On Trivial Solution and Scale Transfer Problems in Graph Regularized NMF

Quanquan Gu[1], **Chris Ding**[2] **and Jiawei Han**[1]
[1]Department of Computer Science, University of Illinois at Urbana-Champaign
[2]Department of Computer Science and Engineering, University of Texas at Arlington
{qgu3,hanj}@illinois.edu, chqding@uta.edu

Abstract

Combining graph regularization with nonnegative matrix (tri-)factorization (NMF) has shown great performance improvement compared with traditional nonnegative matrix (tri-)factorization models due to its ability to utilize the geometric structure of the documents and words. In this paper, we show that these models are not well-defined and suffering from trivial solution and scale transfer problems. In order to solve these common problems, we propose two models for graph regularized nonnegative matrix (tri-)factorization, which can be applied for document clustering and co-clustering respectively. In the proposed models, a Normalized Cut-like constraint is imposed on the cluster assignment matrix to make the optimization problem well-defined. We derive a multiplicative updating algorithm for the proposed models, and prove its convergence. Experiments of clustering and co-clustering on benchmark text data sets demonstrate that the proposed models outperform the original models as well as many other state-of-the-art clustering methods.

1 Introduction

Nonnegative Matrix Factorization (NMF) [Lee and Seung, 2000] [Li and Ding, 2006] [Xu *et al.*, 2003] [Ding *et al.*, 2006] has received increasing interest in the past decade and achieved great success in document clustering and document-word co-clustering [Dhillon *et al.*, 2003] (clustering documents based on their distribution on words, while grouping words based on their distribution on the documents).

Recently, [Cai *et al.*, 2008] proposed a graph regularized nonnegative matrix factorization (GNMF), which incorporates the geometric structure of the documents into NMF by manifold regularization [Belkin *et al.*, 2006]. [Gu and Zhou, 2009] proposed a Dual Regularized Co-Clustering (DRCC) method based on graph regularized semi-nonnegative matrix tri-factorization [Ding *et al.*, 2010]. They showed that not only the documents but also the words are discrete samplings from some manifolds. By manifold regularization [Belkin *et al.*, 2006], the cluster labels of documents are smooth with respect to the intrinsic document manifold, while the cluster labels of words are smooth with respect to the intrinsic word manifold.

However, both GNMF [Cai *et al.*, 2008] and DRCC [Gu and Zhou, 2009] suffer from the scale transfer problem because the regularization term in the objective function is not lower-bounded. More seriously, when the regularization parameter is too large that the regularization term dominates the objective function, the original optimization problem will degenerate into a series of independent subproblems associated with each column of the cluster assignment matrix. In this case, the solution of each row of the cluster assignment matrix will be very similar with each other, resulting in trivial clustering result. We will analyze these two problems in detail in the next section. As far as we know, there is no existing principled way to deal with such kinds of problems. In this paper, to solve the problems in a principled way, we propose two models for graph regularized nonnegative matrix (tri-)factorization, which can be respectively applied to document clustering and document-word co-clustering. We impose a Normalized Cut-type [Shi and Malik, 2000] constraint on the cluster assignment matrix, which makes the optimization problem well-defined. We also derive an iterative multiplicative updating algorithm to solve the proposed models. The convergence of the algorithm is theoretically guaranteed. The main contributions of this paper include: (1) we analyze the common problems existing in many previous graph regularized nonnegative matrix (tri-)factorization models, and (2) we propose a principled way to solve these problems. Experiments of clustering and co-clustering on many benchmark data sets demonstrate that the proposed models overcome the problems and outperform the original models as well as many state-of-the-art clustering methods.

1.1 Notations

Given a document set $\mathcal{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\} \in \mathbb{R}^d$, the goal of clustering is to group the documents into c clusters $\{\mathcal{C}_j\}_{j=1}^c$. And the goal of co-clustering is to group the documents into c clusters $\{\mathcal{C}_j\}_{j=1}^c$, while grouping the words into m clusters $\{\mathcal{W}_j\}_{j=1}^m$. We use a cluster assignment matrix $\mathbf{V} \in \mathbb{R}_+^{n \times c}$ to represent the clustering result of documents, such that if $\arg\max_j V_{ij} = j^*$, then $\mathbf{x}_i$ belongs to cluster $\mathcal{C}_{j^*}$. We denote the ith row of $\mathbf{V}$ by $\mathbf{v}^i$, and the jth column of $\mathbf{V}$ by $\mathbf{v}_j$. Each element in $\mathbf{v}^i$ can be seen as the probability of the ith document belonging to each clusters, which provides *Soft*

Clustering. In the co-clustering scenario, we introduce another clustering assignment matrix $\mathbf{U} \in \mathbb{R}_+^{d\times m}$ to represent the clustering result of words analogously.

2 Related Work

In this section, we will review several methods related to ours, and analyze the potential problems that previous works suffer from.

