Fast Nonnegative Matrix Tri-Factorization for Large-Scale Data Co-Clustering

Hua Wang, Feiping Nie, Heng Huang, Fillia Makedon
Department of Computer Science and Engineering
University of Texas at Arlington, Arlington, Texas 76019, USA
huawangcs@gmail.com, feipingnie@gmail.com, heng@uta.edu, makedon@uta.edu

Abstract

Nonnegative Matrix Factorization (NMF) based co-clustering methods have attracted increasing attention in recent years because of their mathematical elegance and encouraging empirical results. However, the algorithms to solve NMF problems usually involve intensive matrix multiplications, which make them computationally inefficient. In this paper, instead of constraining the factor matrices of NMF to be nonnegative as existing methods, we propose a novel Fast Nonnegative Matrix Tri-factorization (FNMTF) approach to constrain them to be cluster indicator matrices, a special type of nonnegative matrices. As a result, the optimization problem of our approach can be decoupled, which results in much smaller size subproblems requiring much less matrix multiplications, such that our approach works well for large-scale input data. Moreover, the resulted factor matrices can directly assign cluster labels to data points and features due to the nature of indicator matrices. In addition, through exploiting the manifold structures in both data and feature spaces, we further introduce the Locality Preserved FNMTF (LP-FNMTF) approach, by which the clustering performance is improved. The promising results in extensive experimental evaluations validate the effectiveness of the proposed methods.

1 Introduction

Clustering, which partitions a data set into different groups unsupervisedly, is one of the most fundamental topic in statistical learning. Most traditional clustering algorithms are designed for one-side clustering [Nie *et al.*, 2009], *i.e.* cluster either data points or features. However, in many real world applications, the clustering based analysis is interested in two-side clustering results, *i.e.* group the data points and features simultaneously, *e.g.*, “documents” and “words” in document analysis, “users” and “items” in collaborative filtering, “samples” and “genes” in microarray data analysis, *etc*. Typically, instead of being independent, the different clustering tasks on data and features are closely correlated, and it is challenging for traditional clustering algorithms to utilize the data and features interdependence efficiently. Consequently, *co-clustering* techniques, which aim to cluster both data and features simultaneously by leveraging the interrelations between them, have been proposed in recent researches. To name a few, Dhillon [Dhillon, 2001] introduced a bipartite spectral graph partition approach to co-cluster words and documents; Cho *et al*. [Cho *et al.*, 2004] suggested to co-cluster the experimental conditions and genes for microarray data by minimizing the sum-squared-residue; Long *et al*. [Long *et al.*, 2006] presented a relation summary network model to co-cluster the heterogeneous data on a k-partite graph, and so on.

More recently, Ding *et al.* [Ding *et al.*, 2005] explored the relationships between Nonnegative Matrix Factorization (NMF) [Lee and Seung, 1999; 2001] and K-means/spectral clustering, and proposed to use Nonnegative Matrix Tri-factorization (NMTF) [Ding *et al.*, 2006] to co-cluster words and documents at the same time. Due to its mathematical elegance and encouraging empirical results, NMTF method has been further developed to address various aspects of co-clustering [Wang *et al.*, 2008; Li *et al.*, 2010; Gu and Zhou, 2009; Ding *et al.*, 2010]. However, a notorious bottleneck of NMTF based co-clustering approaches is the slow computational speed because of intensive matrix multiplications involved in each iteration step of the solution algorithms, which makes these approaches hard to be applied to large-scale data in real world applications. In this paper, we propose a novel Fast Nonnegative Matrix Tri-factorization (FNMTF) approach to efficiently conduct co-clustering on large-scale data. Our new algorithms are interesting from the following perspectives:

- Instead of enforcing traditional nonnegative constraints on the factor matrices of NMTF, we constrain them to be cluster indicator matrices, a special type of nonnegative matrices. As a result, the clustering results of our approach are readily stored in the resulted factor matrices. However, existing NMF based methods require an extra post-processing step to extract cluster structures from the factor matrices, which often leads to non-unique clustering results.
- Due to the nature of indicator matrices, the optimization problems of our approach can be decoupled into subproblems with much smaller sizes, and the decoupled subproblems involve much less matrix multiplications. Therefore, our approach is computationally efficient and

scale well to large-scale input data.

- Taking into account the manifold structures in both data and feature spaces, we further develop a Locality Preserved FNMTF (LP-FNMTF) approach to incorporate manifold regularizations. Efficient algorithm to optimize the objective with quick convergence is presented.
- Promising experimental results on five benchmark data sets show that our approaches not only are faster than state-of-the-art co-clustering methods but also have competitive clustering performance.

Notations and problem formalization. Throughout this paper, we write matrices as boldface uppercase letters and vectors as boldface lowercase letters. Given a matrix $\mathbf{M} = (m_{ij})$, its i-th row and j-th column are denoted as $\mathbf{m}_{i\cdot}$ and $\mathbf{m}_{\cdot j}$ respectively.

Traditional clustering methods focus on one-side clustering, *i.e.*, clustering the data side based on the similarities along the feature side. In the co-clustering problem, we cluster data points based on the distributions of features, meanwhile cluster features based on the distributions of the data points. Formally, given a data set $\mathcal{X} = \left\{\mathbf{x}_{\cdot i} \in \mathbb{R}^d\right\}_{i=1}^n$, we write $\mathbf{X} = [\mathbf{x}_{\cdot 1}, \ldots, \mathbf{x}_{\cdot n}] = \left[\mathbf{x}_{1\cdot}^T, \ldots, \mathbf{x}_{d\cdot}^T\right]^T$. Our goal is to group the data points $\{\mathbf{x}_{\cdot 1}, \ldots, \mathbf{x}_{\cdot n}\}$ into c clusters $\{\mathcal{C}_j\}_{j=1}^c$, and simultaneously group the features $\{\mathbf{x}_{1\cdot}, \ldots, \mathbf{x}_{d\cdot}\}$ into m clusters $\{\mathcal{W}_j\}_{j=1}^m$.

