Title:
On the Equivalence of Nonnegative Matrix Factorization and K-means - Spectral Clustering

Author:
Ding, Chris
He, Xiaofeng
Simon, Horst D.
Jin, Rong

Publication Date:
06-24-2008

Publication Info:
Lawrence Berkeley National Laboratory

Permalink:
http://escholarship.org/uc/item/6km2b4fz

On the Equivalence of Nonnegative Matrix Factorization and K-means — Spectral Clustering

Chris Ding[*] Xiaofeng He[*] Horst D. Simon[*] Rong Jin[†]

December 4, 2005

Abstract

We provide a systematic analysis of nonnegative matrix factorization (NMF) relating to data clustering. We generalize the usual $X = FG^T$ decomposition to the symmetric $W = HH^T$ and $W = HSH^T$ decompositions. 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. We emphasizes the importance of orthogonality in NMF and soft clustering nature of NMF. These results are verified with experiments on face images and newsgroups.

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 contains negative entries and is difficult to interpret for some applications. In contrast, nonnegative matrix factorization (NMF) [18, 19] restricts the entries in matrix factors to be nonnegative. NMF has been shown recently to be useful for many applications in environment [25], chemometrics [29], pattern recognition [20], multimedia [6], text mining [31, 26] and DNA gene expressions [3]. This is also extended to classification [27]. A number of stuides focus on further developing NMF computational methodologies [15, 22, 26, 5, 21].

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 $\mathbf{x}_i$ is a 2D array of pixels gray level. In text processing, 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.

NMF can be traced back to 1970s (communication from Gene Golub) and has been studied by Paatero [25, 29]. The work of Lee and Seung [18, 19] brought much attention to NMF in machine learning and data mining communities. There appears to have some confusions, however. Lee and Seung emphasizes[18] that NMF factors $\mathbf{f}_k$ contain coherent parts of the original data (images), for example a nose or an eye. Later experiments [16, 20] do not support the parts-of-whole interpretation of NMF. In fact, Hoyer[16] and Li, et al[20] specifically propose sparsification schemes to achieve the parts-of-whole pictures.

[*] Lawrence Berkeley National Laboratory, University of California, Berkeley, CA 94720.

[†] Department of Computer Science and Engineering, Michigan State University, East Lansing, MI 48824.

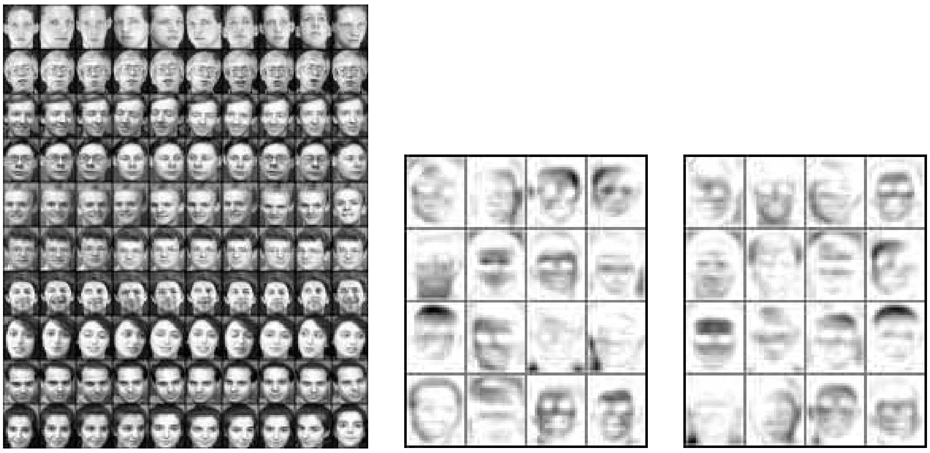

Figure 1: Left: ORL face image dataset. Middle/Right: Computed basis vectors $F = (\mathbf{f}_1, \cdots, \mathbf{f}_{16})$ for 2 runs with random initialization.

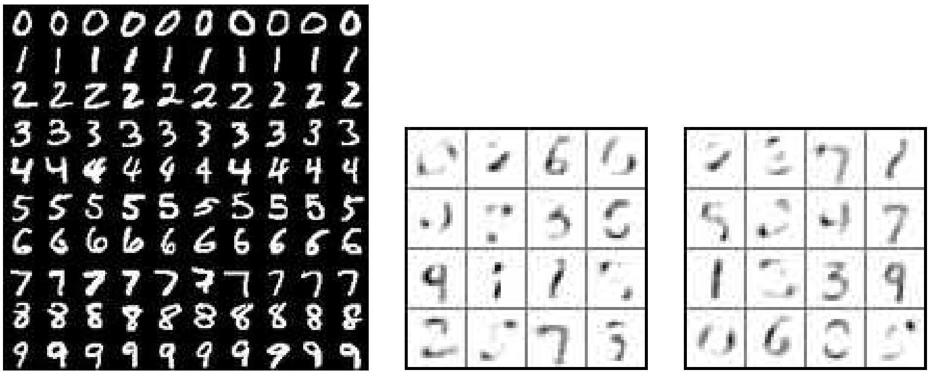

Figure 2: Left: Digits image dataset. Middle/Right: Computed basis $F = (\mathbf{f}_1, \cdots, \mathbf{f}_{16})$ for 2 runs with random initialization.

To further clarify this issue, we perform NMF on image datasets using the multiplicative algorithm. Results for the AT&T ORL dataset are shown in Fig.1 and on Bell Lab digits dataset are shown in Fig.2. These results agree with Hoyer[16] and Li, et al[20].

Furthermore, we note that the basis images are close to the original images in some sense. In face dataset (Fig. 1), many basis images clearly resemble original faces. In digits dataset (Fig. 2), many basis images clearly resemble 1-9 digits. Intuitively, these basis vectors look like *representatives* of clusters.

This observation motivates us to analyze the clustering aspect of NMF. We will show in §4 that the basis images are actually the *cluster centroids* in the K-means (and fuzzy K-means) clustering. In their original paper[18], Lee and Seung emphasizes the difference between NMF and vector quantization (which is identical to the K-means clustering). A number of studies[31, 21, 3], however, show *empirically* the usefulness of NMF for data clustering.

In the rest of this paper, we provide a comprehensive analysis of NMF from the clustering point of view. We begin in §2 with symmetric NMF, i.e., $W = HH^T$ where W is a square matrix of pairwise similarities (could be a kernel). We show that $W = HH^T$ is equivalent to Kernel K-means clustering. In §3, we show that NMF of $X = FG^T$ is equivalent to simultaneous clustering of rows and columns of X.

