Journal of Global Optimization (2022) 84:755–781
https://doi.org/10.1007/s10898-022-01167-7

A novel update rule of HALS algorithm for nonnegative matrix factorization and Zangwill's global convergence

Takehiro Sano[1] · **Tsuyoshi Migita**[1] · **Norikazu Takahashi**[1]

Received: 12 April 2021 / Accepted: 7 April 2022 / Published online: 30 April 2022

Abstract
Nonnegative Matrix Factorization (NMF) has attracted a great deal of attention as an effective technique for dimensionality reduction of large-scale nonnegative data. Given a nonnegative matrix, NMF aims to obtain two low-rank nonnegative factor matrices by solving a constrained optimization problem. The Hierarchical Alternating Least Squares (HALS) algorithm is a well-known and widely-used iterative method for solving such optimization problems. However, the original update rule used in the HALS algorithm is not well defined. In this paper, we propose a novel well-defined update rule of the HALS algorithm, and prove its global convergence in the sense of Zangwill. Unlike conventional globally-convergent update rules, the proposed one allows variables to take the value of zero and hence can obtain sparse factor matrices. We also present two stopping conditions that guarantee the finite termination of the HALS algorithm. The practical usefulness of the proposed update rule is shown through experiments using real-world datasets.

Keywords Nonnegative matrix factorization · Hierarchical alternating least squares algorithm · Global convergence

1 Introduction

Dimensionality reduction methods for large-scale and high-dimensional data have been actively studied in the fields of machine learning and signal processing because of their diverse applications such as feature extraction and visualization (see [8] and references therein). In recent years, Nonnegative Matrix Factorization (NMF) [2, 32] has attracted a great deal of attention as an effective dimensionality reduction method for large-scale nonnegative data, and has been successfully applied to various tasks such as image processing [4, 34], acoustic signal processing [14, 31], network analysis [18, 22, 50], mobile sensor calibration [12] and so on. A key difference between NMF and other dimensionality reduction methods such

This work was supported by JSPS KAKENHI Grant Number JP21H03510.

✉ Norikazu Takahashi
takahashi@okayama-u.ac.jp

[1] Graduate School of Natural Science and Technology, Okayama University, 3–1–1 Tsushima-naka, Kita-ku, Okayama 700–8530, Japan

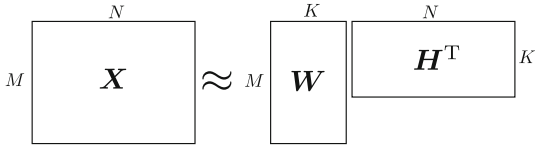

Fig. 1 Nonnegative matrix factorization

as the principal component analysis [51] is that the factor matrices obtained by NMF are nonnegative and tend to be sparse [32]. Thus NMF can learn a parts-based representation of the data [32].

Given an $M \times N$ nonnegative matrix $\boldsymbol{X}$, NMF aims to decompose it into two nonnegative factor matrices $\boldsymbol{W}$ and $\boldsymbol{H}$ of sizes $M \times K$ and $N \times K$, respectively, so that $\boldsymbol{W}\boldsymbol{H}^{\mathrm{T}}$ is approximately equal to $\boldsymbol{X}$, where K is much less than $\min\{M, N\}$ (see Fig. 1). The problem of finding such factor matrices is often formulated as the constrained optimization problem:

$$\begin{aligned} &\text{minimize} \ \ f(\boldsymbol{W}, \boldsymbol{H}) = \frac{1}{2} \left\| \boldsymbol{X} - \boldsymbol{W}\boldsymbol{H}^{\mathrm{T}} \right\|_{\mathrm{F}}^2 \\ &\text{subject to} \ \ \boldsymbol{W} \geq \boldsymbol{0}_{M \times K}, \ \boldsymbol{H} \geq \boldsymbol{0}_{N \times K}, \end{aligned} \tag{1}$$

where $\| \cdot \|_{\mathrm{F}}$ denotes the Frobenius norm of matrices, and $\boldsymbol{0}_{I \times J}$ is the $I \times J$ matrix of all zeros. For matrices $\boldsymbol{P}$ and $\boldsymbol{Q}$ of the same size, $\boldsymbol{P} \geq \boldsymbol{Q}$ means element-wise inequality. The Frobenius norm can be replaced with one of several alternatives such as the I-divergence [33], the Itakura-Saito divergence [14] and others [52]. Also, one or more regularization terms can be added to the objective function in order to enforce desirable properties on the factor matrices [24, 25, 40, 41]. As with many machine learning methods, the ℓ_1-regularization term is often used in NMF.

Various methods for finding a local optimal solution of the optimization problem (1) have been developed so far. Note that finding a global optimal solution is difficult in general because it is known that (1) is NP-hard [49]. Most of the conventional methods update some or all of the elements of one factor matrix at a time because the objective function $f(\boldsymbol{W}, \boldsymbol{H})$ is not jointly convex but convex in $\boldsymbol{W}$ or $\boldsymbol{H}$. For example, the multiplicative update rule (MUR) [33], which is widely known as a simple and easy-to-implement method, alternately updates $\boldsymbol{W}$ and $\boldsymbol{H}$ according to the rule derived from strictly convex functions called the auxiliary functions [33]. An important advantage of the MUR is that the value of the objective function decreases monotonically as long as division by zero does not occur. However, division by zero is certainly possible in the MUR because elements of the factor matrices can become zero. For this reason, convergence of the factor matrices is not guaranteed. In fact, it was shown experimentally that the MUR sometimes fails to converge to a stationary point [19]. To solve this problem, some authors proposed modified MURs [16, 35]. For example, Gillis and Glineur [16] proposed to replace all values less than a positive constant ϵ with ϵ after updating $\boldsymbol{W}$ and $\boldsymbol{H}$ using the original MUR. Their modified MUR was later proved by Takahashi and Hibi [46] to be globally convergent in the sense of Zangwill [53] (see Definition 1 of the present paper) to a stationary point of the corresponding optimization problem:

$$\begin{aligned} &\text{minimize} \ \ f(\boldsymbol{W}, \boldsymbol{H}) = \frac{1}{2} \left\| \boldsymbol{X} - \boldsymbol{W}\boldsymbol{H}^{\mathrm{T}} \right\|_{\mathrm{F}}^2 \\ &\text{subject to} \ \ \boldsymbol{W} \geq \epsilon \boldsymbol{1}_{M \times K}, \ \boldsymbol{H} \geq \epsilon \boldsymbol{1}_{N \times K}, \end{aligned} \tag{2}$$

where $\boldsymbol{1}_{I \times J}$ denotes the $I \times J$ matrix of all ones. Lin [35] proposed a different kind of modified MUR and proved its global convergence to a stationary point of (1). However, this modified MUR is much more complicated than the one mentioned above, and requires a higher computational cost.