We use a partition matrix $\mathbf{G} = \left[\mathbf{g}_{1\cdot}^T, \ldots, \mathbf{g}_{n\cdot}^T\right]^T \in \{0,1\}^{n\times c}$ to represent clustering result of data points, such that $g_{ij} = 1$ if $\mathbf{x}_{\cdot i}$ belongs to cluster $\mathcal{C}_j$ and $g_{ij} = 0$ otherwise. Similarly, we use another partition matrix $\mathbf{F} = \left[\mathbf{f}_{1\cdot}^T, \ldots, \mathbf{f}_{d\cdot}^T\right]^T \in \{0,1\}^{d\times m}$ to represent the clustering results of features. Here, we call $\mathbf{F}$ and $\mathbf{G}$ as cluster indicator matrices, because each row of them, *i.e.*, $\mathbf{f}_{i\cdot} (1 \le i \le d)$ or $\mathbf{g}_{i\cdot} (1 \le i \le n)$, has one and only one element equal to 1 to indicate the cluster membership, while the rest elements are 0. We denote the set of all cluster indicator matrices as Ψ.

2 Fast nonnegative matrix tri-factorization (FNMTF) for co-clustering

In this section, we first briefly review the background of co-clustering using NMTF, which motivates the optimization objective of our approach. After that, an efficient algorithm to solve our objective will be introduced.

2.1 Objective of FNMTF approach

K-means clustering is a standard clustering method in statistical learning, which minimizes the following objective:

$$J_1 = \sum_{j=1}^{c} \sum_{\mathbf{x}_{\cdot i} \in \mathcal{C}_j} \|\mathbf{x}_{\cdot i} - \mathbf{c}_{\cdot j}\|^2 = \sum_{j=1}^{c}\sum_{i=1}^{n} g_{ij} \|\mathbf{x}_{\cdot i} - \mathbf{c}_{\cdot j}\|^2 , \quad s.t. \quad \mathbf{G} \in \Psi^{n\times c}, \tag{1}$$

where $\|\cdot\|$ denotes the Frobenius norm of a matrix, $\mathbf{c}_{\cdot j}$ is the j-th centroid of the data set. Because $\mathbf{G} \in \Psi$ is a cluster indicator matrix, minimizing J_1 is a combinatorial optimization problem, which is hard to be resolved in general. Therefore, the minimization of J_1 is often relaxed to maximize the following objective [Zha *et al.*, 2001; Ding and He, 2004]:

$$J_2 = \mathbf{tr}\left(\mathbf{G}^T\mathbf{X}^T\mathbf{X}\mathbf{G}\right), \quad s.t. \quad \mathbf{G}^T\mathbf{G} = I \ . \tag{2}$$

Note that, $\mathbf{G}$ in J_2 is no longer an indicator matrix, but an arbitrary orthonormal matrix.

Recently, Ding *et al.* [Ding *et al.*, 2005] proved the equivalence between the relaxed objective of K-means clustering in Eq. (2) and the NMF objective when orthonormal constraints are enforced on the factor matrices, which minimizes:

$$\begin{aligned} J_3 &= \|\mathbf{X} - \mathbf{F}\mathbf{G}^T\|^2, \\ s.t. \quad &\mathbf{F} \ge 0, \mathbf{G} \ge 0, \mathbf{F}^T\mathbf{F} = I, \mathbf{G}^T\mathbf{G} = I, \end{aligned} \tag{3}$$

where $\mathbf{X} \in \mathbb{R}_+^{d\times n}$, $\mathbf{F} \in \mathbb{R}_+^{d\times c}$ and $\mathbf{G} \in \mathbb{R}_+^{n\times c}$, and J_3 aims to approximate the nonnegative data matrix $\mathbf{X}$ by the product of $\mathbf{F}$ and $\mathbf{G}$. The orthonormal constraints here ensure the uniqueness (up to a permutation) of the solution, and together with the nonnegative constraints make the resulted $\mathbf{F}$ and $\mathbf{G}$ approximate the K-means clustering results on both features and data points (called as "soft labels") [Ding *et al.*, 2005; 2006]. The latter, simultaneously clustering the rows (features) and the columns (data points) of an input data matrix, is one of the main strength of NMF defined in Eq. (3).

Because the two-factor NMF in Eq. (3) is restrictive, which often gives a rather poor low-rank matrix approximation, one more factor $\mathbf{S} \in \mathbb{R}_+^{m\times c}$ was introduced to absorb the different scales of $\mathbf{X}$, $\mathbf{F}$ and $\mathbf{G}$. This leads to NMTF [Ding *et al.*, 2006]: $\mathbf{X} \approx \mathbf{F}\mathbf{S}\mathbf{G}^T$, where $\mathbf{F} \in \mathbb{R}_+^{d\times m}$ and $\mathbf{G} \in \mathbb{R}_+^{n\times c}$. $\mathbf{S}$ provides increased degrees of freedom such that the low-rank matrix representation remains accurate, while $\mathbf{F}$ gives row clusters and $\mathbf{G}$ gives column clusters. In order to achieve additional flexibility, in clustering scenarios, the nonnegative constraint on $\mathbf{X}$ (thereby the nonnegative constraint on $\mathbf{S}$) can be relaxed [Ding *et al.*, 2010], which leads to the semi-NMTF problem minimizing the following objective:

$$\begin{aligned} J_4 &= \|\mathbf{X} - \mathbf{F}\mathbf{S}\mathbf{G}^T\|^2, \\ s.t. \quad &\mathbf{F} \ge 0, \mathbf{G} \ge 0, \ \mathbf{F}^T\mathbf{F} = I, \ \mathbf{G}^T\mathbf{G} = I \ . \end{aligned} \tag{4}$$