In §4, we show that $X = FG^T$ factorization is identical to K-means clustering under the G orthogonality (among columns of G), and is identical to fuzzy K-means clustering under approximate G orthogonality. In §5, NMF is shown to be equivalent to Laplacian-matrix based spectral clustering and bipartite graph clustering, with W and X replaced by their proper rescaled counter parts: $D^{-1/2}WD^{-1/2}$, $D_r^{-1/2}XD_c^{-1/2}$.

In §6, we propose symmetric $W = HSH^T$ as a more effective factorization than $W = HH^T$. In §7 we generalize the effective multiplicative update algorithm[19] for solving $X = FG^T$ to the case of symmetric NMF of $W = HH^T$ and $W = HSH^T$. There are a number of researches on further developing NMF computational methodologies [27, 26, 21].

In §8, we illustrate some concepts and results of our analysis via experiments on internet newsgroups. Brief summary are given in §9. Overall, our work emphasizes the role of orthogonality, the simultaneous clustering of row and columns for $X \approx FG^T$, and the soft clustering nature of NMF. Preliminary results of this work has appeared in conference[11].

2 Symmetric NMF and Kernel K-means clustering

Here we show that given a symmetric nonnegative matrix W of pairwise similarities, the nonnegative $W = HH^T$ factorization is equivalent to Kernel K-means clustering. In §3, this is generalize to rectangular nonnegative matrices.

K-means clustering is one of most widely used clustering method. Here we review their spectral relaxation formalism[32, 10]. This provides background information and paves the way to symmetric NMF. K-means uses the K cluster centroids, $\mathbf{m}_k = \sum_{i \in C_k} \mathbf{x}_i / n_k$, to characterize the data $X = (\mathbf{x}_1, \cdots, \mathbf{x}_n)$. The objective function is minimizing the sum of squared errors,

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

The solution of the clustering is represented by K non-negative indicator vectors: $H = (\mathbf{h}_1, \cdots, \mathbf{h}_K)$, where the indicator for cluster C_k is

$$\mathbf{h}_k = (0, \cdots, 0, \overbrace{1, \cdots, 1}^{n_k}, 0, \cdots, 0)^T / n_k^{1/2} \tag{3}$$

where $n_k = |C_k|$. Clearly $H^T H = I$. Now Eq.(2) 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_{K} becomes

$$\max_{H^T H = I,\ H \geq 0} J_{\mathrm{W}}(H) = \mathrm{Tr}(H^T W H). \tag{4}$$

The continuous solution for H are given by the principal components of X via an orthogonal transformation[10]. 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

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

The clustering objective function under this mapping, with the help of Eq.(2), can be written as

$$\min J_{\mathrm{K}}(\phi) = \sum_i ||\phi(\mathbf{x}_i)||^2 - \sum_k \frac{1}{n_k} \sum_{i,j \in C_k} \phi(\mathbf{x}_i)^T \phi(\mathbf{x}_j). \tag{5}$$

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.(4). More broadly speaking, W can be any nonnegative pairwise similarity.

In the maximization formulation of Eq.(4), there are two constraints: nonnegativity and orthogonality. If we relax (ignore) nonnegativity and retain orthogonality, the (continuous) solution for H are given by the principal eigenvectors of W (a suitable orthogonal transformation[10] can restore the nonnegativity). If we retain nonnegativity and relax orthogonality, we come to the nonnegative factorization below.

2.1 Nonnegative $W = HH^T$ factorization

We show that the optimization of Eq.(4) can be solved by the symmetrix NMF:

$$W \approx HH^T, \quad H \geq 0. \tag{6}$$

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

$$\min_{H\geq 0} J_1 = ||W - HH^T||^2, \tag{7}$$

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

Theorem 1. $W = HH^T$ factorization is equivalent to Kernel K-means clustering with the strict orthogonality relation Eq.(2) relaxed.

Proof. The maximization of Eq.(4) can be written as

$$\begin{aligned} H &= \arg\min_{H^TH=I,\ H\geq 0} -2\mathrm{Tr}(H^TWH) = \arg\min_{H^TH=I,\ H\geq 0} ||W||^2 - 2\mathrm{Tr}(H^TWH) + ||H^TH||^2 \\ &= \arg\min_{H^TH=I,\ H\geq 0} ||W - HH^T||^2. \end{aligned} \tag{8}$$

In Eq.(8), we add two constants, $||W||^2$ and $||H^TH||^2 = \kappa$. Now 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?

Observation 2. $W = HH^T$ factorization retains approximate H-orthogonality.

proof. One can see that $\min J_1 = ||W - HH^T||^2$ is equivalent to (1) $\max_{H\geq 0} \mathrm{Tr}(H^TWH)$ and (2) $\min_{H\geq 0} ||H^TH||^2$. The first objective is the original optimization objective Eq.(4). For the 2nd objective, we note $||H^TH||^2 = ||H^TH - I||^2 + 2||H||^2 - Tr(I)$. Since $W \approx HH^T$, $||H||^2 = \mathrm{Tr}(HH^T) \approx \mathrm{Tr}W$ is approximately constant. Thus the 2nd objective is reduced to $\min_{H\geq 0} ||H^TH - I||^2$ which ensures approximate H-orthogonality. □

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.

An illustrative example. We demonstrate NLR by a simple example. Fig.3 (left panel) shows a 2D example of 38 points. The similarity between $\mathbf{x}_i, \mathbf{x}_j$ is computed using Gaussian kernel $W_{ij} = \exp(-||\mathbf{x}_i -$

$\mathbf{x}_j||^2/2)$. The data has two dominant clusters. We set $k = 2$ The resulting cluster indicators $\mathbf{h}_1, \mathbf{h}_2$ are shown in Fig.3 (right panel).

If we do a hard clustering by assigning each data point x_i to the cluster $c = \arg\max_k H_{ik}$, we get points in regions $\{A, C, E\}$ as one cluster, and points in regions $\{B, D\}$ as another cluster. This can be see clearly from Fig.3 (right panel). This result is identical to a K-means clustering result.

However, we see that the magnitudes of $\mathbf{h}_1, \mathbf{h}_2$ on points in C are very close to each other. This indicates that a partial (soft) cluster assignment would be more appropriate. Furthermore, the magnitude of $\mathbf{h}_1, \mathbf{h}_2$ on points in E, D are even smaller, indicating they do not belong to either of the dominant clusters. In fact, all points in C, E, D can be considered as outliers. In general, if $\sum_k H_{ik}$ is far below the average value, we may consider $\mathbf{x}_i$ as an outlier. This can be rigorously quantified.

