Orthogonal Nonnegative Matrix Tri-Factorizations for Clustering

Chris Ding
Lawrence Berkeley National Laboratory
Berkeley, CA 94720
chqding@lbl.gov

Tao Li, Wei Peng
School of Computer Science
Florida International University
Miami, FL 33199
taoli,wpeng002@cs.fiu.edu

Haesun Park
College of Computing
Georgia Institute of Technology
Atlanta, GA 30332
hpark@cc.gatech.edu

ABSTRACT

Currently, most research on nonnegative matrix factorization (NMF) focus on 2-factor $X = FG^T$ factorization. We provide a systematic analysis of 3-factor $X = FSG^T$ NMF. While *unconstrained* 3-factor NMF is equivalent to *unconstrained* 2-factor NMF, *constrained* 3-factor NMF brings new features to *constrained* 2-factor NMF. We study the orthogonality constraint because it leads to rigorous clustering interpretation. We provide new rules for updating F, S, G and prove the convergence of these algorithms. Experiments on 5 datasets and a real world case study are performed to show the capability of bi-orthogonal 3-factor NMF on simultaneously clustering rows and columns of the input data matrix. We provide a new approach of evaluating the quality of clustering on words using class aggregate distribution and multi-peak distribution. We also provide an overview of various NMF extensions and examine their relationships.

Categories and Subject Descriptors

I.2 [**Artificial Intelligence**]: Learning; I.5.3 [**Pattern Recognition**]: Clustering

General Terms

Algorithms, Experimentation, Theory

Keywords

nonnegative matrix factorization (NMF), orthogonal factorization, clustering, tri-factorization, multi-peak distribution

1. INTRODUCTION

The nonnegative matrix factorization (NMF) has been shown recently to be useful for many applications in environment, pattern recognition, multimedia, text mining, and DNA gene expressions [3, 5, 15, 20, 27, 32]. This is also extended to classification [30]. NMF can be traced back to 1970s (Notes from G. Golub) and is studied extensively by Paatero [27]. The work of Lee and Seung [18, 19] brought much attention to NMF in machine learning and data mining fields. They suggest that NMF factors contain coherent parts of the original data (images). They emphasize the difference between NMF and vector quantization (which is essentially the K-means clustering). However, later experiments [16, 20] do not support the coherent part interpretation of NMF. In fact, most applications make use of the clustering aspect of NMF, which is de-emphasized by Lee and Seung [18]. A recent theoretical analysis [9] shows the equivalence between NMF and K-means / spectral clustering.

Below we briefly outline NMF which provides notations and further motivations. In general, NMF factorizes input nonnegative data matrix X into 2 nonnegative matrices,

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

where $X \in \mathbb{R}_+^{p\times n}$, $F \in \mathbb{R}_+^{p\times k}$ and $G \in \mathbb{R}_+^{n\times k}$ ($\mathbb{R}_+^{n\times k}$ is the set of all n-by-k matrices whose elements are nonnegative). Generally, the rank of matrices F, G is much lower than the rank of X (i.e., $k \ll \min(p,n)$).

In this paper, we emphasize the orthogonality of matrix factors in NMF. Specifically, we solve the one-sided G-orthogonal NMF,

$$\min_{F\geq 0, G\geq 0} \|X - FG^T\|^2, \ \ s.t.\ G^TG = I. \tag{2}$$

The main advantages are (1) uniqueness of the solution. (2) Clustering interpretations. We will show it is equivalent to K-means clustering.

Furthermore, it is natural to consider imposing orthogonality on both F and G simultaneously in NMF.

$$\min_{F\geq 0, G\geq 0} \|X - FG^T\|^2, \ \ s.t.\ F^TF = I,\ G^TG = I. \tag{3}$$

This corresponds to the simultaneous K-means clustering [9] of the rows and columns of X, where F is the cluster indicator matrix for clustering rows and G is the cluster indicator matrix for clustering columns. However, this double orthogonality is very restrictive and it gives a rather poor matrix low-rank approximation. One needs an extra factor S to absorb the different scales of X, F, G. S provides

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
Copyright 200X ACM X-XXXXX-XX-X/XX/XX ...$5.00.

additional degrees of freedom such that the low-rank matrix representation remains accurate while F gives row clusters and G gives column clusters. Thus we consider the following nonnegative 3-factor decomposition

$$X \simeq FSG^T. \tag{4}$$

For the objective of the function approximation, we optimize

$$\min_{F\geq 0, G\geq 0, S\geq 0} \|X - FSG^T\|^2,\ \ s.t.\ F^TF = I,\ G^TG = I. \tag{5}$$

We note $X \in \mathbb{R}_+^{p\times n}$, $F \in \mathbb{R}_+^{p\times k}$ and $S \in \mathbb{R}_+^{k\times \ell}$ and $G \in \mathbb{R}_+^{n\times \ell}$. This allows the number of row cluster (k) differ from the number of column cluster (ℓ). In most cases, we set $k = \ell$. This form gives a good framework for simultaneously clustering the rows and columns of X. Simultaneously rows and columns clustering using Laplacian matrix has been studied in [7, 35].

NMF is one type of matrix factorizations. There are other types of factorizations [6, 17, 33, 24, 23, 21]. Others include Latent Semantic Indexing [1], scaled PCA [10], generalized SVD [28], etc.

Here are some more notations. We often write $F = (\mathbf{f}_1, \cdots, \mathbf{f}_k)$ and $G = (\mathbf{g}_1, \cdots, \mathbf{g}_k)$. The matrix norm $\|A\| = \sqrt{\sum_{ij} a_{ij}^2}$. In the following, we emphasize the benefit of orthogonality for the uniqueness of the solution in §2. In §3 the benefits of orthogonality are discussed in detail. In §4 the computational algorithm for uni-orthogonal NMF is given. In §5 strong motivations for 3-factor NMF are discussed and the computing algorithm is given. In §7 the detailed proof of algorithm correctness and convergence for uni-orthogonal NMF is presented. In §8 algorithm correctness and convergence for bi-orthogonal NMF are presented. §9 is devoted to experiments on 6 datasets. In particular, §9.3 shows the results on document clustering (columns of X); §9.4 provides a new and systematic analysis on word clustering (rows of X). This differs substantially from the usual (document) clustering. In §10, an interesting case study of clustering system log data is presented. In §11, we provide an overview of various NMF extensions and examine their relationships. The summary is given in §12.

2. UNIQUENESS OF ORTHOGONAL NMF

Generally speaking, for any given solution (F, G) of NMF: $X = FG^T$, there exist large number of matrices (A, B) such that

$$AB^T = I,\ FA \geq 0,\ GB \geq 0. \tag{6}$$

Thus (FA, GB) is also the solution with the same residue $\|X - FG^T\|$. With orthogonality condition, we show that this degree of freedom is eliminated.

Proposition 1. With the orthogonality condition $F^TF = I$ in the NMF, there exist no matrix A, B that satisfies both Eq.(6) and the orthogonality condition $(FA)^T(FA) = I$, except when A, B are permutation matrices, i.e., $A = P, B = P^T, P^TP = I, P_{ij} = 0$ or 1.

