Expert Systems With Applications 174 (2021) 114783

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

A network-based sparse and multi-manifold regularized multiple non-negative matrix factorization for multi-view clustering

Lihua Zhou [a,*], Guowang Du [a], Kevin Lü [b], Lizhen Wang [a]

[a] School of Information Science & Engineering, Yunnan University, Kunming 650091, Yunnan, PR China
[b] Brunel University, Uxbridge UB8 3PH, UK

ARTICLE INFO

Keywords:
Multi-view clustering
Networks
Non-negative matrix factorization
Multi-manifold regularization

ABSTRACT

Multi-view clustering has attracted increasing attention in recent years since many real data sets are usually gathered from different sources or described by different feature types. Amongst various existing multi-view clustering algorithms, those that are based on non-negative matrix factorization (NMF) have exhibited superior performance. However, NMF decomposing original data directly fails to exploit global relationships between data samples and cannot be applied to datasets that are not strictly non-negative. In this paper, a network-based sparse and multi-manifold regularized multiple NMF (NSM_MNMF) for multi-view clustering is proposed, where multi-view data is transformed into multiple networks, and NMF is used to jointly factorize transformed multiple networks for capturing the shared cluster structure embedded in different views. Furthermore, sparse and multi-manifold regularization are incorporated into NMF to keep the intrinsic geometrical information of the multi-view network manifold space. Networks characterize intra-view similarity, and joint factorization reveals inter-view similarity across distinct views, while using NMF to decompose the networks instead of the original data means NSM_MNMF can be applied to datasets that are not strictly non-negative and the clustering results are interpretable. Extensive experiments are conducted on nine real data sets to assess the method proposed, and the results illustrate that NSM_MNMF outperforms other baseline approaches.

1. Introduction

Multi-view clustering (Bickel & Scheffer, 2004) refers to dividing a set of multi-view data gathered from different sources or described by different feature types into clusters such that data samples within the same cluster are more similar than those in different clusters. Many real-world datasets, such as images and multi-variable time series, are composed of multiple views which often contain compatible and supplementary information to each other, jointly uncovering a complete picture of the phenomenon or event of interest from different perspectives (Kumar, Rai, & Daumé, 2011). Thus, clustering performance based on information extracted from multiple views outperforms that from a single view (Zong, Zhang, Zhao, Yu, & Zhao, 2017). In the last few years, multi-view clustering has attracted increasing attention from researchers and has become an important research area in data mining (Zhang, Nie, Li, & Wei, 2019; Zhang, Liu, Shen, Shen, & Shao, 2019; Saini, Bansal, Saha, & Bhattacharyya, 2020).

However, multi-view clustering faces more challenges than classical single-view clustering in which data samples are collected from a single source or represented by one kind of feature type because, as multi-view data may be heterogeneous, different features describe different information from different perspectives (for example, a story can be written in different languages and reported by different news sources), and the similarity between data samples encapsulates complex relationships such as similarity across different views (inter-view similarity) and similarity just within a single view (intra-view similarity). How to integrate automatically different information extracted from different views and disclose the shared structure embedded in different views is the key to clustering multi-view data, which significantly affects clustering effectiveness and efficiency.

To integrate information extracted from different views for clustering multi-view data, a number of algorithms based on NMF have been proposed (Singh & Gordon, 2008; Wendel, Sternig, & Godec, 2011; Liu, Wang, Gao, & Han, 2013; Zhang, Zong, Liu, & Yu, 2015) and have achieved promising performance. These algorithms efficiently discovered the latent clustering structure embedded in multiple views through factorizing matrices formed by multi-view data, and they kept factorized matrices meaningful. Meanwhile, these algorithms imposed various

* Corresponding author.
E-mail address: lhzhou@ynu.edu.cn (L. Zhou).

https://doi.org/10.1016/j.eswa.2021.114783
Received 14 April 2019; Received in revised form 25 January 2021; Accepted 22 February 2021
Available online 27 February 2021

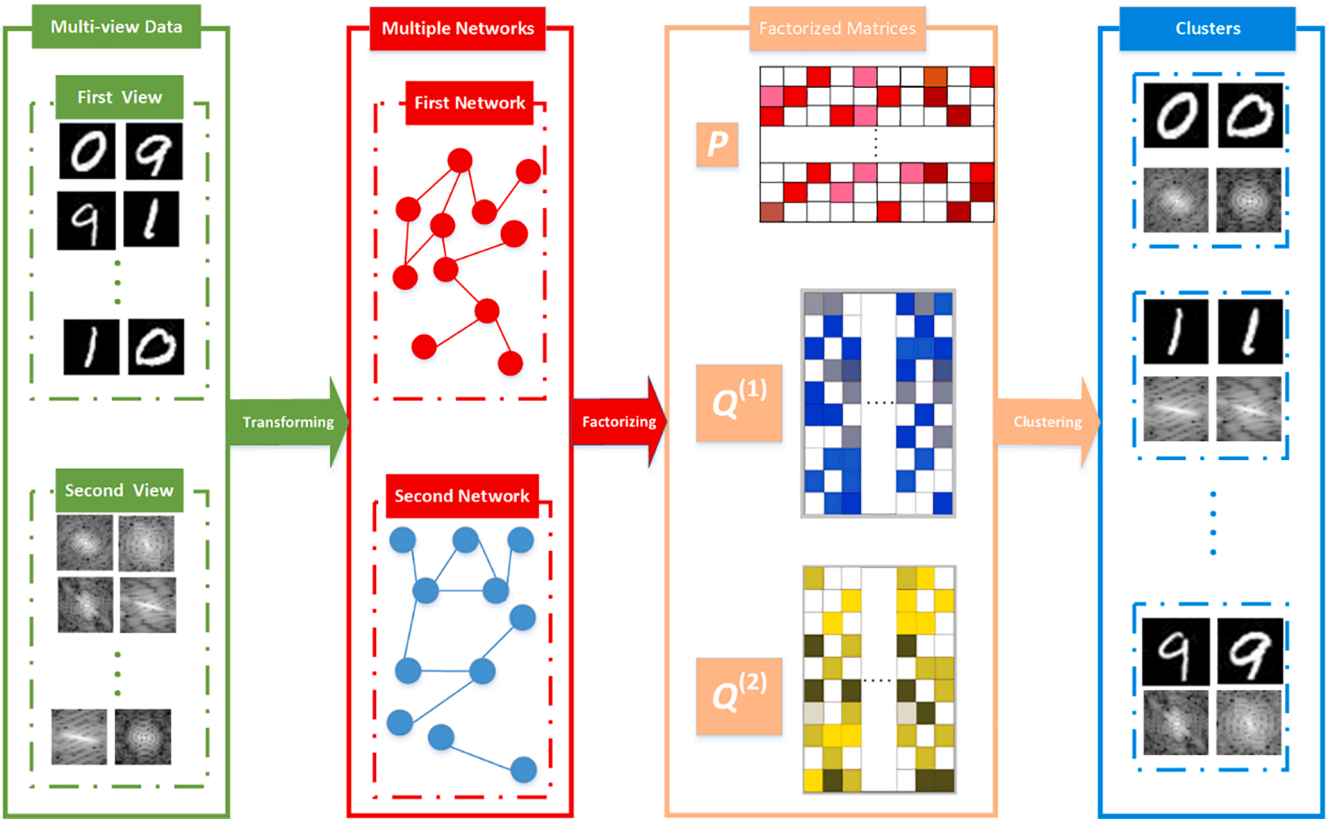

Fig. 1. The framework of NSM_MNMF.

regularizations, such as multi-manifold regularization (Bickel & Scheffer, 2004) and graph dual regularization (Shang, Jiao, & Wang, 2012), on the standard NMF proposed by Lee and Seung (1999) so as to encourage the sparsity of factorization and preserve the inherent geometrical structure embedded in data space. It has been shown that although these additional regularizations are simple, cluster structures discovered by NMF with regularizations have higher qualities than those discovered by standard NMF alone, which indicates that regularizations have a positive effect on clustering.

However, existing NMF-based approaches cannot be applied to datasets that contain negative feature values, such as the Robot Execution Failures dataset (Camarinha-Matos et al., 1996), because NMF is strictly restricted to matrices that are non-negative. Moreover, in the original form of intrinsic features (i.e., data matrix), only the local relationship amongst neighboring data samples can be identified, whilst in general, the global relationship remains unknown (Ferreira & Zhao, 2016), so decomposing the original data directly fails to recognize global relationships amongst data samples. On the other hand, the network-based approaches for clustering are able to characterize the relationship between any pairs or any groups of data samples and can capture arbitrary cluster shapes (Ferreira & Zhao, 2016); thus, network-based approaches have been reported to achieve a promising performance (Zhan et al., 2019), but the clusters obtained lack clear interpretability and pose a difficulty to understand how the samples in each cluster are distributed in different views.

The aim of this paper is to detect the shared cluster structure embedded in multi-view data. To this end, we propose a network-based sparse and multi-manifold regularized multiple NMF (NSM_MNMF), where multi-view data is transformed into multiple networks based on the similarities amongst data samples, and NMF is used to jointly factorize transformed multiple networks for capturing the shared cluster structure embedded in different views. Fig. 1 shows the framework of NSM_MNMF, where a network corresponds to a single view, and in a network, each node denotes a data sample, and each edge represents the similarity between two data samples in the view. The transformation from data domain to topological domain means NSM_MNMF not only gives networks the ability to characterize both local and global relationships amongst data samples but also means it can be applied to datasets that are not strictly non-negative. NMF joint factorization of multiple networks detects the underlying part-based patterns in each view and reveals the potential connections between patterns of different views; thus, the clustering results are interpretable. In addition, to exploit the intrinsic geometric information embedded in the multi-view manifold space and to avoid overfitting on the latent cluster structure as well as controlling the sparseness, we impose multiple regularizations on NMF, including multi-manifold, smooth, and sparse regularization. These constraints make the factorizations drive similar samples of each view towards a common consensus and identify clusters across different views.

The main contributions of this paper are summarized as follows:

(1) An approach based on the NMF of multiple networks for clustering multi-view data, which is referred to as NSM_MNMF, is proposed. NSM_MNMF first transforms multi-view data into multiple networks and then uses NMF to jointly factorize transformed multiple networks for capturing the shared cluster structure embedded in different views. The transformation from data space into network space uncovers the intra-view similarities within a view, while the joint factorization reveals inter-view similarities across different views. In this way,

NSM_MNMF is applicable to datasets that are not strictly non-negative, and the clustering results are interpretable.

(2) A novel method for factorizing multiple networks is proposed, where multiple regularizations are imposed on NSM_MNMF to preserve the intrinsic geometric information embedded in the networks, and efficient updating rules are derived for computing optimal factorized matrices that can minimize decomposition loss. We also prove theoretically the correctness and convergence of the updating rules, and we design and implement an efficient iterative updating algorithm to complete factorizations.

(3) We conducted extensive experiments on nine real multi-view datasets and compared the results of our NSM_MNMF with those of sixteen baseline approaches, including those that decompose the original data directly and those that decompose networks but with different regularizations, to evaluate the performance of the proposed approach. The experimental results demonstrate that NSM_MNMF is much more accurate than the baseline approaches and is robust.

The rest of the paper is arranged as follows. Section 2 offers a brief overview of related work. The details of NSM_MNMF and the convergence proof of the optimization scheme are presented in Section 3. Section 4 provides extensive experiments and results, and in Section 5, conclusions are given.

2. Related work

2.1. Non-negative matrix factorization

NMF (Lee & Seung, 1999) is a matrix factorization algorithm that decomposes a data matrix with non-negative elements into two low-rank matrices such that the product of the two low-rank matrices can approximate the original data matrix.

Let $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n]^{\mathrm{T}} \in \Re^{n\times m}$ be a data matrix all of whose elements are non-negative and of which each row, $\mathbf{x}_i, i = 1, \cdots, n$, denotes the feature vector of a sample. NMF focuses on finding two low rank non-negative matrices, $\mathbf{P} = [p_{ik}] \in \Re^{n\times d}$ and $\mathbf{Q} = \left[q_{jk}\right] \in \Re^{m\times d}, d \ll min\{m, n\}$ such that $\mathbf{X} \approx \mathbf{PQ}^{\mathrm{T}}, \mathrm{s.t.} \mathbf{P} \geqslant 0, \mathbf{Q} \geqslant 0$. This means that the product of $\mathbf{P}$ and $\mathbf{Q}$ can approximate the data matrix $\mathbf{X}$. In general, $\mathbf{P}$ is understood as a basis matrix, and $\mathbf{Q}$ is regarded as a coefficient matrix, and coefficient vectors of $\mathbf{Q}$ are interpreted as the low-rank representation of samples with respect to the new basis $\mathbf{P}$. To quantify the quality of the approximation between $\mathbf{X}$ and $\boldsymbol{PQ}^{\mathrm{T}}$, the square of the Frobenius norm $||\mathbf{X} - \mathbf{PQ}^{\mathrm{T}}||_{\mathrm{F}}^2 = \sum_{i,j}\left(x_{ij} - \sum_{k=1}^{d} p_{ik}q_{jk}\right)^2$ (Paatero & Tapper, 1994), or the "divergence" $D\left(\mathbf{X}||\mathbf{PQ}^{\mathrm{T}}\right) = \sum_{i,j}\left(x_{ij}log\frac{x_{ij}}{\left(\mathbf{PQ}^{\mathrm{T}}\right)_{ij}} - x_{ij} + \left(\mathbf{PQ}^{\mathrm{T}}\right)_{ij}\right)$ (Lee & Seung, 2001) are commonly used as cost functions, where $||\mathbf{X} - \mathbf{PQ}^{\mathrm{T}}||_{\mathrm{F}}^2$ is symmetric but $D\left(\mathbf{X}||\mathbf{PQ}^{\mathrm{T}}\right)$ is not.

NMF is able to learn parts-based representation, which encodes much of the data and makes them easy to interpret, because it allows only additive rather than subtractive combinations in the process of matrix factorization. Through decomposition, samples are mapped into a new space where samples with similar features are close to each other. Therefore, NMF has been widely applied in clustering.

However, Feng, Xiao, Zhou, and Zhuang (2015) summarized that the standard NMF proposed by Lee and Seung (1999) suffers from some inherent shortcomings: (1) it is difficult to control the sparsity of $\boldsymbol{P}$ and $\boldsymbol{Q}$; (2) it is unable to explore the information of the geometric structure and the class label of data; and (3) it is merely optimal under the condition of Gaussian or Poisson noise, and thus it is unfit to handle other noise types.

