Diverse Nonnegative Matrix Factorization for Multi-view Data Representation

Jing Wang, Feng Tian, Hongchuan Yu, *Member, IEEE*, Chang Hong Liu, Kun Zhan, and Xiao Wang*

***Abstract*—Nonnegative matrix factorization (NMF), a method for finding parts-based representation of nonnegative data, has shown remarkable competitiveness in data analysis. Given that real-world datasets are often comprised of multiple features or views which describe data from various perspectives, it is important to exploit diversity from multiple views for comprehensive and accurate data representations. Moreover, real-world datasets often come with high-dimensional features, which demands the efficiency of low-dimensional representation learning approaches. To address these needs, we propose a Diverse Nonnegative Matrix Factorization (DiNMF) approach. It enhances the diversity, reduces the redundancy among multi-view representations with a novel defined diversity term and enables the learning process in linear execution time. We further propose a Locality Preserved DiNMF (LP-DiNMF) for more accurate learning, which ensures diversity from multiple views while preserving the local geometry structure of data in each view. Efficient iterative updating algorithms are derived for both DiNMF and LP-DiNMF, along with proofs of convergence. Experiments on synthetic and real-world datasets have demonstrated the efficiency and accuracy of the proposed methods against the state-of-the-art approaches, proving the advantages of incorporating the proposed diversity term into NMF.**

***Index Terms*—Diversity representation, multi-view learning, nonnegative matrix factorization**

I. Introduction

FINDING a suitable representation is a fundamental problem for many data analysis tasks [1], [2], [3], [4], [5], as a good representation can often reveal the latent structure of data hence facilitate processes such as clustering, classification and recognition. Nonnegative matrix factorization (NMF) [6] is a well-known technique for such representation of data. It is widely studied and applied to real-world data, such as images and texts, because it possesses parts-of-whole interpretations and creates better practical performance.

Several variants of NMF have been proposed to seek for more effective data representation in recent years. For example, Kong et al.,[7] proposed a robust formulation of NMF (RNMF) to deal with large noises by $L_{2,1}$ norm. Cai et al., [8] proposed a graph regularized NMF (GNMF) to model the local manifold structure by constructing an affinity graph. However, the performance of GNMF is known to hinge heavily on the choice of nearest neighbor graph and it is difficult and time consuming to choose a suitable graph. To overcome this limitation, Wang et. al., [9] proposed a multiple graph regularized NMF (MultiGNMF) to approximate intrinsic manifold approximation automatically. Similarly, a relational multi-manifold co-clustering (RMC) approach [10] is proposed to maximally approximate the true intrinsic manifolds of both the sample and feature spaces simultaneously. Li et al., [11] proposed a Locally Constrained A-optimal nonnegative projection method which not only preserves the locally geometrical structure of the data but also incorporates label information as constraints to enhance the discriminating power. Later, Wang et al., [12] proposed two GNMF-based methods to learn the graph that is adaptive to the selected features and learned multiple kernels, respectively. Under the assumption that data samples from different domains have different distributions, but share same feature and class label spaces, Wang et al., [13] proposed a novel NMF-based approach for multiple-domain learning.

Recently, data collected from various sources or represented by different feature extractors are available in many real-world applications [14]. For example, one document may be translated into different languages; web pages can be represented by different features based on both content and hyperlinks; an image or video can be represented by different visual descriptors, such as SIFT [15], HOG [16] and GIST [17]; research communities are formed according to research topics as well as co-authorship links and so on. These heterogeneous features that are represented by different perspectives of data are referred as multiple views [18], [19].

With the increasing amount of multi-view data, approaches employing NMF-based multi-view learning have attracted attention. MultiNMF [20] formulated a joint multi-view NMF learning process with the constraint that encourages representation of each view towards a common consensus. Subsequently, several approaches [21], [22], [23], [24] were proposed based on MultiNMF. Specifically, Zhang et al. [21] developed a multi-manifold NMF (MMNMF) by incorporating the locally geometrical structure of data across multiple views. It regards each view as one manifold and the intrinsic manifold of a dataset as a mixture of the manifolds. Kalayeh et al. [22] proposed a weighted extension of MultiNMF [20] for image annotation, in which two weight matrices are introduced to alleviate the issue of dataset imbalance in real applications. Ou et al. [23] explored the local geometric structure for

This work is supported by EU H2020 project-AniAge (No.691215) and National Natural Science Foundation of China (No. 61702296).

Jing Wang and Feng Tian are with Faculty of Science and Technology, Bournemouth University, BH125BB, UK (Email: {jwang, ftian}@bournemouth.ac.uk).

Hongchuan Yu is with National Centre for Computer Animation, Bournemouth University, BH125BB, UK (Email:hyu@bournemouth.ac.uk).

Chang Hong Liu is with Department of Psychology, Bournemouth University, BH125BB, UK (Email: liuc@bournemouth.ac.uk).

Kun Zhan is with School of Information Science and Engineering, Lanzhou University, Lanzhou 730000, Gansu, China (Email: kzhan@lzu.edu.cn).

Xiao Wang (Corresponding Author) is with Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China (Email: wangxiao_cv@163.com).

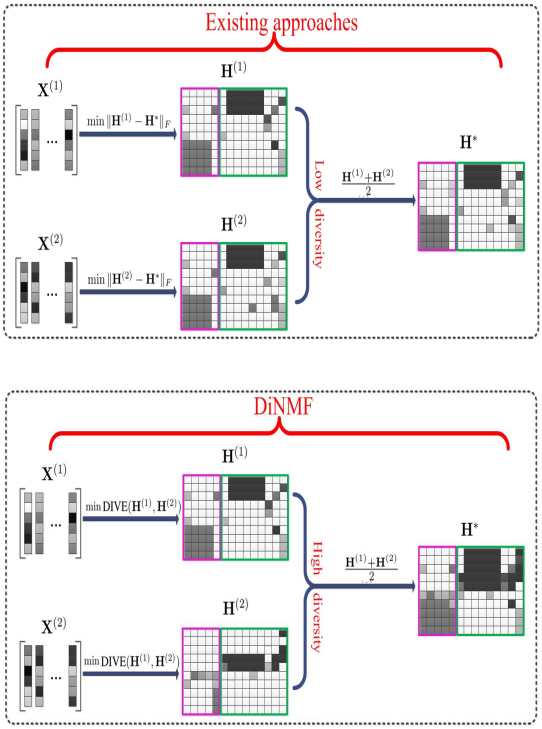


Fig. 1: Comparison of existing NMF-based Multi-view approaches and the proposed DiNMF. A multi-view dataset $\mathbf{X}$ contains two equally important views, i.e., $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$. $\mathbf{H}^{(1)}$ and $\mathbf{H}^{(2)}$ are the corresponding learned representation matrices. $\mathbf{H}^*$ is the final representation. For all matrices, the data vectors are column-wise and the features are row-wise. The ground-truth is shown as group-1 in purple and group-2 in green. By enforcing $\mathbf{H}^{(1)}$ and $\mathbf{H}^{(2)}$ to be close to $\mathbf{H}^*$, the existing approaches learn the data representations of two views jointly to capture the shared underlying common information but cannot ensure their diversity. In contrast, DiNMF is based on a diversity term (DIVE), which captures diverse information among data representations. This ensures that $\mathbf{H}^*$ not only contains common information captured by existing approaches but also preserves some distinct information from each view, thus more comprehensive and accurate.

each view under the patch alignment framework and adopted correntropy-induced metric to measure the reconstruction error of each view to improve the robustness. Wang et al. [24] extended MultiNMF to semi-supervised setting by ensuring that data with same label have same representations and use a single parameter to learn the weight of each view adaptively.

However, one of the main limitations of all these approaches is that the learned data representations from multiple views contain mutually redundant information and lack diverse information. This is because, to a large extent, existing approaches are all to exploit common information shared by multiple views but neglect the diversity among views. The diversity means that each view of the data contain s some distinct information that other views do not have. Taking the diversity into account, we can capture more information of data and achieve more comprehensive and accurate learning, because different views usually describe data from different aspects. Some researches [25], [26], [27] have also shown that the diversity is of importance to multi-view learning. Therefore, it should be beneficial to integrate diversity properties of views into NMF learning.

To achieve this goal, we propose a novel Diverse Nonnegative Matrix Factorization (DiNMF) method. With a novel regularization term, DiNMF encourages the representations from multiple views to be diverse enough to capture comprehensive information, so that a diverse and more accurate data representation is eventually achieved. As illustrated in Figure 1, existing approaches (the upper figure) learn the data representations jointly to capture the underlying common structure shared by two views. They enforce the feature distribution of $\mathbf{H}^{(1)}$ and $\mathbf{H}^{(2)}$ to be similar but fail to take advantage of distinct information of each view. This may lead to unsatisfactory results. It can be seen from the last columns of $\mathbf{H}^{(1)}$ and $\mathbf{H}^{(2)}$ that the feature distributions are nearly same and happen to be similar to columns in the group-1 (purple). Through linear computations, the corresponding column of $\mathbf{H}^*$ will be categorized into a wrong group, i.e., group-1, due to the similarity of feature distribution. On the contrary, DiNMF is based on a novel diversity constraint, i.e., DIVE, which enforces $\mathbf{H}^{(1)}$ and $\mathbf{H}^{(2)}$ to be as diverse as possible. As a result, $\mathbf{H}^*$ contains diverse information for comprehensive learning, since $\mathbf{H}^{(2)}$ captures some distinct information that $\mathbf{H}^{(1)}$ lacks. Moreover, the feature distributions of the two groups are more distinct in-between and this is in line with the ground truth, leading to more accurate learning.

The main contributions of our work are as follows:

1. DiNMF not only ensures the diversity to exploit comprehensive information but also reduces mutually redundancy across multiple representations for more accurate learning. Furthermore, DiNMF is also computationally linear thus has good scalability to large-scale datasets.

2. We further develop Locality Preserved DiNMF (LP-DiNMF) to preserve the locally geometrical structure of the manifolds for multi-view setting, by taking into account the manifold structures in data spaces. This leads to improved clustering accuracy compared with DiNMF.

3. We derive novel and efficient algorithms for both DiNMF and LP-DiNMF to optimize objective functions. The convergence of both algorithms are proved.

4. Experiments on both synthetic and real-world datasets from different domains demonstrate that the proposed methods are not only faster but also achieve more accurate clustering than other state-of-the-art methods.

II. Diverse Nonnegative Matrix Factorization (DiNMF)

In this section, we first briefly review the background of NMF and introduce a straightforward approach to extend the single-view NMF to multi-view setting. After that, we present DiNMF and propose an efficient optimization algorithm for solving the objective function.

