Global Convergence of Modified Multiplicative Updates for Nonnegative Matrix Factorization

Author(s): Norikazu Takahashi and Ryota Hibi

Journal: Computational Optimization and Applications
Volume: 57
Pages: 417–440
Month: March
Year: 2014
DOI: 10.1007/s10589-013-9593-0
Published Version: http://link.springer.com/article/10.1007%2Fs10589-013-9593-0

Computational Optimization and Applications manuscript No.
(will be inserted by the editor)

Global convergence of modified multiplicative updates for nonnegative matrix factorization

Norikazu Takahashi · **Ryota Hibi**

Received: date / Accepted: date

Abstract Nonnegative matrix factorization (NMF) is the problem of approximating a given nonnegative matrix by the product of two nonnegative matrices. The multiplicative updates proposed by Lee and Seung are widely used as efficient computational methods for NMF. However, the global convergence of these updates is not formally guaranteed because they are not defined for all pairs of nonnegative matrices. In this paper, we consider slightly modified versions of the original multiplicative updates and study their global convergence properties. The only difference between the modified updates and the original ones is that the former do not allow variables to take values less than a user-specified positive constant. Using Zangwill's global convergence theorem, we prove that any sequence of solutions generated by either of those modified updates has at least one convergent subsequence and the limit of any convergent subsequence is a stationary point of the corresponding optimization problem. Furthermore, we propose algorithms based on the modified updates that always stop within a finite number of iterations.

Keywords Nonnegative matrix factorization · Multiplicative update · Global convergence · Finite termination

Part of this paper was presented at 18th International Conference on Neural Information Processing, Shanghai, China, in November 2011 [14].

N. Takahashi
Department of Informatics, Kyushu University, 744 Motooka, Nishi-ku, Fukuoka 819–0395 Japan
Institute of Systems, Information Technologies and Nanotechnologies, 2-1-22 Momochihama, Sawara-ku, Fukuoka 814-0001 Japan
Present address: Graduate School of Natural Science and Technology, Okayama University, 3-1-1 Tsushima-naka, Kita-ku, Okayama 700-8530 Japan
E-mail: takahashi@cs.okayama-u.ac.jp

R. Hibi
Department of Informatics, Kyushu University, 744 Motooka, Nishi-ku, Fukuoka 819-0395 Japan
Present address: Mitsubishi UFJ Morgan Stanley Securities Co., Ltd.

1 Introduction

Nonnegative matrix factorization (NMF) [16,17] is the problem of approximating a given large nonnegative matrix V by the product WH of two flat nonnegative matrices W and H. If we consider the columns of V as data vectors, the columns of W and those of H are interpreted as a set of nonnegative basis vectors and a set of nonnegative coefficient vectors, respectively. Each data vector is thus reproduced approximately by a linear combination of the basis vectors with coefficients stored in the corresponding column of H. In this sense, NMF can generate a reduced representation of the original data. Moreover, the basis vectors often represent parts of the object because of the nonnegativity constraints [16]. This is a significant difference between NMF and other factorization methods such as principal component analysis. So far, NMF has been successfully applied to various problems in machine learning, signal processing and so on [2,5,7,15,16,20,22,25].

Usually, NMF is formulated as a constrained optimization problem in which the approximation error has to be minimized with respect to W and H subject to the nonnegativity of these matrices. Lee and Seung [17] considered the cases where the approximation error is measured by the Euclidean distance and the I-divergence, and proposed iterative methods called the multiplicative updates. These updates are widely used as simple and efficient computational methods for NMF because of the following three advantages. First, the updates do not contain parameters like the step size in gradient decent methods, and therefore parameter tuning is not needed. Second, nonnegativity of the matrices W^k and H^k, the solution after k iterations, is automatically satisfied if the initial matrices W^0 and H^0 are chosen to be positive. Third, implementation is easy because the update formulae are very simple.

However, the multiplicative updates of Lee and Seung have a serious drawback that their global convergence is not guaranteed theoretically. By global convergence, we mean that, for any initial solution, the sequence of solutions contains at least one convergent subsequence and the limit of any convergent subsequence is a stationary point of the corresponding optimization problem. The main difficulty in proving global convergence is that the updates, which are expressed in the form of a fraction, are not defined for all pairs of nonnegative matrices. Hence the convergence analysis of the multiplicative updates and their variants is an important research issue in NMF, and many authors have addressed this problem so far [1,10,12,18]. Finesso and Spreij [10] studied convergence properties of the multiplicative update based on the I-divergence minimization and proved, under the assumption that W^k is normalized after each update so that its Frobenius norm becomes one, that the sequences of W^k and W^kH^k always converge. However, their result does not guarantee convergence of the sequence of H^k. Lin [18] considered the case of the Euclidean distance minimization and showed that some modifications to the original multiplicative update can make it well-defined and globally convergent. However, since Lin's modified update is not multiplicative but additive in some cases, this result cannot be directly applied to the original update. Recently, Badeau *et al.* [1] studied local stability of a generalized multiplicative update, which includes the multiplicative updates of Lee and Seung as special cases, using Lyapunov's stability theory and showed that the

local optimal solution of the corresponding optimization problem is asymptotically stable if one of two matrices W^k and H^k is fixed for all k.

The objective of this paper is to show that a slight modification can guarantee global convergence of the multiplicative updates of Lee and Seung [17]. Our attention is focused on the modification proposed by Gillis and Glineur [12]. Their update, which is a modified version of the Euclidean distance-based multiplicative update of Lee and Seung [17], returns a user-specified positive constant if the original update returns a value less than the constant. Note that unlike the updates of Lin [18] and Finesso and Spreij [10], normalization procedure is not involved. Gillis and Glineur proved that their modified multiplicative update decreases the objective function monotonically and that if a sequence of solutions generated by the update has a limit point then it is necessarily a stationary point of the corresponding optimization problem [12]. However, this does not imply global convergence of the update.

In this paper, we consider not only the Euclidean distance-based multiplicative update but also the I-divergence-based one, and prove that their global convergence is guaranteed if they are modified as described by Gillis and Glineur [12]. Our proof is based on Zangwill's global convergence theorem [28, p.91] which is a fundamental result in optimization theory and has played important roles in the convergence analysis of many algorithms in machine learning [21,23,26]. We also propose two algorithms based on the modified updates. They always stop within a finite number of iterations after finding an approximate stationary point of the optimization problem.

There are many other approaches that attempt to solve NMF optimization problems. For example, some authors modified the multiplicative updates of Lee and Seung by adding a small positive constant to the denominators so that they are defined for all nonnegative matrices [3,18]. Also, some authors proposed to apply different optimization techniques to NMF optimization problems [3,7,19]. Furthermore, some authors derived a variety of multiplicative updates by considering various types of divergence between V and WH [1,6,9,27]. Although these updates are potentially superior in some cases, we will not consider them in this paper.

The rest of this paper is organized as follows. In Section 2, we introduce briefly the NMF optimization problems and the multiplicative updates of Lee and Seung. In Section 3, the modified multiplicative updates based on the idea of Gillis and Glineur are first introduced and then convergence theorems for these updates are presented. In addition, algorithms based on the modified multiplicative updates are proposed and their finite termination is proved. In Section 4, the convergence theorems in Section 3 are proved using Zangwill's global convergence theorem. Finally, in Section 5, we conclude the paper with a brief summary.

Part of this paper (Theorem 1 in Section 3) was presented in the authors' conference paper [14]. However, no rigorous proof was given there because of space limitation. In this paper, not only Theorem 1 but also some new results (Theorems 2, 3 and 4) are presented with their complete proofs.

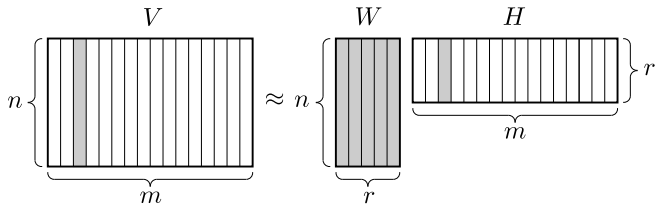

Fig. 1 Nonnegative matrix factorization. A given nonnegative matrix V is approximated by the product of two nonnegative matrices W and H. The j-th column of V is approximated by the linear combination of the columns of W where coefficients are the elements of the j-th column of H.

2 Nonnegative matrix factorization and multiplicative updates

Given a nonnegative matrix $V \in \mathbb{R}_+^{n\times m}$ where $\mathbb{R}_+$ denotes the set of nonnegative real numbers, and a positive integer $r \leq \min(n,m)$, NMF is the problem of finding two nonnegative matrices $W \in \mathbb{R}_+^{n\times r}$ and $H \in \mathbb{R}_+^{r\times m}$ such that V is approximately equal to WH (see Fig.1). Throughout this paper, we assume the following.

Assumption 1 *Each row and column of V has at least one nonzero element.*

Let us consider each column of V as a data vector. If the value of r is sufficiently small, a compact expression for the original data can be obtained through NMF because the total number of elements in the factor matrices W and H is less than that of the original matrix V. Moreover, the columns of W are regarded as a kind of basis for the space spanned by the columns of V because each data vector can be approximated by a linear combination of the columns of W (see Fig. 1).

