# Multiple Incomplete Views Clustering via Weighted Nonnegative Matrix Factorization with $L_{2,1}$ Regularization

Weixiang Shao[1], Lifang He[2(✉)], and Philip S. Yu[1,3]

[1] Department of Computer Science,
University of Illinois at Chicago, Chicago, IL, USA
[2] Institute for Computer Vision, Shenzhen University, Shenzhen, China
lifanghescut@gmail.com
[3] Institute for Data Science, Tsinghua University, Beijing, China

**Abstract.** With the advance of technology, data are often with multiple modalities or coming from multiple sources. Multi-view clustering provides a natural way for generating clusters from such data. Although multi-view clustering has been successfully applied in many applications, most of the previous methods assumed the completeness of each view (*i.e.*, each instance appears in all views). However, in real-world applications, it is often the case that a number of views are available for learning but none of them is complete. The incompleteness of all the views and the number of available views make it difficult to integrate all the incomplete views and get a better clustering solution. In this paper, we propose MIC (Multi-Incomplete-view Clustering), an algorithm based on weighted nonnegative matrix factorization with $L_{2,1}$ regularization. The proposed MIC works by learning the latent feature matrices for all the views and generating a consensus matrix so that the difference between each view and the consensus is minimized. MIC has several advantages comparing with other existing methods. First, MIC incorporates weighted nonnegative matrix factorization, which handles the missing instances in each incomplete view. Second, MIC uses a co-regularized approach, which pushes the learned latent feature matrices of all the views towards a common consensus. By regularizing the disagreement between the latent feature matrices and the consensus, MIC can be easily extended to more than two incomplete views. Third, MIC incorporates $L_{2,1}$ regularization into the weighted nonnegative matrix factorization, which makes it robust to noises and outliers. Forth, an iterative optimization framework is used in MIC, which is scalable and proved to converge. Experiments on real datasets demonstrate the advantages of MIC.

## 1 Introduction

With the advance of technology, real data are often with multiple modalities or coming from multiple sources. Such data is called multi-view data. Different views may emphasize different aspects of the data. Integrating multiple views may help


A. Appice et al. (Eds.): ECML PKDD 2015, Part I, LNAI 9284, pp. 318–334, 2015.
DOI: 10.1007/978-3-319-23528-8_20

improve the clustering performance. For example, one news story may be reported by different news sources, user group can be formed based on users' profiles, user's online social connections, users' transaction history or users' credit score in online shopping recommendation system, one patient can be diagnosed with a certain disease based on different measures, including clinical, imaging, immunologic, serological and cognitive measures. Different from traditional data with a single view, these multi-view data commonly have the following properties:

1. Each view can have its own feature sets, and each view may emphasize different aspects. Different views share some consistency and complementary properties. For example, in online shopping recommendation system, user's credit score has numerical features while users' online social connections provide graph relational features. The credit score emphasizes the creditworthiness of the user, while the social connection emphasizes the social life of the user.
2. Each view may suffer from incompleteness. Due to the nature of the data or the cost of data collection, each available view may suffer from incompleteness of information. For example, not all the news stories are covered by all the news sources, *i.e.*, each news source (view) cannot cover all the news stories. Thus, all the views are incomplete.
3. There may be an arbitrary number of sources. In some applications, the number of available views may be small, while in other applications, it may be quite large.

The above properties raise two fundamental challenges for clustering multi-view data:

1. How to combine various number of views to get better clustering solutions by exploring the consistency and complementary properties of different views.
2. How to deal with the incompleteness of the views, *i.e.*, how to effectively and efficiently get better clustering solutions even all of the views are incomplete.

Multi-view clustering [1,7] provides a natural way for generating clusters from such data. A number of approaches have been proposed for multi-view clustering. Existing multi-view clustering algorithms can be classified into two categories according to [28], distributed approaches and centralized approaches. Distributed approaches, such as [4,15,28] first cluster each view independently from the others, using an appropriate single-view algorithm, and then combine the individual clusterings to produce a final clustering result. Centralized approaches, such as [1,5,24,38] make use of multiple representations simultaneously to mine hidden patterns from the data. In this paper, we mainly focus on the centralized approaches.

Most of the previous studies on multi-view clustering focus on the first challenge. They are all based on the assumption that all of the views are complete, *i.e.*, each instance appears in all views. Few of them addresses how to deal with the second challenge. Recently, there are several methods working on the incompleteness of the views [26,32,34]. They either require the completeness of at

least one base view or cannot be easily extended to more than two incomplete views. However, in real-world applications, it is often the case that more than two views are available for learning and none of them is complete. For example, in document clustering, we can have documents translated into different languages representing multiple views. However, we may not get all the documents translated into each language. Another example is medical diagnosis. Although multiple measurements from a series of medical examinations may be available for a patient, it is not realistic to have each patient complete all the potential examinations, which may result in the incompleteness of all the views. The incompleteness of all the views and the number of available views make it difficult to directly integrate all the incomplete views and get a better clustering solution.

In this paper, we propose MIC (Multi-Incomplete-view Clustering) to handle the situation of multiple incomplete views by integrating the joint weighted nonnegative matrix factorization and $L_{2,1}$ regularization. Weighted nonnegative matrix factorization [20] is a weighted version of nonnegative matrix factorization [25], and has been successfully used in document clustering [35] and recommendation system [16]. $L_{2,1}$ norm of a matrix was first introduced in [9] as rotational invariant $L_1$ norm. Because of its robustness to noise and outliers, $L_{2,1}$ has been widely used in many areas [11,13,18,21]. By integrating weighted nonnegative matrix factorization and $L_{2,1}$ norm, MIC tries to learn a latent subspace where the features of the same instance from different views will be co-regularized to a common consensus, while increasing the robustness of the learned latent feature matrices. The proposed MIC method has several advantages comparing with other state-of-art methods:

1. MIC incorporates weighted nonnegative matrix factorization, which will handle the missing instances in each incomplete view. A weight matrix for each incomplete view is introduced to give the missing instances lower weights than the presented instances in each view.
2. By using a co-regularized approach, MIC pushes the learned latent feature matrices to a common consensus. Because MIC only regularizes the difference between the learned latent feature for each view and the consensus, MIC can be easily extended to more than two incomplete views.
3. MIC incorporates $L_{2,1}$ norm into the weighted nonnegative matrix factorization. $L_{2,1}$ regularization added to the objective function will keep the learned latent feature matrices more robust to noises and outliers, which is naturally perfect for the situation of multiple incomplete views.
4. An iterative optimization framework is used in MIC, which is scalable and proved to converge.