To control the sparsity of $\boldsymbol{P}$ and/or $\boldsymbol{Q}$, Liu, Lu, and Gu (2005) adopted L_1/L_2 regularized NMF (GSNMF) to yield group sparsity such that each of the obtained linear manifolds belongs to a particular class. Sun, Shen, Gao, Ouyang, and Cheng (2017) integrated the decoder with the encoder to form a unified loss function such that the orthogonality constraint naturally imposed by the symmetry between the decoder and the encoder, together with the non-negative constraint, makes sure that the learned representation is sparse.

To discover the latent geometric structure embedded in the data space, Cai, He, Han, and Huang (2011) constructed an affinity graph to encode the latent geometric information on the data distribution and incorporated the geometry as an additional regularization term into NMF (GNMF). The basis of GNMF is the manifold assumption that means if data point $\boldsymbol{x}_i$ is close to $\boldsymbol{x}_j$ in the inherent geometry of the data distribution, then the representations of $\boldsymbol{x}_i$ are also close to those of $\boldsymbol{x}_j$ in the new basis obtained by NMF. According to the hypothesis that if a data point is given birth by several neighboring points on a specific manifold in the original space, it can be reconstructed in a similar way in the low dimensional subspace, Shen and Si (2010) proposed NMF of multiple manifolds (MM_NMF), which used a L_1-graph to preserve the geometric information of multiple manifolds. Shang, Jiao, and Wang (2012) presented a graph dual regularized NMF (DNMF), which constructs nearest neighbor graphs for the data manifold and feature manifold respectively, such that the geometric information embedded in the data and feature spaces can be encoded simultaneously.

To deal with other noise except for the Gaussian or Poisson noise, Kong, Ding, and Huang (2011) extended the standard NMF with $L_{2,1}$-norm ($L_{2,1}$ NMF) for Laplacian noise. Feng, Xiao, Zhou, and Zhuang (2015) extended the standard NMF as a locally weighted sparse graph of regularized NMF (LWSG_NMF), which exploits the local structure information of data with sparse block noise. Peng, Kang, Hu, Cheng, and Cheng (2017) integrated NMF with the principal component model and designed a robust graph of regularized NMF to incorporate fundamental nonlinear structures and capture noise and outliers in the process of factorization.

In addition, Ding et al. (2010) developed the convex and semi-non-negative matrix factorizations (Semi-NMF) by relaxing the non-negativity limitation on the data matrix and the basis matrix. Trigeorgis, Bousmalis, Zafeiriou, and Schuller (2017) proposed Deep Semi-NMF, a deep architecture for Semi-NMF, to automatically learn lower-dimensional hierarchal features for clustering data that is not strictly non-negative.

2.2. Multiple non-negative matrix factorization

The standard NMF and its variants can be applied to only one matrix whose data come from just one view. However, many real-world datasets, such as images of the human face and hand-written digital notes, are represented by different features or views, which often supply information complementary to each other for describing the same set of samples. In order to integrate various heterogeneous features for uncovering the common latent structure shared by multiple views, various joint non-negative matrix factorization algorithms have been developed. For example, jNMF and SNMNMF, proposed by Zhang et al. (2012), decomposed original data matrices sharing the same row

dimension into a common basis matrix and multiple coefficient matrices; MultiNMF (Liu, Wang, Gao, & Han, 2013) and MMNMF (Zong, Zhang, Zhao, Yu, & Zhao, 2017) decomposed original data matrices into multiple basis matrices and multiple coefficient matrices. Yu, Wang, Wang, and Zeng (2020) took account of the uncertain relationship between a sample and a cluster, then investigated an active three-way clustering method via low-rank matrices to improve the clustering accuracy. Huang, Kang, and Xu (2020) proposed an auto-weighted approach that utilized a deep matrix decomposition framework to capture the hierarchical semantics of the input data in a layer-wise way. Zhang, Zhou, Wang, Huang, and He (2021) derived the view-specific representation from data samples by exploiting the local structure within each view and generated the low-rank tensor representation from the view-specific representation to capture the high-order correlation across multiple views, such that the intra-view pairwise information and the inter-view complementary information can be explored. Meanwhile, various constraints have been used to improve clustering performance. SNMNMF (Zhang et al., 2012) enforces the must-link constraints on the adjacency matrices representing the interaction of samples. MultiNMF (Liu, Wang, Gao, & Han, 2013) uses the constraint of consensus coefficient matrix. Wang, Jiang, Wu, and Zhou (2011) introduced a local graph regularization to deal with the inner-view relatedness. Zong et al. (2017) adopted multi-manifold regularization to retain locally the geometrical structure embedded in the multi-view data space. Furthermore, Zhao, Ding, and Fu (2017) presented a deep framework to carry out Semi-NMF, where graph regularizers are introduced to respect the intrinsic geometric structure embedded in each view. SNMNMF encourages smoothness and sparsity, but jNMF, MultiNMF, MMNMF, and Deep Semi-NMF do not.

2.3. The network-based approaches

Network-based approaches have been studied for several clustering tasks, owing to their ability to uncover the hidden structures of data. Ferreira and Zhao (2016) proposed a network-based method to cluster univariate time series data by transforming a time series from a time–space domain to a topological domain and applying community detection algorithms for networks to identify groups of densely connected nodes. Zhan et al. (2019) integrated the graph structures of different views into a global one, such that the correlation of graph structure amongst multiple views can be taken into account. Dai et al. (2019) proposed a reverse nearest neighbor structure-based algorithm (RNN-NSDC) to clustering data that contains clusters of outliers and arbitrary shapes. Kumar et al. (2011) proposed two multi-view spectral clustering algorithms (Co-Reg (Pairwise) and Co-Reg (Centroid)) with co-regularization schemes. The first co-regularization scheme required that the eigenvectors of a view pair should have high pairwise similarity (pairwise co-regularization), while the second one forced the view-specific eigenvectors to look similar by regularizing them towards a common consensus (centroid based co-regularization). Brbić and Kopriva (2018) proposed multi-view low-rank sparse subspace clustering (MLRSSC) algorithms that learned a joint subspace representation by constructing affinity matrix shared among all views. MLRSSC enforced agreement between the affinity matrices of the pairs of views and between the affinity matrices towards a common centroid. Lu, Yan, and Lin (2016) proposed the convex pairwise sparse spectral clustering (PSSC) model to improve sparse spectral clustering (SSC) by exploiting multi-view information of data.

3. A network-based sparse and multi-manifold regularized MNMF (NSM_MNMF)

In this section, we present the detailed introduction of the proposed NSM_MNMF approach, including the method for transforming multi-view data into multiple networks, the design and optimization of the objective function of NSM_MNMF, and the pseudo-code of an algorithm to complete factorization.

3.1. Transforming multi-view data into multiple networks

Let $\mathbf{X}^{(v)} \in \Re^{n\times m^{(v)}}, v = 1, \cdots, n_v$ be the data of the v - th view, where n is the number of samples, and $m^{(v)}$ is the feature dimension in the v - th view. Denote the i - th row of $\mathbf{X}^{(v)}$ as $\mathbf{X}_{i:}^{(v)}$, representing the i - th sample.

To transforming multi-view data into multiple networks represented by a set of affinity matrices $\mathscr{A} = \{\mathbf{A}^{(v)}\}_{v=1}^{n_v}$, a distance function (such as Euclidean distance or Pearson correlation) is firstly used to measure the similarities between samples. Then, each sample is represented as a node, which is connected to its k most similar nodes (k-NN). The process is repeated n_v times, each dealing with a view. $\mathbf{A}^{(v)} = \{\mathbf{A}_{ij}^{(v)}\}_{i,j=1}^{n}$ is the affinity matrix of the v - th view, which can be computed based on the similarities amongst data samples in the v - th view (which is referred to as intra-view similarity with respect to the v - th view). The Algorithm Mv2Mn describes the procedure for computing $\mathscr{A} = \{\mathbf{A}^{(v)}\}_{v=1}^{n_v}$ from $\mathscr{X} = \{\mathbf{X}^{(v)}\}_{v=1}^{n_v}$, where $N_k(\boldsymbol{X}_{i:}^{(v)})$ is the set of k nearest neighbours of $\boldsymbol{X}_{i:}^{(v)}$.

Algorithm Mv2Mn

Input: multi-view data $\mathscr{X} = \{\mathbf{X}^{(v)}\}_{v=1}^{n_v}$

Output: affinity matrices $\mathscr{A} = \{\mathbf{A}^{(v)}\}_{v=1}^{n_v}$

1. For each view $v \in \{1, \cdots, n_v\}$

2.Compute the distance $d\left(\mathbf{X}_{i:}^{(v)}, \mathbf{X}_{j:}^{(v)}\right)$ of each pair of samples $\left(\mathbf{X}_{i:}^{(v)}, \mathbf{X}_{j:}^{(v)}\right), i,j = 1, \cdots, n$

3.Find k-NN of each sample $\mathbf{X}_{i:}^{(v)}$: $//k = \lfloor log_2 n \rfloor + 1$

$N_k\left(\mathbf{X}_{i:}^{(v)}\right) = \left\{\mathbf{X}_{j1:}^{(v)}, \mathbf{X}_{j2:}^{(v)}, \cdots, \mathbf{X}_{jk:}^{(v)} \middle| d\left(\mathbf{X}_{i:}^{(v)}, \mathbf{X}_{j1:}^{(v)}\right) < d\left(\mathbf{X}_{i:}^{(v)}, \mathbf{X}_{j2:}^{(v)}\right) < \cdots < d\left(\mathbf{X}_{i:}^{(v)}, \mathbf{X}_{jk:}^{(v)}\right)\right.$

4.if $\mathbf{X}_{j:}^{(v)} \in N_k\left(\mathbf{X}_{i:}^{(v)}\right)$ or $\mathbf{X}_{i:}^{(v)} \in N_k\left(\mathbf{X}_{j:}^{(v)}\right)$, then $\mathbf{A}_{ij}^{(v)} = exp\left(-\frac{\left\|\mathbf{X}_{i:}^{(v)} - \mathbf{X}_{j:}^{(v)}\right\|_2^2}{d\left(\mathbf{X}_{i:}^{(v)}, \mathbf{X}_{jk:}^{(v)}\right) \times d\left(\mathbf{X}_{j:}^{(v)}, \mathbf{X}_{jk:}^{(v)}\right)}\right)$

5.else $\mathbf{A}_{ij}^{(v)} = 0$

6. End for

3.2. Objective function of NSM_MNMF

Given the set of affinity matrices $\mathscr{A} = \{\mathbf{A}^{(v)}\}_{v=1}^{n_v}$ of n_v views, the aim of NSM_MNMF is to search basis matrix $\mathbf{P} \in \Re_+^{n\times c}$ and coefficient matrices $\{\mathbf{Q}^{(v)}\}_{v=1,\ldots,n_v}$, $\mathbf{Q}^{(v)} \in \Re_+^{n\times c}$, where c is the number of clusters. $\mathbf{P}$ reflects the latent cluster structure shared across multiple views, and the entry $\mathbf{P}_{ij}$ indicates the possibility that i - th sample belongs to the j - th cluster; thus, $\mathbf{P}$ is also referred to as a membership matrix. $\mathbf{Q}^{(v)}$ represents the aggregation strength of samples within each cluster in the v - th view, and the entry $\mathbf{Q}_{ij}^{(v)}$ indicates the coherency of the i - th sample with respect to the j - th cluster in the v - th view.

Let $\mathbf{D}^{(v)}$ be a diagonal matrix, and $\mathbf{D}_{ii}^{(v)} = \sum_l \mathbf{A}_{il}^{(v)}$, $\mathbf{L}^{(v)} = \mathbf{D}^{(v)} - \mathbf{A}^{(v)}$ be the Laplacian matrix of the v - th view. Manifold assumption indicates that contiguous data points in each view locate on or close to a local area of a manifold that is discretely approximated by $\mathbf{L}^{(v)}$. By their nature, multi-view datasets locate on or close to a local area of multiple manifolds (Zong, Zhang, Zhao, Yu, & Zhao, 2017). In this study, we use a consensus manifold (graph Laplacian) $\boldsymbol{L}^*$ and a consensus coefficient matrix $\mathbf{Q}^*$ to represent the latent geometrical structure and aggregation strength shared by multiple views. Then, NSM_MNMF aims to minimize Eq. (1),

$$J\left(\mathbf{P},\mathbf{Q}^{(1)},\cdots,\mathbf{Q}^{(n_v)}\right)=\frac{1}{2}\sum_{v=1}^{n_v}\left\|\mathbf{A}^{(v)}-\mathbf{P}(\mathbf{Q}^{(v)})^{\mathrm{T}}\right\|_{\mathrm{F}}^{2}+\frac{1}{2}\sum_{v=1}^{n_v}\|\mathbf{Q}^{*}-\mathbf{Q}^{(v)}\|_{\mathrm{F}}^{2}+\frac{1}{2}\sum_{v=1}^{n_v}\|\mathbf{L}^{*}-\mathbf{L}^{(v)}\|_{\mathrm{F}}^{2}$$
$$+\frac{\gamma_1}{2}\|\mathbf{P}\|_{\mathrm{F}}^{2}+\frac{\gamma_2}{2}\sum_{v=1}^{n_v}\|\mathbf{Q}^{(v)}\|_{1}+\frac{\lambda_1}{2}\sum_{v=1}^{n_v}\mathrm{Tr}\left((\mathbf{Q}^{(v)})^{\mathrm{T}}\mathbf{L}^{(v)}\mathbf{Q}^{(v)}\right)+\frac{\lambda_2}{2}\mathrm{Tr}\left((\mathbf{Q}^{*})^{\mathrm{T}}\mathbf{L}^{*}\mathbf{Q}^{*}\right) \tag{1}$$
$$s.t.\ \mathbf{P}\geqslant 0\ and\ \mathbf{Q}^{(v)}\geqslant 0,\mathbf{L}^{*}\geqslant 0,\mathbf{Q}^{*}\geqslant 0,\forall \mathrm{v}=1,\cdots,\mathrm{n_v}$$

where $||\cdot||_{\mathrm{F}}^{2}$ represents the Frobenius norm, $\mathrm{Tr}(\cdot)$ denotes the trace of a matrix, and $(\mathbf{Q}^{(v)})^{\mathrm{T}}$ means the transpose of $\mathbf{Q}^{(v)}$. The first item of Eq. (1) is the square fitting error between the affinity matrix and the approximate matrix in each view. The second item of Eq. (1) is designed to minimize the difference between the consensus coefficient matrix and each coefficient matrix. It indicates that coefficient matrices representing aggregation strength in the network space of each view should be regularized towards a shared consensus matrix $\mathbf{Q}^{*}$. The third item of Eq. (1) aims to minimize the difference between the consensus manifold and the manifold of each view. It indicates that Laplacian matrices in multiple views should be regularized towards a shared consensus Laplacian matrix $\mathbf{L}^{*}$. The fourth item of Eq. (1) imposes a smoothness constraint in the membership matrix $\mathbf{P}$, which is used to avoid overfitting on the latent cluster structure. The fifth item of Eq. (1) imposes a sparsity constraint on the coefficient matrix $\mathbf{Q}^{(v)}$ of each view, which is used to control the sparseness of $\mathbf{Q}^{(v)}$. The sixth and last item of Eq. (1) makes sure that the manifold assumptions can be satisfied in each view and the consensus coefficient matrices to preserve the local geometry information of the multi-view network spaces. Thus, the inherent local geometry properties of data, smoothness, and sparseness have been taken into account simultaneously.

