On the Equivalence of Nonnegative Matrix Factorization and Spectral Clustering

Chris Ding[*] Xiaofeng He[*] Horst D. Simon[*]

Abstract

Current nonnegative matrix factorization (NMF) deals with $X = FG^T$ type. We provide a systematic analysis and extensions of NMF to the symmetric $W = HH^T$, and the weighted $W = HSH^T$. We show that (1) $W = HH^T$ is equivalent to Kernel K-means clustering and the Laplacian-based spectral clustering. (2) $X = FG^T$ is equivalent to simultaneous clustering of rows and columns of a bipartite graph. Algorithms are given for computing these symmetric NMFs.

1 Introduction

Standard factorization of a data matrix uses singular value decomposition (SVD) as widely used in principal component analysis (PCA). However, for many dataset such as images and text, the original data matrices are nonnegative. A factorization such as SVD contain negative entries and thus has difficulty for interpretation. Nonnegative matrix factorization (NMF) [7, 8] has many advantages over standard PCA/SVD based factorizations. In contrast to cancellations due to negative entries in matrix factors in SVD based factorizations, the nonnegativity in NMF ensures factors contain coherent parts of the original data (images).

Let $X = (\mathbf{x}_1, \ldots, \mathbf{x}_n) \in \mathbb{R}_+^{p\times n}$ be the data matrix of nonnegative elements. In image processing, each column is a 2D gray level of the pixels. In text mining, each column is a document.

The NMF factorizes X into two nonnegative matrices,

$$X \approx FG^T, \tag{1}$$

where $F = (\mathbf{f}_1, \cdots, \mathbf{f}_k) \in \mathbb{R}_+^{p\times k}$ and $G = (\mathbf{g}_1, \cdots, \mathbf{g}_k) \in \mathbb{R}_+^{n\times k}$. k is a pre-specified parameter. The factorizations are obtained by the least square minimization. A number of researches on further developing NMF computational methodologies [12, 11, 10], and applications on text mining [9, 14, 11].

Here we study NMF in the direction of data clustering. The relationship between NMF and vector quantization, especially the difference, are discussed by Lee and Seung [7] as a motivation for NMF. The clustering aspect of NMF is also studied in [14, 10].

In this paper, we provide a systematic analysis and extensions of NMF and show that NMF is equivalent to Kernel K-means clustering and Laplacian-based spectral clustering.
(1) We study the symmetric NMF of

$$W \approx HH^T \tag{2}$$

where W contains the pairwise similarities or the Kernals. We show that this is equivalent to K-means type clustering and the Laplacian based spectral clustering.
(2) We generalize this to bipartite graph clustering i.e., simultaneously clustering rows and columns of the rectangular data matrix. The result is the standard NMF.
(3) We extend NMFs to weighted NMF:

$$W \approx HSH^T. \tag{3}$$

(4) We derive the algorithms for computing these factorizations.

Overall, our study provides a comprehensive look at the nonnegative matrix fractorization and spectral clustering.

2 Kernel K-means clustering and Symmetric NMF

K-means clustering is one of most widely used clustering method. Here we first briefly introduce the K-means using spectral relaxation [15, 3]. This provides the necessary background information, notations and paves the way to the nonnegative matrix factorization approach in §2.1.

K-means uses K prototypes, the centroids of clusters, to characterize the data. The objective function is to minimize the sum of squared errors,

$$J_{\mathrm{K}} = \sum_{k=1}^{K} \sum_{i\in C_k} ||\mathbf{x}_i - \mathbf{m}_k||^2 = c_2 - \sum_k \frac{1}{n_k} \sum_{i,j\in C_k} \mathbf{x}_i^T\mathbf{x}_j, \tag{4}$$

where $X = (\mathbf{x}_1, \cdots, \mathbf{x}_n)$ is the data matrix, $\mathbf{m}_k = \sum_{i\in C_k} \mathbf{x}_i/n_k$ is the centroid of cluster C_k of n_k points,