The rest of this paper is organized as follows. In the next section, notations and problem formulation are given. The proposed MIC algorithm is then presented in Section 3. Extensive experimental results and analysis are shown in Section 4. Related work is given in Section 5 followed by conclusion in Section 6.

**Table 1.** Summary of the Notations

| Notation | Description |
|---|---|
| $N$ | Total number of instances. |
| $n_v$ | Total number of views. |
| $\mathbf{X}^{(i)}$ | Data matrix for the $i$-th view. |
| $d_i$ | Dimension of features in the $i$-th view. |
| $\mathbf{M}$ | The indicator matrix, where $\mathbf{M}_{i,j} = 1$ indicates $j$-th instance appears in $i$-th view. |
| $\mathbf{W}^{(i)}$ | The diagonal instance weight matrix for the $i$-th view. |
| $\mathbf{U}^{(i)}$ | The latent feature matrix for the $i$-th view. |
| $\mathbf{V}^{(i)}$ | The basis matrix for the $i$-th view. |
| $\mathbf{U}^*$ | The common consensus, latent feature matrix across all the views. |
| $\alpha_i$ | Trade-off parameter between reconstruction error and view disagreement for view $i$. |
| $\beta_i$ | Trade-off parameter between reconstruction error and robustness for view $i$. |

## 2 Problem Formulation and Backgrounds

In this section, we will briefly describe the problem formulation. Then the background knowledge on weighted nonnegative matrix factorization will be introduced.

### 2.1 Problem Formulation

Before we describe the formulation of the problem, we summarize some notations used in this paper in Table 1. Assume we are given a dataset with $N$ instances and $n_v$ views $\{\mathbf{X}^{(i)}, i = 1, 2, ..., n_v\}$, where $\mathbf{X}^{(i)} \in \mathbb{R}^{N \times d_i}$ represents the dataset in view $i$. We define an indicator matrix $\mathbf{M} \in \mathbb{R}^{n_v \times N}$ by,

$$\mathbf{M}_{i,j} = \begin{cases} 1 & \text{if } j\text{-th instance is in the } i\text{-th view.} \\ 0 & \text{otherwise.} \end{cases}$$

where each row of $\mathbf{M}$ represent the instance presence for one view. Most of the previous methods on multi-view clustering assume the completeness of all the views. Every view contains all the instances, *i.e.*, $\mathbf{M}$ is an all one matrix, $\sum_{j=1}^{N} \mathbf{M}_{i,j} = N,\ i = 1, 2, ..., n_v$. However, in most real-world situations, one instance may only appear in some of the views, which may result in the incompleteness of all the views. For each view, the data matrix $\mathbf{X}^{(i)}$ will have a number of rows missing, *i.e.*, $\sum_{j=1}^{N} \mathbf{M}_{i,j} < N,\ i = 1, 2, ..., n_v$.

Our goal is to cluster all the $N$ instances into $K$ clusters by integrating all the $n_v$ incomplete views.

### 2.2 Weighted Nonnegative Matrix Factorization

Let $\mathbf{X} \in \mathbb{R}_+^{N \times M}$ denote the nonnegative data matrix where each row represents an instance and each column represents one attribute. Weighted nonnegative

matrix factorization [20] aims to factorize the data matrix $\mathbf{X}$ into two nonnegative matrices, while giving different weights to the reconstruction errors of different entries. We denote the two nonnegative matrices factors as $\mathbf{U} \in \mathbb{R}_+^{N\times K}$ and $\mathbf{V} \in \mathbb{R}_+^{M\times K}$. Here $K$ is the desired reduced dimension. To facilitate discussions, we call $\mathbf{U}$ the *latent feature matrix* and $\mathbf{V}$ the *basis matrix*. The objective function for general weighted nonnegative matrix factorization can be formulated as below:

$$\min_{\mathbf{U},\mathbf{V}} \|\mathbf{W} * (\mathbf{X} - \mathbf{U}\mathbf{V}^T)\|_F^2, \ \ s.t.\ \mathbf{U} \geq 0, \mathbf{V} \geq 0, \tag{1}$$

where $\|.\|_F$ is the Frobenius norm, $\mathbf{W} \in \mathbb{R}^{N\times M}$ is the weight matrix, $*$ is element-wise production and $\mathbf{U} \geq 0, \mathbf{V} \geq 0$ represent the constraints that all the matrix elements are nonnegative.

## 3 Multi-Incomplete-View Clustering

In this section, we present the Multi-Incomplete-view Clustering (MIC) framework. We model the multi-incomplete-view clustering as a joint weighted nonnegative matrix factorization problem with $L_{2,1}$ regularization. The proposed MIC learns the latent feature matrices for each view and pushes them towards a consensus matrix. Thus, the consensus matrix can be viewed as the shared latent feature matrix across all the views. In the following, we will first describe the construction of the objective function for the proposed method and derive the solution to the optimization problem. Then the whole MIC framework is presented.

### 3.1 Objective Function of MIC

Given $n_v$ views $\{\mathbf{X}^{(i)} \in \mathbb{R}^{N\times d_i}, i = 1, 2, ..., n_v\}$, where each of the views suffers from incompleteness, *i.e.*, $\sum_{j=1}^{N} \mathbf{M}_{i,j} < N$. With more than two incomplete views, we cannot directly apply the existing methods to the incomplete data. One simple solution is to fill the missing instances with average features first, and then apply the existing multi-view clustering methods. However, this approach depends on the quality of the filled instances. For small missing percentages, the quality of the information contained in the filled average features may be good. However, when the number of missing instance increase, the quality of information contained in the filling average features may be bad or even misleading. Thus, simply filling the missing instance will not solve this problem.

Borrowing the similar idea from weighted NMF, we introduce a diagonal weight matrix $\mathbf{W}^{(i)} \in \mathbb{R}^{N\times N}$ for each incomplete views $i$ by

$$\mathbf{W}_{j,j}^{(i)} = \begin{cases} 1 & \text{if } i\text{-th view contains } j\text{-th instance, } i.e., \mathbf{M}_{j,i} = 1. \\ w_i & \text{otherwise.} \end{cases}$$

Note that $\mathbf{W}_{j,j}^{(i)}$ indicates the weight of the $j$-th instance from view $i$, and $w_i$ is the weight of the filled average feature instances for view $i$. In our experiment, $w_i$ is defined as the percentage of the available instances for view $i$:

$$w_i = \frac{\sum_{j=1}^{N} \mathbf{M}_{j,i}}{N}.$$

It can be seen, $\mathbf{W}^{(i)}$ gives lower weights to the missing instances than the presented instances in the $i$-th view. For different views with different incomplete rates, the weights for missing instances are also different. The diagonal weight matrices give higher weights to the missing instances from views with lower incomplete rate.