Another well-known method for solving (1) is the Hierarchical Alternating Least Squares (HALS) algorithm [6, 7], which is much faster than the MUR in many cases, and much simpler than other fast algorithms [17, 20, 26, 28, 36, 54]. The HALS algorithm updates one column of the factor matrices at a time according to the rule derived from the partial derivative of the objective function with respect to the column. The value of the objective function decreases monotonically if the columns of the factor matrices remain nonzero throughout the iterations [27]. However, as with the MUR, elements of the factor matrices can become zero and this may cause division by zero. To solve this problem, some authors proposed modified update rules for the HALS algorithm [6, 15]. The one proposed by Cichocki *et al*. [6] takes the same approach as the modified MURs [16]. It replaces all values less than a positive constant ϵ with ϵ after updating each column of the factor matrices using the original update rule. Although the global convergence to a stationary point of (2) has been proved [29], this update rule cannot obtain sparse factor matrices for the same reason as stated above. In contrast, the update rule given by Gillis [15] not only allows variables to be zero but also avoids division by zero. Furthermore, the value of the objective function decreases monotonically under this update rule. However, the global convergence to a stationary point of (1) is not guaranteed because the level set of the objective function is unbounded.

In this paper, we propose a novel update rule for the HALS algorithm, and prove its global convergence to a stationary point of (1) using Zangwill's global convergence theorem [53]. The proposed update rule is a combination of the original update rule, the update rule of Gillis [15] and a normalization step. The normalization step is elaborately designed to guarantee not only the boundedness of variables but also the closedness of the point-to-set mapping representing the proposed update rule. We also present two stopping conditions that guarantee the finite termination of the HALS algorithm using the proposed update rule. In addition, the practical usefulness of the proposed update rule is shown through experiments using real-world datasets.

There are many variants of NMF. For example, NMF with additional constraints such as orthogonality [9], symmetry [37] and separability [1, 11, 43] have been extensively studied. These variants are important not only from a theoretical viewpoint but also in practice. In fact, they have many applications in document clustering, community detection, dictionary learning and so on. However, we do not consider these variants in this paper because they need their own specialized algorithms.

The remainder of this paper is organized as follows. In Sect. 2, notations and definitions used in later sections are presented. In Sect. 3, the conventional update rules of the HALS algorithm and their convergence property are reviewed. In Sect. 4, a novel update rule of the HALS algorithm is proposed and its global convergence is proved. In Sect. 5, two stopping conditions are presented and the finite termination of the HALS algorithm using these stopping conditions is proved. In Sect. 6, some experimental results are presented to show the practical usefulness of the proposed update rule. Section 7 introduces some variants of the HALS algorithm to which the proposed update rule can be applied. Section 8 concludes this work and discusses a possible future direction.

2 Notations and definitions

The sets of integers, nonnegative integers, and positive integers are denoted by $\mathbb{Z}$, $\mathbb{Z}_+$ and $\mathbb{Z}_{++}$, respectively. Similarly, the sets of real numbers, nonnegative real numbers, and positive

real numbers are denoted by $\mathbb{R}$, $\mathbb{R}_+$ and $\mathbb{R}_{++}$, respectively. The $I \times J$ matrix of all zeros and that of all ones are denoted by $\mathbf{0}_{I\times J}$ and $\mathbf{1}_{I\times J}$, respectively.

For any vector $\boldsymbol{v} = (v_1, v_2, \ldots, v_I)^{\mathrm{T}} \in \mathbb{R}^I$, ℓ_1- and ℓ_2-norms of $\boldsymbol{v}$ are denoted by $\|\boldsymbol{v}\|_1$ and $\|\boldsymbol{v}\|_2$, respectively. The notation $[\boldsymbol{v}]_+$ represents the vector of which the i-th element is given by $\max\{0, v_i\}$ for all i. Similarly, for any vector $\boldsymbol{v} \in \mathbb{R}^I$ and any constant $\epsilon \in \mathbb{R}_{++}$, the notation $[\boldsymbol{v}]_{\epsilon+}$ represents the vector of which the i-th element is given by $\max\{\epsilon, v_i\}$ for all i.

The feasible region of the constrained optimization problem (1) is denoted by $\mathcal{F}$. That is, $\mathcal{F} = \mathbb{R}_+^{M\times K} \times \mathbb{R}_+^{N\times K}$. We call $(\boldsymbol{W}, \boldsymbol{H}) \in \mathbb{R}^{M\times K} \times \mathbb{R}^{N\times K}$ a stationary point of (1) if it satisfies the Karush-Kuhn-Tucker (KKT) conditions:

$$\boldsymbol{W} \geq \mathbf{0}_{M\times K}, \tag{3a}$$
$$\boldsymbol{H} \geq \mathbf{0}_{N\times K}, \tag{3b}$$
$$\nabla_{\boldsymbol{W}} f(\boldsymbol{W}, \boldsymbol{H}) \geq \mathbf{0}_{M\times K}, \tag{3c}$$
$$\nabla_{\boldsymbol{H}} f(\boldsymbol{W}, \boldsymbol{H}) \geq \mathbf{0}_{N\times K}, \tag{3d}$$
$$\nabla_{\boldsymbol{W}} f(\boldsymbol{W}, \boldsymbol{H}) \odot \boldsymbol{W} = \mathbf{0}_{M\times K}, \tag{3e}$$
$$\nabla_{\boldsymbol{H}} f(\boldsymbol{W}, \boldsymbol{H}) \odot \boldsymbol{H} = \mathbf{0}_{N\times K}, \tag{3f}$$

where

$$\nabla_{\boldsymbol{W}} f(\boldsymbol{W}, \boldsymbol{H}) = (\boldsymbol{W}\boldsymbol{H}^{\mathrm{T}} - \boldsymbol{X})\boldsymbol{H},$$
$$\nabla_{\boldsymbol{H}} f(\boldsymbol{W}, \boldsymbol{H}) = (\boldsymbol{H}\boldsymbol{W}^{\mathrm{T}} - \boldsymbol{X}^{\mathrm{T}})\boldsymbol{W},$$

and $\odot$ represents the element-wise product. The set of stationary points of (1) is denoted by $\mathcal{S}$.

Similarly, the feasible region of the constrained optimization problem (2) is denoted by $\mathcal{F}_\epsilon$. That is, $\mathcal{F}_\epsilon = [\epsilon, \infty)^{M\times K} \times [\epsilon, \infty)^{N\times K}$. We call $(\boldsymbol{W}, \boldsymbol{H}) \in \mathbb{R}^{M\times K} \times \mathbb{R}^{N\times K}$ a stationary point of (2) if it satisfies the KKT conditions:

$$\boldsymbol{W} \geq \epsilon\mathbf{1}_{M\times K}, \tag{4a}$$
$$\boldsymbol{H} \geq \epsilon\mathbf{1}_{N\times K}, \tag{4b}$$
$$\nabla_{\boldsymbol{W}} f(\boldsymbol{W}, \boldsymbol{H}) \geq \mathbf{0}_{M\times K}, \tag{4c}$$
$$\nabla_{\boldsymbol{H}} f(\boldsymbol{W}, \boldsymbol{H}) \geq \mathbf{0}_{N\times K}, \tag{4d}$$
$$\nabla_{\boldsymbol{W}} f(\boldsymbol{W}, \boldsymbol{H}) \odot (\boldsymbol{W} - \epsilon\mathbf{1}_{M\times K}) = \mathbf{0}_{M\times K}, \tag{4e}$$
$$\nabla_{\boldsymbol{H}} f(\boldsymbol{W}, \boldsymbol{H}) \odot (\boldsymbol{H} - \epsilon\mathbf{1}_{N\times K}) = \mathbf{0}_{N\times K}, \tag{4f}$$

The set of stationary points of (2) is denoted by $\mathcal{S}_\epsilon$.