A. Objective Function of Non-diverse NMF (NdNMF)

Suppose $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n] \in \mathbf{R}^{m \times n}$ is the nonnegative n data matrix where each column is a data vector and m is the dimensionality of the feature space. NMF aims to find two nonnegative matrix factors $\mathbf{W}$ and $\mathbf{H}$ whose product can well approximate the original matrix:

$$\mathbf{X} \approx \mathbf{WH}. \tag{1}$$

Here the $\mathbf{H} \in \mathbb{R}^{k \times n}$ can be considered as the new representations of data in terms of the basis $\mathbf{W} \in \mathbb{R}^{m \times k}$, where k demotes the desired reduced dimension.

The approximation is quantified by a cost function which can be constructed by distance measures. A popular measure is the square of the Euclidean distance (also known as the Frobenius norm) between two matrices [28]. Thus, NMF aims to minimize the following objective function:

$$\|\mathbf{X} - \mathbf{WH}\|_F^2, \quad s.t. \quad \mathbf{W}, \mathbf{H} \geq 0 \tag{2}$$

This standard NMF can be extended to multi-view setting straightforwardly. Let $\mathbf{X}^{(v)} \in \mathbb{R}^{m^{(v)} \times n}$ be the feature matrix corresponding to the vth view. Similarly, $\mathbf{W}^{(v)}$ and $\mathbf{H}^{(v)}$ are the corresponding basis matrix and representation matrix, respectively. Given V heterogeneous features, we directly integrate all these features together so the objective function (2) becomes

$$\sum_{v=1}^{V} \|\mathbf{X}^{(v)} - \mathbf{W}^{(v)}\mathbf{H}^{(v)}\|_F^2. \quad s.t. \quad \mathbf{W}^{(v)}, \mathbf{H}^{(v)} \geq 0 \tag{3}$$

Obviously, this approach learns each data representation independently and cannot ensure the diversity of different views. To facilitate the subsequent discussion, we call this approach Non-diverse Multi-view Nonnegative Matrix Factorization (NdNMF).

B. Objective Function of DiNMF

A desirable multi-view NMF approach for data analysis needs to satisfy two requirements. First, it should exploit diverse information across multi-view data representations for more comprehensive and accurate learning. Second, it is scalable since the number of data n and dimension of features m could be quite large. In the following, we describe how DiNMF satisfies these two requirements.

Diversity requires that two data vectors be as orthogonal to each other as possible, so that more comprehensive information can be exploited. Let $\mathbf{h}_i^{(v)}$ and $\mathbf{h}_i^{(w)}$ be the ith data representation vectors in two views, i.e, the v-th and w-th views. To ensure the diversity between the two vectors, their dot product should be 0, approximately. To achieve this, we can minimize the following function [29]

$$\|\mathbf{h}_i^{(v)} \circ \mathbf{h}_i^{(w)}\|_0, \tag{4}$$

where $\circ$ designates the product, and $\|\cdot\|_0$ is the l^0 norm which indicates the number of non-zero elements. Due to the non-convexity and discontinuity of l^0 norm, (4) can be relaxed by using l^1 norm as follows,

$$\|\mathbf{h}_i^{(v)} \circ \mathbf{h}_i^{(w)}\|_1 = \sum_{j=1}^{k} |h_{ji}^{(v)}| \cdot |h_{ji}^{(w)}|, \tag{5}$$

where $|\cdot|$ is the absolute value. Since the representations obtained by NMF are non-negative, we can further reformulate (5) as

$$\|\mathbf{h}_i^{(v)} \circ \mathbf{h}_i^{(w)}\|_1 = \sum_{j=1}^{k} h_{ji}^{(v)} \cdot h_{ji}^{(w)}. \tag{6}$$

By extending the calculation of single data vector in (6) to n data vectors setting, we propose the following term to guarantee the diversity among all n data vectors in two views,

$$\begin{aligned} \text{DIVE}(\mathbf{H}^{(v)}, \mathbf{H}^{(w)}) &= \sum_{i=1}^{n}\sum_{j=1}^{k} h_{ji}^{(v)} \cdot h_{ji}^{(w)} \\ &= tr(\mathbf{H}^{(v)}{\mathbf{H}^{(w)}}^T), \end{aligned} \tag{7}$$

where $tr(\cdot)$ is the trace function. Therefore, minimizing (7) will encourage $\mathbf{H}^{(v)}$ and $\mathbf{H}^{(w)}$ to be orthogonal to each other. In other words, the diversity of the representation matrices in two views is guaranteed.

Given a dataset with more views, we incorporate the DIVE into NdNMF to guarantee that data representations in any two views be diverse. Then, the minimization objective function is produced as follows:

$$\begin{aligned} &\sum_{v=1}^{V} \|\mathbf{X}^{(v)} - \mathbf{W}^{(v)}\mathbf{H}^{(v)}\|_F^2 + \alpha \sum_{v \neq w} \text{DIVE}(\mathbf{H}^{(v)}, \mathbf{H}^{(w)}) \\ &\qquad s.t. \quad 1 \leq v, w \leq V, \mathbf{W}^{(v)}, \mathbf{H}^{(v)}, \mathbf{H}^{(w)}, \alpha \geq 0, \end{aligned} \tag{8}$$

where α is a trade-off parameter which controls the weight of DIVE. A smooth regularization term $\|\mathbf{H}^{(v)}\|_F^2$ is added to avoid over-fitting of a view, which leads to the overall objective function as follows:

$$\begin{aligned} &\underbrace{\sum_{v=1}^{V} \|\mathbf{X}^{(v)} - \mathbf{W}^{(v)}\mathbf{H}^{(v)}\|_F^2}_{error} \\ &+ \underbrace{\alpha \sum_{v \neq w} \text{DIVE}(\mathbf{H}^{(v)}, \mathbf{H}^{(w)})}_{diversity} + \underbrace{\beta \sum_{v=1}^{V} \|\mathbf{H}^{(v)}\|_F^2}_{smoothness} \\ &s.t. \quad 1 \leq v, w \leq V, \mathbf{W}^{(v)}, \mathbf{H}^{(v)}, \mathbf{H}^{(w)}, \alpha, \beta \geq 0. \end{aligned} \tag{9}$$

Here β is the weight factor of the smoothness term.

To solve the objective function (9), we develop an efficient optimization algorithm to find the optimal solution of $\mathbf{H}^{(v)}$. After that, we calculate the average value of $\mathbf{H}^{(v)}$ in all views for the final multi-view data representation $\mathbf{H}^*$, i.e., $\mathbf{H}^* = \frac{\sum_{v=1}^{V} \mathbf{H}^{(v)}}{V}$. Following are the details.

C. Solving the Optimization Problem (9)

Since the objective function (9) is not convex with both variables $\mathbf{W}^{(v)}$ and $\mathbf{H}^{(v)}$, it is infeasible to find the global

minimum. Instead, we propose an algorithm to find a local minima by iteratively updating $\mathbf{W}^{(v)}$ with $\mathbf{H}^{(v)}$ fixed and then updating $\mathbf{H}^{(v)}$ with $\mathbf{W}^{(v)}$ fixed.

For each view, the computations of $\mathbf{W}^{(v)}$ and $\mathbf{H}^{(v)}$ are not dependent on other views, so minimizing (9) gives us

$$\begin{aligned}&\|\mathbf{X}^{(v)}-\mathbf{W}^{(v)}\mathbf{H}^{(v)}\|_F^2+\alpha\sum_{w=1;w\neq v}^{V}tr(\mathbf{H}^{(v)}{\mathbf{H}^{(w)}}^T)+\beta\|\mathbf{H}^{(v)}\|_F^2\\&=tr(\mathbf{X}^{(v)}{\mathbf{X}^{(v)}}^T-2\mathbf{X}^{(v)}{\mathbf{H}^{(v)}}^T{\mathbf{W}^{(v)}}^T+\mathbf{W}^{(v)}\mathbf{H}^{(v)}{\mathbf{H}^{(v)}}^T{\mathbf{W}^{(v)}}^T)\\&+\alpha\sum_{w=1,w\neq v}^{V}tr(\mathbf{H}^{(v)}{\mathbf{H}^{(w)}}^T)+\beta tr(\mathbf{H}^{(v)}{\mathbf{H}^{(v)}}^T).\end{aligned}\tag{10}$$

Let $\eta_{ij}^{(v)}$ and $\xi_{ij}^{(v)}$ be the Lagrange multipliers for the constraint $w_{ij}^{(v)}\geq 0$ and $h_{ij}^{(v)}\geq 0$, respectively, and $\boldsymbol{\eta}^{(v)}=[\eta_{ij}^{(v)}]$, $\boldsymbol{\xi}^{(v)}=[\xi_{ij}^{(v)}]$, then the Lagrange function L of (10) is

$$\begin{aligned}L=&tr(\mathbf{X}^{(v)}{\mathbf{X}^{(v)}}^T-2\mathbf{X}^{(v)}{\mathbf{H}^{(v)}}^T{\mathbf{W}^{(v)}}^T\\&+\mathbf{W}^{(v)}\mathbf{H}^{(v)}{\mathbf{H}^{(v)}}^T{\mathbf{W}^{(v)}}^T)+\alpha\sum_{w=1,w\neq v}^{V}tr(\mathbf{H}^{(v)}{\mathbf{H}^{(w)}}^T)\\&+\beta tr(\mathbf{H}^{(v)}{\mathbf{H}^{(v)}}^T)+tr(\boldsymbol{\eta}^{(v)}\mathbf{W}^{(v)})+tr(\boldsymbol{\xi}^{(v)}\mathbf{H}^{(v)}).\end{aligned}\tag{11}$$

Setting the derivative of L to be 0 with respect to $\mathbf{W}^{(v)}$ and $\mathbf{H}^{(v)}$, we have

$$\boldsymbol{\xi}=2{\mathbf{W}^{(v)}}^T\mathbf{X}^{(v)}-2{\mathbf{W}^{(v)}}^T\mathbf{W}^{(v)}\mathbf{H}^{(v)}-\alpha\sum_{w=1,w\neq v}^{V}\mathbf{H}^{(w)}-2\beta\mathbf{H}^{(v)},\tag{12}$$

and

$$\boldsymbol{\eta}=2{\mathbf{W}^{(v)}}^T\mathbf{X}^{(v)}-2{\mathbf{W}^{(v)}}^T\mathbf{W}^{(v)}\mathbf{H}^{(v)}.\tag{13}$$

Following the Karush-Kuhn-Tucker (KKT) condition [30] $\eta_{ij}^{(v)}w_{ij}^{(v)}=0$ and $\xi_{ij}^{(v)}h_{ij}^{(v)}=0$, we get the equations for $w_{ij}^{(v)}$ and $h_{ij}^{(v)}$:

$$(2{\mathbf{W}^{(v)}}^T\mathbf{X}^{(v)}-2{\mathbf{W}^{(v)}}^T\mathbf{W}^{(v)}\mathbf{H}^{(v)}-\alpha\sum_{w=1,w\neq v}^{V}\mathbf{H}^{(w)}-2\beta\mathbf{H}^{(v)})h_{ij}^{(v)}=0,\tag{14}$$

$$(2\mathbf{X}^{(v)}{\mathbf{H}^{(v)}}^T-2\mathbf{W}^{(v)}\mathbf{H}^{(v)}{\mathbf{H}^{(v)}}^T)w_{ij}^{(v)}=0.\tag{15}$$

These equations lead to the following updating rules:

$$h_{ij}^{(v)}\leftarrow h_{ij}^{(v)}\frac{(2{\mathbf{W}^{(v)}}^T\mathbf{X}^{(v)})_{ij}}{(2{\mathbf{W}^{(v)}}^T\mathbf{W}^{(v)}\mathbf{H}^{(v)}+\alpha\sum_{w=1,w\neq v}^{V}\mathbf{H}^{(w)}+2\beta\mathbf{H}^{(v)})_{ij}},\tag{16}$$

$$w_{di}^{(v)}\leftarrow w_{di}^{(v)}\frac{(\mathbf{X}^{(v)}{\mathbf{H}^{(v)}}^T)_{di}}{(\mathbf{W}^{(v)}\mathbf{H}^{(v)}{\mathbf{H}^{(v)}}^T)_{di}}.\tag{17}$$

The procedure to solve (9) is summarized in the Algorithm 1.