Despite its mathematical elegance, Eq. (4) suffers from two problems that impede its practical use. First, similar to Eq. (2), the relaxations on $\mathbf{F}$ and $\mathbf{G}$ make the immediate outputs of Eq. (4) are not cluster labels, which require an additional post-processing step and often lead to non-unique solutions. Second, and more important, Eq. (4) is usually solved by alternately iterative algorithms, and in each iteration step the intensive matrix multiplications are involved [Ding *et al.*, 2005; 2006; 2010; Wang *et al.*, 2008; Li *et al.*, 2010; Gu and Zhou, 2009]. As a result, it is infeasible to apply such algorithms to large-scale real world data due to the expensive computational cost.

In order to tackle these difficulties, instead of solving the relaxed clustering problems as in Eqs. (2–4), we solve the original clustering problem similar to Eq. (1). Specifically, we constrain the factor matrices of NMTF to be cluster indicator matrices and minimize the following objective:

$$J_5 = \|\mathbf{X} - \mathbf{F}\mathbf{S}\mathbf{G}^T\|^2 \quad s.t. \quad \mathbf{F} \in \Psi^{d\times m}, \ \mathbf{G} \in \Psi^{n\times c} \ . \tag{5}$$

We call Eq. (5) as the proposed Fast Nonnegative Matrix Tri-factorization (FNMTF) approach.

Note that, the orthonormal constraints on $\mathbf{F}$ and $\mathbf{G}$ are removed in our objective, as their purposes (unique solution and labeling approximation) are automatically accomplished by the new constraints. Surprisingly, with these new constraints, though more stringent, as shown theoretically shortly in this section and empirically later in Section 4, the computational speed of our approach can be significantly improved.

2.2 An efficient optimization algorithm

Following the standard optimization procedures, we alternately solve the three variables $\mathbf{F}$, $\mathbf{S}$ and $\mathbf{G}$ of J_5 in Eq. (5).

First, fixing $\mathbf{F}$, $\mathbf{G}$, and setting the derivative of J_5 with respect to $\mathbf{S}$ as zero, we have

$$\mathbf{S} = \left(\mathbf{F}^T\mathbf{F}\right)^{-1}\mathbf{F}^T\mathbf{X}\mathbf{G}\left(\mathbf{G}^T\mathbf{G}\right)^{-1} \ . \tag{6}$$

Second, when $\mathbf{F}$ and $\mathbf{S}$ are fixed, the optimization problem to obtain $\mathbf{G}$ can be decoupled and we solve the following simpler problem for each i $(1 \le i \le n)$:

$$\min_{\mathbf{G}\in\Psi} \|\mathbf{x}_{\cdot i} - \mathbf{F}\mathbf{S}\mathbf{g}_{i\cdot}^T\|^2 \ . \tag{7}$$

Because $\mathbf{g}_{i\cdot}$ $(1 \le i \le n) \in \Psi^{1\times c}$ is a cluster indicator vector in which one and only one element is 1 and the rest are zeros, the solution to Eq. (7) can be easily obtained by:

$$g_{ij} = \begin{cases} 1 & j = \arg\min_k \|\mathbf{x}_{\cdot i} - \tilde{\mathbf{f}}_{\cdot k}\|^2, \\ 0 & \text{otherwise}, \end{cases} \tag{8}$$

where $\tilde{\mathbf{F}} = \mathbf{F}\mathbf{S}$ and $\tilde{\mathbf{f}}_{\cdot k}$ is the k-th column of $\tilde{\mathbf{F}}$. Note that, Eq. (8) simply enumerates the c vector norms and seeks the maximum one, without involving any matrix multiplication.

Finally, when $\mathbf{G}$ and $\mathbf{S}$ are fixed, the optimization problem to obtain $\mathbf{F}$ can be similarly decoupled and we solve the following simpler problem for each j $(1 \le j \le d)$:

$$\min_{\mathbf{F}\in\Psi} \|\mathbf{x}_{j\cdot} - \mathbf{f}_{j\cdot}\mathbf{S}\mathbf{G}^T\|^2 \ . \tag{9}$$

Again, since $\mathbf{f}_{j\cdot}$ $(1 \le i \le d) \in \Psi^{1\times m}$ is a cluster indicator vector for feature side, the solution to Eq. (9) is:

$$f_{ij} = \begin{cases} 1 & i = \arg\min_l \|\mathbf{x}_{j\cdot} - \tilde{\mathbf{g}}_{l\cdot}\|^2, \\ 0 & \text{otherwise}, \end{cases} \tag{10}$$

where $\tilde{\mathbf{G}}^T = \mathbf{S}\mathbf{G}^T$ and $\tilde{\mathbf{g}}_{l\cdot}$ is the l-th row of $\tilde{\mathbf{G}}^T$.