[Cai *et al.*, 2008] proposed a graph regularized NMF (GNMF), which adds an additional graph regularizer on NMF, imposing *Manifold Assumption* on the data points. That is, if two documents $\mathbf{x}_i$ and $\mathbf{x}_j$ are close to each other, then their cluster labels $\mathbf{v}^i$ and $\mathbf{v}^j$ should be close as well. This is formulated as follows,

$$\frac{1}{2}\sum_{i,j}||\mathbf{v}^i-\mathbf{v}^j||^2(W_V)_{ij}=\text{tr}(\mathbf{V}^T\mathbf{L}_V\mathbf{V}), \tag{1}$$

where $\mathbf{L}_V=\mathbf{D}_V-\mathbf{W}_V$ is the graph Laplacian on the document graph, $\mathbf{W}_V$ is the adjacency matrix defined on the document graph, $\mathbf{D}_V$ is a diagonal degree matrix of the document graph with $(D_V)_{ii}=\sum_{j=1}^n(W_V)_{ij}$. For example, we can define the adjacency matrix $\mathbf{W}_V$ as

$$(W_V)_{ij}=\begin{cases}1, & \text{if } \mathbf{x}_j\in\mathcal{N}(\mathbf{x}_i) \text{ or } \mathbf{x}_i\in\mathcal{N}(\mathbf{x}_j)\\ 0, & \text{otherwise.}\end{cases} \tag{2}$$

where $\mathcal{N}(\mathbf{x}_i)$ denotes the k-nearest neighbor of $\mathbf{x}_i$. And GNMF minimizes the following objective,

$$\begin{aligned} J &= ||\mathbf{X}-\mathbf{U}\mathbf{V}^T||_F^2+\mu\text{tr}(\mathbf{V}^T\mathbf{L}_V\mathbf{V}),\\ &\text{s.t.} \quad \mathbf{U}\geq 0,\mathbf{V}\geq 0, \end{aligned} \tag{3}$$

where $\mu>0$ is a regularization parameter. Due to the graph regularization term $\text{tr}(\mathbf{V}^T\mathbf{L}_V\mathbf{V})$, GNMF can take into account the geometric information of the data.

[Gu and Zhou, 2009] proposed a dual regularized co-clustering (DRCC) method based on graph regularized (semi-)nonnegative matrix factorization [Ding *et al.*, 2010], which imposes graph regularization on both the word and document cluster assignment matrices, i.e.,

$$\begin{aligned} J_1 &= ||\mathbf{X}-\mathbf{U}\mathbf{S}\mathbf{V}^T||_F^2+\lambda\text{tr}(\mathbf{U}^T\mathbf{L}_U\mathbf{U})+\mu\text{tr}(\mathbf{V}^T\mathbf{L}_V\mathbf{V})\\ &\text{s.t.} \quad \mathbf{U}\geq 0,\mathbf{S}\geq 0,\mathbf{V}\geq 0, \end{aligned} \tag{4}$$

where $\mathbf{S}$ reflects the association between words and documents, $\mathbf{L}_U=\mathbf{D}_U-\mathbf{W}_U$ is the graph Laplacian on the word graph, $\mathbf{W}_U$ is the adjacency matrix of the word graph which can be defined similarly as in Eq. (2), $\mathbf{D}_U$ is the diagonal degree matrix associated with the word graph, with $(D_U)_{ii}=\sum_{j=1}^d(W_U)_{ij}$, the definition of $\mathbf{L}_V$ is the same as above, $\lambda,\mu>0$ are regularization parameters. Hereafter, we refer to the model in Eq. (4) as GNMTF for consistency with GNMF. GNMTF not only considers the geometric structure of the documents as in GNMF, but also takes into account the geometric information of the words.

2.1 Trivial Solution Problem

In this subsection, we show that GNMF [Cai *et al.*, 2008] suffers from trivial solution. When μ approaches ∞, the second term dominates the objective function, Eq. (3) boils down to

$$\begin{aligned} J' &= \text{tr}(\mathbf{V}^T\mathbf{L}_V\mathbf{V})=\sum_{k=1}^c\mathbf{v}_k^T\mathbf{L}_V\mathbf{v}_k,\\ &\text{s.t.} \quad \mathbf{V}\geq 0, \end{aligned} \tag{5}$$

where $\mathbf{v}_k$ denotes the k-th column of $\mathbf{V}$. Eq. (5) can be decomposed into c independent optimization subproblems as follows

$$\begin{aligned} J'' &= \mathbf{v}_k^T\mathbf{L}_V\mathbf{v}_k,\\ &\text{s.t.} \quad \mathbf{v}_k\geq 0, \end{aligned} \tag{6}$$

where $k=1,\ldots,c$. Thus each optimization gets the same solution up to a scale, i.e., $\mathbf{v}_1\propto\ldots\propto\mathbf{v}_c$. This means all elements in each row of $\mathbf{V}$ are identical up to a scale. Hence the cluster assignment by $\mathbf{V}$ tends to assign all the documents to one class, because the cluster assignment is determined by the largest entry in each row. In reality, due to the existence of the first NMF term in Eq. (3), these elements are not precisely identical up to a scale, but rather slightly fluctuate. Figure 1 shows an illustrative toy example. It can be seen that when $\mu=1$, the clustering results is correct. However, when $\mu=10^4$, the clustering results assign all the documents to cluster 1. Since there is a fluctuation caused by the first NMF term, the scale between column 1 and column 2 of $\mathbf{V}$ is not uniformly the same across the rows of $\mathbf{V}$. And when $\mu=10^6$, the problem gets worse. The scale to which column 1 equals to column 2 is uniformly $\frac{0.73}{0.27}$.

The same problem exists in GNMTF [Gu and Zhou, 2009]. This motivates us to propose the new formulation presented in this paper.