Lee and Seung [17] employed the Euclidean distance and the I-divergence for the approximation error between V and WH, and formulated NMF as two types of optimization problems. In the former case, the problem is expressed as

$$\begin{aligned} &\text{minimize } f_{\mathrm{E}}(W,H) = \|V - WH\|^2 \\ &\text{subject to } W \geq 0,\ H \geq 0, \end{aligned} \tag{1}$$

where $\|\cdot\|$ represents the Frobenius norm, that is,

$$\|V - WH\|^2 = \sum_{i=1}^{n}\sum_{j=1}^{m}(V_{ij} - (WH)_{ij})^2,$$

and the inequality $W \geq 0$ (resp. $H \geq 0$) means that all elements of the matrix W (resp. H) are nonnegative. In the latter case, the problem is expressed as

$$\begin{aligned} &\text{minimize } f_{\mathrm{D}}(W,H) = D(V\|WH) \\ &\text{subject to } W \geq 0,\ H \geq 0, \end{aligned} \tag{2}$$

where $D(\cdot\|\cdot)$ is defined by

$$D(V\|WH) = \sum_{i=1}^{n}\sum_{j=1}^{m}\left\{V_{ij}\log\frac{V_{ij}}{(WH)_{ij}} - V_{ij} + (WH)_{ij}\right\}.$$

It is difficult in both cases to find a global optimal solution because the objective functions $f_{\mathrm{E}}(W,H)$ and $f_{\mathrm{D}}(W,H)$ are not convex. In fact, NP-hardness of NMF was proved by Vavasis [24]. Therefore, we have to take the second best way, that is, we try to find a local optimal solution instead of a global one. For this purpose, Lee and Seung [17] proposed the update rule

$$\begin{aligned} H_{aj}^{k+1} &= H_{aj}^{k} \frac{((W^k)^T V)_{aj}}{((W^k)^T W^k H^k)_{aj}}, \\ W_{ia}^{k+1} &= W_{ia}^{k} \frac{(V(H^{k+1})^T)_{ia}}{(W^k H^{k+1}(H^{k+1})^T)_{ia}}, \end{aligned} \tag{3}$$

for the optimization problem (1), and the update rule

$$\begin{aligned} H_{aj}^{k+1} &= H_{aj}^{k} \frac{\sum_{i=1}^{n} W_{ia}^{k} V_{ij}/(W^k H^k)_{ij}}{\sum_{i=1}^{n} W_{ia}^{k}}, \\ W_{ia}^{k+1} &= W_{ia}^{k} \frac{\sum_{j=1}^{m} H_{aj}^{k+1} V_{ij}/(W^k H^{k+1})_{ij}}{\sum_{j=1}^{m} H_{aj}^{k+1}}, \end{aligned} \tag{4}$$

for the optimization problem (2), where k represents the iteration count[1]. The updates like (3) and (4) are called the multiplicative updates because the new estimate is given by the product of the current estimate and some factor. An advantage of these multiplicative updates is that, unlike conventional gradient descent methods, there are no parameters to tune. Another advantage is that positiveness of W^k and H^k is guaranteed for all k under Assumption 1 if the initial matrices W^0 and H^0 are chosen to be positive [19]. For these reasons, the multiplicative updates (3) and (4) are widely used as simple and effective methods for finding local optimal solutions of (1) and (2).

3 Modified multiplicative updates and their global convergence

The most serious drawback of the multiplicative update rules described by (3) and (4) is that the right-hand sides are not defined for all nonnegative matrices W^k and H^k (or H^{k+1}). For example, in the case of Euclidean distance, we cannot obtain H^{k+1} by the update rule (3) when $H^k = 0$, because the denominator of the first equation vanishes.

As mentioned in Section 2, W^k and H^k are positive for all k if the initial matrices W^0 and H^0 are chosen to be positive. Hence the updates can be performed infinitely many times. However, even though the sequence $\{(W^k,H^k)\}_{k=0}^{\infty}$ converges, it is not guaranteed that both $\lim_{k\to\infty} W^k$ and $\lim_{k\to\infty} H^k$ are positive. This means that the update rules may not be defined at $\lim_{k\to\infty}(W^k,H^k)$, which makes it difficult to prove their global convergence using known results such as Zangwill's global convergence theorem [28, p.91].

In this section, we introduce slightly modified versions of the update rules (3) and (4) which are based on the idea of Gillis and Glineur [12], and present convergence

[1] Although it is not explicitly written in their original paper [17] which of H^k and H^{k+1} is used for the computation of W^{k+1}, we consider the latter case throughout this paper as in [18].

theorems. We also propose two algorithms based on the modified updates and prove their finite termination.

3.1 Euclidean distance

In order to prevent elements of matrices W^k and H^k from vanishing, Gillis and Glineur [12] have proposed to modify the update rule (3) as

$$\begin{aligned} H_{aj}^{k+1} &= \max\left(H_{aj}^k \frac{((W^k)^T V)_{aj}}{((W^k)^T W^k H^k)_{aj}}, \varepsilon\right), \\ W_{ia}^{k+1} &= \max\left(W_{ia}^k \frac{(V(H^{k+1})^T)_{ia}}{(W^k H^{k+1}(H^{k+1})^T)_{ia}}, \varepsilon\right), \end{aligned} \tag{5}$$

where ε is any positive constant specified by the user. Each update in (5) returns the positive constant ε if the corresponding original update in (3) returns a value less than ε. This is the only difference between these two update rules. With the modification of the update rule from (3) to (5), we have to modify also the optimization problem (1) as follows:

$$\begin{aligned} &\text{minimize } f_{\mathrm{E}}(W,H) = \|V - WH\|^2 \\ &\text{subject to } W_{ia} \geq \varepsilon,\ H_{aj} \geq \varepsilon, \quad \forall i,a,j\,. \end{aligned} \tag{6}$$

The feasible region of this optimization problem is denoted by X, that is,

$$X = \{(W,H)\,|\,W_{ia} \geq \varepsilon,\ H_{aj} \geq \varepsilon,\ \forall i,a,j\}\,.$$

Karush-Kuhn-Tucker (KKT) conditions [4] for the problem (6) are expressed as follows:[2]

$$W_{ia} \geq \varepsilon, \quad \forall i,a\,, \tag{7}$$

$$H_{aj} \geq \varepsilon, \quad \forall a,j\,, \tag{8}$$

$$(\nabla_W f_{\mathrm{E}}(W,H))_{ia} \geq 0, \quad \forall i,a\,, \tag{9}$$

$$(\nabla_H f_{\mathrm{E}}(W,H))_{aj} \geq 0, \quad \forall a,j\,, \tag{10}$$

$$(\nabla_W f_{\mathrm{E}}(W,H))_{ia}(\varepsilon - W_{ia}) = 0, \quad \forall i,a\,, \tag{11}$$

$$(\nabla_H f_{\mathrm{E}}(W,H))_{aj}(\varepsilon - H_{aj}) = 0, \quad \forall a,j\,, \tag{12}$$

where

$$\begin{aligned} \nabla_W f_{\mathrm{E}}(W,H) &= 2(WH - V)H^T\,, \\ \nabla_H f_{\mathrm{E}}(W,H) &= 2W^T(WH - V)\,. \end{aligned}$$

Therefore, a necessary condition for a point (W,H) to be a local optimal solution of (6) is that the conditions (7)–(12) are satisfied. Hereafter, we call a point (W,H) a stationary point of (6) if it satisfies (7)–(12), and denote the set of all stationary points of (6) by S_{E}.

The global convergence property of the modified update rule (5) is stated as follows.

[2] The conditions (7)–(12) are derived by eliminating Lagrange multipliers in the original KKT conditions.

Theorem 1 *Let $\{(W^k,H^k)\}_{k=0}^{\infty}$ be any sequence generated by the modified update rule (5) with the initial point $(W^0,H^0)\in X$. Then $(W^k,H^k)\in X$ holds for all positive integers k. Moreover, the sequence has at least one convergent subsequence and the limit of any convergent subsequence is a stationary point of the optimization problem (6).*

Proof of Theorem 1 will be given in the next section.

By making use of Theorem 1, we can immediately construct an algorithm that terminates within a finite number of iterations. To do so, we relax the conditions (9)–(12) as

$$(\nabla_W f_{\mathrm{E}}(W,H))_{ia} \geq -\delta_1, \quad \forall i,a, \tag{13}$$

$$(\nabla_H f_{\mathrm{E}}(W,H))_{aj} \geq -\delta_1, \quad \forall j,a, \tag{14}$$

$$W_{ia}-\varepsilon \leq \delta_2 \text{ if } (\nabla_W f_{\mathrm{E}}(W,H))_{ia} > \delta_1, \quad \forall i,a, \tag{15}$$

$$H_{aj}-\varepsilon \leq \delta_2 \text{ if } (\nabla_H f_{\mathrm{E}}(W,H))_{aj} > \delta_1, \quad \forall a,j, \tag{16}$$

where δ_1 and δ_2 are any positive constants specified by the user, and employ these relaxed conditions as a stopping criterion. Let $\bar{S}_{\mathrm{E}}$ be the set of all $(W,H)\in X$ satisfying (13)–(16). Then the proposed algorithm is described as follows:

Algorithm 1

Input: $V\in\mathbb{R}_+^{n\times m}$, $r\in\mathbb{N}$, $\varepsilon>0$, $\delta_1>0$, $\delta_2>0$
Step 1: Choose $(W^0,H^0)\in X$ and set $k=0$.
Step 2: Find (W^{k+1},H^{k+1}) by the update rule (5).
Step 3: If $(W^{k+1},H^{k+1})\in\bar{S}_{\mathrm{E}}$ then return W^{k+1} and H^{k+1}. Otherwise add 1 to k and go to Step 2.