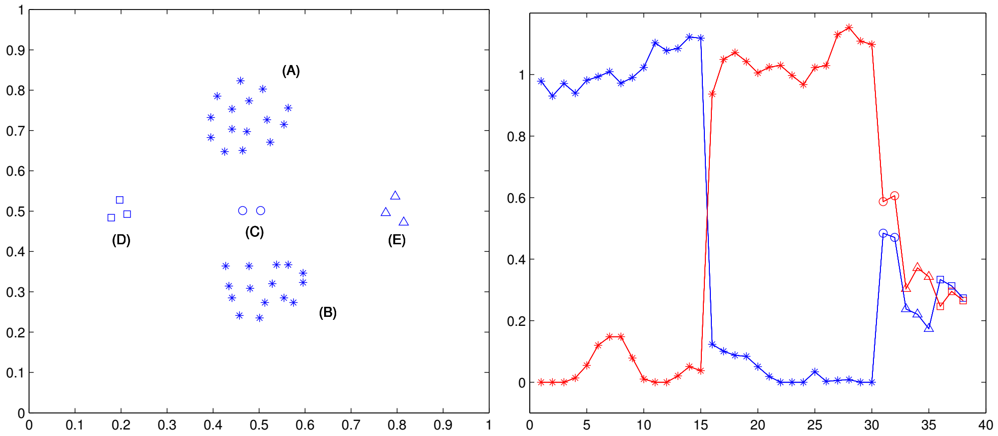

Figure 3: Left: A 2D dataset of 38 data points. Right: Their $H = (\mathbf{h}_1, \mathbf{h}_2)$ values are shown as blue and red curves. Datapoints are ordered by regions $\{B, A, C, E, D\}$, where $B = \{x_1, \cdots, x_{15}\}$, $A = \{x_{16}, \cdots, x_{30}\}$, $C = \{x_{31}, x_{32}\}$, $E = \{x_{33}, x_{34}, x_{35}\}$, $D = \{x_{36}, x_{37}, x_{38}\}$.) H values for points in regions $\{C, E, D\}$ indicate they are fractionally assigned to clusters.

3 NMF and K-means clustering of a bipartite graph

Many application datasets are in the form of rectangular nonnegative matrix, such as a big matrix collecting a set of images or the word-document association matrix for a document set. This rectangle matrix can be viewed as the adjacency matrix $B = (B_{ij})$ of a bipartitie graph, B_{ij} contains the association between row i and column j.

The kernel K-means approach of §2 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.(3). 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.

Let $s(R_k, C_\ell) = \sum_{i \in R_k} \sum_{j \in C_\ell} b_{ij}$ be the sum of weights(similarity) between row cluster R_k and column cluster C_ℓ. K-means type clustering maximizes the within-cluster similarities $s(R_k, C_k)$,

$$\max_{\substack{F^T F = I; \\ G^T G = I; \\ F, G \geq 0}} J_2 = \sum_k \frac{s(R_k, C_k)}{(|R_k| \; |C_k|)^{1/2}} = \text{Tr}(F^T B G). \tag{9}$$

This bipartie graph clustering objective can be obtained more formally: We combine the row and column nodes together as

$$W = \begin{pmatrix} 0 & B \\ B^T & 0 \end{pmatrix}, \; \mathbf{h}_k = \frac{1}{\sqrt{2}} \begin{pmatrix} \mathbf{f}_k \\ \mathbf{g}_k \end{pmatrix}, \; H = \frac{1}{\sqrt{2}} \begin{pmatrix} F \\ G \end{pmatrix} \tag{10}$$

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 of Eq.(4) becomes $\text{Tr}(H^T W H) = 2\text{Tr}(F^T B G)$. On the other hand, J_2 reduces to the standard K-means of Eq.(4) when $W = B$ and row/column become the same.

zzz

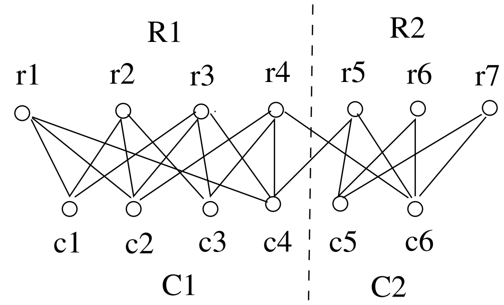

Figure 4: A bipartite graph with r-nodes and c-nodes. The dashed line indicates a possible clustering. K-means type clustering maximizes within-cluster similarities $s(R_1, C_1), s(R_2, C_2)$. Spectral clustering minimizes between-cluster similarities $s(R_1, C_2), s(R_2, C_1)$ and maximizes within-cluster similarities $s(R_1, C_1), s(R_2, C_2)$.

Theorem 3. The K-means bipartie graph clustering problem is equivalent to the following optimization problem,

$$\min_{\substack{F^T F = I; \\ G^T G = I; \\ F, G \geq 0}} J_2 = ||B - F G^T||^2. \tag{11}$$

From Eq.(3), we have $J_2 = \text{Tr} F^T B G$. The optimization problem can be written

$$\min_{\substack{F^T F = I; \\ G^T G = I; \\ F, G \geq 0}} -2\text{Tr}(F^T B G) \quad \Longrightarrow \quad \min_{\substack{F^T F = I; \\ G^T G = I; \\ F, G \geq 0}} ||B||^2 - 2\text{Tr}(F^T B G) + \text{Tr}(F^T F G^T G)$$

In the second equation, we add two constants: $||B||^2$ and $\text{Tr}(F^T F G^T G) = \text{Tr} I = k$. The objective function is identical to $||B - F G^T||^2$. Now we relax orthogonality constraints $F^T F = I; G^T G = I$ to the approximate orthogonality. Thus NMF is equivalent to K-means clustering with relaxed orthogonality constraints. □

Once again the orthogonality plays a crucial role. Here we show that in fact orthogonality are automatically enforced in NMF.

Observation 4. $X = F G^T$ factorization retains G, F orthogonality approximately.
proof. One can see that $J_{\text{NMF}} = ||B||^2 - 2\text{Tr}(F^T B G) + \text{Tr}(F^T F)(G^T G)$ is equivalent to

$$\max_{F, G \geq 0} \text{Tr}(F^T B G), \tag{12}$$

$$\min_{F, G \geq 0} \text{Tr}(F^T F G^T G) \tag{13}$$

The first objective recovers the original optimization objective Eq.(4). We concentrate on 2nd term. Note

$$\text{Tr}(F^T F G^T G) = \text{Tr}(F^T F - I)(G^T G - I) + \text{Tr}(F^T F) + \text{Tr}(G^T G) - \text{Tr}(I).$$