Algorithm 1 The algorithm of DiNMF

Input:
Data for V views $\{\mathbf{X}^{(1)},\mathbf{X}^{(2)},...,\mathbf{X}^{(V)}\}$.
Parameter α and β.
for $v=1$ to V **do**
  Normalizing $\mathbf{X}^{(v)}$
  Initializing $\mathbf{W}^{(v)},\mathbf{H}^{(v)}$
end for
for $v=1$ to V **do**
  while not converging **do**
    Fixing $\mathbf{W}^{(v)}$, updating $\mathbf{H}^{(v)}$ by (16)
    Fixing $\mathbf{H}^{(v)}$, updating $\mathbf{W}^{(v)}$ by (17)
  end while
end for
Calculate the average value of all data representations of each view by $\mathbf{H}^{*}=\frac{\sum_{v=1}^{V}\mathbf{H}^{(v)}}{V}$.
Output: The final representation matrix $\mathbf{H}^{*}$.

D. Convergence of DiNMF

In this section, we prove the convergence of the updating rules (16) and (17). Algorithm 1 is guaranteed to converge to a local minima by the following theorem:

Theorem 1. The objective function (9) is non-increasing under the update rules (16) and (17).

To prove Theorem 1, we need to show that (10) for each view is non-increasing under (16) and (17). Since the second term and the third term of (10) are only related to $\mathbf{H}$, we have exactly the same update formula for $\mathbf{W}$ in DiNMF as in [31]. Here, we only prove (10) is non-increasing under (16). Following [31], we will apply an auxiliary function, which is defined as follows:

Definition 1 A function $G(h,h')$ is an auxiliary function of the function $J(h)$ if $G(h,h')\geq J(h)$ and $G(h,h)=J(h)$ for any h, h'.

The auxiliary function helps because of the following lemma [31],

Lemma 1. If G is an auxiliary function of the objective function J, then J is non-increasing under the update rule

$$h^{t+1}=\arg\min_{h}G(h,h^t).\tag{18}$$

Now, we will show that the update for $\mathbf{H}$ (16) is exactly same as the update (18) with a proper auxiliary function. We rewrite (10) as follows:

$$\begin{aligned}O_1=&\|\mathbf{X}^{(v)}-\mathbf{W}^{(v)}\mathbf{H}^{(v)}\|_F^2\\&+\alpha\sum_{w=1,w\neq v}^{V}\text{DIVE}(\mathbf{H}^{(v)},\mathbf{H}^{(w)})+\beta\|\mathbf{H}^{(v)}\|_F^2\\=&\sum_{i=1}^{m^{(v)}}\sum_{j=1}^{n}(x_{ij}^{(v)}-\sum_{k=1}^{K}w_{ik}^{(v)}h_{kj}^{(v)})^2\\&+\alpha\sum_{w=1,w\neq v}^{V}\sum_{k=1}^{K}\sum_{j=1}^{n}h_{kj}^{(v)}h_{jk}^{(w)}+\beta\sum_{k=1}^{K}\sum_{j=1}^{n}h_{kj}^{(v)}h_{jk}^{(v)}.\end{aligned}\tag{19}$$

Given an element $h_{ab}^{(v)}$ in $\mathbf{H}^{(v)}$, we use $F_{ab}^{(v)}$ to denote the part of O_1 which is only relevant to $h_{ab}^{(v)}$. It is easy to check

that

$$F'_{ab} = (\frac{\partial O_1}{\partial \mathbf{H}})_{ab} =(-2\mathbf{W}^{(v)^T}\mathbf{X}^{(v)} + 2\mathbf{W}^{(v)^T}\mathbf{W}^{(v)}\mathbf{H}^{(v)})_{ab} + (\alpha \sum_{w=1,w\neq v}^{V} \mathbf{H}^{(w)} + 2\beta\mathbf{H}^{(v)})_{ab}, \tag{20}$$

$$F''_{ab} = (2\mathbf{W}^{(v)^T}\mathbf{W}^{(v)})_{aa} + 2\beta\mathbf{I}_{bb}. \tag{21}$$

Since our update is essentially element wise, it is sufficient to show that each F_{ab} is non-increasing under the update rule (16). We prove this by defining the auxiliary function regarding $h_{ab}^{(v)}$ as follows:

Lemma 2. The function

$$G(h_{ab}^{(v)}, h_{ab}^{(v)^t}) = F_{ab}(h_{ab}^{(v)^t}) + F'_{ab}(h_{ab}^{(v)^t})(h^{(v)} - h_{ab}^{(v)^t}) + \frac{2(\mathbf{W}^{(v)^T}\mathbf{W}^{(v)}\mathbf{H}^{(v)})_{ab} + \alpha\sum_{w=1,w\neq v}^{V}\mathbf{H}_{ab}^{(w)} + 2\beta\mathbf{H}_{ab}^{(v)}}{h_{ab}^{(v)^t}}(h^{(v)} - h_{ab}^{(v)^t})^2 \tag{22}$$

is an auxiliary function for F_{ab}, which is the part of O_1 and only relevant to $h_{ab}^{(v)}$.

Proof. Since $G(h^{(v)}, h^{(v)}) = F_{ab}(h^{(v)})$ is obvious, we need only show that $G(h^{(v)}, h_{ab}^{(v)^t}) \geq F_{ab}(h^{(v)})$. To do this, we compare the Taylor series expansion of $F_{ab}(h^{(v)})$:

$$F_{ab}(h^{(v)}) = F_{ab}(h_{ab}^{(v)^t}) + F'_{ab}(h^{(v)} - h_{ab}^{(v)^t}) + F''_{ab}(h^{(v)} - h_{ab}^{(v)^t})^2. \tag{23}$$

Introducing (20) and (21) into (23) and comparing with (22), we can see that, instead of proving that $G(h^{(v)}, h_{ab}^{(v)^t}) \geq F_{ab}(h^{(v)})$, it is equivalent to prove

$$\frac{(\mathbf{W}^{(v)^T}\mathbf{W}^{(v)}\mathbf{H}^{(v)})_{ab} + \beta\mathbf{H}_{ab}^{(v)}}{h_{ab}^{(v)^t}} \geq (\mathbf{W}^{(v)^T}\mathbf{W}^{(v)})_{aa} + \beta\mathbf{I}_{bb}. \tag{24}$$

Since we have

$$(\mathbf{W}^{(v)^T}\mathbf{W}^{(v)}\mathbf{H}^{(v)})_{ab} = \sum_{k=1}^{K}(\mathbf{W}^{(v)^T}\mathbf{W}^{(v)})_{al}h_{lb}^{(v)^t} \geq (\mathbf{W}^{(v)^T}\mathbf{W}^{(v)})_{aa}h_{ab}^{(v)^t} \tag{25}$$

and

$$\beta\mathbf{H}_{ab}^{(v)} = \beta\sum_{j=1}^{n} h_{aj}^{(v)^t}\mathbf{I}_{jb} \geq \beta h_{ab}^{(v)^t}\mathbf{I}_{bb}, \tag{26}$$

(24) holds and $G(h^{(v)}, h_{ab}^{(v)^t}) \geq F_{ab}(h^{(v)})$.

We can now demonstrate the convergence of Theorem 1.

Proof of Theorem 1. Replacing $G(h^{(v)}, h_{ab}^{(v)^t})$ in (18) by (22) results in the update rule

$$h_{ab}^{(v)^{t+1}} = h_{ab}^{(v)^t} - h_{ab}^{(v)^t}\frac{F'_{ab}(h_{ab}^{(v)^t})}{(2\mathbf{W}^{(v)^T}\mathbf{W}^{(v)}\mathbf{H}^{(v)} + \alpha\sum_{w=1,w\neq v}^{V}\mathbf{H}^{(w)} + 2\beta\mathbf{H}^{(v)})_{ab}} = h_{ab}^{(v)^t}\frac{(2\mathbf{W}^{(v)^T}\mathbf{X}^{(v)})_{ab}}{(2\mathbf{W}^{(v)^T}\mathbf{W}^{(v)}\mathbf{H}^{(v)} + \alpha\sum_{w=1,w\neq v}^{V}\mathbf{H}^{(w)} + 2\beta\mathbf{H}^{(v)})_{ab}}. \tag{27}$$

This is exactly the same as (16). Since (22) is an auxiliary function for F_{ab}, F_{ab} is non-increasing under (16) according to Lemma 1.

III. Locality Preserved DiNMF (LP-DiNMF)

Recent research has shown that data are found to lie on a nonlinear low dimensional manifold embedded in a high dimensional ambient space [32], [33], [34]. However, the standard NMF fails to discover such intrinsic geometrical structure of the data space [8]. To find a compact representation which uncovers the hidden semantics and simultaneously respects the intrinsic geometrical structure, we further extend DiNMF to LP-DiNMF so that local geometrical structure could be captured in each view.

