

# Multi-View Multiple Clustering

Shixin Yao<sup>1</sup>, Guoxian Yu<sup>1,3,\*</sup>, Jun Wang<sup>1</sup>, Carlotta Domeniconi<sup>2</sup> and Xiangliang Zhang<sup>3</sup>

<sup>1</sup>College of Computer and Information Sciences, Southwest University, Chongqing, China

<sup>2</sup>Department of Computer Science, George Mason University, VA, USA

<sup>3</sup>CEMSE, King Abdullah University of Science and Technology, Thuwal, SA

{ysx051836, gxyu, kingjun}@swu.edu.cn, carlotta@cs.gmu.edu, xiangliang.zhang@kaust.edu.sa

## Abstract

Multiple clustering aims at exploring alternative clusterings to organize the data in meaningful groups from different perspectives. Existing multiple clustering algorithms are designed for single-view data. We assume that the *individuality* and *commonality* of multi-view data can be leveraged to generate high-quality and diverse clusterings. To this end, we propose a novel multi-view multiple clustering (MVMC) algorithm. MVMC first adapts multi-view self-representation learning to explore the individuality encoding matrices and the shared commonality matrix of multi-view data. It additionally reduces the redundancy (i.e., enhancing the individuality) among the matrices using the Hilbert-Schmidt Independence Criterion (HSIC), and collects shared information by forcing the shared matrix to be smooth across all views. It then uses matrix factorization on the individual matrices, along with the shared matrix, to generate diverse clusterings of high-quality. We further extend multiple co-clustering on multi-view data and propose a solution called multi-view multiple co-clustering (MVMCC). Our empirical study shows that MVMC (MVMCC) can exploit multi-view data to generate multiple high-quality and diverse clusterings (co-clusterings), with superior performance to the state-of-the-art methods.

## 1 Introduction

The goal of clustering is to partition samples into disjoint groups to facilitate the discovery of hidden patterns in the data. Traditional clustering algorithms are designed for single-view data. With the diffusion of the internet of things and of big data, samples can be easily collected from different sources, or observed from different views. For example, a video can be characterized using image signals and audio signals, and a given news can be reported in different languages. Objects with diverse feature views are typically called multi-view data. It's recognized that the integration of information contained in multiple views can achieve consolidated data clustering [Chao *et al.*, 2017]. Many multi-view clustering

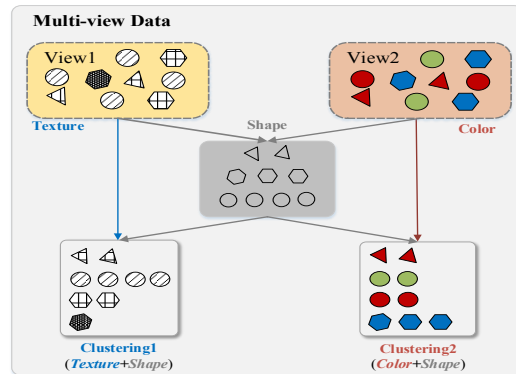


Figure 1: An example of multi-view multiple clustering. Two alternative clusterings (texture+shape and color+shape) can be generated using the *commonality* (shape) and the *individuality* (texture and color) information of the same multi-view objects.

methods have been developed to extract comprehensive information from multiple feature views; examples are co-training based [Kumar and Daumé, 2011], multiple kernel learning [Gönen and Alpaydm, 2011], and subspace learning based [Cao *et al.*, 2015; Luo *et al.*, 2018]. However, the aforementioned clustering methods typically provide a *single* clustering, which may fail to reveal the high-quality and diverse alternative clusterings of the same data.

Figure 1 shows a collection of objects represented by a texture view and a color view. We can group the objects based on their shared shapes. By leveraging the commonality and the individuality of these multi-view objects, we can obtain two alternative clusterings (*texture+shape* and *color+shape*), as shown at the bottom of the figure. From this example, we can see that multi-view data include not only the commonality information for generating high-quality clustering (as multi-view clustering does), but also the individual (or specific) information for generating diverse clusterings (as multiple clustering aims to achieve).

To explore different clusterings of the given data, multiple clustering has emerged as a new branch of clustering in recent years. Some approaches seek diverse clusterings in alternative to those already explored, by enforcing the new ones to be different [Bae and Bailey, 2006; Davidson and Qi, 2008; Yang and Zhang, 2017]; other solutions simultaneously seek

multiple clusterings by reducing their correlation [Caruana *et al.*, 2006; Jain *et al.*, 2008; Dang and Bailey, 2010; Wang *et al.*, 2018], or by seeking orthogonal (or independent) subspaces and clusterings therein [Niu *et al.*, 2010; Ye *et al.*, 2016; Mautz *et al.*, 2018; Wang *et al.*, 2019]. However, these multiple clustering methods are designed for *single-view* data.

Based on the discussed example in Figure 1, we leverage the individuality and the commonality of multi-view data to generate high-quality and diverse clusterings, and we propose an approach called multi-view multiple clustering (MVMC) to achieve this goal. To the best of our knowledge, MVMC is the *first attempt* to encompass both multiple clusterings and multi-view clustering, where the former focuses on generating diverse clusterings from a single data view, and the latter focuses on a single consensus clustering by summarizing the information from different views. MVMC first extends multi-view self-representation learning [Luo *et al.*, 2018] to explore the individuality information encoding matrices and the commonality information matrix shared across views. To obtain more credible commonality information from multiple views, we force the commonality information matrix to be smooth across all views. In addition, we use the Hilbert-Schmidt Independence Criterion (HISC) [Gretton *et al.*, 2005] to enhance the individuality between matrices, and consequently increase the diversity between clusterings. We then use matrix factorization to jointly factorize each individuality matrix (for diversity) and the commonality matrix (for quality) into a clustering indicator matrix and a basis matrix. To simultaneously seek the individual and common data matrices, and diverse clusterings therein, we use an alternating optimization technique to solve the unified objective. In addition, we extend multiple co-clustering [Tokuda *et al.*, 2017; Wang *et al.*, 2018] to the multi-view scenario, and term the extended approach as multi-view multiple co-clustering (MVMCC).

The main contributions of our work are as follows:

- We study how to generate multiple clusterings from multi-view data, which is an interesting and challenging problem, but largely overlooked. The problem we address is different from existing multi-view clustering, which generates a single clustering from multiple views, and also different from multiple clustering, which produces alternative clusterings from single-view data.
- We introduce a unified objective function to simultaneously seek multiple individuality information encoding matrices, and the commonality information encoding matrix. This unified function can leverage the individuality to generate diverse clusterings, and the commonality to boost the quality of the generated clusterings. We further adopt an alternative procedure to optimize the unified objective.
- Extensive experimental results show that MVMC (MVMCC) performs considerably better than existing multiple clustering (co-clustering) algorithms [Cui *et al.*, 2007; Niu *et al.*, 2010; Ye *et al.*, 2016; Yang and Zhang, 2017; Tokuda *et al.*, 2017; Wang *et al.*, 2018] in exploring multiple clusterings and co-clusterings.