Proof. $(FA)^T(FA) = I$ implies $A^TA = I$. Except $A = I$ or permutation matrix, at least one off-diagonal element of A must be negative. Say, $A_{k_1,\ell_1} < 0$. Because of orthogonality, each row of F has exactly one nonzero element. Suppose for i_1-th row of F, the k_1-th element is nonzero. Thus $(FA)_{i_1\ell_1} = \sum_k F_{i_1,k}A_{k,\ell_1} = F_{i_1,k_1}A_{k_1,\ell_1} < 0$. Thus there can be no negative elements in A. □

We note that for any matrix factorization, the freedom of column/row permutation always exists.

3. ORTHOGONAL NMF AND CLUSTERING

Lee and Seung [18] emphasizes the difference between NMF and vector quantization (which is K-means clustering). Later experiments [16, 20] empirically show that NMF has clear clustering effects. Theoretically, NMF is inherently related to kernel K-means clustering.

Theorem 1. Orthogonal NMF,

$$\min_{F\geq 0, G\geq 0} \|X - FG^T\|^2,\ \ s.t.\ G^TG = I. \tag{7}$$

is equivalent to K-means clustering.

This theorem has been previously proved[9] with additional normalization conditions. Here we give a simpler and more general proof, which can easily generalize to bi-orthogonality.

Proof. We write $J = ||X - FG^T||^2 = \mathrm{Tr}(X^TX - 2F^TXG + F^TF)$. The zero gradient condition $\partial L/\partial F = -2XG + 2F = 0$ gives $F = XG$. Thus $J = \mathrm{Tr}(X^TX - G^TX^TXG)$. Since $\mathrm{Tr}(X^TX)$ is a constant, the optimization problem becomes

$$\min_{G\geq 0} \mathrm{Tr}(G^TX^TXG)\ \ s.t.\ G^TG = I. \tag{8}$$

According to Theorem 2 below, this is identical to K-means clustering. ■

We note that Theorem 1 holds even if X and F are not nonnegative, i.e., X and F have mixed-sign entries. This motives generalizing NMF to semi-NMF in §11.1.

Theorem 2[8, 34]. The K-means clustering minimizes

$$J = \sum_{k=1}^{K} \sum_{i \in C_k} ||\mathbf{x}_i - \mathbf{c}_k||^2 = \sum_{k=1}^{K} \sum_{i=1}^{n} G_{ik} ||\mathbf{x}_i - \mathbf{c}_k||^2 \tag{9}$$

where $\mathbf{c}_k$ is the cluster centroid of the k-th cluster. More generally, the Kernel K-means with mapping $\mathbf{x}_i \to \phi(\mathbf{x}_i)$ minimizes

$$J_\phi = \sum_{k=1}^{K} \sum_{i \in C_k} ||\phi(\mathbf{x}_i) - \phi_k||^2 = \sum_{k=1}^{K} \sum_{i=1}^{n} G_{ik} ||\phi(\mathbf{x}_i) - \phi_k||^2 \tag{10}$$

where ϕ_k is the centroid in the feature space, and G is cluster indicator matrix: $G_{ik} = 1$ if $x_i \in C_k$ and $G_{ik} = 0$ otherwise. Each row of G has only one nonzero element. and $G^TG = \mathrm{diag}(|C_1|, \cdots, |C_K|)$, where $|C_k|$ is the number of data points in cluster C_k. Define the normalized $\tilde{G} = G(G^TG)^{-1/2}$ such that $\tilde{G}^T\tilde{G} = I$. Both clusterings can be solved via the optimization problem

$$\max_{\tilde{G}^T\tilde{G} = I,\ \tilde{G} \geq 0} \mathrm{Tr}(\tilde{G}^TW\tilde{G}), \tag{11}$$

where $W_{ij} = \phi(\mathbf{x}_i)^T\phi(\mathbf{x}_j)$ is the kernel. For K-means, $\phi(\mathbf{x}_i) = \mathbf{x}_i$, $W_{ij} = \mathbf{x}_i^T\mathbf{x}_j$.

Now we generalize Theorem 2 to simultaneous row/column K-means clustering.

Theorem 3. Let G be the cluster indicator matrix for K-means clustering of columns of X and F be the cluster indicator matrix for

K-means clustering of rows of X. The simultaneous row/column clustering can be solved by optimizing

$$\min_{\tilde{F},\tilde{G},D\geq 0} \|X-\tilde{F}D\tilde{G}^T\|^2,\ \ s.t.\ \tilde{F}^T\tilde{F}=I,\ \tilde{G}^T\tilde{G}=I,\ D \text{ diagonal}. \tag{12}$$

Proof. The clustering of columns of X is the same as in Theorem 2. Let the rows of X be $(\mathbf{y}_1,\cdots,\mathbf{y}_k)=X^T$. Applying Theorem 2 to rows of X, the simultaneous row and column clustering becomes simultaneous optimizations:

$$\max_{\tilde{F}^T\tilde{F}=I} \operatorname{Tr} \tilde{F}^T XX^T\tilde{F}, \quad \max_{\tilde{G}^T\tilde{G}=I} \operatorname{Tr} \tilde{G}^T X^T X\tilde{G}. \tag{13}$$

We now show that $\tilde{F}$ and $\tilde{G}$, or their un-normalized counterparts, F and G, can be obtained via

$$\min_{F\geq 0,G\geq 0} \|X-FG^T\|^2,\ \ s.t.\ \ F,G \text{ column orthogonal}. \tag{14}$$

From this, let the diagonal matrix $D_F=(||\mathbf{f}_1||,\cdots,||\mathbf{f}_k||)$ and the diagonal matrix $D_G=(||\mathbf{g}_1||,\cdots,||\mathbf{g}_k||)$. We can write $FG^T=(FD_F^{-1})(D_FD_G)(GD_G^{-1})^T$. Thus Eq.(14) is equivalent to Eq.(12).

To show the optimization in Eq.(14) is equivalent to that in Eq.(13), we write $J=\|X-FG^T\|^2=\operatorname{Tr}(X-FG^T)^T(X-FG^T)$. From $\partial T/\partial G=0$, we obtain $G=X^T(F^TF)^{-1}$. Substituting back, we have $J=\operatorname{Tr}(X^TX-\tilde{F}^TXX^T\tilde{F})$, where $\tilde{F}=F(F^TF)^{-1/2}$ satisfies $\tilde{F}^T\tilde{F}=I$. Thus $\min J$ becomes $\max \operatorname{Tr} \tilde{F}^TXX^T\tilde{F}$. This is part of Eq.(13) for $\tilde{F}$. From $\partial T/\partial F=0$, we can show $\min J$ becomes $\max \operatorname{Tr} \tilde{G}^TX^TX\tilde{G}$. This is part of Eq.(13) for $\tilde{G}$. Thus optimization in Eq.(14) is equivalent to that of Eq.(13). □

Without the diagonal factor D in Eq.(12), Theorem 3 has been noted in [9]. The more careful treatment here reveals that the simultaneous row/column K-means clustering allows an extra scale diagonal factor D in the NMF formulation. Generalizing D to full matrix S, we arrive at the bi-orthogonal 3-factor NMF of Eq.(5).

