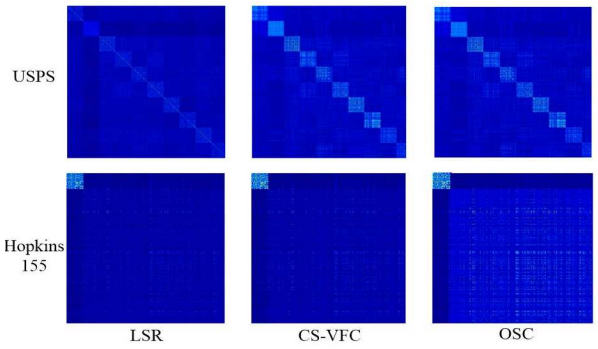


Figure 2: OSC achieves a clearer affinity matrix with more salient block diagonal than LSR and CS-VFC.

are highlighted in **boldface**. Since CASS requires a significant amount of time to process some large datasets, the corresponding results are not available but represented with N/A. All the experiments are done using Matlab 2017 in an Intel Core 2.50GHZ desktop.

Experimental results and analysis

Clustering results Tables 1 and 2 summarize the image clustering results on Extended Yale B and USPS, respectively. On both datasets, OSC not only outperforms LSR which proves the effectiveness of the proposed orderly structured term, but also achieves higher accuracies than all compared methods consistently. Specifically, Table 1 shows that OSC achieves the highest clustering accuracy on all four clustering tasks. For example, OSC outperforms the second best results (SemiLSR) by $1,75\%$ and 4.59% for the 5 and 38 subjects tasks, respectively. Besides, note that CS-VFC performs even worse than SSC in 5 subject tasks, which proves that orderly relationships cannot be effectively used by pairwise approaches. In Table 2, we can see that the clustering accuracies of the first four methods (k-NN, SSC, LRR, and LSR) and TSC are very close, fluctuating between 72.20% and 77.70%. The rest (CASS, SMR, CSVFC and SemiLSR) achieve also similar but better accuracies which vary slightly from 82.40% to 85.52%. Although these methods perform well, OSC still gets the highest accuracy of 88.00%.

Table 3 shows the clustering accuracy on 20 Newsgroups dataset. On this dataset, most approaches achieve promising performance, but OSC still outperforms the rest significantly. Table 4 tabulates the motion segmentation errors of ten methods on the Hopkins 155 database. It shows that OSC makes only 1.31% segmentation error, while the second best is 2.25% by CS-VFC. Arguably, the improvement of OSC on this dataset is moderate. This is mainly because most sequences are actually easy to segment. As a result, even with big improvements on some challenging sequences, the overall improvement is limited, as the reported error is the mean of all 156 segmentation errors.

Affinity matrix analysis. To intuitively demonstrate the clustering results, we closely visualized the corresponding derived affinity matrix, as a clearer block diagonal structure of the affinity matrix leads to higher clustering performance. Figure 2 shows the comparison of OSC with LSR and CS-VFC on a challenging sequence of Hopkins 155 datasets, as well as USPS. Clearly, the results on both figures show that the affinity matrix obtained by OSC has much more salient block diagonal structure compared with those by other methods, which undoubtedly will lead to more accurate results.