## 2 The Proposed Methods

### 2.1 Multi-View Multiple Clustering

Suppose  $\mathbf{X}^v \in \mathbb{R}^{d_v \times n}$  denotes the feature data matrix of the  $v$ -th view, for  $n$  objects in the  $d_v$  dimensional space.  $v \in \{1, 2, \dots, m\}$ , where  $m$  is the number of views.

We aim to generate  $h$  (provided by the user) different clusterings from  $\{\mathbf{X}^v\}_{v=1}^m$  using the shared and individual information embedded in the data matrices. As an application example, one can group the same ‘customers’ (represented by different feature views) from the perspective of purchase capability, loyalty (leave or stay), and fraud (yes or no). Most multi-view clustering approaches in essence focus on the shared and complementary information of multiple data views to generate a consolidated clustering [Chao *et al.*, 2017]. By viewing each subspace as a feature view, an intuitive solution to explore multiple clusterings on multi-view data is to first concatenate different feature views, and then apply subspace-based multiple clustering methods on the concatenated feature vectors [Niu *et al.*, 2010; Ye *et al.*, 2016; Wang *et al.*, 2019].

To find high-quality and diverse multiple clusterings, we should make concrete use of the individuality and commonality of multi-view data. The individuality helps to explore diverse clusterings, while the commonality coordinates the diverse clusterings to capture the common knowledge of multi-view data. To explore the individuality and commonality of multi-view data, we extend the multi-view self-representation learning [Cao *et al.*, 2015; Luo *et al.*, 2018] as follows:

$$J_D(\{\mathbf{D}^k\}_{k=1}^h, \mathbf{U}) = \sum_{k=1}^h \frac{1}{m} \sum_{v=1}^m \|\mathbf{X}^v - \mathbf{X}^v(\mathbf{U} + \mathbf{D}^k)\|_F^2 + \lambda_1 \Phi_1(\{\mathbf{D}^k\}_{k=1}^h) + \lambda_2 \Phi_2(\mathbf{U}) \quad (1)$$

where  $\mathbf{U} \in \mathbb{R}^{n \times n}$  is specified to encode the commonality of the data matrices  $\{\mathbf{X}^v\}_{v=1}^m$ , and  $\mathbf{D}^k \in \mathbb{R}^{n \times n}$  is used to encode the individuality of the  $k$ -th ( $k \in \{1, 2, \dots, h\}$ ) group of views.  $\Phi_1(\{\mathbf{D}^k\}_{k=1}^h)$  and  $\Phi_2(\mathbf{U})$  (defined later) are two constraints used to enhance the individuality and commonality. The multi-view self-representation learning in [Cao *et al.*, 2015; Luo *et al.*, 2018] requires  $h = m$ . In contrast, Eq. (1) does not have this requirement. As a result, the group-wise individuality and diversity are jointly considered, and the number of alternative clusterings can be adjusted by the user. The assumption for the linear representation in Eq. (1) is that a data sample can be expressed as a linear combination of other samples in the subspace. This assumption is widely-used in sparse representation [Wright *et al.*, 2010] and low-rank representation learning [Liu *et al.*, 2013].

[Luo *et al.*, 2018] recently combined  $\mathbf{U}$  and  $\{\mathbf{D}^k\}_{k=1}^m$  into an integrated co-association matrix of samples, and then applied spectral clustering to seek a consistent clustering. Their empirical study shows that the individual information encoded by  $\mathbf{D}^k$  helps producing a robust clustering. However, since  $\mathbf{X}^v$  and  $\mathbf{X}^{v'}$  ( $v' \neq v$ ) describe the same objects using different types of features, the matrix  $\mathbf{D}^k$  resulting from Eq. (1) may still have a large information overlap with  $\mathbf{D}^{k'}$ . As a

result, the expected individuality of  $\mathbf{D}^k$  and  $\mathbf{D}^{k'}$  cannot be guaranteed. This information overlapping is not necessarily an issue for multi-view clustering, which aims at finding a single clustering, but it is for our problem, where multiple clusterings of both high-quality and diversity are expected.

To enhance the diversity between the individuality encoding (representation) matrices  $\{\mathbf{D}^k\}_{k=1}^h$ , we approximately quantify the diversity based on the dependency between these matrices. The smaller the dependency between data matrices is, the larger their diversity is, since the matrices are less correlated. Various measurements can be used to evaluate the dependence between variables. Here, we adopt the Hilbert-Schmidt Independence Criterion (HSIC) [Gretton *et al.*, 2005], for its simplicity, solid theoretical foundation, and capability in measuring both linear and nonlinear dependence between variables. HSIC computes the squared norm of the cross-covariance operator over  $\mathbf{D}^k$  and  $\mathbf{D}^{k'}$  in Hilbert Space to estimate the dependency. The empirical HISC does not have to explicitly compute the joint distribution of  $\mathbf{D}^k$  and  $\mathbf{D}^{k'}$ . It is given by:

$$HSIC(\mathbf{D}^k, \mathbf{D}^{k'}) = (n-1)^{-2} \text{tr}(\mathbf{K}^k \mathbf{H} \mathbf{K}^{k'} \mathbf{H}) \quad (2)$$

where  $\mathbf{K}^k, \mathbf{K}^{k'}, \mathbf{H} \in \mathbb{R}^{n \times n}$ ,  $\mathbf{K}^k$ , and  $\mathbf{K}^{k'}$  are used to measure the kernel induced similarity between the matrices  $\mathbf{D}^k$  and  $\mathbf{D}^{k'}$  respectively.  $\mathbf{H} = \delta_{ij} - 1/n$ ,  $\delta_{ij} = 1$  if  $i = j$ ,  $\delta_{ij} = 0$  otherwise. In this paper, we adopt the inner product kernel to specify  $\mathbf{K}^k = (\mathbf{D}^k)^T \mathbf{D}^k$ ,  $\forall k \in \{1, 2, \dots, h\}$ . Then, we minimize the overall HSIC on  $h$  individuality matrices to reduce their redundancy, and specify  $\Phi_1(\mathbf{D}^k)$  as follows:

$$\begin{aligned} \Phi_1(\{\mathbf{D}^k\}_{k=1}^h) &= \sum_{k=1, k \neq k'}^h HSIC(\mathbf{D}^k, \mathbf{D}^{k'}) \\ &= \sum_{k=1, k \neq k'}^h (n-1)^{-2} \text{tr}(\mathbf{D}^k \mathbf{H} \mathbf{K}^{k'} \mathbf{H} (\mathbf{D}^k)^T) \\ &= \text{tr}(\mathbf{D}^k \tilde{\mathbf{K}}^k (\mathbf{D}^k)^T) \end{aligned} \quad (3)$$

where  $\tilde{\mathbf{K}}^k = (n-1)^{-2} \sum_{k'=1, k' \neq k}^h \mathbf{H} \mathbf{K}^{k'} \mathbf{H}$ .