A. Objective Function of LP-DiNMF Method

Cai et al. [8] imposed graph regularization on NMF. The method is based on the manifold assumption which means that, if two data points $\mathbf{x}_i$ and $\mathbf{x}_j$ are close in the original feature space, the representations of these two data points should be also close to each other. Mathematically, this can be represented by the following form: $\|\mathbf{x}_i - \mathbf{x}_j\| \to 0 \Rightarrow \|\mathbf{h}_i - \mathbf{h}_j\| \to 0$. With multi-view setting, a locality preserved term corresponding to the vth view is defined as:

$$\frac{1}{2}\sum_{i,j=1}^{n}(a_{ij}^{(v)}\|\mathbf{h}_i^{(v)} - \mathbf{h}_j^{(v)}\|^2)) = tr(\mathbf{H}^{(v)}\mathbf{L}^{(v)}\mathbf{H}^{(v)^T}), \tag{28}$$

where $\mathbf{L}^{(v)}$ is the Lagrange matrix $\mathbf{L}^{(v)} = \mathbf{D}^{(v)} - \mathbf{A}^{(v)}$, $\mathbf{A}^{(v)} = (a_{ij}^{(v)})$ is the weight matrix measuring the spatial closeness of data points and $\mathbf{D}^{(v)}$ is a diagonal matrix with $d_{ii}^{(v)} = \sum_j a_{ij}^{(v)}$. One of the most commonly used approaches to define the weight matrix $\mathbf{A}^{(v)}$ on the graph is $0-1$ weighting [8] . If $x_i^{(v)}$ and $x_j^{(v)}$ are one of the nearest neighbors to each other, $a_{ij}^{(v)} = 1$ otherwise $a_{ij}^{(v)} = 0$. Same as [21], we adopt this approach for it is simple to implement and performs well in practice. Combining this locality preserved regularizer with the objective function of DiNMF (9) gives rise to our LP-DiNMF, which minimizes the objective function as follows:

$$\begin{aligned}&\sum_{v=1}^{V}\|\mathbf{X}^{(v)} - \mathbf{W}^{(v)}\mathbf{H}^{(v)}\|_F^2 + \alpha\sum_{v\neq w}\text{DIVE}(\mathbf{H}^{(v)}, \mathbf{H}^{(w)}) \\ &+ \beta\sum_{v=1}^{V}\|\mathbf{H}^{(v)}\|_F^2 + \gamma\sum_{v=1}^{V}tr(\mathbf{H}^{(v)}\mathbf{L}^{(v)}\mathbf{H}^{(v)^T}) \\ &s.t. \quad 1 \leq v, w \leq V, \mathbf{W}^{(v)}, \mathbf{H}^{(v)}, \mathbf{H}^{(w)}, \alpha, \beta, \gamma \geq 0.\end{aligned} \tag{29}$$

Please note that if we set $\alpha = \beta$, the objective function (29) becomes simpler as

$$\begin{aligned}&\sum_{v=1}^{V}\|\mathbf{X}^{(v)} - \mathbf{W}^{(v)}\mathbf{H}^{(v)}\|_F^2 + \alpha\sum_{v=1}^{V}\text{DIVE}(\mathbf{H}^{(v)}, \sum_{w=1}^{V}\mathbf{H}^{(w)}) \\ &+ \gamma\sum_{v=1}^{V}tr(\mathbf{H}^{(v)}\mathbf{L}^{(v)}\mathbf{H}^{(v)^T}) \\ &s.t. \quad 1 \leq v, w \leq V, \mathbf{W}^{(v)}, \mathbf{H}^{(v)}, \mathbf{H}^{(w)}, \alpha, \gamma \geq 0.\end{aligned} \tag{30}$$

The DIVE term in (30) not only works on multi-view setting, but also on the single view. In detail, given different views ($v \neq w$), DIVE enforces the diversity among them. For the single view ($v = w$), DIVE plays an important role to avoid

over-fitting. This demonstrates the full compatibility of our objective function.

B. Solving the Optimization Problem (30)

Note that comparing with (9), the last term of (30) is related to $\mathbf{H}^{(v)}$ only, so we provide the optimization solution for updating $\mathbf{H}^{(v)}$ with $\mathbf{W}^{(v)}$ fixed.

Since updating $\mathbf{W}^{(v)}$ and $\mathbf{H}^{(v)}$ in each view is independent, (30) reduces to minimize the following formulation

$$\begin{aligned}&\|\mathbf{X}^{(v)}-\mathbf{W}^{(v)}\mathbf{H}^{(v)}\|_F^2+\alpha\text{DIVE}(\mathbf{H}^{(v)},\sum_{w=1}^{V}\mathbf{H}^{(w)})\\&+\gamma tr(\mathbf{H}^{(v)}\mathbf{L}^{(v)}{\mathbf{H}^{(v)}}^T).\end{aligned}\tag{31}$$

Let $\varphi_{ij}^{(v)}$ be the Lagrange multipliers for the constraint $h_{ij}^{(v)}\geq 0$ and $\boldsymbol{\varphi}^{(v)}=[\varphi_{ij}^{(v)}]$, the Lagrange function L for each view can be written as

$$\begin{aligned}L=&tr(\mathbf{X}^{(v)}{\mathbf{X}^{(v)}}^T-2\mathbf{X}^{(v)}{\mathbf{H}^{(v)}}^T{\mathbf{W}^{(v)}}^T\\&+\mathbf{W}^{(v)}\mathbf{H}^{(v)}{\mathbf{H}^{(v)}}^T{\mathbf{W}^{(v)}}^T)+\alpha\sum_{w=1,w\neq v}^{V}tr(\mathbf{H}^{(v)}{\mathbf{H}^{(w)}}^T)\\&+\alpha tr(\mathbf{H}^{(v)}{\mathbf{H}^{(v)}}^T)+\gamma tr(\mathbf{H}^{(v)}\mathbf{L}^{(v)}{\mathbf{H}^{(v)}}^T)+tr(\boldsymbol{\varphi}^{(v)}\mathbf{H}^{(v)}).\end{aligned}\tag{32}$$

Requiring that the derivative of L with respect to $\mathbf{H}^{(v)}$ equals to 0 and using the Karush-Kuhn-Tucker (KKT) condition [30] $\varphi_{ij}^{(v)}h_{ij}^{(v)}=0$, we have

$$h_{ij}^{(v)}\leftarrow h_{ij}^{(v)}\frac{(2{\mathbf{W}^{(v)}}^T\mathbf{X}^{(v)}+2\gamma\mathbf{H}^{(v)}\mathbf{A}^{(v)})_{ij}}{(2{\mathbf{W}^{(v)}}^T\mathbf{W}^{(v)}\mathbf{H}^{(v)}+\alpha\mathbf{Q}^{(v)}+2\gamma\mathbf{H}^{(v)}\mathbf{D}^{(v)})_{ij}},\tag{33}$$

where $\mathbf{Q}^{(v)}=\sum_{w=1,w\neq v}^{V}\mathbf{H}^{(w)}+2\mathbf{H}^{(v)}$.

The whole procedure for solving (30) are summarized in the Algorithm 2.

Algorithm 2 The algorithm of LP-DiNMF

Input:
Data for V views $\{\mathbf{X}^{(1)},\mathbf{X}^{(2)},...,\mathbf{X}^{(V)}\}$.
Parameter α and β.
1: Calculate weighting matrix of each view, $\mathbf{A}^{(v)}$
2: Calculate diagonal matrix and Lagrange matrix of each view, $\mathbf{D}^{(v)}$ and $\mathbf{L}^{(v)}$, respectively
3: **for** $v=1$ to V **do**
4: Normalizing $\mathbf{X}^{(v)}$
5: Initializing $\mathbf{W}^{(v)},\mathbf{H}^{(v)}$
6: **while** not converging **do**
7: Fixing $\mathbf{W}^{(v)}$, updating $\mathbf{H}^{(v)}$ by (33)
8: Fixing $\mathbf{H}^{(v)}$, updating $\mathbf{W}^{(v)}$ by (17)
9: **end while**
10: **end for**
11: Calculate the average value of all data representations of each view by $\mathbf{H}^{*}=\frac{\sum_{v=1}^{V}\mathbf{H}^{(v)}}{V}$.
Output: The final representation matrix $\mathbf{H}^{*}$.

C. Convergence of LP-DiNMF

The Algorithm 2 above is guaranteed to converge to a local minima with the following theorem.

Theorem 2. The objective function in (30) is non-increasing under the update rules in (33) and (17).

Same as DiNMF, we omit the proof of (17) here. To prove (30) is non-increasing under (33), we first rewrite (31) as:

$$\begin{aligned}O_2=&\|\mathbf{X}^{(v)}-\mathbf{W}^{(v)}\mathbf{H}^{(v)}\|_F^2+\alpha\sum_{w=1,w\neq v}^{V}\text{DIVE}(\mathbf{H}^{(v)},\mathbf{H}^{(w)})\\&+\alpha\|\mathbf{H}^{(v)}\|_F^2+\gamma tr(\mathbf{H}^{(v)}\mathbf{L}^{(v)}{\mathbf{H}^{(v)}}^T)\\=&\sum_{i=1}^{m^{(v)}}\sum_{j=1}^{n}(x_{ij}^{(v)}-\sum_{k=1}^{K}w_{ik}^{(v)}h_{kj}^{(v)})^2+\alpha\sum_{w=1,w\neq v}^{V}\sum_{k=1}^{K}\sum_{j=1}^{n}h_{kj}^{(v)}h_{jk}^{(w)}\\&+\alpha\sum_{k=1}^{K}\sum_{j=1}^{n}h_{kj}^{(v)}h_{jk}^{(v)}+\gamma\sum_{k=1}^{K}\sum_{j=1}^{n}\sum_{l=1}^{n}h_{kj}^{(v)}L_{jl}^{(v)}h_{lk}^{(v)}.\end{aligned}\tag{34}$$

It is easy to check that

$$\begin{aligned}F_{ab}^{'}=(\frac{\partial O_2}{\partial\mathbf{H}})_{ab}=&(-2{\mathbf{W}^{(v)}}^T\mathbf{X}^{(v)}+2{\mathbf{W}^{(v)}}^T\mathbf{W}^{(v)}\mathbf{H}^{(v)})_{ab}\\&+(\alpha\sum_{w=1,w\neq v}^{V}\mathbf{H}^{(w)}+2\alpha\mathbf{H}^{(v)}+2\gamma\mathbf{H}^{(v)}\mathbf{L}^{(v)})_{ab}\end{aligned}\tag{35}$$

$$F_{ab}^{''}=(2{\mathbf{W}^{(v)}}^T\mathbf{W}^{(v)})_{aa}+2\alpha\mathbf{I}_{bb}+2\gamma\mathbf{L}_{bb}^{(v)}.\tag{36}$$

Again, we prove each F_{ab} is non-increasing under the update rule (33) based on an auxiliary function as following.

Lemma 3. Let $\mathbf{Q}_{ab}$=$\mathbf{H}_{ab}^{(w)}$+2$\mathbf{H}_{ab}^{(v)}$, the function

$$\begin{aligned}G(h_{ab}^{(v)},{h_{ab}^{(v)}}^t)=&F_{ab}({h_{ab}^{(v)}}^t)+F_{ab}^{'}({h_{ab}^{(v)}}^t)(h^{(v)}-{h_{ab}^{(v)}}^t)\\&+\frac{2({\mathbf{W}^{(v)}}^T\mathbf{W}^{(v)}\mathbf{H}^{(v)})_{ab}+\alpha\mathbf{Q}_{ab}+2\gamma(\mathbf{H}^{(v)}\mathbf{D}^{(v)})_{ab}}{{h_{ab}^{(v)}}^t}(h^{(v)}-{h_{ab}^{(v)}}^t)^2\end{aligned}\tag{37}$$

is an auxiliary function for F_{ab} which is the part of O_2 and only relevant to $h_{ab}^{(v)}$.