The procedures to solve J_5 are summarized in Algorithm 1. Due to the nature of alternating optimization, Algorithm 1 is guaranteed to converge to a local minima (existing NMF algorithms [Ding *et al.*, 2005; 2006; 2010] also converges to a local minima because the objectives J_3 and J_4 are not convex in both variables $\mathbf{F}$ and $\mathbf{G}$), and the proof is skipped due to space limit. As can be seen, in step 2 and step 3, the solutions are obtained by enumerating the vector norms, which is definitely much faster than the matrix multiplication used in the existing NMF methods. Upon solution, $\mathbf{G}$ gives the clustering results of data points, and $\mathbf{F}$ gives the clustering results of features directly.

Algorithm 1: Algorithm to solve J_5 in Eq. (5).

```
Input: Data matrix $X = [\mathbf{x}_{\cdot 1}, \ldots, \mathbf{x}_{\cdot n}] \in \mathbb{R}^{d\times n}$.
Initialize $\mathbf{G} \in \Psi^{n\times c}$ and $\mathbf{F} \in \Psi^{d\times m}$ with arbitrary
class indicator matrices;
repeat
    1. calculate $\mathbf{S}$ by Eq. (6) ;
    2. calculate $\mathbf{G}$ by Eq. (8) ;
    3. calculate $\mathbf{F}$ by Eq. (10) ;
until converges;
Output: Indicator matrices $\mathbf{G}$ for data point
        clustering and $\mathbf{F}$ for feature clustering.
```

3 Locality preserved FNMTF (LP-FNMTF)

Recent researches showed that many real world data are sampled from the nonlinear manifolds which are embedded in the high dimensional ambient space [Belkin and Niyogi, 2002]. However, similar to traditional NMF and NMTF, the proposed FNMTF approach assumes that the data points and features are sampled from Euclidean spaces, and fails to discover the intrinsic geometrical and discriminative data and feature structures. Therefore, we further develop our FNMTF approach and propose the Locality Preserved FNMTF (LP-FNMTF) approach to enforce two geometrically based regularizers from both data and feature sides.

3.1 Manifold regularization and the optimization objective of LP-FNMTF method

Because we co-cluster an input data matrix on both data and feature dimensions, we consider two undirected graphs, one constructed from data points, denoted as $\mathcal{G}_d$, and the other one from features, denoted as $\mathcal{G}_f$. The corresponding affinity matrices $\mathbf{W}_d$ and $\mathbf{W}_f$ could be either computed from the input data matrix $\mathbf{X}$ (*e.g.*, as in [Gu and Zhou, 2009]), or obtained from prior knowledge. According to manifold assumption [Belkin and Niyogi, 2002], the regularization terms to measure the smoothness with respect to the intrinsic manifolds of data points and features are given by [Cai *et al.*, 2008; Gu and Zhou, 2009]:

$$\min_{\mathbf{G}\in\Psi} \mathbf{tr}\left(\mathbf{G}^T\mathbf{L}_d\mathbf{G}\right), \quad \text{and} \quad \min_{\mathbf{F}\in\Psi} \mathbf{tr}\left(\mathbf{F}^T\mathbf{L}_f\mathbf{F}\right), \tag{11}$$

where $\mathbf{L}_d = \mathbf{I} - \mathbf{D}_d^{-\frac{1}{2}}\mathbf{W}_d\mathbf{D}_d^{-\frac{1}{2}}$ is the the normalized graph Laplacian of $\mathcal{G}_d$, and $\mathbf{D}_d$ is the degree matrix of $\mathcal{G}_d$; similarly, $\mathbf{L}_f = \mathbf{I} - \mathbf{D}_f^{-\frac{1}{2}}\mathbf{W}_f\mathbf{D}_f^{-\frac{1}{2}}$, and $\mathbf{D}_f$ is the degree matrix of $\mathcal{G}_f$.

Incorporating Eq. (11) into Eq. (5), the objective of LP-FNMTF approach is to minimize:

$$\begin{aligned} J_6 = \|\mathbf{X} - \mathbf{F}\mathbf{S}\mathbf{G}^T\|^2 + \alpha\,\mathbf{tr}\left(\mathbf{G}^T\mathbf{L}_d\mathbf{G}\right) + \beta\,\mathbf{tr}\left(\mathbf{F}^T\mathbf{L}_f\mathbf{F}\right), \\ s.t. \quad \mathbf{F} \in \Psi^{d\times m},\ \mathbf{G} \in \Psi^{n\times c}, \end{aligned} \tag{12}$$

where α and β are regularization parameters to balance the reconstruction error of co-clustering in the first term and labeling smoothness in the data point space and feature space in the second and third terms, respectively.

Because $\mathbf{F}$ and $\mathbf{G}$ are constrained to be cluster indicator matrices, it is difficult to solve Eq. (12) in general. Hence we simplify this problem using the following proposition.

Proposition 1 *Given a symmetric matrix* $\mathbf{A}$ *and its eigen-decomposition* $\mathbf{A}=\mathbf{P}\mathbf{\Sigma}\mathbf{P}^T$*, where* $\mathbf{\Sigma}\in\mathbb{R}^{c\times c}$ *is a diagonal matrix with diagonal elements as the* c *largest eigenvalues, and* $\mathbf{P}$ *is the corresponding eigenvector matrix, the following two optimization problems are equivalent:*

$$(P1):\quad \min_{\mathbf{C}\in\Psi}\textit{tr}\left[\mathbf{C}^T\left(\mathbf{I}-\mathbf{A}\right)\mathbf{C}\right], \tag{13}$$

$$(P2):\quad \min_{\mathbf{C}\in\Psi,\ \mathbf{Q}^T\mathbf{Q}=I}\|\mathbf{C}-\mathbf{B}\mathbf{Q}\|^2, \tag{14}$$