Inspired by subspace-based multi-view learning [Gao *et al.*, 2015; Chao *et al.*, 2017] and manifold regularization [Belkin *et al.*, 2006], we specify  $\Phi_2(\mathbf{U})$  in Eq. (1) to collect more shared information from multiple data views as follows:

$$\Phi_2(\mathbf{U}) = \sum_{v=1}^V \sum_{i,j=1}^n \|\mathbf{u}_i - \mathbf{u}_j\|_2^2 \mathbf{W}_{ij}^v = \text{tr}(\mathbf{U} \tilde{\mathbf{L}} \mathbf{U}^T) \quad (4)$$

where  $\mathbf{W}_{ij}^v$  is the feature similarity between  $\mathbf{x}_i^v$  and  $\mathbf{x}_j^v$ . To compute  $\tilde{\mathbf{W}}^v$ , we simply adopt an  $\epsilon = 5$  nearest neighborhood graph, and use the Gaussian heat kernel (with kernel width set to the standard deviation of the distance between samples) to quantify the similarity between neighborhood samples.  $\tilde{\mathbf{L}} = \sum_{v=1}^m (\mathbf{\Lambda}^v - \mathbf{W}^v)$ , and  $\mathbf{\Lambda}^v$  is a diagonal matrix with  $\Lambda_{ii}^v = \sum_{j=1}^n \mathbf{W}_{ij}^v$ . Minimizing  $\Phi_2(\mathbf{U})$  can guide  $\mathbf{U}$  to encode consistent and complementary information shared across views. In this way, the quality of diverse clusterings can be boosted using enhanced shared information.

Given the equivalence between matrix factorization based clustering and spectral clustering (or  $k$ -means clustering), we adopt the widely used semi-nonnegative matrix factorization [Ding *et al.*, 2010] to explore the  $k$ -th clustering on  $\mathbf{U} + \mathbf{D}^k$  as follows:

$$\mathbf{U} + \mathbf{D}^k = \mathbf{B}^k (\mathbf{R}^k)^T \quad (5)$$

where  $\mathbf{R}^k \in \mathbb{R}^{n \times r_k}$  and  $\mathbf{B}^k \in \mathbb{R}^{n \times r_k}$  ( $r_k$  is the number of sample clusters) are the clustering indicator matrix and the basis matrix, respectively. Here, the  $k$ -th clustering is generated not only with respect to  $\mathbf{D}^k$ , but also to  $\mathbf{U}$ , which encodes the shared information of multi-view data. As such, the explored  $k$ -th clustering (encoded by  $\mathbf{R}^k$ ) not only reflects the individuality of views in the  $k$ -th group, but also captures the commonality of all data views. As a consequence, a high-quality, and yet diverse clustering can be generated.

The above process first explores the individual information matrices and the shared information matrix, and then generates diverse clusterings on the data matrices. A sub-optimal solution may be obtained as a result because the two steps are performed separately. To avoid this, we advocate to simultaneously optimize  $\{\mathbf{D}^k\}_{k=1}^h$  and the diverse clusterings  $\{\mathbf{R}^k\}_{k=1}^h$  therein, and formulate a unified objective function for MVMC as follows:

$$\begin{aligned} J_{MC}(\{\mathbf{D}^k\}_{k=1}^h, \{\mathbf{R}^k\}_{k=1}^h, \mathbf{U}) &= \frac{1}{h} \sum_{k=1}^h \|\mathbf{U} + \mathbf{D}^k - \mathbf{B}^k (\mathbf{R}^k)^T\|_F^2 \\ &+ \lambda_1 \text{tr}(\mathbf{D}^k \tilde{\mathbf{K}}^k (\mathbf{D}^k)^T) + \lambda_2 \text{tr}(\mathbf{U} \tilde{\mathbf{L}} \mathbf{U}^T) \\ \text{s.t. } \mathbf{X}^v &= \mathbf{X}^v (\mathbf{U} + \mathbf{D}^k), v \in \{1, 2, \dots, m\} \end{aligned} \quad (6)$$

By solving Eq. (6), we can simultaneously obtain multiple diverse clusterings of quality by leveraging the commonality and individuality information of multiple views.

Binary matrices  $\{\mathbf{R}^k\}_{k=1}^h$  are hard to directly optimize. As such, we relax the entries of  $\{\mathbf{R}^k\}_{k=1}^h$  to nonnegative numeric values. Since Eq. (6) is not jointly convex for  $\{\mathbf{D}^k\}_{k=1}^h$ ,  $\mathbf{U}$  and  $\{\mathbf{R}^k\}_{k=1}^h$ , it is unrealistic to find the global optimal values for all the variables. Here, we solve Eq. (6) via the alternating optimization method, which alternatively optimizes one variable, while fixing the other variables. The detailed optimization process can be viewed in the supplementary file due to the limitation of space.

## 2.2 Multiple Views Multiple Co-Clusterings

Recently, multiple co-clustering algorithms have also been suggested to explore alternative co-clusterings from the same data [Tokuda *et al.*, 2017; Wang *et al.*, 2018]. Multiple co-clustering methods aim at exploring multiple two-way clusterings, where both samples and features are clustered. In contrast, multiple clustering techniques only explore diverse one-way clusterings, where only samples (or only features) are clustered. Based on the merits of matrix tri-factorization in exploring co-clusters [Wang *et al.*, 2011; 2018], we can seek multiple co-clusterings on multiple views

by optimizing an objective function as follows:

$$\begin{aligned}
 & J_{MCC}(\{\mathbf{D}^v\}_{v=1}^m, \{\mathbf{R}^v\}_{v=1}^m, \{\mathbf{C}^v\}_{v=1}^m, \mathbf{U}) \\
 &= \frac{1}{m} \sum_{v=1}^m \|\mathbf{X}^v(\mathbf{U} + \mathbf{D}^v) - \mathbf{C}^v \mathbf{S}^v (\mathbf{R}^v)^T\|_F^2 \\
 &+ \lambda_1 \text{tr}(\mathbf{D}^v \mathbf{K}(\mathbf{D}^v)^T) + \lambda_2 \text{tr}(\mathbf{U} \tilde{\mathbf{L}} \mathbf{U}^T) \\
 & \text{s.t. } \mathbf{X}^v = \mathbf{X}^v(\mathbf{U} + \mathbf{D}^v), \mathbf{C}^v \geq 0, \mathbf{R}^v \geq 0
 \end{aligned} \tag{7}$$

where  $\mathbf{C}^v \in \mathbb{R}^{d_v \times c_v}$  and  $\mathbf{R}^v \in \mathbb{R}^{n \times r_k}$  correspond to the row-cluster (grouping features) and column-cluster (grouping samples) indicator matrices of the  $h$ -th co-clustering.  $\mathbf{S}^v \in \mathbb{R}^{c_v \times r_v}$  plays the role of absorbing the different scaling factors of  $\mathbf{R}^v$  and  $\mathbf{C}^v$  to minimize the squared error. Here we fix  $h = m$  for MVMCC, since different feature views have different numbers of features. Eq. (7) can be optimized following the similar procedure for optimizing Eq. (6), which is provided in the supplementary file due to page limit.

### 3 Experimental Results and Analysis

#### 3.1 Experimental Setup

In this section, we evaluate our proposed MVMC and MVMCC on five widely-used multi-view datasets [Li *et al.*, 2015; Tao *et al.*, 2018], which are described in Table 1. The datasets have different number of views and are from different domains. *Caltech-7*<sup>1</sup> and *Caltech-20* [Li *et al.*, 2015] are two subsets of Caltech-101, which contains only 7 and 20 classes, respectively. The creation of these subsets is due to the unbalance of the number of data in each class of Caltech-101. Each sample is made of 6 views of the same image. *Mul-fea digits*<sup>2</sup> is comprised of 2,000 data points from 0 to 9 digit classes, with 200 data points for each class. There are six public features available: 76 Fourier coefficients of the character shapes, 216 profile correlations, 64 Karhunenlove coefficients, 240 pixel averages in  $2 \times 3$  windows, 47 Zernike moments, and 6 morphological features. *Wiki article*<sup>3</sup> contains selected sections from the Wikipedia’s featured articles collection. We considered only the 10 most populated categories. It contains two views: text and image. *Corel*<sup>4</sup> [Tao *et al.*, 2018] consists of 5000 images from 50 different categories. Each category has 100 images. The features are color histogram (9), edge direction histogram (18), and WT (9). *Mirflickr*<sup>5</sup> contains 25,000 instances collected from Flickr. Each instance consists of an image and its associated textual tags. To avoid noise, here we remove textual tags that appear less than 20 times in the dataset, and then delete instances without textual tags or semantic labels. This process gives us 16,738 instances.

Multiple clusterings need to quantify the quality and diversity of alternative clusterings. To measure quality, we use the widely-adopted Silhouette Coefficient (SC) and Dunn Index (DI) as the internal index. *Larger* values of SC and

Datasets	$n$	$d_v$	classes	$m$
Caltech-7	1474	[40,48,254,1984,512,928]	7	6
Caltech-20	2386	[40,48,254,1984,512,928]	20	6
Mul-fea digits	2000	[76,216,64,240,47,6]	10	6
Wiki article	2866	[128,10]	10	2
Corel	5000	[9,18,9]	50	3
Mirflickr	16738	[150,500]	24	2

Table 1: Characteristics of multi-view datasets.  $n$  is the number of samples,  $d_v$  is the dimensionality of samples, ‘classes’ is the number of ground-truth clusters, and  $m$  is the number of views.

DI indicate a *high* quality clustering. To quantify the redundancy between alternative clusterings, we use Normalized Mutual Information (NMI) and Jaccard Coefficient (JC) as external indices. *Smaller* values of NMI and JC indicate *smaller* redundancy between alternative clusterings. All these metrics have been used in the multiple clustering literature [Bailey, 2013]. The formal definitions of these metrics, omitted here to save space, can be found in [Bailey, 2013; Yang and Zhang, 2017].

#### 3.2 Discovering Multiple One-way Clusterings and Multiple Co-Clusterings

We compare the one-way multiple clusterings found by MVMC against Dec-kmeans [Jain *et al.*, 2008], MNMF [Yang and Zhang, 2017], OSC [Cui *et al.*, 2007], mSC [Niu *et al.*, 2010], ISAAC [Ye *et al.*, 2016], and MISC [Wang *et al.*, 2019]. We also compare the multiple co-clusterings found by MVMCC against MultiCC [Wang *et al.*, 2018] and MCC-NBMM [Tokuda *et al.*, 2017]. The input parameters of the comparing methods are set as the authors suggested in their papers or shared code. The parameter values of MVMC and MVMCC are  $\lambda_1 = 10$ ,  $\lambda_2 = 100$ , and  $h = 2$  for multiple one-way clusterings, and  $h = m$  for multiple co-clusterings. Since none of the existing multiple clustering algorithms can work on multiple view data, we concatenate the feature vectors of multiple view data and then run these comparing methods on the concatenated vectors to seek alternative clusterings. Our MVMC and MVMCC directly run on the multiple view data, without such feature concatenation.

We use  $k$ -means to generate the reference clustering for MNMF, and then use their respective solutions to generate two alternative clustering ( $\mathcal{C}_1, \mathcal{C}_2$ ). We downloaded the source code of MNMF, ISAAC, MultiCC, MISC, and MCC-NBMM, and implemented the other methods (Dec-kmeans, mSC, and OSC) following the respective original papers. Following the experimental protocol adopted by these methods, we quantify the average clustering quality on  $\mathcal{C}_1$  and  $\mathcal{C}_2$ , and measure the diversity between  $\mathcal{C}_1$  and  $\mathcal{C}_2$ . We fix the number of row-clusters  $r_k$  for each clustering as the respective number of classes of each dataset, as listed in Table 1. For co-clustering, we adopt a widely used technique [Monti *et al.*, 2003] to determine the number of column clusters  $c_k$ . Detailed parameter values are given in the supplementary file.