Proof. In fact, we can see that Lemma 2 is a part of Lemma 3. Similar to the proof of Lemma 2, we incorporate (35) and (36) to the Taylor series expansion of $F_{ab}^{(h(v))}$ (23) and compare it with (37). Since Lemma 2 has been proved with (25) and (26), here we only need to show

$$\frac{2\gamma(\mathbf{H}^{(v)}\mathbf{D}^{(v)})_{ab}}{{h_{ab}^{(v)}}^t}\geq 2\gamma\mathbf{L}_{bb}^{(v)}.\tag{38}$$

Since we have

$$\begin{aligned}(\mathbf{H}^{(v)}\mathbf{D}^{(v)})_{ab}=&{h_{aj}^{(v)}}^t\sum_{j=1}^{n}\mathbf{D}_{jb}^{(v)}\geq{h_{ab}^{(v)}}^t\mathbf{D}_{bb}^{(v)}\\\geq&{h_{ab}^{(v)}}^t(\mathbf{D}^{(v)}-\mathbf{W}^{(v)})_{bb}={h_{ab}^{(v)}}^t\mathbf{L}_{bb}^{(v)},\end{aligned}\tag{39}$$

(37) holds and $G(h^{(v)},{h_{ab}^{(v)}}^t)\geq F_{ab}(h^{(v)})$.

We can now demonstrate the convergence of Theorem 2.

Proof of Theorem 2. Putting $G(h^{(v)},{h_{ab}^{(v)}}^t)$ of (37) into

(18), we get

$$h_{ab}^{(v)^{t+1}} = h_{ab}^{(v)^t} - h_{ab}^{(v)^t} \frac{F'_{ab}(h_{ab}^{(v)^t})}{(2\mathbf{W}^{(v)^T}\mathbf{W}^{(v)}\mathbf{H}^{(v)} + \alpha\mathbf{Q} + 2\gamma\mathbf{H}^{(v)}\mathbf{D}^{(v)})_{ab}}$$
$$= h_{ab}^{(v)^t} \frac{(2\mathbf{W}^{(v)^T}\mathbf{X}^{(v)} + 2\gamma\mathbf{H}^{(v)}\mathbf{A}^{(v)})_{ab}}{(2\mathbf{W}^{(v)^T}\mathbf{W}^{(v)}\mathbf{H}^{(v)} + \alpha\mathbf{Q} + 2\gamma\mathbf{H}^{(v)}\mathbf{D}^{(v)})_{ab}}. \tag{40}$$

This is in line with (33). Since (37) is an auxiliary function for F_{ab}, F_{ab} is non-increasing under (33).

IV. Complexity Analysis for DiNMF and LP-DiNMF

In DiNMF, for each data matrix $\mathbf{X}^{(v)} \in R^{m^{(v)} \times n}$, the complexity of updating $\mathbf{W}^{(v)}$ in (17) is $O(m^{(v)}nk)$. This is same as that of NMF [31]. The cost of updating $\mathbf{H}^{(v)}$ in (16) is $O(m^{(v)}nk + knV)$. Since usually $V \ll m^{(v)}$, assuming the iterative update stops after t iterations, consequently, the overall computation of DiNMF is $O(\sum_{v=1}^{V}(t(m^{(v)}nk)))$. Clearly, its complexity is linear with respect to the number of data points (n) and it can scale well to large datasets. For LP-DiNMF, the overall cost of updating $\mathbf{W}^{(v)}$ and $\mathbf{H}^{(v)}$ is $O(\sum_{v=1}^{V}(tm^{(v)}nk + m^{(v)}n^2)$ because it requires additional $O(m^{(v)}n^2)$ to construct the nearest neighbor graph. The experimental analysis for both complexity is given in the subsection V-G.

V. Experiment

In this section, we carry out extensive experiments on clustering to demonstrate the effectiveness of DiNMF and LP-DiNMF in exploiting the underlying diverse information across multiple views of data.

A. Description of Datasets

We conduct experiments on one synthetic and several real world datasets, which are chosen from different domains, including documents, images and networks. The descriptions of these datasets are summarized in Table I.

TABLE I: Descriptions of the datasets

Datasets		**Size**	**view**	**Cluster**
Synthetic		5000	2	2
Reuters	Reuters-1	600	3	6
	Reuters-2	18578	5	6
Digit		2000	2	10
WebKB	Cornell	195	2	5
	Texas	187	2	5
	Washington	230	2	5
	Winsconsin	265	2	5
Caltech 101 Silhouettes		8641	2	101

• **Synthetic**: We first randomly generate basis matrices $\{\mathbf{W}^{(i)}\}_{i=1}^{2}$ of two views. The dimensions of two matrices are 250 and 800, respectively. The representation matrices $\{\mathbf{H}^{(i)}\}_{i=1}^{2} \in \mathbb{R}^{20\times 5000}$ are generated with the constraint that the corresponding vectors of these two matrices are orthonormal to each other. To ensure that the two data representations not only contain respective distinct information but also share common information, we sample 30% vectors from one representation matrix by adding Gaussian noise with $\mathcal{N}(0, 1)$ and keep these corresponding vectors exactly same in the second view. Thus, we have a dataset that consists of two views, i.e., $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$, where $\mathbf{X}^{(i)} = \mathbf{W}^{(i)}\mathbf{H}^{(i)}$. This dataset is constructed to demonstrate the correctness of the proposed diversity term and also for the computational speed analysis.

• **Reuters**[1]: As in [20], we randomly sample 100 documents each for 6 clusters, and choose English, French and German as three views to form a dataset. We call it **Reuters-1**. Besides, to demonstrate the performance of the proposed methods on large-scale dataset, we also use the original dataset, called **Reuters-2**. It contains feature characteristics of documents that are translated into five languages over 6 categories. In our experiments, we choose one language, English (EN), as the original language source and take the translated documents in the other four languages as the other four sources.

• **UCI Handwritten Digit**[2] : The dataset is composed of 2000 examples from 0 to 9 ten-digit classes. Each example is represented by two kinds of features, pixel averages in 2×3 windows and Zernike moment.

• **WebKB**[3]: It is composed of web pages collected from computer science department websites of four universities: Cornell, Texas, Washington and Wisconsin. The webpages are classified into 7 categories. Here, we choose four most populous categories (course, faculty, project, student) for clustering. A webpage is made of two views: the text on it and the anchor text on the hyperlinks pointing to it.

• **Caltech 101 Silhouettes**[4]: This dataset is based on the Caltech 101 image annotations [35]. It centers and scales each polygon outline of the primary object in the Caltech 101 and render it on a 16 $\times$ 16 pixel image-plane. The outline is rendered as a filled, black polygon on a white background. Since this dataset contains one type of feature only, following [36], we extracted HOG [16] as the second view.

B. Methods to Compare

We compare the proposed approaches with several representative multi-view clustering methods and their variations.

• **Best Single View-NMF** (**BSV**): We run each view of datasets with NMF [31] and the best single view result is reported.

• **Best Single View-GNMF** (**BSVG**): Similar to **BSV**, we run each view of datasets with GNMF [8] and report the best single view results.

• **Feature Concatenation** (**FeatConcate**): It concatenates the features of all views and applies NMF to extract the low dimensional subspace representation.

[1] http://multilingreuters.iit.nrc.ca
[2] http://archive.ics.uci.edu/ml/datasets/Multiple+Features
[3] http://www.cs.cmu.edu/afs/cs.cmu.edu/project/theo-20/www/data/
[4] https://people.cs.umass.edu/ marlin/data.shtml

TABLE II: Comparison of clustering results ((mean ± standard deviation)%) on the datasets