where $\mathbf{Q}$ *is an arbitrary orthonormal matrix and*

$$\mathbf{B}=\mathbf{P}\mathbf{\Sigma}^{1/2}\ . \tag{15}$$

Proof. $(P1)$ is equivalent to $\max_{\mathbf{C}\in\Psi}\mathbf{tr}\left(\mathbf{C}^T\mathbf{A}\mathbf{C}\right)$ that is further equivalent to $\min_{\mathbf{C}\in\Psi}\|\mathbf{C}\mathbf{C}^T-\mathbf{A}\|^2$. By definition the low-rank approximation of $\mathbf{A}$ is given by $\mathbf{A}=\mathbf{B}\mathbf{Q}\left(\mathbf{B}\mathbf{Q}\right)^T$, thus $(P1)$ becomes $\min_{\mathbf{C}\in\Psi,\mathbf{Q}^T\mathbf{Q}=I}\|\mathbf{C}\mathbf{C}^T-\mathbf{B}\mathbf{Q}\left(\mathbf{B}\mathbf{Q}\right)^T\|^2$. $\mathbf{C}$ approximating $\mathbf{B}\mathbf{Q}$ is equivalent to $\mathbf{C}\mathbf{C}^T$ approximating $\mathbf{B}\mathbf{Q}\left(\mathbf{B}\mathbf{Q}\right)^T$. Hence, solving $(P1)$ in Eq. (13) can be reasonably transformed to solve $(P2)$ in Eq. (14), which completes the proof of Proposition 1. ■

Applying Proposition 1 in Eq. (12), the objective of our LP-FNMTF approach is transformed to minimize:

$$J_7=\|\mathbf{X}-\mathbf{F}\mathbf{S}\mathbf{G}^T\|^2+\alpha\|\mathbf{G}-\mathbf{B}_d\mathbf{Q}_d\|^2+\beta\|\mathbf{F}-\mathbf{B}_f\mathbf{Q}_f\|^2,$$
$$s.t.\ \mathbf{F}\in\Psi^{d\times m},\mathbf{G}\in\Psi^{n\times c},\mathbf{Q}_d^T\mathbf{Q}_d=I,\mathbf{Q}_f^T\mathbf{Q}_f=I, \tag{16}$$

where $\mathbf{B}_d$ and $\mathbf{B}_f$ are computed from $\mathbf{L}_d$ and $\mathbf{L}_f$ following the procedures described in Proposition 1.

3.2 Optimization algorithm

Again, we use the alternating iterative method to solve Eq. (16). We first introduce the following theorem.

Theorem 1 *Let* $\mathbf{H}=\mathbf{B}^T\mathbf{C}$ *and the Singular Value Decomposition (SVD) of* $\mathbf{H}$ *be given by* $\mathbf{H}=\mathbf{U}\mathbf{\Lambda}\mathbf{V}^T$*. Fixing* $\mathbf{C}$ *and* $\mathbf{B}$*, the optimum* $\mathbf{Q}$ *to problem* $(P2)$ *defined in Eq.* (14) *is given by* $\mathbf{Q}=\mathbf{U}\mathbf{V}^T$*.*

Proof. When $\mathbf{C}$ is fixed, problem $(P2)$ in Eq. (14) is equivalent to $\max_{\mathbf{Q}^T\mathbf{Q}=I}\mathbf{tr}\left(\mathbf{Q}^T\mathbf{H}\right)$.

We have $\mathbf{tr}\left(\mathbf{Q}^T\mathbf{H}\right)=\mathbf{tr}\left(\mathbf{Q}\mathbf{U}\mathbf{\Lambda}\mathbf{V}^T\right)=\mathbf{tr}\left(\mathbf{\Lambda}\mathbf{V}^T\mathbf{Q}\mathbf{U}\right)=\mathbf{tr}\left(\mathbf{\Lambda}\mathbf{Z}\right)=\sum_i\lambda_{ii}z_{ii}$, where $\mathbf{Z}=\mathbf{V}^T\mathbf{Q}\mathbf{U}$ and λ_{ii} and z_{ii} are the (i,i)-th entry of $\mathbf{\Lambda}$ and $\mathbf{Z}$, respectively.

Note that $\mathbf{Z}$ is orthonormal, *i.e.*, $\mathbf{Z}^T\mathbf{Z}=I$, thus $z_{ii}\le 1$. On the other hand, $\lambda_{ii}\ge 0$ as λ_{ii} is singular value of $\mathbf{H}$. Therefore, $\mathbf{tr}\left(\mathbf{Q}^T\mathbf{H}\right)=\sum_i\lambda_{ii}z_{ii}\le\sum_i\lambda_{ii}$, and when $z_{ii}=1\,(1\le i\le c)$, the equality holds. That is to say, $\mathbf{tr}\left(\mathbf{Q}^T\mathbf{H}\right)$ reaches its maximum when $\mathbf{Z}=I$. Recall that $\mathbf{Z}=\mathbf{V}^T\mathbf{Q}\mathbf{U}$, the solution to $\max_{\mathbf{Q}^T\mathbf{Q}=I}\mathbf{tr}\left(\mathbf{Q}^T\mathbf{H}\right)$ or $(P2)$ is $\mathbf{Q}=\mathbf{U}\mathbf{Z}^T\mathbf{V}^T=\mathbf{U}\mathbf{V}^T$. Theorem 1 is proved. ■

Now, we solve Eq. (16). First, fixing $\mathbf{F}$ and $\mathbf{G}$, by setting the derivative of J_7 with respect to $\mathbf{S}$ as 0, we obtain

$$\mathbf{S}=\left(\mathbf{F}^T\mathbf{F}\right)^{-1}\mathbf{F}^T\mathbf{X}\mathbf{G}\left(\mathbf{G}^T\mathbf{G}\right)^{-1}\ . \tag{17}$$