Table 2 reports the average results (of ten independent runs) and standard deviations of these comparing methods on exploring two alternative one-way clusterings with  $h = 2$ . From Table 2, we can see that MVMC often outperforms the comparing methods across different multi-view datasets, which proves the effectiveness of MVMC on exploring

<sup>1</sup><https://github.com/yeqinglee/mvdata>

<sup>2</sup><https://archive.ics.uci.edu/ml/datasets/Multiple+Features>

<sup>3</sup><http://www.svcl.ucsd.edu/projects/crossmodal/>

<sup>4</sup><http://www.cais.ntu.edu.sg/æchhoi/SVMBMAL/>

<sup>5</sup><http://press.liacs.nl/mirflickr/mirdownload.html>

		Dec-kmeans	ISAAC	MISC	MNMF	mSC	OSC	MVMC
Caltech-7	SC $\uparrow$	0.049 $\pm$ 0.002●	0.235 $\pm$ 0.011○	0.201 $\pm$ 0.002○	0.234 $\pm$ 0.001○	0.163 $\pm$ 0.008○	0.261 $\pm$ 0.004○	0.140 $\pm$ 0.002
	DI $\uparrow$	0.044 $\pm$ 0.000●	0.034 $\pm$ 0.001●	0.048 $\pm$ 0.000●	0.034 $\pm$ 0.000●	0.056 $\pm$ 0.000●	0.066 $\pm$ 0.000	0.062 $\pm$ 0.000
	NMI $\downarrow$	0.024 $\pm$ 0.000●	0.485 $\pm$ 0.023●	0.513 $\pm$ 0.016●	0.022 $\pm$ 0.000●	0.152 $\pm$ 0.002●	0.693 $\pm$ 0.015●	0.006 $\pm$ 0.000
	JC $\downarrow$	0.126 $\pm$ 0.001●	0.363 $\pm$ 0.008●	0.349 $\pm$ 0.001●	0.094 $\pm$ 0.000●	0.136 $\pm$ 0.001●	0.522 $\pm$ 0.046●	0.076 $\pm$ 0.000
Caltech-20	SC $\uparrow$	-0.124 $\pm$ 0.001●	0.085 $\pm$ 0.000○	0.036 $\pm$ 0.000○	-0.169 $\pm$ 0.000●	-0.172 $\pm$ 0.006●	0.196 $\pm$ 0.001○	0.004 $\pm$ 0.000
	DI $\uparrow$	0.026 $\pm$ 0.000●	0.035 $\pm$ 0.000●	0.033 $\pm$ 0.000●	0.009 $\pm$ 0.000●	0.028 $\pm$ 0.000●	0.056 $\pm$ 0.000●	0.183 $\pm$ 0.000
	NMI $\downarrow$	0.056 $\pm$ 0.000●	0.475 $\pm$ 0.011●	0.489 $\pm$ 0.013●	0.052 $\pm$ 0.002●	0.240 $\pm$ 0.003●	0.715 $\pm$ 0.025●	0.027 $\pm$ 0.000
	JC $\downarrow$	0.050 $\pm$ 0.001●	0.222 $\pm$ 0.002●	0.198 $\pm$ 0.002●	0.033 $\pm$ 0.001●	0.074 $\pm$ 0.001●	0.444 $\pm$ 0.004●	0.023 $\pm$ 0.000
Corel	SC $\uparrow$	0.112 $\pm$ 0.014○	-0.052 $\pm$ 0.000●	-0.070 $\pm$ 0.000●	-0.277 $\pm$ 0.000●	-0.128 $\pm$ 0.000●	0.238 $\pm$ 0.002○	-0.016 $\pm$ 0.000
	DI $\uparrow$	0.031 $\pm$ 0.001●	0.032 $\pm$ 0.000●	0.020 $\pm$ 0.000●	0.015 $\pm$ 0.000●	0.019 $\pm$ 0.000●	0.032 $\pm$ 0.000●	0.354 $\pm$ 0.000
	NMI $\downarrow$	0.643 $\pm$ 0.035●	0.204 $\pm$ 0.002●	0.209 $\pm$ 0.002●	0.092 $\pm$ 0.001●	0.394 $\pm$ 0.006●	0.762 $\pm$ 0.018●	0.070 $\pm$ 0.000
	JC $\downarrow$	0.219 $\pm$ 0.004●	0.031 $\pm$ 0.001●	0.029 $\pm$ 0.001●	0.013 $\pm$ 0.000	0.072 $\pm$ 0.001●	0.410 $\pm$ 0.013●	0.010 $\pm$ 0.000
Digits	SC $\uparrow$	-0.133 $\pm$ 0.022●	-0.001 $\pm$ 0.000●	0.061 $\pm$ 0.000	-0.076 $\pm$ 0.000●	-0.117 $\pm$ 0.000●	0.471 $\pm$ 0.013○	0.064 $\pm$ 0.000
	DI $\uparrow$	0.016 $\pm$ 0.000●	0.016 $\pm$ 0.000●	0.018 $\pm$ 0.001●	0.008 $\pm$ 0.000●	0.016 $\pm$ 0.001●	0.062 $\pm$ 0.000●	0.087 $\pm$ 0.000
	NMI $\downarrow$	0.078 $\pm$ 0.000●	0.364 $\pm$ 0.012●	0.399 $\pm$ 0.008●	0.011 $\pm$ 0.000	0.515 $\pm$ 0.022●	0.822 $\pm$ 0.028●	0.008 $\pm$ 0.000
	JC $\downarrow$	0.076 $\pm$ 0.000●	0.279 $\pm$ 0.003●	0.298 $\pm$ 0.000●	0.053 $\pm$ 0.000	0.279 $\pm$ 0.004●	0.656 $\pm$ 0.015●	0.052 $\pm$ 0.000
Wiki article	SC $\uparrow$	0.447 $\pm$ 0.016○	-0.024 $\pm$ 0.000●	-0.031 $\pm$ 0.000●	-0.029 $\pm$ 0.000●	0.108 $\pm$ 0.002○	0.418 $\pm$ 0.012○	0.066 $\pm$ 0.000
	DI $\uparrow$	0.124 $\pm$ 0.000	0.085 $\pm$ 0.000●	0.085 $\pm$ 0.001●	0.083 $\pm$ 0.000●	0.095 $\pm$ 0.001●	0.135 $\pm$ 0.001○	0.122 $\pm$ 0.000
	NMI $\downarrow$	0.803 $\pm$ 0.019●	0.042 $\pm$ 0.000●	0.041 $\pm$ 0.000●	0.006 $\pm$ 0.000	0.212 $\pm$ 0.001●	0.783 $\pm$ 0.052●	0.006 $\pm$ 0.000
	JC $\downarrow$	0.593 $\pm$ 0.006●	0.078 $\pm$ 0.001●	0.078 $\pm$ 0.000●	0.056 $\pm$ 0.001	0.113 $\pm$ 0.002●	0.535 $\pm$ 0.014●	0.052 $\pm$ 0.000
Mirflickr	SC $\uparrow$	-0.004 $\pm$ 0.000○	-0.092 $\pm$ 0.000●	-0.028 $\pm$ 0.000○	-0.058 $\pm$ 0.000●	-0.093 $\pm$ 0.000●	0.017 $\pm$ 0.000○	-0.038 $\pm$ 0.000
	DI $\uparrow$	0.061 $\pm$ 0.002●	0.062 $\pm$ 0.005●	0.071 $\pm$ 0.000●	0.053 $\pm$ 0.001●	0.064 $\pm$ 0.001●	0.059 $\pm$ 0.002●	0.173 $\pm$ 0.005
	NMI $\downarrow$	0.427 $\pm$ 0.012●	0.016 $\pm$ 0.000●	0.021 $\pm$ 0.000●	0.014 $\pm$ 0.000●	0.216 $\pm$ 0.006●	0.575 $\pm$ 0.011●	0.005 $\pm$ 0.000
	JC $\downarrow$	0.878 $\pm$ 0.022●	0.047 $\pm$ 0.000●	0.037 $\pm$ 0.000●	0.023 $\pm$ 0.000	0.073 $\pm$ 0.000●	0.368 $\pm$ 0.011●	0.022 $\pm$ 0.000

