Algorithms for Orthogonal Nonnegative Matrix Factorization

Seungjin Choi

***Abstract*— Nonnegative matrix factorization (NMF) is a widely-used method for multivariate analysis of nonnegative data, the goal of which is decompose a data matrix into a basis matrix and an encoding variable matrix with all of these matrices allowed to have only nonnegative elements. In this paper we present simple algorithms for orthogonal NMF, where orthogonality constraints are imposed on basis matrix or encoding matrix. We develop multiplicative updates directly from the true gradient (natural gradient) in Stiefel manifold, whereas existing algorithms consider additive orthogonality constraints. Numerical experiments on face image data for a image representation task show that our orthogonal NMF algorithm preserves the orthogonality, while the goodness-of-fit (GOF) is minimized. We also apply our orthogonal NMF to a clustering task, showing that it works better than the original NMF, which is confirmed by experiments on several UCI repository data sets.**

I. Introduction

NONNEGATIVE matrix factorization (NMF) is a widely-used multivariate analysis method which is proven to be useful in learning a faithful representation of nonnegative data [10], [18]. Examples of nonnegative data include images, spectrograms, documents, and so on. Multiplicative updates proposed by Lee and Seung [10], [11] popularized NMF in diverse applications, including face recognition [15], [23], [24], audio and sound processing [4], [9], [22], medical imaging [1], [13], EEG classification for brain computer interface [12], document clustering [21], [25], bioinformatics [3], and so on. Earlier work on positive matrix factorization can be also found in [18].

NMF allows only non-subtractive combinations of nonnegative basis vectors to approximate the original nonnegative data, possibly providing a parts-based representation [10]. Incorporating extra constraints such as locality was shown to improve the decomposition, identifying better local features or providing more sparse representation [15]. Sparseness constraints were explicitly incorporated into NMF [8], [19]. Orthogonality constraints were implicitly or explicitly considered in NMF [6], [14], [27], [28]

In this paper we pay our attention to *orthogonal NMF* where orthogonality constraints between nonnegative basis vectors are incorporated into NMF multiplicative updates. We revisit several existing algorithms for orthogonal NMF and present a new algorithm which is directly derived from the true gradient (natural gradient) in Stiefel manifold, whereas existing algorithms consider additive orthogonality constraints. We also develop orthogonal NMF algorithms in the same manner, where orthogonality is imposed on encoding variables.

Seungjin Choi is with the Department of Computer Science, Pohang University of Science and Technology, Korea (phone: +82-54-279-2259; fax: +82-54-279-2299; email: seungjin@postech.ac.kr).

The rest of this paper is organized as follows. The next section describes NMF, especially in the case where least squares error measure is used. Two different derivations of multiplicative updates are illustrated, one of which is used to derive our orthogonal NMF algorithm. Sec. III reviews recognition model-based methods for NMF where a nonnegative projection matrix is estimated, instead of learning the basis matrix. Sec. IV presents our main contribution, deriving the orthogonal NMF algorithms where multiplicative updates preserve orthogonality between nonnegative basis vectors or between encoding variables. Numerical experiments on face image data are provided in Sec. V, stressing that orthogonal NMF yields more localized parts-based representation, compared to the original NMF. We also provide empirical results on several UCI repository data sets, emphasizing that the orthogonal NMF with imposing orthogonal constraints on encoding variables provide better clustering performnce. Finally conclusions are drawn in Sec. VI.

II. Nonnegative Matrix Factorization

Suppose that N observed data points, $\{\boldsymbol{x}_t\}$, $t = 1, \ldots, N$ are available. Denote the data matrix by $\boldsymbol{X} = [\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N] \in \mathbb{R}^{m \times N}$. NMF seeks a decomposition of the nonnegative data matrix $\boldsymbol{X} \geq 0$ that is of the form:

$$\boldsymbol{X} \approx \boldsymbol{A}\boldsymbol{S}, \tag{1}$$

where $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ contains basis vectors in its columns and $\boldsymbol{S} \in \mathbb{R}^{n \times N}$ is the associated encoding variable matrix. Both matrices $\boldsymbol{A}$ and $\boldsymbol{S}$ are restricted to have only nonnegative elements in the decomposition, i.e., $\boldsymbol{A} \geq 0$ and $\boldsymbol{S} \geq 0$.