[*] Lawrence Berkeley National Laboratory, University of California, Berkeley, CA 94720. {chqding, xhe, hdsimon}@lbl.gov

and $c_2 = \sum_i ||\mathbf{x}_i||^2$. The solution of the clustering is represented by K non-negative indicator vectors:

$$H = (\mathbf{h}_1, \cdots, \mathbf{h}_K),\ \mathbf{h}_k^T \mathbf{h}_\ell = \delta_{k\ell}. \tag{5}$$

where

$$\mathbf{h}_k = (0, \cdots, 0, \overbrace{1, \cdots, 1}^{n_k}, 0, \cdots, 0)^T / n_k^{1/2} \tag{6}$$

Now Eq.(4) becomes $J_\mathrm{K} = \mathrm{Tr}(X^T X) - \mathrm{Tr}(H^T X^T X H)$. The first term is a constant. Let $W = X^T X$. Thus $\min J_\mathrm{K}$ becomes

$$\max_{H^T H = I,\ H \geq 0} J_\mathrm{W}(H) = \mathrm{Tr}(H^T W H). \tag{7}$$

The pairwise similarity matrix $W = X^T X$ is the standard inner-product linear Kernel matrix. It can be extended to any other kernels. This is done using a nonlinear transformation (a mapping) to the higher dimensional space

$$\mathbf{x}_i \to \phi(\mathbf{x}_i)$$

The clustering objective function under this mapping, with the help of Eq.(4), can be written as

$$\min J_\mathrm{K}(\phi) = \sum_i ||\phi(\mathbf{x}_i)||^2 - \sum_k \frac{1}{n_k} \sum_{i,j \in C_k} \phi(\mathbf{x}_i)^T \phi(\mathbf{x}_j). \tag{8}$$

The first term is a constant for a given mapping function $\phi(\cdot)$ and can be ignored. Let the kernel matrix $W_{ij} = \phi(\mathbf{x}_i)^T \phi(\mathbf{x}_j)$. Using the cluster indicators H, the kernel K-means clustering is reduced to Eq.(7).

The objective function in Eq.(7). can be symbolically written as

$$J_\mathrm{W} = \sum_k \frac{1}{n_k} \sum_{i,j \in C_k} w_{ij} = \mathrm{Tr}(H^T W H). \tag{9}$$

Kernel K-means aims at maximizing within-cluster similarities. The advantage of Kernel K-means is that it can describe data distributions more complicated than Gaussion distributions.

2.1 Nonnegative relaxation of Kernel K-means

We show the spectral relaxation of Eq.(7) can be solved by the matrix factorization

$$W \approx HH^T, \quad H \geq 0. \tag{10}$$

Casting this in an optimization framework, an appropriate objective function is

$$\min_{H \geq 0} J_1 = ||W - HH^T||^2, \tag{11}$$

where the matrix norm $||A||^2 = \sum_{ij} a_{ij}^2$, the Frobeneus norm.

Theorem 1. NMF of $W = HH^T$ is equivalent to Kernel K-means clustering with the the strict orthogonality relation Eq.(5) relaxed.

Proof. The maximization of Eq.(7) can be written as

$$\begin{aligned} H &= \arg \min_{H^T H = I,\ H \geq 0} -2\mathrm{Tr}(H^T W H) \\ &= \arg \min_{H^T H = I,\ H \geq 0} -2\mathrm{Tr}(H^T W H) + ||H^T H||^2 \\ &= \arg \min_{H^T H = I,\ H \geq 0} (||W||^2 - 2\mathrm{Tr}(H^T W H) + ||H^T H||^2) \\ &= \arg \min_{H^T H = I,\ H \geq 0} ||W - HH^T||^2. \end{aligned} \tag{12}$$

Relaxing the orthogonality $H^T H = I$ completes the proof. □

A question immediately arises. Will the orthogonality $H^T H = I$ get lost?