Because $FG^T \approx X$, the scale of F, G are constrained. Thus $\mathrm{Tr}(F^TF) = ||F||^2$ and $\mathrm{Tr}(G^TG) = ||G||^2$ are approximately constants. Therefore, the minimization problem of Eq.(13) becomes minimization of $\mathrm{Tr}(F^TF - I)(G^TG - I) = \sum_{k\ell}(F^TF - I)_{k\ell}(G^TG - I)_{k\ell}$, which is bounded from above by

$$\sqrt{\sum_{k\ell}(F^TF - I)^2_{k\ell}}\sqrt{\sum_{k\ell}(G^TG - I)^2_{k\ell}} = ||F^TF - I|| \cdot ||G^TG - I||,$$

using Cauchy's inequality $\sum_i a_i b_i \le \sqrt{\sum_i a_i^2}\sqrt{\sum_i b_i^2}$. Minimization of a function $f(x)$ can be approximated by the minimization of the upper bound of $f(x)$. Thus, the minimization problem of Eq.(13) becomes

$$\min_{F,G\ge 0}(||F^TF - I|| \cdot ||G^TG - I||) \Rightarrow (\min_{F\ge 0}||F^TF - I||)\ (\min_{G\ge 0}||G^TG - I||) \tag{14}$$

Therefore, the orthogonality of F are approximately enforced. So does the orthogonality of G.

4 NMF as soft clustering

In §3 we show NMF is doing a simultaneous the rows and columns clustering of the rectangle matrix B. When the orthogonality constraints of F, G are relaxed, it can be viewed soft clustering because we can interprete the rows of F as posterior probability for row clustering and the rows of G as posterior probability for column clustering. In this section, we give a more detailed analysis on the soft clustering of NMF. We show that NMF is closely approximated by a fuzzy K-means clustering.

For simplicity, we normalize each column of the rectangle matrix B to 1. This is achieved by $X \equiv BD_r^{-1}$, where D_r is the diagonal matrix containing the row sums of B. We denote the NMF of this column normalized matrix $X = (\mathbf{x}_1, \cdots, \mathbf{x}_n)$ as

$$X \approx CH^T, \quad C = (\mathbf{c}_1, \cdots, \mathbf{c}_k),\ H = (\mathbf{h}_1, \cdots, \mathbf{h}_k).$$

and with the consistent normalization

$$\sum_{j=1}^{p} C_{jk} = 1,\ \sum_{r=1}^{k} H_{ir} = 1.$$

because $\sum_{i=1}^p (CH^T)_{ij} = \sum_{i=1}^p \sum_{r=1}^k C_{ir}H_{jr} = 1$, is consistent with $\sum_i X_{ij} = 1$.

The advantage of this normalization are: (1) H is a meaningful posterior probability because its row sums are all 1. (2) Because the centroids of a subset of $\{\mathbf{x}_l\}$ must be column-sum-to-1, columns of C can be interpreted as cluster centroids. (For these reasons, in this section we write $X = CH^T$ instead of $X = FG^T$.)

First we consider the case where $\mathbf{h}_k$ are mutually orthogonal. This corresponding to *hard clustering*, i.e., each object belongs to exactly one cluster and $\mathbf{c}_k$ are exactly the cluster centroids.

Theorem 5(Hard clustering). H-orthogonal NMF is identical to K-means clustering.

Proof. We have

$$J_{\text{NMF}} = \|X - CH^T\|^2 = \sum_{i=1}^{n}\left\|\mathbf{x}_i - \sum_{k=1}^{\kappa}\mathbf{c}_k h_{ik}\right\|^2 = \sum_{i=1}^{n}\left\|\sum_{k=1}^{\kappa} h_{ik}(\mathbf{x}_i - \mathbf{c}_k)\right\|^2 \tag{15}$$

The orthogonality condition implies that on each row of H, only one element is nonzero. Thus

$$J_{\text{NMF}} = \sum_{i=1}^{n}\sum_{k=1}^{\kappa} h_{ik}^2\|\mathbf{x}_i - \mathbf{c}_k\|^2. \tag{16}$$

Now since $h_{ik} = 0, 1$, this becomes the standard K-means clustering. □

Next, we consider the general case where $\mathbf{h}_k$ are not necessarily orthogonal, i.e., each object could belong to several clusters fractionally. We have the following results

Theorem 6(NMF Bound). J_{NMF} can be approximated by a fuzzy K-means clustering :

$$J_{\text{NMF}} \approx J_F^{(f=2)}, \quad |J_{\text{NMF}} - J_F^{(f=2)}| \leq J_K \tag{17}$$

where the fuzzy K-means clustering objective function[2] is

$$J_F^{(f)} = \sum_{i=1}^{n} \sum_{k=1}^{\kappa} h_{ik}^f \|\mathbf{x}_i - \mathbf{c}_k\|^2, \quad J_K = J_F^{(f=1)}, \quad \sum_{k=1}^{\kappa} h_{ik} = 1. \tag{18}$$

Proof. From Eq.(15), we have

$$J_{\text{NMF}} = \sum_{i=1}^{n} \left[\sum_{k=1}^{\kappa} h_{ik}^2 \|\mathbf{x}_i - \mathbf{c}_k\|^2 + \sum_{k \neq \ell} h_{ik} h_{i\ell} (\mathbf{x}_i - \mathbf{c}_k)^T (\mathbf{x}_i - \mathbf{c}_\ell) \right] = J_{\text{NMF}}^{(1)} + J_{\text{NMF}}^{(2)}.$$

The first term is $J_{\text{NMF}}^{(1)} = J_F^{(f=2)}$. To analyze the second term, we note that due to sign cancellation, we expect $J_{\text{NMF}}^{(2)}$ fluctuated around zero, i.e.,

$$J_{\text{NMF}}^{(2)} = \sum_{i=1}^{n} \sum_{k \neq \ell} h_{ik} h_{i\ell} (\mathbf{x}_i - \mathbf{c}_k)^T (\mathbf{x}_i - \mathbf{c}_\ell) \approx n \sum_{k \neq \ell} \langle h_{ik} h_{i\ell} (\mathbf{x}_i - \mathbf{c}_k)^T (\mathbf{x}_i - \mathbf{c}_\ell) \rangle \approx 0. \tag{19}$$

Using the inequality $2(\mathbf{x}_i - \mathbf{c}_k)^T (\mathbf{x}_i - \mathbf{c}_\ell) \leq ||\mathbf{x}_i - \mathbf{c}_k||^2 + ||\mathbf{x}_i - \mathbf{c}_\ell||^2$, we have

$$\begin{aligned} |J_{\text{NMF}}^{(2)}| &\leq \sum_{i=1}^{n} \sum_{k \neq \ell} h_{ik} h_{i\ell} |(\mathbf{x}_i - \mathbf{c}_k)^T (\mathbf{x}_i - \mathbf{c}_\ell)| & (20) \\ &\leq \frac{1}{2} \sum_{i=1}^{n} \sum_{k \neq \ell} h_{ik} h_{i\ell} (||\mathbf{x}_i - \mathbf{c}_k||^2 + ||\mathbf{x}_i - \mathbf{c}_\ell||^2) & (21) \\ &\leq \frac{1}{2} \sum_{i=1}^{n} \left[\sum_{k=1}^{\kappa} h_{ik} ||\mathbf{x}_i - \mathbf{c}_k||^2 + \sum_{\ell=1}^{\kappa} h_{i\ell} ||\mathbf{x}_i - \mathbf{c}_\ell||^2 \right] = J_K. & (22) \end{aligned}$$