We consider the squared Euclidean distance as a discrepancy measure between the data $\boldsymbol{X}$ and the model $\boldsymbol{A}\boldsymbol{S}$, leading to the following error function

$$\mathcal{E} = \|\boldsymbol{X} - \boldsymbol{A}\boldsymbol{S}\|^2 = \sum_i \sum_j \left\{ X_{ij} - [\boldsymbol{A}\boldsymbol{S}]_{ij} \right\}^2, \tag{2}$$

where $X_{ij} = [\boldsymbol{X}]_{ij}$ represents (i,j)-element of $\boldsymbol{X}$. Multiplicative updating rules for $\boldsymbol{A}$ and $\boldsymbol{S}$ (referred to as LS-NMF [11]) are given by

$$\boldsymbol{S} \leftarrow \boldsymbol{S} \odot \frac{\boldsymbol{A}^\top \boldsymbol{X}}{\boldsymbol{A}^\top \boldsymbol{A}\boldsymbol{S}}, \tag{3}$$

$$\boldsymbol{A} \leftarrow \boldsymbol{A} \odot \frac{\boldsymbol{X}\boldsymbol{S}^\top}{\boldsymbol{A}\boldsymbol{S}\boldsymbol{S}^\top}, \tag{4}$$

where $\odot$ denotes the Hadamard product (elementwise product) and $\frac{\boldsymbol{A}}{\boldsymbol{B}}$ represents the elementwise division, i.e., $\left[\frac{\boldsymbol{A}}{\boldsymbol{B}}\right]_{ij} = \frac{A_{ij}}{B_{ij}}$.

A. Derivation 1

The core idea used in deriving multiplicative updates (3) and (4), which is introduced in [11], is as follows. Suppose that the gradient of an error function has a decomposition that is of the form

$$\nabla \mathcal{E} = [\nabla \mathcal{E}]^+ - [\nabla \mathcal{E}]^-, \tag{5}$$

where $[\nabla \mathcal{E}]^+ > 0$ and $[\nabla \mathcal{E}]^- > 0$. Then the updating rule that is of the form

$$\Theta \leftarrow \Theta \odot \frac{[\nabla \mathcal{E}]^-}{[\nabla \mathcal{E}]^+}, \tag{6}$$

preserves the nonnegativity of the parameter Θ, while $\nabla \mathcal{E} = 0$ when the convergence is achieved.

In LS-NMF, gradients with respect to $\boldsymbol{A}$ with $\boldsymbol{S}$ fixed are given by

$$[\nabla_A \mathcal{E}]^+ = \boldsymbol{A}\boldsymbol{S}\boldsymbol{S}^\top, \quad [\nabla_A \mathcal{E}]^- = \boldsymbol{X}\boldsymbol{S}^\top,$$

leading to (4) through the relation (6). Gradients with respect to $\boldsymbol{S}$ with $\boldsymbol{A}$ fixed are also given by

$$[\nabla_S \mathcal{E}]^+ = \boldsymbol{A}^\top \boldsymbol{A}\boldsymbol{S}, \quad [\nabla_S \mathcal{E}]^- = \boldsymbol{A}^\top \boldsymbol{X},$$

yielding (3).

B. Derivation 2

Alternatively we can derive the multiplicative updates (3) and (4) from KKT conditions. To this end, we consider the Lagrangian $\mathcal{L}$ that is of the form

$$\mathcal{L} = \frac{1}{2}\|\boldsymbol{X} - \boldsymbol{A}\boldsymbol{S}\|^2 - \mathrm{tr}\left\{\boldsymbol{\Lambda}\boldsymbol{A}^\top\right\} - \mathrm{tr}\left\{\boldsymbol{\Omega}\boldsymbol{S}^\top\right\}, \tag{7}$$

where $\boldsymbol{\Lambda}$ and $\boldsymbol{\Omega}$ are Lagrangian multiplier matrices.

The KKT conditions require:

$$\frac{1}{2}\frac{\partial \|\boldsymbol{X} - \boldsymbol{A}\boldsymbol{S}\|^2}{\partial S_{ij}} = \Omega_{ij}, \tag{8}$$

$$\frac{1}{2}\frac{\partial \|\boldsymbol{X} - \boldsymbol{A}\boldsymbol{S}\|^2}{\partial A_{ij}} = \Lambda_{ij}, \tag{9}$$

as optimality conditions and

$$\Omega_{ij} S_{ij} = 0, \tag{10}$$

$$\Lambda_{ij} A_{ij} = 0, \tag{11}$$

as complementary slackness conditions. Incorporating (8) and (9) into (10) and (11), respectively leads to

$$\left[\boldsymbol{A}^\top \boldsymbol{A}\boldsymbol{S} - \boldsymbol{A}^\top \boldsymbol{X}\right]_{ij} S_{ij} = 0, \tag{12}$$

$$\left[\boldsymbol{A}\boldsymbol{S}\boldsymbol{S}^\top - \boldsymbol{X}\boldsymbol{S}^\top\right]_{ij} A_{ij} = 0, \tag{13}$$

which suggests iterative algorithms (3) and (4).

