

On the clustering aspect of nonnegative matrix factorization

Andri Mirzal

Grad. School of Information Science and Technology,
Hokkaido University, Kita 14 Nishi 9, Kita-Ku,
Sapporo, Japan
andri@complex.eng.hokudai.ac.jp

Masashi Furukawa

Grad. School of Information Science and Technology,
Hokkaido University, Kita 14 Nishi 9, Kita-Ku,
Sapporo, Japan
mack@complex.eng.hokudai.ac.jp

Abstract—This paper provides a theoretical explanation on the clustering aspect of nonnegative matrix factorization (NMF). We prove that even without imposing orthogonality nor sparsity constraint on the basis and/or coefficient matrix, NMF still can give clustering results, thus providing a theoretical support for many works, e.g., Xu et al. [1] and Kim et al. [2], that show the superiority of the standard NMF as a clustering method.

Keywords—bound-constrained optimization, clustering method, non-convex optimization, nonnegative matrix factorization

I. INTRODUCTION

NMF is a matrix approximation technique that factorizes a nonnegative matrix into a pair of other nonnegative matrices of much lower rank:

$$\mathbf{A} \approx \mathbf{B}\mathbf{C}, \quad (1)$$

where $\mathbf{A} \in \mathbb{R}_+^{M \times N} = [\mathbf{a}_1, \dots, \mathbf{a}_N]$ denotes the feature-by-item data matrix, $\mathbf{B} \in \mathbb{R}_+^{M \times K} = [\mathbf{b}_1, \dots, \mathbf{b}_K]$ denotes the basis matrix, $\mathbf{C} \in \mathbb{R}_+^{K \times N} = [\mathbf{c}_1, \dots, \mathbf{c}_N]$ denotes the coefficient matrix, and K denotes the number of factors which usually chosen so that $K \ll \min(M, N)$. There are also other variants of NMF like semi-NMF, convex NMF, and symmetric NMF. Detailed discussions can be found in, e.g., [3] and [4].

The nonnegativity constraints and the reduced dimensionality define the uniqueness and power of NMF. The nonnegativity constraints allow only nonsubtractive linear combinations of the basis vectors \mathbf{b}_k to construct the data vectors \mathbf{a}_n , thus providing the parts-based interpretations as shown in [5], [6], [7]. And the reduced dimensionality provides NMF with the clustering aspect and data compression capabilities.

The most important NMF's application is in the data clustering, as some works have shown that it is a superior method compared to the standard clustering methods like spectral methods and K -means algorithm. In particular, Xu et al. [1] showed that NMF outperforms standard spectral methods in finding the document clustering in two text corpora, TDT2 and Reuters. And Kim et al. [2] showed that NMF and sparse NMF are much more superior methods compared to the K -means algorithm in both a synthetic dataset (which is well separated) and a real dataset (TDT2).

If sparsity constraints are imposed to columns of \mathbf{C} , the clustering aspect of NMF is intuitive since in the extreme case where there is only one nonzero entry per column, NMF will be equivalent to the K -means algorithm employed to the data vectors \mathbf{a}_n [8], and the sparsity constraints can be thought as the relaxation to the strict orthogonality constraints on rows of

\mathbf{C} (an equivalent explanation can also be stated for imposing sparsity on rows of \mathbf{B}).

However, as reported by Xu et al. [1] and Kim et al. [2], even without imposing sparsity constraints, NMF still can give very promising clustering results. But the authors didn't give any theoretical analysis on why the standard NMF—NMF without sparsity nor orthogonality constraint—can give such good results. So far the best explanation for this remarkable fact is only qualitative: the standard NMF produces non-orthogonal latent semantic directions (the basis vectors) that are more likely to correspond to each of the clusters than those produced by the spectral methods, thus the clustering induced from the latent semantic directions of the standard NMF are better than clustering by the spectral methods [1]. Therefore, this work attempts to provide a theoretical support for the clustering aspect of the standard NMF.