Metrics	Methods	Synthetic	Reuters-1	Reuters-2	Digit	WebKB				Caltech101 Silhouettes
						Cornell	Texas	Washington	Winsconsin	
AC	NMF	74.45±0.24	45.58±2.66	45.29±1.57	64.30±3.55	42.21±2.19	53.90±2.14	55.30±2.02	48.00±1.89	24.61±0.64
	BSVG	74.11±0.10	46.23±1.88	45.19±0.96	61.15±2.63	43.15±1.23	54.62±1.95	55.74±0.97	49.75±0.75	25.03±0.35
	FeatConcate	78.30±0.12	49.37±2.61	43.27±2.41	48.66±1.27	40.64±1.61	53.32±2.76	54.87±2.82	45.28±1.68	26.06±2.28
	ColNMF	78.33±0.08	41.53±1.68	48.64±1.47	66.90±3.72	40.26±3.79	47.60±2.71	51.17±0.76	41.30±1.49	25.29±0.58
	MultiNMF	71.08±1.22	51.45±1.55	44.37±0.91	61.70±1.80	44.51±2.55	55.29±2.84	50.17±1.50	47.40±0.93	25.89±0.97
	MMNMF	88.00±0.17	52.23±2.01	43.76±0.47	70.65±3.22	46.77±1.22	55.87±2.69	50.09±1.36	49.87±1.12	26.66±0.89
	RMKMC	90.36±0.10	39.48±3.08	40.28±1.19	58.97±1.06	39.99±0.66	54.48±0.92	59.18±0.64	54.21±0.87	25.31±0.51
	CoRegSPC	79.06±1.34	54.13±1.66	N/A	58.41±3.89	42.26±0.50	57.54±0.52	59.74±0.42	48.91±0.26	24.64±0.68
	RMSC	70.26±1.14	48.46±2.32	N/A	67.50±0.00	38.56±0.96	42.67±0.61	41.39±0.91	36.26±1.61	26.51±0.41
	NdNMF	76.20±1.45	45.38±3.25	43.75±0.51	62.16±0.63	47.90±1.10	57.54±0.84	54.74±1.04	56.30±0.91	25.68±0.21
	DiNMF	*98.88±0.46*	*54.17±2.44*	*50.43±1.27*	*72.35±2.05*	*54.46±1.75*	*57.77±1.07*	*59.30±1.15*	*62.03±2.62*	*28.60±0.31*
	LP-DiNMF	**99.96±0.48**	**55.08±2.12**	**51.00±1.83**	**95.20±1.73**	**60.10±2.92**	**59.25**±3.09	**65.39±1.74**	**67.02±1.90**	**29.13±0.26**
NMI	BSV	60.95±0.14	33.75±2.00	29.04±1.14	59.84±0.98	13.59±1.82	24.32±1.95	28.59±1.34	25.10±1.94	50.72±0.31
	BSVG	53.16±0.47	34.06±2.65	28.44±0.53	60.34±2.39	21.73±2.45	24.46±1.73	29.27±1.70	25.49±1.27	51.71±0.08
	FeatConcate	52.51±0.00	31.39±2.25	28.25±2.91	43.30±2.82	16.57±1.52	22.60±2.23	30.49±1.28	21.62±1.85	53.03±0.41
	ColNMF	54.65±0.68	31.55±2.45	32.09±2.21	62.95±1.96	17.42±1.73	20.56±2.69	20.36±2.56	30.17±1.24	52.44±0.82
	MultiNMF	53.29±0.45	36.38±0.81	29.02±1.39	55.44±0.79	12.41±0.36	20.28±1.28	21.57±0.75	13.58±0.90	50.00±0.38
	MMNMF	67.83±0.79	38.71±3.01	29.24±1.23	*73.95±1.83*	26.38±2.90	28.15±2.83	33.52±1.35	29.98±2.82	50.16±0.37
	RMKMC	55.72±0.68	22.03±3.47	22.18±0.14	58.79±3.41	11.79±2.97	17.24±2.34	30.27±1.17	28.42±0.77	52.33±0.26
	CoRegSPC	52.34±0.94	35.77±1.77	N/A	50.68±2.69	14.34±0.00	21.42±0.28	23.54±0.09	15.32±0.00	52.31±0.35
	RMSC	41.89±0.55	29.75±1.61	N/A	62.00±0.00	14.52 ±1.15	17.95 ±0.89	18.85±0.95	14.48±0.24	53.83±0.32
	NdNMF	44.08±0.72	21.95±4.09	22.95±1.61	56.65±2.83	22.81±1.57	20.66±3.14	29.75±3.02	33.18±2.71	52.55±0.84
	DiNMF	*80.23±0.79*	*21.50±2.31*	*29.27±1.50*	*66.60±1.63*	*33.31±1.78*	*37.84±2.51*	*38.20±1.43*	*43.00±1.79*	*54.60±0.02*
	LP-DiNMF	**83.12**±0.46	**38.52±1.49**	**32.14±1.22**	**90.45±1.27**	**32.65±3.04**	**40.57±2.51**	**42.39±1.39**	**43.03±2.80**	**55.41±0.34**
Purity	BSV	79.45±0.24	47.41±2.41	48.35±1.01	67.63±2.97	45.77±1.44	64.55±1.09	67.20±1.13	51.70±1.96	43.58±0.22
	BSVG	80.11±0.10	48.48±2.32	51.73±0.70	69.19±3.83	53.41±2.51	66.88±0.70	67.87 ±1.40	62.55±1.43	44.37±0.13
	FeatConcate	79.30±0.15	50.20±2.80	48.33±3.53	51.55±2.73	49.54±2.81	64.18±1.19	66.35±2.62	73.21±2.91	45.89±0.31
	ColNMF	78.33±0.05	44.14±2.14	44.57±0.46	71.11±1.07	49.44±1.16	62.74±1.13	68.43±1.44	61.00±1.25	45.26±0.35
	MultiNMF	70.08±1.22	53.60±1.31	54.50±1.14	63.21±1.22	45.18±1.09	64.01±1.07	67.83±1.28	59.55±1.21	44.89±0.48
	MMNMF	88.00±0.17	53.70±2.39	53.43±0.75	74.15±2.73	57.95±2.80	67.43±3.04	70.43±0.98	67.66±2.09	41.94±0.29
	RMKMC	90.16±0.62	40.63±3.28	40.22±0.32	63.87±3.43	46.59±2.70	60.99±1.66	68.91±1.45	65.81±1.64	45.15±0.24
	CoRegSPC	77.54±1.32	**55.47±1.70**	N/A	58.66±1.43	44.97±0.25	64.54±0.52	65.48±0.22	52.72±0.36	45.35±0.58
	RMSC	69.36±0.37	49.98±2.25	N/A	68.60±0.78	49.28±1.65	60.67±2.61	63.26±1.11	56.23±0.53	47.30±0.41
	NdNMF	76.06±0.45	46.57±3.35	47.25±1.57	64.23±2.02	57.36±2.01	63.03±1.66	67.57±2.78	68.40±2.91	44.89±0.23
	DiNMF	*98.36±0.52*	54.65±2.04	*54.55±1.27*	*74.45±0.49*	*63.82±2.37*	*69.51±1.39*	*70.96±2.03*	*73.48±2.11*	*47.69±0.52*
	LP-DiNMF	**98.96±0.42**	*55.15±1.51*	**56.39±1.79**	**95.20±1.27**	**67.13±2.29**	**72.62±3.39**	**72.78±1.01**	**74.19±2.18**	**48.37±0.30**

• **ColNMF** [37]: It simultaneously factors data matrices of multiple views to different basis matrices with the shared consensus coefficient matrix.

• **MultiNMF** [20]: It searches for a compatible clustering solutions across multiple views by minimizing the differences between data representation matrices of each view and the consensus matrix.

• **MMNMF** [21]: It preserves the locally geometrical structure of the manifolds for multi-view clustering with regarding that the intrinsic manifold of the dataset is embedded in a convex hull of all the views' manifolds, and incorporates such an intrinsic manifold and an intrinsic coefficient matrix with a multi-manifold regularizer.

• **RMKMC** [38]: This multi-view k-means approach integrates heterogeneous features of data and utilizes the common cluster indicator to do clustering across multiple views. $l_{2,1}$-norm is employed to improve the robustness.

• **CoRegSPC** [39]: This pairwise multi-view spectral clustering method co-regularizes the clustering hypotheses to enforce corresponding data points in each view to have the same cluster membership.

• **RMSC** [40]: This is a multi-view spectral clustering method based on low rank and sparse decomposition of the transition matrix.

• **NdNMF**: It conducts each view independently using standard NMF [6], and then applies k-means on the combination of new representations of each view.

C. Settings

For each compared method, we set the parameters according to original papers where the approaches were first proposed. As BSVG, MMNMF and LP-DiNMF require construction of the nearest neighbor graph, we set the number of nearest neighbor equal to the number of classes of the data k, as suggested in [21]. For DiNMF and LP-DiNMF, we normalize the data first and then initialize both $\mathbf{W}^{(v)}$ and $\mathbf{H}^{(v)}$ for each view in the range [0,1]. Similar to [41], [42], the regularization parameters (α, β in (9) and α, γ in (30)) are chosen from $\{0.0001,0.001,0.01,0.1,1,10,100,1000\}$. To avoid randomness, we run each method 10 times with different initializations and report the average results and their standard deviations. The clustering results are evaluated by three widely adopted evaluation metrics, including accuracy (AC) [43], normalized mutual information (NMI) [43] and Purity [44]. Each metric favors different properties in clustering, and hence we report results on these measures to perform a more comprehensive evaluation. For all these metrics, the higher value indicates better clustering quality.

D. Clustering Results

Table II demonstrates the average results and standard deviations for each method on the datasets. Note that, the results of CoRegSPC and RMSC on Reuters-2 are not available (N/A) since they demand huge memory. In each row of the table, the best result is highlighted in **boldface** and the second best result in *italic*. It is clear to see that both DiNMF and LP-DiNMF consistently outperform the other methods, sometimes even very significantly, which demonstrates the advantage of our approaches in terms of clustering performance. Compared with NdNMF, DiNMF improves performances more than 5% on all datasets in terms of AC, NMI and Purity, which proves the effectiveness of the proposed diversity constraints. We also notice that directly concatenating all the features (i.e., FeatConcate) is not an ideal approach since it always performs worse than the best single view (BSV). Moreover, LP-DiNMF performs better than DiNMF on all the datasets. This indicates that exploiting the geometric structures in data spaces indeed can improve the cluster performance, also verifies the manifold assumption and confirms the correctness of our approaches.

E. Analysis of Redundancy Rate

To verify that DIVE reduces the redundancy information among multiple representations, we propose a redundancy rate (RED) metric as follows:

$$\text{RED}(\mathbf{H}^{(1)},...,\mathbf{H}^{(V)}) = \frac{\sum_{i=1}^{n}\sum_{v=1,v\neq w}^{V} cos^2(\mathbf{h}_i^{(v)},\mathbf{h}_i^{(w)})}{V(V-1)n}. \tag{41}$$

It assesses the average sum of similarity of all n data vectors in all pairs of views and ranges from 0 to 1, where 0 means a completely complementary result, and 1 vice versa.

We compare the redundancy rate of the proposed approaches against MultiNMF, MMNMF and NdNMF, which are all under the framework of NMF and then take the same approach to obtain the final multi-view representation matrix $\mathbf{H}^*(=\frac{\sum_{v=1}^{V}\mathbf{H}^{(v)}}{V})$. The results of comparison are reported in Table III.

It can be seen that MultiNMF always gets the highest rate followed by MMNMF and NdNMF, while it is less than 20% for DiNMF in all cases. This demonstrates the effectiveness of the proposed DIVE that enforces the complementarity across multiple views. However, LP-DiNMF does not always achieve stable and low redundancy rate. For example, it gets the lowest redundancy rate in Texas with 0.1222 compared with other approaches, but a higher rate (0.1852) than DiNMF in Winsconsin. This is because the representations of multiple views in LP-DiNMF are co-regularized by both the manifold structure and the diversity term. There is a tradeoff between the two regularization terms. Thus, different from DiNMF which is only regularized by the diversity term, LP-DiNMF is less likely to get the lowest rate.

To have a visual perception of redundancy, we take the Digit (2 views) and Reuters-1 (3 views) as examples and demonstrate the redundancy rate of each data vector in details, as shown in Figure 2. The horizontal axis represents the number of data points and the vertical axis means the scaled redundancy rate. For each approach, the scaled redundancy rate is the percentage of its true redundancy rates over that of all five approaches. Each method is represented by one color. The wider area a color occupies, the more redundant information an approach has. Figure 2 shows that DiNMF (marked in purple) occupies the narrowest area, while MultiNMF occupies the widest area in both Digit and Reuters datasets. The results of Figure 2 is inline with Table III, which proves that DiNMF

TABLE III: Comparison of redundancy rate

Methods	Synthetic	Reuters-1	Digit	Cornell	Texas	Washington	Winsconsin
MultiNMF	0.9986	0.9970	0.5826	0.8503	0.8472	0.8229	0.8521
MMNMF	0.5998	0.4800	0.4437	0.3440	0.4318	0.3598	0.3698
NdNMF	0.4637	0.2658	0. 2755	0.2395	0.2077	0.2683	0.1122
DiNMF	**0.1838**	**0.1087**	**0.1931**	**0.0651**	0.1873	**0.0609**	**0.0783**
LP-DiNMF	0.3509	0.1266	0.2663	0.0894	**0.1222**	0.1013	0.1852