Orderly relationship analysis. To further validate the effectiveness of the orderly relationships for the interpretation of data structure, we examined $\mathbf{H}$ for LSR and OSC in Figure 3. Since the performance on each dataset has similar tendency, due to page restriction, we only show the experimental results on the first 5 clusters of Extended Yale B for the following analyses. For clearer visualization, we only demonstrate 3 groups of images and sample 10 (out of 64) images from each selected group as example. In Figure 3, the different groups are represented by different colors. Two examples of true relations among images, i.e., $S(a,b) > S(a,c)$ and $S(a,c) > S(a,d)$, were given by observing these images in terms of different subjects and expressions, respectively. All representations displayed after the dimensionalities of their features are reduced to 2-D by PCA. It can be seen that the original image $\mathbf{X}_a$ is closer to $\mathbf{X}_d$ than to $\mathbf{X}_c$, although the ground truth should be $S(a,c) > S(a,d)$. LSR however gives $D(\mathbf{H}_a,\mathbf{H}_c) > D(\mathbf{H}_a,\mathbf{H}_d)$, which experimentally proves that LSR cannot maintain the orderly structure. Consequently, it leads $\mathbf{H}_a$ to be close to $\mathbf{H_d}$ which belongs to a different group. In contrast, OSC effectively enforces $D(\mathbf{H}_a,\mathbf{H}_b) < D(\mathbf{H}_a,\mathbf{H}_c) < D(\mathbf{H}_a,\mathbf{H}_d)$ with $S(a,b) > S(a,c) > S(a,d)$. Apparently, both $\mathbf{H}_a$ and $\mathbf{H}_b$

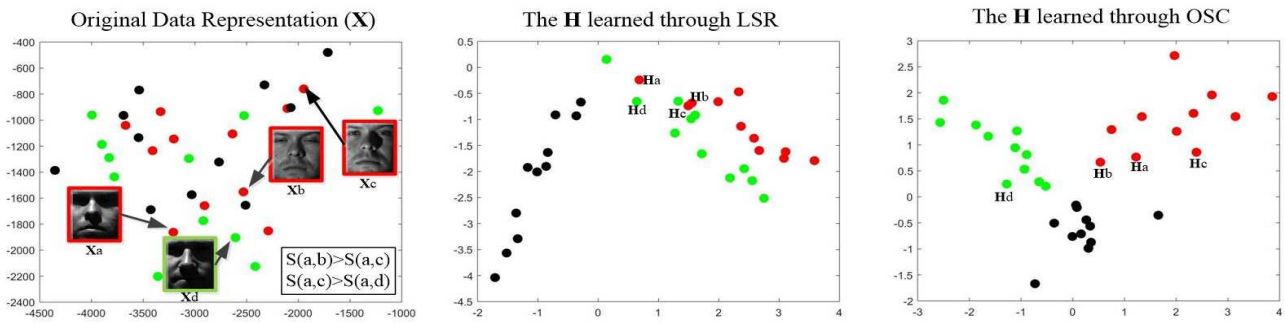
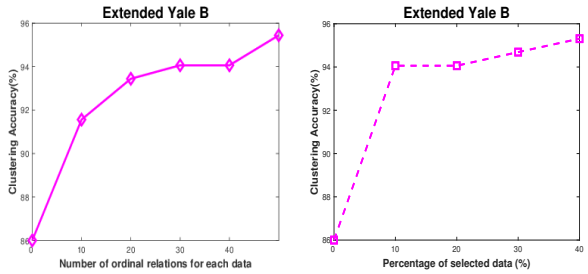


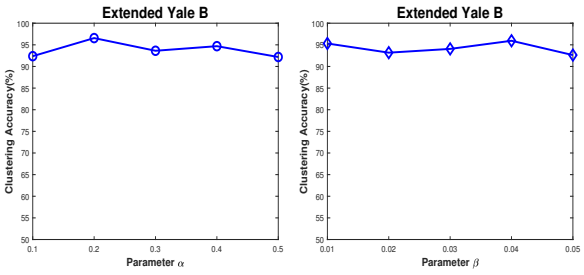
Figure 3: An example of the learned representations of LSR and OSC on the dataset Extended Yale B. LSR fails to learn orderly structure among representations, while OSC effectively cluster data into three groups with orderly structure preserved.



(a) (b)

Figure 4: Clustering accuracy of OSC $w.r.t$ orderly relationships.

represent the faces of the same subject with the same expression, which are definitely of the highest relativity. Since $\mathbf{H}_d$ represents a face of a different subject, it is the least related to $\mathbf{H}_a$. Therefore, the $\mathbf{H}$ learned through OSC demonstrates a clearer structure of data.



(a) (b)

Figure 5: Parameters analysis.