Second, fixing $\mathbf{F}$, $\mathbf{G}$ and $\mathbf{S}$, we can decouple Eq. (16) into two following subproblems:

$$\min_{\mathbf{Q}_d^T\mathbf{Q}_d=I}\|\mathbf{G}-\mathbf{B}_d\mathbf{Q}_d\|^2,\quad \min_{\mathbf{Q}_f^T\mathbf{Q}_f=I}\|\mathbf{F}-\mathbf{B}_f\mathbf{Q}_f\|^2\ . \tag{18}$$

Applying Theorem 1, $\mathbf{Q}_d=\mathbf{U}_d\mathbf{V}_d^T$ where $\mathbf{U}_d$ and $\mathbf{V}_d$ are obtained by SVD on $\mathbf{B}_d^T\mathbf{G}$; $\mathbf{Q}_f=\mathbf{U}_f\mathbf{V}_f^T$ where $\mathbf{U}_f$ and $\mathbf{V}_f$ are obtained by SVD on $\mathbf{B}_f^T\mathbf{F}$.

Third, we fix $\mathbf{S}$, $\mathbf{F}$ and $\mathbf{Q}_d$ to update $\mathbf{G}$. Because $\mathbf{G}$ is a cluster indicator matrix, Eq. (16) is decoupled to the following simpler problems for each $1\le i\le n$:

$$\min_{\mathbf{G}\in\Psi}\|\mathbf{x}_{\cdot i}-\tilde{\mathbf{F}}\mathbf{g}_{i\cdot}^T\|^2+\alpha\|\mathbf{g}_{i\cdot}-\left(\tilde{\mathbf{b}}_d\right)_{i\cdot}\|^2, \tag{19}$$

where $\tilde{\mathbf{F}}=\mathbf{F}\mathbf{S}$, $\left(\tilde{\mathbf{b}}_d\right)_{i\cdot}$ denotes the i-th row of $\tilde{\mathbf{B}}_d=\mathbf{B}_d\mathbf{Q}_d$. Thus, the solution can be obtained by

$$g_{ij}=\begin{cases}1 & j=\arg\min_k\left(\|\mathbf{x}_{\cdot i}-\tilde{\mathbf{f}}_{\cdot k}\|^2-2\alpha\left(\tilde{\mathbf{B}}_d\right)_{ik}\right),\\ 0 & \text{otherwise}\ .\end{cases} \tag{20}$$

Finally, when fixing $\mathbf{S}$, $\mathbf{G}$ and $\mathbf{Q}_f$, let $\tilde{\mathbf{B}}_f=\mathbf{B}_f\mathbf{Q}_f$, $\tilde{\mathbf{G}}^T=\mathbf{S}\mathbf{G}^T$ and $\tilde{\mathbf{g}}_{l\cdot}$ is the l-th row of $\tilde{\mathbf{G}}^T$, we similarly obtain $\mathbf{F}$ as:

$$f_{ij}=\begin{cases}1 & i=\arg\min_l\left(\|\mathbf{x}_{j\cdot}-\tilde{\mathbf{g}}_{l\cdot}\|^2-2\beta\left(\tilde{\mathbf{B}}_f\right)_{jl}\right),\\ 0 & \text{otherwise.}\end{cases} \tag{21}$$

The procedures to solve J_7 are summarized in Algorithm 2. Again, because step 4 and step 5 only involve vector norm enumeration without matrix multiplication, our algorithm is more computationally efficient. Empirical results show that the convergence of our algorithm is fast, which make it feasible to solve the large-scale real world problems via our approach in practice.

Algorithm 2: Algorithm to solve J_7 in Eq. (16).

```
Input: Data matrix $X = [\mathbf{x}_{\cdot 1}, \ldots, \mathbf{x}_{\cdot n}] \in \mathbb{R}^{d\times n}$.
1. Initialize $\mathbf{G} \in \Psi^{n\times c}$ and $\mathbf{F} \in \Psi^{d\times m}$ with
arbitrary class indicator matrices;
2. Calculate $\mathbf{B}_d$ and $\mathbf{B}_f$ from $\mathbf{L}_d$ and $\mathbf{L}_f$ following
the description of Proposition 1;
repeat
    1. Calculate $\mathbf{S}$ by Eq. (17) ;
    2. Calculate $\mathbf{Q}_d = \mathbf{U}_d\mathbf{V}_d^T$ where $\mathbf{U}_d$ and $\mathbf{V}_d$ are
    obtained by SVD on $\mathbf{B}_d^T\mathbf{G}$ ;
    3. Calculate $\mathbf{Q}_f = \mathbf{U}_f\mathbf{V}_f^T$ where $\mathbf{U}_f$ and $\mathbf{V}_f$ are
    obtained by SVD on $\mathbf{B}_f^T\mathbf{F}$ ;
    4. Calculate $\mathbf{G}$ by Eq. (20) ;
    5. Calculate $\mathbf{F}$ by Eq. (21) ;
until Converges;
Output: Class indicator matrices $\mathbf{G}$ and $\mathbf{F}$ for data
        and feature clustering tasks, respectively.
```

4 Experiments

In this section, we evaluate the proposed FNMTF and LP-FNMTF approaches, and compare them against state-of-the-art (co-)clustering methods, including Semi-NMF (SNMF) [Ding *et al.*, 2010], Orthogonal NMTF (ONMTF) [Ding *et al.*, 2006], Graph regularized NMF (GNMF) [Cai *et al.*, 2008]

Table 1: Description of experimental data sets

Data sets	# sample	# feature	# classes
Coil20	1140	1024	20
WebKB	4199	1000	4
WebACE	2340	1000	20
CSTR	476	1000	4
RCV1	193844	1979	103

and Dual Regularized Co-clustering (DRCC) [Gu and Zhou, 2009] methods. We also report the clustering results by K-means and NMF [Lee and Seung, 2001] methods as baselines.