Table 2: Quality and Diversity of the various competing methods on finding multiple clusterings.  $\uparrow(\downarrow)$  indicates the direction of preferred values for the corresponding measure. ●/○ indicates whether MVMC is statistically (according to pairwise  $t$ -test at 95% significance level) superior/inferior to the other method.

		MCC-NBMM	MultiCC	MVMCC
Caltech-7	SC $\uparrow$	-0.100 $\pm$ 0.002●	-0.103 $\pm$ 0.006●	0.198 $\pm$ 0.004
	DI $\uparrow$	0.034 $\pm$ 0.000●	0.011 $\pm$ 0.000●	0.047 $\pm$ 0.000
	NMI $\downarrow$	0.376 $\pm$ 0.014●	0.005 $\pm$ 0.000	0.005 $\pm$ 0.000
	JC $\downarrow$	0.185 $\pm$ 0.003●	0.087 $\pm$ 0.000	0.083 $\pm$ 0.000
Caltech-20	SC $\uparrow$	-0.134 $\pm$ 0.000●	-0.229 $\pm$ 0.012●	0.080 $\pm$ 0.000
	DI $\uparrow$	0.026 $\pm$ 0.000●	0.011 $\pm$ 0.000●	0.156 $\pm$ 0.008
	NMI $\downarrow$	0.325 $\pm$ 0.010●	0.021 $\pm$ 0.000	0.026 $\pm$ 0.000
	JC $\downarrow$	0.150 $\pm$ 0.002●	0.056 $\pm$ 0.000●	0.029 $\pm$ 0.000
Corel	SC $\uparrow$	-0.087 $\pm$ 0.002●	-0.172 $\pm$ 0.012●	-0.017 $\pm$ 0.000
	DI $\uparrow$	0.024 $\pm$ 0.000●	0.015 $\pm$ 0.000●	0.152 $\pm$ 0.002
	NMI $\downarrow$	0.377 $\pm$ 0.013●	0.164 $\pm$ 0.002●	0.070 $\pm$ 0.000
	JC $\downarrow$	0.176 $\pm$ 0.004●	0.044 $\pm$ 0.000●	0.010 $\pm$ 0.000
Digits	SC $\uparrow$	-0.243 $\pm$ 0.024●	-0.214 $\pm$ 0.013●	0.144 $\pm$ 0.002
	DI $\uparrow$	0.014 $\pm$ 0.000	0.003 $\pm$ 0.000●	0.018 $\pm$ 0.000
	NMI $\downarrow$	0.286 $\pm$ 0.006●	0.010 $\pm$ 0.000○	0.207 $\pm$ 0.003
	JC $\downarrow$	0.166 $\pm$ 0.001●	0.060 $\pm$ 0.000○	0.115 $\pm$ 0.000
Wiki article	SC $\uparrow$	-0.0694 $\pm$ 0.000●	-0.058 $\pm$ 0.000●	0.064 $\pm$ 0.000
	DI $\uparrow$	0.079 $\pm$ 0.000	0.041 $\pm$ 0.001●	0.078 $\pm$ 0.000
	NMI $\downarrow$	0.287 $\pm$ 0.005●	0.007 $\pm$ 0.000	0.006 $\pm$ 0.000
	JC $\downarrow$	0.123 $\pm$ 0.002●	0.054 $\pm$ 0.000	0.052 $\pm$ 0.000
Mirflickr	SC $\uparrow$	-0.095 $\pm$ 0.000●	-0.194 $\pm$ 0.002●	0.064 $\pm$ 0.000
	DI $\uparrow$	0.052 $\pm$ 0.000●	0.065 $\pm$ 0.000●	0.151 $\pm$ 0.003
	NMI $\downarrow$	0.017 $\pm$ 0.000●	0.081 $\pm$ 0.000●	0.005 $\pm$ 0.000
	JC $\downarrow$	0.052 $\pm$ 0.000●	0.052 $\pm$ 0.000●	0.022 $\pm$ 0.000

Table 3: Quality and Diversity of the various competing methods on finding multiple co-clusterings. ●/○ indicates whether MVMCC is statistically (according to pairwise  $t$ -test at 95% significance level) superior/inferior to the other method.

alternative clusterings on multi-view data. MVMC always gives the best result on the diversity metrics (NMI and JC). This fact suggests it can find two alternative clusterings with high diversity. MVMC occasionally has a lower quality value (SC and DI) than some of the comparing methods. This is explainable, since obtaining alternative clusterings with high-diversity and high-quality is a well-known dilemma, and MVMC achieves a much larger diversity than the comparing methods. Although the comparing methods employ different techniques to explore alternative clusterings in the subspaces, they almost always lose to MVMC. The cause is that

the concatenated long feature vectors override the intrinsic structures of different views. This also explains why the comparing methods have lower values on diversity metrics (NMI and JC). In practice, because of the long concatenated feature vectors, the comparing methods generally suffer from a long runtime and cannot be applied on multi-view datasets with high-dimensional feature views. For example, we tried experiments on a large text dataset with five views, 18,758 samples, and more than 10,000 features per view. Since the concatenated data matrix has more than 100,000 features, almost all the comparing methods cannot complete after a long time, or even cannot run on a moderate server. In contrast, our MVMC is rather efficient, it does not need to concatenate features, and is directly applicable on each view.