The parameter $\gamma_1\geqslant 0$ inhibits the growth of $\mathbf{P}$, the parameter $\gamma_2\geqslant 0$ controls the sparseness of $\mathbf{Q}^{(v)}$, and the parameters $\lambda_1,\lambda_2\geqslant 0$ are used to adjust the weights of the multi-manifold regularizations. γ_1, γ_2, λ_1 and λ_2 are called trade-off parameters.

3.3. Updating rules

It is impracticable to obtain a global optimal solution with an algorithm because the objective function of Eq. (1) is convex for just one variable individually rather than jointly convex for all variables together. Thus, all variables can only be optimized one by one. In this study, an iterative update procedure consisting of the following four steps is repeated until convergence for minimizing $J$of Eq. (1): (1) fixing $\mathbf{L}^{*}$, $\mathbf{Q}^{*}$, and $\mathbf{P}$ to minimize J over $\mathbf{Q}^{(v)}$; (2) fixing $\mathbf{L}^{*}$, $\mathbf{Q}^{*}$, and $\mathbf{Q}^{(v)}$ to minimize J over $\mathbf{P}$; and (3) fixing $\mathbf{P}$, $\mathbf{Q}^{(v)}$, and $\mathbf{Q}^{*}$ to minimize J over $\mathbf{L}^{*}$; and (4) fixing $\mathbf{P}$, $\mathbf{Q}^{(v)}$, and $\mathbf{L}^{*}$ to minimize J over $\mathbf{Q}^{*}$.

3.3.1. Fixing $\boldsymbol{L}^{*}$, $\boldsymbol{Q}^{*}$,and $\boldsymbol{P}$, to minimize J over $\boldsymbol{Q}^{(v)}$

When $\mathbf{L}^{*}$, $\mathbf{Q}^{*}$, and $\mathbf{P}$ are kept unchanged, for each given view v, the computation of the coefficient matrix $\mathbf{Q}^{(v)}$ is independent of $\mathbf{Q}^{(v')},\mathrm{v}'\neq\mathrm{v}$; thus, just $J_{\mathbf{Q}^{(v)}}$ needs to be minimized where $J_{\mathbf{Q}^{(v)}}=\left\|\mathbf{A}^{(v)}-\mathbf{P}(\mathbf{Q}^{(v)})^{\mathrm{T}}\right\|_{\mathrm{F}}^{2}+\|\mathbf{Q}^{*}-\mathbf{Q}^{(v)}\|_{\mathrm{F}}^{2}+\gamma_2\|\mathbf{Q}^{(v)}\|_{1}+\lambda_1\mathrm{Tr}\left((\mathbf{Q}^{(v)})^{\mathrm{T}}\mathbf{L}^{(v)}\mathbf{Q}^{(v)}\right)$ with the constraint $\mathbf{Q}^{(v)}\geqslant 0$. Let $\boldsymbol{\omega}$ be the Lagrange multiplier matrix for the constraint $\mathbf{Q}^{(v)}\geqslant 0$, we can write the Lagrange as:

$$\begin{aligned}H_{\mathbf{Q}^{(v)}}&=J_{\mathbf{Q}^{(v)}}+\mathrm{Tr}(\boldsymbol{\omega}\mathbf{Q}^{(v)})\\&=\mathrm{Tr}\left(\mathbf{A}^{(v)}(\mathbf{A}^{(v)})^{\mathrm{T}}-2\mathbf{A}^{(v)}\mathbf{Q}^{(v)}\mathbf{P}^{\mathrm{T}}+\mathbf{P}(\mathbf{Q}^{(v)})^{\mathrm{T}}\mathbf{Q}^{(v)}\mathbf{P}^{\mathrm{T}}\right)+\mathrm{Tr}\left((\mathbf{Q}^{*})^{\mathrm{T}}\mathbf{Q}^{*}-2(\mathbf{Q}^{*})^{\mathrm{T}}\mathbf{Q}^{(v)}+(\mathbf{Q}^{(v)})^{\mathrm{T}}\mathbf{Q}^{(v)}\right)\\&\quad+\gamma_2\|\mathbf{Q}^{(v)}\|_{1}+\lambda_1\mathrm{Tr}\left((\mathbf{Q}^{(v)})^{\mathrm{T}}\mathbf{L}^{(v)}\mathbf{Q}^{(v)}\right)+\mathrm{Tr}(\boldsymbol{\omega}\mathbf{Q}^{(v)})\end{aligned} \tag{2}$$

Based on the Karush-Kuhn-Tucker (KKT) conditions, the partial derivative of $H_{\mathbf{Q}^{(v)}}$ on $\mathbf{Q}^{(v)}$ can be derived as follows:

$$\begin{aligned}\frac{\partial H_{\mathbf{Q}^{(v)}}}{\partial \boldsymbol{Q}^{(v)}}&=-(\mathbf{A}^{(v)})^{\mathrm{T}}\mathbf{P}+\mathbf{Q}^{(v)}\mathbf{P}^{\mathrm{T}}\mathbf{P}-\mathbf{Q}^{*}+\mathbf{Q}^{(v)}+\gamma_2\mathbf{I}+\lambda_1\mathbf{L}^{(v)}\mathbf{Q}^{(v)}+\boldsymbol{\omega}\\&=-(\mathbf{A}^{(v)})^{\mathrm{T}}\mathbf{P}+\mathbf{Q}^{(v)}\mathbf{P}^{\mathrm{T}}\mathbf{P}-\mathbf{Q}^{*}+\mathbf{Q}^{(v)}+\gamma_2\mathbf{I}+\lambda_1(\mathbf{D}^{(v)}-\mathbf{A}^{(v)})\mathbf{Q}^{(v)}+\boldsymbol{\omega}\\\boldsymbol{\omega}_{l,k}\boldsymbol{Q}_{l,k}^{(v)}&=0,\forall 1\leqslant \mathrm{l}\leqslant \mathrm{n},\forall 1\leqslant \mathrm{k}\leqslant \mathrm{c}\end{aligned} \tag{3}$$

where $\mathbf{I}$ is an identity matrix. Correspondingly, the following update rule for $\mathbf{Q}^{(v)}$ element-wise can be derived and shown in Eq. (4).

$$(\mathbf{Q}^{(v)})_{lk}=(\mathbf{Q}^{(v)})_{lk}\frac{\left((\mathbf{A}^{(v)})^{\mathrm{T}}\mathbf{P}+\mathbf{Q}^{*}+\lambda_1\mathbf{A}^{(v)}\mathbf{Q}^{(v)}\right)_{lk}}{\left(\mathbf{Q}^{(v)}\mathbf{P}^{\mathrm{T}}\mathbf{P}+\mathbf{Q}^{(v)}+\gamma_2\mathbf{I}+\lambda_1\mathbf{D}^{(v)}\mathbf{Q}^{(v)}\right)_{lk}} \tag{4}$$

Theorem 1:. Based on the updating rules of Eq. (4), the objective function J of Eq. (1) is non-increasing.

The proof of Theorem 1 is presented in the Appendix A.

3.3.2. Fixing $\mathbf{L}^{*}$, $\mathbf{Q}^{*}$, and $\mathbf{Q}^{(v)}$, to minimize J over $\mathbf{P}$

When $\mathbf{P}$ is updated, the items that are irrelevant to $\mathbf{P}$ may be neglected; thus, just $J_{\mathbf{P}}=\frac{1}{2}\left(\sum_{v=1}^{n_v}\left\|\mathbf{A}^{(v)}-\mathbf{P}\left(\mathbf{Q}^{(v)}\right)^{\mathrm{T}}\right\|_{\mathrm{F}}^{2}+\gamma_1\|\mathbf{P}\|_{\mathrm{F}}^{2}\right)$ with the constraint $\mathbf{P}\geqslant 0$ needs to be minimized. Let ψ be the Lagrange multiplier matrix for the constraint $\mathbf{P}\geqslant 0$, then the Lagrange can be written as:

Table 1
Statistics of the nine datasets (where "–" means no information in this view).

	Robot Execution Failures (Time series)					3-Source (Text)	WebKB (Text)	BBCSport (Text)	Digit (Image)
	LP1	LP2	LP3	LP4	LP5				
Number of instances	88	47	47	117	164	169	203	282	2000
Number of views	6	6	6	6	6	3	3	3	2
Number of classes	4	5	4	3	5	6	4	5	10
Dimensions of view 1	15	15	15	15	15	3560	1703	2582	76
Dimensions of view 2	15	15	15	15	15	3631	230	2544	240
Dimensions of view 3	15	15	15	15	15	3068	230	2465	–
Dimensions of view 4	15	15	15	15	15	–	–	–	–
Dimensions of view 5	15	15	15	15	15	–	–	–	–
Dimensions of view 6	15	15	15	15	15	–	–	–	–

$$H_{\mathbf{P}} = \frac{1}{2}\left(\sum_{v=1}^{n_v}\left\|\mathbf{A}^{(v)} - \mathbf{P}(\mathbf{Q}^{(v)})^{\mathrm{T}}\right\|_{\mathrm{F}}^2 + \gamma_1\|\mathbf{P}\|_{\mathrm{F}}^2\right) + \mathrm{Tr}(\psi\mathbf{P})$$
$$= \frac{1}{2}\left(\sum_{v=1}^{n_v}\mathrm{Tr}(\mathbf{A}^{(v)}(\mathbf{A}^{(v)})^{\mathrm{T}} - 2\mathbf{A}^{(v)}\mathbf{Q}^{(v)}\mathbf{P}^{\mathrm{T}} + \mathrm{P}(\mathbf{Q}^{(v)})^{\mathrm{T}}\mathbf{Q}^{(v)}\mathbf{P}^{\mathrm{T}}) + \gamma_1\mathrm{Tr}(\mathbf{P}^{\mathrm{T}}\mathbf{P})\right) + \mathrm{Tr}(\psi\mathbf{P}) \tag{5}$$

Based on the Karush-Kuhn-Tucker (KKT) conditions, the partial derivative of $H_{\mathbf{P}}$ on $\mathbf{P}$ can be derived as follows:

$$\frac{\partial H_{\mathbf{P}}}{\partial \mathbf{P}} = \sum_{v=1}^{n_v}\left(-\mathbf{A}^{(v)}\mathbf{Q}^{(v)} + \mathbf{P}(\mathbf{Q}^{(v)})^{\mathrm{T}}\mathbf{Q}^{(v)}\right) + \gamma_1\mathbf{P} + \psi$$
$$\psi_{i,k}\mathbf{P}_{i,k} = 0, \forall 1\leqslant i\leqslant n, \forall 1\leqslant k\leqslant c \tag{6}$$

Correspondingly, the update rule for $\mathbf{P}$ element-wise can be derived and shown in Eq. (7).

$$\mathbf{P}_{ik} = \mathbf{P}_{ik}\frac{\left(\sum_{v=1}^{n_v}\mathbf{A}^{(v)}\mathbf{Q}^{(v)}\right)_{ik}}{\left(\sum_{v=1}^{n_v}\mathbf{P}(\mathbf{Q}^{(v)})^{\mathrm{T}}\mathbf{Q}^{(v)} + \gamma_1\mathbf{P}\right)_{ik}} \tag{7}$$

Theorem 2:. Based on the updating rules of Eq. (7), the objective function J of Eq. (1) is non-increasing.

The proof of Theorem 2 is omitted because it is similar to that of Theorem 1.