II. CLUSTERING ASPECT OF NMF

To compute \mathbf{B} and \mathbf{C} , usually eq. 1 is rewritten into a minimization problem in the Frobenius norm criterion.

$$\min_{\mathbf{B}, \mathbf{C}} J(\mathbf{B}, \mathbf{C}) = \frac{1}{2} \|\mathbf{A} - \mathbf{B}\mathbf{C}\|_F^2 \text{ s.t. } \mathbf{B} \geq \mathbf{0}, \mathbf{C} \geq \mathbf{0}. \quad (2)$$

In addition to the usual Frobenius norm criterion, the family of Bregman divergences—which Frobenius norm and Kullback-Leibler divergence are part of it—can also be used as the affinity measures. Detailed discussion on the Bregman divergences for NMF can be found in [9].

Sometimes it is more practical and intuitive to decompose $J(\mathbf{B}, \mathbf{C})$ into a series of smaller objectives.

$$\min_{\mathbf{B}, \mathbf{C}} J(\mathbf{B}, \mathbf{C}) \equiv \left(\min_{\mathbf{B}, \mathbf{c}_1} J_1(\mathbf{B}, \mathbf{c}_1), \dots, \min_{\mathbf{B}, \mathbf{c}_N} J_N(\mathbf{B}, \mathbf{c}_N) \right), \quad (3)$$

where

$$\min_{\mathbf{B}, \mathbf{c}_n} J_n(\mathbf{B}, \mathbf{c}_n) = \frac{1}{2} \|\mathbf{a}_n - \mathbf{B}\mathbf{c}_n\|_2^2, \quad n \in [1, N]. \quad (4)$$

Minimizing J_n is known to be the nonnegative least square (NLS) problem, and some fast NMF algorithms are developed based on solving the NLS subproblems, e.g., alternating NLS with block principal pivoting algorithm [10], active set method [11], and projected quasi-Newton algorithm [12]. Decomposing NMF problem into NLS subproblems also transforms the non-convex optimization in eq. 3 to the convex optimization subproblems in eq. 4. Even though eq. 4 is not strictly convex,

for two-block case, any limit point of the sequence $\{\mathbf{B}^t, \mathbf{C}^t\}$, where t is the updating step, is a stationary point [13].

The objective in eq. 4 aims to simultaneously find the suitable basis vectors such that the latent factors are revealed, and the coefficient vector \mathbf{c}_n such that a linear combination of the basis vectors ($\mathbf{B}\mathbf{c}_n$) is close to \mathbf{a}_n . In clustering term this can be rephrased as: to simultaneously find the cluster centers and the cluster assignments.

To investigate the clustering aspect of NMF, four possibilities of NMF settings are discussed: (1) imposing orthogonality constraints on both rows of \mathbf{C} and columns of \mathbf{B} , (2) imposing orthogonality constraints on rows of \mathbf{C} , (3) imposing orthogonality constraints on columns of \mathbf{B} , and (4) no orthogonality constraint is imposed. The last case is the standard NMF which its clustering aspect is the focus of this paper as many works reported that it is a very effective clustering method.

A. Orthogonality constraints on both \mathbf{B} and \mathbf{C}

The following theorems proves that imposing column-orthogonality constraints on \mathbf{B} and row-orthogonality constraints on \mathbf{C} lead to the simultaneous clustering of similar items and related features.

Theorem 1. *Minimizing the following objective*

$$\begin{aligned} \min_{\mathbf{B}, \mathbf{C}} J_a(\mathbf{B}, \mathbf{C}) &= \frac{1}{2} \|\mathbf{A} - \mathbf{B}\mathbf{C}\|_F^2 \\ \text{s.t. } \mathbf{B} &\geq \mathbf{0}, \mathbf{C} \geq \mathbf{0}, \mathbf{B}^T \mathbf{B} = \mathbf{I}, \mathbf{C}\mathbf{C}^T = \mathbf{I} \end{aligned} \quad (5)$$

is equivalent to applying ratio association to $\mathcal{G}(\mathbf{A}^T \mathbf{A})$ and $\mathcal{G}(\mathbf{A}\mathbf{A}^T)$, where $\mathbf{A}^T \mathbf{A}$ and $\mathbf{A}\mathbf{A}^T$ are the item affinity matrix and the feature affinity matrix respectively, thus leads to simultaneous clustering of similar items and related features.

