Nonnegative Matrix Factorization with Orthogonality Constraints

Jiho Yoo and Seungjin Choi
Department of Computer Science and Engineering
Pohang University of Science and Technology (POSTECH)
Pohang, Republic of Korea
{zentasis,seungjin}@postech.ac.kr

Received 3 May 2010; Revised 24 May 2010; Accepted 26 May 2010

Nonnegative matrix factorization (NMF) is a popular method for multivariate analysis of nonnegative data, which is to decompose a data matrix into a product of two factor matrices with all entries restricted to be nonnegative. NMF was shown to be useful in a task of clustering (especially document clustering), but in some cases NMF produces the results inappropriate to the clustering problems. In this paper, we present an algorithm for orthogonal nonnegative matrix factorization, where an orthogonality constraint is imposed on the nonnegative decomposition of a term-document matrix. The result of orthogonal NMF can be clearly interpreted for the clustering problems, and also the performance of clustering is usually better than that of the NMF. We develop multiplicative updates directly from true gradient on Stiefel manifold, whereas existing algorithms consider additive orthogonality constraints. Experiments on several different document data sets show our orthogonal NMF algorithms perform better in a task of clustering, compared to the standard NMF and an existing orthogonal NMF.

Categories and Subject Descriptors: H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval–*Clustering*

General Terms: Document Clustering, Nonnegative Matrix Factorization

Additional Key Words and Phrases: Multiplicative Updates, Orthogonal Nonnegative Matrix Factorization, Stiefel Manifold

1. INTRODUCTION

Nonnegative matrix factorization (NMF) is a multivariate analysis method which is proven to be useful in learning a faithful representation of nonnegative data such as images, spectrograms, and documents [Lee and Seung 1999]. NMF seeks a decomposition of a nonnegative data matrix into a product of basis and encoding matrices with all of these matrices restricted to have only nonnegative elements. NMF

allows only non-subtractive combinations of nonnegative basis vectors to approximate the original nonnegative data, possibly providing a parts-based representation [Lee and Seung 1999]. Incorporating extra constraints such as locality and orthogonality was shown to improve the decomposition, identifying better local features or providing more sparse representation [Li et al. 2001]. Orthogonality constraints were imposed on NMF [Ding et al. 2006], where nice clustering interpretation was studied in the framework of NMF.

One of prominent applications of NMF is document clustering [Xu et al. 2003; Shahnaz et al. 2006], where a decomposition of a term-document matrix is considered. NMF finds bases representing the significant terms in each cluster, and encodings representing the cluster where each document belongs. However, the bases found by NMF form a convex cone containing data points, and in some cases it does not correspond to the clusters we wish to find. On the other hand, the orthogonal NMF, which is the NMF with orthogonality constraints on the encoding matrix, can find the basis matrix and encoding matrix having clear interpretation in the clustering problems. Each basis found by orthogonal NMF indicates the direction to the center of each cluster in the data, and the encoding value for a document clearly indicates the corresponding cluster for the document. In fact, orthogonal NMF is shown to be equivalent to the k-means clustering in the sense of the objective function [Ding et al. 2005].

Orthogonality can be imposed on NMF by using standard Lagrangian multiplier method [Ding et al. 2006]. In this paper we take a different approach which exploits the geometric structure of the constraint space. The rectangular matrices constrained to be orthogonal forms a Stiefel manifold, and the learning on the Stiefel manifold has been studied on the differential geometry community [Smith 1993; Edelman et al. 1998]. We employ the natural gradient on the Stiefel manifold to derive a multiplicative update algorithm which preserves both the orthogonality and the nonnegativity. Experiments on several different document data sets show our orthogonal NMF algorithms perform better in a task of clustering, compared to the standard NMF and an existing orthogonal NMF. This is an extension of our earlier work which was presented in [Choi 2008; Yoo and Choi 2008].

The rest of this paper is organized as follows. Section 2 explains about the document clustering problem and how NMF can be applied for the problem. Probabilistic interpretation of NMF model shows the relationship between the factorization result and the clustering interpretation. Section 3 shows how the orthogonality constraints on NMF brings better clustering results, and reviews the existing orthogonal NMF algorithm [Ding et al. 2006]. Section 4 introduces the computation of gradient and geodesics on Stiefel manifold, and then proposes our orthogonal NMF algorithm which uses the gradient on the Stiefel manifold. Section 5 shows the evaluation of the proposed algorithm on the document clustering problems with document datasets. Finally, Section 6 concludes the paper.

2. NMF FOR DOCUMENT CLUSTERING

In the vector-space model of text data, each document is represented by an m-dimensional vector $\boldsymbol{x}_t \in \mathbb{R}^M$, where M is the number of terms in the dictionary. Given

N documents, we construct a term-document matrix $\boldsymbol{x}_t \in \mathbb{R}^{M \times N}$ where X_{ij} corresponds to the significance of term t_i in document d_j that is calculated by

$$X_{ij} = \mathrm{TF}_{ij} \log\left(\frac{N}{\mathrm{DF}_i}\right),$$

where TF_{ij} denotes the frequency of term t_i in document d_j and DF_i represents the number of documents containing term t_i. Elements X_{ij} are always nonnegative and equal zero only when corresponding terms do not appear in the document.