A simple objective function to combine multiple incomplete views can be:

$$\min_{\{\mathbf{U}^{(i)}\},\{\mathbf{V}^{(i)}\}} \mathcal{O} = \sum_{i=1}^{n_v} (\|\mathbf{W}^{(i)}(\mathbf{X}^{(i)} - \mathbf{U}^{(i)}{\mathbf{V}^{(i)}}^T)\|_F^2 \ s.t.\ \mathbf{U}^{(i)} \geq 0,\ \mathbf{V}^{(i)} \geq 0,\ i = 1, 2, ..., n_v, \tag{2}$$

where $\mathbf{U}^{(i)}$ and $\mathbf{V}^{(i)}$ are the latent feature matrix and basis matrix for the $i$-th view.

However, Eq. (2) only decomposes the different views independently without taking advantages of the relationship between the views. In order to make use of the relation between different views, we push the latent feature matrices for different views towards a common consensus by adding additional term $R$ to Eq. (2) to minimize the disagreement between different views and the common consensus.

$$\begin{aligned} \min_{\{\mathbf{U}^{(i)}\},\{\mathbf{V}^{(i)}\},\mathbf{U}^*} \sum_{i=1}^{n_v} \Big( \|\mathbf{W}^{(i)}(\mathbf{X}^{(i)} - \mathbf{U}^{(i)}{\mathbf{V}^{(i)}}^T)\|_F^2 + \alpha_i R(\mathbf{U}^{(i)}, \mathbf{U}^*) \Big) \\ s.t.\ \mathbf{U}^* \geq 0, \mathbf{U}^{(i)} \geq 0, \mathbf{V}^{(i)} \geq 0, \quad i = 1, 2, ..., n_v, \end{aligned} \tag{3}$$

where $\mathbf{U}^* \in \mathbb{R}^{N*K}$ is the consensus latent feature matrix across all the views, and $\alpha_i$ is the trade-off parameter between reconstruction error and disagreement between view $i$ and the consensus. In this paper we define $R$ as the square of Frobenius norm of the weighted difference between the latent feature matrices:

$$R(\mathbf{U}^{(i)}, \mathbf{U}^*) = \|\mathbf{W}^{(i)}(\mathbf{U}^{(i)} - \mathbf{U}^*)\|_F^2.$$

Additionally, considering the nature of incomplete views, we added $L_{2,1}$ regularization into Eq. 3, which is robust to noises and outliers and widely used in many applications [10,17,37].

Formally, the objective function of MIC is as follows:

$$\begin{aligned} \min_{\{\mathbf{U}^{(i)}\},\{\mathbf{V}^{(i)}\},\mathbf{U}^*} \mathcal{O} = \sum_{i=1}^{n_v} (\|\mathbf{W}^{(i)}(\mathbf{X}^{(i)} - \mathbf{U}^{(i)}{\mathbf{V}^{(i)}}^T)\|_F^2 + \alpha_i \|\mathbf{W}^{(i)}(\mathbf{U}^{(i)} - \mathbf{U}^*)\|_F^2 + \beta_i \|\mathbf{U}^{(i)}\|_{2,1}) \\ s.t.\ \mathbf{U}^{(i)} \geq 0,\ \mathbf{V}^{(i)} \geq 0,\ \mathbf{U}^* \geq 0, \quad i = 1, 2, ..., n_v. \end{aligned} \tag{4}$$

where $\beta_i$ is the trade-off between robustness and accuracy of reconstruction for the $i$-th view, $\|\cdot\|_{2,1}$ is the $L_{2,1}$ norm and defined as:

$$\|\mathbf{U}\|_{2,1} = \sum_{i=1}^{N} \left( \sum_{k=1}^{K} |\mathbf{U}_{i,k}|^2 \right)^{1/2}$$

### 3.2 Optimization

In the following, we give the solution to Eq. 4. For the sake of convenience, we will see both $\alpha_i$ and $\beta_i$ as positive in the derivation, and denote $\tilde{\mathbf{W}}^{(i)} = {\mathbf{W}^{(i)}}^T \mathbf{W}^{(i)}$. As we see, minimizing Eq. 4 is with respect to $\{\mathbf{U}^{(i)}\}$, $\{\mathbf{V}^{(i)}\}$ and $\mathbf{U}^*$, and we cannot give a closed-form solution. We propose an alternating scheme to optimize the objective function. Specifically, the following two steps are repeated until convergence: (1) fixing $\{\mathbf{U}^{(i)}\}$ and $\{\mathbf{V}^{(i)}\}$, minimize $\mathcal{O}$ over $\mathbf{U}^*$, (2) fixing $\mathbf{U}^*$, minimize $\mathcal{O}$ over $\{\mathbf{U}^{(i)}\}$ and $\{\mathbf{V}^{(i)}\}$.

**Fixing $\{\mathbf{U}^{(i)}\}$ and $\{\mathbf{V}^{(i)}\}$, minimize $\mathcal{O}$ over $\mathbf{U}^*$.** With $\{\mathbf{U}^{(i)}\}$ and $\{\mathbf{V}^{(i)}\}$ fixed, we need to minimize the following objective function:

$$\mathcal{J}(\mathbf{U}^*) = \sum_{i=1}^{n_v} \alpha_i \|\mathbf{W}^{(i)}(\mathbf{U}^{(i)} - \mathbf{U}^*)\|_F^2 \quad s.t.\ \mathbf{U}^* \geq 0 \tag{5}$$

We take the derivative of the objective function $\mathcal{J}$ in Eq. 5 over $\mathbf{U}^*$ and set it to 0:

$$\frac{\partial \mathcal{J}}{\partial \mathbf{U}^*} = \sum_{i=1}^{n_v} 2\alpha_i \tilde{\mathbf{W}}^{(i)} \mathbf{U}^* - 2\alpha_i \tilde{\mathbf{W}}^{(i)} \mathbf{U}^{(i)} = 0 \tag{6}$$