III. Nonnegative Projections

In this section we consider a recognition model, $\boldsymbol{Y} = \boldsymbol{W}^\top \boldsymbol{X}$, where $\boldsymbol{X} \geq 0$ and $\boldsymbol{W} \geq 0$. We give an overview of two nonnegative projection methods, including projective NMF [28] and nonnegative Hebbian rule [27]. Then we address an orthogonality issue in these algorithms that were not clearly discussed in [27], [28].

A. Projective NMF

A common derivation of subspace analysis [17], [26] involves a linear transform which minimizes the reconstruction error. In other words, subspace analysis seeks a linear transform $\boldsymbol{W} \in \mathbb{R}^{m \times n}$ which involves the following optimization problem:

$$\arg\min_{\boldsymbol{W}} \|\boldsymbol{X} - \boldsymbol{W}\boldsymbol{W}^\top \boldsymbol{X}\|^2. \tag{14}$$

Yuan and Oja [28] proposed projective NMF where nonnegativity constraint on $\boldsymbol{W}$ was incorporated into subspace analysis, leading to the following optimization problem:

$$\arg\min_{\boldsymbol{W} \geq 0} \|\boldsymbol{X} - \boldsymbol{W}\boldsymbol{W}^\top \boldsymbol{X}\|^2. \tag{15}$$

The gradient of $\|\boldsymbol{X} - \boldsymbol{W}\boldsymbol{W}^\top \boldsymbol{X}\|^2$ with respect to $\boldsymbol{W}$ is given by

$$\begin{aligned}\frac{\partial}{\partial \boldsymbol{W}}&\left[\|\boldsymbol{X} - \boldsymbol{W}\boldsymbol{W}^\top \boldsymbol{X}\|^2\right] \\ &= -4\boldsymbol{X}\boldsymbol{X}^\top \boldsymbol{W} + 2\boldsymbol{W}\boldsymbol{W}^\top \boldsymbol{X}\boldsymbol{X}^\top \boldsymbol{W} \\ &\quad +2\boldsymbol{X}\boldsymbol{X}^\top \boldsymbol{W}\boldsymbol{W}^\top \boldsymbol{W},\end{aligned} \tag{16}$$

which suggests the following multiplicative updating rule:

$$\boldsymbol{W} \leftarrow \boldsymbol{W} \odot \frac{\boldsymbol{X}\boldsymbol{X}^\top \boldsymbol{W}}{\boldsymbol{W}\boldsymbol{W}^\top \boldsymbol{X}\boldsymbol{X}^\top \boldsymbol{W} + \boldsymbol{X}\boldsymbol{X}^\top \boldsymbol{W}\boldsymbol{W}^\top \boldsymbol{W}}. \tag{17}$$

It was shown in [28] that projective NMF gives more localized and sparse parts-based representation of images, compared to the standard NMF. Empirical results also showed that column vectors of $\boldsymbol{W}$ are close to be orthogonal each other.

B. Nonnegative Hebbian Rule

Oja's rule [16] is a widely-used PCA method, extracting the largest principal component in a single neuron linear network described by $y_t = \boldsymbol{w}^\top \boldsymbol{x}_t$. The updating rule for the weight vector $\boldsymbol{w}$ is given by

$$\boldsymbol{w}_{t+1} = \boldsymbol{w}_t + \eta\{y_t \boldsymbol{x}_t - y_t^2 \boldsymbol{w}_t\}, \tag{18}$$

where $\eta > 0$ is a learning rate.

The normalized Hebbian rule is given by

$$\widetilde{\boldsymbol{w}}_{t+1} = \boldsymbol{w}_t + \eta\, y_t \boldsymbol{x}_t, \tag{19}$$

$$\boldsymbol{w}_{t+1} = \frac{\widetilde{\boldsymbol{w}}_{t+1}}{\|\widetilde{\boldsymbol{w}}_{t+1}\|}. \tag{20}$$

Oja's rule (18) is a consequence of combining Hebbian update (19) and a linear approximation of the normalization (20). That is,

$$\begin{aligned}\boldsymbol{w}_{t+1} &= \{\boldsymbol{w}_t + \eta\, y_t \boldsymbol{x}_t\}\left\{1 + 2\eta y_t^2 + \mathcal{O}(\eta^2)\right\}^{-\frac{1}{2}} \\ &\approx \{\boldsymbol{w}_t + \eta\, y_t \boldsymbol{x}_t\}\left\{1 - \eta\, y_t^2\right\} \\ &= \boldsymbol{w}_t + \eta\{y_t \boldsymbol{x}_t - y_t^2 \boldsymbol{w}_t\} + \mathcal{O}(\eta^2),\end{aligned}$$

which is Oja's rule (18).