NMF seeks a decomposition of $\boldsymbol{X} \in \mathbb{R}^{M \times N}$ that is of the form

$$\boldsymbol{X} \approx \boldsymbol{U}\boldsymbol{V}^{\mathsf{T}}, \tag{1}$$

where $\boldsymbol{U} \in \mathbb{R}^{M \times K}$ and $\boldsymbol{V} \in \mathbb{R}^{N \times K}$ are restricted to be nonnegative matrices as well and K corresponds to the number of clusters when NMF is used for clustering. Matrices $\boldsymbol{U}$ and $\boldsymbol{V}$, in general, are interpreted as follows.

– When columns in $\boldsymbol{X}$ are treated as data points in m-dimensional space, columns in $\boldsymbol{U}$ are considered as *basis vectors* (or *factor loadings*) and each row in $\boldsymbol{V}$ is *encoding* that represents the extent to which each basis vector is used to reconstruct each data vector.
– Alternatively, when rows in $\boldsymbol{X}$ are data points in N-dimensional space, columns in $\boldsymbol{V}$ correspond to basis vectors and each row in $\boldsymbol{U}$ represents encoding.

Applying NMF to a term-document matrix for document clustering, each column of $\boldsymbol{X}$ is treated as a data point in m-dimensional space. In such a case, the factorization (1) is interpreted as follows.

– U_{ij} corresponds to the degree to which term t_i belongs to cluster c_j. In other words column j of $\boldsymbol{U}$, denoted by $\boldsymbol{u}_j$, is associated with a prototype vector (center) for cluster c_j.
– V_{ij} corresponds to the degree document d_i is associated with cluster j. With appropriate normalization, V_{ij} is proportional to a posterior probability of cluster c_j given document d_i. More details on probabilistic interpretation of NMF for document clustering are summarized in Sec. 2.2.

2.1 Multiplicative updates for NMF

We consider the squared Euclidean distance as a discrepancy measure between the data $\boldsymbol{X}$ and the model $\boldsymbol{U}\boldsymbol{V}^{\mathsf{T}}$, leading to the following least squares error function

$$\mathcal{E} = \frac{1}{2}\left\|\boldsymbol{X} - \boldsymbol{U}\boldsymbol{V}^{\mathsf{T}}\right\|^2, \tag{2}$$

where $\|\cdot\|$ represents the Frobenius norm of a matrix. NMF involves the following optimization:

$$\underset{U \geq 0,\, V \geq 0}{\arg\min}\ \mathcal{E} = \frac{1}{2}\left\|\boldsymbol{X} - \boldsymbol{U}\boldsymbol{V}^{\mathsf{T}}\right\|^2, \tag{3}$$

for the nonnegative input matrix $\boldsymbol{X}$. Gradient descent learning (which is additive

update) can be applied to determine a solution to (3), however, nonnegativity for $\boldsymbol{U}$ and $\boldsymbol{V}$ is not preserved without further operations at iterations.

On the other hand, a multiplicative method developed in [Lee and Seung 2001] provides a simple algorithm for (3). We give a slightly different approach from [Lee and Seung 2001] to derive the same multiplicative algorithm. Suppose that the gradient of an error function has a decomposition that is of the form

$$\nabla\mathcal{E} = [\nabla\mathcal{E}]^{+} - [\nabla\mathcal{E}]^{-}, \tag{4}$$

where $[\nabla\mathcal{E}]^{+} > 0$ and $[\nabla\mathcal{E}]^{-} > 0$. Then multiplicative update for parameters has Θ the form

$$\Theta \leftarrow \Theta \odot \left(\frac{[\nabla\mathcal{E}]^{-}}{[\nabla\mathcal{E}]^{+}}\right)^{\cdot\eta}, \tag{5}$$

where $\odot$ represents Hadamard product (elementwise product), $(\div)$ represents elementwise division, $(\cdot)^{\cdot\eta}$ denotes the elementwise power and η is a learning rate ($0 < \eta \leq 1$). It can be easily seen that the multiplicative update (5) preserves the nonnegativity of the parameter Θ, while $\nabla\mathcal{E} = 0$ when the convergence is achieved.

Derivatives of the error function (2) with respect to $\boldsymbol{U}$ with $\boldsymbol{V}$ fixed and with respect to $\boldsymbol{V}$ with $\boldsymbol{U}$ fixed, are given by

$$\nabla_U\mathcal{E} = [\nabla_U\mathcal{E}]^{+} - [\nabla_U\mathcal{E}]^{-} = \boldsymbol{U}\boldsymbol{V}^{\top}\boldsymbol{V} - \boldsymbol{X}\boldsymbol{V}, \tag{6}$$

$$\nabla_V\mathcal{E} = [\nabla_V\mathcal{E}]^{+} - [\nabla_V\mathcal{E}]^{-} = \boldsymbol{V}\boldsymbol{U}^{\top}\boldsymbol{U} - \boldsymbol{X}^{\top}\boldsymbol{U}. \tag{7}$$

With these gradient calculations, the rule (5) with $\eta = 1$ yields the well-known Lee and Seung's multiplicative updates [Lee and Seung 2001]

$$\boldsymbol{U} \leftarrow \boldsymbol{U} \odot \frac{\boldsymbol{X}\boldsymbol{V}}{\boldsymbol{U}\boldsymbol{V}^{\top}\boldsymbol{V}}, \tag{8}$$

$$\boldsymbol{V} \leftarrow \boldsymbol{V} \odot \frac{\boldsymbol{X}^{\top}\boldsymbol{U}}{\boldsymbol{V}\boldsymbol{U}^{\top}\boldsymbol{U}}. \tag{9}$$

2.2 Probabilistic interpretation and normalization

Probabilistic interpretation of NMF, as in probabilistic latent semantic indexing (PLSI), was given in [Gaussier and Goutte 2005] where equivalence between PLSI and NMF (with I-divergence) was shown.