Since $\tilde{\mathbf{W}}^{(i)}$ is a positive diagonal matrix and $\alpha_i$ is a positive constant, $\sum_{i=1}^{n_v} \alpha_i \tilde{\mathbf{W}}^{(i)}$ is invertible. Solving Eq. 6, we have an exact solution for $\mathbf{U}^*$:

$$\mathbf{U}^* = \left( \sum_{i=1}^{n_v} \alpha_i \tilde{\mathbf{W}}^{(i)} \right)^{-1} \left( \sum_{i=1}^{n_v} \alpha_i \tilde{\mathbf{W}}^{(i)} \mathbf{U}^{(i)} \right) \geq 0 \tag{7}$$

**Fixing $\mathbf{U}^*$, minimize $\mathcal{O}$ over $\{\mathbf{U}^{(i)}\}$ and $\{\mathbf{V}^{(i)}\}$.** With $\mathbf{U}^*$ fixed, the computation of $\mathbf{U}^{(i)}$ and $\mathbf{V}^{(i)}$ does not depend on $\mathbf{U}^{(i')}$ or $\mathbf{V}^{(i')}, i' \neq i$. Thus for each view $i$, we need to minimize the following objective function:

$$\min_{\mathbf{U}^{(i)}, \mathbf{V}^{(i)}} \|\mathbf{W}^{(i)}(\mathbf{X}^{(i)} - \mathbf{U}^{(i)} {\mathbf{V}^{(i)}}^T)\|_F^2 + \alpha_i \|\mathbf{W}^{(i)}(\mathbf{U}^{(i)} - \mathbf{U}^*)\|_F^2 + \beta_i \|\mathbf{U}^{(i)}\|_{2,1} \quad s.t.\ \mathbf{U}^{(i)} \geq 0, \mathbf{V}^{(i)} \geq 0 \tag{8}$$

We will iteratively update $\mathbf{U}^{(i)}$ and $\mathbf{V}^{(i)}$ using the following multiplicative updating rules. We repeat the two steps iteratively until the objective function in Eq. 8 converges.

**(1) Fixing $\mathbf{U}^*$ and $\mathbf{V}^{(i)}$, minimize $\mathcal{O}$ over $\mathbf{U}^{(i)}$.** For each $\mathbf{U}^{(i)}$, we need to minimize the following objective function:

$$\mathcal{J}(\mathbf{U}^{(i)}) = \|\mathbf{W}^{(i)}(\mathbf{X}^{(i)} - \mathbf{U}^{(i)}{\mathbf{V}^{(i)}}^T)\|_F^2 + \alpha_i\|\mathbf{W}^{(i)}(\mathbf{U}^{(i)} - \mathbf{U}^*)\|_F^2 + \beta_i\|\mathbf{U}^{(i)}\|_{2,1} \quad s.t.\ \ \mathbf{U}^{(i)} \geq 0 \tag{9}$$

The derivative of $\mathcal{J}(\mathbf{U}^{(i)})$ with respect to $\mathbf{U}^{(i)}$ is

$$\frac{\partial \mathcal{J}}{\partial \mathbf{U}^{(i)}} = -2\tilde{\mathbf{W}}^{(i)}\mathbf{X}^{(i)}\mathbf{V}^{(i)} + 2\tilde{\mathbf{W}}^{(i)}\mathbf{U}^{(i)}{\mathbf{V}^{(i)}}^T\mathbf{V}^{(i)} + 2\alpha_i\tilde{\mathbf{W}}^{(i)}\mathbf{U}^{(i)} - 2\alpha_i\tilde{\mathbf{W}}^{(i)}\mathbf{U}^* + \beta_i\mathbf{D}^{(i)}\mathbf{U}^{(i)} \tag{10}$$

Here $\mathbf{D}^{(i)}$ is a diagonal matrix with the $j$-th diagonal element given by

$$\mathbf{D}^{(i)}_{j,j} = \frac{1}{\|\mathbf{U}^{(i)}_{j,:}\|_2}, \tag{11}$$

where $\mathbf{U}^{(i)}_{j,:}$ is the $j$-th row of matrix $\mathbf{U}^{(i)}$, and $\|\cdot\|_2$ is the $L_2$ norm.

Using the Karush-Kuhn-Tucker (KKT) complementary condition [3] for the nonnegativity of $\mathbf{U}^{(i)}$, we get

$$(-2\tilde{\mathbf{W}}^{(i)}\mathbf{X}^{(i)}\mathbf{V}^{(i)} + 2\tilde{\mathbf{W}}^{(i)}\mathbf{U}^{(i)}{\mathbf{V}^{(i)}}^T\mathbf{V}^{(i)} + 2\alpha_i\tilde{\mathbf{W}}^{(i)}\mathbf{U}^{(i)} - 2\alpha_i\tilde{\mathbf{W}}^{(i)}\mathbf{U}^* + \beta_i\mathbf{D}^{(i)}\mathbf{U}^{(i)})_{j,k}\mathbf{U}^{(i)}_{j,k} = 0 \tag{12}$$

Based on this equation, we can derive the updating rule for $\mathbf{U}^{(i)}$:

$$\mathbf{U}^{(i)}_{j,k} \leftarrow \mathbf{U}^{(i)}_{j,k}\sqrt{\frac{\left(\tilde{\mathbf{W}}^{(i)}\mathbf{X}^{(i)}\mathbf{V}^{(i)} + \alpha_i\tilde{\mathbf{W}}^{(i)}\mathbf{U}^*\right)_{j,k}}{\left(\mathbf{U}^{(i)}{\mathbf{V}^{(i)}}^T\mathbf{V}^{(i)} + \alpha_i\tilde{\mathbf{W}}^{(i)}\mathbf{U}^{(i)} + 0.5\beta_i\mathbf{D}^{(i)}\mathbf{U}^{(i)}\right)_{j,k}}} \tag{13}$$

**(2) Fixing $\mathbf{U}^{(i)}$ and $\mathbf{U}^*$, minimize $\mathcal{O}$ over $\mathbf{V}^{(i)}$.** For each $\mathbf{V}^{(i)}$,we need to minimize the following objective function:

$$\mathcal{J}(\mathbf{V}^{(i)}) = \|\mathbf{W}^{(i)}(\mathbf{X}^{(i)} - \mathbf{U}^{(i)}{\mathbf{V}^{(i)}}^T)\|_F^2 \quad s.t.\ \ \mathbf{V}^{(i)} \geq 0 \tag{14}$$

The derivative of $\mathcal{J}(\mathbf{V}^{(i)})$ with respect to $\mathbf{V}^{(i)}$ is

$$\frac{\partial \mathcal{L}}{\partial \mathbf{V}^{(i)}} = 2\mathbf{V}^{(i)}{\mathbf{U}^{(i)}}^T\tilde{\mathbf{W}}^{(i)}\mathbf{U}^{(i)} - 2{\mathbf{X}^{(i)}}^T\tilde{\mathbf{W}}^{(i)}\mathbf{U}^{(i)} \tag{15}$$