Proof:

$$\begin{aligned} \|\mathbf{A} - \mathbf{B}\mathbf{C}\|_F^2 &= \text{tr}((\mathbf{A} - \mathbf{B}\mathbf{C})^T (\mathbf{A} - \mathbf{B}\mathbf{C})) \\ &= \text{tr}(\mathbf{A}^T \mathbf{A} - 2\mathbf{C}^T \mathbf{B}^T \mathbf{A} + \mathbf{I}). \end{aligned} \quad (6)$$

The Lagrangian function:

$$\begin{aligned} L_a(\mathbf{B}, \mathbf{C}) &= J_a(\mathbf{B}, \mathbf{C}) - \text{tr}(\mathbf{\Gamma}_B \mathbf{B}^T) - \text{tr}(\mathbf{\Gamma}_C \mathbf{C}) + \\ &\quad \text{tr}(\mathbf{\Lambda}_B (\mathbf{B}^T \mathbf{B} - \mathbf{I})) + \text{tr}(\mathbf{\Lambda}_C (\mathbf{C}\mathbf{C}^T - \mathbf{I})), \end{aligned} \quad (7)$$

where $\mathbf{\Gamma}_B \in \mathbb{R}_+^{M \times K}$, $\mathbf{\Gamma}_C \in \mathbb{R}_+^{N \times K}$, $\mathbf{\Lambda}_B \in \mathbb{R}_+^{K \times K}$, and $\mathbf{\Lambda}_C \in \mathbb{R}_+^{K \times K}$ are the Lagrange multipliers. By the Karush-Kuhn-Tucker (KKT) optimality conditions we get:

$$\nabla_{\mathbf{B}} L_a = \mathbf{B} - \mathbf{A}\mathbf{C}^T - \mathbf{\Gamma}_B + 2\mathbf{B}\mathbf{\Lambda}_B = \mathbf{0}, \quad (8)$$

$$\nabla_{\mathbf{C}} L_a = \mathbf{C} - \mathbf{B}^T \mathbf{A} - \mathbf{\Gamma}_C^T + 2\mathbf{\Lambda}_C \mathbf{C} = \mathbf{0}, \quad (9)$$

with complementary slackness:

$$\mathbf{\Gamma}_B \otimes \mathbf{B} = \mathbf{0}, \quad \mathbf{\Gamma}_C^T \otimes \mathbf{C} = \mathbf{0}, \quad (10)$$

where \otimes denotes component-wise multiplications. Assume $\mathbf{\Gamma}_B = \mathbf{0}$, $\mathbf{\Lambda}_B = \mathbf{0}$, $\mathbf{\Gamma}_C = \mathbf{0}$, and $\mathbf{\Lambda}_C = \mathbf{0}$ (at the stationary point these assumptions are reasonable since the

complementary slackness conditions hold and the Lagrange multipliers can be assigned to zeros), we get:

$$\mathbf{B} = \mathbf{A}\mathbf{C}^T \text{ and} \quad (11)$$

$$\mathbf{C} = \mathbf{B}^T \mathbf{A}. \quad (12)$$

Substituting eq. 11 into eq. 6, we get:

$$\min_{\mathbf{C}} J_a(\mathbf{C}) \equiv \max_{\mathbf{C}} \text{tr}(\mathbf{C}\mathbf{A}^T \mathbf{A}\mathbf{C}^T). \quad (13)$$

Similarly, substituting eq. 12 into eq. 6, we get:

$$\min_{\mathbf{B}} J_a(\mathbf{B}) \equiv \max_{\mathbf{B}} \text{tr}(\mathbf{B}^T \mathbf{A}\mathbf{A}^T \mathbf{B}). \quad (14)$$