Let us consider the joint probability of term and document, $p(t_i, d_j)$, which is factorized by

$$\begin{aligned} p(t_i,\ d_j) &= \sum_k\ p(t_i, d_j \mid c_k)p(c_k) \\ &= \sum_k\ p(t_i \mid c_k)p(d_j \mid c_k)p(c_k), \end{aligned} \tag{10}$$

where $p(c_k)$ is the prior probability for cluster c_k. Elements of the term-document matrix, X_{ij}, can be treated as $p(t_i, d_j)$, provided X_{ij} are divided by $\mathbf{1}^{\top}\boldsymbol{X}\mathbf{1}$ such that $\Sigma_i\Sigma_j X_{ij} = 1$ where $\mathbf{1} = [1, \ldots, 1]^{\top}$ with appropriate dimension.

Relating (10) to the factorization (1), U_{ik} corresponds to $p(t_i \mid c_k)$, representing the significance of term t_i in cluster c_k. Applying sum-to-one normalization to each column of $\boldsymbol{U}$, i.e., $\boldsymbol{U}\boldsymbol{D}_U^{-1}$ where $\boldsymbol{D}_U \equiv \text{diag}\,(\mathbf{1}^\top \boldsymbol{U})$, we have an exact relation

$$[\boldsymbol{U}\boldsymbol{D}_U^{-1}]_{ik} = p(t_i \mid c_k)\,.$$

Assume that $\boldsymbol{X}$ is normalized such that $\Sigma_i \Sigma_j X_{ij} = 1$. We define a scaling matrix $\boldsymbol{D}_V \equiv \text{diag}\,(\mathbf{1}^\top \boldsymbol{V})$. Then the factorization (1) can be rewritten as

$$\boldsymbol{X} = (\boldsymbol{U}\boldsymbol{D}_U^{-1})(\boldsymbol{D}_U\boldsymbol{D}_V)(\boldsymbol{V}\boldsymbol{D}_V^{-1})^\top\,. \tag{11}$$

Comparing (11) with the factorization (10), one can see that each element of the diagonal matrix $\boldsymbol{D} \equiv \boldsymbol{D}_U\boldsymbol{D}_V$ corresponds to cluster prior $p(c_k)$. In the case of unnormalized $\boldsymbol{X}$, the prior matrix $\boldsymbol{D}$ absorbs the scaling factor, therefore in practice, the data matrix does not have to be normalized in advance.

In a task of clustering, we need to calculate the posterior of cluster for a given document $p(c_k \mid d_j)$. Applying Bayes' rule, the posterior of cluster is given by the document likelihood and cluster prior probability. That is, $p(c_k \mid d_j)$ is given by

$$\begin{aligned} p(c_k \mid d_j) &\propto p(d_j \mid c_k)p(c_k) \\ &= [\boldsymbol{D}(\boldsymbol{V}\boldsymbol{D}_V^{-1})^\top]_{kj} \\ &= [(\boldsymbol{D}_U\boldsymbol{D}_V)(\boldsymbol{D}_V^{-1}\boldsymbol{V}^\top)]_{kj} \\ &= [\boldsymbol{D}_U\boldsymbol{V}^\top]_{kj}\,. \end{aligned} \tag{12}$$

It follows from (12) that $(\boldsymbol{V}\boldsymbol{D}_U)^\top$ yields the posterior probability of cluster, requiring the normalization of $\boldsymbol{V}$ using the diagonal matrix $\boldsymbol{D}_U$. Thus, we assign document d_j to cluster k^* if

$$k^* = \arg\max_k [\boldsymbol{V}\boldsymbol{D}_U]_{jk}. \tag{13}$$

Document clustering by NMF was first developed in [Xu et al. 2003]. Here we use only different normalization and summarize the algorithm below.

