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

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

The pairwise similarity matrix $W = X^TX$ 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

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

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

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

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 factorization of Kernel K-means

We show that the optimization 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. $W = HH^T$ factorization is equivalent to Kernel K-means clustering with the strict orthogonality relation Eq.(5) relaxed.

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

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

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

If the nonnegativity condition is relaxed (ignored), the solution to H are the k eigenvectors with the largest eigenvalues and orthogonality is retained. Now we keep the nonnegativity of H. Will the orthogonality get lost?

Theorem 2. $W = HH^T$ factorization retains H orthogonality approximately.

proof. One can see that $\min J_1 = ||W - HH^T||^2$ is equivalent to

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

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

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

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

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 $\mathrm{Tr}(H^TWH) = 0$. More precisely, since $W \approx HH^T$,

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

where $|\mathbf{h}| = \sum_i |h_i| = \sum_i h_i$ is the L_1-norm of vector $\mathbf{h}$. This means $||\mathbf{h}_\ell|| > 0$. Therefore, optimization of Eq.(14) with the nonzero constraint Eq.(16) 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, minimization of Eq.(15) with the nonzero constraint Eq.(16) leads to the column equalization condition

$$||\mathbf{h}_1|| \approx ||\mathbf{h}_2|| \approx \cdots \approx ||\mathbf{h}_k||. \tag{18}$$

This assures the approximate balance of cluster sizes. □

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 (a form of a contingency table), such as the word-document association matrix in text mining or the DNA gene expression profiles. This kind of datasets can be represented by a bipartitie graph; the input data matrix B is the graph adjacency matrix contains the association among row and column objects.

The relation between symmetric NMF and K-means can be easily extended to bipartitie graph; here we simultaneously cluster the rows and columns of B. Let the rows of B be $(\mathbf{y}_1, \cdots, \mathbf{y}_k) = B^T$ and the indicator matrix $F = (\mathbf{f}_1, \cdots, \mathbf{f}_k)$ for these row-clusters. Let the columns of B be $(\mathbf{x}_1, \cdots, \mathbf{x}_k) = B$ and the indicator matrix $G = (\mathbf{g}_1, \cdots, \mathbf{g}_k)$ for these column-clusters. According Eq.(7), simultaneous row and column K-means clustering becomes simultaneous optimizations:

$$\max_{F^TF=I,\ F\geq 0} \mathrm{Tr}\ F^TBB^TF, \quad \max_{G^TG=I,\ G\geq 0} \mathrm{Tr}\ G^TB^TBG. \tag{19}$$

This simultaneous clustering can be formulated more compactly. We combine the row and column nodes together as $W = \begin{pmatrix} 0 & B \\ B^T & 0 \end{pmatrix}$, $\mathbf{h}_k = \frac{1}{\sqrt{2}}\begin{pmatrix} \mathbf{f}_k \\ \mathbf{g}_k \end{pmatrix}$, $H = \frac{1}{\sqrt{2}}\begin{pmatrix} F \\ G \end{pmatrix}$ 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 K-means type clustering objective function is

$$\max_{\substack{F^TF\,=\,I;\\ G^TG\,=\,I;\\ F,\,G\,\geq\,0}} J_2 = \frac{1}{2}\mathrm{Tr}\begin{pmatrix} F \\ G \end{pmatrix}^T \begin{pmatrix} 0 & B \\ B^T & 0 \end{pmatrix}\begin{pmatrix} F \\ G \end{pmatrix} = \mathrm{Tr}\ F^TBG. \tag{20}$$

The solution of this quadratic optimization is given by the first K eigenvectors of $\begin{pmatrix} 0 & B \\ B^T & 0 \end{pmatrix}\begin{pmatrix} \mathbf{f}_k \\ \mathbf{g}_k \end{pmatrix} = \lambda_k \begin{pmatrix} \mathbf{f}_k \\ \mathbf{g}_k \end{pmatrix}$ which is equivalent to $B\mathbf{g}_k = \lambda_k\mathbf{f}_k, B^T\mathbf{f}_k = \lambda_k\mathbf{g}_k$. Upon substition, the equations are $B^TB\mathbf{g}_k = \lambda_k^2\mathbf{g}_k, BB^T\mathbf{f}_k = \lambda_k^2\mathbf{f}_k$, which define the solutions of optimizations in Eq.(19). This proves that Eq.(20) is the objective for simultaneous row and column K-means clustering.