Therefore, minimizing J_a is equivalent to simultaneously optimizing:

$$\max_{\mathbf{C}} \text{tr}(\mathbf{C}\mathbf{A}^T \mathbf{A}\mathbf{C}^T) \text{ s.t. } \mathbf{C}\mathbf{C}^T = \mathbf{I}, \text{ and} \quad (15)$$

$$\max_{\mathbf{B}} \text{tr}(\mathbf{B}^T \mathbf{A}\mathbf{A}^T \mathbf{B}) \text{ s.t. } \mathbf{B}^T \mathbf{B} = \mathbf{I}. \quad (16)$$

Eq. 15 and eq.16 are the ratio association objectives (see [14] for details on various graph cuts objectives) applied to $\mathcal{G}(\mathbf{A}^T \mathbf{A})$ and $\mathcal{G}(\mathbf{A}\mathbf{A}^T)$ respectively. Thus minimizing J_a leads to the simultaneous clustering of similar items and related features. ■

B. Orthogonality constraints on \mathbf{C}

When the orthogonality constraints are imposed only on rows of \mathbf{C} , it is no longer clear whether columns of \mathbf{B} will lead to the feature clustering. The following theorem shows that without imposing the orthogonality constraints on \mathbf{B} , the resulting \mathbf{B} can still lead to the feature clustering.

Theorem 2. *Minimizing the following objective*

$$\begin{aligned} \min_{\mathbf{B}, \mathbf{C}} J_b(\mathbf{B}, \mathbf{C}) &= \frac{1}{2} \|\mathbf{A} - \mathbf{B}\mathbf{C}\|_F^2 \\ \text{s.t. } \mathbf{B} &\geq \mathbf{0}, \mathbf{C} \geq \mathbf{0}, \mathbf{C}\mathbf{C}^T = \mathbf{I} \end{aligned} \quad (17)$$

is equivalent to applying ratio association to $\mathcal{G}(\mathbf{A}^T \mathbf{A})$, and also leads to the feature clustering indicator matrix \mathbf{B} which is approximately column-orthogonal.

Proof:

$$\begin{aligned} \|\mathbf{A} - \mathbf{B}\mathbf{C}\|_F^2 &= \text{tr}((\mathbf{A} - \mathbf{B}\mathbf{C})^T (\mathbf{A} - \mathbf{B}\mathbf{C})) \\ &= \text{tr}(\mathbf{A}^T \mathbf{A} - 2\mathbf{B}^T \mathbf{A}\mathbf{C}^T + \mathbf{C}^T \mathbf{B}^T \mathbf{B}\mathbf{C}). \end{aligned} \quad (18)$$

The Lagrangian function:

$$\begin{aligned} L_b(\mathbf{B}, \mathbf{C}) &= J_b(\mathbf{B}, \mathbf{C}) - \text{tr}(\mathbf{\Gamma}_B \mathbf{B}^T) - \text{tr}(\mathbf{\Gamma}_C \mathbf{C}) + \\ &\quad \text{tr}(\mathbf{\Lambda}_C (\mathbf{C}\mathbf{C}^T - \mathbf{I})). \end{aligned} \quad (19)$$

By applying the KKT conditions, we get:

$$\mathbf{B} = \mathbf{A}\mathbf{C}^T \text{ and} \quad (20)$$

$$\mathbf{C} = \mathbf{B}^T \mathbf{A}. \quad (21)$$

By substituting eq. 20 and eq. 21 into eq. 18, minimizing J_b is equivalent to simultaneously optimizing:

$$\max_{\mathbf{C}} \text{tr}(\mathbf{C}\mathbf{A}^T\mathbf{A}\mathbf{C}^T) \text{ s.t. } \mathbf{C}\mathbf{C}^T = \mathbf{I}, \quad (22)$$

$$\max_{\mathbf{B}} \text{tr}(\mathbf{B}^T\mathbf{A}\mathbf{A}^T\mathbf{B}), \text{ and} \quad (23)$$

$$\min_{\mathbf{B}} \text{tr}(\mathbf{A}^T\mathbf{B}\mathbf{B}^T\mathbf{B}\mathbf{A}) \equiv \min_{\mathbf{B}} \text{tr}(\mathbf{B}^T\mathbf{B}\mathbf{B}^T\mathbf{B}). \quad (24)$$