Algorithm outline: Document clustering by NMF

(1) Construct a term-document matrix $\boldsymbol{X}$.
(2) Apply NMF to $\boldsymbol{X}$, yielding $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{V}^\top$.
(3) Normalize $\boldsymbol{U}$ and $\boldsymbol{V}$:

$$\boldsymbol{U} \leftarrow \boldsymbol{U}\boldsymbol{D}_U^{-1}\,,$$
$$\boldsymbol{V} \leftarrow \boldsymbol{V}\boldsymbol{D}_U\,,$$

where $\boldsymbol{D}_U = \mathbf{1}^\top \boldsymbol{U}$.
(4) Assign document d_j to cluster k^* if

$$k^* = \arg\max_k V_{jk}.$$

3. ORTHOGONAL NMF FOR DOCUMENT CLUSTERING

3.1 Orthogonality for clustering

NMF usually works well for the clustering problems by finding a convex cone which contains all the data points. However, there exist some cases where the axes of the

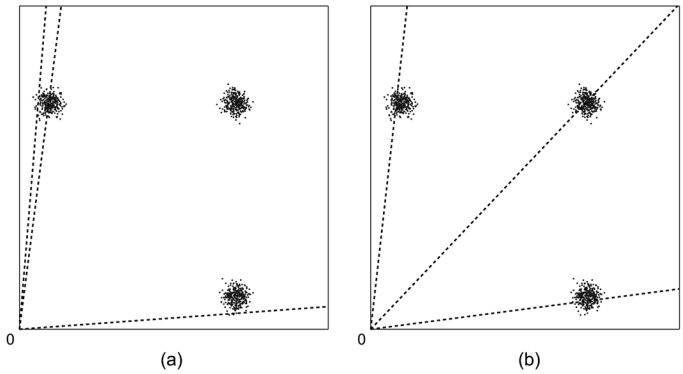

Figure 1. A synthetic example comparing NMF and orthogonal NMF in a clustering problem which has three underlying clusters. The dots represent the data points and the dotted lines represent the direction of the basis vectors obtained by the NMF or orthogonal NMF algorithms. (a) NMF fails to find three clusters correctly in this case, because the algorithm builds a convex cone containing all the data points by the learnt basis vectors. (b) Orthogonal NMF correctly finds underlying clusters, because the algorithm matches exactly one basis for each data point.

convex cone do not correspond to the clusters we want to find. Figure 1(a) shows an example of such cases. The example data clearly consists of three clusters, however, NMF can reconstruct these data with the combinations of only two bases, so the remaining basis goes toward a meaningless direction. Although NMF can reconstruct all the data points with very small error, the clustering fails in this case.

We can obtain results more appropriate for the clustering problems by imposing orthogonality constraints on the encoding matrix, that is $\boldsymbol{V}^\top \boldsymbol{V} = \boldsymbol{I}$ with the identity matrix $\boldsymbol{I}$ with appropriate dimension. In this orthogonal NMF, the encoding matrix should satisfy both orthogonality and nonnegativity constraints. As a result, we can find the encoding matrix which has a form of a cluster indicator matrix, where only one nonzero element exists in each row. Note that if a row has more than two nonzero elements, the matrix cannot satisfy the orthogonality constraint because some non-diagonal elements of $\boldsymbol{V}^\top \boldsymbol{V}$ also become nonzero. By applying orthogonal NMF, each data point is represented by only one basis vector, even if the point can be represented by some combination of several basis vectors. Because of this characteristic, orthogonal NMF can produce the results more appropriate for the clustering problems (Fig. 1(b)). Moreover, to minimize the error in these constraints, each basis vector indicates the center of the corresponding cluster, which is good for the interpretation for the clustering.

3.2 Orthogonal NMF and the previous approach

In general, orthogonal NMF involves a decomposition (1) as in NMF, but requires that $\boldsymbol{U}$ or $\boldsymbol{V}$ satisfies the orthogonality constraint such that $\boldsymbol{U}^\top \boldsymbol{U} = \boldsymbol{I}$ or $\boldsymbol{V}^\top \boldsymbol{V} = \boldsymbol{I}$ [Choi 2008]. As stated in the previous section, we consider the case where $\boldsymbol{V}^\top \boldsymbol{V} = \boldsymbol{I}$ is incorporated into the optimization (3). In this section, we review the existing orthogonal NMF method presented in [Ding et al. 2006].

Orthogonal NMF with $\boldsymbol{V}^{\mathsf{T}}\boldsymbol{V} = \boldsymbol{I}$ is formulated as the following optimization problem:

$$\arg\min_{U,V} \quad \mathcal{E} = \frac{1}{2}\left\|\boldsymbol{X} - \boldsymbol{U}\boldsymbol{V}^{\mathsf{T}}\right\|^2$$
$$\text{subject to} \quad \boldsymbol{V}^{\mathsf{T}}\boldsymbol{V} = \boldsymbol{I},\ \boldsymbol{U} \geq 0,\ \boldsymbol{V} \geq 0\,. \tag{14}$$

In [Ding et al. 2006], the constrained optimization problem (14) is solved by introducing a Lagrangian with a penalty term

$$\mathcal{L} = \frac{1}{2}\left\|\boldsymbol{X} - \boldsymbol{U}\boldsymbol{V}^{\mathsf{T}}\right\|^2 + \frac{1}{2}\mathrm{tr}\left\{\Lambda(\boldsymbol{V}^{\mathsf{T}}\boldsymbol{V} - \boldsymbol{I})\right\}, \tag{15}$$

where Λ is a symmetric matrix containing Lagrangian multipliers. The gradient of $\mathcal{L}$ with respect to the matrix $\boldsymbol{V}$ can be calculated by

$$\nabla_V \mathcal{L} = \boldsymbol{X}^{\mathsf{T}}\boldsymbol{U} + \boldsymbol{U}\boldsymbol{V}^{\mathsf{T}}\boldsymbol{V} + \boldsymbol{V}\Lambda. \tag{16}$$