The straightforward extension of (19) and (20) to multiple outputs, i.e., $\boldsymbol{y} = \boldsymbol{W}^\top \boldsymbol{x}$, is given by

$$\widetilde{\boldsymbol{W}}_{t+1} = \boldsymbol{W}_t + \eta\, \boldsymbol{x}_t \boldsymbol{y}_t^\top, \tag{21}$$

$$\boldsymbol{W}_{t+1} = \widetilde{\boldsymbol{W}}_{t+1} \left[\widetilde{\boldsymbol{W}}_{t+1}^\top \widetilde{\boldsymbol{W}}_{t+1}\right]^{-\frac{1}{2}}. \tag{22}$$

In the same manner, we combine (21) and (22) with a linear approximation of the normalization, which leads to

$$\begin{aligned}
&\boldsymbol{W}_{t+1} \\
&= \widetilde{\boldsymbol{W}}_{t+1} \left[\boldsymbol{I} + \eta\, \boldsymbol{y}_t \boldsymbol{x}_t^\top \boldsymbol{W}_t + \eta\, \boldsymbol{W}_t^\top \boldsymbol{x}_t \boldsymbol{y}_t + \mathcal{O}(\eta^2)\right]^{-\frac{1}{2}} \\
&\approx \widetilde{\boldsymbol{W}}_{t+1} \left[\boldsymbol{I} - \frac{1}{2}\eta \left\{\boldsymbol{y}_t \boldsymbol{x}_t^\top \boldsymbol{W}_t + \boldsymbol{W}_t^\top \boldsymbol{x}_t \boldsymbol{y}_t\right\}\right] \\
&\approx \boldsymbol{W}_t + \eta \left\{\boldsymbol{x}_t \boldsymbol{x}_t^\top \boldsymbol{W}_t - \boldsymbol{W}_t \boldsymbol{W}_t^\top \boldsymbol{x}_t \boldsymbol{x}_t^\top \boldsymbol{W}_t\right\}.
\end{aligned} \tag{23}$$

The batch version of (23) is given by

$$\boldsymbol{W} \leftarrow \boldsymbol{W} + \eta \left\{\boldsymbol{X}\boldsymbol{X}^\top \boldsymbol{W} - \boldsymbol{W}\boldsymbol{W}^\top \boldsymbol{X}\boldsymbol{X}^\top \boldsymbol{W}\right\}. \tag{24}$$

As in projective NMF, the nonnegative Hebbian rule was proposed by Yang and Laaksonen [27]. The multiplicative update for nonnegative Hebbian learning is given by

$$\boldsymbol{W} \leftarrow \boldsymbol{W} \odot \frac{\boldsymbol{X}\boldsymbol{X}^\top \boldsymbol{W}}{\boldsymbol{W}\boldsymbol{W}^\top \boldsymbol{X}\boldsymbol{X}^\top \boldsymbol{W}}. \tag{25}$$

Note that nonnegative Hebbian rule (25) is very close to projective NMF (17), where only difference is the term $\boldsymbol{X}\boldsymbol{X}^\top \boldsymbol{W}\boldsymbol{W}^\top \boldsymbol{W}$ in the denominator of (17) that is known to have little effect in learning [27].

IV. Algorithms for Orthogonal NMF

We present two orthogonal NMF algorithms, where one is for $\boldsymbol{A}^\top \boldsymbol{A} = \boldsymbol{I}$ and the other is for $\boldsymbol{S}\boldsymbol{S}^\top = \boldsymbol{I}$.

A. Orthogonal NMF: $\boldsymbol{A}^\top \boldsymbol{A} = \boldsymbol{I}$

One way to impose the orthogonality constraint on $\boldsymbol{A}$ is to consider the Lagrangian $\widetilde{\mathcal{E}}$

$$\widetilde{\mathcal{E}} = \frac{1}{2}\left\|\boldsymbol{X} - \boldsymbol{A}\boldsymbol{S}\right\|^2 + \frac{1}{2}\mathrm{tr}\left\{\boldsymbol{\Omega}\left(\boldsymbol{A}^\top \boldsymbol{A} - \boldsymbol{I}\right)\right\}, \tag{26}$$

where $\boldsymbol{\Omega} \in \mathbb{R}^{n \times n}$ is the Lagrangian multiplier matrix. Incorporating the optimal value of $\boldsymbol{\Omega}$, Ding *et al.* [6] proposed the following algorithm

$$\boldsymbol{S} \leftarrow \boldsymbol{S} \odot \frac{\boldsymbol{A}^\top \boldsymbol{X}}{\boldsymbol{A}^\top \boldsymbol{A}\boldsymbol{S}}, \tag{27}$$

$$\boldsymbol{A} \leftarrow \boldsymbol{A} \odot \left(\frac{\boldsymbol{X}\boldsymbol{S}^\top}{\boldsymbol{A}\boldsymbol{A}^\top \boldsymbol{X}\boldsymbol{S}^\top}\right)^{\cdot\frac{1}{2}}, \tag{28}$$

where $(\cdot)^{\cdot\frac{1}{2}}$ denotes the elementwise square root. Updates (27) and (28) are referred to as Ding-Ti-Peng-Park (DTPP) algorithm.