Using the KKT complementary condition for the nonnegativity of $\mathbf{V}^{(i)}$, we get

$$(\mathbf{V}^{(i)}{\mathbf{U}^{(i)}}^T\tilde{\mathbf{W}}^{(i)}\mathbf{U}^{(i)} - {\mathbf{X}^{(i)}}^T\tilde{\mathbf{W}}^{(i)}\mathbf{U}^{(i)})_{j,k}\mathbf{V}^{(i)}_{j,k} = 0 \tag{16}$$

Based on this equation, we can derive the updating rule for $\mathbf{V}^{(i)}$:

$$\mathbf{V}^{(i)}_{j,k} \leftarrow \mathbf{V}^{(i)}_{j,k}\sqrt{\frac{({\mathbf{X}^{(i)}}^T\tilde{\mathbf{W}}^{(i)}\mathbf{U}^{(i)})_{j,k}}{(\mathbf{V}^{(i)}{\mathbf{U}^{(i)}}^T\tilde{\mathbf{W}}^{(i)}\mathbf{U}^{(i)})_{j,k}}} \tag{17}$$

**Algorithm 1.** Multi-Incomplete-view Clustering (MIC)

**Input:** Nonnegative data matrices for incomplete views $\{\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, ..., \mathbf{X}^{(n_v)}\}$, indicator matrix $\mathbf{M}$, parameters $\{\alpha_1, \alpha_2, ..., \alpha_{n_v}, \beta_1, \beta_2, ..., \beta_{n_v}\}$, number of clusters K.
**Output:** Basis matrices $\{\mathbf{V}^{(1)}, \mathbf{V}^{(2)}, ..., \mathbf{V}^{(n_v)}\}$, latent feature matrices $\{\mathbf{U}^{(1)}, \mathbf{U}^{(2)}, ..., \mathbf{U}^{(n_v)}\}$, consensus matrix $\mathbf{U}^*$ and clustering results.

Fill the missing instances in each incomplete view with average feature values.
Normalize each view $\mathbf{X}^{(i)}$ such that $\|\mathbf{X}^{(i)}\|_1 = 1$.
Initialize $\mathbf{U}^{(i)}$ and $\mathbf{V}^{(i)}$ for $1 \leq i \leq n_v$.
**repeat**
  Fixing $\mathbf{U}^{(i)}s$ and $\mathbf{V}^{(i)}$s, update $\mathbf{U}^*$ by Eq. 7.
  **for** $i = 1$ **to** $n_v$ **do**
    **repeat**
      Fixing $\mathbf{U}^*$ and $\mathbf{V}^{(i)}$, update $\mathbf{U}^{(i)}$ by Eq. 13.
      Fixing $\mathbf{U}^{(i)}$ and $\mathbf{U}^*$, update $\mathbf{V}^{(i)}$ by Eq. 17.
      Normalize $\mathbf{V}^{(i)}$ and $\mathbf{U}^{(i)}$ by Eq. 18.
    **until** Eq. 8 converges.
  **end for**
**until** Eq. 4 converges.
Apply k-means on $\mathbf{U}^*$ to get the clustering result.

It is worth noting that to prevent $\mathbf{V}^{(i)}$ from having arbitrarily large values (which may lead to arbitrarily small values of $\mathbf{U}^{(i)}$), it is common to put a constraint on each basis matrix $\mathbf{V}^{(i)}$ [14], s.t. $\|\mathbf{V}^{(i)}_{:,k}\|_1 = 1, \ \forall \ 1 \leq k \leq K$. However, the updated $\mathbf{V}^{(i)}$ may not satisfy the constraint. We need to normalize $\mathbf{V}^{(i)}$ and change $\mathbf{U}^{(i)}$ to make the constraint satisfied and keep the accuracy of the approximation $\mathbf{X}^{(i)} \approx \mathbf{U}^{(i)}{\mathbf{V}^{(i)}}^T$:

$$\mathbf{V}^{(i)} \leftarrow \mathbf{V}^{(i)}\mathbf{Q}^{(i)-1}, \mathbf{U}^{(i)} \leftarrow \mathbf{U}^{(i)}\mathbf{Q}^{(i)} \tag{18}$$

Here, $\mathbf{Q}^{(i)}$ is a diagonal matrix with the $k$-th diagonal element given by $\mathbf{Q}^{(i)}_{k,k} = \sum_j^{d_i} \mathbf{V}^{(i)}_{j,k}$.

The whole procedure is summarized in Algorithm 1. We will first fill the missing instances with average feature values in each incomplete view. Then we normalize the data and initialize the latent feature matrices and basis matrices. We apply the iterative alternating optimization procedure until the objective function converges. $k$-means is then applied to the learned consensus latent feature matrix to get the clustering solution.

## 4 Experiments and Results

### 4.1 Comparison Methods

We compare the proposed MIC method with several state-of-art methods. The details of comparison methods are as follows:

- **MIC:** MIC is the clustering framework proposed in this paper, which applies weighted joint nonnegative matrix with $L_{2,1}$ regularization. If not stated, the co-regularization parameter set $\{\alpha_i\}$ and the robust parameter set $\{\beta_i\}$ are all set to 0.01 for all the views throughout the experiment.
- **Concat:** Feature concatenation is one straightforward way to integrate all the views. We first fill the missing instances with the average features for each view. Then we concatenate the features of all the views, and run $k$-means directly on this concatenated view representation.
- **MultiNMF:** MultiNMF [27] is one of the most recent multi-view clustering methods based on joint nonnegative matrix factorization. MultiNMF added constraints to original nonnegative matrix factorization that pushes clustering solution of each view towards a common consensus.
- **ConvexSub:** The subspace-based multi-view clustering method developed by [17]. In the experiments, we set $\beta = 1$ for all the views. We run the ConvexSub method using a range of $\gamma$ values as in the original paper, and present the best results obtained.
- **PVC:** Partial multi-view clustering [26] is one of the state-of-art multi-view clustering methods, which deals with incomplete views. PVC works by establishing a latent subspace where the instances corresponding to the same example in different views are close to each other. In our experiment, we set the parameter $\lambda$ to 0.01 as in the original paper.
- **CGC:** CGC [6] is the most recent work that deals with many-to-many instance relationship, which can be used in the situation of incomplete views. In order to run the CGC algorithm, for every pair of incomplete views, we generate the mapping between the instances that appears in both views. In the experiment, the parameter $\lambda$ is set to 1 as in the original paper.