3.3.3. Fixing $\mathbf{P}$, $\mathbf{Q}^{(v)}$ and $\mathbf{Q}^*$, to minimize J over $\mathbf{L}^*$

When $\mathbf{L}^*$ is updated, just $J_{\mathbf{L}^*} = \frac{1}{2}\sum_{v=1}^{n_v}\left\|\mathbf{L}^* - \mathbf{L}^{(v)}\right\|_{\mathrm{F}}^2 + \frac{\lambda_2}{2}\mathrm{Tr}\left((\mathbf{Q}^*)^{\mathrm{T}}\mathbf{L}^*\mathbf{Q}^*\right)$ with the constraint $\mathbf{L}^*\geqslant 0$ needs to be minimized. Let $\boldsymbol{\upsilon}$ be the Lagrange multiplier matrix for the constraint $\mathbf{L}^*\geqslant 0$, then the Lagrange can be written as:

$$H_{\mathbf{L}^*} = J_{\mathbf{L}^*} + \mathrm{Tr}(\boldsymbol{\upsilon}\mathbf{L}^*)$$
$$= \frac{1}{2}\sum_{v=1}^{n_v}\mathrm{Tr}((\mathbf{L}^*)^{\mathrm{T}}\mathbf{L}^* - 2(\mathbf{L}^*)^{\mathrm{T}}\mathbf{L}^{(v)} + (\mathbf{L}^{(v)})^{\mathrm{T}}\mathbf{L}^{(v)}) + \frac{\lambda_2}{2}\mathrm{Tr}((\mathbf{Q}^*)^{\mathrm{T}}\mathbf{L}^*\mathbf{Q}^*) + \mathrm{Tr}(\boldsymbol{\upsilon}\mathbf{L}^*) \tag{8}$$

According to the Karush-Kuhn-Tucker (KKT) conditions, the partial derivative of $H_{\mathbf{L}^*}$ on $\mathbf{L}^*$ can be derived as follows:

$$\frac{\partial H_{\mathbf{L}^*}}{\partial \mathbf{L}^*} = n_v\mathbf{L}^* - \sum_{v=1}^{n_v}\mathbf{L}^{(v)} + \lambda_2\mathbf{Q}^*(\mathbf{Q}^*)^{\mathrm{T}} + \boldsymbol{\upsilon} = n_v\mathbf{L}^* - \sum_{v=1}^{n_v}\left(\mathbf{D}^{(v)} - \mathbf{A}^{(v)}\right) + \lambda_2\mathbf{Q}^*(\mathbf{Q}^*)^{\mathrm{T}} + \boldsymbol{\upsilon}$$
$$\boldsymbol{\upsilon}_{i,k}\mathbf{L}^*_{i,k} = 0, \forall 1\leqslant i\leqslant n, \forall 1\leqslant k\leqslant c \tag{9}$$

Thus, the update rule for $\mathbf{L}^*$ element-wise, is:

$$(\mathbf{L}^*)_{ik} = (\mathbf{L}^*)_{ik}\frac{\left(\sum_{v=1}^{n_v}\mathbf{D}^{(v)}\right)_{ik}}{\left(n_v\mathbf{L}^* + \sum_{v=1}^{n_v}\mathbf{A}^{(v)} + \lambda_2\mathbf{Q}^*(\mathbf{Q}^*)^{\mathrm{T}}\right)_{ik}} \tag{10}$$

Theorem 3:. Based on the updating rules of Eq. (10), the objective function J of Eq. (1) is non-increasing.

The proof of Theorem 3 is omitted because it is also similar to that of Theorem 1.

3.3.4. Fixing $\mathbf{P}$, $\mathbf{Q}^{(v)}$ and $\mathbf{L}^*$, to minimize J over $\mathbf{Q}^*$

When $\mathbf{Q}^*$ is computed, just $J_{\mathbf{Q}^*} = \frac{1}{2}\sum_{v=1}^{n_v}\left\|\mathbf{Q}^* - \mathbf{Q}^{(v)}\right\|_{\mathrm{F}}^2 + \frac{\lambda_2}{2}\mathrm{Tr}\left((\mathbf{Q}^*)^{\mathrm{T}}\mathbf{L}^*\mathbf{Q}^*\right)$ with the constraint $\mathbf{Q}^*\geqslant 0$ needs to be minimized. Let $\boldsymbol{\xi}$ be the Lagrange multiplier matrix for the constraint $\mathbf{Q}^*\geqslant 0$, then the Lagrange can be written as:

$$H_{\mathbf{Q}^*} = \frac{1}{2}\sum_{v=1}^{n_v}\|\mathbf{Q}^* - \mathbf{Q}^{(v)}\|_F^2 + \frac{\lambda_2}{2}\mathrm{Tr}((\mathbf{Q}^*)^T\mathbf{L}^*\mathbf{Q}^*) + \mathrm{Tr}(\xi\mathbf{Q}^*)$$
$$= \frac{1}{2}\sum_{v=1}^{n_v}\mathrm{Tr}\left((\mathbf{Q}^*)^T\mathbf{Q}^* - 2(\mathbf{Q}^*)^T\mathbf{Q}^{(v)} + (\mathbf{Q}^{(v)})^T\mathbf{Q}^{(v)}\right) + \frac{\lambda_2}{2}\mathrm{Tr}\left((\mathbf{Q}^*)^T\mathbf{L}^*\mathbf{Q}^*\right) + \mathrm{Tr}(\xi\mathbf{Q}^*) \quad (11)$$

the partial derivative of $H_{\mathbf{Q}^*}$ on $\mathbf{Q}^*$ is:

$$\frac{\partial H_{\mathbf{Q}^*}}{\partial \mathbf{Q}^*} = n_v\mathbf{Q}^* - \sum_{v=1}^{n_v}\mathbf{Q}^{(v)} + \lambda_2\mathbf{L}^*\mathbf{Q}^* + \xi$$
$$\xi_{i,k}\mathbf{Q}^*_{i,k} = 0, \forall 1\leqslant i\leqslant n, \forall 1\leqslant k\leqslant c \quad (12)$$

and the following update rule can be derived as follows:

$$(\mathbf{Q}^*)_{ik} = (\mathbf{Q}^*)_{ik}\frac{\left(\sum_{v=1}^{n_v}\mathbf{Q}^{(v)}\right)_{ik}}{(n_v\mathbf{Q}^* + \lambda_2\mathbf{L}^*\mathbf{Q}^*)_{ik}} \quad (13)$$

Theorem 4:. Based on the updating rules of Eq. (13), the objective function J of Eq. (1) is non-increasing.

The proof of Theorem 4 is omitted because it is also similar to that of Theorem 1.

It is easy to see that $\left(\mathbf{Q}^{(v)}\right)_{lk}$, $\mathbf{P}_{ik}$, $(\mathbf{L}^*)_{ik}$ and $(\mathbf{Q}^*)_{ik}$ remain non-negative after each update because all operations do not involve any negative elements.

3.4. The algorithm NSM_MNMF

The procedures for calculating and updating $\mathbf{P}$, $\mathbf{Q}^{(v)}$, $\mathbf{L}^*$, and $\mathbf{Q}^*$ are performed in Algorithm NSM_MNMF. When the value of $J(\mathbf{P}, \mathbf{Q}^{(1)}, \cdots, \mathbf{Q}^{(n_v)})$ in the Eq. (1) is no longer decreasing (this can be guaranteed by Theorem 1 ~ Theorem 4), the calculated $\mathbf{P}$ reflects the latent cluster structure shared by multiple views, and the entry $\mathbf{P}_{ij}$ indicates the possibility that the i - th sample belongs to the j - th cluster; therefore, the cluster label of sample i can be assigned by computing $argmax_k\mathbf{P}_{ik}$.

The pseudo-code of the Algorithm NSM_MNMF is shown below:

Algorithm NSM_MNMF

Input: multi-view data $\mathscr{X} = \{\mathbf{X}^{(v)}\}_{v=1}^{n_v}$, trade-off parameters $\gamma_1, \gamma_2, \lambda_1, \lambda_2$, the number of clusters c.

Output: Membership matrix $\mathbf{P}$, coefficient matrix $\mathbf{Q}^{(1)}, \mathbf{Q}^{(2)}, \cdots, \mathbf{Q}^{(n_v)}$, consensus Laplacian matrix $\mathbf{L}^*$, and consensus connectivity matrix $\mathbf{Q}^*$.

1. Compute the affinity matrices $\mathscr{A} = \mathrm{Mv2Mn}(\mathscr{X})$ // $\mathscr{A} = \{\mathbf{A}^{(v)}\}_{v=1}^{n_v}$
2. Initialize $\mathbf{P}, \mathbf{Q}^{(1)}, \mathbf{Q}^{(2)}, \ldots, \mathbf{Q}^{(n_v)}$, $\mathbf{L}^*$, $\mathbf{Q}^*$ with random values uniformly selected from the interval [0,1]
3. Repeat
4. For $v = 1$ to n_v
5. Fixed $\mathbf{L}^*$, $\mathbf{Q}^*$, $\mathbf{P}, \mathbf{Q}^{(1)}, \cdots, \mathbf{Q}^{(v-1)}, \mathbf{Q}^{(v+1)}, \cdots, \mathbf{Q}^{(n_v)}$, update $\mathbf{Q}^{(v)}$ with Eq. (4)
6. End for
7. Fixed $\mathbf{L}^*$, $\mathbf{Q}^*$, $\mathbf{Q}^{(1)}, \mathbf{Q}^{(2)}, \cdots, \mathbf{Q}^{(n_v)}$, update $\mathbf{P}$ with Eq. (7)
8. Fixed $\mathbf{P}$, $\mathbf{Q}^{(1)}, \mathbf{Q}^{(2)}, \cdots, \mathbf{Q}^{(n_v)}$, $\mathbf{Q}^*$, update $\mathbf{L}^*$ with Eq. (10)
9. Fixed $\mathbf{P}$, $\mathbf{Q}^{(1)}, \mathbf{Q}^{(2)}, \cdots, \mathbf{Q}^{(n_v)}$, $\mathbf{L}^*$, update $\mathbf{Q}^*$ with Eq. (13)
10. Until Eq. (1) convergence. // the value of J is no longer decreasing or maximal iterations is exhausted

here are n_v views and n samples; each sample is described by $m^{(v)}$ features, so the time complexity of Algorithm Mv2Mn is $O(n_v n^2 m^{(v)})$. Let t_{loop} be the number of iterations for the outer loop (Line 3–10 in Algorithm NSM_MNMF), then the execution time for factorizing jointly n_v affinity matrices is $O(t_{\mathrm{loop}} n_v n^2 c)$. Thus, the total execution time of Algorithm NSM_MNMF is $O(n_v n^2 m^{(v)} + t_{\mathrm{loop}} n_v n^2 c)$. It is linear with the number of views and clusters but is squared with respect to the number of samples.

4. Experiments and results

This section presents the extensive experiments conducted to identify the latent clustering structure shared by multiple views, which were carried out to evaluate:

W1: whether factorizing networks can get more rational clustering results than factorizing original data
W2: whether regularizations of NSM_MNMF can preserve the intrinsic geometrical structure of data
W3: whether joint factorization can capture both intra-view and inter-view similarities amongst nodes
W4: whether NSM_MNMF can achieve fast converge, and how sensitive NSM_MNMF is to regularization parameters.

4.1. Datasets

We used nine datasets in the experiments. The nine datasets consist of five time series datasets on Robot Execution Failures (LP1 ~ LP5[1]), three text datasets (3-Source,[2] WebKb,[3] and BBCSport[4]), and one handwritten digit dataset (Digit[5]). Table 1 sums up the main statistics of the nine datasets.

Robot Execution Failures dataset: The time series datasets, consisting of LP1 ~ LP5, describe five different learning problems respectively. Each dataset contains measurements of three forces (Fx, Fy, Fz) and three torques (Tx, Ty, Tz) within 315 ms after the failure detection of a robot, and the evolution of each force or each torque is characterized by 15 force/torque values. The evolution of each force or torque is treated as one view. Each time series is annotated with one of the class labels, such as normal, collision, or obstruction.

3-Sources: a text dataset on news stories reported in three well-known online news sources, where each news story was marked with one of the six topical labels: business, entertainment, health, politics, sport, and technology; 169 new stories reported in all three sources were selected for experiments, where each source is treated as one view.

WebKB: composed of 203 web-pages collected from 4 universities in the US. Each web-page is characterized by the content of the page, the anchor text of the hyper-link, and the text in its title. Each page is annotated with one of the four topical labels: course, faculty, project, and student.

BBCSport: a text dataset of articles of sports news collected from the BBC Sport website. The dataset consists of 282 documents; each document was divided into three parts and was marked with one of the five

[1] http://archive.ics.uci.edu/ml
[2] http://mlg.ucd.ie/datasets
[3] http://www.cs.cmu.edu/afs/cs.cmu.edu/project/theo-20/www/data/
[4] http://mlg.ucd.ie/datasets/segment.html
[5] http://archive.ics.uci.edu/ml/datasets.html

Table 2
The parameters of NSM_MNMF.

		Robot Execution Failures ($k = 7$)					3-Source ($k = 7$)	WebKB ($k = 7$)	BBCSport ($k = 7$)	Digit ($k = 100$)
		LP1	LP2	LP3	LP4	LP5				
NSM_MNMF	λ_1	1	1e−3	0.01	0.1	1	100	1e+3	10	1
	λ_2	100	1	1	10	0.01	1	100	0.1	1e−3
	γ_1	100	0.1	1e−3	1e+3	1e−3	1e−3	1	100	1e−3
	γ_2	1e−3	100	1e+3	1e−3	0.01	100	100	0.01	1e−3

Table 3
RI of algorithms on nine datasets.

RI	Robot Execution Failures					3-Source	WebKB	BBCSport	Digit
	LP1	LP2	LP3	LP4	LP5				
MultiNMF	–	–	–	–	–	0.75	0.43	0.63	0.95
MultiNMF (Graph)	0.76	0.69	0.63	0.59	0.67	0.80	0.57	0.86	0.96
GMultiNMF	–	–	–	–	–	0.75	0.73	0.60	0.97
GMultiNMF (Graph)	0.76	0.69	0.63	0.62	0.69	0.79	0.68	0.80	**0.98**
SNMNMF	–	–	–	–	–	0.71	0.39	0.64	0.92
SNMNMF (Graph)	0.77	0.72	0.63	0.66	0.70	0.87	0.70	0.89	0.97
MNMF	0.88	0.76	0.69	0.75	0.68	0.80	0.61	0.84	0.97
CMNMF	0.85	0.80	0.77	0.76	0.75	0.80	0.61	0.84	0.97
GMNMF	0.86	0.80	0.81	0.85	**0.80**	0.83	0.61	0.86	0.97
Co-Reg (Pairwise)	0.79	0.83	0.78	0.73	0.77	0.83	0.76	0.78	0.95
Co-Reg (Centroid)	0.75	0.84	0.67	0.71	0.76	0.86	0.67	0.79	0.94
MLRSSC (Pairwise)	0.86	0.81	0.76	0.68	0.70	0.84	0.71	0.91	0.94
MLRSSC (Centroid)	0.87	0.79	0.80	0.67	0.76	0.83	0.72	0.88	0.94
KMLRSSC (Pairwise)	0.80	0.83	0.77	0.49	0.75	0.81	0.69	0.72	0.93
KMLRSSC (Centroid)	0.85	0.79	0.75	0.68	0.73	0.80	0.69	0.71	0.94
CSMSC	0.72	0.83	0.78	0.77	0.78	0.76	0.68	0.87	0.93
NSM_MNMF	**0.88**	**0.86**	**0.83**	**0.88**	0.77	**0.88**	**0.82**	**0.95**	**0.98**

Table 4
AC of algorithms on nine datasets.