The Lagrangian multiplier Λ in above equation is approximated to be $\Lambda = \boldsymbol{V}^{\mathsf{T}}\boldsymbol{X}^{\mathsf{T}}\boldsymbol{U} - \boldsymbol{U}^{\mathsf{T}}\boldsymbol{U}$, hence the gradient becomes

$$\begin{aligned}\nabla_V \mathcal{L} &= -\boldsymbol{X}^{\mathsf{T}}\boldsymbol{U} + \boldsymbol{V}\boldsymbol{U}^{\mathsf{T}}\boldsymbol{U} + \boldsymbol{V}(\boldsymbol{V}^{\mathsf{T}}\boldsymbol{X}^{\mathsf{T}}\boldsymbol{U} - \boldsymbol{U}^{\mathsf{T}}\boldsymbol{U}) \\ &= -\boldsymbol{X}^{\mathsf{T}}\boldsymbol{U} + \boldsymbol{V}\boldsymbol{U}^{\mathsf{T}}\boldsymbol{X}^{\mathsf{T}}\boldsymbol{U}\end{aligned} \tag{17}$$

Applying the relation (5) to the above gradient leads the following multiplicative update rule for the matrix $\boldsymbol{V}$,

$$\boldsymbol{V} \leftarrow \boldsymbol{V} \odot \frac{\boldsymbol{X}^{\mathsf{T}}\boldsymbol{U}}{\boldsymbol{V}\boldsymbol{U}^{\mathsf{T}}\boldsymbol{X}^{\mathsf{T}}\boldsymbol{U}}\,. \tag{18}$$

The update rule for the unconstrained matrix U is the same as (8). In the remaining of the paper, we will call this algorithm as DTPP.

The probabilistic interpretation of the result of orthogonal NMF is the same to the case of NMF, because both methods are based on the same two-factor decomposition model. We can determine the clusters for given data by applying (13) to the learned factor matrix $\boldsymbol{V}$ after an appropriate normalization.

4. ORTHOGONAL NMF ON STIEFEL MANIFOLD

In this section, we present a new method for orthogonal NMF based on the structure of the manifold arisen from constrained matrices. The constraint surface which is the set of $N \times K$ orthonormal matrices $\boldsymbol{V}$ such that $\boldsymbol{V}^{\mathsf{T}}\boldsymbol{V} = \boldsymbol{I}$ is known as the Stiefel manifold [Stiefel 1936]. We will introduce how the important quantities on the Stiefel manifold can be computed, and then derive the multiplicative update rule for the orthogonal NMF by using the gradient on the Stiefel manifold.

4.1 Gradient and geodesic on the Stiefel manifold

Minimizing (2) where $\boldsymbol{V}$ is constrained to the set of $N \times K$ matrices such that $\boldsymbol{V}^{\mathsf{T}}\boldsymbol{V} = \boldsymbol{I}$ was well studied in [Smith 1993; Edelman et al. 1998; Nishimori and Akaho 2005]. Here we present the equations for computing the gradients and geodesics, which are the important quantities in developing algorithms on the Stiefel manifold.

An equation defining tangents to the Stiefel manifold at a point $\boldsymbol{V}$ is obtained by differentiating $\boldsymbol{V}^{\top}\boldsymbol{V} = \boldsymbol{I}$, yielding

$$\boldsymbol{V}^{\top}\Delta + \Delta^{\top}\boldsymbol{V} = 0, \tag{19}$$

i.e., $\boldsymbol{V}^{\top}\Delta$ is *skew-symmetric*. The canonical metric on the Stiefel manifold [Edelman et al. 1998] is given by

$$g_c(\Delta, \Delta) = \mathrm{tr}\left\{\Delta^{\top}\left(I - \frac{1}{2}\boldsymbol{V}\boldsymbol{V}^{\top}\right)\Delta\right\}, \tag{20}$$

whereas the Euclidean metric is given by

$$g_e(\Delta, \Delta) = \mathrm{tr}\{\Delta^{\top}\Delta\}. \tag{21}$$

We define the partial derivatives of $\mathcal{E}$ with respect to the elements of $\boldsymbol{V}$ as

$$[\nabla_V \mathcal{E}]_{ij} = \frac{\partial \mathcal{E}}{\partial V_{ij}}. \tag{22}$$

For the function $\mathcal{E}$ (2) (with $\boldsymbol{U}$ fixed) defined on the Stiefel manifold, the gradient of $\mathcal{E}$ at $\boldsymbol{V}$ is defined to be the tangent vector $\tilde{\nabla}_V\mathcal{E}$ such that

$$\begin{aligned} g_e(\nabla_V \mathcal{E}, \Delta) &= \mathrm{tr}\ \{(\nabla_V \mathcal{E})^{\top}\Delta\} \\ &= g_c\ (\tilde{\nabla}_V\mathcal{E}\,,\ \Delta) \\ &= \mathrm{tr}\ \left\{(\tilde{\nabla}_V\mathcal{E})^{\top}\left(\boldsymbol{I} - \frac{1}{2}\boldsymbol{V}\boldsymbol{V}^{\top}\right)\Delta\right\}, \end{aligned} \tag{23}$$

for all tangent vectors Δ at $\boldsymbol{V}$.