Theorem 2 *For any positive constants ε, δ_1 and δ_2, Algorithm 1 stops within a finite number of iterations.*

Proof Let $\{(W^{k_l},H^{k_l})\}_{l=1}^{\infty}$ be any convergent subsequence of the sequence $\{(W^k,H^k)\}_{k=0}^{\infty}$ generated by the modified update rule (5), and $(\bar{W},\bar{H})\in X$ be the limit of the subsequence. Then, by Theorem 1, $(\bar{W},\bar{H})$ satisfies

$$\begin{aligned}
(\nabla_W f_{\mathrm{E}}(\bar{W},\bar{H}))_{ia} &\geq 0, \quad \forall i,a,\\
(\nabla_H f_{\mathrm{E}}(\bar{W},\bar{H}))_{aj} &\geq 0, \quad \forall a,j,\\
(\nabla_W f_{\mathrm{E}}(\bar{W},\bar{H}))_{ia}(\varepsilon-\bar{W}_{ia}) &= 0, \quad \forall i,a,\\
(\nabla_H f_{\mathrm{E}}(\bar{W},\bar{H}))_{aj}(\varepsilon-\bar{H}_{aj}) &= 0, \quad \forall a,j.
\end{aligned}$$

Recall that $\nabla_W f_{\mathrm{E}}(W,H)$ and $\nabla_H f_{\mathrm{E}}(W,H)$ are continuous for all $(W,H)\in X$. For all (i,a) such that $(\nabla_W f_{\mathrm{E}}(\bar{W},\bar{H}))_{ia}=0$, there exists a positive integer L_{ia}^1 such that

$$|(\nabla_W f_{\mathrm{E}}(W^{k_l},H^{k_l}))_{ia}| \leq \delta_1, \quad \forall l \geq L_{ia}^1.$$

For all (i,a) such that $(\nabla_W f_{\mathrm{E}}(\bar{W},\bar{H}))_{ia}>0$, there exists a positive integer L_{ia}^1 such that

$$(\nabla_W f_{\mathrm{E}}(W^{k_l},H^{k_l}))_{ia} \geq -\delta_1 \text{ and } W_{ia}^{k_l}-\varepsilon \leq \delta_2, \quad \forall l \geq L_{ia}^1,$$

because $\bar{W}_{ia} = \varepsilon$ holds. For all (a, j) such that $(\nabla_H f_{\mathrm{E}}(\bar{W}, \bar{H}))_{aj} = 0$, there exists a positive integer L^2_{aj} such that

$$|(\nabla_W f_{\mathrm{E}}(W^{k_l}, H^{k_l}))_{ia}| \leq \delta_1, \quad \forall l \geq L^2_{ia}.$$

For all (a, j) such that $(\nabla_H f_{\mathrm{E}}(\bar{W}, \bar{H}))_{aj} > 0$, there exists a positive integer L^2_{aj} such that

$$(\nabla_H f_{\mathrm{E}}(W^{k_l}, H^{k_l}))_{aj} \geq -\delta_1 \text{ and } H^{k_l}_{aj} - \varepsilon \leq \delta_2, \quad \forall l \geq L^2_{aj},$$

because $\bar{H}_{aj} = \varepsilon$ holds. From these considerations, we immediately see that there exists a positive integer L such that the stopping criterion of Algorithm 1 is satisfied for all (W^{k_l}, H^{k_l}) with $l \geq L$. This means that Algorithm 1 always stops within a finite number of iteration. □

3.2 I-divergence

As in the case of Euclidean distance, we modify the update rule (4) as

$$\begin{aligned} H^{k+1}_{aj} &= \max\left(H^k_{aj} \frac{\sum_{i=1}^n W^k_{ia} V^k_{ij} / (W^k H^k)_{ij}}{\sum_{\mu=1}^n W^k_{\mu a}}, \varepsilon \right), \\ W^{k+1}_{ia} &= \max\left(W^k_{ia} \frac{\sum_{j=1}^m H^{k+1}_{aj} V_{ij} / (W^k H^{k+1})_{ij}}{\sum_{\nu=1}^m H^k_{a\nu}}, \varepsilon \right), \end{aligned} \tag{17}$$

where ε is any positive constant specified by the user. The modified update rule corresponds to modifying the optimization problem (2) as follows:

$$\begin{aligned} &\text{minimize } f_{\mathrm{D}}(W, H) = D(V \,\|\, WH) \\ &\text{subject to } W_{ia} \geq \varepsilon,\ H_{aj} \geq \varepsilon, \quad \forall i, a, j. \end{aligned} \tag{18}$$

The feasible region of this optimization problem is X as in the case of (6). KKT conditions for the problem (18) are expressed as follows:[3]

$$W_{ia} \geq \varepsilon, \quad \forall i, a, \tag{19}$$
$$H_{aj} \geq \varepsilon, \quad \forall a, j, \tag{20}$$
$$(\nabla_W f_{\mathrm{D}}(W, H))_{ia} \geq 0, \quad \forall i, a, \tag{21}$$
$$(\nabla_H f_{\mathrm{D}}(W, H))_{aj} \geq 0, \quad \forall a, j, \tag{22}$$
$$(\nabla_W f_{\mathrm{D}}(W, H))_{ia} (\varepsilon - W_{ia}) = 0, \quad \forall i, a, \tag{23}$$
$$(\nabla_H f_{\mathrm{D}}(W, H))_{aj} (\varepsilon - H_{aj}) = 0, \quad \forall a, j, \tag{24}$$

where

$$(\nabla_W f_{\mathrm{D}}(W, H))_{ia} = \sum_{j=1}^m \left\{ H_{aj} - \frac{V_{ij} H_{aj}}{(WH)_{ij}} \right\},$$

[3] The conditions (19)–(24) are derived by eliminating Lagrange multipliers in the original KKT conditions.

$$(\nabla_H f_{\mathrm{D}}(W,H))_{aj} = \sum_{i=1}^{n} \left\{ W_{ia} - \frac{V_{ij} W_{ia}}{(WH)_{ij}} \right\}.$$

Therefore, a necessary condition for a point (W,H) to be a local optimal solution of (18) is that the conditions (19)–(24) are satisfied. Hereafter, we call a point (W,H) a stationary point of (18) if it satisfies (19)–(24), and denote the set of all stationary points of (18) by S_{D}.

The global convergence property of the modified update rule (17) is stated as follows.

Theorem 3 *Let $\{(W^k,H^k)\}_{k=0}^{\infty}$ be any sequence generated by the modified update rule (17) with the initial point $(W^0,H^0) \in X$. Then $(W^k,H^k) \in X$ holds for all positive integers k. Moreover, the sequence has at least one convergent subsequence and the limit of any convergent subsequence is a stationary point of the optimization problem (18).*

Proof of Theorem 3 will be given in the next section.

By making use of Theorem 3, we can easily construct an algorithm that terminates within a finite number of iterations. To do so, we relax the conditions (21)–(24) as

$$(\nabla_W f_{\mathrm{D}}(W,H))_{ia} \geq -\delta_1, \quad \forall i,a, \tag{25}$$

$$(\nabla_H f_{\mathrm{D}}(W,H))_{aj} \geq -\delta_1, \quad \forall a,j, \tag{26}$$

$$W_{ia} - \varepsilon \leq \delta_2 \text{ if } (\nabla_W f_{\mathrm{D}}(W,H))_{ia} > \delta_1, \quad \forall i,a, \tag{27}$$

$$H_{aj} - \varepsilon \leq \delta_2 \text{ if } (\nabla_H f_{\mathrm{D}}(W,H))_{aj} > \delta_1, \quad \forall a,j, \tag{28}$$

where δ_1 and δ_2 are any positive constants specified by the user, and employ these relaxed conditions as a stopping criterion. Let $\bar{S}_{\mathrm{D}}$ be the set of all $(W,H) \in X$ satisfying (25)–(28). Then the proposed algorithm is described as follows:

Algorithm 2

Input: $V \in \mathbb{R}_+^{n\times m}$, $r \in \mathbb{N}$, $\varepsilon > 0$, $\delta_1 > 0$, $\delta_2 > 0$
Step 1: Choose $(W^0,H^0) \in X$ and set $k = 0$.
Step 2: Find (W^{k+1},H^{k+1}) by the update rule (17).
Step 3: If $(W^{k+1},H^{k+1}) \in \bar{S}_{\mathrm{D}}$ then return W^{k+1} and H^{k+1}. Otherwise add 1 to k and go to Step 2.

Theorem 4 *For any positive constants ε, δ_1 and δ_2, Algorithm 2 stops within a finite number of iterations.*

We omit the proof of Theorem 4 because it is almost same as the proof of Theorem 2.

3.3 Related works

The modified update rule (5) was first proposed by Gillis and Glineur [12], as stated above. They proved not only that $f_{\mathrm{E}}(W^k,H^k)$ is nonincreasing under (5) but also that

if a sequence of solutions generated by (5) has a limit point then it is necessarily a stationary point of the optimization problem (6), but these facts are not sufficient to prove global convergence of (5). As a matter of fact, we cannot rule out, for example, the existence of a sequence $\{(W^k,H^k)\}_{k=0}^{\infty}$ such that $f_{\mathrm{E}}(W^k,H^k)$ takes the same value for all k and the sequence visits a finite number of distinct points periodically. However, on the other hand, in another paper [13], they showed through numerical experiments that (5) works better than the original update rule (3) in some cases. This indicates that (5) is important not only from a theoretical point of view but also in practice.

Lin [18] proposed a modified version of (3) and proved that any sequence $\{(W^k,H^k)\}_{k=0}^{\infty}$ generated by the modified rule has at least one convergent subsequence and their limits are stationary points of the optimization problem (1). However, Lin's update rule considerably differs from the original one because of many extra operations. In particular, in the case where $(\nabla_H f_{\mathrm{E}}(W^k,H^k))_{aj}$ is negative and H_{aj}^k is less than a user-specified small positive constant, Lin's update rule is not multiplicative but additive. Also, the matrix W^k must be normalized after each update in order to guarantee that the sequence $\{(W^k,H^k)\}_{k=0}^{\infty}$ is in a bounded set. In contrast, the normalization is not required in the modified update rule (5). Nevertheless, the boundedness of the sequence $\{(W^k,H^k)\}_{k=0}^{\infty}$ generated by (5) is guaranteed as shown in the next section.

Finesso and Spreij [10] studied the convergence properties of the multiplicative update (4) by interpreting it as an alternating minimization procedure [8]. Under the assumption that the matrix W^k is normalized after each update, they proved that any sequence $\{(W^k,H^k)\}_{k=0}^{\infty}$ generated by (4) satisfies the following properties: 1) W^k converges to a nonnegative matrix. 2) For each triple (i,a,j), $W_{ia}^k H_{aj}^k$ converges to a nonnegative number. 3) For each pair (a,j), H_{aj}^k converges to a nonnegative number if $\lim_{k\to\infty}\sum_{i=1}^n W_{ia}^k > 0$ [10, Theorem 6.1]. However, they said nothing about the convergence of H_{aj}^k for the case where $\lim_{k\to\infty}\sum_{i=1}^n W_{ia}^k = 0$.

Badeau *et al.* [1] studied the local stability of a generalized multiplicative update, which includes (3) and (4) as special cases, using Lyapunov's stability theory and showed that the local optimal solution of the corresponding optimization problem is asymptotically stable if one of two factor matrices W and H is fixed to a nonnegative constant matrix.