Theorem 2. NMF of $W = HH^T$ retains near-orthogonality of $H^T H = I$.

proof. One can see $\min J_1$ is equivalent to

$$\max_{H \geq 0} \mathrm{Tr}(H^T W H), \tag{13}$$

$$\min_{H \geq 0} ||H^T H||^2. \tag{14}$$

The first objective recovers the original optimization objective Eq.(7). We concentrate on 2nd term. Note

$$||H^T H||^2 = \sum_{\ell k} (H^T H)_{\ell k}^2 = \sum_{\ell \neq k} (\mathbf{h}_\ell^T \mathbf{h}_k)^2 + \sum_k (\mathbf{h}_k^T \mathbf{h}_k)^2.$$

Minimizing the first term is equivalent to enforcing the orthogonality among $\mathbf{h}_\ell$: $\mathbf{h}_\ell^T \mathbf{h}_k \approx 0$. Minimizing the second term is equivalent to

$$\min\ ||\mathbf{h}_1||^4 + \cdots + ||\mathbf{h}_K||^4. \tag{15}$$

However, H cannot be all zero, otherwise we would have $\mathrm{Tr}(H^T W H) = 0$. More precisely, since $W \approx HH^T$, we have

$$\sum_{ij} w_{ij} \approx \sum_{ij} (HH^T)_{ij} = \sum_{kij} h_{ik} h_{jk} = \sum_k |\mathbf{h}_k|^2, \tag{16}$$

where $|\mathbf{h}| = \sum_i |h_i| = \sum_i h_i$ is the L_1 of vector $\mathbf{h}$. This means $||\mathbf{h}_\ell|| > 0$. Therefore, optimization of Eq.(14) with consideration of Eq.(13) implies H has near orthogonal columns, i.e.,

$$\mathbf{h}_\ell^T \mathbf{h}_k \approx \begin{cases} 0 & \text{if } l \neq k, \\ ||\mathbf{h}_k||^2 > 0 & \text{if } l = k. \end{cases} \tag{17}$$

Furthermore, given that $||\mathbf{h}_\ell|| > 0$, minimization of Eq.(15) will naturely lead to the column equalization condition

$$||\mathbf{h}_1|| = \cdots = ||\mathbf{h}_k||. \tag{18}$$

□

The near-orthogonality of columns of H is important for data clustering. An exact orthogonality implies that each row of H can have only one nonzero element, which implies that each data object belongs only to 1 cluster. This is hard clustering, such as in K-means . The near-orthogonality condition relaxes this a bit, i.e., each data object could belong fractionally to more than 1 cluster. This is soft clustering. A completely non-orthogonality among columns of H does not have a clear clustering interpretation.

3 Bipartite graph K-means clustering and NMF

A large number of datasets in today's applications are in the form of rectangular nonnegative matrix, such as the word-document association matrix in text mining or the DNA gene expression profiles. These types of datasets can be conveniently represented by a bipartitie graph where the graph adjacency matrix B contains the association among row and column objects, which is the input data matrix $B = X$.

The above kernel K-means approach can be easily extended to bipartitie graph. Let $\mathbf{f}_k$ be the indicator for the k-th row cluster. $\mathbf{f}_k$ has the same form of $\mathbf{h}_k$ as in Eq.(6). Put them together we have the indicator matrix $F = (\mathbf{f}_1, \cdots, \mathbf{f}_k)$. Analogously, we define the indicator matrix $G = (\mathbf{g}_1, \cdots, \mathbf{g}_k)$ for column-clusters.

We combine the row and column nodes together as

$$W = \begin{pmatrix} 0 & B \\ B^T & 0 \end{pmatrix}, \; \mathbf{h}_k = \frac{1}{\sqrt{2}} \begin{pmatrix} \mathbf{f}_k \\ \mathbf{g}_k \end{pmatrix}, \; H = \frac{1}{\sqrt{2}} \begin{pmatrix} F \\ G \end{pmatrix} \tag{19}$$

where the factor $1/\sqrt{2}$ allows the simultaneous normalizations $\mathbf{h}_k^T \mathbf{h}_k = 1$, $\mathbf{f}_k^T \mathbf{f}_k = 1$, and $\mathbf{g}_k^T \mathbf{g}_k = 1$. The Kernel K-means type clustering objective function becomes

$$\max_{F,G} J_{\mathrm{K}}^{\mathrm{B}} = \frac{1}{2} \mathrm{Tr} \begin{pmatrix} F \\ G \end{pmatrix}^T \begin{pmatrix} 0 & B \\ B^T & 0 \end{pmatrix} \begin{pmatrix} F \\ G \end{pmatrix} \tag{20}$$