Solving (23) for $\tilde{\nabla}_V\mathcal{E}$ such that $\boldsymbol{V}^{\top}\tilde{\nabla}_V\mathcal{E}$ is skew-symmetric yields

$$\tilde{\nabla}_V\mathcal{E} = \nabla_V\mathcal{E} - \boldsymbol{V}(\nabla_V\mathcal{E})^{\top}\boldsymbol{V}. \tag{24}$$

This gradient indicates the steepest direction the function $\mathcal{E}$ ascends in the Stiefel manifold.

The natural movement on the manifold can be determined by the geodesic of the manifold. The geodesic on the Stiefel manifold from a point $\boldsymbol{V}$ toward a direction ∇ with length t can be calculated as

$$\varphi(\boldsymbol{V}, \Delta, t) = \exp\{t(\boldsymbol{D}\boldsymbol{V}^{\top} - \boldsymbol{V}\boldsymbol{D}^{\top})\}\boldsymbol{V} \tag{25}$$

where $\boldsymbol{D} = (\boldsymbol{I} - \frac{1}{2}\boldsymbol{V}\boldsymbol{V}^{\top})\Delta$ [Edelman et al. 1998; Nishimori and Akaho 2005]. To minimize an objective function, we can follow the geodesic toward the opposite direction of the gradient of the function, which can be calculated by

$$\varphi(\boldsymbol{V}, -\nabla_V\mathcal{E}, t) = \exp\{t(\nabla_V\mathcal{E}\boldsymbol{V}^{\top} - \boldsymbol{V}\nabla_V\mathcal{E}^{\top})\}\boldsymbol{V}. \tag{26}$$

If we move along the geodesic, the learning does not escape from the Stiefel manifold, so we can find an exact orthogonal solution.

4.2 Multiplicative updates on Stiefel manifold

The most natural way to minimize the function while $\boldsymbol{V}$ satisfies the orthogonality constraints is following the geodesic on the Stiefel manifold along the direction of the

negative gradient (26). However, our problem has additional nonnegative constraints on the matrix $\boldsymbol{V}$, so we cannot simply follow the geodesics to find our solution. The formula for the geodesics is based on the calculation of the matrix exponential, which is not constrained to be nonnegative and not possible to decompose like (4). Therefore, we cannot guarantee that the solution from the geodesic search remains nonnegative.

Different to the case of geodesics, the gradient on the Stiefel manifold (24) consists of only sum and product of the nonnegative matrices, so we can build a multiplicative update rule from the gradient to preserve the nonnegativity of the factor matrix. The gradient of some manifold lies on the tangent plane of the manifold, therefore if we move along the gradient the resulting point slightly escapes from the manifold. In the case of the Stiefel manifold, we can rescale the size of each column to be 1 to prevent the solution to diverge [Nishimori and Akaho 2005].

The gradient on the Stiefel manifold for the objective function (14) can be computed by using (24) as

$$\begin{aligned}\tilde{\nabla}_V \mathcal{E} &= (-\boldsymbol{X}^\top \boldsymbol{U} + \boldsymbol{V}\boldsymbol{U}^\top \boldsymbol{U}) - \boldsymbol{V}(-\boldsymbol{X}^\top \boldsymbol{U} + \boldsymbol{V}\boldsymbol{U}^\top \boldsymbol{U})^\top \boldsymbol{V} \\ &= \boldsymbol{V}\boldsymbol{U}^\top \boldsymbol{X}\boldsymbol{V} - \boldsymbol{X}^\top \boldsymbol{U} \\ &= [\tilde{\nabla}_V \mathcal{E}]^+ - [\tilde{\nabla}_V \mathcal{E}]^- . \end{aligned} \tag{27}$$

Invoking the relation (5) with replacing ∇_V by $\tilde{\nabla}_V$ yields

$$\boldsymbol{V} \leftarrow \boldsymbol{V} \odot \frac{\boldsymbol{X}^\top \boldsymbol{U}}{\boldsymbol{V}\boldsymbol{U}^\top \boldsymbol{X}\boldsymbol{V}}. \tag{28}$$

We have to rescale the size of columns of $\boldsymbol{V}$ to be 1 by using $\boldsymbol{V} \leftarrow \boldsymbol{V}\boldsymbol{D}_V^{-1}$, where $\boldsymbol{D}_V = \mathrm{diag}(\mathbf{1}^\top \boldsymbol{D})$. The updating rule for the unconstrained matrix $\boldsymbol{U}$ is the same as (8). In the remaining of the paper, we will call this algorithm as ONMF.

5. EXPERIMENTS

We tested the ONMF algorithm on the six standard document datasets (*CSTR*, *k1a*, *k1b*, *re0*, and *re1*) and compared the performance with the standard NMF and the Ding et al.'s orthogonal NMF (DTPP) [Ding et al. 2006]. The statistics of the datasets are summarized in Table I.

We applied the stemming and stop-word removal for each dataset, and select 1,000 terms based on the mutual information with the class labels. Normalized-cut weighting [Xu et al. 2003] is applied to the input data matrix.