4 Proofs of Theorems 1 and 3

We will prove Theorems 1 and 3 in this section. The first parts of these theorems apparently follow from the update rules (5) and (17). In order to prove the second parts, we make use of Zangwill's global convergence theorem [28, p.91], which is a fundamental result in optimization theory. Let A be a point-to-point mapping[4] from X into itself and S be a subset of X. Then Zangwill's global convergence theorem claims the following: if the mapping A satisfies the following three conditions then, for any initial point $(W^0,H^0)\in X$, the sequence $\{(W^k,H^k)\}_{k=0}^{\infty}$ generated by A contains at least

[4] Although A is assumed to be a point-to-set mapping in the original version of Zangwill's global convergence theorem, we consider in this paper its special case where A is a point-to-point mapping.

one convergent subsequence and the limit of any convergent subsequence belongs to S.

1. All points in the sequence $\{(W^k,H^k)\}_{k=0}^{\infty}$ belong to a compact set in X.
2. There is a function $z : X \to \mathbb{R}$ satisfying the following two conditions.
   (a) If $(W,H) \notin S$ then $z(A(W,H)) < z(W,H)$.
   (b) If $(W,H) \in S$ then $z(A(W,H)) \leq z(W,H)$.
3. The mapping A is continuous in $X \setminus S$.

In the following, we will first prove Theorem 1 by showing that these conditions are satisfied when the mapping A is defined by (5) and S is set to S_{E}. We will next prove Theorem 2 by showing that these conditions are satisfied when the mapping A is defined by (17) and S is set to S_{D}.

4.1 Proof of Theorem 1

Let us rewrite (5) as

$$H^{k+1} = A_1(W^k,H^k),$$
$$W^{k+1} = A_2(W^k,H^{k+1}),$$

or, more simply,

$$(W^{k+1},H^{k+1}) = A(W^k,H^k),$$

where the mapping A is defined by

$$A(W,H) = (A_2(W,A_1(W,H)),A_1(W,H)).$$

Let us also set $S = S_{\mathrm{E}}$. Since the mapping A is continuous in X, the third condition of Zangwill's global convergence theorem is satisfied. We will thus show in the following that A also satisfies the remaining two conditions.

The following lemma guarantees that the first condition is satisfied.

Lemma 1 *For any initial point $(W^0,H^0) \in X$, the sequence $\{(W^k,H^k)\}_{k=0}^{\infty}$ generated by the mapping A belongs to a compact set in X.*

Proof Let (W,H) be any point in X. Then we have

$$\begin{aligned}
H_{aj}\frac{(W^TV)_{aj}}{(W^TWH)_{aj}} &= H_{aj}\frac{\sum_{i=1}^n W_{ia}V_{ij}}{\sum_{l=1}^r (W^TW)_{al}H_{lj}} \\
&= H_{aj}\frac{\sum_{i=1}^n W_{ia}V_{ij}}{\sum_{l=1}^r (\sum_{i=1}^n W_{ia}W_{il})H_{lj}} \\
&= H_{aj}\frac{\sum_{i=1}^n W_{ia}V_{ij}}{\sum_{i=1}^n W_{ia}^2H_{aj} + \sum_{l=1,l\neq a}^r (\sum_{i=1}^n W_{ia}W_{il})H_{lj}} \\
&= \frac{\sum_{i=1}^n W_{ia}V_{ij}}{\sum_{i=1}^n W_{ia}^2 + \sum_{l=1,l\neq a}^r (\sum_{i=1}^n W_{ia}W_{il})(H_{lj}/H_{aj})} \\
&< \frac{\sum_{i=1}^n W_{ia}V_{ij}}{\sum_{i=1}^n W_{ia}^2}
\end{aligned}$$

$$= \frac{\sum_{i=1}^{n} \left(W_{ia} / \sqrt{\sum_{\mu=1}^{n} W_{\mu a}^2} \right) V_{ij}}{\sqrt{\sum_{i=1}^{n} W_{ia}^2}}$$
$$\leq \frac{\sqrt{\sum_{i=1}^{n} V_{ij}^2}}{\varepsilon \sqrt{n}},$$

from which the inequality

$$(A_1(W,H))_{aj} \leq \max\left(\frac{\sqrt{\sum_{i=1}^{n} V_{ij}^2}}{\varepsilon \sqrt{n}}, \varepsilon \right)$$

holds for any pair (a, j). Note that the right-hand side is a constant which depends on neither W nor H. Similarly, we have

$$W_{ia} \frac{(VH^T)_{ia}}{(WHH^T)_{ia}} < \frac{\sqrt{\sum_{j=1}^{m} V_{ij}^2}}{\varepsilon \sqrt{m}},$$

from which the inequality

$$(A_2(W,H))_{ia} \leq \max\left(\frac{\sqrt{\sum_{j=1}^{m} V_{ij}^2}}{\varepsilon \sqrt{m}}, \varepsilon \right)$$

holds for any pair (i, a). Note that the right-hand side is a constant which depends on neither W nor H. Hence $A(W,H)$ belongs to a compact set in X. This means that, for any initial point $(W^0, H^0) \in X$, the sequence $\{(W^k, H^k)\}_{k=0}^{\infty}$ generated by the mapping A belongs to a compact set in X. □

The last step is to prove that the second condition of Zangwill's global convergence theorem is also satisfied. To do this, we first need to introduce two auxiliary functions for f_{E}. Let $(\hat{W}, \hat{H})$ be any point in X. Let the function $g_{\mathrm{E}}^{\hat{W}} : [\varepsilon, \infty)^{r \times m} \times [\varepsilon, \infty)^{r \times m} \to \mathbb{R}$ be defined by

$$g_{\mathrm{E}}^{\hat{W}}(H, H') = f_{\mathrm{E}}(\hat{W}, H') + \sum_{a=1}^{r} \sum_{j=1}^{m} g_{\mathrm{E}aj}^{\hat{W}}(H_{aj}, H'),$$

where the function $g_{\mathrm{E}aj}^{\hat{W}} : [\varepsilon, \infty) \times [\varepsilon, \infty)^{r \times m} \to \mathbb{R}$ is defined by

$$g_{\mathrm{E}aj}^{\hat{W}}(H_{aj}, H') = 2(\hat{W}^T(\hat{W}H' - V))_{aj}(H_{aj} - H'_{aj}) + \frac{(\hat{W}^T\hat{W}H')_{aj}}{H'_{aj}}(H_{aj} - H'_{aj})^2. \quad (29)$$

Similarly, let the function $h_{\mathrm{E}}^{\hat{H}} : [\varepsilon, \infty)^{n \times r} \times [\varepsilon, \infty)^{n \times r} \to \mathbb{R}$ be defined by

$$h_{\mathrm{E}}^{\hat{H}}(W, W') = f_{\mathrm{E}}(W, \hat{H}) + \sum_{i=1}^{n} \sum_{a=1}^{r} h_{\mathrm{E}ia}^{\hat{H}}(W_{ia}, W'), \quad (30)$$

where the function $h^{\hat{H}}_{\mathrm{E}ia} : [\varepsilon,\infty) \times [\varepsilon,\infty)^{n\times r} \to \mathbb{R}$ is defined by

$$h^{\hat{H}}_{\mathrm{E}ia}(W_{ia},W') = 2((W'\hat{H}-V)\hat{H}^T)_{ia}(W_{ia}-W'_{ia}) + \frac{(W'\hat{H}\hat{H}^T)_{ia}}{W'_{ia}}(W_{ia}-W'_{ia})^2 .$$

The functions $g^{\hat{W}}_{\mathrm{E}}$ and $h^{\hat{H}}_{\mathrm{E}}$ are essentially the same as the auxiliary functions considered by Lee and Seung [17], though mathematical expressions are slightly different. However, note that the domains of $g^{\hat{W}}_{\mathrm{E}}$ and $h^{\hat{H}}_{\mathrm{E}}$ are restricted to $[\varepsilon,\infty)^{r\times m} \times [\varepsilon,\infty)^{r\times m}$ and $[\varepsilon,\infty)^{n\times r} \times [\varepsilon,\infty)^{n\times r}$, respectively, in the present paper. This is an important difference between our functions and theirs.

In the following, we give five lemmas which are needed to prove that the second condition of Zangwill's global convergence theorem is satisfied. Although some of them can be immediately obtained from some of the results given by Lee and Seung [17], we will provide proofs for all lemmas in order to make this paper self-contained.

Lemma 2 *For any $\hat{W} \in [\varepsilon,\infty)^{n\times r}$, the function $g^{\hat{W}}_{\mathrm{E}}$ satisfies the following two conditions:*

$$g^{\hat{W}}_{\mathrm{E}}(H,H) = f_{\mathrm{E}}(\hat{W},H), \quad \forall H \in [\varepsilon,\infty)^{r\times m}, \tag{31}$$

$$g^{\hat{W}}_{\mathrm{E}}(H,H') \geq f_{\mathrm{E}}(\hat{W},H), \quad \forall H,H' \in [\varepsilon,\infty)^{r\times m}. \tag{32}$$

Also, for any $\hat{H} \in [\varepsilon,\infty)^{r\times m}$, the function $h^{\hat{H}}_{\mathrm{E}}$ satisfies the following two conditions:

$$h^{\hat{H}}_{\mathrm{E}}(W,W) = f_{\mathrm{E}}(W,\hat{H}), \quad \forall W \in [\varepsilon,\infty)^{n\times r},$$

$$h^{\hat{H}}_{\mathrm{E}}(W,W') \geq f_{\mathrm{E}}(W,\hat{H}), \quad \forall W,W' \in [\varepsilon,\infty)^{n\times r}.$$

Proof We prove only the first part because the second one can be proved in the same way. Since $g^{\hat{W}}_{\mathrm{E}aj}(H_{aj},H) = 0$ holds for all $H \in [\varepsilon,\infty)^{r\times m}$ and indices a and j, the first condition (31) is satisfied. To see that the second condition (32) is also satisfied, we first rewrite $f_{\mathrm{E}}(\hat{W},H)$ using the Taylor series expansion as

$$f_{\mathrm{E}}(\hat{W},H) = f_{\mathrm{E}}(\hat{W},H') + \sum_{a=1}^{r}\sum_{j=1}^{m} 2(\hat{W}^T(\hat{W}H'-V))_{aj}(H_{aj}-H'_{aj}) + \sum_{a=1}^{r}\sum_{b=1}^{r}\sum_{j=1}^{m} (\hat{W}^T\hat{W})_{ab}(H_{aj}-H'_{aj})(H_{bj}-H'_{bj}).$$