Proposition 2. The bi-orthogonal 3-factor NMF is equivalent to

$$\max_{G^TG=I,\ G\geq 0} \operatorname{Tr}[G^T(F^TX)^T(F^TX)G],\ \ \text{fixing } F, \tag{15}$$

and alternatively,

$$\max_{F^TF=I,\ F\geq 0} \operatorname{Tr}[F^T(XG)(XG)^TF],\ \ \text{fixing } G. \tag{16}$$

Proof. Expanding $J_4=\operatorname{Tr}(X^TX-2X^TFSG^T+S^TS)$. Setting the derivative $\partial J_4/\partial S=0$, we obtain

$$S=F^TXG,\ \text{ or }\ S_{\ell k}=\mathbf{f}_\ell^TX\mathbf{g}_k=\frac{1}{|R_\ell|^{1/2}\,|C_k|^{1/2}}\sum_{i\in R_\ell}\sum_{J\in C_k}X_{ij}. \tag{17}$$

where $|R_\ell|$ is the size of the ℓ-th row cluster, and $|C_k|$ is the size of the k-th column cluster. $S_{\ell k}$ represents properly normalized within-cluster sum of weights ($\ell=k$) and between-cluster sum of weights ($\ell\neq k$). The meaning of 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.

Substituting $S=F^TXG$ into J_4, we have $J_4=\operatorname{Tr}(X^TX-G^TX^TFF^TXG)$. This leads to optimization in Eqs.(15,16). ■

Now, applying Theorem 2 to Eqs.(15,16), we have

Theorem 4. In the bi-orthogonal 3-factor NMF, G gives the solution for kernel K-means clustering of columns of X using kernel $W=X^TFF^TX$ (inner product of the projection of X into the subspace spanned by F). Similarly, F gives the solution for clustering rows of X using kernel $W=XGG^TX^T$ (inner product of the projection of X into the subspace spanned by G).

4. COMPUTING UNI-ORTHOGONAL NMF

We are interested in solving the following one-side orthogonal (F-orthogonal) NMF

$$\min_{F\geq 0,G\geq 0} \|X-FG^T\|^2,\ \ s.t.\ F^TF=I. \tag{18}$$

We solve it using an iterative update algorithm. For this F-orthogonal optimization problem Eq.(18) the update rules are

$$G_{jk}\leftarrow G_{jk}\frac{(X^TF)_{jk}}{(GF^TF)_{jk}}, \tag{19}$$

$$F_{ik}\leftarrow F_{ik}\frac{(XG)_{ik}}{(FF^TXG)_{ik}}. \tag{20}$$

The correctness and convergence proofs involve optimization theory, auxiliary function and several matrix inequalities. The lengthy proofs are given in §7.

Alternatively, we optimize the G-orthogonal NMF

$$\min_{F\geq 0,G\geq 0} \|X-FG^T\|^2,\ \ s.t.\ G^TG=I. \tag{21}$$

The update rules are

$$G_{jk}\leftarrow G_{jk}\frac{(X^TF)_{jk}}{(GG^TX^TF)_{jk}}, \tag{22}$$

$$F_{ik}\leftarrow F_{ik}\frac{(XG)_{ik}}{(FG^TG)_{ik}}. \tag{23}$$

Update rules Eqs.(19,23) are standard NMF rules[19]. Update rules Eqs.(22,20) are results of this paper (see §7).

Initialization. Using the relation to K-means clustering (Theorem 1), we initialize F,G for the G-orthogonal NMF of Eq.(21) as the following. We do K-means clustering of columns of X. From this we obtain the cluster centroids $C=(\mathbf{c}_1,\cdots,\mathbf{c}_k)$ and set $F=C$. From the cluster memberships, we obtain G. We set $G\leftarrow G+0.2$. We initialize the F-orthogonal NMF of Eq.(18) similarly.

5. COMPUTING BI-ORTHOGONAL NMF

First, we emphasize the role of orthogonal in 3-factor NMF. Considering the unconstrained 3-factor NMF

$$\min_{F\geq 0,G\geq 0,S\geq 0} \|X-FSG^T\|^2, \tag{24}$$

we note that this 3-factor NMF can be reduced to the unconstrained 2-factor NMF by mapping $F\leftarrow FS$. Another way to say this is that the degree of freedom of FSG^T is the same as FG^T.

Therefore, 3-factor NMF is interesting only when it can not be transformed into 2-factor NMF. This happens when certain constraints are applied to the 3-factor NMF. However, not all constrained 3-factor NMF differ from their 2-factor NMF counterpart. For example, the following 1-sided orthogonal 3-factor NMF

$$\min_{F\geq 0,G\geq 0,S\geq 0} \|X-FSG^T\|^2,\ \ F^TF=I \tag{25}$$

is no different from its 2-factor counterpart Eq.(18), because the mapping $F \leftarrow FS$ reduces one to another.

It is clear that

$$\min_{F\geq 0, G\geq 0, S\geq 0} ||X - FSG^T||^2, \;\; s.t. \; F^T F = I, \; G^T G = I. \tag{26}$$

has no corresponding 2-factor counterpart. We call it the bi-orthogonal tri-factorization and is the focus of this paper. It can be computed using the following update rules

$$G_{jk} \leftarrow G_{jk} \frac{(X^T FS)_{jk}}{(GG^T X^T FS)_{jk}}, \tag{27}$$

$$F_{ik} \leftarrow F_{ik} \frac{(XGS^T)_{ik}}{(FF^T XGS^T)_{ik}}. \tag{28}$$

$$S_{ik} \leftarrow S_{ik} \frac{(F^T XG)_{ik}}{(F^T FSG^T G)_{ik}}. \tag{29}$$

These rules are obtained as the following:

Update G. Clearly, fixing (FS), updating G is identical to Eq.(2) (replacing D_n by I). The updating rule is given by Eq.(22). Replacing F by FS, update rule Eq.(22) becomes Eq.(27)

Update F. Similarly, fixing SG^T, the rule updating F is obtained from Eq.(20). Replacing G by GS^T, we obtain the updating rule of Eq.(28).

Update S. Fixing F, G, we update S using Eq.(29). The correctness and convergence of these update rules are proved in §8.

Initialization. Using the relation to simultaneous K-means clustering of rows and columns (Theorem 4), we initialize F, G, S as the following. (A) We do K-means clustering of columns of X. From this we obtain the cluster memberships which gives G. We set $G \leftarrow G + 0.2$. (B) We do K-means clustering of rows of X and obtain the cluster memberships as F. We set $F \leftarrow F + 0.2$. (C) We initialize S using Eq.(17).

6. SYMMETRIC 3-FACTOR NMF: $W = HSH^T$

An important special case is that the input X contains a matrix of pairwise similarities: $X = X^T = W$. In this case, $F = G = H$. We optimize the symmetric NMF:

$$\min_{H\geq 0, S\geq 0} ||X - HSH^T||^2, \;\; s.t. \; H^T H = I. \tag{30}$$

This can be computed using

$$H_{jk} \leftarrow H_{jk} \frac{(W^T HS)_{jk}}{(HH^T W^T HS)_{jk}}, \tag{31}$$

$$S_{ik} \leftarrow S_{ik} \frac{(H^T WH)_{ik}}{(H^T HSH^T H)_{ik}}. \tag{32}$$

7. UNI-ORTHOGONAL NMF: CORRECTNESS AND CONVERGENCE

We wish to solve the following optimization problem

$$\min_{F\geq 0} J_3(F) = ||X - FG^T||^2, \quad s.t. \; F^T F = I, \tag{33}$$

where X is nonnegative input and G is fixed. We prove that the update algorithm of Eq.(20) correctly solves this problem.