In our experiments, we use 5 data sets to evaluate the compared methods, which are summarized in Table 1. The first four are widely used as benchmarks in clustering literatures [Ding *et al.*, 2006; Cai *et al.*, 2008; Gu and Zhou, 2009; Ding *et al.*, 2010]. The last one has very large sample size and feature size [Chen *et al.*, 2010]. In order to run the experiments on contemporary computers, for RCV1 data set, following previous studies, we remove the keywords (features) appearing less than 100 times in the corpus, which results in 1979 (out of 47236) keywords in our experiments.

To evaluate the clustering results, we adopt three widely used standard metrics: clustering accuracy [Cai *et al.*, 2008], normalized mutual information (NMI) [Cai *et al.*, 2008] and cluster purity [Ding *et al.*, 2006].

4.1 Clustering results

Experiments setup. Because each clustering algorithm has one or more parameters to be tuned, in order to compare fairly, we run these algorithms under different parameter settings, and select the best average result for each one. We set the number of clusters as the true number of classes for all clustering algorithms on all data sets.

For co-clustering methods, including ONMTF, DRCC and our two methods, the number of feature clusters is set to be the same as that of data clusters, *i.e.*, $m = c$.

For manifold regularized methods, including GNMF, DRCC and our LP-FNMTF methods, we construct nearest-neighbor graph following [Gu and Zhou, 2009], where the neighborhood size for graph construction is set by searching the grid of $\{1, 2, \ldots, 10\}$, and the regularization parameters (*i.e.*, α and β in Eq. (16)) are set by searching the grid of $\{0.1, 1, 10, 100, 500, 1000\}$.

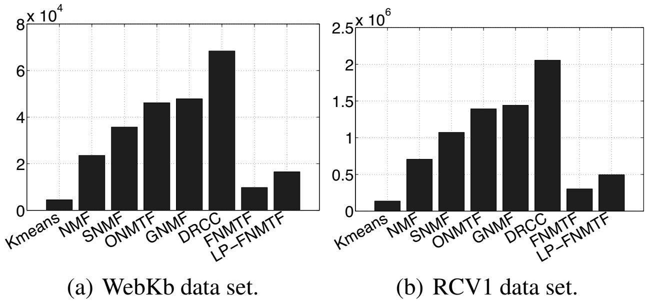
Given the number of clusters, no parameter selection is needed for K-means, NMF and SNMF methods.

Results. Under each parameter setting of every method in comparison, we repeat clustering 50 times, and the average result is computed. We report the best average result for each method on five data sets in Table 2.

From the results in Table 2, we can see that the proposed FNMTF and LP-FNMTF methods consistently outperform the other compared methods, sometimes very significantly, which demonstrate the advantage of our approaches in terms of clustering performance. A more careful examination on the results shows that, the co-clustering methods, including ONMTF, DDRC and our proposed FNMTF and LP-FNMTF methods, generally achieve better clustering results, which are consistent with the widely accepted hypothesis that clustering of features can help clustering of data points. Finally, LP-FNMTF method is superior to FNMTF method on all the data sets except CSTR. This indicates that exploiting the geometric structures in data and feature spaces indeed can improve the cluster performance, which verifies manifold assumption and confirms the correctness of our algorithms.

Table 3: Average iteration numbers to converge of compared clustering methods.

Data	Coil20	WebKB	WebACE	CSTR	RCV1
Kmeans	26.4	29.2	28.3	24.2	87.1
NMF	30.3	40.2	38.5	30.1	92.5
SNMF	37.7	45.5	44.1	36.3	98.6
ONMTF	41.2	50.3	49.2	40.3	104.4
GNMF	50.2	62.1	61.1	46.8	112.5
DRCC	51.6	64.4	62.3	47.9	129.1
FNMTF	14.1	16.5	15.9	14.3	45.2
LP-FNMTF	15.2	17.2	16.3	15.6	48.1



(a) WebKb data set.

(b) RCV1 data set.

Figure 1: Convergence time (ms) of compared clustering methods on WebKB and RCV1 data sets.

4.2 Studies of computational speeds

In this subsection, we evaluate the computational speeds of the compared (co-)clustering methods. All our experiments are performed on a Dell PowerEdge 2900 server, which has two quad-core Intel Xeon 5300 sequence CPU processors at 3.0 GHz and 48G bytes memory.

We first examine the convergence rates of the compared methods. We repeat clustering 50 times by each method with its optimal parameters on each data set. The average iteration numbers of each method on each data set are reported in Table 3. The results show that the proposed FNMTF and LP-FNMTF approaches require much less iterations to converge, therefore they are more computationally efficient.

In addition, we also report the average convergence time of the compared methods on WebKB data and RCV1 data sets as in Figure 1. From the results, we can see that our FNMTF and LP-FNMTF methods are only slower than K-means method while much faster than all other state-of-the-art clustering methods. These results are consistent with our theoretical analysis that our methods are implemented on subproblems with much smaller sizes and use much less matrix multiplications. Therefore, our approaches are suitable for clustering on large-scale data. The results on three other smaller data sets are not shown due to space limit, from which the same observations can be seen.