Many iterative algorithms for solving (1) have been proposed so far. Such an algorithm starts with an initial point $(\boldsymbol{W}^{(0)}, \boldsymbol{H}^{(0)}) \in \mathcal{F}$ and generates a sequence of points $\{(\boldsymbol{W}^{(t)}, \boldsymbol{H}^{(t)})\}_{t=0}^{\infty} \subset \mathcal{F}$ that is expected to converge to a stationary point of (1). Following to Zangwill [53], we define the global convergence of an iterative algorithm for solving (1) as follows.

Definition 1 (Global Convergence) An iterative algorithm for solving (1) is said to be globally convergent to $\mathcal{S}$ if any sequence $\{(\boldsymbol{W}^{(t)}, \boldsymbol{H}^{(t)})\}_{t=0}^{\infty} \subset \mathcal{F}$ generated by the algorithm has at least one convergent subsequence and the limit of any convergent subsequence belongs to $\mathcal{S}$.

Note that Definition 1 does not mean the convergence of the whole sequence $\{(\boldsymbol{W}^{(t)}, \boldsymbol{H}^{(t)})\}_{t=0}^{\infty}$ to a stationary point. Nevertheless, the notion of global convergence as defined above is of

great practical importance because the finite termination of the algorithm is guaranteed if we relax the KKT conditions in a proper way and use them as the stopping condition [29, 46, 47].

Using Zangwill's global convergence theorem [53], we can obtain a theorem that gives a sufficient condition for an iterative algorithm for solving (1) to be globally convergent to $\mathcal{S}$. Before presenting the theorem, we introduce two important notions: point-to-set mappings and their closedness. We consider every iterative algorithm for solving (1) as an iterative process of defining a set of candidate points in the next iteration from the point in the current iteration, and selecting one from the candidate points in some way. Each algorithm is thus characterized by how to define the set of candidate points, which is represented by a point-to-set mapping from $\mathcal{F}$ to its subsets. For point-to-set mappings from $\mathcal{F}$ to its subsets, their closedness is defined as follows.

Definition 2 (Closed Mapping) A point-to-set mapping A from $\mathcal{F}$ to its subsets is said to be closed on $\mathcal{D} \subseteq \mathcal{F}$ if, for any sequence $\{(\boldsymbol{P}^{(t)}, \boldsymbol{Q}^{(t)})\}_{t=0}^{\infty} \subset \mathcal{F}$ that converges to $(\boldsymbol{P}^{(\infty)}, \boldsymbol{Q}^{(\infty)}) \in \mathcal{D}$ and any sequence $\{(\boldsymbol{U}^{(t)}, \boldsymbol{V}^{(t)})\}_{t=0}^{\infty} \subset \mathcal{F}$ such that $(\boldsymbol{U}^{(t)}, \boldsymbol{V}^{(t)}) \in A(\boldsymbol{P}^{(t)}, \boldsymbol{Q}^{(t)})$ for all $t \in \mathbb{Z}_+$ and it converges to $(\boldsymbol{U}^{(\infty)}, \boldsymbol{V}^{(\infty)}) \in \mathcal{F}$, their limits satisfy $(\boldsymbol{U}^{(\infty)}, \boldsymbol{V}^{(\infty)}) \in A(\boldsymbol{P}^{(\infty)}, \boldsymbol{Q}^{(\infty)})$.

It is often the case that the set $A(\boldsymbol{W}, \boldsymbol{H})$ consists of only one point in $\mathcal{F}$ for any $(\boldsymbol{W}, \boldsymbol{H}) \in \mathcal{F}$. In this case, A can be considered as a point-to-point mapping from $\mathcal{F}$ to itself, and the closedness defined above can be considered as the continuity of A.

Now we are ready to present a theorem that can be obtained as a direct consequence of Zangwill's global convergence theorem [53].

Theorem 1 *Let A be the point-to-set mapping from $\mathcal{F}$ to its subsets that represents an iterative algorithm for solving (1). If A satisfies the following conditions then the algorithm is globally convergent to $\mathcal{S}$.*

1. *Any sequence $\{(\boldsymbol{W}^{(t)}, \boldsymbol{H}^{(t)})\}_{t=0}^{\infty}$ generated by the mapping A in such a way that $(\boldsymbol{W}^{(0)}, \boldsymbol{H}^{(0)}) \in \mathcal{F}$ and $(\boldsymbol{W}^{(t+1)}, \boldsymbol{H}^{(t+1)}) \in A(\boldsymbol{W}^{(t)}, \boldsymbol{H}^{(t)})$ for all $t \in \mathbb{Z}_+$ is contained in a compact subset of $\mathcal{F}$.*
2. *The mapping A does not increase the value of f. To be more specific, for any point $(\boldsymbol{W}, \boldsymbol{H}) \in \mathcal{F}$, the following statements hold true.*
   (a) *If $(\boldsymbol{W}, \boldsymbol{H}) \notin \mathcal{S}$ then $f(\boldsymbol{U}, \boldsymbol{V}) < f(\boldsymbol{W}, \boldsymbol{H})$ for all $(\boldsymbol{U}, \boldsymbol{V}) \in A(\boldsymbol{W}, \boldsymbol{H})$.*
   (b) *If $(\boldsymbol{W}, \boldsymbol{H}) \in \mathcal{S}$ then $f(\boldsymbol{U}, \boldsymbol{V}) \leq f(\boldsymbol{W}, \boldsymbol{H})$ for all $(\boldsymbol{U}, \boldsymbol{V}) \in A(\boldsymbol{W}, \boldsymbol{H})$.*
3. *The mapping A is closed on $\mathcal{F} \setminus \mathcal{S}$.*

The global convergence of iterative algorithms for solving (2) and the closedness of point-to-set mappings from $\mathcal{F}_\epsilon$ to its subsets can be defined in the same way as above. Also, if we replace $\mathcal{F}$ and $\mathcal{S}$ in Theorem 1 with $\mathcal{F}_\epsilon$ and $\mathcal{S}_\epsilon$, respectively, we obtain a theorem that gives a sufficient condition for algorithms for solving (2) to be globally convergent to $\mathcal{S}_\epsilon$.

Zangwill's global convergence theorem is well known as a powerful framework for proving the global convergence of iterative algorithms. For example, it was used in proving the global convergence of the concave-convex procedure [45], the decomposition method for support vector machines [48], and the modified MUR for NMF [47].

3 HALS algorithm

In this section, we review the HALS algorithm [6] for solving the optimization problem (1) and some of its variants. We also review their convergence property.