Following the standard theory of constrained optimization, we introduce the Lagrangian multipliers λ (a symmetric matrix of size $K \times K$) and minimize the Lagrangian function

$$L_3(F) = ||X - FG^T||^2 + \text{Tr}[\lambda(F^T F - I)]. \tag{34}$$

Note $||X - FG^T||^2 = Tr(X^T X - 2F^T XG + G^T GF^T F)$. The gradient is

$$\frac{\partial L}{\partial F} = -2XG + 2FG^T G + 2F\lambda. \tag{35}$$

The KKT complementarity condition for the nonnegativity of F_{ik} gives

$$(-2XG + 2FG^T G + 2F\lambda)_{ik} F_{ik} = 0. \tag{36}$$

This is the fixed point relation that local minima for S must satisfy.

The standard approach is to solve the coupled equations Eq.(36) and constraint $F^T F = I$ for F, λ. using a nonlinear method such as Newton's method. There are nk variables for F and $k(k+1)/2$ for λ and the same number of equations. This system of nonlinear equations is generally difficult to solve.

As a contribution, here we provide a much simpler algorithm Eq.(20) to compute the solution. There are two issues: (1) convergence of the algorithm; (2) correctness of the converged solution.

Correctness.

We show that given an initial guess of F, successive update of

$$F_{ik} \leftarrow F_{ik} \frac{(XG)_{ik}}{[F(G^T G + \lambda)]_{ik}}. \tag{37}$$

will converge to a local minima of the problem. The correctness is assured by the fact that at convergence, the solution will satisfy Eq.(36). We will show that the Lagrangian multiplier λ is given by Eq.(44). Substituting, we recover the update rule of Eq.(20).

Convergence.

The convergence is guaranteed by the monotonicity theorem

Theorem 5. The Lagrangian function L_3 is monotonically decreasing (non-increasing) under the update rule Eq.(37), assuming $G^T G + \lambda \geq 0$. Because L_3 is obviously bounded from below, the successive iteration will converge.

Let us first prove the following proposition which plays a key role in the proof of Theorems 5 and 7.

Proposition 6. For any matrices $A \in \mathbb{R}_+^{n\times n}, B \in \mathbb{R}_+^{k\times k}, S \in \mathbb{R}_+^{n\times k}, S' \in \mathbb{R}_+^{n\times k}$, and A, B are symmetric, the following inequality holds

$$\sum_{i=1}^{n} \sum_{p=1}^{k} \frac{(AS'B)_{ip} S_{ip}^2}{S'_{ip}} \geq \text{Tr}(S^T ASB). \tag{38}$$

Proof. Let $S_{ip} = S'_{ip} u_{ip}$. Using the explicit index, the difference $\Delta =$ LHS-RHS can be written as

$$\Delta = \sum_{i,j=1}^{n} \sum_{p,q=1}^{k} A_{ij} S'_{jq} B_{qp} S'_{ip} (u_{ip}^2 - u_{ip} u_{jq}).$$

Because A, B are symmetric, this is equal to

$$= \sum_{i,j=1}^{n} \sum_{p,q=1}^{k} A_{ij} S'_{jq} B_{qp} S'_{ip} \left(\frac{u_{ip}^2 + u_{jq}^2}{2} - u_{ip} u_{jq}\right)$$

$$= \frac{1}{2} \sum_{i,j=1}^{n} \sum_{p,q=1}^{k} A_{ij} S'_{jq} B_{qp} S'_{ip} (u_{ip}^2 - u_{jq}^2)^2 \geq 0.$$

□

When $B = I$, and S is a column vector, this result reduces to the one shown in [19]. Now we are ready to prove Theorem 2.

Proof of Theorem 5.

We use the auxiliary function approach [19]. A function $Z(H,\tilde{H})$ is called an auxiliary function of $L(H)$ if it satisfies

$$Z(H,\tilde{H}) \geq L(H), \quad Z(H,H) = L(H), \tag{39}$$

for any $H,\tilde{H}$. Define

$$H^{(t+1)} = \arg\min_{H} Z(H,H^{(t)}). \tag{40}$$

By construction, $L(H^{(t)}) = Z(H^{(t)},H^{(t)}) \geq Z(H^{(t+1)},H^{(t)}) \geq L(H^{(t+1)})$. Thus $L(H^{(t)})$ is monotonic decreasing (non-increasing). The key is to find appropriate $Z(H,\tilde{H})$ and its global minima.

We write L_3 of Eq.(34) as

$$L_3(F) = \mathrm{Tr}[-2F^TXG + (G^TG+\lambda)F^TF],$$

where we ignore the constraints X^TX and $\mathrm{Tr}\lambda$. Now we show that the following function

$$Z(F,F') = -\sum_{ik} 2(F^TXG)_{ik} + \sum_{ik} \frac{(F'(G^TG+\lambda)]_{ik}F_{ik}^2}{F'_{ik}} \tag{41}$$

is an auxiliary function of $L_3(F)$. First, it is obvious that when $F' = F$ the equality holds $Z(F,F') = L_3(F)$. Second, the inequality holds $Z(F,F') \geq L_3(F)$, because: the second term in $Z(F,F')$ is always bigger than the second term in $L_3(F)$, due to Proposition 6. Thus the conditions of Eq.(39) are satisfied.

Now according to Eq.(40), we need to find the *global* minimum of $f(F) = Z(F,F')$ fixing F'. A local minima is given by

$$0 = \frac{\partial Z(F,F')}{\partial F_{ik}} = -2(XG)_{ik} + 2\frac{[F'(G^TG+\lambda)]_{ik}F_{ik}}{F'_{ik}}.$$

Solving for F_{ik}, the minima is

$$F_{ik} = F'_{ik}\frac{(XG)_{ik}}{[F(G^TG)+\lambda]_{ik}}$$

We can show the Hessian matrix $\partial^2 Z(F,F')/\partial F_{ik}\partial F_{j\ell}$ is positive definite. Thus this is a convex function and the minima is also the global minima.

Now according to Eq.(40), $F^{(t+1)} = F$ and $F' = F^{(t)}$, we recover Eq.(37). □

It remains to determine the Lagrangian multiplier λ and make sure $G^TG + \lambda \geq 0$. In Eq.(37), summing over index i, we have $(-F^TXG + G^TG + \lambda)_{kk} = 0$. Therefore we obtain the diagonal elements of the Lagrangian multipliers

$$\lambda_{kk} = (F^TXG - G^TG)_{kk} \tag{42}$$

The off-diagonal elements of the Lagrangian multipliers are approximately obtained by setting $\partial L/\partial F_{ik} = 0$ (ignoring the nonnegativity constraint on F). From Eq.(35), and some algebra, we obtain

$$\lambda_{k\ell} = (F^TXG - G^TG)_{k\ell}, \; k \neq \ell \tag{43}$$

Combining Eqs.(42, 43), we have a compact solution for the Lagrangian multipliers

$$\lambda = F^TXG - G^TG. \tag{44}$$

Clearly, the condition in Theorem 2, $G^TG + \lambda \geq 0$ is satisfied. Substituting this in Eq.(37) we obtain the update rule of Eq.(20).