Note that the step in eq. 24 is justifiable since \mathbf{A} is a constant matrix. By using the fact $\text{tr}(\mathbf{X}^T\mathbf{X}) = \|\mathbf{X}\|_F^2$, eq. 24 can be rewritten as:

$$\min_{\mathbf{B}} \|\mathbf{B}^T\mathbf{B}\|_F^2 = \min_{\mathbf{B}} \left(\sum_i (\mathbf{b}_i^T \mathbf{b}_i)^2 + \sum_{i \neq j} (\mathbf{b}_i^T \mathbf{b}_j)^2 \right). \quad (25)$$

The objective in eq. 22 is equivalent to eq. 15 and eventually leads to the clustering of similar items. So the remaining problem is how to prove that optimizing eq. 23 and 25 simultaneously will lead to the feature clustering indicator matrix \mathbf{B} which is approximately column-orthogonal.

Eq. 23 resembles eq. 14, but without orthogonality nor upper bound constraint, so one can easily optimize eq. 23 by setting \mathbf{B} to an infinity matrix. However, this violates eq. 25 which favors small \mathbf{B} . Conversely, one can optimize eq. 25 by setting \mathbf{B} to a null matrix, but again this violates eq. 23. Therefore, these two objectives create implicit lower and upper bound constraints on \mathbf{B} , and eq. 23 and eq. 25 can be rewritten into:

$$\max_{\mathbf{B}} \text{tr}(\mathbf{B}^T \hat{\mathbf{A}} \mathbf{B}), \text{ and} \quad (26)$$

$$\min_{\mathbf{B}} \left(\underbrace{\sum_i (\mathbf{b}_i^T \mathbf{b}_i)^2}_{j_{b1}} + \underbrace{\sum_{i \neq j} (\mathbf{b}_i^T \mathbf{b}_j)^2}_{j_{b2}} \right) \quad (27)$$

$$\text{s.t. } \mathbf{0} \leq \mathbf{B} \leq \Upsilon_{\mathbf{B}},$$

where $\hat{\mathbf{A}}$ denotes the feature affinity matrix and $\Upsilon_{\mathbf{B}}$ denotes the upperbound constraints on \mathbf{B} . Now we have box-constraint objectives which are known to behave well and are guaranteed to converge to the stationary point [15].

Even though the objectives are now transformed into box-constraint optimization problems, since there is no column-orthogonality constraint, maximizing eq. 26 can be easily done by setting each entry of \mathbf{B} to the corresponding largest possible value (in graph term this means to only create one partition on $\mathcal{G}(\hat{\mathbf{A}})$). But this scenario results in the maximum value of eq. 27, which violates the objective. Conversely, minimizing eq. 27 to the smallest possible value (minimizing j_{b1} implies minimizing j_{b2} , but not vice versa) violates eq. 26.

Thus, the most reasonable scenario is: setting j_{b2} as small as possible and balancing j_{b1} with eq. 26. This scenario is the relaxed ratio association applied to $\mathcal{G}(\hat{\mathbf{A}})$, and as long as vertices of $\mathcal{G}(\hat{\mathbf{A}})$ are clustered, simultaneous optimizing eq. 26 and eq. 27 leads to the clustering of related features. Moreover, as j_{b2} is minimum, \mathbf{B} is approximately column-orthogonal. ■

C. Orthogonality constraints on \mathbf{B}

Theorem 3. Minimizing the following objective

$$\min_{\mathbf{B}, \mathbf{C}} J_c(\mathbf{B}, \mathbf{C}) = \frac{1}{2} \|\mathbf{A} - \mathbf{B}\mathbf{C}\|_F^2 \quad (28)$$

$$\text{s.t. } \mathbf{B} \geq \mathbf{0}, \mathbf{C} \geq \mathbf{0}, \mathbf{B}^T\mathbf{B} = \mathbf{I}$$

is equivalent to applying ratio association to $\mathcal{G}(\mathbf{A}\mathbf{A}^T)$, and also leads to the item clustering indicator matrix \mathbf{C} which is approximately row-orthogonal.