Table 3 reports the results of MVMCC, MultiCC, and MCC-NBMM on exploring multiple co-clusterings, whose number is equal to the number of views  $h = m$  (unlike  $h = 2$  for Table 2), because of the feature view heterogeneity. For this evaluation, we report the average quality and diversity values of all pairwise alternative co-clusterings of  $h$  clusterings. We can see that MVMCC significantly outperforms the two state-of-the-art multiple co-clustering methods across different evaluation metrics and datasets. MultiCC sometimes obtains a better diversity than MVMCC. This is because it directly optimizes the diversity on the sample-cluster and feature-cluster matrices, while our MVMCC indirectly optimizes the diversity, mainly using the sample-cluster matrices. These results prove the effectiveness of our solution on exploring multiple co-clusterings on multi-view data.

Following the experimental setup in Table 2, we conduct additional experiments to investigate the contribution of the shared matrix  $\mathbf{U}$  on improving the quality of multiple clusterings. For this investigation, we introduce a variant of MVMC( $n\mathbf{U}$ ), which only uses  $\{\mathbf{D}^k\}_{k=1}^h$  to generate multiple clusterings and disregards the shared information matrix  $\mathbf{U}$ . From

		SC $\uparrow$	DI $\uparrow$	NMI $\downarrow$	JC $\downarrow$
Digits	MVMC(nU)	0.061	0.076	0.008	0.050
	MVMC	0.064	0.087	0.008	0.052
Wiki article	MVMC(nU)	0.059	0.088	0.005	0.051
	MVMC	0.066	0.122	0.006	0.052

Table 4: Comparison results with/without the shared information matrix  $\mathbf{U}$  for discovering multiple clusterings.

the reported results on Digits and Wikipedia datasets in Table 4, we can see that the diversity (NMI and JC) of the two alternative clusterings remains almost the same when the shared information matrix  $\mathbf{U}$  is excluded. However, the quality (SC and DI) is clearly reduced when  $\mathbf{U}$  is excluded. This contrast proves that  $\mathbf{U}$  indeed improves the quality of multiple clusterings, and justifies our motivation to seek  $\mathbf{U}$  across views and leverage it with the individuality information matrix  $\mathbf{D}^v$  for multiple clusterings.

In summary, we can conclude that the diversity information matrix helps generating diverse clusterings, and the commonality information matrix of multi-view data can improve the quality of the clusterings. These experimental results also confirm our assumption that the diversity and commonality of multi-view data can be leveraged to generate diverse clusterings with high-quality.

### 3.3 Parameter Analysis

$\lambda_1$  and  $\lambda_2$  are two important input parameters of MVMC (M-CMCC) for seeking the individuality and commonality information of multi-view data, which consequently affect the quality and diversity of multiple clusterings. We investigate the sensitivity of MVMC to these parameters by varying  $\lambda_1$  (it controls diversity) and  $\lambda_2$  (it controls quality) in the range  $[10^{-3}, 10^{-2}, \dots, 10^3]$ . Figure 2 reports the Quality (DI) and Diversity (1-NMI, the larger the better) of MVMC on the Caltech-7 dataset. We have several interesting observations: (i) diversity (1-NMI) increases as  $\lambda_1$  increases, but not as much as with the increase of  $\lambda_2$  (see Figure 2(b)); quality (DI) increases as  $\lambda_2$  increases, but not as much as with the increase of  $\lambda_1$ . (ii) The synchronous increase of  $\lambda_1$  and  $\lambda_2$  does not necessarily give the highest results in quality and diversity. (iii) When both  $\lambda_1$  and  $\lambda_2$  are fixed to a small value, both quality and diversity are reduced. This fact suggests that both diversity and commonality information of multi-view data should be used for the exploration of alternative clusterings. These observations again confirm the known dilemma between diversity and quality of multiple clusterings. The values  $\lambda_1 = [10, 100]$  and  $\lambda_2 = [10, 100]$  often provide the best balance between quality and diversity.

We vary  $h$  from 2 to  $2 * m$  on the Caltech-7 dataset to explore the variation of average quality and diversity of multiple clusterings generated by MVMC. In Figure 3, as  $h$  increases, the average quality of multiple clusterings decreases gradually with small fluctuations, and the average diversity fluctuates in a small range. These patterns are due to the quality-diversity dilemma of multiple clusterings. Increasing the number of alternative and diverse clusterings results in a sacrifice of quality. Overall, we can find that our MVMC can explore  $h \geq 2$  alternative clusterings with quality and diversity.

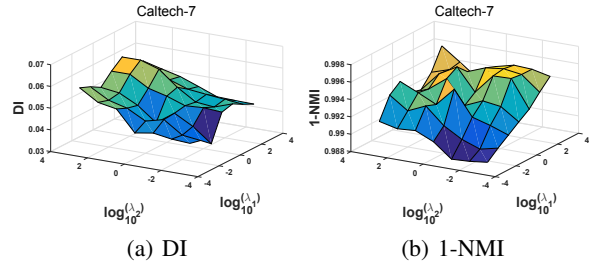


Figure 2: Quality (SI) and Diversity(1-NMI) of MVMC vs.  $\lambda_1$  and  $\lambda_2$  on the Caltech-7 dataset.

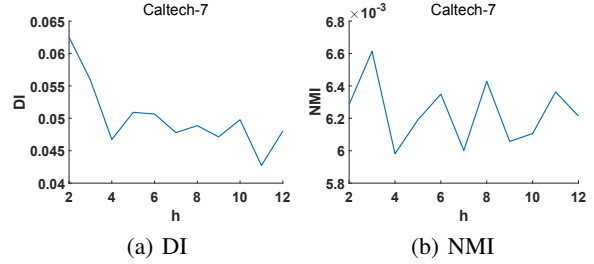


Figure 3: Quality (SI) and Diversity(NMI, the lower the better) of MVMC vs.  $h$  from 2 to  $2 * m$  on the Caltech-7 dataset.

The time complexity of our MVMC is  $O(tn^2d(h^2v + 2hv + 2h))$ , where  $t$  is the number of iterations for optimization. The experiments are conducted on a server with Ubuntu 16.04, Intel Xeon8163 with 1TB RAM; all methods are implemented in Matlab2014a. The detailed runtime analysis can be viewed in the supplementary file due to the limitation of space.

## 4 Conclusion

In this paper, we proposed an approach to generate multiple clusterings (co-clusterings) from multi-view data, which is an interesting but largely overlooked clustering topic encompassing both multi-view clustering and multiple clusterings. Our approach leverages the diversity and commonality of multi-view data to generate multiple clusterings, and outperforms state-of-the-art multiple clustering solutions. Our study confirms the existence of individuality and commonality of multi-view data, and their contribution for generating diverse clusterings with good quality. In the future, we plan to find a principled way to automatically determine the number of alternative clusterings, and explore weighting schemes of data views. The code of MVMC (MVMCC) is available at <http://mlda.swu.edu.cn/codes.php?name=MVMC>.