We use the clustering accuracy (CA) and normalized mutual information (NMI) to compare the performance of different clustering algorithms. To compute the clustering accuracy, we first applied Kuhn-Munkres maximal matching algorithm [Lovasz and Plummer 1986] to find the appropriate matching between the clustering result and the target labels. If we denote the true label for the document n to be c_n, and the matched label $\tilde{c}_n$, CA can be computed by

$$\mathrm{CA} = \frac{\sum_{n=1}^{N} \delta(c_n, \tilde{c}_n)}{N},$$

where $\delta(x, y) = 1$ for $x = y$ and $\delta(x, y) = 0$ for $x \neq y$.

Table I. Document dataset details are summarized, where for each dataset, the number of classes, the number of documents, the number of terms and the maximum/minimum cluster size are described.

Datasets	# classes	# documents	cluster size	
			max	min
CSTR	4	602	214	96
k1a	20	2340	494	9
k1b	6	2340	1389	60
re0	13	1504	608	11
re1	25	1657	371	10
wap	20	1560	341	5

NMI is based on the mutual information (MI) between the set of estimated clusters $\mathcal{C}$ and the set of ground-truth clusters $\tilde{\mathcal{C}}$, which can be calculated by

$$\mathrm{NMI}(\mathcal{C}, \tilde{\mathcal{C}}) = \frac{1}{\max(H(\mathcal{C}), H(\tilde{\mathcal{C}}))} \sum_i \sum_j \frac{|C_i \cap \tilde{C}_j|}{N} \log_2 \frac{N|C_i \cap \tilde{C}_j|}{|C_i||\tilde{C}_j|},$$

where C_i is the set of documents grouped into the i-th cluster, $\tilde{C}_i$ is the set of documents in the i-th ground truth cluster, and $H(\cdot)$ denotes the entropy.

Because the algorithms gave different results depending on the initial conditions, we calculated the mean value of CA and NMI over 100 runs with different initial conditions. For the CA (Table II), ONMF brought the best performance over the three algorithms except the *cstr* dataset. For the NMI (Table III), the overall tendency was similar to the CA, but NMF showed the best performance for the *wap* and *k1a* datasets. Imposing orthogonality with ONMF usually leads better clustering performances in our experiments.

The orthogonality of the matrix $\boldsymbol{V}$ is also measured by using the difference between the $\boldsymbol{V}^\top \boldsymbol{V}$ and the identity matrix, that is, $|\boldsymbol{V}^\top \boldsymbol{V} - \boldsymbol{I}|$. The changes of the orthogonality

Table II. Mean and standard deviations of clustering accuracies (CA) averaged over 100 trials for standard NMF, Ding et al.'s orthogonal NMF (DTPP), and proposed orthogonal NMF (ONMF) for six document datasets. For each case, the highest performance is marked with the boldface letters. The results significantly worse than the best performance are marked with *, where Wilcoxon rank-sum test with p-value 0.01 was used.

	NMF	DTPP	ONMF
CSTR	0.7568 ± 0.0796	**0.7844 ± 0.0911**	0.7268 ± 0.1017
wap	0.4744 ± 0.0289*	0.4281 ± 0.0796*	**0.4917 ± 0.0796**
k1a	0.4773 ± 0.0289*	0.4311 ± 0.0169*	**0.4907 ± 0.0402**
k1b	0.7896 ± 0.0744	0.6087 ± 0.0357*	**0.8109 ± 0.0636**
re0	0.3624 ± 0.0123*	0.3384 ± 0.0364*	**0.3691 ± 0.0175**
re1	0.4822 ± 0.0421*	0.4452 ± 0.0224*	**0.5090 ± 0.0458**

Table III. Mean and standard deviations of normalized mutual information (NMI) averaged over 100 trials for standard NMF, Ding et al.'s orthogonal NMF (DTPP), and proposed orthogonal NMF (ONMF) for six document datasets. The best performance is marked with the boldface letters. The results significantly worse than the best performance are marked with *, where Wilcoxon rank-sum test with p-value 0.01 was used.

	NMF	DTPP	ONMF
CSTR	0.6111 ± 0.1160*	0.6714 ± 0.0676	0.5316* ± 0.1484
wap	0.5658 ± 0.0138	0.5129 ± 0.0111*	0.5647 ± 0.0148
k1a	0.5716 ± 0.0117	0.5155 ± 0.0142*	0.5660 ± 0.0188
k1b	0.6260 ± 0.0535*	0.4817 ± 0.0197*	0.6758 ± 0.0579
re0	0.3169 ± 0.0126*	0.3106 ± 0.0183*	0.3252 ± 0.0157
re1	0.5172 ± 0.0195*	0.5117 ± 0.0140*	0.5319 ± 0.0256

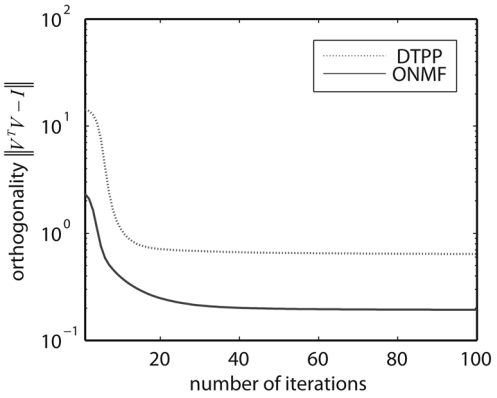

Figure 2. The convergence of orthogonality $\|\boldsymbol{V}^{\top}\boldsymbol{V}-\boldsymbol{I}\|$ of Ding et al.'s orthogonal NMF (DTPP) and our orthogonal NMF (ONMF) for the CSTR dataset.

over the iterations are measured and averaged for 100 trials. ONMF obtained better orthogonality values than DTPP for the most of the datasets. The change of orthogonality for the *CSTR* dataset is shown in Figure 2 for an example.