Following the Kernel K-means clustering optimization of Eq.(11), F, G are obtained by minimizing

$$J_1^B = \left\| \begin{pmatrix} 0 & B \\ B^T & 0 \end{pmatrix} - \begin{pmatrix} F \\ G \end{pmatrix} \begin{pmatrix} F \\ G \end{pmatrix}^T \right\|^2 \tag{21}$$

Note that

$$J_1^B = 2||B - FG^T||^2 + ||F^T F||^2 + ||G^T G||^2.$$

Minimization of the second and third terms, following the analysis of Eqs.(14, 17), is equivalent to enforcing the orthogonality contraints

$$\mathbf{f}_\ell^T \mathbf{f}_k \approx 0, \;\; \mathbf{g}_\ell^T \mathbf{g}_k \approx 0, \;\; \ell \neq k. \tag{22}$$

Thus $\min J_1^B$ becomes

$$\min_{F,G} J_2 = ||B - FG^T||^2 \tag{23}$$

subject to the orthogonality constraints in Eq.(22). This is the form of NMF proposed by Lee and Seung. Therefore, we have shown that the bipartite graph kernel K-means clustering leads to NMF for nonnegative rectangular matrix.

The orthogonality constraints play an important role. If the orthogonality holds vigourously, we can show directly that NMF of $||B - FG^T||^2$ is equivalent to simultaneously K-means clustering of rows and columns of B.

To show this, we first prove it for clustering of columns of $B = (\mathbf{b}_1, \cdots, \mathbf{b}_n) = (b_{ij})$, under the normalization

$$\sum_{i=1}^{p} b_{ij} = 1, \; \sum_{r=1}^{k} g_{ir} = 1, \; \sum_{j=1}^{p} f_{jk} = 1. \tag{24}$$

For any given data, normalization of X is first applied. The second normalization indicates that the i-th row of G are the posterior probabilities for $\mathbf{b}_i$ belonging to k clusters; they should add up to 1. The 3rd normalization $\sum_j (\mathbf{f}_k)_j$ is a standard length normalize of the vector $\mathbf{f}_k$. Since we approximate $B \approx FG^T$, the normalization of FG^T should be consistent with the normalization of B. Indeed, $\sum_{i=1}^{p} (FG^T)_{ij} = \sum_{i=1}^{p} \sum_{r=1}^{k} F_{ir} G_{jr} = 1$, consistent with $\sum_i B_{ij} = 1$.

With this self-consistent normalization and imposing strict orthogonalilty on G: $\mathbf{g}_\ell^T \mathbf{g}_k = 0$, $\ell \neq k$, we call the resulting factorizaton as Orthogonal NMF.

Theorem 3. Orthogonal NMF is identical to K-means clustering.

Proof. We have

$$J_2 = \|B - FG^T\|^2 = \sum_{i=1}^{n} \left\| \mathbf{b}_i - \sum_{k=1}^{\kappa} g_{ik} \mathbf{f}_k^T \right\|^2, \tag{25}$$

because the Frobenious norm of a matrix is equivalent to sum of column norms. Now due to the normalization of G, we have

$$\begin{aligned} J_2 &= \left\| \sum_{k=1}^{\kappa} g_{ik} (\mathbf{b}_i - \mathbf{f}_k) \right\|^2 = \sum_{k=1}^{\kappa} g_{ik}^2 ||\mathbf{b}_i - \mathbf{f}_k||^2 \\ &= \sum_{k=1}^{\kappa} g_{ik} ||\mathbf{b}_i - \mathbf{f}_k||^2 = \sum_{k=1}^{\kappa} \sum_{i \in C_k} ||\mathbf{b}_i - \mathbf{f}_k||^2 \end{aligned}$$