In contrast, we directly use the result of the true gradient in Stiefel manifold which is a parameter space with the orthogonality constraint $\boldsymbol{A}^\top \boldsymbol{A} = \boldsymbol{I}$. In Stiefel manifold, the true gradient (or natural gradient) [7] is computed by

$$\begin{aligned}
\widetilde{\nabla}_A \mathcal{E} &= \nabla_A \mathcal{E} - \boldsymbol{A}\left[\nabla_A \mathcal{E}\right]^\top \boldsymbol{A} \\
&= \left[\widetilde{\nabla}_A \mathcal{E}\right]^+ - \left[\widetilde{\nabla}_A \mathcal{E}\right]^- \\
&= \left[\boldsymbol{A}\boldsymbol{S}\boldsymbol{X}^\top \boldsymbol{A}\right] - \left[\boldsymbol{X}\boldsymbol{S}^\top\right].
\end{aligned} \tag{29}$$

Thus the multiplicative updates for our ONMF are of the form

$$\boldsymbol{S} \leftarrow \boldsymbol{S} \odot \frac{\boldsymbol{A}^\top \boldsymbol{X}}{\boldsymbol{A}^\top \boldsymbol{A}\boldsymbol{S}}, \tag{30}$$

$$\boldsymbol{A} \leftarrow \boldsymbol{A} \odot \frac{\boldsymbol{X}\boldsymbol{S}^\top}{\boldsymbol{A}\boldsymbol{S}\boldsymbol{X}^\top \boldsymbol{A}}, \tag{31}$$

where multiplicative update for $\boldsymbol{S}$ is the same as the one for the standard NMF. Algorithms discussed in this paper are summarized in Table I, where the ONMF algorithm in this case is referred to as ONMF-A.

B. Orthogonal NMF: $\boldsymbol{S}\boldsymbol{S}^\top = \boldsymbol{I}$

We can easily derive the ONMF algorithm with preserving $\boldsymbol{S}\boldsymbol{S}^\top = \boldsymbol{I}$, which is referred to as ONMF-S, in the same manner. We compute the true gradient in Stiefel manifold where $\boldsymbol{S}\boldsymbol{S}^\top = \boldsymbol{I}$ is satisfied

$$\begin{aligned}
\widetilde{\nabla}_S \mathcal{E} &= \nabla_S \mathcal{E} - \boldsymbol{S}\left[\nabla_S \mathcal{E}\right]^\top \boldsymbol{S} \\
&= \left[\widetilde{\nabla}_S \mathcal{E}\right]^+ - \left[\widetilde{\nabla}_S \mathcal{E}\right]^- \\
&= \left[\boldsymbol{S}\boldsymbol{X}^\top \boldsymbol{A}\boldsymbol{S}\right] - \left[\boldsymbol{A}^\top \boldsymbol{X}\right],
\end{aligned} \tag{32}$$

which leads to the multiplicative updating rule for $\boldsymbol{S}$ that has the form

$$\boldsymbol{S} \leftarrow \boldsymbol{S} \odot \frac{\boldsymbol{A}^\top \boldsymbol{X}}{\boldsymbol{S}\boldsymbol{X}^\top \boldsymbol{A}\boldsymbol{S}}. \tag{33}$$

The updating rule for $\boldsymbol{A}$ is identical to the one in the standard NMF.

V. Numerical Experiments

A. Experiment 1

We use ORL face image dataset [20] which consists of 400 face images (40 people and 10 images for each person) of size 45×35. Each face image is converted to a vector $\boldsymbol{x}_t \in \mathbb{R}^{1575}$, leading to the data matrix $\boldsymbol{X} \in \mathbb{R}^{1575 \times 400}$. We apply the standard NMF, our ONMF, and Ding-Ti-Peng-Park (DTPP) algorithm [6] to this $\boldsymbol{X}$, for comparative study in terms of the orthogonality and the goodness-of-fit (GOF). The intrinsic dimension $n = 25$ was chosen to learn 25 bases (associated with column vectors of $\boldsymbol{A}$). These bases are reshaped back into 45×35 images which are shown in Fig. 1, where basis images computed by our ONMF algorithm are much more localized parts, compared to the standard NMF. We measure the orthogonality of bases by $\|\boldsymbol{A}^\top \boldsymbol{A} - \boldsymbol{I}\|$ and GOF by $\|\boldsymbol{X} - \boldsymbol{A}\boldsymbol{S}\|$. Results are shown in Fig. 2, where our ONMF algorithm preserves the orthogonality slightly better than DTPP algorithm, while its GOF is comparable to the standard NMF.