AC	Robot Execution Failures					3-Source	WebKB	BBCSport	Digit
	LP1	LP2	LP3	LP4	LP5				
MultiNMF	–	–	–	–	–	0.54	0.54	0.46	0.88
MultiNMF (Graph)	0.63	0.43	0.38	0.57	0.40	0.57	0.61	0.78	0.89
GMultiNMF	–	–	–	–	–	0.49	0.75	0.52	0.93
GMultiNMF (Graph)	0.65	0.43	0.47	0.57	0.45	0.60	0.73	0.73	**0.95**
SNMNMF	–	–	–	–	–	0.65	0.53	0.58	0.72
SNMNMF (Graph)	0.69	0.57	0.62	0.74	0.46	0.78	0.74	0.87	0.92
MNMF	0.88	0.63	0.45	0.75	0.40	0.56	0.47	0.78	0.93
CMNMF	0.84	0.60	0.62	0.77	0.48	0.57	0.46	0.78	0.92
GMNMF	0.85	0.60	0.70	0.89	**0.60**	0.65	0.52	0.80	0.93
Co-Reg (Pairwise)	0.71	0.63	0.63	0.78	0.56	0.66	0.68	0.64	0.86
Co-Reg (Centroid)	0.57	0.70	0.44	0.75	0.55	0.72	0.61	0.59	0.85
MLRSSC (Pairwise)	0.85	0.62	0.61	0.73	0.52	0.65	0.62	0.83	0.79
MLRSSC (Centroid)	0.86	0.58	0.66	0.71	0.53	0.66	0.63	0.79	0.81
KMLRSSC (Pairwise)	0.72	0.60	0.64	0.61	0.50	0.61	0.57	0.53	0.77
KMLRSSC (Centroid)	0.84	0.59	0.60	0.74	0.52	0.60	0.57	0.51	0.78
CSMSC	0.56	0.65	0.63	0.83	0.60	0.72	0.71	0.77	0.74
NSM_MNMF	**0.88**	**0.68**	**0.70**	**0.92**	0.55	**0.78**	**0.79**	**0.94**	0.94

topical labels: athletes, cricket, football, rugby, and tennis.

Digit: an image dataset from the UCI repository. The dataset consists of 2000 images of hand-written digits (0–9). In total, 76 Fourier coefficients and the 240 pixel averages in 2×3 windows were selected as two views for our experiments.

In this study, the similarities between samples are measured by Euclidean distance in the Robot Execution Failures and Digit datasets, while cosine similarity is used for the other three datasets (3-Sources, WebKB, and BBCSport).

4.2. Evaluation metrics

Three metrics, namely, the accuracy (AC), normalized mutual information (NMI), and RandIndex (RI) (Halkidi et al., 2001), were employed to assess and compare the performance of the proposed algorithm in this study. AC was used to compute the clustering accuracy that measured the percentage of the correct clustering results; NMI was used to measure the mutual dependence information between the clustering results and the ground truth. When data samples are partitioned perfectly, the NMI score is 1, and when the clustering labels and the ground truth labels are independent, the NMI score is 0; RI is a measure of the similarity between the clustering results and the ground truth, reflecting the purity within each cluster. The scores of the three metrics can be computed according to the difference between the obtained cluster labels and those offered by the datasets. For these three metrics (AC, NMI, and RI), a value does close to 1 means a good clustering result.

Table 5
NMI of algorithms on nine datasets.

NMI	Robot Execution Failures					3-Source	WebKB	BBCSport	Digit
	LP1	LP2	LP3	LP4	LP5				
MultiNMF	–	–	–	–	–	0.45	0.10	0.17	0.80
MultiNMF (Graph)	0.37	0.31	0.23	0.33	0.31	0.59	0.18	0.70	0.82
GMultiNMF	–	–	–	–	–	0.44	0.35	0.29	0.86
GMultiNMF (Graph)	0.37	0.29	0.29	0.41	0.35	0.50	0.42	0.61	**0.89**
SNMNMF	–	–	–	–	–	0.45	0	0.24	0.66
SNMNMF (Graph)	0.41	0.37	0.32	0.39	0.27	0.62	0.43	0.70	0.84
MNMF	0.69	0.53	0.34	0.60	0.36	0.59	0.19	0.68	0.85
CMNMF	0.70	0.54	0.48	0.67	0.37	0.59	0.18	0.68	0.86
GMNMF	0.71	0.57	0.53	0.67	0.45	0.62	0.18	0.68	0.86
Co-Reg(Pairwise)	0.61	0.51	0.44	0.56	0.37	0.56	0.46	0.45	0.79
Co-Reg (Centroid)	0.45	**0.61**	0.31	0.56	0.41	0.66	0.30	0.47	0.77
MLRSSC (Pairwise)	0.72	0.51	0.42	0.47	0.38	0.57	0.38	0.76	0.75
MLRSSC (Centroid)	0.73	0.45	0.49	0.45	**0.46**	0.58	0.40	0.68	0.77
KMLRSSC (Pairwise)	0.68	0.51	0.41	0.09	0.45	0.52	0.33	0.33	0.75
KMLRSSC (Centroid)	0.71	0.50	0.41	0.48	0.40	0.51	0.34	0.33	0.76
CSMSC	0.38	0.51	0.44	0.55	0.41	0.63	0.28	0.67	0.73
NSM_MNMF	**0.78**	0.55	**0.54**	**0.72**	0.42	**0.71**	**0.54**	**0.83**	0.87

4.3. Baseline algorithms

In this study, sixteen algorithms were used to compare with our NSM_MNMF:

(1) MultiNMF (Liu, Wang, Gao, & Han, 2013): joint non-negative matrix factorization, decomposes the original data $\mathscr{X} = \{\mathbf{X}^{(v)}\}_{v=1}^{n_v}$ directly into multiple basis matrices and multiple coefficient matrices

(2) MultiNMF (Graph): a variant of MultiNMF, decomposes affinity matrices $\mathscr{A} = \{\mathbf{A}^{(v)}\}_{v=1}^{n_v}$

(3) GMultiNMF (Wang, Kong, Fu, Li, & Zhang, 2015): considers the inner-view relatedness of data, and decomposes the original data $\mathscr{X} = \{\mathbf{X}^{(v)}\}_{v=1}^{n_v}$ directly

(4) GMultiNMF (Graph): a variant of GMultiNMF, decomposes affinity matrices $\mathscr{A} = \{\mathbf{A}^{(v)}\}_{v=1}^{n_v}$

(5) SNMNMF (Zhang et al., 2012): the sparse regularized NMF, enforces the must-link constraints on the affinity matrices representing interaction of samples, and decomposes the original data $\mathscr{X} = \{\mathbf{X}^{(v)}\}_{v=1}^{n_v}$ into a common basis matrix and multiple coefficient matrices

(6) SNMNMF (Graph): a variant of SNMNMF, decomposes affinity matrices $\mathscr{A} = \{\mathbf{A}^{(v)}\}_{v=1}^{n_v}$

(7) MNMF (Du et al., 2018): decomposes affinity matrices $\mathscr{A} = \{\mathbf{A}^{(v)}\}_{v=1}^{n_v}$, without any regularization being imposed on NMF;

(8) CMNMF: a variant of MNMF with L_2 norm constraint, whose objective function is defined as: $J_{\mathrm{CMNMF}} = \frac{1}{2}$

$$\left(\sum_{v=1}^{n_v}\left\|\mathbf{A}^{(v)} - \mathbf{P}\left(\mathbf{Q}^{(v)}\right)^{\mathrm{T}}\right\|_{\mathrm{F}}^{2} + \lambda_1\|\mathbf{P}\|_{\mathrm{F}}^{2} + \lambda_2\sum_{v=1}^{n_v}\|\mathbf{Q}^{(v)}\|_{\mathrm{F}}^{2}\right) s.t. \mathbf{P}, \mathbf{Q}^{(v)} \geqslant 0$$

(9) GMNMF: a variant of CMNMF with local graph regularization $\mathbf{L}^{(v)}$, whose objective function is defined as: $J_{\mathrm{GMNMF}} = \frac{1}{2}$

$$\left(\sum_{v=1}^{n_v}\left\|\mathbf{A}^{(v)} - \mathbf{P}\left(\mathbf{Q}^{(v)}\right)^{\mathrm{T}}\right\|_{F}^{2} + \lambda_1\|\mathbf{P}\|_{\mathrm{F}}^{2} + \lambda_2\sum_{v=1}^{n_v}\|\mathbf{Q}^{(v)}\|_{\mathrm{F}}^{2}\right) + \frac{\gamma}{2}\mathrm{Tr}\left(\mathbf{P}\sum_{v=1}^{n_v}\mathbf{L}^{(v)}\mathbf{P}^{\mathrm{T}}\right) s.t. \mathbf{P}, \mathbf{Q}^{(v)} \geqslant 0$$

(10) Co-Reg (Pairwise) (Kumar et al., 2011): co-regularized multi-view spectral clustering with pairwise co-regularization ensures that the eigenvectors $\mathbf{U}^{(v)}$ and $\mathbf{U}^{(w)}$ of a view pair (v, w) have high pairwise similarity

(11) Co-Reg (Centroid) (Kumar et al., 2011): co-regularized multi-view spectral clustering with centroid co-regularization ensures the view-specific eigenvectors look similar by regularizing them towards a common consensus

(12) MLRSSC (Pairwise) (Brbić & Kopriva, 2018): multi-view low-rank sparse subspace clustering based on pairwise similarities, encourages similarity between pairs of representation matrices;

(13) MLRSSC (Centroid) (Brbić & Kopriva, 2018): centroid-based MLRSSC enforces representations across different views towards a common centroid

(14) KMLRSSC (Pairwise) (Brbić & Kopriva, 2018): the kernel extension of MLRSSC (Pairwise) solves the problem in a Reproducing Kernel Hilbert Space (RKHS)

(15) KMLRSSC (Centroid) (Brbić & Kopriva, 2018): the kernel extension of MLRSSC (Centroid) solves the problem in a RKHS

(16) CSMSC (Lu, Yan, & Lin, 2016): convex sparse multi-view spectral clustering with pairwise regularization

The variants of MultiNMF, GMultiNMF, and SNMNMF, i.e., MultiNMF (Graph), GMultiNMF (Graph), and SNMNMF (Graph), were designed to evaluate **W1**. By comparing the results of decomposition from the original data and the networks, we can evaluate whether factorizing networks can get more rational clustering results than factorizing the original data.

CMNMF, GMNMF, Co-reg, MLRSSC, and CSMSC were used to evaluate **W2**. Comparing the results of decomposition under different constraints, it is possible to evaluate whether regularizations of NSM_MNMF can preserve more intrinsic geometrical structures.

The comparison between the results of MultiNMF and GMultiNMF with those of SNMNMF, MNMF, and NSM_MNMF is helpful to evaluate

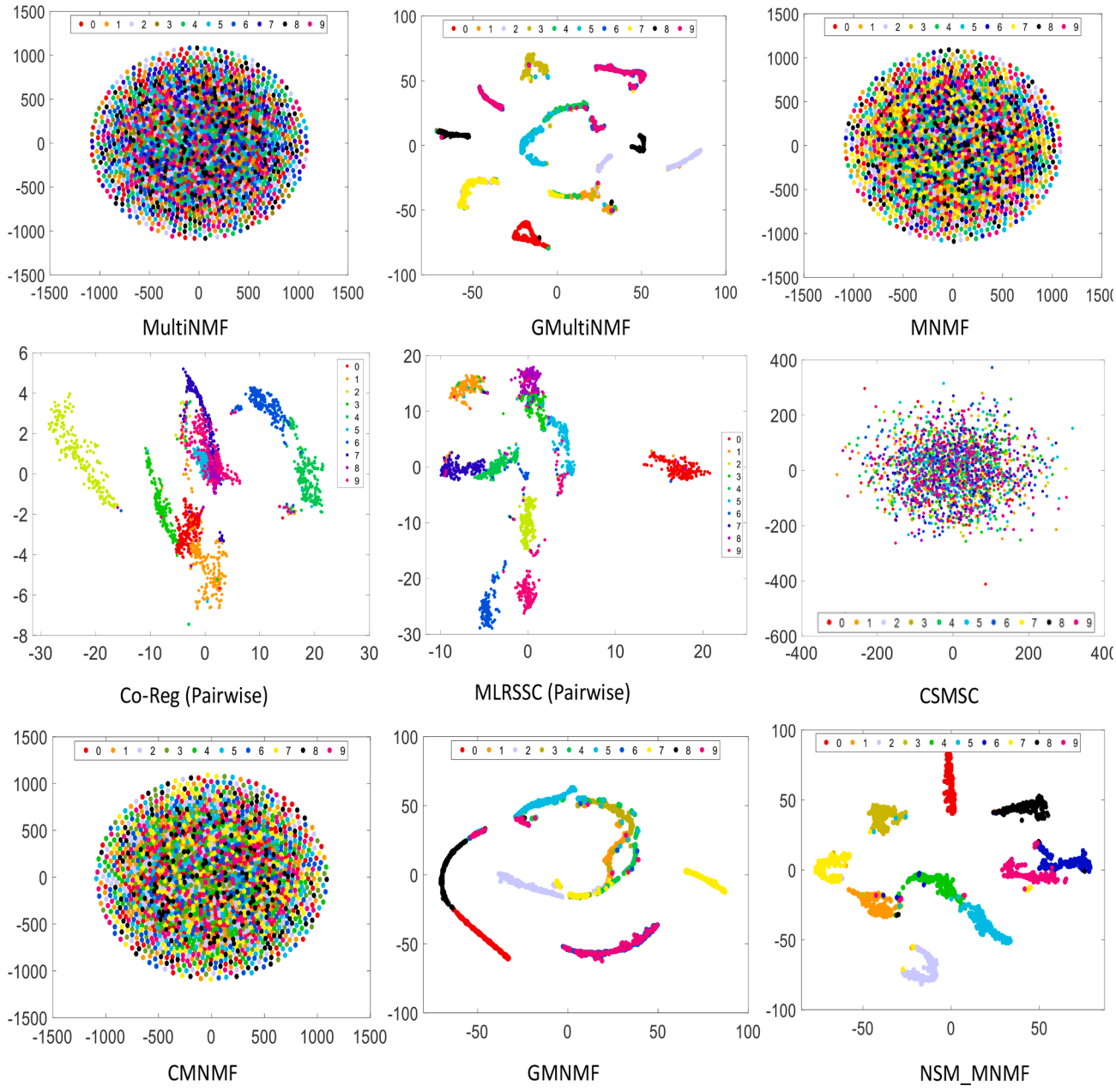

Fig. 2. The visualization of clustering results obtained by different NMF.