Let the k-th columns of $\boldsymbol{W}$ and $\boldsymbol{H}$ be denoted by $\boldsymbol{w}_k$ and $\boldsymbol{h}_k$, respectively. Then the problem (1) is rewritten as follows:

$$\begin{aligned} &\text{minimize} \ \frac{1}{2}\left\|\boldsymbol{X}-\sum_{k=1}^{K}\boldsymbol{w}_k\boldsymbol{h}_k^{\mathrm{T}}\right\|_{\mathrm{F}}^2 \\ &\text{subject to} \ \boldsymbol{w}_k \geq \boldsymbol{0}_{M\times 1}, \quad \boldsymbol{h}_k \geq \boldsymbol{0}_{N\times 1}, \quad k=1,2,\ldots,K. \end{aligned} \tag{5}$$

The HALS algorithm, which can be viewed as a special case of the block coordinate descent (BCD) method [27], updates $2K$ column vectors $\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_K$ and $\boldsymbol{h}_1, \boldsymbol{h}_2, \ldots, \boldsymbol{h}_K$ one by one in a fixed order so that the value of the objective function of (5) decreases monotonically. When updating $\boldsymbol{w}_k$, the HALS algorithm considers all other variables as constants and solves the following subproblem:

$$\begin{aligned} &\text{minimize} \ p_k(\boldsymbol{w}_k)=\frac{1}{2}\left\|\boldsymbol{R}_k^{\mathrm{T}}-\boldsymbol{h}_k\boldsymbol{w}_k^{\mathrm{T}}\right\|_{\mathrm{F}}^2 \\ &\text{subject to} \ \boldsymbol{w}_k \geq \boldsymbol{0}_{M\times 1} \end{aligned} \tag{6}$$

where

$$\boldsymbol{R}_k=\boldsymbol{X}-\sum_{\tilde{k}=1,\tilde{k}\neq k}^{K}\boldsymbol{w}_{\tilde{k}}\boldsymbol{h}_{\tilde{k}}^{\mathrm{T}}.$$

If $\boldsymbol{h}_k \neq \boldsymbol{0}_{N\times 1}$, the objective function $p_k(\boldsymbol{w}_k)$ is strictly convex and minimized at $\boldsymbol{w}_k=\boldsymbol{R}_k\boldsymbol{h}_k/\|\boldsymbol{h}_k\|_2^2$. Hence the subproblem (6) has the unique optimal solution $\boldsymbol{w}_k=\left[\boldsymbol{R}_k\boldsymbol{h}_k/\|\boldsymbol{h}_k\|_2^2\right]_+$ [27, Theorem 2]. Similarly, when updating $\boldsymbol{h}_k$, the HALS algorithm considers all other variables as constants and solves the following subproblem:

$$\begin{aligned} &\text{minimize} \ q_k(\boldsymbol{h}_k)=\frac{1}{2}\left\|\boldsymbol{R}_k-\boldsymbol{w}_k\boldsymbol{h}_k^{\mathrm{T}}\right\|_{\mathrm{F}}^2 \\ &\text{subject to} \ \boldsymbol{h}_k \geq \boldsymbol{0}_{N\times 1}. \end{aligned} \tag{7}$$

Taking into account the correspondence between variables and constants in (6) and those in (7), we can say that the subproblem (7) has the unique optimal solution $\boldsymbol{h}_k=\left[\boldsymbol{R}_k^{\mathrm{T}}\boldsymbol{w}_k/\|\boldsymbol{w}_k\|_2^2\right]_+$ if $\boldsymbol{w}_k \neq \boldsymbol{0}_{M\times 1}$. Based on these analyses, the update rule described by

$$\boldsymbol{w}_k \leftarrow \left[\frac{\boldsymbol{R}_k\boldsymbol{h}_k}{\|\boldsymbol{h}_k\|_2^2}\right]_+, \tag{8}$$

$$\boldsymbol{h}_k \leftarrow \left[\frac{\boldsymbol{R}_k^{\mathrm{T}}\boldsymbol{w}_k}{\|\boldsymbol{w}_k\|_2^2}\right]_+ \tag{9}$$

is obtained [7, 23, 27]. In this paper, we call the algorithm based on this update rule the HALS algorithm [7] though it is also called the rank-one residue iteration algorithm [23].

For the HALS algorithm, the following result is known.

Theorem 2 *(Kim et al. [27]) If the columns of $\boldsymbol{W}$ and $\boldsymbol{H}$ remain nonzero throughout the iterations, every limit point of the sequence $\left\{(\boldsymbol{W}^{(t)},\boldsymbol{H}^{(t)})\right\}_{t=0}^{\infty}$ generated by the HALS algorithm belongs to $\mathcal{S}$.*

Note that the global convergence of the HALS algorithm is not guaranteed by this theorem. There are two issues to consider. First, the assumption that the columns of $\boldsymbol{W}$ and $\boldsymbol{H}$ remain nonzero throughout the iterations may not always be valid. Once $\boldsymbol{w}_k$ becomes zero for example, $\boldsymbol{h}_k$ cannot be updated because the right-hand side of (9) becomes an indeterminate form. Second, even though the assumption is valid, it may occur that the sequence generated by the HALS algorithm has no limit point.

A simple way to avoid indeterminate forms is to use

$$\boldsymbol{w}_k \leftarrow \left[\frac{\boldsymbol{R}_k \boldsymbol{h}_k}{\|\boldsymbol{h}_k\|_2^2}\right]_{\epsilon+}, \tag{10}$$

$$\boldsymbol{h}_k \leftarrow \left[\frac{\boldsymbol{R}_k^{\mathrm{T}} \boldsymbol{w}_k}{\|\boldsymbol{w}_k\|_2^2}\right]_{\epsilon+} \tag{11}$$

instead of (8) and (9), where ϵ is a small positive constant. This update rule was introduced by Cichocki *et al*. [6] to avoid the numerical instability, but later proved to be globally convergent as shown in the following theorem.

Theorem 3 *(Kimura and Takahashi [29]) The HALS algorithm using the update rule described by (10) and (11) is globally convergent to* $\mathcal{S}_\epsilon$.

Note that the update rule described by (10) and (11) does not perform NMF but positive matrix factorization [39]. In addition, the limit of any convergent subsequence is not a stationary point of (1) but one of (2) as shown in Theorem 3. Hence this update rule produces only dense factor matrices. One may claim that sparse factor matrices will be obtained if we replace all ϵ in the factor matrices with zeros and that the pair of the resulting sparse factor matrices will be close to $\mathcal{S}$. However, it is not clear whether this claim always holds true or not.

Another simple way to avoid indeterminate forms is to use

$$\boldsymbol{w}_k \leftarrow \frac{[\boldsymbol{R}_k \boldsymbol{h}_k + \delta \boldsymbol{w}_k]_+}{\|\boldsymbol{h}_k\|_2^2 + \delta}, \tag{12}$$

$$\boldsymbol{h}_k \leftarrow \frac{\left[\boldsymbol{R}_k^{\mathrm{T}} \boldsymbol{w}_k + \delta \boldsymbol{h}_k\right]_+}{\|\boldsymbol{w}_k\|_2^2 + \delta} \tag{13}$$

instead of (8) and (9), where δ is a positive constant. This update rule is derived from auxiliary functions of $p_k(\boldsymbol{w}_k)$ and $q_k(\boldsymbol{h}_k)$ [15]. Details will be shown in the proof of Lemma 2. For this update rule, the following result is known.

Theorem 4 *(Gillis [15]) Every limit point of the sequence* $\left\{(\boldsymbol{W}^{(t)}, \boldsymbol{H}^{(t)})\right\}_{t=0}^{\infty}$ *generated by the HALS algorithm using the update rule described by (12) and (13) belongs to* $\mathcal{S}$.

Just like Theorem 2 for the original HALS algorithm, Theorem 4 says nothing about the global convergence of the update rule described by (12) and (13) to $\mathcal{S}$. The existence of a limit point is not guaranteed even though the objective function value decreases monotonically along the sequence $\left\{(\boldsymbol{W}^{(t)}, \boldsymbol{H}^{(t)})\right\}_{t=0}^{\infty}$ generated by the update rule, because the level set of the objective function $f(\boldsymbol{W}, \boldsymbol{H})$ is unbounded.