Thus $|J_{\text{NMF}} - J_F| = |J_{\text{NMF}}^{(2)}| \leq J_K$. □

In fuzzy K-means clustering, cluster centroids are updated as

$$\mathbf{c}_k = \frac{\sum_{i=1}^n h_{ik}^f \mathbf{x}_i}{\sum_{i=1}^n h_{ik}^f}, \quad h_{ik} = \left[\sum_{\ell=1}^{\kappa} \left[\frac{\|\mathbf{x}_i - \mathbf{c}_k\|}{\|\mathbf{x}_i - \mathbf{c}_\ell\|} \right]^{\frac{2}{f-1}} \right]^{-1}. \tag{23}$$

Clearly $J_F^{(f=1)}$ reduces to standard K-means clustering where h_{ik} is discretized into $0, 1$.

To summerize, the soft clustering nature of NMF is close to $f = 2$ fuzzy K-means clustering and K-means clustering provides a good initialization for NMF.

5 NMF and Spectral clustering

In recent years spectral clustering using the Laplacian of the graph emerges as solid approach for data clustering [14, 28, 12, 24, 1, 8, 23] (complete references in [9]). Here we focus on the spectral clustering objective functions. There are three objectives: the Ratio Cut [14], the Normalized Cut [28], and the MinMax Cut [12]. We are interested in the multi-way clustering objective functions,

$$J = \sum_{1\le p<q\le K} \frac{s(C_p, C_q)}{\rho(C_p)} + \frac{s(C_p, C_q)}{\rho(C_q)} = \sum_{k=1}^{K} \frac{s(C_k, \bar{C}_k)}{\rho(C_k)} \tag{24}$$

$$\rho(C_k) = \begin{cases} |C_k| & \text{for} \quad \text{Ratio Cut} \\ \sum_{i\in C_k} d_i & \text{for} \quad \text{Normalized Cut} \\ s(C_k, C_k) & \text{for} \quad \text{MinMax Cut} \end{cases} \tag{25}$$

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[13] of Normalized-Cut and MinMaxCut. We focus on Normalized Cut.

Theorem 7. Normalized Cut using pairwise similarity matrix W is equivalent to Kernel K-means clustering with the kernel matrix

$$\widetilde{W} = D^{-1/2} W D^{-1/2}. \tag{26}$$

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

Corallary 8. Normalized Cut using similarity W is equivalent to nonnegative matrix factorization

$$\min_{H\ge 0} J_3 = ||\widetilde{W} - HH^T||^2. \tag{27}$$

Proof of Theorem 7. Let $\mathbf{h}_k$ be the cluster indicators as in Eq.(3). 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

$$J_{\text{NC}} = \sum_{\ell=1}^{K} \frac{\mathbf{h}_\ell^T (D - W)\mathbf{h}_\ell}{\mathbf{h}_\ell^T D \mathbf{h}_\ell} = \sum_{\ell=1}^{K} \mathbf{z}_\ell^T (I - \widetilde{W})\mathbf{z}_\ell \tag{29}$$

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

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

This is identical to the Kernel K-means clustering of Eq.(4) and can be solved by the nonnegative factorization of $\widetilde{W} = ZZ^T$. 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. □

We note besides nonnegative factorization solution for Eq.(30), another approach [30] is semi-definite programming by noting that $\text{Tr}(Z^T \widetilde{W} Z) = \text{Tr}(\widetilde{W} Z Z^T) = \text{Tr}(\widetilde{W} Y)$. This becomes maximization of a linear function of $f(Y)$, where $Y \equiv ZZ^T$ is restricted to semidefinite positive matrix. On another aspect, we have emphasized that NMF allows approximate posterior interpretation of H. A strict posterior interpretation of H can be guarantteed[17] by rigorously enforces the probability condition $\sum_k H_{ik} = 1$.

5.1 Spectral bipartite graph clustering and NMF

In the Laplacian based bipartite graph clustering[33, 7, 13] the clustering objective function for Normalized cut for $K = 2$ is

$$J^{\mathrm{B}}_{\mathrm{NC}} = \frac{s(R_1, C_2) + s(R_2, C_1)}{2s(R_1, C_1) + s(R_1, C_2) + s(R_2, C_1)} + \frac{s(R_1, C_2) + s(R_2, C_1)}{2s(R_2, C_2) + s(R_1, C_2) + s(R_2, C_1)} \tag{32}$$

Note $s(R_k, C_\ell)$ is defined near Eq.(9) Let D_r be a diagonal matrix containing row sums of B, and D_c be a diagonal matrix containing column sums of B. Then $2s(R_1, C_1)+s(R_1, C_2)+s(R_2, C_1) = \mathbf{f}_1^T D_r \mathbf{f}_1 + \mathbf{g}_1^T D_c \mathbf{g}_1$. Thus

$$J^{\mathrm{B}}_{\mathrm{NC}} = {\scriptstyle K} - \sum_{k=1}^{K} = \frac{2\mathbf{f}_k^T B \mathbf{g}_k}{\mathbf{f}_k^T D_r \mathbf{f}_k + \mathbf{g}_k^T D_c \mathbf{g}_k}$$

This can be immediately generalized to $K \geq 2$. Let

$$Z = (\mathbf{z}_1, \cdots, \mathbf{z}_k),\ \mathbf{z}_k = \begin{pmatrix} D_r^{1/2}\mathbf{f}_k \\ D_c^{1/2}\mathbf{g}_k \end{pmatrix} \Big/ \left\| \begin{matrix} D_r^{1/2}\mathbf{f}_k \\ D_c^{1/2}\mathbf{g}_k \end{matrix} \right\|,\ \mathbf{z}_k = \frac{1}{\sqrt{2}} \begin{pmatrix} \mathbf{x}_k \\ \mathbf{y}_k \end{pmatrix},\ Z = \frac{1}{\sqrt{2}} \begin{pmatrix} X \\ Y \end{pmatrix} \tag{33}$$

and set

$$\widetilde{B} = D_r^{-1/2} B D_c^{-1/2}. \tag{34}$$

we have

$$J^{\mathrm{B}}_{\mathrm{NC}} = K - \mathrm{Tr}\left[Z^T \begin{pmatrix} 0 & \widetilde{B} \\ \widetilde{B}^T & 0 \end{pmatrix} Z \right] = K - \frac{1}{2}\mathrm{Tr}\left[\begin{pmatrix} X \\ Y \end{pmatrix}^T \begin{pmatrix} 0 & \widetilde{B} \\ \widetilde{B}^T & 0 \end{pmatrix} \begin{pmatrix} X \\ Y \end{pmatrix} \right] = K - \mathrm{Tr}\ X^T \widetilde{B} Y. \tag{35}$$