The 2nd equality is due to the orthogonality condition of G which implies that on each row of G, only one element is nonzero. This also implies $g_{ik} = 0, 1$. Thus $g_{ik}^2 = g_{ik}$, giving the 3rd equality. Thus the final result is the standard K-means clustering of $\{\mathbf{b}_i\}_{i=1}^n$. $\mathbf{f}_k$ is the cluster centroid. ■

In NMF of optimizing $\min \|B - FG^T\|$, rows and columns are treated in equal footing, since we could equally write $J_2 = \|B^T - GF^T\|$. Thus clustering of columns of B is happening *simultaneously* as the clustering the rows of B.

4 Spectral clustering and NMF

In recent years spectral clustering using the Laplacian of the graph emerges as solid approach for data clustering. Here we focus on the clustering objective functions. There are three clustering objective functions. the Ratio Cut [6], the Normalized Cut [13], and the MinMax Cut [4]. We are interested in the multi-way clustering objective functions,

$$J = \sum_{1\le p<q\le K} \frac{s(C_p, C_q)}{\rho(C_p)} + \frac{s(C_p, C_q)}{\rho(C_q)} = \sum_{k=1}^{K} \frac{s(C_k, \bar{C}_k)}{\rho(C_k)} \tag{26}$$

$$\rho(C_k) = \begin{cases} |C_k| & \text{for} \quad \text{Ratio Cut} \\ \sum_{i\in C_k} d_i & \text{for} \quad \text{Normalized Cut} \\ s(C_k, C_k) & \text{for} \quad \text{MinMax Cut} \end{cases} \tag{27}$$

where $\bar{C}_k$ is the complement of subset C_k in graph G.

Here we show that the minimization of these objective functions can be equivalently carried out via the nonnegative matrix factorizations. The proof follows the multi-way spectral relaxation of Ratio Cut clustering objective function[1], and Normalized Cut and MinMax Cut clustering objective functions[5].

For concreteness and simplicity, here we outline the proof for the case of Normalized Cut. Let $\mathbf{h}_k$ be the cluster indicators as in Eq.(6). One can easily see that

$$s(C_k, \bar{C}_k) = \sum_{i\in C_k}\sum_{j\in \bar{C}_k} w_{ij} = \mathbf{h}_\ell^T (D - W)\mathbf{h}_\ell \tag{28}$$

and $\sum_{i\in C_k} d_i = \mathbf{h}_\ell^T D \mathbf{h}_\ell$. Define the scaled cluster indicator vector $\mathbf{z}_\ell = D^{1/2}\mathbf{h}_\ell / ||D^{1/2}\mathbf{h}_\ell||$, which obey the orthonormal condition $\mathbf{z}_\ell^T \mathbf{z}_k = \delta_{\ell k}$, or $Z^T Z = I$, where $Z = (\mathbf{z}_1, \cdots, \mathbf{z}_K)$. Substituting into the Normalized Cut objective function, we have

$$J_{\rm NC} = \sum_{\ell=1}^{K} \frac{\mathbf{h}_\ell^T (D - W)\mathbf{h}_\ell}{\mathbf{h}_\ell^T D \mathbf{h}_\ell} = \sum_{\ell=1}^{K} \mathbf{z}_\ell^T (I - \widetilde{W})\mathbf{z}_\ell$$

where

$$\widetilde{W} = D^{-1/2} W D^{-1/2}. \tag{29}$$

The first term is a constant. Thus the minimization problem becomes

$$\max_{Z^T Z = I,\ Z\ge 0} \mathrm{Tr}(Z^T \widetilde{W} Z) \tag{30}$$

Now if we remove the restriction that Z takes the discrete values, and allow Z to be any continuous value, this is the multi-way spectral relaxation of the Normalized Cut[5]. The solution of the maximization problem is given by the k principal eigenvectors of the matrix $\widetilde{W}$.