0.19	0.39	0.81	2.80	2.40	3.00	2.62
0.55	0.41	0.90	2.22	2.39	2.34	2.53
0.48	0.96	0.48	2.15	2.56	2.34	2.26
1.51	1.86	1.89	0.60	0.14	0.52	0.35
1.55	1.89	1.64	0.70	0.25	0.60	0.87

(a)

0.06	0.94
0.07	0.93
0.14	0.86
0.86	0.14
0.93	0.07
0.87	0.13
0.87	0.13

(b)

0.62	0.38
0.62	0.38
0.62	0.38
0.66	0.34
0.66	0.34
0.66	0.34
0.66	0.34

(c)

0.73	0.27
0.73	0.27
0.73	0.27
0.73	0.27
0.73	0.27
0.73	0.27
0.73	0.27

(d)

Figure 1: (a) is a word-document matrix. The first 3 columns is one cluster, the last 4 columns is another cluster; (b) is $\mathbf{V}$ learned by GNMF with $\mu=1$; (c) is $\mathbf{V}$ learned by GNMF with $\mu=10^4$; (d) is $\mathbf{V}$ learned by GNMF with $\mu=10^6$. Note that we normalize each row of $\mathbf{V}$ by ℓ_1 norm for better viewing.

2.2 Scale Transfer Problem

Another problem that GNMF suffers from is scale transfer problem in the optimization. Suppose $\{\mathbf{U}^*, \mathbf{V}^*\}$ is the optimal solution for Eq. (3) with optimal objective function value, i.e., $J(\mathbf{U}^*, \mathbf{V}^*)$. For any real scalar $\alpha > 1$, the scale transferred solution $\{\alpha\mathbf{U}^*, \frac{\mathbf{V}^*}{\alpha}\}$, will lead to a "smaller" objective function value,

$$J(\alpha\mathbf{U}^*, \frac{\mathbf{V}^*}{\alpha}) < J(\mathbf{U}^*, \mathbf{V}^*). \quad (7)$$

Thus, the ultimate solution is $\mathbf{U}^* = \infty, \mathbf{V}^* = 0$ while $\mathbf{U}^*\mathbf{V}^{*T}$ remains a fixed value.

The same problem exists in GNMTF [Gu and Zhou, 2009]. To resolve this problem, [Gu and Zhou, 2009] proposed to use ℓ_2 normalization on columns of $\mathbf{U}$ and $\mathbf{V}$ in each iteration during the optimization, and compensate the norms of $\mathbf{U}$ and $\mathbf{V}$ to $\mathbf{S}$. However, this is not a principled way to solve the problem. We will see that this problem can be resolved in the new formulation proposed in this paper.

3 The Proposed Models

In this section, we present two Nonnegative Matrix (Tri-)Factorization models with Graph Regularization, followed by its optimization algorithm. We also prove the convergence of the optimization algorithm.

3.1 Formulation

The key idea to resolve the problem is adding a Normalized Cut [Shi and Malik, 2000] type constraint on $\mathbf{V}$, which leads to the following minimization problem

$$\begin{aligned} J_2 &= ||\mathbf{X} - \mathbf{U}\mathbf{V}^T||_F^2 - \mu\text{tr}(\mathbf{V}^T\mathbf{W}_V\mathbf{V}), \\ &\text{s.t.} \quad \mathbf{U} \geq 0, \mathbf{V} \geq 0, \mathbf{V}^T\mathbf{D}_V\mathbf{V} = \mathbf{I}, \end{aligned} \quad (8)$$

where $\mathbf{I}$ is the identity matrix with proper size, $\mathbf{W}_V$ is exactly the adjacency matrix defined on the document graph, μ is a positive regularization parameter balancing the reconstruction error in the first NMF term and the label smoothness in the second term. Note that we omit the term $\mu\mathbf{V}^T\mathbf{D}_V\mathbf{V}$ as it is a constant, i.e., $\mu\mathbf{I}$. Hence the regularization term turns out to be $-\mu\text{tr}(\mathbf{V}^T\mathbf{W}_V\mathbf{V})$. We can see that the difference between Eq. (8) and Eq. (3) is the additional constraint $\mathbf{V}^T\mathbf{D}\mathbf{V} = \mathbf{I}$. The difference is seemingly small, but plays an essential role. With this additional constraint, the optimization problem in Eq. (8) is well-defined and no longer suffers from either scale transfer problem or trivial solution. For example, let μ approach ∞, then the second term dominates the objective function, and Eq. (8) boils down to maximize

$$\begin{aligned} J_2' &= \text{tr}(\mathbf{V}^T\mathbf{W}_V\mathbf{V}), \\ &\text{s.t.} \quad \mathbf{V} \geq 0, \mathbf{V}^T\mathbf{D}_V\mathbf{V} = \mathbf{I}, \end{aligned} \quad (9)$$

which can be seen as nonnegative relaxed Normalized Cut.