W3 because MultiNMF and GMultiNMF produce different basis matrices, while SNMNMF, MNMF, and NSM_MNMF jointly factorize matrices such that they produce only a shared basis matrix.

The **W4** was evaluated by comparing the results obtained by carrying out NSM_MNMF under different regularization parameters (Section 4.6 and 4.7).

For baselines, MultiNMF,[6] MultiGNMF,[7] SNMNMF,[8] Co-reg,[9] MLRSSC,[10] and CSMSC,[11] we executed the source codes released by the authors and tuned their parameters to obtain clusters with the highest metric values.

[6] https://github.com/SunMuxin/Multi-view/tree/master/multiviewNMF
[7] https://github.com/DUT-DIPLab/Graph-Multi-NMF-Feature-Clustering
[8] http://page.amss.ac.cn/shihua.zhang/software.html
[9] https://github.com/dugzzuli/CoregularizedSC
[10] https://github.com/Geovhbn/MLRSSC
[11] https://github.com/dugzzuli/CSMSC

4.4. Clustering performance

The parameters of NSM_MNMF are presented in Table 2, while the parameters of the baseline algorithms are presented in Appendix B. MultiNMF (graph), GMultiNMF (graph), and SNMNMF (graph) use the same parameters as MultiNMF, GMultiNMF, and SNMNMF. Since the clusters discovered by NMF-based algorithms were related to the initial values chosen randomly, each algorithm was run 20 times for each dataset, and the average accuracy is reported. The clustering results are presented in Tables 3–5, where bold numbers represent the best results. Note that no results exist for MultiNMF, GMultiNMF, and SNMNMF on LP1 ∼ LP5 datasets because these datasets are not strictly non-negative, meaning that the three algorithms cannot be used.

From Tables 3–5, we made the following observations, which answer the questions **W1** ∼ **W3:**

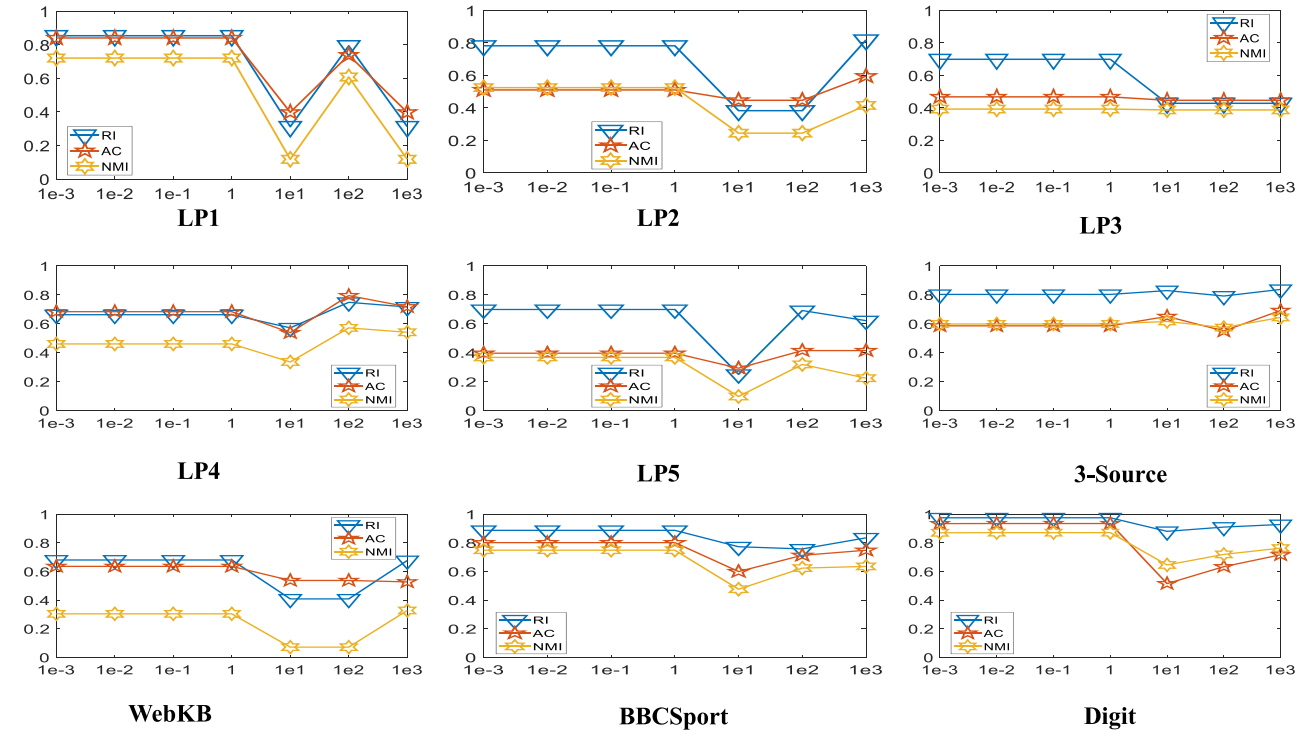

Fig. 3. Performance of the NSM_MNMF w.r.t. parameters λ_2 ($\lambda_1 = \gamma_1 = \gamma_2 = 0.001$).

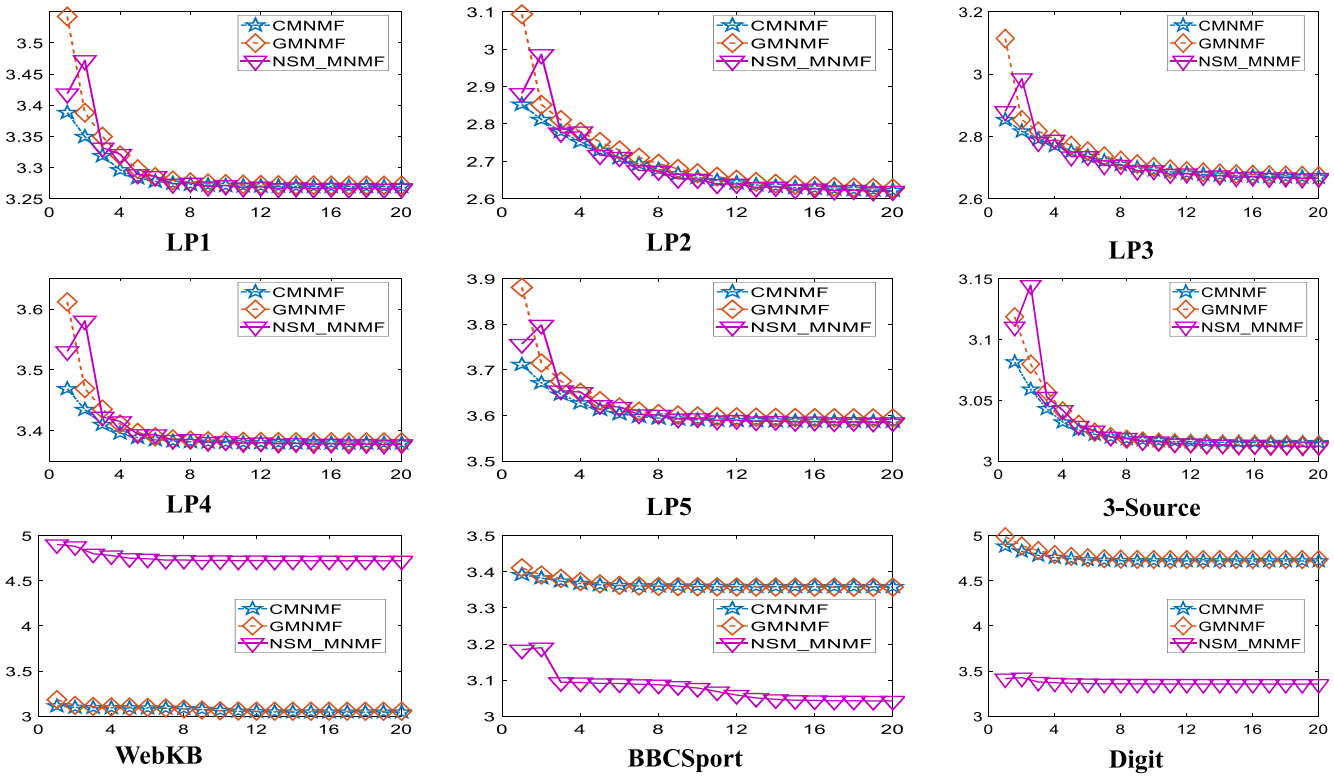

Fig. 4. Convergence curve of CMNMF, GMNMF and NSM_MNMF algorithms.

- Amongst all methods, NSM_MNMF achieves the highest measures with respect to the three metrics (RI, AC, NMI) in most datasets. This shows that NSM_MNMF can capture the global intra-view relationships within a view and inter-view relationships across multiple views; meanwhile, NSM_MNMF also preserves the intrinsic geometrical information embedded in the network space.
- Amongst MultiNMF, GMultiNMF, and SNMNMF and their variants, methods factorizing the networks outperform those factorizing the

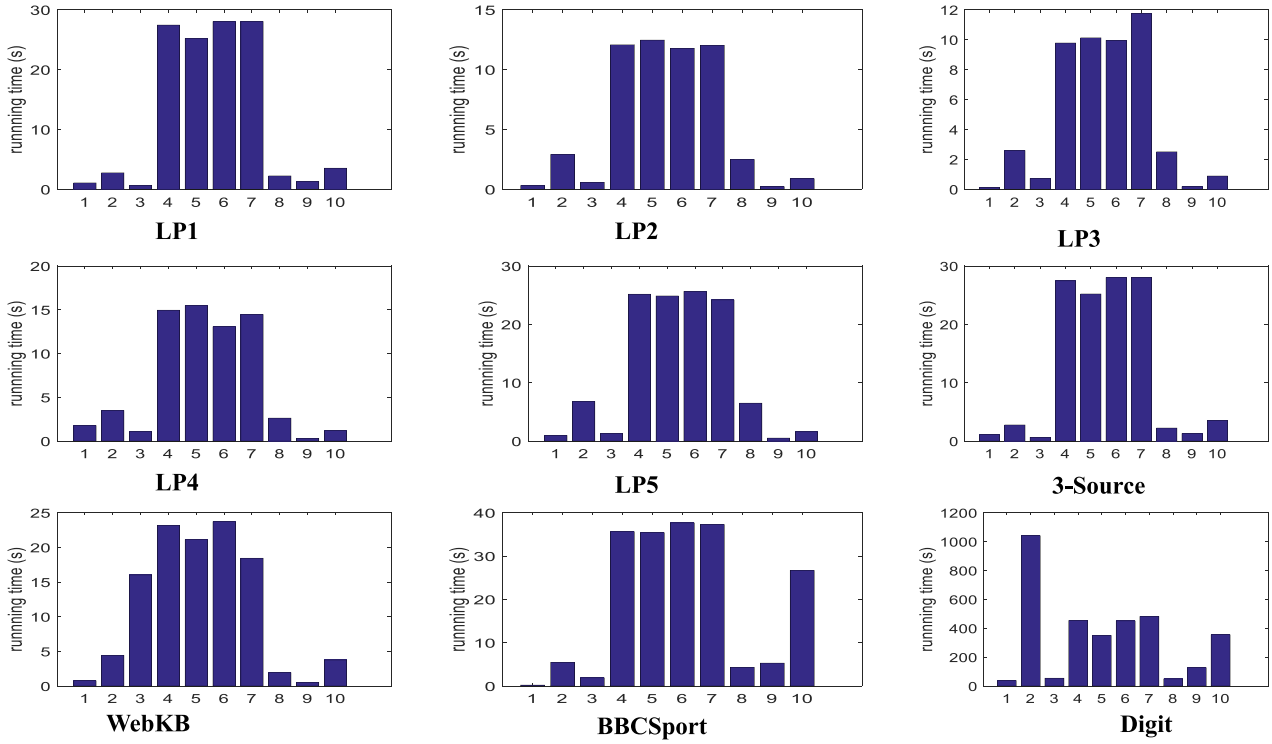

Fig. 5. The comparison of consumed time of algorithms on all datasets.

original data only. This shows that in clustering multi-view data, utilizing global relationships amongst nodes characterized by networks can improve the clustering performance.

- Amongst NSM_MNMF, MultiNMF, GMultiNMF, MNMF, Cog-Reg, MLRSSC, and CSMSC and their variants, methods with graph regularization in general are superior to those without graph regularization or those that use the L_2 norm. This can be explained because graph regularization preserves the intrinsic geometrical information embedded in the dataset or network space.
- Amongst all the methods, those producing a shared basis matrix surpass methods producing different basis matrices, which suggests that joint factorization can capture both intra-view and inter-view similarities amongst nodes.

4.5. Visualization of clustering results

We used t-SNE (Maaten & Hinton, 2008) to visualize clustering results, i.e., each sample is mapped as a point in a two-dimensional space, and each cluster is denoted as a color. A good clustering method is expected to produce clusters such that samples within the same cluster (i.e., points with the same color) are close to each other, while samples belonging to different clusters (points with different colors) are separated far apart in the two-dimensional space. As a representative case, Fig. 2 presents the visualization of clusters obtained by different approaches in the Digit dataset. From Fig. 2, we can observe that NSM_MNMF can obtain clearer visualization results compared with the other baseline methods, as the points with the same color are close to each other, and the points with different colors are far from each other. GmultiNMF, GMNMF, Co-Reg (Pairwise), and MLRSSC (Pairwise) do not perform as well as NSM_MNMF because although the points with different colors are far from each other, the points with the same color are not condensed enough. Meanwhile, in the results of MultiNMF, MNMF, CSMSC, and CMNMF, all points with different colors are mixed together such that it is difficult to distinguish different clusters. This further verifies the effectiveness of the proposed NSM_MNMF method.

4.6. Parameter study

There are four regularization parameters in the NSM_MNMF algorithm: λ_1,λ_2, γ_1 and γ_2, where λ_1 and λ_2 are utilized to adjust the weights of the multi-manifold regularization, γ_1 is applied to suppress the growth of $\mathbf{P}$, and γ_2 controls the sparseness of $\mathbf{Q}^{(v)}$. To investigate the influence of these parameters on clustering results, the Algorithm NSM_MNMF is executed under different parameters. Each parameter ranges from 1e to 3 to 1e+3 with step 0.001, and we chose a value from this interval for a parameter with the fixed values of the other three parameters.