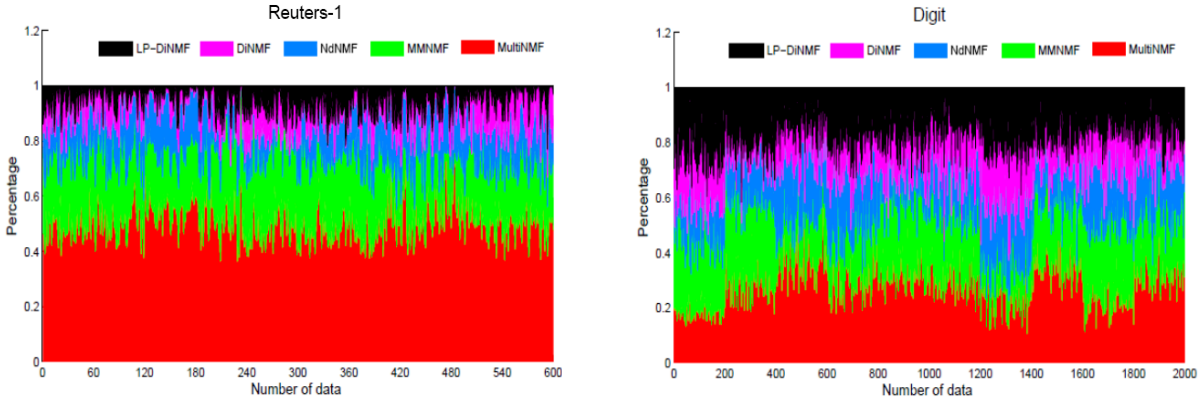
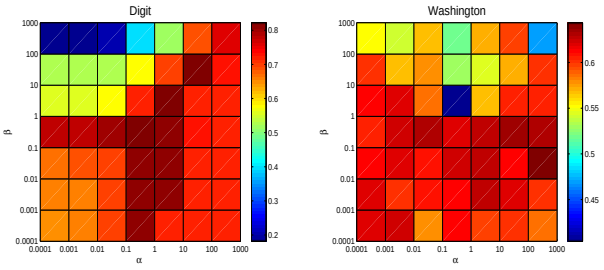
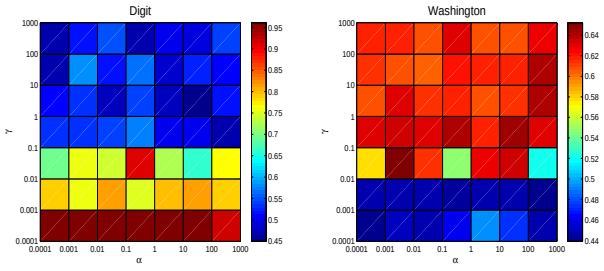


Fig. 2: Comparison of redundancy rate on Reuters-1 and Digit dataset

effectively exploits the diverse information across multiple views.



(a) DiNMF



(b) LP-DiNMF

Fig. 3: The effect of parameter α and β in DiNMF and α and γ in LP-DiNMF. Different colors means different accuracies and the color close to red indicates high accuracy.

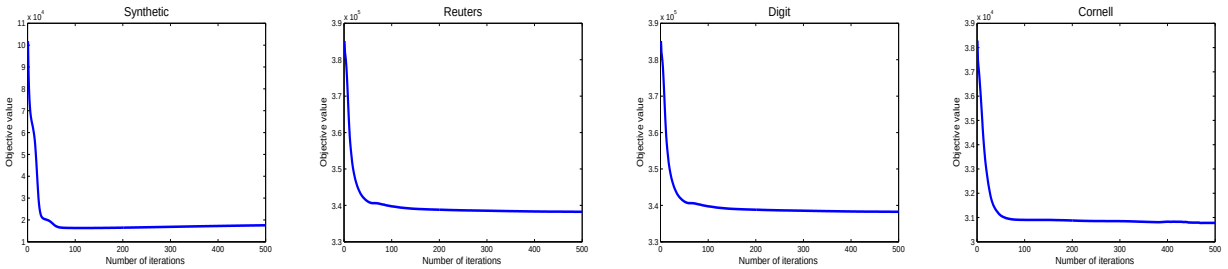
F. Parameter Study

We tested the effect of the parameters α and β of DiNMF, as well as α and γ of LP-DiNMF. In DiNMF α and β affect the diversity and smoothness, while in LP-DiNMF, α and γ adjust the effects of the diversity and graph regularization term. For both methods, we picked the value of each parameter from $\{0.0001, 0.001, 0.01, 0.1, 1, 10, 100, 1000\}$. Taking the Digit and Washington as examples, we can find that DiNMF in Figure 3(a) achieves more than 70% accuracy in Digit and 60% in Washington for α and β in most cases, demonstrating that the the performance of DiNMF is relatively robust to parameter tuning. Figure 3(b) shows that LP-DiNMF is relatively stable with varying α, but significantly affected by γ. This further verified the importance of preserving manifold structure.

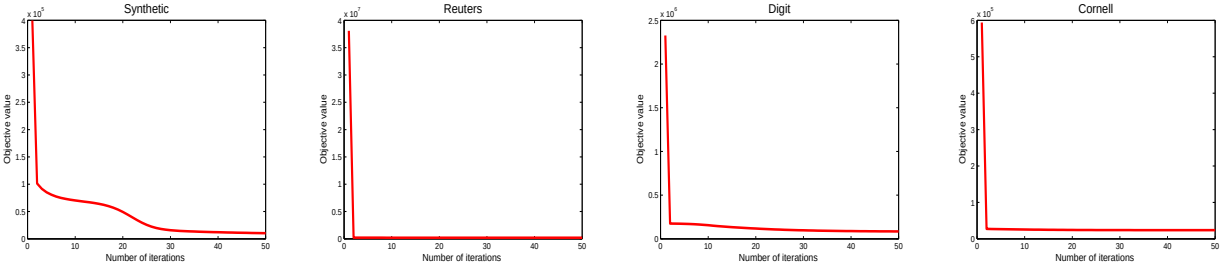
G. Study of Computational Speed

We have proven the convergence of our update rules and analyzed the computational complexity of DiNMF and LP-DiNMF against MMNMF in previous sections. Here our experiments demonstrate their convergence curves in Figure 4 and computational time in Figure 5. All our experiments are conducted on a PC with two octa-core Intel Xeon CPU processors at 2.5 GHz and 256G bytes memory.

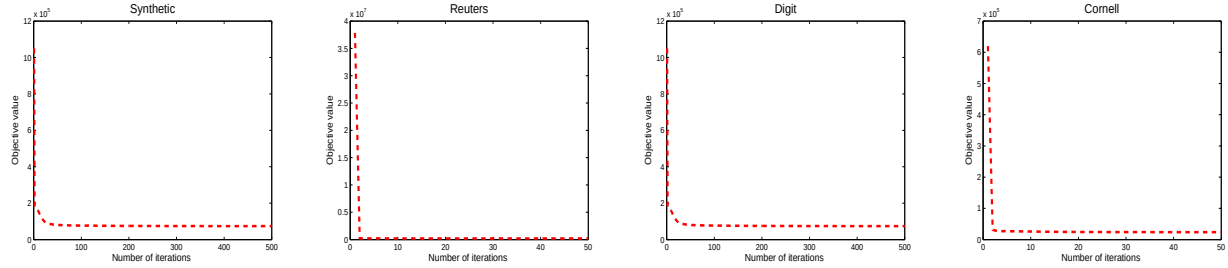
Because the results of different networks datasets (Cornell, Texas, Washington and Winsconsin) have similar convergency, here we just took one network (Cornell) as an example. Figure 4 shows the convergence curve of the three methods on Synthetic, Reuters, Digit and Cornell. For each figure, the



(a) MMNMF

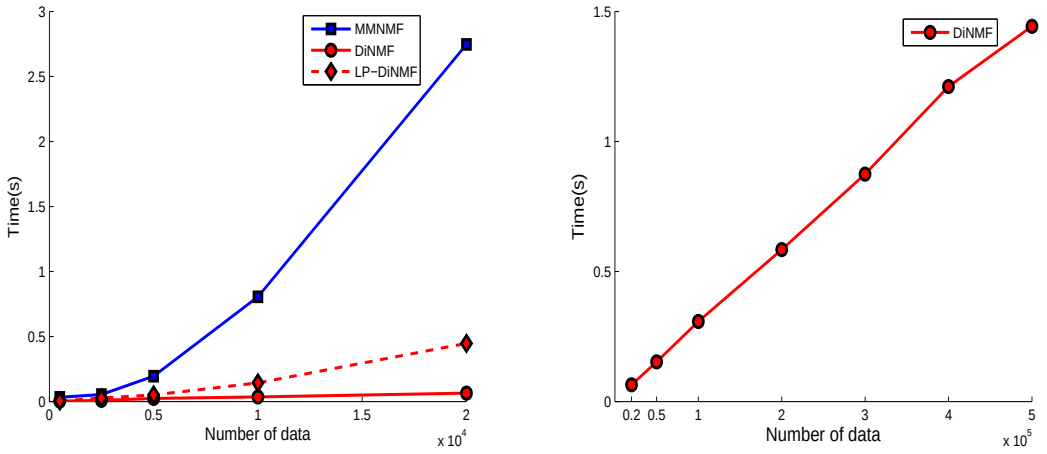


(b) DiNMF



(c) LP-DiNMF

Fig. 4: Comparison of convergence speed (Note that different scales of axes are used for clearer illustration)



(a) (b)

Fig. 5: Running time of DiNMF v.s. MMNMF on Synthetic dataset.

horizontal axis is the number of iterations and the vertical axis is the value of objective function. We can see that MMNMF (Figure 4(a)) needs around 100 iterations for each dataset, while DiNMF (Figure 4(b)) is the most efficient, since the objective function values are non-increasing and drop sharply within a small number of iterations (10 iterations) in all cases. Although LP-DiNMF (Figure 4(c)) requires nearly 100 iterations for the Synthetic and Digit database, its objective values drop faster than that of MMNMF. This empirically proves our convergence theory.

As discussed in section IV, DiNMF has linear time complexity with the number of data points. Here, we verify this claim on the Synthetic dataset. Figure 5 reports the average running time of each iteration of three methods on the Synthetic dataset. The default setting is 5000 data points, 2 clusters, and 2 views. During the experiment, we fix the number of clusters and views but change the number of data. Figure 5 (a) shows the running time of three methods in terms of varying data points within $\{0.05, 0.25, 0.5, 1, 1.5, 2\} \times 10^4$. Clearly, DiNMF is linear in execution time, and MMNMF costs significantly more time than DiNMF and LP-DiNMF. To better demonstrate DiNMF's linearity and good scalability to large datasets, we increase the amount of data to a large scale, i.e., $\{0.2, 0.5, 1, 2, 3, 4, 5\} \times 10^5$ and report corresponding running time each in Figure 5 (b). Clearly, the results are in line with the analysis in subsection IV.

VI. Conclusion

In this paper, we have advanced the frontier of NMF by proposing a novel idea that explores diverse information among multi-view representations. To achieve this, we have proposed a Diverse Nonnegative Matrix Factorization (DiNMF) approach for more comprehensive and accurate multi-view learning. With a novel diversity regularization term, DiNMF explicitly enforces the orthogonality of different data representations. Importantly, DiNMF converges linearly and scales well with large-scale data. Taking a step further, we have extended DiNMF by incorporating manifold information and proposed Locality Preserved DiNMF (LP-DiNMF) method. Extensive experiments conducted on both synthetic and benchmark datasets have demonstrated promising results of our methods, which conform to our theoretical analysis. For future work, we aim to study diversity in Nonnegative Tensor Factorization, the nature generalization of NMF to higher dimensions, with which a wider range of applications can be expected.