Note that since the off-diagonal elements of the Lagrangian multipliers $\Lambda = (\lambda_{k\ell})$ are obtained approximately, the final solution for F does not satisfy $F^TF = I$ exactly. This is in fact an advantage. If $F^TF = I$ holds exactly, due to nonnegativity, each row of F can only has one nonzero elements. In the multiplicative updating algorithm, a zero element will lock its self at zero permanently. The slight deviation from exact $F^TF = I$ allows all elements in a row to be nonzero (although most are very small) and the final pattern of nonzeros could change as the updating process evolve.

So far we assume G is fixed. Given F, we can update G using the standard rule of Eq.(19). We can alternatively update F,G, and residue $J(F,G)$ will monotonically decrease

$$J(F^{(0)},G^{(0)}) \geq J(F^{(1)},G^{(0)}) \geq J(F^{(1)},G^{(1)}) \cdots.$$

In summary, we have proved that the minimization of Eq.(18) can be solved by the updating rules of Eqs.(19,20).

Alternative Update Algorithm

We can show that the Langranigan function $L_3(F)$. has another auxiliary function

$$Z(F,F') = -\sum_{ik} 2(XG)_{ik}F'_{ik}(1+\log\frac{F_{ik}}{F'_{ik}}) + \sum_{ik} \frac{(F'(G^TG+\lambda)]_{ik}F_{ik}^2}{F'_{ik}}, \tag{45}$$

because the first term in $Z(F,F')$ is always smaller than the first term in $L_3(F)$, due to the inequality $z \geq 1 + \log(z)$, $\forall z > 0$, and we set $z = F_{ik}/F'_{ik}$. From this auxiliary function, we can drive the following update rule

$$F_{ik} \leftarrow F_{ik}\sqrt{\frac{(XG)_{ik}}{[F(G^TG+\lambda)]_{ik}}}. \tag{46}$$

in contrast to Eq.(37).

8. 3-FACTOR NMF: CORRECTNESS AND CONVERGENCE

In 3-factor NMF, the key is the factor S in the middle. Factors F,G can be dealt with in the same way as in 2-factor NMF.

Theorem 7. Let F,G be any fixed matrices,

$$\begin{aligned} J_5(S) &= ||X - FSG^T||^2 \\ &= \mathrm{Tr}(X^TX - 2G^TX^TFS + F^TFSG^TGS^T) \end{aligned} \tag{47}$$

is monotonically decreasing under the update rule of Eq.(29).

Proof. First we prove the correctness. Following the same approach in §7, The KKT complementarity condition for the nonnegativity if S_{ik} gives

$$(-F^TXG + FF^TSG^TG)_{ik}S_{ik} = 0. \tag{48}$$

At convergence, the solution from the update rule Eq.(29) satisfies Eq.(48). This proves the correctness of update rule Eq.(29).

Next, we consider the convergence of the update rule Eq.(29).

Theorem 8. The objective function $J_5(S)$ is non-increasing under the update rule Eq.(29).

Proof. We use the auxiliary function approach in the proof of Theorem 5 near Eqs.(39, 40). Now we show that

$$Z(S,S') = ||X||^2 - 2\mathrm{Tr}(F^T XGS) + \sum_{ik} \frac{(F^T FS'G^T G)_{ik} S_{ik}^2}{S'_{ik}} \qquad (49)$$

is an auxiliary function of $J_5(S)$. The third term in $Z(S,S')$ is always bigger than the third term in $J_5(S)$, due to Proposition 6 in §7 Eq.(38). The second in $Z(S,S')$ is identical to the second term in $J_5(S)$. Thus the condition $Z(S,S') \geq J_5(S)$ holds. The equality condition $Z(S,S) = J_5(S)$ holds obviously. Therefore $Z(S,S')$ is an auxiliary function of $J_5(S)$.

According to Eq.(40), $S^{(t+1)}$ is given by the minimum of $J(S,S')$ while fixing $S' = S^{(t)}$. The minimum is obtained by setting

$$0 = \frac{\partial Z(S,S')}{\partial S_{ik}} = -2(F^T XG)_{ik} + 2\frac{(F^T FS'G^T G)_{ik} S_{ik}}{S'_{ik}}$$

which is equal to

$$S_{ik} = S'_{ik} \frac{(F^T XG)_{ik}}{(F^T FS'G^T G)_{ik}}$$

According to Eq.(40), $S^{(t+1)} = S$ and $S' = S^{(t)}$. We recover Eq.(29). Under this update rule, $J_5(S)$ decreases monotonically. □

Alternative Update Algorithm

We can also show that $Z(S,S') =$

$$||X||^2 - \sum_{ik} 2(F^T XG)_{ik} S'_{ik}(1 + \log \frac{S_{ik}}{S'_{ik}}) + \sum_{ik} \frac{(F^T FS'G^T G)_{ik} S_{ik}^2}{S'_{ik}} \qquad (50)$$

is another auxiliary function of $J_5(S)$. From this auxiliary function, we can derive an alternative update rule of for $J_5(S)$:

$$S_{ik} \leftarrow S_{ik} \sqrt{\frac{(F^T XG)_{ik}}{(F^T FS'G^T G)_{ik}}} \qquad (51)$$

in constract to Eq.(29).

9. EXPERIMENTS

In this section, we apply the bi-orthogonal 3-factor NMF (BiOR-NM3F) clustering algorithm to cluster documents and compare its performance with other standard clustering algorithms. In our experiments, documents are represented using the binary vector-space model where each document is a binary vector in the term space.

9.1 Datasets

We use a variety of datasets, most of which are frequently used in the information retrieval research. Table 1 summarizes the characteristics of the datasets.

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

WebKB4. The WebKB dataset contains webpages gathered from university computer science departments. There are about 8280 documents and they are divided into 7 categories: student, faculty, staff, course, project, department and other. The raw text is about 27MB. Among these 7 categories, student, faculty, course and project are four most populous entity-representing categories. The associated subset is typically called **WebKB4**. In this paper, we perform experiments on the 4-category dataset.

Datasets	# documents	# class
CSTR	476	4
WebKB4	4199	4
Reuters-top 10	2,900	10
WebAce	2,340	20
Newsgroups	20,000	20

Table 1: Document Datasets Descriptions.

Reuters. The Reuters-21578 Text Categorization Test collection contains documents collected from the Reuters newswire in 1987. It is a standard text categorization benchmark and contains 135 categories. In our experiments, we use a subset of the data collection which includes the 10 most frequent categories among the 135 topics and we call it **Reuters-top 10**.

WebAce. This is from WebACE project and has been used for document clustering [2, 13]. The dataset contains 2340 documents consisting news articles from Reuters new service via the Web in October 1997. These documents are divided into 20 classes.

Newsgroups. The 20 newsgroups dataset contains approximately 20,000 articles evenly divided among 20 Usenet newsgroups. The raw text size is 26MB.

To pre-process the datasets, we remove the stop words using a standard stop list, all HTML tags are skipped and all header fields except subject and organization of the posted articles are ignored. In all our experiments, we first select the top 1000 words by mutual information with class labels. The feature selection is done with the rainbow package [25].