4 New update rule and its global convergence

In this section, we propose a new update rule of the HALS algorithm and prove that it is globally convergent to $\mathcal{S}$.

4.1 Proposed update rule

The update rule we propose in this paper is described by

$$\boldsymbol{w}_k \leftarrow \frac{[\boldsymbol{R}_k \boldsymbol{h}_k + \delta \boldsymbol{w}_k]_+}{\|\boldsymbol{h}_k\|_2^2 + \delta}, \tag{14}$$

$$\boldsymbol{w}_k \leftarrow \begin{cases} \boldsymbol{w}_k / \|\boldsymbol{w}_k\|_2, & \text{if } \boldsymbol{w}_k \neq \boldsymbol{0}_{M\times 1}, \\ \boldsymbol{u}_k, & \text{otherwise}, \end{cases} \tag{15}$$

$$\boldsymbol{h}_k \leftarrow \left[\boldsymbol{R}_k^{\mathrm{T}} \boldsymbol{w}_k\right]_+ \tag{16}$$

where δ is a positive constant and $\boldsymbol{u}_k$ is an arbitrary nonnegative unit vector. It is clear that division by zero never occurs in the proposed update rule. The first formula (14) is the same as (12). The second formula (15) is the normalization procedure for $\boldsymbol{w}_k$. The third formula (16) is used instead of (9) because $\|\boldsymbol{w}_k\|_2^2 = 1$ always holds when $\boldsymbol{h}_k$ is updated. The normalization procedure plays an important role when we prove that any sequence generated by the proposed update rule is contained in a compact subset of $\mathcal{F}$.

In this paper, we focus our attention on the case where the columns of $\boldsymbol{W}$ are normalized, but the alternative case where the columns of $\boldsymbol{H}$ are normalized can be dealt with in the same way. Here we should note that the modified MUR [35] also uses a normalization procedure, but this is slightly different from ours. It uses $\boldsymbol{0}_{M\times 1}$ instead of $\boldsymbol{u}_k$ in (15).

A formal statement of the proposed update rule is presented in Algorithm 1. Note that Step 4 is added to facilitate the global convergence analysis, though it is not necessary for practical purpose. Note also that Steps 2 and 3 can be replaced with

$$\boldsymbol{w}_k \leftarrow \left[\boldsymbol{w}_k + \frac{\left(\boldsymbol{X} - \boldsymbol{W}\boldsymbol{H}^{\mathrm{T}}\right)\boldsymbol{h}_k}{\|\boldsymbol{h}_k\|_2^2 + \delta}\right]_+$$

and Step 6 can be replaced with

$$\boldsymbol{h}_k \leftarrow \left[\boldsymbol{h}_k + \left(\boldsymbol{X} - \boldsymbol{W}\boldsymbol{H}^{\mathrm{T}}\right)^{\mathrm{T}} \boldsymbol{w}_k\right]_+$$

for an efficient implementation (see Cichocki and Fan [5] for more details) . It is easy to see that the proposed update rule has the same computational complexity per iteration as the original update rule. The following theorem establishes the global convergence of the proposed update rule.

Theorem 5 *The HALS algorithm using the update rule shown in Algorithm 1 is globally convergent to* $\mathcal{S}$.

This theorem can be proved by using Theorem 1. Details are shown in the next subsection.

4.2 Proof of Theorem 5

We prove Theorem 5 by using Theorem 1. Let the point-to-set mapping representing Algorithm 1 be denoted by A. Also, let the point-to-set mappings corresponding to Steps 3, 4, 5

Algorithm 1 Proposed Update Rule for NMF

Input: $\boldsymbol{X} \in \mathbb{R}_+^{M\times N}$, $(\boldsymbol{W}, \boldsymbol{H}) \in \mathcal{F}$, $\delta \in \mathbb{R}_{++}$
Output: Updated $(\boldsymbol{W}, \boldsymbol{H}) \in \mathcal{F}$
1: Set $k \leftarrow 1$.
2: Set $\boldsymbol{R}_k \leftarrow \boldsymbol{X} - \sum_{\tilde{k}=1,\tilde{k}\neq k}^{K} \boldsymbol{w}_{\tilde{k}}\boldsymbol{h}_{\tilde{k}}^{\mathrm{T}}$.
3: Set $\boldsymbol{w}_k \leftarrow [\boldsymbol{R}_k\boldsymbol{h}_k + \delta\boldsymbol{w}_k]_+/(\|\boldsymbol{h}_k\|_2^2 + \delta)$.
4: Set $\boldsymbol{h}_k \leftarrow \boldsymbol{h}_k\|\boldsymbol{w}_k\|_2$. (This step is not necessary for practical purpose.)
5: If $\boldsymbol{w}_k \neq \boldsymbol{0}_{M\times 1}$ then set $\boldsymbol{w}_k \leftarrow \boldsymbol{w}_k/\|\boldsymbol{w}_k\|_2$. Otherwise set $\boldsymbol{w}_k \leftarrow \boldsymbol{u}_k$ where $\boldsymbol{u}_k$ is an arbitrary nonnegative unit vector.
6: Set $\boldsymbol{h}_k \leftarrow [\boldsymbol{R}_k^{\mathrm{T}}\boldsymbol{w}_k]_+$.
7: If $k = K$ return $(\boldsymbol{W}, \boldsymbol{H})$ and stop. Otherwise set $k \leftarrow k + 1$ and go to Step 2.

and 6 of Algorithm 1 be denoted by D_k^{W}, S_k^{H}, S_k^{W} and D_k^{H}, respectively. Then A is expressed as

$$A = D_K^{\mathrm{H}} \circ S_K^{\mathrm{W}} \circ S_K^{\mathrm{H}} \circ D_K^{\mathrm{W}} \circ \cdots \circ D_1^{\mathrm{H}} \circ S_1^{\mathrm{W}} \circ S_1^{\mathrm{H}} \circ D_1^{\mathrm{W}}$$

where $\circ$ denotes the composition of mappings. The mappings D_k^{W}, S_k^{H} and D_k^{H} are given by

$$\begin{aligned}
D_k^{\mathrm{W}}(\boldsymbol{W}, \boldsymbol{H}) = \{(\boldsymbol{U}, \boldsymbol{V}) \in \mathcal{F} \mid \boldsymbol{u}_k &= [\boldsymbol{R}_k\boldsymbol{h}_k + \delta\boldsymbol{w}_k]_+/(\|\boldsymbol{h}_k\|_2^2 + \delta),\\
\boldsymbol{u}_{\tilde{k}} &= \boldsymbol{w}_{\tilde{k}} \text{ for all } \tilde{k} \neq k, \boldsymbol{V} = \boldsymbol{H}\},\\
S_k^{\mathrm{H}}(\boldsymbol{W}, \boldsymbol{H}) = \{(\boldsymbol{U}, \boldsymbol{V}) \in \mathcal{F} \mid \boldsymbol{U} &= \boldsymbol{W}, \boldsymbol{v}_k = \boldsymbol{h}_k\|\boldsymbol{w}_k\|_2, \boldsymbol{v}_{\tilde{k}} = \boldsymbol{h}_{\tilde{k}} \text{ for all } \tilde{k} \neq k\},\\
D_k^{\mathrm{H}}(\boldsymbol{W}, \boldsymbol{H}) = \{(\boldsymbol{U}, \boldsymbol{V}) \in \mathcal{F} \mid \boldsymbol{U} &= \boldsymbol{W}, \boldsymbol{v}_k = [\boldsymbol{R}_k\boldsymbol{w}_k]_+, \boldsymbol{v}_{\tilde{k}} = \boldsymbol{h}_{\tilde{k}} \text{ for all } \tilde{k} \neq k\},
\end{aligned}$$