TABLE I

SUMMARY OF ALGORITHMS.

Algorithm	Update rule
NMF [11]	$\boldsymbol{A} \leftarrow \boldsymbol{A} \odot \frac{\boldsymbol{X}\boldsymbol{S}^\top}{\boldsymbol{A}\boldsymbol{S}\boldsymbol{S}^\top}, \quad \boldsymbol{S} \leftarrow \boldsymbol{S} \odot \frac{\boldsymbol{A}^\top \boldsymbol{X}}{\boldsymbol{A}^\top \boldsymbol{A}\boldsymbol{S}}$
PNMF [28]	$\boldsymbol{W} \leftarrow \boldsymbol{W} \odot \frac{\boldsymbol{X}\boldsymbol{X}^\top \boldsymbol{W}}{\boldsymbol{W}\boldsymbol{W}^\top \boldsymbol{X}\boldsymbol{X}^\top \boldsymbol{W} + \boldsymbol{X}\boldsymbol{X}^\top \boldsymbol{W}\boldsymbol{W}^\top \boldsymbol{W}}$
NHL [27]	$\boldsymbol{W} \leftarrow \boldsymbol{W} \odot \frac{\boldsymbol{X}\boldsymbol{X}^\top \boldsymbol{W}}{\boldsymbol{W}\boldsymbol{W}^\top \boldsymbol{X}\boldsymbol{X}^\top \boldsymbol{W}}$
DTPP [6]	$\boldsymbol{A} \leftarrow \boldsymbol{A} \odot \left(\frac{\boldsymbol{X}\boldsymbol{S}^\top}{\boldsymbol{A}\boldsymbol{A}^\top \boldsymbol{X}\boldsymbol{S}^\top}\right)^{\cdot \frac{1}{2}}, \quad \boldsymbol{S} \leftarrow \boldsymbol{S} \odot \frac{\boldsymbol{A}^\top \boldsymbol{X}}{\boldsymbol{A}^\top \boldsymbol{A}\boldsymbol{S}}$
ONMF-A (proposed method)	$\boldsymbol{A} \leftarrow \boldsymbol{A} \odot \frac{\boldsymbol{X}\boldsymbol{S}^\top}{\boldsymbol{A}\boldsymbol{S}\boldsymbol{X}^\top \boldsymbol{A}}, \quad \boldsymbol{S} \leftarrow \boldsymbol{S} \odot \frac{\boldsymbol{A}^\top \boldsymbol{X}}{\boldsymbol{A}^\top \boldsymbol{A}\boldsymbol{S}}$
ONMF-S (proposed method)	$\boldsymbol{A} \leftarrow \boldsymbol{A} \odot \frac{\boldsymbol{X}\boldsymbol{S}^\top}{\boldsymbol{A}\boldsymbol{S}\boldsymbol{S}^\top}, \quad \boldsymbol{S} \leftarrow \boldsymbol{S} \odot \frac{\boldsymbol{A}^\top \boldsymbol{X}}{\boldsymbol{S}\boldsymbol{X}^\top \boldsymbol{A}\boldsymbol{S}}$

B. Experiment 2

We apply the standard NMF and our ONMF-S to a clustering task. We perform the sum-to-one normalization in the following way

$$\boldsymbol{A} \leftarrow \boldsymbol{A}\boldsymbol{D}_A^{-1}, \quad \boldsymbol{S} \leftarrow \boldsymbol{D}_A \boldsymbol{S},$$

where $\boldsymbol{D}_A = \mathrm{diag}(\boldsymbol{1}^\top \boldsymbol{A})$. Then, a data point $\boldsymbol{x}_j$ is assigned to cluster i^* if

$$i^* = \arg\max_i S_{ij}.$$

In fact, it was shown in [5] that NMF with $\boldsymbol{S}\boldsymbol{S}^\top = \boldsymbol{I}$ is equivalent to k-means clustering in the sense that they involve the same objective function. That is why ONMF-S is expected to provide better clustering performance than the standard NMF. We evaluate the clustering performance of the standard NMF and ONMF-S in terms of classification accuracy, using three UCI data sets (Iris, Wine, Breast cancer) [2]. Table II outlines clustering performance averaged over 100 independent runs (with different initial conditions), showing that ONMF-S indeed provides better clustering performance than the standard NMF.

TABLE II

COMPARISON OF CLUSTERING PERFORMANCE IN TERMS OF A CLASSIFICATION ACCURACY.