Similar with the strategy adopted in Eq. (8), we add two additional constraints on $\mathbf{U}$ and $\mathbf{V}$ to Eq. (4), resulting in the following minimization problem,

$$\begin{aligned} J_3 &= ||\mathbf{X} - \mathbf{U}\mathbf{S}\mathbf{V}^T||_F^2 - \lambda\text{tr}(\mathbf{U}^T\mathbf{W}_U\mathbf{U}) - \mu\text{tr}(\mathbf{V}^T\mathbf{W}_V\mathbf{V}) \\ &\text{s.t.} \quad \mathbf{U} \geq 0, \mathbf{S} \geq 0, \mathbf{V} \geq 0 \\ &\quad \mathbf{U}^T\mathbf{D}_U\mathbf{U} = \mathbf{I}, \mathbf{V}^T\mathbf{D}_V\mathbf{V} = \mathbf{I}, \end{aligned} \quad (10)$$

where $\lambda, \mu \geq 0$ are regularization parameters balancing the reconstruction error of co-clustering in the first term, and the label smoothness of the words and documents in the second and third terms. The additional two constraints, i.e., $\mathbf{U}^T\mathbf{D}_U\mathbf{U} = \mathbf{I}$ and $\mathbf{V}^T\mathbf{D}_V\mathbf{V} = \mathbf{I}$, are essential to make the optimization in Eq. (10) well-defined and not prone to trivial solutions.

Since the optimization problem in Eq. (8) can be seen as a special case of that in Eq. (10) when absorbing $\mathbf{S}$ to $\mathbf{U}$ and setting $\lambda = 0$, in the sequel, we will develop the optimization algorithm to solve J_3, which is a more general problem. The derived optimization algorithm can be adapted to optimize J_2 very straightforwardly.

3.2 Optimization Algorithm

Optimizing Eq. (10) with respect to $\mathbf{U}$ is equivalent to optimizing

$$\begin{aligned} J_3' &= ||\mathbf{X} - \mathbf{U}\mathbf{S}\mathbf{V}^T||_F^2 - \lambda\text{tr}(\mathbf{U}^T\mathbf{W}_U\mathbf{U}) \\ &\text{s.t.} \quad \mathbf{U} \geq 0, \mathbf{U}^T\mathbf{D}_U\mathbf{U} = \mathbf{I}. \end{aligned} \quad (11)$$

Since $\mathbf{U}^T\mathbf{D}_U\mathbf{U} = \mathbf{I}$, we introduce the Lagrangian multiplier $\mathbf{\Lambda} \in \mathbb{R}^{m\times m}$, thus the Lagrangian function is

$$\begin{aligned} L(\mathbf{U}) &= ||\mathbf{X} - \mathbf{U}\mathbf{S}\mathbf{V}^T||_F^2 - \lambda\text{tr}(\mathbf{U}^T\mathbf{W}_U\mathbf{U}) \\ &+ \text{tr}(\mathbf{\Lambda}(\mathbf{U}^T\mathbf{D}_U\mathbf{U} - \mathbf{I})). \end{aligned} \quad (12)$$

The gradient of $L(\mathbf{U})$ with respect to $\mathbf{U}$ is

$$\begin{aligned} \frac{\partial L(\mathbf{U})}{\partial \mathbf{U}} &= -2\mathbf{X}\mathbf{V}\mathbf{S}^T + 2\mathbf{U}\mathbf{S}\mathbf{V}^T\mathbf{V}\mathbf{S}^T \\ &- 2\lambda\mathbf{W}_U\mathbf{U} + 2\mathbf{D}_U\mathbf{U}\mathbf{\Lambda}. \end{aligned} \quad (13)$$

Using the Karush-Kuhn-Tucker complementarity condition [Boyd and Vandenberghe, 2004] $(\frac{\partial L(\mathbf{U})}{\partial \mathbf{U}})_{ij}\mathbf{U}_{ij} = 0$, we get

$$(-\mathbf{X}\mathbf{V}\mathbf{S}^T + \mathbf{U}\mathbf{S}\mathbf{V}^T\mathbf{V}\mathbf{S}^T - \lambda\mathbf{W}_U\mathbf{U} + \mathbf{D}_U\mathbf{U}\mathbf{\Lambda})_{ij}\mathbf{U}_{ij} = 0. \quad (14)$$

Since $\mathbf{\Lambda}$ may take mixed signs, we introduce $\mathbf{\Lambda} = \mathbf{\Lambda}^+ - \mathbf{\Lambda}^-$, where $\mathbf{\Lambda}_{ij}^+ = (|\mathbf{\Lambda}_{ij}| + \mathbf{\Lambda}_{ij})/2$ and $\mathbf{\Lambda}_{ij}^- = (|\mathbf{\Lambda}_{ij}| - \mathbf{\Lambda}_{ij})/2$. Then we get the following updating formula

$$\mathbf{U}_{ij} \leftarrow \mathbf{U}_{ij}\sqrt{\frac{[\mathbf{X}\mathbf{V}\mathbf{S}^T + \lambda\mathbf{W}_U\mathbf{U} + \mathbf{D}_U\mathbf{U}\mathbf{\Lambda}^-]_{ij}}{[\mathbf{U}\mathbf{S}\mathbf{V}^T\mathbf{V}\mathbf{S}^T + \mathbf{D}_U\mathbf{U}\mathbf{\Lambda}^+]_{ij}}}. \quad (15)$$

It remains to determine the Lagrangian multiplier $\mathbf{\Lambda}$. Following the similar trick used in [Ding *et al.*, 2006], we obtain

$$\mathbf{\Lambda} = \mathbf{U}^T\mathbf{X}\mathbf{V}\mathbf{S}^T - \mathbf{U}^T\mathbf{U}\mathbf{S}\mathbf{V}^T\mathbf{V}\mathbf{S}^T + \lambda\mathbf{U}^T\mathbf{W}_U\mathbf{U}. \quad (16)$$