and the mapping $S_k^{\mathrm{W}}(\boldsymbol{W}, \boldsymbol{H})$ is given by

$$S_k^{\mathrm{W}}(\boldsymbol{W}, \boldsymbol{H}) = \{(\boldsymbol{U}, \boldsymbol{V}) \in \mathcal{F} \mid \boldsymbol{u}_k = \boldsymbol{w}_k/\|\boldsymbol{w}_k\|_2, \boldsymbol{u}_{\tilde{k}} = \boldsymbol{w}_{\tilde{k}} \text{ for all } \tilde{k} \neq k, \boldsymbol{V} = \boldsymbol{H}\}$$

if $\boldsymbol{w}_k \neq \boldsymbol{0}_{M\times 1}$, and

$$S_k^{\mathrm{W}}(\boldsymbol{W}, \boldsymbol{H}) = \{(\boldsymbol{U}, \boldsymbol{V}) \in \mathcal{F} \mid \|\boldsymbol{u}_k\|_2 = 1, \boldsymbol{u}_{\tilde{k}} = \boldsymbol{w}_{\tilde{k}} \text{ for all } \tilde{k} \neq k, \boldsymbol{V} = \boldsymbol{H}\},$$

otherwise. Note that the set $D_k^{\mathrm{W}}(\boldsymbol{W}, \boldsymbol{H})$ consists of only one point in $\mathcal{F}$, which is represented as a continuous function of $(\boldsymbol{W}, \boldsymbol{H})$. The same can be said for $S_k^{\mathrm{H}}(\boldsymbol{W}, \boldsymbol{H})$ and $D_k^{\mathrm{H}}(\boldsymbol{W}, \boldsymbol{H})$.

We now prove that the proposed update rule satisfies the second condition in Theorem 1. Let us begin with the definition and an important property of the auxiliary function [33] because it plays an important role in our proof.

Definition 3 (Auxiliary Function [33]) For a function $g : \mathbb{R}_+ \to \mathbb{R}$, a two-variable function $\bar{g} : \mathbb{R}_+ \times \mathbb{R}_+ \to \mathbb{R}$ is called an auxiliary function of g if the following conditions hold:

1. $\bar{g}(x, x) = g(x)$ for all $x \in \mathbb{R}_+$,
2. $\bar{g}(x, y) \geq g(x)$ for all $x, y \in \mathbb{R}_+$.

Lemma 1 *Let $\bar{g} : \mathbb{R}_+ \times \mathbb{R}_+ \to \mathbb{R}$ be an auxiliary function of $g : \mathbb{R}_+ \to \mathbb{R}$. If the inequality $\bar{g}(a, b) \leq \bar{g}(b, b)$ holds for nonnegative numbers a and b then $g(a) \leq g(b)$. In particular, if $\bar{g}(a, b) < \bar{g}(b, b)$ then $g(a) < g(b)$.*

Proof If $\bar{g}(a, b) \leq \bar{g}(b, b)$, we have

$$g(a) \leq \bar{g}(a, b) \leq \bar{g}(b, b) = g(b). \tag{17}$$

The first inequality follows from the second condition in Definition 3 and the equality follows from the first condition in Definition 3. If $\bar{g}(a, b)$ is strictly less than $\bar{g}(b, b)$, it is clear from (17) that $g(a) < g(b)$. □

Using Lemma 1, we obtain the following three lemmas.

Lemma 2 *The objective function $f(\boldsymbol{W}, \boldsymbol{H})$ is nonincreasing under the proposed update rule shown in Algorithm 1.*

Proof The objective function is nonincreasing under the composite mapping $S_k^{\mathrm{W}} \circ S_k^{\mathrm{H}}$ for all k because the value of $\boldsymbol{w}_k \boldsymbol{h}_k^{\mathrm{T}}$ does not change before and after the composite mapping is performed. Also, the objective function is nonincreasing under D_k^{H} for all k because $\boldsymbol{h}_k = [\boldsymbol{R}_k^{\mathrm{T}} \boldsymbol{w}_k]_+$ is the unique optimal solution of (7) when $\|\boldsymbol{w}_k\|_2 = 1$. So it suffices for us to show that the objective function is nonincreasing under D_k^{W} for all k.

When the mapping D_k^{W} is performed, only $\boldsymbol{w}_k = [w_{1k}, w_{2k}, \ldots, w_{Mk}]^{\mathrm{T}}$ is updated. We thus consider all variables other than $\boldsymbol{w}_k$ as constants, and show that the value of $p_k(\boldsymbol{w}_k)$, the objective function of (6), does not increase. Note that $p_k(\boldsymbol{w}_k)$ is rewritten as

$$p_k(\boldsymbol{w}_k) = \sum_{m=1}^{M} p_{mk}(w_{mk})$$

where

$$\begin{aligned} p_{mk}(x) &= \frac{1}{2} \left\| (\boldsymbol{r}_m^{\mathrm{r}})^{\mathrm{T}} - \boldsymbol{h}_k x \right\|_2^2 \\ &= \frac{1}{2} \|\boldsymbol{h}_k\|_2^2 x^2 - \boldsymbol{r}_m^{\mathrm{r}} \boldsymbol{h}_k x + \frac{1}{2} \|\boldsymbol{r}_m^{\mathrm{r}}\|_2^2 \end{aligned} \tag{18}$$

and $\boldsymbol{r}_m^{\mathrm{r}}$ is the m-th row of $\boldsymbol{R}_k$. For the function $p_{mk}(x)$, we define a two-variable function $\bar{p}_{mk}(x, y)$ as follows:

$$\bar{p}_{mk}(x, y) = p_{mk}(x) + \frac{\delta}{2}(x - y)^2 \tag{19}$$

where δ is a positive constant used in Algorithm 1. It is clear that $\bar{p}_{mk}(x, y)$ is an auxiliary function of $p_{mk}(x)$ and strongly convex in both x and y (but not jointly) [15, 42]. For each value of y, the minimum point x^* of $\bar{p}_{mk}(x, y)$ in $\mathbb{R}_+$ is uniquely determined as

$$x^* = \frac{\left[\boldsymbol{r}_m^{\mathrm{r}} \boldsymbol{h}_k + \delta y\right]_+}{\|\boldsymbol{h}_k\|_2^2 + \delta}. \tag{20}$$