The point of departure from spectral relaxation of Normalized Cut is to recognize that the maximization problem of Eq.(30) is identical to the maximization problem of Eq.(7) with the same orthogonality constraints. Following the same discussions there, the maximization problem of Eq.(30) is equivalent to

$$\min_{Z\ge 0} J_3 = ||\widetilde{W} - ZZ^T||^2, \tag{31}$$

Once the solution $\widehat{Z}$ for $\min J_3(Z)$ is obtained, we can recover H by optimizing

$$\min_{H\ge 0} \sum_\ell \left\| \hat{\mathbf{z}}_\ell - \frac{D^{1/2}\mathbf{h}_\ell}{||D^{1/2}\mathbf{h}_\ell||} \right\|^2, \tag{32}$$

The solution can be easily shown to be $\mathbf{h}_k = D^{-1/2}\hat{\mathbf{z}}_k$. This gives the solution to Normalized Cut via the NMF approach. This can be extended to RatioCut and MinMaxCut. Summarizing, we have proved that
Theorem 4. NMF is equivalent to spectral clustering.

5 Weighted Nonnegative $W = HSH^T$

In both Kernel K-means and spectral clustering, we assume the pairwise similarity matrix W are semi positive definite. For kernel matrices, this is true. But a large number of similarity matrices is nonnegative, but not s.p.d. This motivates us to propose the following more general NMF:

$$\min_H J_5 = ||W - HSH^T||^2, \tag{33}$$

When the similarity matrix W is indefinite, W has negative eigenvalues. HH^T will not provide a good approximation, because HH^T can not obsorb the subspace associated with negative eigenvalues. However, HSH^T can obsorb subspaces associated with both positive and negative eigenvalues, i.e., the indefiniteness of W is passed on to S. This distinction is well-known in linear algebra where matrix factorizations have Cholesky factorization $A = LL^T$ if matrix A is s.p.d. Otherwise, one does $A = LDL^T$ factorization, where the diagonal matrix D takes care of the negeative eigenvalues.

The second reason for nonnegative $W = HSH^T$ is that the extra degrees of freedom provided by S allow H to be more closer to the form of cluster indicators. This benefits occur for both s.p.d. W and indefinite W.

The third reason for nonnegative $W = HSH^T$ is that S provides a good characterization of the quality of the clustering. Generally speaking, given a fixed W and number of clusters κ, the residue of the matrix approximation $J_5^{\text{opt}} = \min ||W - HSH^T||^2$ will be smaller than $J_1^{\text{opt}} = \min ||W - HH^T||^2$. Futhermore, the K-by-K matrix S has a special meaning. To see this, let us assume H are vigorous cluster indicators, i.e., $H^T H = I$. Setting the derivative $\partial J_5 / \partial S = 0$, we obtain

$$S = H^T W H, \text{ or } S_{\ell k} = \mathbf{h}_\ell^T W \mathbf{h}_k = \frac{\sum_{i \in C_\ell} \sum_{j \in C_k} w_{ij}}{\sqrt{n_\ell n_k}} \tag{34}$$

S represents properly normalized within-cluster sum of weights ($\ell = k$) and between-cluster sum of weights ($\ell \neq k$). For this reason, we call this type of NMF as weighted NMF. The usefulness of weighted NMF is that if the clusters are well-separated, we would see the off-diagonal elemens of S are much smaller than the diagonal elements of S.

The fourth reason is the consistency between standard $W = HH^T$ and $B = FG^T$. Since we can define a kernel as $W = B^T B$. Thus the factorization $W \approx B^T B \approx (FG^T)^T (FG^T) = G(F^T F) G^T$. Let $S = F^T F$, we obtain the weighted NMF.

6 Algorithms for computing symmetric NMF

We briefly outline the algorithms for computing symmetric factorizations $W = HH^T$ and $W = HSH^T$. For $W = HH^T$, the updating rule is

$$H_{ik} \leftarrow H_{ik} \left(1 - \beta + \beta \frac{(WH)_{ik}}{(HH^T H)_{ik}} \right). \tag{35}$$

where $0 < \beta \leq 1$. In practice, we find $\beta = 1/2$ is a good choice. A faster algorithm[1]

$$H \leftarrow \max \left(WH(H^T H)^{-1}, 0 \right). \tag{36}$$

can be used in the first stage of the iteration. Algorithmic issues of symmtric `NMF` is also studied in [2].