However, since $\mathbf{U}^T\mathbf{D}_U\mathbf{U} - \mathbf{I}$ is symmetric, we have $\text{tr}(\mathbf{\Lambda}(\mathbf{U}^T\mathbf{D}_U\mathbf{U} - \mathbf{I})) = \text{tr}((\mathbf{U}^T\mathbf{D}_U\mathbf{U} - \mathbf{I})\mathbf{\Lambda}^T) = \text{tr}(\mathbf{\Lambda}^T(\mathbf{U}^T\mathbf{D}_U\mathbf{U} - \mathbf{I}))$. Therefore only the symmetric part of $\mathbf{\Lambda}$ contributes to $L(\mathbf{U})$, i.e., $\mathbf{\Lambda}$ should also be symmetric. Hence we use $\mathbf{\Lambda}' = \frac{\mathbf{\Lambda}+\mathbf{\Lambda}^T}{2}$ instead of $\mathbf{\Lambda}$.

Similarly, we can obtain the updating formula for $\mathbf{V}$ as

$$\mathbf{V}_{ij} \leftarrow \mathbf{V}_{ij}\sqrt{\frac{[\mathbf{X}^T\mathbf{U}\mathbf{S} + \mu\mathbf{W}_V\mathbf{V} + \mathbf{D}_V\mathbf{V}\mathbf{\Xi}'^-]_{ij}}{[\mathbf{V}\mathbf{S}^T\mathbf{U}^T\mathbf{U}\mathbf{S} + \mathbf{D}_V\mathbf{V}\mathbf{\Xi}'^+]_{ij}}}, \quad (17)$$

where $\mathbf{\Xi}' = \frac{\mathbf{\Xi}+\mathbf{\Xi}^T}{2}$ and

$$\mathbf{\Xi} = \mathbf{V}^T\mathbf{X}^T\mathbf{U}\mathbf{S} - \mathbf{V}^T\mathbf{V}\mathbf{S}^T\mathbf{U}^T\mathbf{U}\mathbf{S} + \mu\mathbf{V}^T\mathbf{W}_V\mathbf{V}. \quad (18)$$

And the updating formula for $\mathbf{S}$ is

$$\mathbf{S}_{ij} \leftarrow \mathbf{S}_{ij}\sqrt{\frac{[\mathbf{U}^T\mathbf{X}\mathbf{V}]_{ij}}{[\mathbf{U}^T\mathbf{U}\mathbf{S}\mathbf{V}^T\mathbf{V}]_{ij}}}. \quad (19)$$

The iteration in Eq.(15), (17) and (19) terminates until convergence.

3.3 Convergence Analysis

In this section, we prove the convergence of the updating formulas in Eq. (15), Eq. (17) and Eq.(19), using the auxiliary function approach [Lee and Seung, 2000].

Definition 3.1. *[Lee and Seung, 2000] $Z(h, h')$ is an auxiliary function for $F(h)$ if the conditions*

$$Z(h, h') \geq F(h), Z(h, h) = F(h),$$

are satisfied.

Lemma 3.2. *[Lee and Seung, 2000] If Z is an auxiliary function for F, then F is non-increasing under the update*

$$h^{(t+1)} = \arg\min_h Z(h, h^{(t)}).$$

Theorem 3.3. *Let*

$$\begin{aligned} J_3(\mathbf{U}) &= tr(-2\mathbf{U}^T\mathbf{X}\mathbf{V}\mathbf{S}^T + \mathbf{U}\mathbf{S}\mathbf{V}^T\mathbf{V}\mathbf{S}^T\mathbf{U}^T \\ &- \lambda\mathbf{U}^T\mathbf{W}_U\mathbf{U} + \mathbf{\Lambda}\mathbf{U}^T\mathbf{D}_U\mathbf{U}), \end{aligned} \quad (20)$$

by ignoring the term $-tr(\mathbf{\Lambda})$. Then the following function

$$\begin{aligned} & Z(\mathbf{U}, \mathbf{U}') \\ = & -2\sum_{ij}(\mathbf{X}\mathbf{V}\mathbf{S}^T)_{ij}\mathbf{U}'_{ij}(1 + \log\frac{\mathbf{U}_{ij}}{\mathbf{U}'_{ij}}) \\ + & \sum_{ij}\frac{(\mathbf{U}'\mathbf{S}\mathbf{V}^T\mathbf{V}\mathbf{S}^T)_{ij}\mathbf{U}^2_{ij}}{\mathbf{U}'_{ij}} + \sum_{ij}\frac{(\mathbf{D}_U\mathbf{U}'\mathbf{\Lambda}^+)_{ij}\mathbf{U}^2_{ij}}{\mathbf{U}'_{ij}} \\ - & \lambda\sum_{ijk}(\mathbf{W}_U)_{jk}\mathbf{U}'_{ji}\mathbf{U}'_{ki}(1 + \log\frac{\mathbf{U}_{ji}\mathbf{U}_{ki}}{\mathbf{U}'_{ji}\mathbf{U}'_{ki}}) \\ - & \sum_{ijkl}(\mathbf{\Lambda}^-)_{kj}(\mathbf{D}_U)_{jl}\mathbf{U}'_{ji}\mathbf{U}'_{lk}(1 + \log\frac{\mathbf{U}_{ji}\mathbf{U}_{lk}}{\mathbf{U}'_{ji}\mathbf{U}'_{lk}}) \end{aligned}$$

is an auxiliary function for $J_3(\mathbf{U})$. Furthermore, it is a convex function in $\mathbf{U}$ and its global minimum is

$$\mathbf{U}_{ij} = \mathbf{U}'_{ij}\sqrt{\frac{[\mathbf{X}\mathbf{V}\mathbf{S}^T + \lambda\mathbf{W}_U\mathbf{U}' + \mathbf{D}_U\mathbf{U}'\mathbf{\Lambda}^-]_{ij}}{[\mathbf{U}'\mathbf{S}\mathbf{V}^T\mathbf{V}\mathbf{S}^T + \mathbf{D}_U\mathbf{U}'\mathbf{\Lambda}^+]_{ij}}}. \quad (21)$$