Therefore, by Lemma 1, we have

$$p_{mk}(x^*) \leq p_{mk}(y).$$

Substituting $y = w_{mk}$ into this inequality, we have

$$p_{mk}\left(\frac{\left[\boldsymbol{r}_m^{\mathrm{r}} \boldsymbol{h}_k + \delta w_{mk}\right]_+}{\|\boldsymbol{h}_k\|_2^2 + \delta} \right) \leq p_{mk}(w_{mk})$$

from which we have

$$p_k\left(\frac{[\boldsymbol{R}_k \boldsymbol{h}_k + \delta \boldsymbol{w}_k]_+}{\|\boldsymbol{h}_k\|_2^2 + \delta} \right) = \sum_{m=1}^{M} p_{mk}\left(\frac{\left[\boldsymbol{r}_m^{\mathrm{r}} \boldsymbol{h}_k + \delta w_{mk}\right]_+}{\|\boldsymbol{h}_k\|_2^2 + \delta} \right)$$

$$\leq \sum_{m=1}^{M} p_{mk}(w_{mk})$$
$$= p_k(\boldsymbol{w}_k).$$

This means that $f(\boldsymbol{W}, \boldsymbol{H})$ is nonincreasing under D_k^{W}. □

Lemma 3 *A point $(\boldsymbol{W}^*, \boldsymbol{H}^*)$ is a stationary point of (1) if and only if $\boldsymbol{w}_k$ is a stationary point of (6) with $\boldsymbol{h}_k = \boldsymbol{h}_k^*$ for $k = 1, 2, \ldots, K$ and $\boldsymbol{h}_k$ is a stationary point of (7) with $\boldsymbol{w}_k = \boldsymbol{w}_k^*$ for $k = 1, 2, \ldots, K$.*

Proof We omit the proof because it is similar to [29, Lemma 3]. □

Lemma 4 *For any $(\boldsymbol{W}, \boldsymbol{H}) \in \mathcal{F}$, the following statements hold true.*

1. *If $(\boldsymbol{W}, \boldsymbol{H}) \notin \mathcal{S}$ then $f(\boldsymbol{U}, \boldsymbol{V}) < f(\boldsymbol{W}, \boldsymbol{H})$ for all $(\boldsymbol{U}, \boldsymbol{V}) \in A(\boldsymbol{W}, \boldsymbol{H})$.*
2. *If $(\boldsymbol{W}, \boldsymbol{H}) \in \mathcal{S}$ then $f(\boldsymbol{U}, \boldsymbol{V}) \leq f(\boldsymbol{W}, \boldsymbol{H})$ for all $(\boldsymbol{U}, \boldsymbol{V}) \in A(\boldsymbol{W}, \boldsymbol{H})$.*

Proof It is clear from Lemma 2 that the second statement holds true. Thus we only have to consider the first statement. Let $(\boldsymbol{W}, \boldsymbol{H})$ be any point in $\mathcal{F} \setminus \mathcal{S}$. It follows from Lemma 3 that there exists at least one k such that i) $\boldsymbol{w}_k$ is not a stationary point of (6) or ii) $\boldsymbol{h}_k$ is not a stationary point of (7).

In the first case, there exists at least one m such that $p'_{mk}(w_{mk}) < 0$ if $w_{mk} = 0$ and $p'_{mk}(w_{mk}) \neq 0$ if $w_{mk} > 0$, where $p_{mk}(x)$ is given by (18). For such an m, the auxiliary function $\bar{p}_{mk}(x, y)$ of $p_{mk}(x)$, which is given by (19), satisfies

$$\frac{\partial \bar{p}_{mk}}{\partial x}(w_{mk}, w_{mk}) = p'_{mk}(w_{mk}) + \delta(w_{mk} - w_{mk}) = p'_{mk}(w_{mk})$$

which is negative if $w_{mk} = 0$ and nonzero if $w_{mk} > 0$. This means that $x = w_{mk}$ is not the unique minimum point of $\bar{p}_{mk}(x, w_{mk})$. Hence $\bar{p}_{mk}(x^*, w_{mk}) < \bar{p}_{mk}(w_{mk}, w_{mk})$ where x^* is the unique minimum point given by (20). From this inequality and Lemma 1, we have $p_{mk}(x^*) < p_{mk}(w_{mk})$ which implies that

$$p_k\left(\frac{[\boldsymbol{R}_k \boldsymbol{h}_k + \delta \boldsymbol{w}_k]_+}{\|\boldsymbol{h}_k\|_2^2 + \delta}\right) < p_k(\boldsymbol{w}_k).$$

Therefore, $f(\boldsymbol{W}, \boldsymbol{H})$ strictly decreases under the mapping A.

In the second case, we can show in the same way as above that $f(\boldsymbol{W}, \boldsymbol{H})$ strictly decreases under the mapping A. □

We next prove that the proposed update rule satisfies the first condition in Theorem 1. To do so, for any point $(\boldsymbol{W}^{(0)}, \boldsymbol{H}^{(0)})$ in $\mathcal{F}$, we define the set $\mathcal{L}_{(\boldsymbol{W}^{(0)}, \boldsymbol{H}^{(0)})}$ as follows:

$$\mathcal{L}_{(\boldsymbol{W}^{(0)}, \boldsymbol{H}^{(0)})} = \{(\boldsymbol{W}, \boldsymbol{H}) \in \mathcal{F} \mid f(\boldsymbol{W}, \boldsymbol{H}) \leq f(\boldsymbol{W}^{(0)}, \boldsymbol{H}^{(0)}), \|\boldsymbol{w}_k\|_2 = 1 \text{ for all } k\}.$$

Note that this is not a level set of f because of the conditions that $\|\boldsymbol{w}_k\|_2 = 1$ for all k. The next lemma shows the boundedness of this set.

Lemma 5 *The set $\mathcal{L}_{(\boldsymbol{W}^{(0)}, \boldsymbol{H}^{(0)})}$ is bounded for any $(\boldsymbol{W}^{(0)}, \boldsymbol{H}^{(0)}) \in \mathcal{F}$.*

Proof Let $(\boldsymbol{W}, \boldsymbol{H})$ be any point in $\mathcal{L}_{(\boldsymbol{W}^{(0)}, \boldsymbol{H}^{(0)})}$. It suffices for us to show that $\|\boldsymbol{h}_k\|_2$ is bounded for $k = 1, 2, \ldots, K$. Because $q_k(\boldsymbol{h}_k)$ is convex, the inequality

$$q_k(\boldsymbol{h}_k) \geq q_k(\boldsymbol{v}) + \nabla q_k(\boldsymbol{v}_k)^{\mathrm{T}}(\boldsymbol{h}_k - \boldsymbol{v})$$

$$= q_k(\boldsymbol{v}) - \left(\boldsymbol{R}_k^{\mathrm{T}}\boldsymbol{w}_k - \|\boldsymbol{w}_k\|_2^2\boldsymbol{v}\right)^{\mathrm{T}}(\boldsymbol{h}_k - \boldsymbol{v})$$
$$= q_k(\boldsymbol{v}) - \left(\boldsymbol{R}_k^{\mathrm{T}}\boldsymbol{w}_k - \boldsymbol{v}\right)^{\mathrm{T}}(\boldsymbol{h}_k - \boldsymbol{v})$$

holds for any $\boldsymbol{v} \in \mathbb{R}^N$ [3]. Substituting $\boldsymbol{v} = \boldsymbol{R}_k^{\mathrm{T}}\boldsymbol{w}_k + \mathbf{1}_{N\times 1}$, we have

$$q_k(\boldsymbol{h}_k) \geq \frac{1}{2}\|\boldsymbol{R}_k - \boldsymbol{w}_k(\boldsymbol{R}_k^{\mathrm{T}}\boldsymbol{w}_k + \mathbf{1}_{N\times 1})^{\mathrm{T}}\|_{\mathrm{F}}^2 + \mathbf{1}_{N\times 1}^{\mathrm{T}}(\boldsymbol{h}_k - \boldsymbol{R}_k^{\mathrm{T}}\boldsymbol{w}_k - \mathbf{1}_{N\times 1}).$$