Proof: By following the proof of theorem 2, minimizing J_c is equivalent to simultaneously optimizing:

$$\max_{\mathbf{B}} \text{tr}(\mathbf{B}^T\mathbf{A}\mathbf{A}^T\mathbf{B}) \text{ s.t. } \mathbf{B}^T\mathbf{B} = \mathbf{I}, \quad (29)$$

$$\max_{\mathbf{C}} \text{tr}(2\mathbf{C}^T\mathbf{A}^T\mathbf{A}\mathbf{C}), \text{ and} \quad (30)$$

$$\min_{\mathbf{C}} \text{tr}(\mathbf{C}^T\mathbf{C}\mathbf{A}^T\mathbf{A}\mathbf{C}^T\mathbf{C}) \equiv \min_{\mathbf{C}} \text{tr}(\mathbf{C}\mathbf{C}^T\mathbf{C}\mathbf{C}^T). \quad (31)$$

Eq. 29 is equivalent to eq. 16 and leads to the clustering of related features. And optimizing eq. 30 and Eq. 31 simultaneously is equivalent to:

$$\max_{\mathbf{C}} \text{tr}(\mathbf{C}\tilde{\mathbf{A}}\mathbf{C}^T), \text{ and} \quad (32)$$

$$\min_{\mathbf{C}} \left(\underbrace{\sum_i (\check{\mathbf{c}}_i \check{\mathbf{c}}_i^T)^2}_{j_{c1}} + \underbrace{\sum_{i \neq j} (\check{\mathbf{c}}_i \check{\mathbf{c}}_j^T)^2}_{j_{c2}} \right) \quad (33)$$

$$\text{s.t. } \mathbf{0} \leq \mathbf{C} \leq \Upsilon_{\mathbf{C}},$$

where $\tilde{\mathbf{A}}$ denotes the item affinity matrix, $\check{\mathbf{c}}_i$ denotes the i -th row of \mathbf{C} , and $\Upsilon_{\mathbf{C}}$ denotes the upperbound constraints on \mathbf{C} .

As in the proof of theorem 2, the most reasonable scenario in simultaneously optimizing eq. 32 and eq. 33 is by setting j_{c2} as small as possible and balancing j_{c1} with eq. 32. This leads to the clustering of similar items, and as j_{c2} is minimum, \mathbf{C} is approximately row-orthogonal. ■

D. No orthogonality constraint on both \mathbf{B} and \mathbf{C}

In this section we prove that applying the standard NMF to the feature-by-item data matrix eventually leads to the simultaneous feature and item clustering.

Theorem 4. Minimizing the following objective

$$\min_{\mathbf{B}, \mathbf{C}} J_d(\mathbf{B}, \mathbf{C}) = \frac{1}{2} \|\mathbf{A} - \mathbf{B}\mathbf{C}\|_F^2 \quad (34)$$

$$\text{s.t. } \mathbf{B} \geq \mathbf{0}, \mathbf{C} \geq \mathbf{0},$$

leads to the feature clustering indicator matrix \mathbf{B} and the item clustering indicator matrix \mathbf{C} which are approximately column- and row-orthogonal respectively.

Proof: By following the proof of theorem 2, minimizing J_d is equivalent to simultaneously optimizing:

$$\max_{\mathbf{B}, \mathbf{C}} \text{tr}(\mathbf{B}^T\mathbf{A}\mathbf{C}^T), \text{ and} \quad (35)$$

$$\min_{\mathbf{B}, \mathbf{C}} \text{tr}(\mathbf{B}^T\mathbf{B}\mathbf{C}\mathbf{C}^T). \quad (36)$$