Proof. For the space limit, we omit it. It will be presented in the longer version of this paper. □

Theorem 3.4. *Updating $\mathbf{U}$ using Eq. (15) will monotonically decrease the value of the objective in Eq. (10), hence it converges.*

Proof. By Lemma 3.2 and Theorem 3.3, we can get that $J_3(\mathbf{U}^0) = Z(\mathbf{U}^0, \mathbf{U}^0) \geq Z(\mathbf{U}^1, \mathbf{U}^0) \geq J_3(\mathbf{U}^1) \geq \ldots$. Thus $J_3(\mathbf{U})$ is monotonically decreasing. Since $J_3(\mathbf{U})$ is obviously bounded below, the theorem is proved. □

The convergence of the updating formulas of $\mathbf{V}$ and $\mathbf{S}$ can be proved similarly. For the space limit, we omit it.

4 Experiments

In this section, We compare our methods with K-means [Bishop, 2006], Normalized Cut (NCut) [Shi and Malik, 2000], NMF [Lee and Seung, 2000], ONMTF [Ding *et al.*, 2006], GNMF [Cai *et al.*, 2008] and GNMTF [Gu and Zhou, 2009]. For GNMF and GNMTF, we test the un-normalized version, and the version doing ℓ_2 normalization on columns of $\mathbf{V}$ (and $\mathbf{U}$) in each iteration, which are referred to as GNMF_C2 and GNMTF_C2. We refer to the proposed model in Eq. (8) as IGNMF and the model in Eq. (10) as IGNMTF.

4.1 Evaluation Metrics

To evaluate the clustering results, we adopt the performance measures used in [Xu *et al.*, 2003]. These performance measures are the standard measures widely used for clustering.

Clustering Accuracy is defined as follows:

$$Acc = \frac{\sum_{i=1}^n \delta(map(r_i), l_i)}{n}, \quad (22)$$

where r_i denotes the cluster label of $\mathbf{x}_i$, and l_i denotes the true class label, n is the total number of documents, $\delta(x, y)$ is the delta function that equals one if $x = y$ and equals zero otherwise, and $map(r_i)$ is the permutation mapping function that maps each cluster label r_i to the equivalent label from the data set.

Normalized Mutual Information is used for determining the quality of clusters. Given a clustering result, the NMI is estimated by

$$NMI = \frac{\sum_{i=1}^c\sum_{j=1}^c n_{i,j}\log\frac{n_{i,j}}{n_i\hat{n}_j}}{\sqrt{(\sum_{i=1}^c n_i\log\frac{n_i}{n})(\sum_{j=1}^c \hat{n}_j\log\frac{\hat{n}_j}{n})}}, \quad (23)$$

where n_i denotes the number of data contained in the cluster $\mathcal{C}_i(1 \leq i \leq c)$, $\hat{n}_j$ is the number of data belonging to the $\mathcal{L}_j(1 \leq j \leq c)$, and $n_{i,j}$ denotes the number of data that are in the intersection between the cluster $\mathcal{C}_i$ and the class $\mathcal{L}_j$. The larger the NMI is, the better the clustering result will be.

4.2 Data Sets

In our experiment, we use four data sets which are widely used as benchmark data sets in clustering literature [Ding *et al.*, 2006] [Cai *et al.*, 2005].

CSTR consists of the abstracts of technical reports published in the Department of Computer Science at a university. The data set contained 476 abstracts, which were divided into four research areas: Natural Language Processing (NLP), Robotics/Vision, Systems and Theory.

News4 is selected from the famous 20-newsgroups data set[1]. The topic *rec* containing *autos, motorcycles, baseball* and *hockey* was selected from the version 20news-18828.

WebKB4 contains webpages gathered from university computer science departments. There are about 8280 documents and they are divided into 7 categories: student, faculty, staff, course, project, department and others, among which student, faculty, course and project are four largest entity-representing categories.

[1] http://people.csail.mit.edu/jrennie/20Newsgroups/

TDT2 consists of articles collected during the first half of 1998 and taken from 6 sources, including 2 newswires, 2 radio programs and 2 television programs. In this experiment, those documents appearing in two or more categories were removed. We use 10 out of 96 semantic categories.

4.3 Parameter Settings

In order to compare these algorithms fairly, we run these algorithms under different parameter settings, and select the best average result to compare with each other. We set the number of clusters equal to the true number of classes for all the data sets and clustering algorithms.

For NCut [Shi and Malik, 2000], the scale parameter of Gaussian kernel for constructing adjacency matrix is set by the grid $\{10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 10^2, 10^3\}$.

For ONMTF, the number of word clusters is set to be the same as the number of document clusters, i.e., the true number of classes, according to [Ding *et al.*, 2006].