Then we have

$$g^{\hat{W}}_{\mathrm{E}}(H,H') - f_{\mathrm{E}}(\hat{W},H) = \sum_{j=1}^{m}\sum_{a=1}^{r}\sum_{b=1}^{r} M^{(j)}_{ab}\left(\frac{H_{aj}-H'_{aj}}{H'_{aj}}\right)\left(\frac{H_{bj}-H'_{bj}}{H'_{bj}}\right), \tag{33}$$

where

$$M^{(j)}_{ab} = \delta_{ab}(\hat{W}^T\hat{W}H')_{aj}H'_{bj} - (\hat{W}^T\hat{W})_{ab}H'_{aj}H'_{bj}, \tag{34}$$

and δ_{ab} represents the Kronecker's delta. We next show that the matrices $M^{(j)} = [M^{(j)}_{ab}]$ $(j = 1,2,\ldots,m)$ are positive semi-definite for all $\hat{W} \in [\varepsilon,\infty)^{n\times r}$ and $H' \in$

$[\varepsilon,\infty)^{r\times m}$. If this is true, the right-hand side of (33) is nonnegative for all $H,H' \in [\varepsilon,\infty)^{r\times m}$. Since the right-hand side of (34) can be rewritten as

$$\begin{aligned} M^{(j)}_{ab} &= \delta_{ab}\sum_{l=1}^{r}(\hat{W}^T\hat{W})_{al}H'_{lj}H'_{bj} - (\hat{W}^T\hat{W})_{ab}H'_{aj}H'_{bj} \\ &= \begin{cases} \sum_{l=1,l\neq a}^{r}(\hat{W}^T\hat{W})_{al}H'_{lj}H'_{aj}, & \text{if } a=b \\ -(\hat{W}^T\hat{W})_{ab}H'_{aj}H'_{bj}, & \text{if } a\neq b, \end{cases} \end{aligned}$$

the matrix $M^{(j)}$ satisfies

$$M^{(j)}_{aa} = \sum_{l=1,l\neq a}^{r}|M^{(j)}_{al}|, \quad a=1,2,\ldots,r,$$

which means that $M^{(j)}$ is real, symmetric and diagonally dominant with positive diagonal elements. Therefore, $M^{(j)}$ $(j=1,2,\ldots,m)$ are positive semi-definite. □

Lemma 3 *Let $(\hat{W},\hat{H})$ be any point in X. Then $g_{\mathrm{E}}^{\hat{W}}(H,\hat{H})$, which is considered as a function of H, is strictly convex in $[\varepsilon,\infty)^{r\times m}$. Also, $h_{\mathrm{E}}^{\hat{H}}(W,\hat{W})$, which is considered as a function of W, is strictly convex in $[\varepsilon,\infty)^{n\times r}$.*

Proof The second-order partial derivatives of $g_{\mathrm{E}}^{\hat{W}}(H,\hat{H})$ are given by

$$\frac{\partial^2 g_{\mathrm{E}}^{\hat{W}}(H,\hat{H})}{\partial H_{aj}\partial H_{a'j'}} = \begin{cases} \frac{(\hat{W}^T\hat{W}\hat{H})_{aj}}{\hat{H}_{aj}}, & \text{if } (a,j)=(a',j') \\ 0, & \text{otherwise}, \end{cases}$$

where $(a,j),(a',j') \in \{1,2,\ldots,r\}\times\{1,2,\ldots,m\}$. Since $(\hat{W}^T\hat{W}\hat{H})_{aj}/\hat{H}_{aj}$ is a positive constant, $g_{\mathrm{E}}^{\hat{W}}(H,\hat{H})$ is strictly convex in $[\varepsilon,\infty)^{r\times m}$. The second part can be proved in the same way. □

Lemma 4 *Let $(\hat{W},\hat{H})$ be any point in X. The optimization problem*

$$\begin{aligned} &\textit{minimize } \ g_{\mathrm{E}}^{\hat{W}}(H,\hat{H}) \\ &\textit{subject to } H_{aj}\geq\varepsilon, \quad \forall a,j \end{aligned} \tag{35}$$

has a unique optimal solution which is given by $A_1(\hat{W},\hat{H})$. Also, the optimization problem

$$\begin{aligned} &\textit{minimize } \ h_{\mathrm{E}}^{\hat{H}}(W,\hat{W}) \\ &\textit{subject to } W_{ia}\geq\varepsilon, \quad \forall i,a \end{aligned} \tag{36}$$

has a unique optimal solution which is given by $A_2(\hat{W},\hat{H})$.

Proof It suffices for us to show that for any pair (a,j), the optimization problem

$$\begin{aligned} &\text{minimize } \ g_{\mathrm{E}aj}^{\hat{W}}(H_{aj},\hat{H}) \\ &\text{subject to } H_{aj}\geq\varepsilon \end{aligned} \tag{37}$$

has a unique optimal solution which is given by $(A_1(\hat{W},\hat{H}))_{aj}$ and that for any pair (i,a), the optimization problem

$$\begin{array}{l}\text{minimize } h^{\hat{H}}_{\mathrm{E}ia}(W_{ia},\hat{W})\\ \text{subject to } W_{ia}\geq\varepsilon\end{array} \tag{38}$$

has a unique optimal solution which is given by $(A_2(\hat{W},\hat{H}))_{aj}$. In the following, we consider only the first part because the second part can be proved similarly. Since $g^{\hat{W}}_{\mathrm{E}aj}(H_{aj},\hat{H})$ is strictly convex in $[\varepsilon,\infty)$, the equation $\mathrm{d}g^{\hat{W}}_{\mathrm{E}aj}(H_{aj},\hat{H})/\mathrm{d}H_{aj}=0$ has at most one solution in $[\varepsilon,\infty)$. By solving this equation, we have

$$H_{aj}=\hat{H}_{aj}\frac{(\hat{W}^TV)_{aj}}{(\hat{W}^T\hat{W}\hat{H})_{aj}}\,.$$

Let the right-side hand be denoted by H^*_{aj} which is a nonnegative number. If $H^*_{aj}\geq\varepsilon$ then H^*_{aj} is apparently the optimal solution of (37). If $H^*_{aj}<\varepsilon$ then ε is the optimal solution of (37) because $g^{\hat{W}}_{\mathrm{E}aj}(H_{aj},\hat{H})$ is strictly monotone increasing in $[\varepsilon,\infty)$. Therefore the optimal solution of (37) is identical with $(A_1(\hat{W},\hat{H}))_{aj}$. □

Lemma 5 *The inequality $f_{\mathrm{E}}(A(\hat{W},\hat{H}))\leq f_{\mathrm{E}}(\hat{W},\hat{H})$ holds for all $(\hat{W},\hat{H})\in X$.*

Proof By Lemmas 2 and 4, we have

$$f_{\mathrm{E}}(\hat{W},A_1(\hat{W},\hat{H}))\leq g^{\hat{W}}_{\mathrm{E}}(A_1(\hat{W},\hat{H}),\hat{H})\leq g^{\hat{W}}_{\mathrm{E}}(\hat{H},\hat{H})=f_{\mathrm{E}}(\hat{W},\hat{H}),\ \ \forall(\hat{W},\hat{H})\in X$$

and

$$f_{\mathrm{E}}(A_2(\hat{W},\hat{H}),\hat{H})\leq h^{\hat{H}}_{\mathrm{E}}(A_2(\hat{W},\hat{H}),\hat{W})\leq h^{\hat{H}}_{\mathrm{E}}(\hat{W},\hat{W})=f_{\mathrm{E}}(\hat{W},\hat{H}),\ \ \forall(\hat{W},\hat{H})\in X\,.$$

From these two inequalities, we have

$$f_{\mathrm{E}}(A(\hat{W},\hat{H}))=f_{\mathrm{E}}(A_2(\hat{W},A_1(\hat{W},\hat{H})),A_1(\hat{W},\hat{H}))\leq f_{\mathrm{E}}(\hat{W},A_1(\hat{W},\hat{H}))\leq f_{\mathrm{E}}(\hat{W},\hat{H}) \tag{39}$$

which completes the proof. □

Lemma 6 *$(\hat{W},\hat{H})\in S_{\mathrm{E}}$ if and only if $\hat{H}$ and $\hat{W}$ are the optimal solutions of (35) and (36), respectively.*

Proof It suffices for us to show that $(\hat{W},\hat{H})\in S_{\mathrm{E}}$ if and only if $\hat{H}_{aj}$ is the optimal solution of (37) for any pair (a,j) and $\hat{W}_{ia}$ is the optimal solution of (38) for any pair (i,a). By the definition (29) of $g^{\hat{W}}_{\mathrm{E}aj}(H_{aj},\hat{H})$, we have

$$\left.\frac{\partial g^{\hat{W}}_{\mathrm{E}aj}(H_{aj},\hat{H})}{\partial H_{aj}}\right|_{H_{aj}=\hat{H}_{aj}}=(\nabla_H f_{\mathrm{E}}(\hat{W},\hat{H}))_{aj}\,.$$

Since $g^{\hat{W}}_{\mathrm{E}aj}(H_{aj},\hat{H})$ is strictly convex in $[\varepsilon,\infty)$, the necessary and sufficient condition for $\hat{H}_{aj}$ to be the optimal solution of (37) is given by

$$(\nabla_H f_{\mathrm{E}}(\hat{W},\hat{H}))_{aj}\begin{cases}=0,\ \text{if}\ \hat{H}_{aj}>\varepsilon\\ \geq 0,\ \text{if}\ \hat{H}_{aj}=\varepsilon\,,\end{cases}$$

which is equivalent to the set of conditions (8) and (10). By the definition (30) of $h^{\hat{H}}_{\mathrm{E}ia}(W_{ia},\hat{W})$, we have

$$\left.\frac{\partial h^{\hat{H}}_{\mathrm{E}ia}(W_{ia},\hat{W})}{\partial W_{ia}}\right|_{W_{ia}=\hat{W}_{ia}} = (\nabla_W f_{\mathrm{E}}(\hat{W},\hat{H}))_{ia}, \quad \forall i,a\,.$$

Hence the necessary and sufficient condition for $\hat{W}$ to be the optimal solution of (36) is given by

$$(\nabla_W f_{\mathrm{E}}(\hat{W},\hat{H}))_{ia}\begin{cases} = 0, \text{ if } \hat{W}_{ia} > \varepsilon \\ \geq 0, \text{ if } \hat{W}_{ia} = \varepsilon \end{cases} \quad \forall i,a\,,$$

which is equivalent to the set of conditions (7) and (9). □

From Lemmas 4–6, we derive the following lemma which claims that the second condition of Zangwill's global convergence theorem is satisfied by setting $z = f_{\mathrm{E}}$.