By substituting $\mathbf{B} = \mathbf{A}\mathbf{C}^T$ and $\mathbf{C} = \mathbf{B}^T\mathbf{A}$ into the above equations, we get:

$$\max_{\mathbf{B}} \text{tr}(\mathbf{B}^T \hat{\mathbf{A}} \mathbf{B}), \text{ and} \quad (37)$$

$$\min_{\mathbf{B}} \text{tr}(\mathbf{B}^T \mathbf{B} \mathbf{B}^T \mathbf{A} \mathbf{A}^T \mathbf{B}) \equiv \min_{\mathbf{B}} \text{tr}(\mathbf{B}^T \mathbf{B} \mathbf{B}^T \mathbf{B}) \quad (38)$$

for feature clustering, and:

$$\max_{\mathbf{C}} \text{tr}(\mathbf{C} \tilde{\mathbf{A}} \mathbf{C}^T), \text{ and} \quad (39)$$

$$\min_{\mathbf{C}} \text{tr}(\mathbf{C} \mathbf{A}^T \mathbf{A} \mathbf{C}^T \mathbf{C} \mathbf{C}^T) \equiv \min_{\mathbf{C}} \text{tr}(\mathbf{C} \mathbf{C}^T \mathbf{C} \mathbf{C}^T) \quad (40)$$

for item clustering. Therefore, minimizing J_d is equivalent to simultaneously optimizing:

$$\max_{\mathbf{B}} \text{tr}(\mathbf{B}^T \hat{\mathbf{A}} \mathbf{B}), \quad (41)$$

$$\min_{\mathbf{B}} \left(\sum_i (\mathbf{b}_i^T \mathbf{b}_i)^2 + \sum_{i \neq j} (\mathbf{b}_i^T \mathbf{b}_j)^2 \right), \quad (42)$$

$$\max_{\mathbf{C}} \text{tr}(\mathbf{C} \tilde{\mathbf{A}} \mathbf{C}^T), \text{ and} \quad (43)$$

$$\min_{\mathbf{C}} \left(\sum_i (\tilde{\mathbf{c}}_i \tilde{\mathbf{c}}_i^T)^2 + \sum_{i \neq j} (\tilde{\mathbf{c}}_i \tilde{\mathbf{c}}_j^T)^2 \right), \quad (44)$$

$$\text{s.t. } \mathbf{0} \leq \mathbf{B} \leq \mathbf{Y}_B, \text{ and } \mathbf{0} \leq \mathbf{C} \leq \mathbf{Y}_C,$$

which will lead to the feature clustering indicator matrix \mathbf{B} and the item clustering indicator matrix \mathbf{C} that are approximately column- and row-orthogonal respectively. ■

III. UNIPARTITE AND DIRECTED GRAPH CASES

The affinity matrix \mathbf{W} induced from a unipartite (undirected) graph is a symmetric matrix, which is a special case of the rectangular affinity matrix \mathbf{A} . Therefore, by following the discussion in section II, it can be shown that the standard NMF applied to \mathbf{W} leads to the clustering indicator matrix which is almost orthogonal.

The affinity matrix \mathbf{V} induced from a directed graph is an asymmetric square matrix. Since columns and rows of \mathbf{V} correspond to the same set of vertices with the same order, as the clustering problem is concerned, \mathbf{V} can be replaced by $\mathbf{V} + \mathbf{V}^T$ which is a symmetric matrix. Then the standard NMF can be applied to this matrix to get the clustering indicator matrix which is almost orthogonal.

IV. RELATED WORKS

Ding et al. [8] provides the theoretical analysis on the equivalences between orthogonal NMF to K -means clustering for both rectangular data matrices and symmetric matrices. However as their proofs utilize the zero gradient conditions, the hidden assumptions (setting the Lagrange multipliers to zeros) are not revealed there. Actually it can be easily shown that their approach is the KKT conditions applied to the unconstrained version of eq. 2. Thus there is no guarantee that minimizing eq. 2 by using the zero gradient conditions leads to the stationary point located on the nonnegative orthant as required by the objective.