Fig. 3 presents the values of the RI, AC, and NMI with respect to different λ_2 on nine datasets. The trends of the RI, AC, and NMI with respect to different λ_1, γ_1, and γ_2, are similar to the one with respect to λ_2, so we do not present them.

As we can see from Fig. 3, NSM_MNMF performs in a relatively stable manner on all datasets when $0\leqslant\lambda_2\leqslant1$. Based on the experiments with respect to λ_1, γ_1, and γ_2, we suggest that the appropriate ranges for the four parameters are $0\leqslant\lambda_1\leqslant0.01$, $0\leqslant\lambda_2\leqslant1$, $0\leqslant\gamma_1\leqslant0.01$, and $0\leqslant\gamma_2\leqslant0.01$ respectively, which can ensure that the NSM_MNMF performance is relatively stable on all datasets.

4.7. Algorithm convergence

In this sub-section, we examine the convergences of CMNMF, GMNMF, and NSM_MNMF. The convergence curves of three algorithms are shown in Fig. 4, where the abscissa denotes the iteration number, and the ordinate represents the log-value of the objective function. It can be seen that in all datasets, the objective values of all algorithms decrease quickly within 10 iterations and converge within 50 iterations. Meanwhile, after convergence, the log-value of the objective function of NSM_MNMF is less than those of CMNMF and GMNMF except on WebKB. This indicates that the objective function and updating rules designed for NSM_MNMF serve their purpose.

4.8. Consumed time

Fig. 5 presents the execution times of ten algorithms on nine datasets, where the numbers 1–10 under the horizontal axis denote MultiNMF (graph), GMultiNMF (graph), SNMNMF, MNMF, CMNMF, GMNMF, NSM_MNMF, Cog-Reg, MLRSSC, and CSMSC respectively. The number of iterations was set as 500 for all algorithms, while the number of internal iterations of MultiNMF and GMultiNMF was set as 1.

It can be seen from Fig. 5 that MultiNMF runs fastest on all datasets; the execution times of MNMF, CMNMF, GMNMF, and NSM_MNMF have no significant difference, and in general, they are more time consuming than other algorithms. This indicates that NMFs with a shared basis matrix take more time than NMFs with different basis matrices. One exception is that GMultiNMF (graph) spends the most time on the Digit dataset. Cog-Reg, MLRSSC, and CSMSC are faster than MNMF, CMNMF, GMNMF, and NSM_MNMF.

5. Conclusion

The challenge for clustering multi-view data is how to capture the intra-view similarities amongst data samples within a view and the inter-view similarities across distinct views, and how to make the clustering results interpretable. In this study, we developed an approach of network-based sparse and multi-manifold regularized multiple NMF (NSM_MNMF) for clustering multi-view data. This approach first transforms multi-view data into multiple networks and then jointly factorizes these networks. We also proposed a novel method to factorize multiple networks, where multiple regularizations are imposed on NMF and efficient updating rules are derived for computing optimal factorized matrices. Furthermore, we conducted extensive experiments on nine real multi-view datasets and compared them with sixteen baseline approaches using three evaluation metrics; NSM_MNMF achieves the highest measures in terms of RI, AC, and NMI in most datasets, which demonstrates that NSM_MNMF can capture the global intra-view relationships and inter-view relationships. Meanwhile, NSM_MNMF also preserves the intrinsic geometrical information embedded in the network space. The combination of NMF and network representation enables NSM_MNMF not only to be equipped with the ability to characterize both local and global relationships amongst samples but also inherits the high interpretability of NMF, meanwhile enabling NSM_MNMF to be applied to datasets that are not strictly non-negative and, therefore, extending the applicability of NMF.

As for future work direction, we will continue to study how to select regularization parameters automatically, such as by bi-level optimization or attention networks. In addition, we consider to use the method of deep architecture to extend NSM_MNMF to automatically learn lower-dimensional hierarchal and non-linear features embedded in the multi-view data for further improving the performance of clustering.

CRediT authorship contribution statement

Lihua Zhou: Conceptualization, Methodology, Writing - original draft. **Guowang Du:** Software, Validation, Data curation. **Kevin Lü:** Writing - review & editing. **Lizhen Wang:** Supervision, Methodology.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (61762090, 62062066, and 61966036), the Program for Innovation Research Team (in Science and Technology) in University of Yunnan Province (IRTSTYN), and the National Social Science Foundation of China (18XZZ005).

Appendix A. (the proof of the Theorem 1):

The objective function J of NSM_MNMF in Eq. (1) is certainly bounded from below by zero.

We will follow the similar procedure described in Lee and Seung (2001) to prove Theorem 1. It begins with the definition of the auxiliary function.

Definition. *$G(q^{(v)}, q^{'(v)})$ is an auxiliary function for $F(q^{(v)})$ if the conditions $G(q^{(v)}, q^{'(v)}) \geqslant F(q^{(v)}), \mathrm{G}(\mathrm{q}^{(\mathrm{v})}, \mathrm{q}^{(\mathrm{v})}) = \mathrm{F}(\mathrm{q}^{(\mathrm{v})})$ are satisfied.*

Lemma 1:. *If $G(q^{(v)}, q^{'(v)})$ is an auxiliary function of $F(q^{(v)})$, then $F(q^{(v)})$ is non-increasing under the update: $q^{(v)(t+1)} = \underset{q^{(v)}}{argmin} G(q^{(v)}, q^{(v)(t)})$.*

Proof: .. *Considering any element $q_{ab}^{(v)}$ in $\mathbf{Q}^{(v)}$, we use $F_{ab}^{(v)}$ to denote the part of $J_{\mathbf{Q}^{(v)}}$ which is relevant only to $q_{ab}^{(v)}$. It is easy to check that:*

$$F_{ab}^{'(v)} = \left(\frac{\partial J_{\mathbf{Q}^{(v)}}}{\partial \mathbf{Q}^{(v)}}\right)_{ab} = \left(-(\mathbf{A}^{(v)})^{\mathrm{T}}\mathbf{P} + \mathbf{Q}^{(v)}\mathbf{P}^{\mathrm{T}}\mathbf{P} - \mathbf{Q}^{*} + \mathbf{Q}^{(v)} + \gamma_2\mathbf{I} + \lambda_1\mathbf{L}^{(v)}\mathbf{Q}^{(v)}\right)_{ab}$$

$$F_{ab}^{''(v)} = \left(\mathbf{P}^{\mathrm{T}}\mathbf{P} + \mathbf{I}\right)_{bb} + \lambda_1(\mathbf{L}^{(v)})_{aa}$$

Lemma 2:. *Function*

$G\left(q_{ab}^{(v)}, q_{ab}^{(v)(t)}\right) = F_{ab}^{(v)}\left(q_{ab}^{(v)(t)}\right) + F_{ab}^{'(v)}\left(q_{ab}^{(v)(t)}\right)\left(q_{ab}^{(v)} - q_{ab}^{(v)(t)}\right) + \frac{\left(\mathbf{Q}^{(v)}\mathbf{P}^{\mathrm{T}}\mathbf{P} + \mathbf{Q}^{(v)} + \lambda_1\mathbf{D}^{(v)}\mathbf{Q}^{(v)}\right)_{ab} + \gamma_2}{2q_{ab}^{(v)(t)}}\left(q_{ab}^{(v)} - q_{ab}^{(v)(t)}\right)^2$ is an auxiliary function for $F_{ab}^{(v)}$, the part of $J_{\mathbf{Q}^{(v)}}$ which is relevant only to $q_{ab}^{(v)}$.

Proof.. *It is obvious that $G\left(q_{ab}^{(v)}, q_{ab}^{(v)}\right) = F_{ab}^{(v)}\left(q_{ab}^{(v)}\right)$, so we need only show that $G\left(q_{ab}^{(v)}, q_{ab}^{(v)(t)}\right) \geqslant F_{ab}^{(v)}\left(q_{ab}^{(v)}\right)$. To do this, we compare $G\left(q_{ab}^{(v)}, q_{ab}^{(v)(t)}\right)$ with the Taylor series expansion of $F_{ab}^{(v)}(q^{(v)})$:*

$$F_{ab}^{(v)}\left(q_{ab}^{(v)}\right) = F_{ab}^{(v)}\left(q_{ab}^{(v)(t)}\right) + F_{ab}^{'(v)}\left(q_{ab}^{(v)(t)}\right)\left(q_{ab}^{(v)} - q_{ab}^{(v)(t)}\right) + \frac{\left(\mathbf{P}^{\mathrm{T}}\mathbf{P} + \mathbf{I}\right)_{bb} + \lambda_1(\mathbf{L}^{(v)})_{aa}}{2}\left(q_{ab}^{(v)} - q_{ab}^{(v)(t)}\right)^2$$

To find that $G\left(q_{ab}^{(v)}, q_{ab}^{(v)(t)}\right) \geqslant F_{ab}^{(v)}\left(q_{ab}^{(v)}\right)$ is equivalent to

$$\frac{\left(\mathbf{Q}^{(v)}\mathbf{P}^{\mathrm{T}}\mathbf{P}+\mathbf{Q}^{(v)}+\lambda_1\mathbf{D}^{(v)}\mathbf{Q}^{(v)}\right)_{ab}+\gamma_2}{q_{ab}^{(v)(t)}}\geqslant\left(\mathbf{P}^{\mathrm{T}}\mathbf{P}+\mathbf{I}\right)_{bb}+\lambda_1\left(\mathbf{L}^{(v)}\right)_{aa}$$

Because $\left(\mathbf{Q}^{(v)}\mathbf{P}^{\mathrm{T}}\mathbf{P}\right)_{ab}=\sum_{l=1}^{k}q_{al}^{(v)(t)}\left(\mathbf{P}^{\mathrm{T}}\mathbf{P}\right)_{lb}\geqslant q_{ab}^{(v)(t)}\left(\mathbf{P}^{\mathrm{T}}\mathbf{P}\right)_{bb}$, $\frac{\left(\mathbf{Q}^{(v)}\right)_{ab}}{q_{ab}^{(v)(t)}}=(\mathbf{I})_{bb}$,

$$\lambda_1\left(\boldsymbol{D}^{(v)}\boldsymbol{Q}^{(v)}\right)_{ab}=\lambda_1\sum_{l=1}^{M}\mathbf{D}_{al}^{(v)}q_{lb}^{(v)(t)}\geqslant\lambda_1\mathbf{D}_{aa}^{(v)}q_{ab}^{(v)(t)}\geqslant\lambda_1\left(\mathbf{D}^{(v)}-\mathbf{A}^{(v)}\right)_{aa}q_{ab}^{(v)(t)}=\lambda_1\mathbf{L}_{aa}^{(v)}q_{ab}^{(v)(t)}$$

Thus, $\frac{\left(\boldsymbol{Q}^{(v)}\mathbf{P}^{\mathrm{T}}\mathbf{P}+\mathbf{Q}^{(v)}+\lambda_1\mathbf{D}^{(v)}\mathbf{Q}^{(v)}\right)_{ab}}{q_{ab}^{(v)(t)}}\geqslant\left(\mathbf{P}^{\mathrm{T}}\mathbf{P}+\mathbf{I}\right)_{bb}+\lambda_1\left(\mathbf{L}^{(v)}\right)_{aa}$.

Also because $\frac{\left(\mathbf{Q}^{(v)}\mathbf{P}^{\mathrm{T}}\mathbf{P}+\mathbf{Q}^{(v)}+\lambda_1\mathbf{D}^{(v)}\mathbf{Q}^{(v)}\right)_{ab}+\gamma_2}{q_{ab}^{(v)(t)}}\geqslant\frac{\left(\mathbf{Q}^{(v)}\mathbf{P}^{\mathrm{T}}\mathbf{P}+\mathbf{Q}^{(v)}+\lambda_1\mathbf{D}^{(v)}\mathbf{Q}^{(v)}\right)_{ab}}{q_{ab}^{(v)(t)}}$, thus

$$\frac{\left(\mathbf{Q}^{(v)}\mathbf{P}^{\mathrm{T}}\mathbf{P}+\mathbf{Q}^{(v)}+\lambda_1\mathbf{D}^{(v)}\mathbf{Q}^{(v)}\right)_{ab}+\gamma_2}{q_{ab}^{(v)(t)}}\geqslant\left(\mathbf{P}^{\mathrm{T}}\mathbf{P}+\mathbf{I}\right)_{bb}+\lambda_1\left(\mathbf{L}^{(v)}\right)_{aa}$$

It means $G\left(q_{ab}^{(v)},q_{ab}^{(v)(t)}\right)\geqslant F_{ab}^{j}\left(q_{ab}^{(v)}\right)$.