Standard K-means maximizes the within-cluster similarities. For bipartite graph, J_2 minimizes the bipartite within-cluster similarities $s(R_k, C_k)$,

$$J_2 = \sum_k \frac{s(R_k, C_k)}{(|R_k|\ |C_k|)^{1/2}}, \ s(R_k, C_k) = \sum_{i\in R_k}\sum_{j\in C_k} b_{ij}. \tag{21}$$

Clearly, without nonnegative constraint, the solution is given by the first K left and right singular vectors of the SVD of B. We focus on the nonnegative case.

Theorem 3. The simultaneous row and column K-means clustering J_2 is equivalent to the following optimization problem,

$$\min_{\substack{F^TF\,=\,I;\\ G^TG\,=\,I;\\ F,\,G\,\geq\,0}} ||B - FG^T||^2. \tag{22}$$

Proof. We have $\max_{F,G} J_2 \Rightarrow \min_{F,G} -\mathrm{Tr}(F^TBG) \Rightarrow \min_{F,G} ||B||^2 - 2\mathrm{Tr}(F^TBG) + \mathrm{Tr}(F^TFG^TG)$. Here we add two constants: $||B||^2$ and $\mathrm{Tr}(F^TFG^TG) = \mathrm{Tr}I = k$. The objective function is identical to $||B - FG^T||^2$. □

Now we relax vigourous orthogonality contraints $F^TF = I; G^TG = I$ to the approximate orthogonality. Therefore, NMF is equivalent to K-means clustering with relaxed orthogonality contraints.

The orthogonality constraints play an important role. In the above, we assume both F and G are orthogonal. If one of them is orthogonal, we can explicitly write $||B - FG^T||^2$ as a K-means clustering objective function. To show this, we impose the normalization

$$\sum_{i=1}^{p} b_{ij} = 1, \ \sum_{r=1}^{k} g_{ir} = 1, \ \sum_{j=1}^{p} f_{jk} = 1. \tag{23}$$

For any given data B, column L_1 normalization of B is 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 $\sum_i B_{ij} = 1$. Let $B = (\mathbf{x}_1, \cdots, \mathbf{x}_n)$ and $\mathbf{g}_\ell^T\mathbf{g}_k = 0,\ \ell \neq k$, we have

Theorem 4. NMF with orthogonal G is identical to K-means clustering of the columns of B.

Proof. We have

$$J_2 = \|B - FG^T\|^2 = \sum_{i=1}^{n} \left\| \mathbf{x}_i - \sum_{k=1}^{K} g_{ik}\mathbf{f}_k \right\|^2 . \tag{24}$$

Due to the row normalization of G, each term becomes

$$\left\| \sum_{k=1}^{K} g_{ik}(\mathbf{x}_i - \mathbf{f}_k) \right\|^2 = \sum_{k=1}^{K} g_{ik}^2 ||\mathbf{x}_i - \mathbf{f}_k||^2 = \sum_{k=1}^{K} g_{ik} ||\mathbf{x}_i - \mathbf{f}_k||^2$$

The orthogonality condition of G implies that in each row of G, only one element is nonzero and $g_{ik} = 0, 1$. Thus $g_{ik}^2 = g_{ik}$. Summing over i, $J_2 = \sum_{k=1}^{K} \sum_{i \in C_k} \|\mathbf{x}_i - \mathbf{f}_k\|^2$, which is the K-means clustering with $\mathbf{f}_k$ as the cluster centroid. ■

4 Spectral clustering and NMF

In recent years spectral clustering using the Laplacian of the graph emerges as solid approach for data clustering (see references in [2]). Here we focus on the spectral clustering objective functions. There are three objectives: the Ratio Cut [6], the Normalized Cut [13], and the MinMax Cut [4]. We are interested in the multi-way clustering objective functions,

$$J = \sum_{1 \le p < q \le K} \frac{s(C_p, C_q)}{\rho(C_p)} + \frac{s(C_p, C_q)}{\rho(C_q)} = \sum_{k=1}^{K} \frac{s(C_k, \bar{C}_k)}{\rho(C_k)} \tag{25}$$

$$\rho(C_k) = \begin{cases} |C_k| & \text{for} \quad \text{Ratio Cut} \\ \sum_{i \in C_k} d_i & \text{for} \quad \text{Normalized Cut} \\ s(C_k, C_k) & \text{for} \quad \text{MinMax Cut} \end{cases} \tag{26}$$

where $\bar{C}_k$ is the complement of subset C_k in graph G, $s(A, B) = \sum_{i \in A} \sum_{j \in B} w_{ij}$, and $d_i = \sum_j w_{ij}$.