	NMF	ONMF
Iris	0.7653	**0.7995**
Wine	0.7125	**0.8746**
Breast cancer (wdbc)	0.7343	**0.8028**

VI. CONCLUSIONS

We have presented simple multiplicative updates for orthogonal NMF where orthogonality between nonnegative basis vectors are imposed in learning the decomposition. The core idea was to directly use the true gradient in Stiefel manifold in developing multiplicative updates for ONMF. Empirical results have confirmed that the orthogonality was preserved and our algorithm provided localized parts-based representations.

Acknowledgments: This work was supported by Korea Ministry of Knowledge Economy under the ITRC support program supervised by the IITA (IITA-2008-C1090-0801-0045) and National Core Research Center for Systems Bio-Dynamics. The author thanks to Mr. Jiho Yoo for his help with Experiment 2.

REFERENCES

[1] J. H. Ahn, S. Kim, J. H. Oh, and S. Choi, "Multiple nonnegative-matrix factorization of dynamic PET images," in *Proceedings of Asian Conference on Computer Vision*, 2004.

[2] A. Asuncion and D. Newman, "UCI machine learning repository," 2007. [Online]. Available: http://www.ics.uci.edu/∼mlearn/MLRepository.html

[3] J. P. Brunet, P. Tamayo, T. R. Golub, and J. P. Mesirov, "Metagenes and molecular pattern discovery using matrix factorization," *Proceedings of the National Academy of Sciences, USA*, vol. 101, no. 12, pp. 4164–4169, 2004.

[4] Y. C. Cho and S. Choi, "Nonnegative features of spectro-temporal sounds for classfication," *Pattern Recognition Letters*, vol. 26, no. 9, pp. 1327–1336, 2005.

[5] C. Ding, X. He, and H. D. Simon, "On the equivalence of nonnegative matrix factorization and spectral clustering," in *Proceedings of the SIAM International Conference on Data Mining*, Newport Beach, CA, 2005, pp. 606–610.

[6] C. Ding, T. Li, W. Peng, and H. Park, "Orthogonal nonnegative matrix tri-factorizations for clustering," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Philadelphia,PA, 2006.

[7] A. Edelman, T. Arias, and S. T. Smith, "The geometry of algorithms with orthogonality constraints," *SIAM J. Matrix Anal. Appl.*, vol. 20, no. 2, pp. 303–353, 1998.

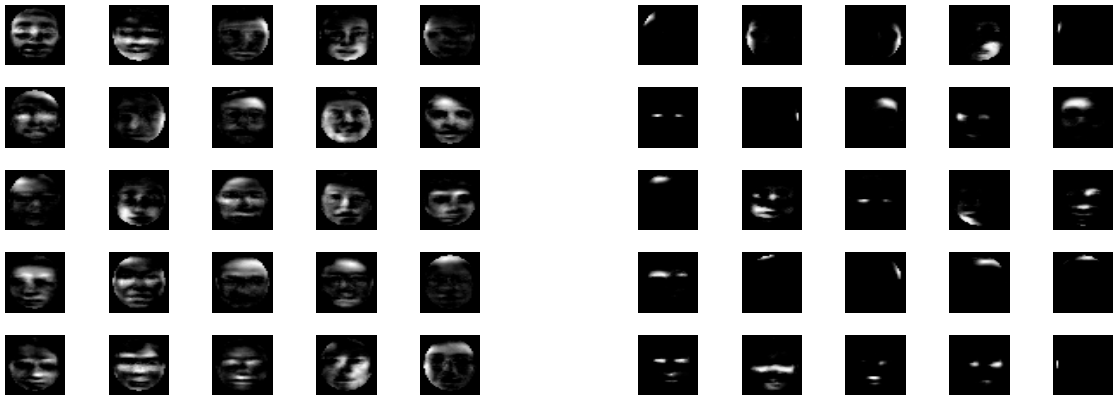
[8] P. O. Hoyer, "Non-negative matrix factorization with sparseness constraints," *Journal of Machine Learning Research*, vol. 5, pp. 1457–1469, 2004.

[9] M. Kim and S. Choi, "Monaural music source separation: Nonnegativity, sparseness, and shift-invariance," in *Proceedings of the International Conference on Independent Component Analysis and Blind Signal Separation*. Charleston, South Carolina: Springer, 2006, pp. 617–624.