For weighted `NMF` $W = HSH^T$, the update rules are

$$S_{ik} \leftarrow S_{ik} \frac{(H^T W H)_{ik}}{(H^T H S H^T H)_{ik}}. \tag{37}$$

$$H_{ik} \leftarrow H_{ik} \left(1 - \beta + \beta \frac{(WHS)_{ik}}{(HSH^T HS)_{ik}} \right). \tag{38}$$

Acknowledgement. This work is supported partially by U.S. Department of Energy, Office of Science, under contract DE-AC03-76SF00098 through a LBNL LDRD grant.

References

[1] P.K. Chan, M.Schlag, and J.Y. Zien. Spectral k-way ratio-cut partitioning and clustering. *IEEE Trans. CAD-Integrated Circuits and Systems*, 13:1088–1096, 1994.

[2] M. Catral, L. Han, M. Neumann, and R.J. Plemmons. On reduced rank nonnegative matrix factorizations for symmetric matrices. *Linear Algebra and Its Applications*, to appear.

[3] C. Ding and X. He. K-means clustering and principal component analysis. *LBNL-52983. Int'l Conf. Machine Learning (ICML2004)*, 2004.

[4] C. Ding, X. He, H. Zha, M. Gu, and H. Simon. A min-max cut algorithm for graph partitioning and data clustering. *Proc. IEEE Int'l Conf. Data Mining*, 2001.

[5] M. Gu, H. Zha, C. Ding, X. He, and H. Simon. Spectral relaxation models and structure analysis for k-way graph clustering and bi-clustering. *Penn State Univ Tech Report CSE-01-007*, 2001.

[6] L. Hagen and A.B. Kahng. New spectral methods for ratio cut partitioning and clustering. *IEEE. Trans. on Computed Aided Desgin*, 11:1074–1085, 1992.

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

[8] D.D. Lee and H.S. Seung. Algorithms for non-negatvie matrix factorization. *Advances in Neural Information Processing Systems*, 13. 2001.

[9] S.Z. Li, X. Hou, H. Zhang, Q. Cheng. Learning spatially localized, parts-based representation. In *Proc. IEEE Computer Vision and Pattern Recognition*, pp:207–212, Hawaii, 2001.

[10] T. Li and S. Ma. IFD: Iterative feature and data clustering. In *Proc SIAM Int'l conf. on Data Mining (SDM 2004)*, pages 472–476, 2004.

[11] V. P. Pauca, F. Shahnaz, M.W. Berry, and R. J. Plemmons. Text mining using non-negative matrix factorization. In *Proc SIAM Int'l conf. Data Mining (SDM 2004)*, pages 452–456, 2004.

[12] F. Sha, L.K. Saul, and D.D. Lee. Multiplicative updates for nonnegative quadratic programming in support vector machines. *Advances in Neural Information Processing Systems 15*, pp:1041–1048. MIT Press, Cambridge, MA, 2003.

[13] J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE. Trans. on Pattern Analysis and Machine Intelligence*, 22:888–905, 2000.

[14] W. Xu, X. Liu, and Y. Gong. Document clustering based on non-negative matrix factorization. In *Proc. SIGIR pp.267–273, 2003.*

[15] H. Zha, C. Ding, M. Gu, X. He, and H.D. Simon. Spectral relaxation for K-means clustering. *Advances in Neural Information Processing Systems 14 (NIPS'01)*, pages 1057–1064, 2002.

[1] For the nonsymmetric NMF of Eq.(1), the algorithm is $F \leftarrow \max \left(BG(G^T G)^{-1}, 0 \right)$, $G \leftarrow \max \left(B^T F (F^T F)^{-1}, 0 \right)$. Without nonnegative constraints, these algorithms converge respectively to *global* optimal solutions of J_1 in Eq.(11) and J_2 in Eq.(23).