Applying the standard NMF to the symmetric matrix leads to almost orthogonal matrix was previously proven by Ding

et al. [16]. But due to the used approach, the theorem cannot be extended to the rectangular matrices which so far are the usual form of the data (practical applications of NMF seemed exclusively for rectangular matrices). Therefore, their results cannot be used to explain the abundant experimental results that show the power of the standard NMF in clustering, latent factors identification, learning the parts of objects, and producing sparse matrices even without explicit sparsity constraint [5].

V. CONCLUSION

By using the strict KKT optimality conditions, we showed that even without explicitly imposing orthogonality nor sparsity constraint NMF produces approximately column-orthogonal basis matrix and row-orthogonal coefficient matrix which lead to the simultaneous feature and item clustering. This result, therefore, gives the theoretical explanation on some experimental results that show the power of the standard NMF as a clustering tool which are reported to be better than the spectral methods [1] and K -means algorithm [2].

REFERENCES

- [1] W. Xu, X. Liu and Y. Gong, "Document clustering based on non-negative matrix factorization," Proc. ACM SIGIR, pp. 267-73, 2003.
- [2] J. Kim and H. Park, "Sparse nonnegative matrix factorization for clustering," CSE Technical Reports, Georgia Institute of Technology, 2008.
- [3] T. Li and C. Ding, "The relationships among various nonnegative matrix factorization methods for clustering," Proc. ACM 6th Int'l Conf. on Data Mining, pp. 362-71, 2006.
- [4] C. Ding, T. Li, and M.I. Jordan, "Convex and Semi-Nonnegative Matrix Factorizations," IEEE Transactions on Pattern Analysis and Machine Intelligence, pp. 45-55, 2010.
- [5] D. Lee and H. Seung, "Learning the parts of objects by non-negative matrix factorization," Nature, 401(6755), pp. 788-91, 1999.
- [6] S.Z. Li, X.W. Hou, H.J. Zhang, and Q.S. Cheng, "Learning spatially localized, parts-based representation," Proc. IEEE Comp. Soc. Conf. on Computer Vision and Pattern Recognition, pp. 207-12, 2001.
- [7] P.O. Hoyer, "Non-negative Matrix Factorization with Sparseness Constraints," The Journal of Machine Learning Research, Vol. 5, pp. 1457-69, 2004.
- [8] C. Ding, T. Li, W. Peng, and H. Park, "Orthogonal nonnegative matrix t-factorizations for clustering," Proc. 12th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining, pp. 126-35, 2006.
- [9] I.S. Dhillon and S. Sra, "Generalized nonnegative matrix approximation with Bregman divergences," UTCS Technical Reports, The University of Texas at Austin, 2005.
- [10] J. Kim and H. Park, "Toward faster nonnegative matrix factorization: A new algorithm and comparisons," Proc. 8th IEEE International Conference on Data Mining, pp. 353-62, 2008.
- [11] H. Kim and H. Park, "Nonnegative matrix factorization based on alternating nonnegativity constrained least squares and active set method," SIAM. J. Matrix Anal. & Appl., Vol. 30(2), pp. 713-30, 2008.
- [12] D. Kim, S. Sra, and I.S. Dhillon, "Fast projection-based methods for the least squares nonnegative matrix approximation problem," Stat. Anal. Data Min., Vol. 1(1), pp. 38-51, 2008.
- [13] L. Grippo and M. Sciandrone, "On the convergence of the block nonlinear Gauss-Seidel method under convex constraints," Operation Research Letters, Vol. 26, pp. 127-36, 2000.
- [14] I. S. Dhillon, Y. Guan, and B. Kulis, "Weighted Graph Cuts without eigenvectors: A multilevel approach," IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 29, No. 11, pp. 1944-57, 2007.
- [15] P.H. Calamai and J.J. More, "Projected gradient methods for linearly constrained problems," Mathematical Programming, Vol. 39, pp. 93-116, 1987.
- [16] C. Ding, X. He, and H.D. Simon, "On the equivalence of nonnegative matrix factorization and spectral clustering," Proc. SIAM Data Mining Conference, pp. 606-10, 2005.