9.2 Evaluation Measures

The above document datasets are standard labeled corpora widely used in the information retrieval literature. We view the labels of the datasets as the objective knowledge on the structure of the datasets. To measure the clustering performance, we use purity [36] and Adjusted Rand Index (ARI) [26] as our performance measures. We expect these measures would provide us with good insights on how our algorithm works.

Purity measures the extent to which each cluster contained data points from primarily one class [36]. The purity of a clustering solution is obtained as a weighted sum of individual cluster purity values and is given by

$$Purity = \sum_{i=1}^{K} \frac{n_i}{n} P(S_i), P(S_i) = \frac{1}{n_i} max_j(n_i^j),$$

where S_i is a particular cluster of size n_i, n_i^j is the number of documents of the i-th input class that were assigned to the j-th cluster, K is the number of clusters and n is the total number of points [1]. In general, the larger the values of purity, the better the clustering solution is.

[1] $P(S_i)$ is also called the individual cluster purity.

Entropy measures how classes distributed on various clusters [36]. The entropy of the entire clustering solution is computed as:

$$Entropy = -\frac{1}{n\log_2 m}\sum_{i=1}^{K}\sum_{j=1}^{m} n_i^j \log_2 \frac{n_i^j}{n_i}, \qquad (52)$$

where m is the number of original labels, K is the number of clusters. Generally, the smaller the entropy value, the better the clustering quality is.

The Rand Index is defined as the number of pairs of objects which are both located in the same cluster and the same class, or both in different clusters and different classes, divided by the total number of objects [29]. Adjusted Rand Index which adjusts Rand Index is set between $[0,1]$ [26]. The higher the Adjusted Rand Index, the more resemblance between the clustering results and the labels.

9.3 Document Clustering Result Analysis

We compare our bi-orthogonal 3-factor NMF (BiOR-NM3F) clustering with the K-means algorithm. The comparisons are shown in Table 2. Each entry is the corresponding performance measure value of the algorithm on the row dataset. From Table 2, we observe that our BiOR-NM3F clustering achieves better purity results than K-means on CSTR, WebKB4, Reuters and Newsgroup datasets. In particular, on Newsgroup, the improvement of the purity value is significant (from 0.330 to 0.507). The purity results of K-means are slightly better than BiOR-NM3F on WebAce: the difference is only 0.005.

The performance of purity and ARI is consistent in our comparison, i.e., higher purity values usually correspond to higher ARI values. However, there exist slight differences in the relative performance of purity and entropy in our comparison, i.e., higher purity values do not necessarily correspond to lower entropy values (e.g., on Reuters dataset). This is because the entropy measure takes into account the entire distribution of the documents in a particular cluster and not just the largest class as in the computation of the purity.

In summary, the comparison shows that BiOR-NM3F is a viable and competitive algorithm in document clustering domain, especially considering that BiOR-NM3F is performing document clustering and words clustering simultaneously, while K-means is performing document clustering only.

Datasets	BiOR-NM3F			K-means		
	Purity	Entropy	ARI	Purity	Entropy	ARI
CSTR	0.754	0.402	0.436	0.712	0.412	0.189
WebKB4	0.583	0.372	0.428	0.534	0.442	0.418
Reuters	0.558	0.976	0.510	0.545	0.726	0.506
WebAce	0.541	0.889	0.449	0.546	0.868	0.452
Newsgroups	0.507	1.233	0.179	0.330	1.488	0.149

Table 2: Performance Comparisons of clustering algorithms. Each entry is the corresponding performance value of the algorithm on the row dataset.

9.4 Word Clustering Result Analysis

Our BiOR-NM3F algorithm performs clustering of words simultaneously, where the factor F is the cluster indicator for words.

In this section, we describe the experimental results on word clustering. We consider two clustering strategies: i) hard clustering where a word is assigned to a single cluster and ii) soft clustering where a word can be assigned to several clusters. We analyse hard clustering using class conditional word distribution. We analyse soft clustering using multi-peak distribution. To our knowledge, both of these two analysis methods are new.

9.4.1 Hard Clustering Evaluation

Quantitatively, we can view the i-th row of the cluster indicator F as the posterior probability that word i belongs to each of the K word clusters. For hard clustering, we assign a word to the cluster that has the largest probability value.

Word clustering has no clear *a prior* labels to compare with. We resolve this difficulty by considering the class conditional word distribution. For each document class (with known labels), we compute the aggregate word distribution, the frequency of word occurring in different documents in the class. For hard clustering, we assign each word to the class with highest probability in the aggregate distribution. We expect this assignment would provide a reasonable criterion for evaluating word clustering, i.e., we expect the word clustering results match this assignment. We also use purity, entropy and ARI for evaluating the match. Table 3 shows the hard clustering results on words.

Datasets	Purity	Entropy	ARI
CSTR	0.718	0.490	0.478
WebKB4	0.666	0.668	0.379
Reuters	0.479	0.983	0.272
WebAce	0.599	0.857	0.479
Newsgroups	0.602	0.886	0.275

Table 3: Performance of hard clustering on words. Each entry is the corresponding performance value of the word clustering on the row dataset.

9.4.2 Soft Clustering Evaluation

Note that in general, the cluster indicator F for words is not exactly orthogonal. This is because the off-diagonal Lagrangian multipliers of Eq.(43) are obtained ignoring the non-negativity constraints of G. This slight deviation from rigorous orthogonality produces a benefit of soft clustering.

Here we also provide a systematic analysis of the soft clustering of words. Quantitatively, we view i-th row of F as the posterior probability that word i belongs to each of the K word clusters. Let this row of F be $(p_1,\cdots,p_k)$, which has been normalized to $\sum_k p_k = 1$. Suppose a word has a posterior distribution of

$$(0.96, 0, 0.04, \cdots, 0);$$

it is obvious that 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 $(0.48, 0.48, 0.04, \cdots, 0)$; obviously

Words	Robotics/Vision	Systems	Theory	NLP
1-Peak words				
Polynomial	0.011	0.004	0.966	0.019
Multiprocessor	0.024	0.934	0.021	0.021
complexity	0.023	0.019	0.896	0.062
cache	0.017	0.953	0.015	0.015
set	0.022	0.009	0.890	0.079
object	0.726	0.056	0.031	0.187
train	0.027	0.011	0.024	0.938
reason	0.040	0.008	0.018	0.934
camera	0.898	0.019	0.042	0.041
collapse	0.036	0.015	0.916	0.033
parallel	0.036	0.901	0.031	0.031
compiler	0.060	0.834	0.053	0.053
latency	0.055	0.848	0.049	0.048
robot	0.892	0.040	0.022	0.045
lexical	0.055	0.022	0.049	0.874
study	0.087	0.164	0.673	0.076
track	0.858	0.026	0.058	0.058
percept	0.856	0.018	0.042	0.084
active	0.700	0.075	0.056	0.168
sensor	0.858	0.026	0.058	0.058
2-Peak words				
recognition	0.557	0.004	0.025	0.414
visual	0.668	0.004	0.008	0.320
learn	0.577	0.005	0.034	0.384
human	0.534	0.035	0.020	0.411
representation	0.377	0.011	0.077	0.535
action	0.465	0.023	0.026	0.486
interface	0.428	0.422	0.038	0.113
computation	0.156	0.018	0.433	0.393
information	0.301	0.107	0.180	0.415
3-Peak words				
system	0.378	0.220	0.031	0.372
process	0.335	0.353	0.016	0.296
describe	0.321	0.233	0.060	0.386
user	0.336	0.389	0.020	0.256
perform	0.377	0.352	0.060	0.211
4-Peak words				
present	0.319	0.315	0.188	0.178
algorithm	0.191	0.480	0.177	0.152
implement	0.194	0.435	0.114	0.257
paper	0.183	0.279	0.323	0.215