Thus minimizing J_{NC} becomes $\max \mathrm{Tr}(X^T \widetilde{B} Y)$. Now, repeating the same analysis as in §2, the solution to $\min J^{\mathrm{B}}_{\mathrm{NC}}$ becomes

$$\min_{X^T X = I,\ Y^T Y = I} ||\widetilde{B} - XY^T||^2 \tag{36}$$

Let the solution for this NMF be $X = (\hat{\mathbf{x}}_1, \cdots, \hat{\mathbf{x}}_k)$, $Y = (\hat{\mathbf{y}}_1, \cdots, \hat{\mathbf{y}}_k)$. Once X, Y are obtained, we need to recover $F = (\mathbf{f}_1, \cdots, \mathbf{f}_k)$, $G = (\mathbf{g}_1, \cdots, \mathbf{g}_k)$, by optimizing

$$\min_{\{\mathbf{f}_k, \mathbf{g}_k\}} \left\| \begin{pmatrix} \hat{\mathbf{x}}_k \\ \hat{\mathbf{y}}_k \end{pmatrix} - \sqrt{2} \begin{pmatrix} D_r^{1/2}\mathbf{f}_k \\ D_c^{1/2}\mathbf{g}_k \end{pmatrix} \Big/ \left\| \begin{matrix} D_r^{1/2}\mathbf{f}_k \\ D_c^{1/2}\mathbf{g}_k \end{matrix} \right\| \right\|^2 \tag{37}$$

The solution can be easily shown to be

$$\mathbf{f}_k = D_r^{-1/2}\hat{\mathbf{x}}_k,\ \mathbf{g}_k = D_c^{-1/2}\hat{\mathbf{y}}_k, \tag{38}$$

This gives the spectral bipartite graph clustering via the NMF.

6 Nonnegative $W = HSH^T$ Factorization

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:

$$\min_H J_5 = ||W - HSH^T||^2, \tag{39}$$

where $W \in \mathbb{R}_+^{n\times n}$, $H \in \mathbb{R}_+^{n\times k}$, and $S \in \mathbb{R}_+^{k\times k}$ is not necessarily diagonal. 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 absorb 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 $J_1^{\text{opt}} = \min ||W - HH^T||^2$. Furthermore, 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

$$S = H^T W H, \text{ or } S_{\ell k} = \mathbf{h}_\ell^T W \mathbf{h}_k = \frac{\sum_{i\in C_\ell}\sum_{j\in C_k} w_{ij}}{\sqrt{n_\ell n_k}} \tag{40}$$

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 elements 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.

7 Algorithms for computing $W = HH^T$ and $W = HSH^T$

In this section, we brief outline the algorithms for computing the factorizations. Since all these are non-convex functions, the algorithms seek to find a local minima, similar in nature as K-means clustering and spectral clustering. The algorithms have the same style as the NMF multiplicative updating rules of Lee and Seung [18, 19].

7.1 Symmetric NMF $W = HH^T$

The updating rule is

$$H_{ik} \leftarrow H_{ik}\left(1 - \beta + \beta\frac{(WH)_{ik}}{(HH^T H)_{ik}}\right). \tag{41}$$

where $0 < \beta \leq 1$. In practical application, we find $\beta = 1/2$ is a good choice.

To derive the update rule of Eq.(41), we expand the objective function $J = ||W - HH^T||^2$ and obtain

$$\frac{\partial J}{\partial H} = -4WH + 4HH^T H$$

By the standard optimization theory via Lagrangian multipliers, the first order KKT complementarity slackness condition is $\left(\frac{\partial J}{\partial X_{ik}}\right) X_{ik} = 0$, evaluated at the local minima X_{ij}^*. For symmetric NMF, this leads to

$$(-4WH + 4HH^T H)_{ik} H_{ik} = 0. \tag{42}$$

This is a fixed point relation that H_{ik} must satisfy at convergence. There are many ways to iteratively update H_{ik}. We use gradient decent, $H_{ik} \leftarrow H_{ik} - \epsilon_{ik} \frac{\partial J}{\partial H_{ik}}$, and set $\epsilon_{ik} = \beta h_{ik}/(4HH^T H)_{ik}$ to obtain the update rule of Eq.(41), which satisfies the fixed point equation of Eq.(42).

Updating symmetric NMF using the nonsymmetric NMF rules of is studied in [4].

7.2 Weighted symmetric NMF $W = HSH^T$

The update rules are

$$S_{ik} \leftarrow S_{ik} \frac{(H^T W H)_{ik}}{(H^T HSH^T H)_{ik}}. \tag{43}$$

$$H_{ik} \leftarrow H_{ik} \left(1 - \beta + \beta \frac{(WHS)_{ik}}{(HSH^T HS)_{ik}} \right). \tag{44}$$

The derivation follows $W = HH^T$ case. The KKT complementarity condition $\left(\frac{\partial J}{\partial S_{ik}}\right) S_{ik} = 0$ gives

$$(-2H^T WH + 2H^T HSH^T H)_{ik} S_{ik} = 0 \tag{45}$$

for S_{ik}. Using gradient descent, we obtain update rule Eq.43, which satisfies the fixed point equation Eq.45.

Similary, the KKT complementarity condition $\left(\frac{\partial J}{\partial H_{ik}}\right) H_{ik} = 0$ gives

$$(-4WHS + 4HSH^T HS)_{ik} H_{ik} = 0, \tag{46}$$

for H_{ik}. Using gradient descent, we obtain update rule Eq.44, which satisfies the fixed point equation Eq.46.

8 Experiments on Internet Newsgroups

We perform experiments on Internet newsgroups articles to illustrate issues studied earlier regarding to orthogonality, soft clustering, etc. For comparison, we do K-means clustering on the same datasets. Our main purposes are to demonstrate (1) NMF performs substantially better than standard K-means . (2) NMF performs soft clustering of rows and columns simultaneously.