Proof of theorem 1. *$q_{ab}^{(v)(t+1)}$ can be obtained by minimizing $G\left(q_{ab}^{(v)},q_{ab}^{(v)(t)}\right)$. To do this, we compute the partial derivative of the auxiliary function $G\left(q_{ab}^{(v)},q_{ab}^{(v)(t)}\right)$ to $q_{ab}^{(v)}$ and let it be equal to* 0:

$$\frac{\partial G\left(q_{ab}^{(v)},q_{ab}^{(v)(t)}\right)}{\partial q_{ab}^{(v)}}=F_{ab}^{'(v)}\left(q_{ab}^{(v)(t)}\right)+\frac{\left(\mathbf{Q}^{(v)}\mathbf{P}^{\mathrm{T}}\mathbf{P}+\mathbf{Q}^{(v)}+\lambda_1\mathbf{D}^{(v)}\mathbf{Q}^{(v)}\right)_{ab}+\gamma_2}{q_{ab}^{(v)(t)}}\left(q_{ab}^{(v)}-q_{ab}^{(v)(t)}\right)$$

$$=\left(-\left(\mathbf{A}^{(v)}\right)^{\mathrm{T}}\mathbf{P}+\mathbf{Q}^{(v)}\mathbf{P}^{\mathrm{T}}\mathbf{P}-\mathbf{Q}^{*}+\mathbf{Q}^{(v)}+\gamma_2\mathbf{I}+\lambda_1\mathbf{L}^{(v)}\mathbf{Q}^{(v)}\right)_{ab}-\left(\mathbf{Q}^{(v)}\mathbf{P}^{\mathrm{T}}\mathbf{P}+\mathbf{Q}^{(v)}+\lambda_1\mathbf{D}^{(v)}\mathbf{Q}^{(v)}\right)_{ab}-\gamma_2$$

$$+\frac{\left(\mathbf{Q}^{(v)}\mathbf{P}^{\mathrm{T}}\mathbf{P}+\mathbf{Q}^{(v)}+\lambda_1\mathbf{D}^{(v)}\mathbf{Q}^{(v)}\right)_{ab}+\gamma_2}{q_{ab}^{(v)(t)}}q_{ab}^{(v)}=0$$

$$\left(-\left(\mathbf{A}^{(v)}\right)^{\mathrm{T}}\mathbf{P}-\mathbf{Q}^{*}-\lambda_1\mathbf{A}^{(v)}\mathbf{Q}^{(v)}\right)_{ab}+\frac{\left(\mathbf{Q}^{(v)}\mathbf{P}^{\mathrm{T}}\mathbf{P}+\mathbf{Q}^{(v)}+\lambda_1\mathbf{D}^{(v)}\mathbf{Q}^{(v)}\right)_{ab}+\gamma_2}{q_{ab}^{(v)(t)}}q_{ab}^{(v)}=0$$

Thus $q_{ab}^{(v)}=q_{ab}^{(v)(t)}\frac{\left(\left(\boldsymbol{A}^{(v)}\right)^{\mathrm{T}}\mathbf{P}+\mathbf{Q}^{*}+\lambda_1\mathbf{A}^{(v)}\mathbf{Q}^{(v)}\right)_{ab}}{\left(\boldsymbol{Q}^{(v)}\boldsymbol{P}^{\mathrm{T}}\mathbf{P}+\mathbf{Q}^{(v)}+\lambda_1\mathbf{D}^{(v)}\mathbf{Q}^{(v)}\right)_{ab}+\gamma_2}$.

Since $G\left(q_{ab}^{(v)},q_{ab}^{(v)(t)}\right)$ is an auxiliary function of $F_{ab}^{(v)}$, $F_{ab}^{(v)}$ is non-increasing under this update rule.

Appendix B. (the parameters of the baseline algorithms):

		Robot Execution Failures ($k=7$)					3-Source($k=7$)	WebKB($k=7$)	BBCSport($k=7$)	Digit($k=100$)
		LP1	LP2	LP3	LP4	LP5				
MultiNMF	λ_v	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
GMultiNMF	λ_f	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
	μ	10	10	10	10	10	10	10	10	10
SNMNMF	λ_1	1e-3	100	1e+3	0.1	100	10	1e-3	10	1e-3
	λ_2	0.1	10	10	10	100	0.1	0.1	0.1	1e-3
	γ_1	1e-5	1e-4	1e-4	0.01	0.1	0.01	1e-5	0.01	1e-3
	γ_2	1e-3	1e-4	1e-4	0.1	0.01	1e-5	0.1	1e-5	1e-3
CMNMF	λ_1	0.1	100	10	1e-3	10	1e-3	1e-3	1e-3	1e-3
	λ_2	100	1	10	1e-3	10	1e-3	1e-3	1e-3	1e-3
GMNMF	λ	1e-3	10	10	10	100	1e-3	1e-3	1e-3	0.1
	γ	1e-3	1e-3	10	10	1e-3	1e-3	1e-3	1e-3	0.1
Co-Reg (Pairwise)	λ	0.1	1e-5	1e-4	0.01	0.1	1e-5	0.01	0.01	0.1
Co-Reg (Centroid)	λ	0.1	1e-3	1e-3	0.1	0.1	0.1	1e-3	1e-5	1e-3
MLRSSC(Pairwise)	μ	100	100	100	100	100	100	1e+4	100	100
	β_1	0.3	0.3	0.3	0.3	0.3	0.3	0.9	0.3	0.5
	β_2	0.7	0.7	0.7	0.7	0.7	0.7	0.1	0.7	0.5
	λ	0.4	0.4	0.4	0.4	0.4	0.3	0.7	0.4	0.7
MLRSSC(Centroid)	μ	100	100	100	100	100	100	1e+4	100	10
	β_1	0.1	0.1	0.1	0.1	0.1	0.1	0.5	0.1	0.9
	β_2	0.9	0.9	0.9	0.9	0.9	0.9	0.5	0.9	0.1
	λ	0.7	0.7	0.7	0.7	0.7	0.7	0.3	0.7	0.7
KMLRSSC(Pairwise)	μ	1e+3	1e+3	1e+3	1e+3	1e+3	1e+3	1e+4	1e+3	1e+4
	β_1	0.3	0.3	0.3	0.3	0.3	0.3	0.5	0.3	0.7
	β_2	0.7	0.7	0.7	0.7	0.7	0.7	0.5	0.7	0.3
	λ	0.4	0.4	0.4	0.4	0.4	0.3	0.9	0.4	0.7
KMLRSSC(Centroid)	μ	1e+4	1e+4	1e+4	1e+4	1e+4	1e+4	1e+4	1e+4	1e+4
	β_1	0.3	0.3	0.3	0.3	0.3	0.3	0.5	0.3	0.7
	β_2	0.7	0.7	0.7	0.7	0.7	0.7	0.5	0.7	0.3
	λ	0.7	0.7	0.7	0.7	0.7	0.7	0.3	0.7	0.5

(*continued on next page*)

(*continued*)

		Robot Execution Failures ($k = 7$)					3-Source($k = 7$)	WebKB($k = 7$)	BBCSport($k = 7$)	Digit($k = 100$)
		LP1	LP2	LP3	LP4	LP5				
CSMSC	α	0.01	1e-4	0.1	0.1	1e-5	0.01	1e-4	0.01	1e-3
	β	1e-5	1e-5	1e-4	1e-4	1e-5	1e-3	1e-5	1e-5	1e-4
NSM_MNMF	λ_1	1	1e-3	0.01	0.1	1	100	1e+3	10	1
	λ_2	100	1	1	10	0.01	1	100	0.1	1e-3
	γ_1	100	0.1	1e-3	1e+3	1e-3	1e-3	1	100	1e-3
	γ_2	1e-3	100	1e+3	1e-3	0.01	100	100	0.01	1e-3

References

Bickel, S., & Scheffer, T. 2004. Multi-view clustering., The Fourth IEEE International Conference on Data Mining (ICDM 2004): 19-26. Brighton, UK.

Brbić, M., & Kopriva, I. (2018). Multi-view low-rank sparse subspace clustering. *Pattern Recognition, 73*, 247–258.

Cai, D., He, X., Han, J., & Huang, T. S. (2011). Graph regularized nonnegative matrix factorization for data representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 33*(8), 1548–1560.

Camarinha-Matos, L. M., Seabra Lopes, L., & Barata, J. 1996. Integration and learning in supervision of flexible assembly systems. IEEE Transactions on Robotics and Automation, 12(2): 202–219.

Dai, Q.-Z., Xiong, Z.-Y., Xie, J., Wang, X.-X., Zhang, Y.-F., & Shang, J.-X. (2019). A novel clustering algorithm based on the natural reverse nearest neighbor structure. *Information Systems, 84*, 1–16. https://doi.org/10.1016/j.is.2019.04.001

Ding, C. H. Q., Tao Li, & Jordan, M. I. (2010). Convex and semi-nonnegative matrix factorizations. *IEEE Trans. Pattern Anal. Mach. Intell., 32*(1), 45–55. https://doi.org/10.1109/TPAMI.2008.277

Du, G., Zhou, L., Wang, L., & Chen, H. (2018). Multivariate time series clustering via multi-relational community detection in networks, Asia-Pacific Web (APWeb) and Web-Age Information Management (WAIM) Joint International Conference on Web and Big Data (APWeb-WAIM 2018): 138-145. Macau, China.

Feng, Y., Xiao, J., Zhou, K., & Zhuang, Y. (2015). A locally weighted sparse graph regularized non-negative matrix factorization method. *Neurocomputing, 169*, 68–76.

Ferreira, L. N., & Zhao, L. (2016). Time series clustering via community detection in networks. *Information Sciences, 326*, 227–242. https://doi.org/10.1016/j.ins.2015.07.046

Halkidi, M., Batistakis, Y., & Vazirgiannis, M. (2001). On clustering validation techniques. *Journal of Intelligent Information Systems Integrating Artificial Intelligence & Database Technologies, 17*(2–3), 107–145.

Huang, S., Kang, Z., & Xu, Z. (2020). Auto-weighted multi-view clustering via deep matrix decomposition. *Pattern Recognition, 97*, 107015. https://doi.org/10.1016/j.patcog.2019.107015

Kong, D., Ding, C., & Huang, H. (2011). Robust nonnegative matrix factorization using L21-norm, the 20th ACM international conference on Information and knowledge management (CIKM 2011): 673-682. Glasgow, Scotland, UK.

Kumar, A., Rai, P., & Daumé, H. I. (2011). Co-regularized multi-view spectral clustering. Advances in Neural Information Processing Systems, 24: 1413–1421.

Lee, D. D., & Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature, 401*(6755), 788–791. https://doi.org/10.1038/44565

Lee, D. D., & Seung, H. S. (2001). Algorithms for non-negative matrix factorization. *Advances in Neural Information Processing Systems*, 556–562.

Liu, J., Wang, C., Gao, J., & Han, J. (2013). Multi-view clustering via joint nonnegative matrix factorization, The 13th SIAM International Conference on Data Mining: 252–260. San Diego, California, USA.

Liu, X., Lu, H., & Gu, H. (2005). Group sparse non-negative matrix factorization for multi-manifold learning. *Intelligence, 27*(12), 1945–1959.

Lu, C., Yan, S., & Lin, Z. (2016). Convex sparse spectral clustering: single-view to multi-view. *IEEE Transactions on Image Processing, 25*(6), 2833–2843. https://doi.org/10.1109/TIP.2016.2553459

Maaten, L. V. D., & Hinton, G. (2008). Visualizing data using t-SNE. Journal of Machine Learning Research, 9(11): 2579–2605.

Paatero, P., & Tapper, U. (1994). Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics, 5*(2), 111–126. https://doi.org/10.1002/env.3170050203

Peng, C., Kang, Z., Hu, Y., Cheng, J., & Cheng, Q. (2017). Robust graph regularized nonnegative matrix factorization for clustering. *Acm Transactions on Knowledge Discovery from Data, 11*(3), 1–30.

Saini, N., Bansal, D., Saha, S., & Bhattacharyya, P. (2020). Multi-objective multi-view based search result clustering using differential evolution framework. *Expert Systems with Applications, 114299*.

Shang, F., Jiao, L. C., & Wang, F. (2012). Graph dual regularization non-negative matrix factorization for co-clustering. *Pattern Recognition, 45*(6), 2237–2250.

Shen, B., & Si, L. 2010. Nonnegative matrix factorization clustering on multiple manifolds, The 24th AAAI Conference on Artificial Intelligence (AAAI 2010): 575–580. Atlanta, Georgia, USA.

Singh, A. P., & Gordon, G. J. (2008). *Relational learning via collective matrix factorization,* ***The 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2008)****: 650–658*. USA: Las Vegas.

Sun, B.-J., Shen, H., Gao, J., Ouyang, W., & Cheng, X. (2017). A Non-negative Symmetric Encoder-Decoder Approach for Community Detection, The 2017 ACM Conference on Information and Knowledge Management (CIKM 2017): 597-606. Singapore.

Trigeorgis, G., Bousmalis, K., Zafeiriou, S., & Schuller, B. W. (2017). A deep matrix factorization method for learning attribute representations. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 39*(3), 417–429.

Wang, Y., Jiang, Y., Wu, Y., & Zhou, Z.-H. (2011). *Local and structural consistency for multi-manifold clustering,* ***The 22nd International Joint Conference on Artificial Intelligence (IJCAI 2011)****: 1559–1564*. Barcelona: Catalonia, Spain.

Wang, Z., Kong, X., Fu, H., Li, M., & Zhang, Y. (2015). *Feature extraction via multi-view non-negative matrix factorization with local graph regularization,* ***The 2015 IEEE International Conference on Image Processing (ICIP 2015)****: 3500–3504*. Canada: Quebec.

Wendel, A., Sternig, S., & Godec, M. (2011). Non-negative matrix factorization in multimodality data for segmentation and label prediction. In *16th Computer Vision Winter Workshop (CVWW 2011)* (pp. 1–8). Autriche: Mitterberg.

Yu, H., Wang, X., Wang, G., & Zeng, X. (2020). An active three-way clustering method via low-rank matrices for multi-view data. *Information Sciences, 507*(1), 823–839.

Zhan, K., Niu, C., Chen, C., Nie, F., Zhang, C., & Yang, Y. (2019). Graph structure fusion for multiview clustering. *IEEE Transtations on Knowledge and Data Engineering, 31*(10), 1984–1993.

Zhang, R., Nie, F., Li, X., & Wei, X. (2019). Feature selection with multi-view data: A survey. *Information Fusion, 50*(2019), 158–167.

Zhang, G.-Y., Zhou, Y.-R., Wang, C.-D., Huang, D., & He, X.-Y. (2021). Joint representation learning for multi-view subspace clustering. *Expert Systems with Application, 166*, Article 113913.

Zhang, S., Liu, C. C., Li, W., Hui, S., Laird, P. W., & Zhou, X. J. (2012). Discovery of multi-dimensional modules by integrative analysis of cancer genomic data. *Nucleic Acids Research, 40*(19), 9379–9391.

Zhang, X., Zong, L., Liu, X., & Yu, H. (2015). *Constrained NMF-based multi-view clustering on unmapped data,* ***The 29th AAAI Conference on Artificial Intelligence (AAAI 2015)****: 3174–3180*. Texas, USA: Austin.

Zhang, Z., Liu, L., Shen, F., Shen, H. T., & Shao, L. (2019). Binary multi-view clustering. *IEEE Transactions on Pattern Analysis & Machine Intelligence, 41*(7), 1774–1782.

Zhao, H., Ding, Z., & Fu, Y. 2017. Multi-View Clustering via Deep Matrix Factorization, 31st AAAI Conference on Artificial Intelligence (AAAI 2017): 2921-2927. San Francisco, USA.

Zong, L., Zhang, X., Zhao, L., Yu, H., & Zhao, Q. (2017). Multi-view clustering via multi-manifold regularized non-negative matrix factorization. *Neural Networks, 88*, 74–89.