Lemma 7 *Let $(\hat{W},\hat{H})$ be any point in X. If $(\hat{W},\hat{H}) \in S_{\mathrm{E}}$ then $A(\hat{W},\hat{H}) = (\hat{W},\hat{H})$. Otherwise $f_{\mathrm{E}}(A(\hat{W},\hat{H})) < f_{\mathrm{E}}(\hat{W},\hat{H})$. That is, S_{E} is identical with the set of fixed points of the mapping A.*

Proof We first consider the case where $(\hat{W},\hat{H}) \in S_{\mathrm{E}}$. By Lemma 6, $\hat{H}$ and $\hat{W}$ are unique optimal solutions of (35) and (36), respectively. By Lemma 4, this implies $A_1(\hat{W},\hat{H}) = \hat{H}$ and $A_2(\hat{W},\hat{H}) = \hat{W}$. Therefore, we have

$$A(\hat{W},\hat{H}) = (A_2(\hat{W},A_1(\hat{W},\hat{H})),A_1(\hat{W},\hat{H})) = (A_2(\hat{W},\hat{H}),\hat{H}) = (\hat{W},\hat{H})\,.$$

We next consider the case where $(\hat{W},\hat{H}) \notin S_{\mathrm{E}}$. In this case, by Lemma 6, at least one of the following statements must be false: 1) $\hat{H}$ is the unique optimal solution of (35). 2) $\hat{W}$ is the unique optimal solution of (36). Suppose that the statement 1) does not hold true. Then, by Lemma 4, we have $g^{\hat{W}}_{\mathrm{E}}(A_1(\hat{W},\hat{H}),\hat{H}) < g^{\hat{W}}_{\mathrm{E}}(\hat{H},\hat{H})$ which implies that the second inequality of (39) holds as a strict inequality. Therefore, $f_{\mathrm{E}}(A(\hat{W},\hat{H}))$ is strictly less than $f_{\mathrm{E}}(\hat{W},\hat{H})$. Suppose next that the statement 1) holds true but 2) does not. Then, by Lemma 4, we have $A_1(\hat{W},\hat{H}) = \hat{H}$ and $h^{\hat{H}}_{\mathrm{E}}(A_2(\hat{W},\hat{H}),\hat{W}) < h^{\hat{H}}_{\mathrm{E}}(\hat{W},\hat{W})$. From these facts and (39), we have

$$f_{\mathrm{E}}(A(\hat{W},\hat{H})) = f_{\mathrm{E}}(A_2(\hat{W},A_1(\hat{W},\hat{H})),A_1(\hat{W},\hat{H})) = f_{\mathrm{E}}(A_2(\hat{W},\hat{H}),\hat{H}) < f_{\mathrm{E}}(\hat{W},\hat{H})\,.$$

Therefore, $f_{\mathrm{E}}(A(\hat{W},\hat{H}))$ is strictly less than $f_{\mathrm{E}}(\hat{W},\hat{H})$. □

4.2 Proof of Theorem 3

As in the proof of Theorem 1, let us rewrite (17) as

$$\begin{aligned} H^{k+1} &= A_1(W^k,H^k)\,, \\ W^{k+1} &= A_2(W^k,H^{k+1})\,, \end{aligned}$$

or, more simply,

$$(W^{k+1},H^{k+1}) = A(W^k,H^k)\,,$$

where the mapping A is defined by

$$A(W,H) = (A_2(W,A_1(W,H)),A_1(W,H))\,.$$

Let us also set $S = S_{\mathrm{D}}$. Since the mapping A is continuous in X, the third condition of Zangwill's global convergence theorem is satisfied. We will thus show in the following that A also satisfies the remaining two conditions.

The following lemma guarantees that the first condition is satisfied.

Lemma 8 *For any initial point $(W^0,H^0) \in X$, the sequence $\{(W^k,H^k)\}_{k=0}^{\infty}$ generated by the mapping A belongs to a compact set in X.*

Proof Let (W,H) be any point in X. Then we have

$$\begin{aligned}
H_{aj}\frac{\sum_{i=1}^n W_{ia}V_{ij}/(WH)_{ij}}{\sum_{i=1}^n W_{ia}} &= H_{aj}\sum_{i=1}^n \frac{W_{ia}V_{ij}}{\sum_{\mu=1}^n W_{\mu a}\sum_{l=1}^r W_{il}H_{lj}} \\
&= H_{aj}\sum_{i=1}^n \frac{W_{ia}V_{ij}}{\sum_{\mu=1}^n W_{\mu a}(W_{ia}H_{aj}+\sum_{l=1,l\neq a}^r W_{il}H_{lj})} \\
&= \sum_{i=1}^n \frac{W_{ia}V_{ij}}{\sum_{\mu=1}^n W_{\mu a}\{W_{ia}+\sum_{l=1,l\neq a}^r W_{il}(H_{lj}/H_{aj})\}} \\
&< \sum_{i=1}^n \frac{W_{ia}V_{ij}}{(\sum_{\mu=1}^n W_{\mu a})W_{ia}} \\
&= \sum_{i=1}^n \frac{V_{ij}}{\sum_{\mu=1}^n W_{\mu a}} \\
&\leq \frac{\sum_{i=1}^n V_{ij}}{\varepsilon n}\,,
\end{aligned}$$

from which the inequality

$$(A_1(W,H))_{aj} \leq \max\left(\frac{\sum_{i=1}^n V_{ij}}{\varepsilon n},\varepsilon\right)$$

holds for any pair (a,j). Note that the right-hand side is a constant which depends on neither W nor H. Similarly, we have

$$W_{ia}\frac{\sum_{j=1}^m H_{aj}V_{ij}/(WH)_{ij}}{\sum_{j=1}^m H_{aj}} < \frac{\sum_{j=1}^m V_{ij}}{\varepsilon m}\,,$$

from which the inequality

$$(A_2(W,H))_{ia} \leq \max\left(\frac{\sum_{j=1}^m V_{ij}}{\varepsilon m},\varepsilon\right)$$

holds for any pair (i,a). Note that the right-hand side is a constant which depends on neither W nor H. Hence $A(W,H)$ belongs to a compact set in X. This means that, for any initial point $(W^0,H^0) \in X$, the sequence $\{(W^k,H^k)\}_{k=0}^{\infty}$ generated by the mapping A belongs to a compact set in X. □

The last step is to prove that the second condition of Zangwill's global convergence theorem is also satisfied. To do this, we first need to introduce two auxiliary functions f_{D}. Let $(\hat{W},\hat{H})$ be any point in X. Let the function $g_{\mathrm{D}}^{\hat{W}} : [\varepsilon,\infty)^{r\times m} \times [\varepsilon,\infty)^{r\times m} \to \mathbb{R}$ be defined by

$$g_{\mathrm{D}}^{\hat{W}}(H,H') = \sum_{i=1}^{n}\sum_{j=1}^{m}\left\{V_{ij}\log V_{ij} - V_{ij} + \frac{V_{ij}}{(\hat{W}H')_{ij}}\sum_{a=1}^{r}\hat{W}_{ia}H'_{aj}\log\frac{\hat{W}_{ia}H'_{aj}}{(\hat{W}H')_{ij}}\right\} + \sum_{a=1}^{r}\sum_{j=1}^{m} g_{\mathrm{D}aj}^{\hat{W}}(H_{aj},H'),$$

where the function $g_{\mathrm{D}aj}^{\hat{W}} : [\varepsilon,\infty)\times[\varepsilon,\infty)^{r\times m}$ is defined by

$$g_{\mathrm{D}aj}^{\hat{W}}(H_{aj},H') = H_{aj}\sum_{i=1}^{n}\hat{W}_{ia} - H'_{aj}\sum_{i=1}^{n}\left\{\frac{\hat{W}_{ia}V_{ij}}{(\hat{W}H')_{ij}}\log(\hat{W}_{ia}H_{aj})\right\}.$$

Similarly, let the function $g_{\mathrm{D}}^{\hat{H}} : [\varepsilon,\infty)^{n\times r}\times[\varepsilon,\infty)^{n\times r} \to \mathbb{R}$ be defined by

$$h_{\mathrm{D}}^{\hat{H}}(W,W') = \sum_{i=1}^{n}\sum_{j=1}^{m}\left\{V_{ij}\log V_{ij} - V_{ij} + \frac{V_{ij}}{(W'\hat{H})_{ij}}\sum_{a=1}^{r}W'_{ia}\hat{H}_{aj}\log\frac{W'_{ia}\hat{H}_{aj}}{(W'\hat{H})_{ij}}\right\} + \sum_{i=1}^{n}\sum_{a=1}^{r} h_{\mathrm{D}ia}^{\hat{H}}(W_{ia},W'),$$

where the function $h_{\mathrm{D}ia}^{\hat{H}} : [\varepsilon,\infty)\times[\varepsilon,\infty)^{n\times r} \to \mathbb{R}$ is defined by

$$h_{\mathrm{D}ia}^{\hat{H}}(W_{ia},W') = \sum_{j=1}^{m}\left\{W_{ia}\hat{H}_{aj} - V_{ij}\frac{W'_{ia}\hat{H}_{aj}}{(W'\hat{H})_{ij}}\log(W_{ia}\hat{H}_{aj})\right\}.$$

The functions $g_{\mathrm{D}}^{\hat{W}}$ and $h_{\mathrm{D}}^{\hat{H}}$ are essentially the same as the auxiliary functions considered by Lee and Seung [17], though mathematical expressions are slightly different. In the following, we give five lemmas which are needed to prove that the second condition of Zangwill's global convergence theorem is satisfied. Although Lemmas 9 and 10 below can be immediately obtained from some of the results given by Lee and Seung [17], we will provide proofs for these lemmas in order to make this paper self-contained. On the other hand, as for Lemmas 11, 12 and 13, we omit proofs because they are similar to those for Lemmas 4, 5 and 6.