A 20-newsgroup dataset is obtained from CMU: www.cs.cmu.edu/afs/ cs/project/theo-11/www/naive-bayes.html. B contains the word-document matrix. 500 words are selected according to the mutual information. `tf.idf` term weighting is used. we normalize each document to 1 in L_2-norm. We use two sets of 5-newsgroup combinations:

A		B	
NG2:	comp.graphics	NG2:	comp.graphics
NG9:	rec.motorcycles	NG3:	comp.os.ms-windows
NG10:	rec.sport.baseball	NG8:	rec.autos
NG15:	sci.space	NG13:	sci.electronics
NG18:	talk.politics.mideast	NG19:	talk.politics.misc

Dataset A is moderately overlapping and dataset B is strongly overlapping. To accumulate sufficient statistics, we generate 5 random datasets for each 5-newsgroup combinations: 100 documents were randomly

Sample	1	2	3	4	5
Dataset A					
K-means	0.748	0.790	0.815	0.862	0.873
NMF	0.876	0.916	0.912	0.902	0.884
Dataset B					
K-means	0.531	0.491	0.576	0.632	0.697
NMF	0.612	0.590	0.608	0.652	0.711

Table 1: Clustering accuracy for K-means and NMF for 5 random samples.

sampled from each newsgroup. K-means and NMF are applied to these 5 random sampled datasets. For K-means , we run 10 trials with random starts and select the run with one with the best objective function value. NMF results uses K-means results as initial guess. We discuss results on $W = HSH^T$ and $B = FG^T$ separately.

8.1 Results using $W = HSH^T$

We compute cosine similarity between documents by setting $W = B^T B$. The experiments is done as explained above. We first discuss clustering accuracy, using the known class labels. The results for clustering accuracy for dataset A and B are listed in Table 1. We see that NMF consistently improve over K-means , sometimes quite substantially.

Next, we consider the orthogonality of H, whose importance w.r.t. clustering is emphasized in §2 - §4. We compute the normalized orthogonality, $(H^TH)_{\text{nm}} = D^{-1/2}(H^TH)D^{-1/2}$, where $D = \text{diag}(H^TH)$. Thus the diagonal is normalized to 1, and derivation can be clearly judged. The computed $(H^TH)_{\text{nm}}$ are given below:

$$(H^TH)_{\text{nm}} = \begin{bmatrix} 1 & 0.321 & 0.305 & 0.355 & 0.283 \\ & 1 & 0.294 & 0.293 & 0.304 \\ & & 1 & 0.240 & 0.259 \\ & & & 1 & 0.238 \\ & & & & 1 \end{bmatrix}, \quad S = \begin{bmatrix} .1625 & .0017 & .0009 & .0022 & .0023 \\ & .2234 & .0010 & .0026 & .0022 \\ & & .2017 & .0014 & .0008 \\ & & & .2576 & .0027 \\ & & & & .2410 \end{bmatrix}$$

The average of off-diagonal elements is 0.289. Thus the solution is about 29% off the perfect orthogonality. The factor S in $W = HSH^T$ is given above. S is close to a diagonal matrix. However, the small values are important to make HSH^T a better matrix approximation of W.

8.2 Results using $B = FG^T$

The most interesting aspect of $B = FG^T$ factorization is the simultaneous clustering of document and words. We first consider the orthogonality of F, G which characterizes the level of soft clustering. The normalized orthogonality as in §8.1 are computed and given below

$$(F^TF)_{nm} = \begin{bmatrix} 1 & 0.224 & 0.227 & 0.283 & 0.266 \\ & 1 & 0.133 & 0.228 & 0.268 \\ & & 1 & 0.160 & 0.179 \\ & & & 1 & 0.260 \\ & & & & 1 \end{bmatrix}, \quad (G^TG)_{nm} = \begin{bmatrix} 1 & 0.107 & 0.189 & 0.130 & 0.126 \\ & 1 & 0.063 & 0.088 & 0.114 \\ & & 1 & 0.078 & 0.097 \\ & & & 1 & 0.095 \\ & & & & 1 \end{bmatrix}$$

One can see that off-diagonal elements are generally small. Comparing to $(H^TH)_{nm}$ in $W = HSH^T$ factorization, F, G are more orthogonal than H. This is mainly due to the fact that matrix B is much more sparse than W. Furthermore, F is less orthogonal than G, indicating word clusters are more overlaped than document clusters. The consequence of this difference is partially exhibited below.

1-peak	2-peak	3-peak	4-peak	5-peak
238	123	75	45	19

Table 2: Number of words in different category. Total words is 500.

8.3 Word clustering analysis

Since NMF of the word-document matrix is doing simultaneous clustering of documents and words, let us examine word clustering in some details. We believe this type of word clustering analysis is new; To my knowledge, most text analysis and IR research emphasize document clustering; word clustering is far less frequently discussed.

Since there are no known labels for word clustering, we look into the distribution of word cluster indicators. By the meaning of word, we can approximately assess its relevance to each cluster (newsgroup). We focus on dataset A.

We view the i-th row of word cluster indicator F as the posterior probability that word i belongs to each of the κ word clusters. Let this row of F be $(p_1, \cdots, p_\kappa)$, which has been normalized to $\sum_k p_k = 1$.

Suppose a word has a posterior distribution of $(0.9, 0.1, 0, 0, 0)$; it is obvious this word is cleanly clustered into one cluster. We say this word has a 1-peak distribution. Suppose another word has a posterior distribution of $(0.4, 0.6, 0, 0, 0)$; this word is clustered into two cluster. We say this word has a 2-peak distribution. Now we wish to characterize each word as belonging to 1-peak, 2-peak, 3-peak, etc, We set five prototype distributions:

$$(1,0,0,0,0),\ (\frac{1}{2},\frac{1}{2},0,0,0),\ (\frac{1}{3},\frac{1}{3},\frac{1}{3},0,0),\ (\frac{1}{4},\frac{1}{4},\frac{1}{4},\frac{1}{4},0),\ (\frac{1}{5},\frac{1}{5},\frac{1}{5},\frac{1}{5},\frac{1}{5}),$$

and all permutations. For example, $(1, 0, 0, 0, 0)$ is equivalent to $(0, 1, 0, 0, 0)$. For each word, we assign it to the closest prototype distributions based on symmetrized Kullbak-Leibler distance, In practice, we first sort the row such that the components decrease from left to right, and then assign it to the closest prototype. In Table 2, we show a typical result of this categorization.

Generally speaking, the less number of peaks a word has, the more unique content the word holds, which in turn makes the word more clearly related to the topic of its assigned cluster. This is clearly seen in the 1-peak words in Table 3, where we list top-ranked words in 1-peak, 2-peak and 5-peak categories. The 2-peak words generally have a meaning that fits two clusters. In contrast, the 5-peak words generally have no specific content. All these results are consistent with our understanding and are expected from a systematic content analysis.