For GNMF, GNMF_C2 and IGNMF, the neighborhood size k in Eq. (2) is set to 10 according to the observation in [Cai *et al.*, 2008], and the regularization parameter is set by searching the grid $\{0.1, 1, 10, 50, 100, 500, 1000\}$.

For GNMTF, GNMTF_C2 and IGNMTF, the number of word clusters is set the same as the number of document clusters, i.e., the true number of classes, as in ONMTF. According to the observation in [Gu and Zhou, 2009], the neighborhood size k of the document graph as well as the word graph is set to 10. We also set $\lambda = \mu$ and tune it by the grid $\{0.1, 1, 10, 50, 100, 500, 1000\}$.

Under each parameter setting of each method mentioned above, we repeat clustering 20 times, and the average result is computed. We report the best average result for each method.

4.4 Clustering Results

Table 1 shows the clustering accuracy of all the algorithms on all the data sets, while Table 2 shows the normalized mutual information.

Table 1: Clustering Accuracy (%)

Data Sets	CSTR	News4	WebKB4	TDT2
Kmeans	76.34	81.58	69.73	76.97
NCut	65.97	60.56	67.16	65.97
NMF	70.97	85.22	69.55	85.28
GNMF	73.11	85.02	72.89	87.04
GNMF_C2	78.35	84.75	71.38	88.03
IGNMF	**85.29**	**87.93**	**75.19**	**91.71**
ONMTF	77.00	83.99	68.85	82.54
GNMTF	75.81	73.34	70.83	87.04
GNMTF_C2	81.76	88.53	71.37	84.04
IGNMTF	**85.90**	**90.94**	**73.81**	**92.28**

We can see that GNMF_C2 is better than GNMF on 2 out of 4 data sets. This implies that to some extent normalization can resolve the problems suffered by GNMF, and in turn improve the performance of GNMF. IGNMF outperforms GNMF and GNMF_C2 consistently. This indicates the strength of the proposed model, which benefits from its well-defined optimization problem.

Table 2: Normalized Mutual Information (%)

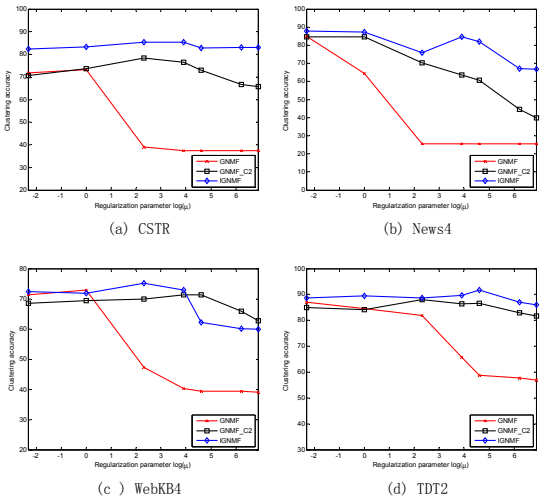
Data Sets	CSTR	News4	WebKB4	TDT2
Kmeans	65.31	71.29	46.65	86.09
NCut	57.61	72.12	44.37	55.73
NMF	67.80	71.98	43.73	88.69
GNMF	67.49	71.07	46.86	89.09
GNMF_C2	66.01	70.90	47.61	90.06
IGNMF	**69.76**	**72.68**	**48.46**	**92.19**
ONMTF	67.16	70.53	45.52	86.09
GNMTF	62.94	58.10	45.64	90.45
GNMTF_C2	63.14	74.71	**46.45**	87.22
IGNMTF	**72.49**	**76.87**	44.87	**90.98**

Similarly, GNMTF_C2 outperforms GNMTF on 3 out of 4 data sets according to the clustering accuracy. This implies that the scale transfer problem is even more serious in GNMTF, since there are two regularizers, i.e. $\mathrm{tr}(\mathbf{U}^T\mathbf{L}_U\mathbf{U})$ and $\mathrm{tr}(\mathbf{V}^T\mathbf{L}_V\mathbf{V})$ in the objective function. Again, IGNMTF is superior to GNMTF and GNMTF_C2, because the scale transfer problem is resolved in a principled way in IGNMTF.

GNMTF type methods are superior to GNMF type methods on 3 out of 4 data sets. Similar observation was observed in [Gu and Zhou, 2009]. This verifies the manifold assumption on the words again.

4.5 Study on the Regularization Parameter

In this subsection, we first investigate the sensitivity of GNMF, GNMF_C2 (GNMF with $\ell 2$ normalization on columns of $\mathbf{V}$) and IGNMF with respect to the regularization parameter μ. We vary the value of μ, and plot the average clustering accuracy in Figure 2. For better viewing, we scale μ by $\log(\mu)$.



(a) CSTR (b) News4

(c) WebKB4 (d) TDT2

Figure 2: Clustering accuracy (%) of GNMF type methods with respect to the regularization parameter μ.

Figure 2 shows that when μ is larger than 1, the performance of GNMF declines sharply due to the trivial solution problem. For example, on the CSTR data set, when $\mu = 50, 100, 500, 1000$, the clustering accuracy are all 37.39%.