6. CONCLUSIONS

We have addressed orthogonal NMF which is better suited to the task of clustering, compared to NMF. We have developed a multiplicative update algorithm for orthogonal NMF, exploiting gradient on a Stiefel manifold. Numerical experiments on several document datasets confirmed the performance gain over the standard NMF as well as an existing orthogonal NMF algorithm. The method can be extended to a nonnegative 3-factor decomposition, which has recently applied to the problem of co-clustering [Yoo and Choi 2010].

ACKNOWLEDGMENTS

This work was supported by National Research Foundation of Korea (KRF-2008-313-D00939), National IT Industry Promotion Agency (NIPA) under the program of Software Engineering Technologies Development and Experts Education, and NRF WCU Program (Project No. R31-2008-000-10100-0).

REFERENCES

CHOI, S. 2008. Algorithms for orthogonal nonnegative matrix factorization. In *Proceedings of the International Joint Conference on Neural Networks (IJCNN)*. Hong Kong.

DING, C., HE, X., AND SIMON, H. D. 2005. On the equivalence of nonnegative matrix factorization and spectral clustering. In *Proceedings of the SIAM International Conference on Data Mining (SDM)*. Newport Beach, CA, 606–610.

DING, C., LI, T., PENG, W., AND PARK, H. 2006. Orthogonal nonnegative matrix tri-factorizations for clustering. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*. Philadelphia, PA.

EDELMAN, A., ARIAS, T., AND SMITH, S. T. 1998. The geometry of algorithms with orthogonality constraints. *SIAM J. Matrix Anal. Appl. 20*, 2, 303–353.

GAUSSIER, E. AND GOUTTE, C. 2005. Relation between PLSA and NMF and implications. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*. Salvador, Brazil.

LEE, D. D. AND SEUNG, H. S. 1999. Learning the parts of objects by non-negative matrix factorization. *Nature 401*, 788–791.

LEE, D. D. AND SEUNG, H. S. 2001. Algorithms for non-negative matrix factorization. In *Advances in Neural Information Processing Systems (NIPS)*. Vol. 13. MIT Press.

LI, S. Z., HOU, X. W., ZHANG, H. J., AND CHENG, Q. S. 2001. Learning spatially localized parts-based representation. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*. Kauai, Hawaii, 207–212.

LOVASZ, L. AND PLUMMER, M. 1986. *Matching Theory*. Akademiai Kiado.

NISHIMORI, Y. AND AKAHO, S. 2005. Learning algorithms utilizing quasi-geodesic flows on the Stiefel manifold. *Neurocomputing 67*, 106–135.

SHAHNAZ, F., BERRY, M., PAUCA, P., AND PLEMMONS, R. 2006. Document clustering using nonnegative matrix factorization. *Information Processing and Management 42*, 373–386.

SMITH, S. T. 1993. Geometric optimization methods for adaptive filtering. Ph.D. thesis, Harvard University.

STIEFEL, E. 1935-1936. Richtungsfelder und fernparallelismus in n-dimensionalem mannig faltigkeiten. *Commentarii Math. Helvetici 8*, 305–353.

XU, W., LIU, X., AND GONG, Y. 2003. Document clustering based on non-negative matrix factorization. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*. Toronto, Canada.

YOO, J. AND CHOI, S. 2008. Orthogonal nonnegative matrix factorization: Multiplicative updates on Stiefel manifolds. In *Proceedings of the 9th International Conference on Intelligent Data Engineering and Automated Learning (IDEAL)*. Daejeon, Korea.

YOO, J. AND CHOI, S. 2010. Orthogonal nonnegative matrix-tri-factorization for co-clustering: Multiplicative updates on stiefel manifolds. *Information Processing and Management*. in press.

Jiho Yoo received BS in Computer Science and Mathematics from POSTECH. Currently he is a PhD candidate at department of computer science in POSTECH. He is interested in the nonnegative matrix factorizations and its applications to document clustering, collaborative prediction, and musical source separation.

Seungjin Choi received the B.S. and M.S. degrees in Electrical Engineering from Seoul National University, Korea, in 1987 and 1989, respectively, and the Ph.D. degree in electrical engineering from the University of Notre Dame, Indiana, in 1996. He was a Visiting Assistant Professor in the Department of Electrical Engineering at University of Notre Dame, Indiana, during the Fall semester of 1996. He was with the Laboratory for Artificial Brain Systems, RIKEN, Japan, in 1997 and was an Assistant Professor in the School of Electrical and Electronics Engineering, Chungbuk National University from 1997 to 2000. He is currently a Professor of Computer Science at Pohang University of Science and Technology, Korea. His primary research interests include machine learning, Bayesian inference, and probabilistic models.