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

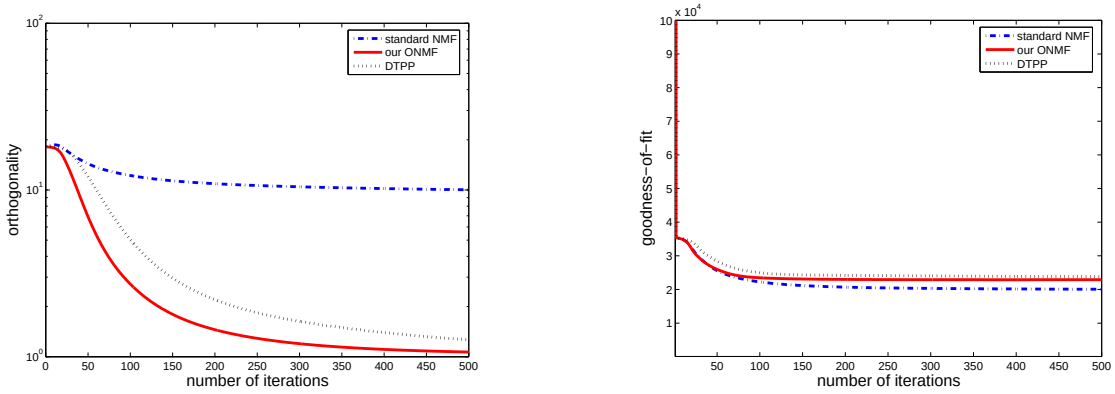
[11] ——, "Algorithms for non-negative matrix factorization," in *Advances in Neural Information Processing Systems*, vol. 13. MIT Press, 2001.

[12] H. Lee, A. Cichocki, and S. Choi, "Nonnegative matrix factorization for motor imagery EEG classification," in *Proceedings of the International Conference on Artificial Neural Networks*. Athens, Greece: Springer, 2006.



(a) (b)

Fig. 1. Basis images computed by: (a) NMF; (b) our ONMF.



(a) (b)

Fig. 2. Comparison of NMF, our ONMF, and DTPP in terms of: (a) orthogonality; (b) goodness-of-fit.

[13] J. S. Lee, D. D. Lee, S. Choi, and D. S. Lee, "Application of non-negative matrix factorization to dynamic positron emission tomography," in *Proceedings of the International Conference on Independent Component Analysis and Blind Signal Separation*, San Diego, California, 2001, pp. 629–632.

[14] H. Li, T. Adali, W. Wang, D. Emge, and A. Cichocki, "Non-negative matrix factorization with orthogonality constraints and its applications to Raman spectroscopy," *Journal of VLSI Signal Processing*, vol. 48, pp. 83–97, 2007.

[15] S. Z. Li, X. W. Hou, H. J. Zhang, and Q. S. Cheng, "Learning spatially localized parts-based representation," in *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, Kauai, Hawaii, 2001, pp. 207–212.

[16] E. Oja, "A simplified neuron model as a principal component analyzer," *J. Mathematical Biogology*, vol. 15, pp. 267–273, 1982.

[17] ——, "Neural networks, principal component analysis, and subspaces," *International Journal of Neural Systems*, vol. 1, pp. 61–68, 1989.

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

[19] A. Pascual-Montano, J. M. Carazo, K. K. D. Lehmann, and R. D. Pascual-Margui, "Nonsmooth nonnegtive matrix facotorization (nsNMF)," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 28, no. 3, pp. 403–415, 2006.

[20] F. Samaria and A. Harter, "Parameterisation of a stochastic model for human face identification," in *Proceedings of the 2nd IEEE Workshop on Applications of Computer Vision*, Sarasota, FL, 1994.

[21] F. Shahnaz, M. Berry, P. Pauca, and R. Plemmons, "Document clustering using nonnegative matrix factorization," *Information Processing and Management*, vol. 42, pp. 373–386, 2006.

[22] P. Smaragdis and J. C. Brown, "Non-negative matrix factorization for polyphonic music transcription," in *Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, NY, 2003, pp. 177–180.

[23] Y. Wang and Y. Jia, "Non-negative matrix factorization framework for face recognition," *Journal of Pattern Recognition and Artificial Intelligence*, vol. 19, no. 4, pp. 495–511, 2005.

[24] Y. Wang, Y. Jia, C. Hu, and M. Turk, "Fisher non-negative matrix factorization for learning local features," in *Proceedings of the Asian Conference on Computer Vision*, Jeju Island, Korea, 2004.

[25] W. Xu, X. Liu, and Y. Gong, "Document clustering based on non-negative matrix factorization," in *Proceedings of the ACM SIGIR International Conference on Research and Development in Information Retrieval*, Toronto, Canada, 2003.

[26] B. Yang, "Projection approximation subsapce tracking," *IEEE Trans. Signal Processing*, vol. 43, no. 1, pp. 95–107, 1995.

[27] Z. Yang and J. Laaksonen, "Multiplicative updates for non-negative projections," *Neurocomputing*, vol. 71, pp. 363–373, 2007.

[28] Z. Yuan and E. Oja, "Projective nonnegative matrix factorization for image compression and feature extraction," in *Proceedings of Scandinavian Conference on Image Analysis*, 2005, pp. 333–342.