Since the number of orderly relationships may influence the construction of the affinity matrix (and thus the clustering performance), we also analyzed this in Figure 4 from two aspects. We first selected 10% data from Extended Yale B and varied the relations associated with each selected data from 0 to 50 with 10 interval as in Figure 4(a). We then varied percentages of data within $\{0\%, 1\%, 5\%, 10\%, 20\%, 50\%\}$ and fixed 30 orderly relations for each selected data in Figure 4(b). Both figures show that the clustering accuracy generally increases with more supervisions. This further proves the effectiveness of OSC on incorporating the orderly relationships.

Parameters and convergence analysis. We tested the effect of parameters α and β of OSC with δ being fixed as 0.002. First, we fixed $\beta = 0.05$ to test α with varying from 0.1 to 0.5 with 0.1 interval, and then fixed $\alpha = 0.2$ to test β varying from 0.001 to 0.005. Figure 5 shows that the clustering accuracy is relatively stable when α and β vary. This well demonstrates the robustness of OSC.

We also experimentally demonstrated the convergence of OSC 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 objective function values are non-increasing and drop sharply within 5 iterations, which empirically validates its convergence.

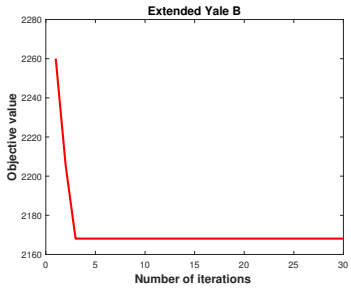


Figure 6: Convergence curves.

Conclusion

In this paper we propose orderly subspace clustering (OSC), which for the first time explores the orderly relations among data points to learn subspace representations. OSC incorporates a novel ranking-based regularizer on the representations into the learning objective to make the subspace rep-

resentations truly reflect the data structure. As a regularizer, the proposed mechanism of orderly relation preserving can also be utilized in other subspace clustering methods. We have done experiments on several benchmark datasets and demonstrated that OSC can not only preserve the true data structure which beyond what existing methods can offer, but also achieve more accurate clustering against the state-of-the-arts. In the future, we will explore a more solid analysis in theory on the preservation of the orderly relationship.

Acknowledgements

This work was supported by JST CREST (No. JPMJCR1304), JSPS KAKENHI (No. 18J12201) and National Natural Science Foundation of China (No. 61503281).

References

Amid, E., and Ukkonen, A. 2015. Multiview triplet embedding: Learning attributes in multiple maps. In *International Conference on Machine Learning*, 1472–1480.

Breitenbach, M., and Grudic, G. Z. 2005. Clustering through ranking on manifolds. In *Proceedings of the 22nd international conference on Machine learning*, 73–80. ACM.

Cao, X.; Zhang, C.; Fu, H.; Liu, S.; and Zhang, H. 2015. Diversity-induced multi-view subspace clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 586–594.

Chang, S.; Qi, G.-J.; Aggarwal, C. C.; Zhou, J.; Wang, M.; and Huang, T. S. 2014. Factorized similarity learning in networks. In *Data Mining (ICDM), 2014 IEEE International Conference on*, 60–69. IEEE.

Elhamifar, E., and Vidal, R. 2009. Sparse subspace clustering. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, 2790–2797. IEEE.

Fang, X.; Xu, Y.; Li, X.; Lai, Z.; and Wong, W. K. 2015. Robust semi-supervised subspace clustering via non-negative low-rank representation. *IEEE transactions on cybernetics*.

Feng, J.; Lin, Z.; Xu, H.; and Yan, S. 2014. Robust subspace segmentation with block-diagonal prior. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3818–3825.

Georghiades, A. S.; Belhumeur, P. N.; and Kriegman, D. J. 2001. 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.

Gong, C.; Tao, D.; Liu, W.; Liu, L.; and Yang, J. 2017. Label propagation via teaching-to-learn and learning-to-teach. *IEEE transactions on neural networks and learning systems* 28(6):1452–1465.

Heckel, R., and Bölcskei, H. 2015. Robust subspace clustering via thresholding. *IEEE Transactions on Information Theory* 61(11):6320–6342.