It is worth to note that MultiNMF and ConvexSub are two recent methods for multi-view clustering. Both of them assumes the completeness of all the available views. PVC is among the first works that does not assume the completeness of any view. However, PVC can only works with two incomplete views. For the sake of comparison, all the views are considered with equivalent importance in the evaluation of all the multi-view algorithms. The results evaluated by two metrics, the normalized mutual information (NMI) and the accuracy (AC). Since we use $k$-means to get the clustering solution at the end of the algorithm, we run $k$-means 20 times and report the average performance.

### 4.2 Dataset

In this paper, three different real-world datasets are used to evaluate the proposed method MIC. Among the three datasets, the first one is handwritten digit data, the second one is text data, while the last one is flower image data. The important statistics of them are summarized in Table 2.

- **Handwritten Dutch Digit Recognition (Digit)**: This data contains 2000 handwritten numerals ("0"-"9") extracted from a collection of Dutch

**Table 2.** Statics of the data

| Data | size | # views | # clusters |
|---|---|---|---|
| Digit | 2000 | 5 | 10 |
| 3Sources | 416 | 3 | 6 |
| Flowers | 1360 | 3 | 17 |

**Table 3.** Incomplete rates for 3Sources

| Data | V1 | V2 | V3 | size |
|---|---|---|---|---|
| BBC-Reuters | 13.51% | 27.76% | - | 407 |
| BBC-Guardian | 12.87% | 25.25% | - | 404 |
| Reuters-Guardian | 23.44% | 21.35% | - | 384 |
| 3Sources | 15.38% | 29.33% | 27.40% | 416 |

utility maps [12]. The following feature spaces (views) with different vector-based features are available for the numbers: (1) 76 Fourier coefficients of the character shapes, (2) 216 profile correlations, (3) 64 Karhunen-Love coefficients, (4) 240 pixel averages in $2 \times 3$ windows, (5) 47 Zernike moments. All these features are conventional vector-based features but in different feature spaces.

- **3-Source Text data (3Sources)**[1] It is collected from three online news sources: BBC, Reuters, and The Guardian, where each news source can be seen as one view for the news stories. In total there are 948 news articles covering 416 distinct news stories from the period February to April 2009. Of these distinct stories, 169 were reported in all three sources, 194 in two sources, and 53 appeared in a single news source. Each story was manually annotated with one of the six topical labels: business, entertainment, health, politics, sport, technology.
- **Oxford Flowers Data (Flowers)**: The Oxford Flower Dataset is composed of 17 flower categories, with 80 images for each category [30]. Each image is described by different visual features using color, shape, and texture. In this paper, we use the $\chi^2$ distance matrices for different flower features (color, shape, texture) as three different views.

Both Digit and Flowers data are complete. We randomly delete instances from each view to make the views incomplete. To simplify the situation, we delete the same number of instances for all the views, and run the experiment under different incomplete percentages from 0% (all the views are complete) to 50% (all the views have 50% instances missing). It is also worth to note that 3Sources is naturally incomplete. Also since PVC can only with with two incomplete views, in order to compare PVC with other methods, we take any two of the three incomplete views and run experiments on them. We also report the results on all the three incomplete views. The statistics of 3Sources data are summarized in Table 3.

### 4.3 Results

The results for Digit data and Flower data are shown in Figs. 1-4. We report the results for various incomplete rates (from 0% to 50% with 10% as interval). Table 4 contains the results for 3Sources data.

[1] http://mlg.ucd.ie/datasets/3sources.html

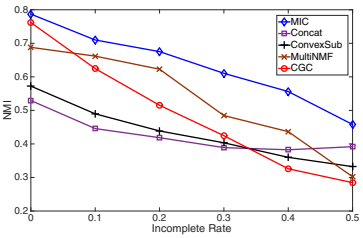



**Fig. 1.** NMIs for Digit.

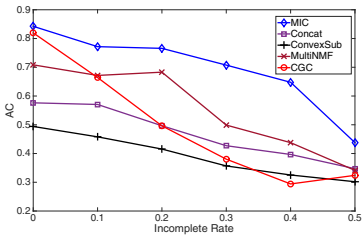



**Fig. 2.** ACs for Digit.

From Figs. 1 and 2 for Digit data, we can see that the proposed MIC method outperforms all the other methods in all the scenarios, especially for relatively large incomplete rates (about 12% higher than other methods in NMI and about 20% higher in AC for incomplete rates 30% and 40% ). It is worth to note that when the incomplete rate is 0, CGC is the second best method in both NMI and AC, which is very close to MIC. However, as the incomplete rate increases, the performance of CGC drops quickly. One of the possible reasons is that CGC works on the similarity matrices/kernels, as the incomplete rate increases, estimated similarity/kernel matrices are not accurate. Also, as the incomplete rate increases, fewer instance mappings between views are available. Combining these two factors, the performance of CGC drops for incomplete views. We can also observe that for incomplete views (incomplete rate $> 0$), multiNMF gives the second best performance (still at lease 5% lower in NMI and at lease 8% lower in AC).

In Table 4, we can also observe that the proposed method outperforms all the other methods in both NMI and AC. MultiNMF and ConvexSub perform the best among the compared techniques.

From Figs. 3 and 4 for Flowers data, we can observe that in most of the cases, MIC outperforms all the other methods. It is worth to note that when all the views are complete, the performances of ConvexSub and MultiNMF are almost the same as MIC. As the incomplete rate increases, MIC starts to show the advantages over other methods. However, when the incomplete rate is too large (*e.g.*, 50%), the performance of MIC is almost the same as ConvexSub and MultiNMF.