Table 4: Words Multi-Peak Distribution for CSTR dataset.

this word is clustered into two clusters. We say this word has a 2-peak distribution. In general, we wish to characterize each word as belonging to 1-peak, 2-peak, 3-peak etc. For K word clusters, we set K prototype distributions:

$$(1,0,\cdots,0),(\frac{1}{2},\frac{1}{2},\cdots,0),\cdots,(\frac{1}{K},\cdots,\frac{1}{K}).$$

For each word, we assign it to the closest prototype distribution based on the Euclidean distance, allowing all possible permutations of the clusters. For example, $(1,0,0,\cdots,0)$ is equivalent to $(0,1,0,\cdots,0)$. In practice, we first sort the row such that the components decrease from the left to the right, and then assign it to the closest prototype. Generally speaking, the less peaks of the posterior distribution of the word, the more unique content of the word has. To further illustrate the soft clustering evaluation, we take a closer look at the CSTR dataset. Table 4 lists several words in 1-peak, 2-peak, 3-peak and 4-peak categories respectively. We see that these words are meaningful and are often representatives of the associated document clusters. For example, *multiprocessor* and *cache* are 1-peak words that are associated with the **Systems** cluster; *polynomial* and *complexity* are 1-peak words related to the **Theory** cluster; *recognition* and *learning* are 2-peak words associated with **Robotics/Vision** and **NLP** clusters; *Interface* is a 2-peak words associated with **Robotics/Vision** and **System** clusters; *system*, *process*, and *user* are 3-peak words associated with **Robotics/Vision**, **System** and **NLP** clusters; *present*, *algorithm* and *paper* are 4-peak words. To summarize, the word clustering is capable of distinguishing the contents of words. The results of peak words are consistent with what we would expect from a systematic content analysis. This aspect of tri-factorization shows a unique capability that most other clustering algorithms are lacking.

10. A CASE STUDY ON SYSTEM LOG DATA

In this section, we present a case study of applying our clustering technique to system log data. In system management applications, to perform automated analysis of the historical data across multiple components when problems occur, we need to cluster the log messages with disparate formats to automatically infer the common set of semantic situations and obtain a brief description for each situation [22].

The log files used in our experiments are collected from several different machines with different operating systems using logdump2td (NT data collection tool) developed at IBM T.J. Watson Research Center. The data in the log files describe the status of each component and record system operational changes, such as the starting and stopping of services, detection of network applications, software configuration modifications, and software execution errors. The raw log files contain a free-format ASCII description of the event. In our experiment, we apply clustering algorithms to group the messages into different semantic situations. To preprocess text messages, we remove stop words and skip HTML labels. The raw log messages have been manually labeled with its semantic situation by domain experts [22]. The set of semantic situations include **start**, **stop**, **dependency**, **create**, **connection**, **report**, **request**, **configuration**, and **other**. The detailed explanations of these situations can be found in [4].

Algorithms	Purity	Entropy	ARI
BiOR-NM3F	0.806	0.303	0.856
K-means	0.684	0.491	0.572

Table 5: Clustering Results on System Log Data

We obtain good message clustering results as shown in Table 5. The performance of BiOR-NM3F is better than K-means on all three measures. Table 6 shows the words in 1-peak, 2-peak, 3-peak and 4-peak categories for the log data respectively. We can derive meaningful common situations from the word cluster results. For example, situation **start** can be described by 1-peak words such as *started*, *starting*, and *service*, and 2-peak words such as *version*. The situation **configure** can be described by 1-peak words such as *configuration*, two-peak words such as *product*, and 3-peak words such as *professional*.

Words	Start	Create	Configure	Dependency	Report	Connection	Request	Other	Stop
1-Peak words									
configure	0.014	0.018	0.880	0.002	0.012	0.004	0.020	0.020	0.030
respond	0.019	0.023	0.028	0.002	0.016	0.821	0.028	0.021	0.042
network	0.032	0.038	0.047	0.004	0.026	0.703	0.046	0.032	0.072
create	0.009	0.926	0.013	0.001	0.007	0.003	0.013	0.015	0.023
service	0.704	0.025	0.015	0.022	0.161	0.003	0.015	0.013	0.052
start	0.918	0.012	0.015	0.003	0.009	0.003	0.015	0.012	0.023
contact	0.024	0.024	0.030	0.807	0.017	0.006	0.029	0.021	0.042
fault	0.044	0.053	0.064	0.005	0.613	0.014	0.063	0.054	0.100
stop	0.031	0.038	0.047	0.004	0.026	0.010	0.046	0.034	0.764
restart	0.034	0.041	0.050	0.004	0.028	0.011	0.049	0.042	0.751
blank	0.002	0.003	0.003	0.000	0.002	0.000	0.003	0.982	0.005
fault	0.029	0.035	0.043	0.004	0.743	0.009	0.042	0.032	0.063
start	0.706	0.041	0.050	0.004	0.028	0.011	0.049	0.041	0.070
inventory	0.019	0.023	0.028	0.002	0.016	0.821	0.028	0.022	0.041
2-Peak words									
exist	0.013	0.016	0.020	0.535	0.055	0.297	0.019	0.013	0.035
product	0.035	0.011	0.513	0.001	0.043	0.003	0.252	0.011	0.131
version	0.454	0.020	0.024	0.004	0.416	0.005	0.024	0.023	0.040
complete	0.022	0.013	0.608	0.001	0.009	0.003	0.284	0.010	0.050
root	0.018	0.022	0.027	0.002	0.015	0.538	0.317	0.020	0.041
3-Peak words									
fail	0.011	0.013	0.046	0.416	0.043	0.308	0.015	0.010	0.148
professional	0.059	0.071	0.347	0.007	0.049	0.019	0.260	0.063	0.135
4-Peak words									
timeout	0.096	0.117	0.143	0.012	0.080	0.093	0.141	0.115	0.213
detection	0.077	0.093	0.114	0.010	0.318	0.025	0.112	0.081	0.170

Table 6: Word Multi-Peak Distributions for Log Data

The case study on clustering log message files for computing system management provides a successful story of applying the cluster model in real applications. The log messages are relatively short with a large vocabulary size [31]. Hence they are usually represented as sparse high-dimensional vectors. In addition, the log generation mechanisms implicitly create some associations between the terminologies and the situations. Our clustering method explicitly models the data and word assignments and is also able to exploit the association between data and features. The synergy of these factors leads to the good application on system management.

11. NMF RELATED FACTORIZATIONS

Besides 3-factor extension in this paper, there are many other NMF extensions. Here we provide an overview.

First, we consider different types of nonnegative factorizations. The standard NMF can be written as

$$\text{NMF:} \quad X_+ \approx F_+ G_+$$

using an intuitive notation for $X, F, G \geq 0$.