Since there are 178 documents in the third class, we can see that $178/476$ is exactly 37.39%. This means that GNMF assigns all the documents in this data set to the third cluster, which is the trivial solution problem we analyzed in Section 2. Similar phenomenons can be observed in the other 3 data sets. GNMF_C2 is able to overcome this problem to certain extent. However, the price of normalization is potential performance decrease. IGNMF performs the best and is not very sensitive to the regularization parameter μ. More importantly, it does not sacrifice its performance to solve the problem.

Second, we study the sensitivity of GNMTF, GNMTF_C2 (GNMTF with $\ell 2$ normalization on columns of $\mathbf{U}$ and $\mathbf{V}$), and IGNMTF with respect to the regularization parameter $\mu(=\lambda)$. We vary the value of $\mu(=\lambda)$, and plot the average clustering accuracy with respect to $\mu(=\lambda)$ in Figure 3. Note that the horizontal axis represents $\log(\mu)(=\log(\lambda))$.

Figure 3 illustrates that when μ is larger than 0.1, the performance of GNMTF drops off precipitously due to the trivial solution problem. We can see that the problem is even more serious in GNMTF as we pointed above. If we take a closer look, we can also observe that GNMTF assigns all the documents to one cluster when μ is very large, e.g. $\mu = 50, 100, 500, 1000$. GNMTF_C2 is able to overcome this problem to some extent, which is consistent with the experimental results in [Gu and Zhou, 2009]. IGNMTF performs best and is very stable with respect to the regularization parameter $\mu(=\lambda)$. This strengthens the advantage of the new formulation in Eq. (10).

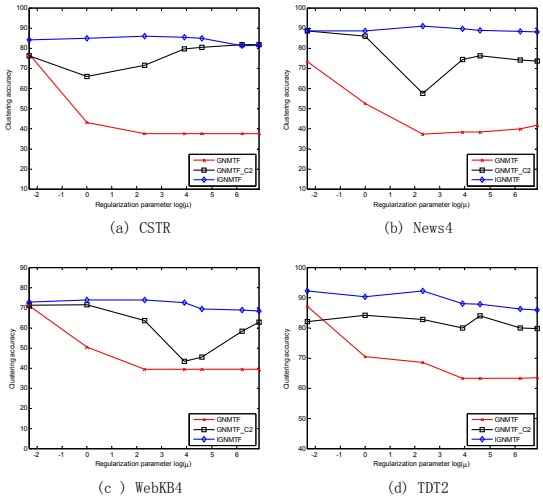


Figure 3: Clustering accuracy (%) of GNMTF type methods with respect to the regularization parameter $\mu(=\lambda)$.

5 Conclusions

In this paper, we propose two models for graph regularized nonnegative matrix factorization and graph regularized nonnegative matrix tri-factorization respectively, which overcome the scale transfer problem and trivial solution problem in GNMF and GNMTF. The proposed models can be solved via multiplicative updating algorithm, and its convergence is theoretically guaranteed. Experiments of clustering on many benchmark data sets demonstrate that the proposed models outperform the original models and many other state-of-the-art clustering methods.

Acknowledgments

The work was supported in part by NSF IIS-09-05215, U.S. Air Force Office of Scientific Research MURI award FA9550-08-1-0265, and the U.S. Army Research Laboratory under Cooperative Agreement Number W911NF-09-2-0053 (NS-CTA).

References

[Belkin *et al.*, 2006] Mikhail Belkin, Partha Niyogi, and Vikas Sindhwani. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *Journal of Machine Learning Research*, 7:2399–2434, 2006.

[Bishop, 2006] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[Boyd and Vandenberghe, 2004] Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge University Press, Cambridge, 2004.

[Cai *et al.*, 2005] Deng Cai, Xiaofei He, and Jiawei Han. Document clustering using locality preserving indexing. *IEEE Trans. Knowl. Data Eng.*, 17(12):1624–1637, 2005.

[Cai *et al.*, 2008] Deng Cai, Xiaofei He, Xiaoyun Wu, and Jiawei Han. Non-negative matrix factorization on manifold. In *ICDM*, pages 63–72, 2008.

[Dhillon *et al.*, 2003] Inderjit S. Dhillon, Subramanyam Mallela, and Dharmendra S. Modha. Information-theoretic co-clustering. In *KDD*, pages 89–98, 2003.

[Ding *et al.*, 2006] Chris H. Q. Ding, Tao Li, Wei Peng, and Haesun Park. Orthogonal nonnegative matrix t-factorizations for clustering. In *KDD*, pages 126–135, 2006.

[Ding *et al.*, 2010] Chris H. Q. Ding, Tao Li, and Michael I. Jordan. Convex and semi-nonnegative matrix factorizations. *IEEE Trans. Pattern Anal. Mach. Intell.*, 32(1):45–55, 2010.

[Gu and Zhou, 2009] Quanquan Gu and Jie Zhou. Co-clustering on manifolds. In *KDD*, pages 359–368, 2009.

[Lee and Seung, 2000] Daniel D. Lee and H. Sebastian Seung. Algorithms for non-negative matrix factorization. In *NIPS*, pages 556–562, 2000.

[Li and Ding, 2006] Tao Li and Chris H. Q. Ding. The relationships among various nonnegative matrix factorization methods for clustering. In *ICDM*, pages 362–371, 2006.

[Shi and Malik, 2000] Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(8):888–905, 2000.

[Xu *et al.*, 2003] Wei Xu, Xin Liu, and Yihong Gong. Document clustering based on non-negative matrix factorization. In *SIGIR*, pages 267–273, 2003.