### 4.4 Parameter Study

There are two sets of parameters in the proposed methods: $\{\alpha_i\}$, trade-off parameter between reconstruction error and view disagreement and $\{\beta_i\}$, trade-off parameter between the reconstruction error and robustness. Here we explore the effects of the view disagreement trade-off parameter and the robust trade-off parameter to clustering performance. We first fix $\{\beta_i\}$ to 0.01, run MIC with various $\{\alpha_i\}$ values (from $10^{-7}$ to 100). Then fix $\{\alpha_i\}$ to 0.01, run MIC with various $\{\beta_i\}$ values (from $10^{-7}$ to 100). Due to the limit of space, we only report

**Table 4.** Results on 3Sources Text Data

| Methods | BBC-Reuters | | BBC-Guardian | | Reuters-Guardian | | Three-Source | |
|---|---|---|---|---|---|---|---|---|
| | NMI | AC | NMI | AC | NMI | AC | NMI | AC |
| Concat | 0.2591 | 0.3465 | 0.2526 | 0.3599 | 0.2474 | 0.3633 | 0.2757 | 0.3429 |
| ConvexSub | 0.3309 | 0.3913 | 0.3576 | 0.4584 | 0.3450 | 0.4370 | 0.3653 | 0.4504 |
| PVC | 0.2931 | 0.4252 | 0.2412 | 0.4334 | 0.2488 | 0.4145 | – | – |
| CGC | 0.2336 | 0.4167 | 0.2470 | 0.3857 | 0.2682 | 0.4530 | 0.2875 | 0.4279 |
| MultiNMF | 0.3687 | 0.4517 | 0.3647 | 0.4693 | 0.3487 | 0.4281 | 0.4131 | 0.4756 |
| MIC | **0.3814** | **0.4912** | **0.3813** | **0.4988** | **0.3800** | **0.4612** | **0.4512** | **0.5631** |

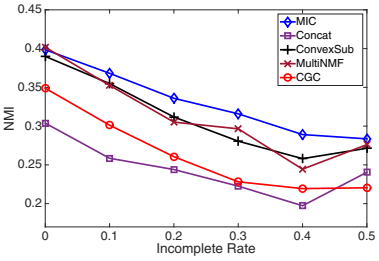



**Fig. 3.** NMIs for Flowers.

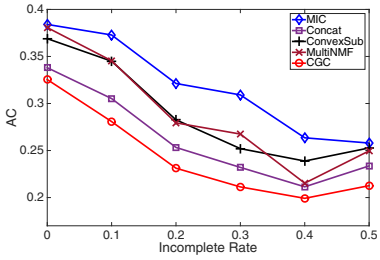
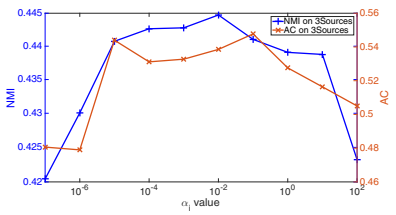



**Fig. 4.** ACs for Flowers.

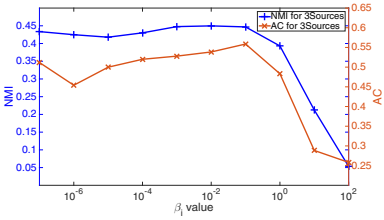
the results on 3Souces data with all the three views in Fig. 5. From Fig. 5, we can find that MIC achieves stably good performance when $\alpha_i$ is around $10^{-2}$ and $\beta_i$ is from $10^{-5}$ to $10^{-1}$.

### 4.5 Convergence Study

The three updates rules for $\mathbf{U}^*$, $\{\mathbf{V}^{(i)}\}$ and $\{\mathbf{U}^{(i)}\}$ are iterative. In the supplemental material, we prove that each update will decrease the value objective function and the whole process will converge to a local minima solution. Fig. 6 shows the convergence curve together with its performance for Digit data with

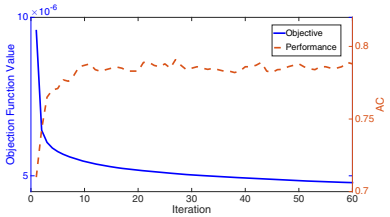



Performance of MIC v.s. $\alpha_i$

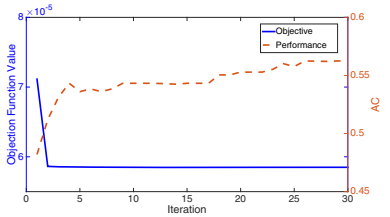



Performance of MIC v.s. $\beta_i$

**Fig. 5.** Parameter study on 3Sources.




Digit data with 10% incomplete rate.




3Sources data with all three views.

**Fig. 6.** Convergence and corresponding performance curve.

10% incomplete rate and 3Sources data using all the three views. The blue solid line shows the value of the objective function and the red dashed line indicates the accuracy of the method. As can be seen, for Digit data, the algorithm will converge after 30 iterations. For 3Sources data, after less than 10 iterations, the algorithm will converge.

## 5 Related Work

There are two areas of related works upon which the proposed model is built. Multi-view learning [2,22,29], is proposed to learn from instances which have multiple representations in different feature spaces. Specifically, Multi-view clustering [1,28] is most related to our work. For example, [1] developed and studied partitioning and agglomerative, hierarchical multi-view clustering algorithms for text data. [23,24] are among the first works proposed to solve the multi-view clustering problem via spectral projection. Linked Matrix Factorization [33] is proposed to explore clustering of a set of entities given multiple graphs. Recently, [34] proposed a kernel based approach which allows clustering algorithms to be applicable when there exists at least one complete view with no missing data. As far as we know, [26,32] are the only two works that do not require the completeness of any view. However, both of the methods can only work with two incomplete views.

Nonnegative matrix factorization [25] is the second area that is related to our work. NMF has been successfully used in unsupervised learning [31,36]. Different variations were proposed in the last decade. For example, [8] posed a three factor NMF and added orthogonal constrains for rigorous clustering interpretation. [19] introduced sparsity constraints on the latent feature matrix, which will give more sparse latent representations. [20] proposed a weighted version of NMF, which gives different weights to different entries in the data. Recently, [6,27] propsed to use NMF to clustering data from multiple views/sources. However, they cannot deal with multiple incomplete views. The proposed MIC, which uses weighted joint NMF to handle the incompleteness of the views and maintain the robustness by introducing the $L_{2,1}$ regularization.

## 6 Conclusion

In this paper, we study the problem of clustering on data with multiple incomplete views, where each view suffers from incompleteness of instances. Based on weighted NMF, the proposed MIC method learns the latent feature matrices for all the incomplete views and pushes them towards a common consensus. To achieve the goal, we use a joint weighted NMF algorithm to learn not only the latent feature matrix for each view but also minimize the disagreement between the latent feature matrices and the consensus matrix. By giving missing instances from each view lower weights, MIC minimizes the negative influences from the missing instances. It also maintains the robustness to noises and outliers by introducing the $L_{2,1}$ regularization. Extensive experiments conducted on three datasets demonstrate the effectiveness of the proposed MIC method on data with multiple incomplete views comparing with other state-of-art methods.