To summarize, during the simultaneous clustering of documents and words, NMF is capable of distinguishing the contents of words which fit different clusters. The results on 1-peak, 2-peak, etc., indicate that NMF has a unique capability that many other clustering methods are lacking.

9 Summary

We prove that NMF is equivalent to Kernel K-means clustering and spectral clustering when the orthogonality condition is relaxed.

Our theoretical results clarify the meaning of factorization matrices: (a) the basis vectors (columns of

cluster names	1-peak words
space	satellite orbit mission incoming alaska launch
mideast	israel arab muslim jew palestinian turkey
motercycle	motorcycle bmw chain biker rider bike
baseball	game pitcher basebal pitch catcher sox
graphics	format graphic code pixel video viewer

2-peak words						5-peak words					
	space	*midest*	*motor*	*baseball*	*graphics*		*space*	*midest*	*motor*	*baseball*	*graphics*
plane	.04	.46	.00	.00	.50	ago	.19	.28	.13	.21	.19
fly	.46	.01	.00	.46	.06	area	.28	.27	.21	.12	.11
project	.57	.00	.02	.00	.42	full	.22	.14	.27	.22	.16
scienc	.38	.12	.00	.00	.50	give	.16	.20	.19	.30	.16
water	.12	.42	.38	.08	.00	kind	.22	.19	.12	.16	.32
monitor	.37	.07	.12	.00	.44	lot	.19	.12	.18	.31	.20
fast	.11	.00	.35	.07	.47	school	.20	.26	.14	.21	.20
young	.00	.35	.15	.50	.00	small	.19	.32	.12	.14	.23
model	.35	.00	.06	.15	.44	thing	.19	.12	.29	.26	.13
net	.36	.06	.06	.14	.37	write	.25	.21	.20	.22	.12

Table 3: Top ranked words in 1-peak, 2-peak and 5-peak distributions. Posterior probabilities for 2-peak and 5-peak words are shown.

C in $X = CH^T$) are cluster centroids ; (2) the data projections (rows of H) are cluster indicator vectors.

We emphasize that the relaxation of orthogonality condition is, in fact, a useful feature of NMF for soft clustering. It enhance the posterior probability interpretation of the cluster indicator matrix H, which provides a soft clustering framework and overcomes some of the problems of hard clustering. Both the example in Fig.3 and the word posterior distribution (§8.3) illustrate this soft clustering aspect.

Overall, we provide an unified understanding of two aspects of kernel K-means (cf. Eq.4) and Laplacian matrix-based spectral clustering (cf. Eq.30): (a) if the orthogonality is retained while nonnegativity is relaxed, we obtain eigenvector solutions; (b) if the orthogonality is relaxed while nonnegativity is strictly enforced, we obtain NMF.

Acknowledgments. This work is partially supported by U.S. Department of Energy, Office of Science, under contract number DE-AC03-76SF00098.

References

[1] F. R. Bach and M. I. Jordan. Learning spectral clustering. *Neural Info. Processing Systems 16 (NIPS 2003)*, 2003.

[2] J. C. Bezdek. *Pattern Recognition with Fuzzy Objective Function Algorithms.* Plenum Press, New York, 1981.

[3] J.-P. Brunet, P. Tamayo, T.R. Golub, and J.P. Mesirov. Metagenes and molecular pattern discovery using matrix factorization. *Proc. Nat'l Academy of Sciences USA*, 102(12):4164–4169, 2004.

[4] M. Catral, Lixing Han, Michael Neumann, and Robert J. Plemmons. On reduced rank nonnegative matrix factorizations for symmetric matrices. *Linear Algebra and Its Applications*, to appear.

[5] M. Chu, F. Diele, R. Plemmons, and S. Ragni. Optimality, computation, and interpretations of nonnegative matrix factorizations. October 2004.

[6] M. Cooper and J. Foote. Summarizing video using non-negative similarity matrix factorization. In *Proc. IEEE Workshop on Multimedia Signal Processing*, pages 25–28, 2002.

[7] I. Dhillon and D. Modha. Concept decomposition for large sparse text data using clustering. *Machine Learning*, 42:143–175, 2001.

[8] I.S. Dhillon, Y. Guan, and B. Kulis. Kernel k-means: spectral clustering and normalized cuts. *Proc. ACM Int'l Conf Knowledge Disc. Data Mining (KDD 2004)*, 2004.

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

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

[11] C. Ding, X. He, and H.D. Simon. On the equivalence of nonnegative matrix factorization and spectral clustering. *Proc. SIAM Data Mining Conf*, 2005.

[12] 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 (ICDM)*, pages 107–114, 2001.

[13] 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.

[14] 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.

[15] P. O. Hoyer. Non-negative sparse coding. In *P. 2002 IEEE Workshop on Neural Networks for Signal Processing*, pages 557–565, 2002.

[16] P. O. Hoyer. Non-negative matrix factorization with sparseness constraints. *J. Machine Learning Research*, 5:1457–1469, 2004.

[17] R. Jin, C. Ding, and F. Kang. A probabilistic approach for optimizing spectral clustering. In *Advances in Neural Information Processing Systems*, 2005.

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

[19] 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.

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

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

[22] W. Liu and J. Yi. Existing and new algorithms for non-negative matrix factorization. 2003.

[23] M. Meila and J. Shi. A random walks view of spectral segmentation. *AI-statistics Workshop*, 2001.

[24] A.Y. Ng, M.I. Jordan, and Y. Weiss. On spectral clustering: Analysis and an algorithm. *Proc. Neural Info. Processing Systems (NIPS 2001)*, 2001.

[25] P. Paatero and U. Tapper. Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics*, 5:111–126, 1994.

[26] 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 on Data Mining (SDM 2004)*, pages 452–456, 2004.

[27] F. Sha, L.K. Saul, and D.D. Lee. Multiplicative updates for nonnegative quadratic programming in support vector machines. In *Advances in Neural Information Processing Systems 15*, pages 1041–1048. 2003.

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

[29] Y.-L. Xie, P.K. Hopke, and P. Paatero. Positive matrix factorization applied to a curve resolution problem. *Journal of Chemometrics*, 12(6):357–364, 1999.

[30] E.P. Xing and M.I. Jordan. On semidefinite relaxation for normalized k-cut and connections to spectral clustering. *University of California Berkeley Tech Report CSD-03-1265*, 2003.

[31] W. Xu, X. Liu, and Y. Gong. Document clustering based on non-negative matrix factorization. In *Proc. ACM Conf. Research development in IR(SIRGIR)*, pages 267–273, 2003.

[32] 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.

[33] H. Zha, X. He, C. Ding, M. Gu, and H.D. Simon. Bipartite graph partitioning and data clustering. *Proc. Int'l Conf. Information and Knowledge Management (CIKM 2001)*, 2001.