Lemma 9 *Let $(\hat{W},\hat{H})$ be any point in X. The function $g_{\mathrm{D}}^{\hat{W}}$ satisfies the following two conditions:*

$$g_{\mathrm{D}}^{\hat{W}}(H,H) = f_{\mathrm{D}}(\hat{W},H), \quad \forall H\in[\varepsilon,\infty)^{r\times m}, \tag{40}$$

$$g_{\mathrm{D}}^{\hat{W}}(H,H') \geq f_{\mathrm{D}}(\hat{W},H), \quad \forall H,H'\in[\varepsilon,\infty)^{r\times m}. \tag{41}$$

Also, the function $h_{\mathrm{D}}^{\hat{H}}$ satisfies the following two conditions:

$$h_{\mathrm{D}}^{\hat{H}}(W,W) = f_{\mathrm{D}}(W,\hat{H}), \quad \forall W\in[\varepsilon,\infty)^{n\times r},$$

$$h_{\mathrm{D}}^{\hat{H}}(W,W') \geq f_{\mathrm{D}}(W,\hat{H}), \quad \forall W,W'\in[\varepsilon,\infty)^{n\times r}.$$

Proof We prove only the first part because the second part can be proved in the same way. For any $\hat{W} \in [\varepsilon,\infty)^{n\times r}$ and $H \in [\varepsilon,\infty)^{r\times m}$, $g_{\mathrm{D}}^{\hat{W}}(H,H)$ can be transformed as

$$\begin{aligned}
g_{\mathrm{D}}^{\hat{W}}(H,H) &= \sum_{i=1}^{n}\sum_{j=1}^{m}\left\{V_{ij}\log V_{ij} - V_{ij} + \frac{V_{ij}}{(\hat{W}H)_{ij}}\sum_{a=1}^{r}\hat{W}_{ia}H_{aj}\log\frac{\hat{W}_{ia}H_{aj}}{(\hat{W}H)_{ij}}\right\} \\
&\quad + \sum_{a=1}^{r}\sum_{j=1}^{m}\left\{H_{aj}\sum_{i=1}^{n}\hat{W}_{ia} - H_{aj}\sum_{i=1}^{n}\frac{\hat{W}_{ia}V_{ij}}{(\hat{W}H)_{ij}}\log(\hat{W}_{ia}H_{aj})\right\} \\
&= \sum_{i=1}^{n}\sum_{j=1}^{m}(V_{ij}\log V_{ij} - V_{ij}) - \sum_{i=1}^{n}\sum_{j=1}^{m}V_{ij}\log(\hat{W}H)_{ij} + \sum_{i=1}^{n}\sum_{j=1}^{m}(\hat{W}H)_{ij} \\
&= \sum_{i=1}^{n}\sum_{j=1}^{m}\left\{V_{ij}\log\frac{V_{ij}}{(\hat{W}H)_{ij}} - V_{ij} + (\hat{W}H)_{ij}\right\} \\
&= f_{\mathrm{D}}(\hat{W},H)\,.
\end{aligned}$$

Thus the condition (40) holds true. In order to show (41), we consider

$$g_{\mathrm{D}}^{\hat{W}}(H,H') - f_{\mathrm{D}}(\hat{W},H) = \sum_{i=1}^{n}\sum_{j=1}^{m}V_{ij}\left\{\log(\hat{W}H)_{ij} - \log\frac{H_{aj}}{H'_{aj}} - \frac{\hat{W}_{ia}H'_{aj}}{(\hat{W}H')_{ij}}\log(\hat{W}H')_{ij}\right\}. \tag{42}$$

From the concavity of the log function,

$$\log(\hat{W}H)_{ij} = \log\left(\sum_{a=1}^{r}\hat{W}_{ia}H_{aj}\right) \geq \sum_{a=1}^{r}\mu_a\log\left(\frac{\hat{W}_{ia}H_{aj}}{\mu_a}\right) \tag{43}$$

for any $H \in [\varepsilon,\infty)^{r\times m}$ and any set of positive numbers $\mu_1,\mu_2,\ldots,\mu_r$ such that $\sum_{a=1}^{r}\mu_a = 1$. By substituting $\mu_a = (\hat{W}_{ia}H'_{aj})/(\hat{W}H')_{ij}$ for $a = 1,2,\ldots,r$ into (43), we have

$$\begin{aligned}
\log(\hat{W}H)_{ij} &= \log\left(\sum_{a=1}^{r}\hat{W}_{ia}H_{aj}\right) \\
&\geq \sum_{a=1}^{r}\frac{\hat{W}_{ia}H'_{aj}}{(\hat{W}H')_{ij}}\log\left(\hat{W}_{ia}H_{aj}\cdot\frac{(\hat{W}H')_{ij}}{\hat{W}_{ia}H'_{aj}}\right) \\
&\geq \sum_{a=1}^{r}\frac{\hat{W}_{ia}H'_{aj}}{(\hat{W}H')_{ij}}\left\{\log(\hat{W}H')_{ij} + \log\frac{H_{aj}}{H'_{aj}}\right\},
\end{aligned}$$

which implies that the right-hand side of (42) is nonnegative for all $H,H' \in [\varepsilon,\infty)^{r\times m}$. □

Lemma 10 *Let $(\hat{W},\hat{H})$ be any point in X. Then $g_{\mathrm{D}}^{\hat{W}}(H,\hat{H})$, which is considered as a function of H, is strictly convex in $[\varepsilon,\infty)^{r\times m}$. Also, $h_{\mathrm{D}}^{\hat{H}}(W,\hat{W})$, which is considered as a function of W, is strictly convex in $[\varepsilon,\infty)^{n\times r}$.*

Proof The second-order partial derivatives of $g_{\mathrm{D}}^{\hat{W}}(H,\hat{H})$ are given by

$$\frac{\partial^2 g_{\mathrm{D}}^{\hat{W}}(H,\hat{H})}{\partial H_{aj}\partial H_{a'j'}} = \begin{cases} \frac{\hat{H}_{aj}}{H_{aj}^2}\sum_{i=1}^n \frac{\hat{W}_{ia}V_{ij}}{(\hat{W}\hat{H})_{ij}}, & \text{if } (a,j)=(a',j') \\ 0, & \text{otherwise}, \end{cases}$$

where $(a,j),(a',j') \in \{1,2,\ldots,r\}\times\{1,2,\ldots,m\}$. Note here that $(\hat{H}_{aj}/H_{aj}^2)\sum_{i=1}^n(\hat{W}_{ia}V_{ij}/(\hat{W}\hat{H})_{ij})$ is positive for all $H_{aj} \in [\varepsilon,\infty)$ because of Assumption 1. Therefore, $g_{\mathrm{D}}^{\hat{W}}(H,\hat{H})$ is strictly convex in $[\varepsilon,\infty)^{r\times m}$. The second part can be proved in the same way. □

Lemma 11 *Let $(\hat{W},\hat{H})$ be any point in X. The optimization problem*

$$\begin{array}{l} \textit{minimize } g_{\mathrm{D}}^{\hat{W}}(H,\hat{H}) \\ \textit{subject to } H_{aj} \geq \varepsilon, \quad \forall a,j \end{array} \tag{44}$$

has a unique optimal solution which is given by $A_1(\hat{W},\hat{H})$. Also, the optimization problem

$$\begin{array}{l} \textit{minimize } h_{\mathrm{D}}^{\hat{H}}(W,\hat{W}) \\ \textit{subject to } W_{ia} \geq \varepsilon, \quad \forall i,a \end{array} \tag{45}$$

has a unique optimal solution which is given by $A_2(\hat{W},\hat{H})$.

Lemma 12 *The inequality $f_{\mathrm{D}}(A(\hat{W},\hat{H})) \leq f_{\mathrm{D}}(\hat{W},\hat{H})$ holds for all $(\hat{W},\hat{H}) \in X$.*

Lemma 13 *$(\hat{W},\hat{H}) \in S_{\mathrm{D}}$ if and only if $\hat{H}$ and $\hat{W}$ are the optimal solutions of (44) and (45), respectively.*

From Lemmas 11–13, we derive the following lemma which claims that the second condition of Zangwill's global convergence theorem is satisfied by setting $z = f_{\mathrm{D}}$. The proof is omitted because it is similar to that for Lemma 7.

Lemma 14 *Let $(\hat{W},\hat{H})$ be any point in X. If $(\hat{W},\hat{H}) \in S_{\mathrm{D}}$ then $A(\hat{W},\hat{H}) = (\hat{W},\hat{H})$. Otherwise $f_{\mathrm{D}}(A(\hat{W},\hat{H})) < f_{\mathrm{D}}(\hat{W},\hat{H})$. That is, S_{D} is identical with the set of fixed points of the mapping A.*

5 Conclusion

We have shown that the global convergence of the multiplicative updates proposed by Lee and Seung is established if they are slightly modified as discussed by Gillis and Glineur. Their idea is just to prevent each variable from becoming smaller than a user-specified positive constant ε, but this slight modification guarantees the boundedness of solutions without normalization. Using Zangwill's global convergence theorem, we have proved that any sequence of solutions generated by the modified updates has at least one convergent subsequence and the limit of any convergent subsequence is a stationary point of the corresponding optimization problem. Furthermore, we have developed two algorithms based on the modified updates which always stop within a finite number of iterations after finding an approximate stationary point.

One may be concerned with the fact that matrices obtained by the modified updates are always dense. However, when sparse matrices are preferable, we only have to replace all ε in the obtained matrices with zero. If ε is set to a small positive number, this replacement will not affect the results significantly. It is in fact proved that setting the entries of W and H equal to ε to zero gives a solution which is $\mathcal{O}(\varepsilon)$ close to a stationary point of the original problem, and that the objective function is affected by an additive factor of at most $\mathcal{O}(\varepsilon)$ [11].