**Acknowledgments.** This work is supported in part by NSF (CNS-1115234), NSFC (61272050, 61273295, 61472089), Google Research Award, the Pinnacle Lab at Singapore Management University, Huawei grants, and the Science Foundation of Guangdong Province (2014A030313556, 2014A030308008).

## References

1. Bickel, S., Scheffer, T.: Multi-view clustering. In: ICDM, pp. 19–26 (2004)
2. Blum, A., Mitchell, T.: Combining labeled and unlabeled data with co-training. In: COLT, New York, NY, USA, pp. 92–100 (1998)
3. Boyd, S., Vandenberghe, L.: Convex Optimization. Cambridge University Press, New York (2004)
4. Bruno, E., Marchand-Maillet, S.: Multiview clustering: a late fusion approach using latent models. In: SIGIR. ACM, New York (2009)
5. Chaudhuri, K., Kakade, S.M., Livescu, K., Sridharan, K.: Multi-view clustering via canonical correlation analysis. In: ICML, New York, NY, USA (2009)
6. Cheng, W., Zhang, X., Guo, Z., Wu, Y., Sullivan, P.F., Wang, W.: Flexible and robust co-regularized multi-domain graph clustering. In: SIGKDD, pp. 320–328. ACM (2013)
7. de Sa, V.R.: Spectral clustering with two views. In: ICML Workshop on Learning with Multiple Views (2005)
8. Ding, C., Li, T., Peng, W., Park, H.: Orthogonal nonnegative matrix T-factorizations for clustering. In: SIGKDD, pp. 126–135. ACM (2006)
9. Ding, C., Zhou, D., He, X., Zha, H.: R1-PCA: rotational invariant L1-norm principal component analysis for robust subspace factorization. In: ICML, pp. 281–288. ACM (2006)
10. Ding, W., Wu, X., Zhang, S., Zhu, X.: Feature selection by joint graph sparse coding. In: SDM, Austin, Texas, pp. 803–811, May 2013
11. L. Du, X. Li, and Y. Shen. Robust nonnegative matrix factorization via half-quadratic minimization. In: ICDM, pp. 201–210 (2012)
12. Duin, R.P.: Handwritten-Numerals-Dataset

13. Evgeniou, A., Pontil, M.: Multi-task Feature Learning. Advances in Neural Information Processing Systems **19**, 41 (2007)
14. Févotte, C.: Majorization-minimization algorithm for smooth itakura-saito nonnegative matrix factorization. In: ICASSP, pp. 1980–1983. IEEE (2011)
15. Greene, D., Cunningham, P.: A matrix factorization approach for integrating multiple data views. In: Buntine, W., Grobelnik, M., Mladenić, D., Shawe-Taylor, J. (eds.) ECML PKDD 2009, Part I. LNCS, vol. 5781, pp. 423–438. Springer, Heidelberg (2009)
16. Gu, Q., Zhou, J., Ding, C.: Collaborative filtering: weighted nonnegative matrix factorization incorporating user and item graphs. In: SDM. SIAM (2010)
17. Guo, Y.: Convex subspace representation learning from multi-view data. In: AAAI, Bellevue, Washington, USA (2013)
18. Huang, H., Ding, C.: Robust tensor factorization using R1 norm. In: CVPR, pp. 1–8. IEEE (2008)
19. Kim, J., Park, H.: Sparse Nonnegative Matrix Factorization for Clustering (2008)
20. Kim, Y., Choi, S.: Weighted nonnegative matrix factorization. In: International Conference on Acoustics, Speech and Signal Processing, pp. 1541–1544 (2009)
21. Kong, D., Ding, C., Huang, H.: Robust nonnegative matrix factorization using L21-norm. In: CIKM, New York, NY, USA, pp. 673–682 (2011)
22. Kriegel, H.P., Kunath, P., Pryakhin, A., Schubert, M.: MUSE: multi-represented similarity estimation. In: ICDE, pp. 1340–1342 (2008)
23. Kumar, A., Daume III, H.: A co-training approach for multi-view spectral clustering. In: ICML, New York, NY, USA, pp. 393–400, June 2011
24. Kumar, A., Rai, P., Daumé III, H.: Co-regularized multi-view spectral clustering. In: NIPS, pp. 1413–1421 (2011)
25. Lee, D., Seung, S.: Learning the Parts of Objects by Nonnegative Matrix Factorization. Nature **401**, 788–791 (1999)
26. Li, S., Jiang, Y., Zhou, Z.: Partial multi-view clustering. In: AAAI, pp. 1968–1974 (2014)
27. Liu, J., Wang, C., Gao, J., Han, J.: Multi-view clustering via joint nonnegative matrix factorization. In: SDM (2013)
28. Long, B., Philip, S.Y., (Mark) Zhang, Z.: A general model for multiple view unsupervised learning. In: SDM, pp. 822–833. SIAM (2008)
29. Nigam, K., Ghani, R.: Analyzing the effectiveness and applicability of co-training. In CIKM, pp. 86–93. ACM, New York (2000)
30. Nilsback, M.-E., Zisserman, A.: A visual vocabulary for flower classification. In: CVPR, vol. 2, pp. 1447–1454 (2006)
31. Shahnaz, F., Berry, M., Pauca, V.P., Plemmons, R.: Document Clustering Using Nonnegative Matrix Factorization. Information Processing & Management **42**(2), 373–386 (2006)
32. Shao, W., Shi, X., Yu, P.: Clustering on multiple incomplete datasets via collective kernel learning. In: ICDM (2013)
33. Tang, W., Lu, Z., Dhillon, I.S.: Clustering with multiple graphs. In: ICDM, Miami, Florida, USA, pp. 1016–1021, December 2009
34. Trivedi, A., Rai, P., Daumé III, H., DuVall, S.L.: Multiview clustering with incomplete views. In: NIPS 2010: Workshop on Machine Learning for Social Computing, Whistler, Canada (2010)
35. Wang, D., Li, T., Ding, C.: Weighted feature subset non-negative matrix factorization and its applications to document understanding. In: ICDM (2010)

36. Xu, W., Liu, X., Gong, Y.: Document clustering based on non-negative matrix factorization. In: SIGIR, pp. 267–273 (2003)
37. Zhang, X., Yu, Y., White, M., Huang, R., Schuurmans, D.: Convex sparse coding, subspace learning, and semi-supervised extensions. In: AAAI (2011)
38. Zhou, D., Burges, C.: Spectral clustering and transductive learning with multiple views. In: ICML, pp. 1159–1166. ACM, New York (2007)