The classic matrix factorization is Principal Component Analysis (PCA) which uses the singular value decomposition $X \approx U \Sigma V^T$, where we allow U, V to have mixed-signs; the input data could have mixed-signs. absorbing Σ into U, we can write

$$\text{PCA:} \quad X_\pm \approx U_\pm V_\pm$$

However, even if X have mixed-signs, we could enforce G to be nonnegative (since G can be interpreted as cluster indicators, as in §3). This is called semi-NMF [11]:

$$\text{semi-NMF:} \quad X_\pm \approx F_\pm G_+$$

Theorem 1 provides the basis for this semi-NMF formulation.

Both NMF and semi-NMF have clustering capabilities which are generally better than the K-means. In fact, PCA is effectively doing K-means clustering[8, 34]. Let G be the cluster indicators for the k clusters then (1) $GG^T \simeq VV^T$; (ii) the principal directions, UU^T, project data points into the subspace spanned by the k cluster centroids.

11.1 NMF and PLSI

So far, the cost function we used for computing NMF is the sum of squared errors, $||X - FG^T||^2$. Another cost function KL divergence:

$$J_{\text{NMF-KL}} = \sum_{i=1}^{m} \sum_{j=1}^{n} X_{ij} \left[\log \frac{X_{ij}}{(FG^T)_{ij}} - X_{ij} + (FG^T)_{ij} \right] \qquad (53)$$

PLSI [14] maximizes the likelihood

$$J_{\text{PLSI}} = \sum_{i=1}^{m} \sum_{j=1}^{n} X(w_i, d_j) \log P(w_i, d_j) \qquad (54)$$

where the joint occurrence probability is factorized (i.e., parameterized or approximated) as

$$P(w_i, d_j) = \sum_k P(w_i|z_k) P(z_k) P(d_j|z_k) \qquad (55)$$

In [12], it is shown that Objective function of PLSI is identical to the objective function of NMF, i.e., $J_{\text{PLSI}} = -J_{\text{NMF-KL}} + constant$ by setting $(FG^T)_{ij} = P(w_i, d_j)$. Therefore, the NMF update algorithm and the EM algorithm in training PLSI are alternative methods to optimize the same objective function.

12. SUMMARY

We study computational algorithms for orthogonal 2-factor NMF and 3-factor NMF. The bi-orthogonal 3-factor NMF provides a strong capability of simultaneously clustering rows and columns. We derive new updating rules and prove the convergence of these algorithms. Experiments show the usefulness of this approach. We also provide a new approach of evaluating the quality of word clustering. In addition, we also present an overview of various NMF extensions and examine their relationships.

Acknowledgments

C. Ding is partially supported by the US Dept of Energy, Office of Science. T. Li is partially supported by a 2005 IBM Faculty Award, and the National Science Foundation CAREER Award under grant no. NSF IIS-0546280. W. Peng is supported by a Florida International University Presidential Graduate Fellowship.

13. REFERENCES

[1] M.W. Berry, S.T. Dumais, and Gavin W. O'Brien. Using linear algebra for intelligent information retrieval. *SIAM Review*, 37:573–595, 1995.

[2] D. Boley. Principal direction divisive partitioning. *Data mining and knowledge discovery*, 2:325–344, 1998.

[3] J.-P. Brunet, P. Tamayo, T.R. Golub, and J.P. Mesirov. Metagenes and molecular pattern discovery using matrix factorization. *Proc. Natl Academy of Sciences USA*, 102(12):4164–4169, 2004.

[4] M. Chessell. Specification: Common base event, 2003. http://www-128.ibm.com /developerworks/webservices/library/ws-cbe/.

[5] 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.

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

[7] I. S. Dhillon. Co-clustering documents and words using bipartite spectral graph partitioning. *Proc. ACM Int'l Conf Knowledge Disc. Data Mining (KDD 2001)*, 2001.

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

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

[10] C. Ding, X. He, H. Zha, and H. Simon. Unsupervised learning: self-aggregation in scaled principal component space. *Proc. 6th European Conf. Principles of Data Mining and Knowledge Discovery (PKDD)*, pages 112–124, 2002.

[11] C. Ding, T. Li, and M. Jordan. Convex and semi-nonnegative matrix factorizations for clustering and low-dimension representation. Technical Report LBNL-60428, Lawrence Berkeley National Laboratory, University of California, Berkeley, 2006.

[12] C. Ding, T. Li, and W. Peng. Nonnegative matrix factorization and probabilistic latent semantic indexing: Equivalence, chi-square statistic, and a hybrid method. In *Proc. of National Conf. on Artificial Intelligence (AAAI-06)*, 2006.

[13] E. Han, D. Boley, M. Gini, R. Gross, K. Hastings, G. Karypis, V. Kumar, B. Mobasher, and J. Moore. WebACE: A web agent for document categorization and exploration. In *Proceedings of the 2nd International Conference on Autonomous Agents (Agents'98)*. ACM Press, 1998.

[14] T. Hofmann. Probabilistic latent semantic analysis. In *Proceedings of the 15th Annual Conference on Uncertainty in Artificial Intelligence (UAI-99)*, pages 289–296, 1999.

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

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

[17] G. Karypis and E.-H. Han. Concept indexing: A fast dimensionality reduction algorithm with applications to document retrieval and categorization. *Proc. 9th Int'l Conf. Information and Knowledge Management (CIKM 2000)*, 2000.

[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. 2001.

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

[21] T. Li. A general model for clustering binary data. In *KDD*, pages 188–197, 2005.

[22] T. Li, F. Liang, S. Ma, and W. Peng. An integrated framework on mining log files for computing system management. In *KDD*, 2005.

[23] T. Li, S. Ma, and M. Ogihara. Document clustering via adaptive subspace iteration. In *SIGIR*, pages 218–225, 2004.

[24] B. Long, Z. Zhang, and P.S. Yu. Co-clustering by block value decomposition. In *Proc. SIGKDD Int'l Conf. on Knowledge Discovery in Data Mining (KDD'05), pp.635–640.*

[25] A.K. McCallum. Bow: A toolkit for statistical language modeling, text retrieval, classification and clustering. http://www.cs.cmu.edu/ mccallum/bow, 1996.

[26] G.W. Milligan and M.C. Cooper. A study of the comparability of external criteria for hierarchical cluster analysis. *Multivar Behav Res*, 21:846–850, 1986.

[27] 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.

[28] H. Park and P. Howland. Generalizing discriminant analysis using the generalized singular value decomposition. *IEEE. Trans. on Pattern Analysis and Machine Intelligence*, 26:995 – 1006, 2004.

[29] W.M. Rand. Objective criteria for the evaluation of clustering methods. *J Am Stat Assoc*, 66:846–850, 1971.

[30] 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.

[31] J. Stearley. Toward informatic analysis of syslogs. In *Proceedings of IEEE International Conference on Cluster Computing*, 2004.

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

[33] D. Zeimpekis and E. Gallopoulos. Clsi: A flexible approximation scheme from clustered term-document matrices. *Proc. SIAM Data Mining Conf*, pages 631–635, 2005.

[34] 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.

[35] 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.

[36] Y. Zhao and G. Karypis. Empirical and theoretical comparisons of selected criterion functions for document clustering. *Machine Learning*, 55(3):311–331, 2004.