Table 2: Clustering results measured by accuracy/NMI/purity of the compared methods.

Data	Metrics	Kmeans	NMF	SNMF	ONMTF	GNMF	DRCC	FNMTF	LP-FNMTF
Coil20	Accuracy	0.495	0.487	0.527	0.635	0.665	0.680	0.696	**0.725**
	NMI	0.489	0.479	0.511	0.561	0.549	0.566	0.584	**0.621**
	Purity	0.441	0.437	0.468	0.478	0.481	0.483	0.497	**0.512**
WebKB	Accuracy	0.698	0.668	0.621	0.685	0.717	0.725	0.774	**0.792**
	NMI	0.467	0.427	0.418	0.455	0.458	0.487	0.501	**0.522**
	Purity	0.601	0.595	0.604	0.664	0.672	0.674	0.696	**0.712**
WebACE	Accuracy	0.526	0.514	0.527	0.635	0.665	0.680	0.696	**0.714**
	NMI	0.519	0.512	0.538	0.587	0.556	0.571	0.604	**0.611**
	Purity	0.479	0.481	0.491	0.487	0.511	0.493	0.517	**0.532**
CSRT	Accuracy	0.763	0.759	0.699	0.771	0.742	0.812	**0.894**	0.847
	NMI	0.654	0.668	0.614	0.673	0.635	0.681	**0.753**	0.722
	Purity	0.612	0.587	0.614	0.645	0.637	0.656	**0.701**	0.682
RCV1	Accuracy	0.168	0.156	0.173	0.196	0.201	0.213	0.238	**0.241**
	NMI	0.274	0.274	0.283	0.301	0.312	0.318	0.339	**0.342**
	Purity	0.134	0.121	0.139	0.146	0.152	0.159	0.172	**0.179**

5 Conclusions

In this work, we proposed a novel Fast Nonnegative Matrix Tri-factorization (FNMTF) method to simultaneously cluster both data side and feature side of an input data matrix. We adopt the idea of NMF/NMTF based co-clustering methods, but constrain the factor matrices to be cluster indicator matrices, a special type of nonnegative matrices. Through the new constraints, the optimization problem of our method is decoupled into a number of much smaller subproblems that require much less matrix multiplication than the existing NMF based co-clustering algorithms, which makes our approaches of particular use for real world large-scale data. We further developed our method to incorporate manifold information and proposed Locality Preserved FNMTF (LP-FNMTF) method. We conducted extensive experiments on benchmark data sets that demonstrate promising results of our methods, which is consistent with our theoretical analysis.

Acknowledgments

This research was supported by NSF-CNS 0923494, NSF-IIS 1041637, NSF-CNS 1035913.

References

[Belkin and Niyogi, 2002] M. Belkin and P. Niyogi. Laplacian eigenmaps and spectral techniques for embedding and clustering. In *NIPS*, 2002.

[Cai *et al.*, 2008] D. Cai, X. He, X. Wu, and J. Han. Nonnegative matrix factorization on manifold. In *ICDM*, 2008.

[Chen *et al.*, 2010] W.Y. Chen, Y. Song, H. Bai, C.J. Lin, and E.Y. Chang. Parallel spectral clustering in distributed systems. *IEEE TPAMI*, 2010.

[Cho *et al.*, 2004] H. Cho, I.S. Dhillon, Y. Guan, and S. Sra. Minimum sum-squared residue co-clustering of gene expression data. In *SDM*, 2004.

[Dhillon, 2001] I.S. Dhillon. Co-clustering documents and words using bipartite spectral graph partitioning. In *SIGKDD*, 2001.

[Ding and He, 2004] C. Ding and X. He. K-means clustering via principal component analysis. In *ICML*, 2004.

[Ding *et al.*, 2005] C. Ding, X. He, and H.D. Simon. On the equivalence of nonnegative matrix factorization and spectral clustering. In *SDM*, 2005.

[Ding *et al.*, 2006] C. Ding, T. Li, W. Peng, and H. Park. Orthogonal nonnegative matrix tri-factorizations for clustering. In *SIGKDD*, 2006.

[Ding *et al.*, 2010] C. Ding, T. Li, and M.I. Jordan. Convex and semi-nonnegative matrix factorizations. *IEEE TPAMI*, 32(1):45–55, 2010.

[Gu and Zhou, 2009] Q. Gu and J. Zhou. Co-clustering on manifolds. In *SIGKDD*, 2009.

[Lee and Seung, 1999] D.D. Lee and H.S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999.

[Lee and Seung, 2001] D.D. Lee and H.S. Seung. Algorithms for non-negative matrix factorization. In *NIPS*, 2001.

[Li *et al.*, 2010] T. Li, V. Sindhwani, C. Ding, and Y. Zhang. Bridging Domains with Words: Opinion Analysis with Matrix Tri-factorizations. In *SDM*, 2010.

[Long *et al.*, 2006] B. Long, X. Wu, Z.M. Zhang, and P.S. Yu. Unsupervised learning on k-partite graphs. In *SIGKDD*, 2006.

[Nie *et al.*, 2009] Feiping Nie, Dong Xu, Ivor W. Tsang, and Changshui Zhang. Spectral embedded clustering. In *IJCAI*, pages 1181–1186, 2009.

[Wang *et al.*, 2008] F. Wang, T. Li, and C. Zhang. Semi-supervised clustering via matrix factorization. In *SDM*, 2008.

[Zha *et al.*, 2001] Hongyuan Zha, Xiaofeng He, Chris Ding, Horst Simon, and Ming Gu. Spectral relaxation for k-means clustering. In *NIPS*, 2001.