The approach presented in this paper may be applied to various multiplicative algorithms for NMF or other optimization problems. Developing a unified framework for the global convergence analysis of multiplicative updates is a topic for future research.

Acknowledgements This work was partially supported by JSPS KAKENHI Grant Numbers 24560076 and 23310104, and by the project "R&D for cyber-attack predictions and rapid response technology by means of international cooperation" of the Ministry of Internal Affairs and Communications, Japan.

References

1. Badeau, R., Bertin, N., Vincent, E.: Stability analysis of multiplicative update algorithms and application to nonnegative matrix factorization. IEEE Transactions on Neural Networks **21**(12), 1869–1881 (2010)
2. Berry, M.W., Browne, M.: Email surveillance using non-negative matrix factorization. Computational and Mathematical Organization Theory **11**, 249–264 (2005)
3. Berry, M.W., Browne, M., Langville, A.N., Pauca, V.P., Plemmons, R.J.: Algorithms and applications for approximate nonnegative matrix factorization. Computational Statistics and Data Analysis **52**, 155–173 (2007)
4. Boyd, S., Vandenberghe, L.: Convex Optimization. Cambridge University Press, Cambridge, UK (2004)
5. Brunet, J.P., Tamayo, P., Golub, T.R., Mesirov, J.P.: Metagenes and molecular pattern discovery using matrix factorization. Proceedings of National Academy of Science **101**(12), 4164–4169 (2004)
6. Cai, D., He, X., Han, J., Huang, T.S.: Graph regularized nonnegative matrix factorization for data representation. IEEE Transactions on Pattern Analysis and Machine Intelligence **33**(8), 1548–1560 (2011)
7. Cichocki, A., Zdunek, R., Phan, A.H., Amari, S.I.: Nonnegative Matrix and Tensor Factorizations. John Wiley & Sons, West Sussex, U.K. (2009)
8. Csiszár, I., Tusnády, G.: Information geometry and alternating minimization procedures. Statistics and Decisions, Supplemental Issue pp. 205–237 (1984)
9. Févotte, C., Bertin, N., Durrieu, J.L.: Nonnegative matrix factorization with the Itakura-Saito divergence: With application to music analysis. Neural Computation **21**(3), 793–830 (2009)
10. Finesso, L., Spreij, P.: Nonnegative matrix factorization and I-divergence alternating minimization. Linear Algebra and its Applications **416**, 270–287 (2006)
11. Gillis, N.: Nonnegative matrix factorization: Complexity, algorithms and applications. Ph.D. thesis, Université Catholique de Louvain, Louvain-la-Neuve (2011)
12. Gillis, N., Glineur, F.: Nonnegative factorization and the maximum edge biclique problem. ArXiv e-prints (2008)
13. Gillis, N., Glineur, F.: Accelerated multiplicative updates and hierarchical ALS algorithms for nonnegative matrix factorization. Neural Computation **24**(4), 1085–1105 (2012)
14. Hibi, R., Takahashi, N.: A modified multiplicative update algorithm for Euclidean distance-based nonnegative matrix factorization and its global convergence. In: Proceedings of 18th International Conference on Neural Information Processing, Part-II, pp. 655–662 (2011)
15. Holzapfel, A., Stylianou, Y.: Musical genre classification using nonnegative matrix factorization-based features. IEEE Transactions on Audio, Speech, and Language Processing **16**(2), 424–434 (2008)

16. Lee, D.D., Seung, H.S.: Learning the parts of objects by non-negative matrix factorization. Nature **401**, 788–792 (1999)
17. Lee, D.D., Seung, H.S.: Algorithms for non-negative matrix factorization. In: T.K. Leen, T.G. Dietterich, V. Tresp (eds.) Advances in Neural Information Processing Systems, vol. 13, pp. 556–562 (2001)
18. Lin, C.J.: On the convergence of multiplicative update algorithms for nonnegative matrix factorization. IEEE Transactions on Neural Networks **18**(6), 1589–1596 (2007)
19. Lin, C.J.: Projected gradient methods for non-negative matrix factorization. Neural Computation **19**(10), 2756–2779 (2007)
20. Lu, W., Sun, W., Lu, H.: Robust watermarking based on DWT and nonnegative matrix factorization. Computers and Electrical Engineering **35**, 183–188 (2009)
21. Sha, F., Min, Y., Saul, L.K., Lee, D.D.: Multiplicative updates for nonnegative quadratic programming. Neural Computation **19**, 2004–2031 (2007)
22. Shahnaz, F., Berry, M.W., Pauca, V.P., Plemmons, R.J.: Document clustering using nonnegative matrix factorization. Information Processing and Management **42**, 373–386 (2006)
23. Takahashi, N., Nishi, T.: Global convergence of decomposition learning methods for support vector machines. IEEE Transactions on Neural Networks **17**(6), 1362–1369 (2006)
24. Vavasis, S.A.: On the complexity of nonnegative matrix factorization. SIAM Journal on Optimization **20**(3), 1364–1377 (2009)
25. Wang, R.S., Zhang, S., Wang, Y., Zhang, X.S., Chen, L.: Clustering complex networks and biological networks by nonnegative matrix factorization with various similarity measures. Neurocomputing **72**, 134–141 (2008)
26. Wu, C.F.J.: On the convergence properties of the EM algorithm. The Ananls of Statistics **11**(1), 95–103 (1983)
27. Yang, Z., Oja, E.: Unified development of multiplicative algorithm for linear and quadratic nonnegative matrix factorization. IEEE Transactions on Neural Networks **22**(12), 1878–1891 (2011)
28. Zangwill, W.I.: Nonlinear programming: A unified approach. Prentice-Hall, Englewood Cliffs, NJ (1969)

Erratum Sheet

Norikazu Takahashi and Ryota Hibi, "Global convergence of modified multiplicative updates for nonnegative matrix factorization", Computational Optimization and Applications, vol.57, pp.417–440, 2014.

1. Eq.(42) on Page 437 is not correct. It should be replaced with

$$g_{\mathrm{D}}^{\hat{W}}(H,H')-f_{\mathrm{D}}(\hat{W},H)=\sum_{i=1}^{n}\sum_{j=1}^{m}V_{ij}\left\{\log(\hat{W}H)_{ij}-\sum_{a=1}^{r}\frac{\hat{W}_{ia}H'_{aj}}{(\hat{W}H')_{ij}}\log\frac{H_{aj}}{H'_{aj}}-\log(\hat{W}H')_{ij}\right\}$$

which can be derived as follows:

$$\begin{aligned}
&g_{\mathrm{D}}^{\hat{W}}(H,H')-f_{\mathrm{D}}(\hat{W},H)\\
&=\sum_{i=1}^{n}\sum_{j=1}^{m}\left\{V_{ij}\log V_{ij}-V_{ij}+\frac{V_{ij}}{(\hat{W}H')_{ij}}\sum_{a=1}^{r}\hat{W}_{ia}H'_{aj}\log\frac{\hat{W}_{ia}H'_{aj}}{(\hat{W}H')_{ij}}\right\}\\
&\quad+\sum_{a=1}^{r}\sum_{j=1}^{m}\left\{H_{aj}\sum_{i=1}^{n}\hat{W}_{ia}-H'_{aj}\sum_{i=1}^{n}\frac{\hat{W}_{ia}V_{ij}}{(\hat{W}H')_{ij}}\log(\hat{W}_{ia}H_{aj})\right\}\\
&\quad-\sum_{i=1}^{n}\sum_{j=1}^{m}\left\{V_{ij}\log\frac{V_{ij}}{(\hat{W}H)_{ij}}-V_{ij}+(\hat{W}H)_{ij}\right\}\\
&=\sum_{i=1}^{n}\sum_{j=1}^{m}\left\{V_{ij}\log V_{ij}-V_{ij}+V_{ij}\sum_{a=1}^{r}\frac{\hat{W}_{ia}H'_{aj}}{(\hat{W}H')_{ij}}\log\frac{\hat{W}_{ia}H'_{aj}}{(\hat{W}H')_{ij}}\right\}\\
&\quad+\sum_{i=1}^{n}\sum_{j=1}^{m}\left\{\sum_{a=1}^{r}\hat{W}_{ia}H_{aj}-V_{ij}\sum_{a=1}^{r}\frac{\hat{W}_{ia}H'_{aj}}{(\hat{W}H')_{ij}}\log(\hat{W}_{ia}H_{aj})\right\}\\
&\quad-\sum_{i=1}^{n}\sum_{j=1}^{m}\left\{V_{ij}\log\frac{1}{(\hat{W}H)_{ij}}-V_{ij}+(\hat{W}H)_{ij}\right\}\\
&=\sum_{i=1}^{n}\sum_{j=1}^{m}V_{ij}\left\{\log(\hat{W}H)_{ij}+\sum_{a=1}^{r}\frac{\hat{W}_{ia}H'_{aj}}{(\hat{W}H')_{ij}}\log\frac{\hat{W}_{ia}H'_{aj}}{(\hat{W}H')_{ij}\hat{W}_{ia}H_{aj}}\right\}\\
&=\sum_{i=1}^{n}\sum_{j=1}^{m}V_{ij}\left\{\log(\hat{W}H)_{ij}-\sum_{a=1}^{r}\frac{\hat{W}_{ia}H'_{aj}}{(\hat{W}H')_{ij}}\log\frac{H_{aj}}{H'_{aj}}-\sum_{a=1}^{r}\frac{\hat{W}_{ia}H'_{aj}}{(\hat{W}H')_{ij}}\log(\hat{W}H')_{ij}\right\}\\
&=\sum_{i=1}^{n}\sum_{j=1}^{m}V_{ij}\left\{\log(\hat{W}H)_{ij}-\sum_{a=1}^{r}\frac{\hat{W}_{ia}H'_{aj}}{(\hat{W}H')_{ij}}\log\frac{H_{aj}}{H'_{aj}}-\log(\hat{W}H')_{ij}\right\}.
\end{aligned}$$

(Last updated: September 3, 2014)