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[5] of NormalizedCut and MinMaxCut. We focus on Normalized Cut.

Theorem 5. Normalized Cut using pairwise similarity matrix W is equivalent to Kernel K-means clustering with the kernel matrix

$$\widetilde{W} = D^{-1/2} W D^{-1/2}. \tag{27}$$

where $D = \mathrm{diag}(d_1, \cdots, d_n)$.

Corallary 5. Normalized Cut using similarity W is equivalent to nonnegative matrix factorization

$$\min_{H \ge 0} J_3 = ||\widetilde{W} - HH^T||^2. \tag{28}$$

Proof of Theorem 5. Let $\mathbf{h}_k$ be the cluster indicators as in Eq.(6). One can easily see that

$$s(C_k, \bar{C}_k) = \sum_{i \in C_k} \sum_{j \in \bar{C}_k} w_{ij} = \mathbf{h}_\ell^T (D - W) \mathbf{h}_\ell \tag{29}$$

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

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

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

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

This is identical to the Kernel K-means clustering of Eq.(7). Once the solution $\widehat{Z}$ is obtained, we can recover H by optimizing

$$\min_{H \ge 0} \sum_{\ell} \left\| \hat{\mathbf{z}}_\ell - \frac{D^{1/2}\mathbf{h}_\ell}{||D^{1/2}\mathbf{h}_\ell||} \right\|^2 . \tag{31}$$

The exact solution are $\mathbf{h}_k = D^{-1/2}\hat{\mathbf{z}}_k$, or $H = D^{-1/2} Z$. Thus row i of Z is multiplied by a constant $d_i^{-1/2}$. The relative weight across different cluster in the same row remain same. Thus H represents the same clustering as Z does. □

Theorem 5 show the spectral clustering are directly related to Kernel K-means clustering, which is equivalence to NMF by Theorem 1. Thus NMF, Kernel K-means clustering and spectral clustering are unified in a simple way: they are different prescriptions of the same problem with slightly different constraints.

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:

$$\min_{H} J_5 = ||W - HSH^T||^2, \tag{32}$$

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 K, 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

$$S = H^T W H, \text{ or } S_{\ell k} = \mathbf{h}_\ell^T W \mathbf{h}_k = \frac{\sum_{i \in C_\ell} \sum_{j \in C_k} w_{ij}}{\sqrt{n_\ell n_k}} \tag{33}$$

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 Symmetric NMF Algorithms

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

$$H_{ik} \leftarrow H_{ik} \left(1 - \beta + \beta \frac{(WH)_{ik}}{(HH^T H)_{ik}} \right). \tag{34}$$

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

$$H \leftarrow \max \left(WH(H^T H)^{-1}, 0 \right). \tag{35}$$

can be used in the first stage of the iteration. Algorithmic issues of symmtric NMF is also studied in [1].

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

$$S_{ik} \leftarrow S_{ik} \frac{(H^T W H)_{ik}}{(H^T HSH^T H)_{ik}}. \tag{36}$$

$$H_{ik} \leftarrow H_{ik} \left(1 - \beta + \beta \frac{(WHS)_{ik}}{(HSH^T HS)_{ik}} \right). \tag{37}$$

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

References

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

[2] C. Ding. A tutorial on spectral clustering. *Int'l Conf. Machine Learning (ICML2004)*, 2004.

[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. In T. G. Dietterich and V. Tresp, editors, *Advances in Neural Information Processing Systems*, volume 13. The MIT Press, 2001.

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

[10] T. Li and S. Ma. IFD: Iterative feature and data clustering. In *Proceedings of the 2004 SIAM International conference on Data Mining (SDM 2004)*, pages 472–476, Lake Buena Vista, Florida, 2004.

[11] V. P. Pauca, F. Shahnaz, M. W. Berry, R. J. Plemmons. Text mining using non-negative matrix factorization. In *Proceedings of the 2004 SIAM International conference on Data Mining (SDM 2004)*, pages 452–456, Lake Buena Vista, Florida, 2004.

[12] F. Sha, L. K. Saul, and D. D. Lee. Multiplicative updates for nonnegative quadratic programming in support vector machines. In S. Thrun S. Becker and K. Obermayer, editors, *Advances in Neural Information Processing Systems 15*, pages 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, Y. Gong. Document clustering based on non-negative matrix factorization. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 267–273, Toronto, Canada, 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.(22).