References

[1] Y.-H. Xiao, Z.-F. Zhu, Y. Zhao, and Y.-C. Wei, "Class-driven nonnegative matrix factorization for image representation," *Journal of Computer Science and Technology*, vol. 28, no. 5, pp. 751–761, 2013.

[2] D. Cai, X. He, X. Wu, and J. Han, "Non-negative matrix factorization on manifold," in *Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on*. IEEE, 2008, pp. 63–72.

[3] Q.-L. Lin, B. Sheng, Y. Shen, Z.-F. Xie, Z.-H. Chen, and L.-Z. Ma, "Fast image correspondence with global structure projection," *Journal of Computer Science and Technology*, vol. 27, no. 6, pp. 1281–1288, 2012.

[4] X. He, S. Yan, Y. Hu, P. Niyogi, and H.-J. Zhang, "Face recognition using laplacianfaces," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 27, no. 3, pp. 328–340, 2005.

[5] X. Niyogi, "Locality preserving projections," in *Neural information processing systems*, vol. 16, 2004, p. 153.

[6] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, 1999.

[7] D. Kong, C. Ding, and H. Huang, "Robust nonnegative matrix factorization using l21-norm," in *Proceedings of the 20th ACM international conference on Information and knowledge management*. ACM, 2011, pp. 673–682.

[8] D. Cai, X. He, J. Han, and T. S. Huang, "Graph regularized nonnegative matrix factorization for data representation," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 33, no. 8, pp. 1548–1560, 2011.

[9] J. J.-Y. Wang, H. Bensmail, and X. Gao, "Multiple graph regularized nonnegative matrix factorization," *Pattern Recognition*, vol. 46, no. 10, pp. 2840–2847, 2013.

[10] P. Li, J. Bu, C. Chen, Z. He, and D. Cai, "Relational multimanifold coclustering," *IEEE Transactions on cybernetics*, vol. 43, no. 6, pp. 1871–1881, 2013.

[11] P. Li, J. Bu, C. Chen, C. Wang, and D. Cai, "Subspace learning via locally constrained a-optimal nonnegative projection," *Neurocomputing*, vol. 115, pp. 49–62, 2013.

[12] J. J.-Y. Wang, J. Z. Huang, Y. Sun, and X. Gao, "Feature selection and multi-kernel learning for adaptive graph regularized nonnegative matrix factorization," *Expert Systems with Applications*, vol. 42, no. 3, pp. 1278–1286, 2015.

[13] J. J.-Y. Wang and X. Gao, "Beyond cross-domain learning: Multiple-domain nonnegative matrix factorization," *Engineering Applications of Artificial Intelligence*, vol. 28, pp. 181–189, 2014.

[14] C. Xu, D. Tao, and C. Xu, "A survey on multi-view learning," *arXiv preprint arXiv:1304.5634*, 2013.

[15] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.

[16] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, vol. 1. IEEE, 2005, pp. 886–893.

[17] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *International journal of computer vision*, vol. 42, no. 3, pp. 145–175, 2001.

[18] C.-T. Nguyen, D.-C. Zhan, and Z.-H. Zhou, "Multi-modal image annotation with multi-instance multi-label lda," in *Proceedings of the Twenty-Third international joint conference on Artificial Intelligence*. AAAI Press, 2013, pp. 1558–1564.

[19] C. Xu, D. Tao, and C. Xu, "Multi-view self-paced learning for clustering," in *Proceedings of the 24th International Conference on Artificial Intelligence. AAAI Press*, 2015, pp. 3974–3980.

[20] J. Liu, C. Wang, J. Gao, and J. Han, "Multi-view clustering via joint nonnegative matrix factorization," in *Proc. of SDM*, vol. 13. SIAM, 2013, pp. 252–260.

[21] X. Zhang, L. Zhao, L. Zong, X. Liu, and H. Yu, "Multi-view clustering via multi-manifold regularized nonnegative matrix factorization," in *Data Mining (ICDM), 2014 IEEE International Conference on*. IEEE, 2014, pp. 1103–1108.

[22] M. M. Kalayeh, H. Idrees, and M. Shah, "Nmf-knn: Image annotation using weighted multi-view non-negative matrix factorization," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2014, pp. 184–191.

[23] W. Ou, S. Yu, G. Li, J. Lu, K. Zhang, and G. Xie, "Multi-view non-negative matrix factorization by patch alignment framework with view consistency," *Neurocomputing*, 2016.

[24] J. Wang, X. Wang, F. Tian, C. H. Liu, H. Yu, and Y. Liu, "Adaptive multi-view semi-supervised nonnegative matrix factorization," in *International Conference on Neural Information Processing*. Springer, 2016, pp. 435–444.

[25] A. Blum and T. Mitchell, "Combining labeled and unlabeled data with co-training," in *Proceedings of the eleventh annual conference on Computational learning theory*. ACM, 1998, pp. 92–100.

[26] K. Chaudhuri, S. M. Kakade, K. Livescu, and K. Sridharan, "Multi-view clustering via canonical correlation analysis," in *Proceedings of the 26th annual international conference on machine learning*. ACM, 2009, pp. 129–136.

[27] W. Wang and Z.-H. Zhou, "Analyzing co-training style algorithms," in *Machine Learning: ECML 2007*. Springer, 2007, pp. 454–465.

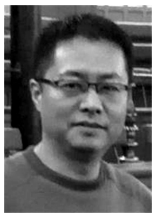
[28] P. Paatero and U. Tapper, "Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values," *Environmetrics*, vol. 5, no. 2, pp. 111–126, 1994.
[29] X. Guo, "Exclusivity regularized machine," *arXiv preprint arXiv:1603.08318*, 2016.
[30] S. P. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge university press, 2004.
[31] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Advances in neural information processing systems*, 2001, pp. 556–562.
[32] M. Belkin, P. Niyogi, and V. Sindhwani, "Manifold regularization: A geometric framework for learning from labeled and unlabeled examples," *The Journal of Machine Learning Research*, vol. 7, pp. 2399–2434, 2006.
[33] F. Shang, L. Jiao, J. Shi, and J. Chai, "Robust positive semidefinite l-isomap ensemble," *Pattern Recognition Letters*, vol. 32, no. 4, pp. 640–649, 2011.
[34] P. Li, J. Bu, B. Xu, B. Wang, and C. Chen, "Locally discriminative spectral clustering with composite manifold," *Neurocomputing*, vol. 119, pp. 243–252, 2013.
[35] L. Fei-Fei, R. Fergus, and P. Perona, "One-shot learning of object categories," *IEEE transactions on pattern analysis and machine intelligence*, vol. 28, no. 4, pp. 594–611, 2006.
[36] X. Cai, F. Nie, W. Cai, and H. Huang, "Heterogeneous image features integration via multi-modal semi-supervised learning model," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 1737–1744.
[37] A. P. Singh and G. J. Gordon, "Relational learning via collective matrix factorization," in *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2008, pp. 650–658.
[38] X. Cai, F. Nie, and H. Huang, "Multi-view k-means clustering on big data," in *Proceedings of the Twenty-Third international joint conference on Artificial Intelligence*. AAAI Press, 2013, pp. 2598–2604.
[39] A. Kumar, P. Rai, and H. Daume, "Co-regularized multi-view spectral clustering," in *Advances in Neural Information Processing Systems*, 2011, pp. 1413–1421.
[40] R. Xia, Y. Pan, L. Du, and J. Yin, "Robust multi-view spectral clustering via low-rank and sparse decomposition," in *AAAI Conference on Artificial Intelligence*, 2014.
[41] H. Wang, F. Nie, and H. Huang, "Multi-view clustering and feature learning via structured sparsity." in *ICML (3)*, 2013, pp. 352–360.
[42] J. Huang, F. Nie, H. Huang, and C. Ding, "Robust manifold nonnegative matrix factorization," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 8, no. 3, p. 11, 2014.
[43] H. Liu, Z. Wu, X. Li, D. Cai, and T. S. Huang, "Constrained nonnegative matrix factorization for image representation," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 34, no. 7, pp. 1299–1311, 2012.
[44] C. Ding, T. Li, W. Peng, and H. Park, "Orthogonal nonnegative matrix t-factorizations for clustering," in *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2006, pp. 126–135.

Jing Wang received the B.Eng. degree in electronics and information technology from Anhui University of Technology, Ma'anshan, China, in 2010 and the M.Sc. degree in multimedia information technology from City University of Hong Kong, Hong Kong, in 2012. She is currently pursuing the Ph.D. degree with the Faculty of Science and Technology, Bournemouth University, Bournemouth, U.K.

Her current research interests include machine learning, computer vision and data mining.

Feng Tian received the Ph.D. degree in Mechanical Engineering from Xi'an Jiaotong University, Xi'an, China in 1997.

He is currently an Associate Professor of Media Technology in Bournemouth University, Bournemouth, U.K. He was an Assistant Professor in Nanyang Technological University in Singapore. His current research interests include computer graphics, computer animation, games technology, augmented reality, image processing.



Hongchuan Yu (M'00) received the Ph.D. degree in Computer Vision from Chinese Academy of Sciences, Beijing, China, in 2000.

He is currently a Senior Lecturer of Computer Graphics in National Centre for Computer Animation, Bournemouth University, Bournemouth, U.K. As investigator, he has secured over 2 million in research grants from EU FP7, EU H2020, Royal Society etc. He has published more than 60 academic articles in reputable journals and conferences. His current research interests include image processing, pattern recognition and computer graphics.

Dr. Yu is a fellow of High Education of Academy United Kingdom. He regularly served as PC members/referees for IEEE journals and conferences, including IEEE Transactions on Pattern Analysis and Machine Intelligence, IEEE Transactions on Image Processing, IEEE Transactions on Visualization and Computer Graphics.

Chang Hong Liu received the Ph.D. degree in cognitive psychology from the University of Toronto, Toronto, Canada, in 1995.

He is currently a Professor of Psychology at Bournemouth University, Bournemouth, U.K. His current research interests include human face recognition, perception of facial expression and attractiveness, attention, memory and cognition.

Kun Zhan received the B.S. and Ph.D. degrees from School of Information Science and Engineering, Lanzhou University, China, in 2005 and 2010, respectively. He was a visiting student with the Department of Electrical and Computer Engineering, Dalhousie University, Halifax, Canada, from 2009 to 2010. Currently, He works in Lanzhou University. His main research interests include machine learning and neural networks.

Xiao Wang received the Ph.D. degree from the School of Computer Science and Technology at Tianjin University, Tianjin, China, in 2016.

He is currently a postdoctoral with the Department of Computer Science and Technology, Tsinghua University, Beijing, China. He got the China Scholarship Council Fellowship in 2014 and visited Washington University in St. Louis, USA, as a joint training student from 2014 to 2015. His current research interests include complex network analysis,