Hu, H.; Lin, Z.; Feng, J.; and Zhou, J. 2014. Smooth representation clustering. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, 3834–3841. IEEE.

Kendall, M. G. 1948. Rank correlation methods.

Liu, H.; Wu, Z.; Li, X.; Cai, D.; and Huang, T. S. 2012. Constrained nonnegative matrix factorization for image representation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 34(7):1299–1311.

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

Liu, H.; Ji, R.; Wu, Y.; and Liu, W. 2016. Towards optimal binary code learning via ordinal embedding. In *AAAI*, 1258–1265.

Liu, W.; Xu, D.; Tsang, I.; and Zhang, W. 2018. Metric learning for multi-output tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Liu, G.; Lin, Z.; and Yu, Y. 2010. Robust subspace segmentation by low-rank representation. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, 663–670.

Liu, W.; Tsang, I. W.; and Müller, K.-R. 2017. An easy-to-hard learning paradigm for multiple classes and multiple labels. *The Journal of Machine Learning Research* 18(1):3300–3337.

Lu, C.-Y.; Min, H.; Zhao, Z.-Q.; Zhu, L.; Huang, D.-S.; and Yan, S. 2012. Robust and efficient subspace segmentation via least squares regression. In *Computer Vision–ECCV 2012*. Springer. 347–360.

Lu, C.; Feng, J.; Lin, Z.; and Yan, S. 2013. Correlation adaptive subspace segmentation by trace lasso. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, 1345–1352. IEEE.

Rao, S.; Tron, R.; Vidal, R.; and Ma, Y. 2010. Motion segmentation in the presence of outlying, incomplete, or corrupted trajectories. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 32(10):1832–1845.

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

Song, D.; Meyer, D. A.; and Tao, D. 2015. Top-k link recommendation in social networks. In *Data Mining (ICDM), 2015 IEEE International Conference on*, 389–398. IEEE.

Terada, Y., and Luxburg, U. 2014. Local ordinal embedding. In *International Conference on Machine Learning*, 847–855.

Wagstaff, K.; Cardie, C.; Rogers, S.; Schrödl, S.; et al. 2001. Constrained k-means clustering with background knowledge. In *ICML*, volume 1, 577–584.

Wang, J.; Wang, X.; Tian, F.; Liu, C. H.; and Yu, H. 2017a. Constrained low-rank representation for robust subspace clustering. *IEEE transactions on cybernetics* 47(12):4534–4546.

Wang, S.; Tang, J.; Aggarwal, C.; Chang, Y.; and Liu, H. 2017b. Signed network embedding in social media. In *Proceedings of the 2017 SIAM International Conference on Data Mining*, 327–335. SIAM.

Wang, C.-P.; Zhang, J.-S.; Du, F.; and Shi, G. 2018a. Symmetric low-rank representation with adaptive distance penalty for semi-supervised learning. *Neurocomputing*.

Wang, J.; Tian, F.; Liu, W.; Wang, X.; Zhang, W.; and Yamanishi, K. 2018b. Ranking preserving nonnegative matrix factorization. 2776–2782.

Wu, F.; Hu, Y.; Gao, J.; Sun, Y.; and Yin, B. 2015. Ordered subspace clustering with block-diagonal priors. *Cybernetics, IEEE Transactions on*.

Zhang, C.; Hu, Q.; Fu, H.; Zhu, P.; and Cao, X. 2017. Latent multi-view subspace clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4279–4287.

Zhang, C.; Fu, H.; Hu, Q.; Cao, X.; Xie, Y.; Tao, D.; and Xu, D. 2018. Generalized latent multi-view subspace clustering. *IEEE transactions on pattern analysis and machine intelligence*.

Zhou, C.; Zhang, C.; Li, X.; Shi, G.; and Cao, X. 2014. Video face clustering via constrained sparse representation. In *Multimedia and Expo (ICME), 2014 IEEE International Conference on*, 1–6. IEEE.