## Acknowledgments

This work is supported by NSFC (61872300 and 61873214), Fundamental Research Funds for the Central Universities (XDJK2019B024), NSF of CQ CSTC (cstc2018jcyjAX0228 and cstc2016jcyjA0351), King Abdullah University of Science and Technology (KAUST), Saudi Arabia.

## References

- [Bae and Bailey, 2006] Eric Bae and James Bailey. Coala: A novel approach for the extraction of an alternate clustering of high quality and high dissimilarity. In *ICDM*, pages 53–62, 2006.
- [Bailey, 2013] James Bailey. Alternative clustering analysis: A review. In Aggarwal Charu and Reddy Chandan, editors, *Data Clustering: Algorithms and Applications*, pages 535–550. CRC Press, 2013.
- [Belkin *et al.*, 2006] Mikhail Belkin, Partha Niyogi, and Vikas Sindhwani. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *JMLR*, 7(11):2399–2434, 2006.
- [Cao *et al.*, 2015] Xiaochun Cao, Changqing Zhang, Huazhu Fu, Si Liu, and Hua Zhang. Diversity-induced multi-view subspace clustering. In *CVPR*, pages 586–594, 2015.
- [Caruana *et al.*, 2006] Rich Caruana, Mohamed Elhawary, Nam Nguyen, and Casey Smith. Meta clustering. In *ICDM*, pages 107–118, 2006.
- [Chao *et al.*, 2017] Guoqing Chao, Shiliang Sun, and Jinbo Bi. A survey on multi-view clustering. *arXiv preprint arXiv:1712.06246*, 2017.
- [Cui *et al.*, 2007] Ying Cui, Xiaoli Z Fern, and Jennifer G Dy. Non-redundant multi-view clustering via orthogonalization. In *ICDM*, pages 133–142, 2007.
- [Dang and Bailey, 2010] Xuan Hong Dang and James Bailey. Generation of alternative clusterings using the cami approach. In *SDM*, pages 118–129, 2010.
- [Davidson and Qi, 2008] Ian Davidson and Zijie Qi. Finding alternative clusterings using constraints. In *ICDM*, pages 773–778, 2008.
- [Ding *et al.*, 2010] Chris HQ Ding, Tao Li, and Michael I Jordan. Convex and semi-nonnegative matrix factorizations. *TPAMI*, 32(1):45–55, 2010.
- [Gao *et al.*, 2015] Hongchang Gao, Feiping Nie, Xuelong Li, and Heng Huang. Multi-view subspace clustering. In *ICCV*, pages 4238–4246, 2015.
- [Gönen and Alpaydm, 2011] Mehmet Gönen and Ethem Alpaydm. Multiple kernel learning algorithms. *JMLR*, 12(7):2211–2268, 2011.
- [Gretton *et al.*, 2005] Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf. Measuring statistical dependence with hilbert-schmidt norms. In *ALT*, pages 63–77, 2005.
- [Jain *et al.*, 2008] Prateek Jain, Raghu Meka, and Inderjit S Dhillon. Simultaneous unsupervised learning of disparate clusterings. *Statistical Analysis and Data Mining*, 1(3):195–210, 2008.
- [Kumar and Daumé, 2011] Abhishek Kumar and Hal Daumé. A co-training approach for multi-view spectral clustering. In *ICML*, pages 393–400, 2011.
- [Li *et al.*, 2015] Yeqing Li, Feiping Nie, Heng Huang, and Junzhou Huang. Large-scale multi-view spectral clustering via bipartite graph. In *AAAI*, 2015.
- [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. *TPAMI*, 35(1):171–184, 2013.
- [Luo *et al.*, 2018] Shirui Luo, Changqing Zhang, Wei Zhang, and Xiaochun Cao. Consistent and specific multi-view subspace clustering. In *AAAI*, pages 3730–3737, 2018.
- [Mautz *et al.*, 2018] Dominik Mautz, Wei Ye, Claudia Plant, and Christian Böhm. Discovering non-redundant k-means clusterings in optimal subspaces. In *KDD*, pages 1973–1982, 2018.
- [Monti *et al.*, 2003] Stefano Monti, Pablo Tamayo, Jil-l Mesirov, and Todd Golub. Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data. *Machine learning*, 52(1-2):91–118, 2003.
- [Niu *et al.*, 2010] Donglin Niu, Jennifer G Dy, and Michael I Jordan. Multiple non-redundant spectral clustering views. In *ICML*, pages 831–838, 2010.
- [Tao *et al.*, 2018] Hong Tao, Chenping Hou, Xinwang Liu, Tongliang Liu, Dongyun Yi, and Jubo Zhu. Reliable multi-view clustering. In *AAAI*, pages 4123–4130, 2018.
- [Tokuda *et al.*, 2017] Tomoki Tokuda, Junichiro Yoshimoto, Yu Shimizu, and et al. Multiple co-clustering based on nonparametric mixture models with heterogeneous marginal distributions. *PLoS ONE*, 12(10):e0186566, 2017.
- [Wang *et al.*, 2011] Hua Wang, Feiping Nie, Heng Huang, and Fillia Makedon. Fast nonnegative matrix tri-factorization for large-scale data co-clustering. In *IJCAI*, pages 1553–1558, 2011.
- [Wang *et al.*, 2018] Xing Wang, Guoxian Yu, Carlotta Domeniconi, Jun Wang, Zhiwen Yu, and Zili Zhang. Multiple co-clusterings. In *ICDM*, pages 1308–1313, 2018.
- [Wang *et al.*, 2019] Xing Wang, Guoxian Yu, Carlotta Domeniconi, Jun Wang, Guoqiang Xiao, and Maozu Guo. Multiple independent subspace clusterings. In *AAAI*, pages 1–8, 2019.
- [Wright *et al.*, 2010] John Wright, Yi Ma, Julien Mairal, Guillermo Sapiro, Thomas S Huang, and Shuicheng Yan. Sparse representation for computer vision and pattern recognition. *Proceedings of the IEEE*, 98(6):1031–1044, 2010.
- [Yang and Zhang, 2017] Sen Yang and Lijun Zhang. Non-redundant multiple clustering by nonnegative matrix factorization. *Machine Learning*, 106(5):695–712, 2017.
- [Ye *et al.*, 2016] Wei Ye, Samuel Maurus, Nina Hubig, and Claudia Plant. Generalized independent subspace clustering. In *ICDM*, pages 569–578, 2016.