Hence, the inequality $q_k(\boldsymbol{h}_k) \leq f(\boldsymbol{W}^{(0)}, \boldsymbol{H}^{(0)})$ implies that

$$\mathbf{1}_{N\times 1}^{\mathrm{T}}(\boldsymbol{h}_k - \boldsymbol{R}_k^{\mathrm{T}}\boldsymbol{w}_k - \mathbf{1}_{N\times 1}) \leq f(\boldsymbol{W}^{(0)}, \boldsymbol{H}^{(0)})$$

from which we have

$$\begin{aligned}\|\boldsymbol{h}_k\|_2 &\leq \|\boldsymbol{h}_k\|_1 \\ &\leq f(\boldsymbol{W}^{(0)}, \boldsymbol{H}^{(0)}) + \mathbf{1}_{N\times 1}^{\mathrm{T}}\boldsymbol{R}_k^{\mathrm{T}}\boldsymbol{w}_k + N \\ &\leq f(\boldsymbol{W}^{(0)}, \boldsymbol{H}^{(0)}) + \mathbf{1}_{N\times 1}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\mathbf{1}_{M\times 1} + N.\end{aligned}$$

This completes the proof. □

Using Lemma 5, we obtain the following lemma.

Lemma 6 *Any sequence* $\left\{(\boldsymbol{W}^{(t)}, \boldsymbol{H}^{(t)})\right\}_{t=0}^{\infty}$ *generated by Algorithm 1 is contained in a compact subset of* $\mathcal{F}$.

Proof We easily see from Step 5 of Algorithm 1 that $\|\boldsymbol{w}_k^{(t)}\|_2 = 1$ for all k and $t \in \mathbb{Z}_{++}$, where $\boldsymbol{w}_k^{(t)}$ is the k-th column of $\boldsymbol{W}^{(t)}$. Also, it follows from Lemma 2 that $f(\boldsymbol{W}^{(t)}, \boldsymbol{H}^{(t)}) \leq f(\boldsymbol{W}^{(0)}, \boldsymbol{H}^{(0)})$ for all $t \in \mathbb{Z}_+$. Therefore $(\boldsymbol{W}^{(t)}, \boldsymbol{H}^{(t)}) \in \mathcal{L}_{(\boldsymbol{W}^{(0)}, \boldsymbol{H}^{(0)})}$ for all $t \in \mathbb{Z}_{++}$. Because $\mathcal{L}_{(\boldsymbol{W}^{(0)}, \boldsymbol{H}^{(0)})}$ is bounded as shown in Lemma 5, the sequence $\{(\boldsymbol{W}^{(t)}, \boldsymbol{H}^{(t)})\}_{t=0}^{\infty}$ is contained in a compact subset of $\mathcal{F}$. □

We finally prove that the proposed update rule satisfies the third condition in Theorem 1. The next lemma shows the closedness of the point-to-set mappings $S_1^{\mathrm{W}}, S_2^{\mathrm{W}}, \ldots, S_K^{\mathrm{W}}$.

Lemma 7 *The point-to-set mappings* $S_1^{\mathrm{W}}, S_2^{\mathrm{W}}, \ldots, S_K^{\mathrm{W}}$ *are closed on* $\mathcal{F}$.

Proof Let $\{(\boldsymbol{W}^{(t)}, \boldsymbol{H}^{(t)})\}_{t=0}^{\infty}$ and $\{(\boldsymbol{U}^{(t)}, \boldsymbol{V}^{(t)})\}_{t=0}^{\infty}$ be any two convergent sequences in $\mathcal{F}$ that satisfy $(\boldsymbol{U}^{(t)}, \boldsymbol{V}^{(t)}) \in S_k^{\mathrm{W}}(\boldsymbol{W}^{(t)}, \boldsymbol{H}^{(t)})$ for all $t \in \mathbb{Z}_+$. Let $(\boldsymbol{W}^{(\infty)}, \boldsymbol{H}^{(\infty)})$ and $(\boldsymbol{U}^{(\infty)}, \boldsymbol{V}^{(\infty)})$ be the limits of these two sequences. It is clear from the definition of S_k^{W} that $\|\boldsymbol{u}_k^{(t)}\|_2 = 1$ for all $t \in \mathbb{Z}_+$, $\boldsymbol{u}_{\tilde{k}}^{(t)} = \boldsymbol{w}_{\tilde{k}}^{(t)}$ for all $\tilde{k} \neq k$ and $t \in \mathbb{Z}_+$, and $\boldsymbol{V}^{(t)} = \boldsymbol{H}^{(t)}$ for all $t \in \mathbb{Z}_+$. We first consider the case where $\boldsymbol{w}_k^{(\infty)} \neq \mathbf{0}_{M\times 1}$. In this case, $S_k^{\mathrm{W}}(\boldsymbol{W}^{(\infty)}, \boldsymbol{H}^{(\infty)})$ consists only of the point

$$\left(\left(\boldsymbol{w}_1^{(\infty)}, \ldots, \boldsymbol{w}_{k-1}^{(\infty)}, \frac{\boldsymbol{w}_k^{(\infty)}}{\|\boldsymbol{w}_k^{(\infty)}\|_2}, \boldsymbol{w}_{k+1}^{(\infty)}, \ldots, \boldsymbol{w}_K^{(\infty)}\right), \boldsymbol{H}^{(\infty)}\right)$$

and $\{(\boldsymbol{U}^{(t)}, \boldsymbol{V}^{(t)})\}_{t=0}^{\infty}$ converges to it. We next consider the case where $\boldsymbol{w}_k^{(\infty)} = \mathbf{0}_{M\times 1}$. In this case, $S_k^{\mathrm{W}}(\boldsymbol{W}^{(\infty)}, \boldsymbol{H}^{(\infty)})$ is the set of all $(\boldsymbol{W}, \boldsymbol{H}) \in \mathcal{F}$ such that $\|\boldsymbol{w}_k\|_2 = 1$, $\boldsymbol{w}_{\tilde{k}} = \boldsymbol{w}_{\tilde{k}}^{(\infty)}$ for all $\tilde{k} \neq k$ and $\boldsymbol{H} = \boldsymbol{H}^{(\infty)}$. Also, $(\boldsymbol{U}^{(\infty)}, \boldsymbol{V}^{(\infty)})$ satisfies $\|\boldsymbol{u}_k^{(\infty)}\|_2 = 1$, $\boldsymbol{u}_{\tilde{k}}^{(\infty)} = \boldsymbol{w}_{\tilde{k}}^{(\infty)}$ for all $\tilde{k} \neq k$ and $\boldsymbol{V}^{(\infty)} = \boldsymbol{H}^{(\infty)}$. Therefore, we have $(\boldsymbol{U}^{(\infty)}, \boldsymbol{V}^{(\infty)}) \in S_k^{\mathrm{W}}(\boldsymbol{W}^{(\infty)}, \boldsymbol{H}^{(\infty)})$. □

Given a point $(\boldsymbol{W}^{(0)}, \boldsymbol{H}^{(0)}) \in \mathcal{F}$, we define $\mathcal{L}^1_{(\boldsymbol{W}^{(0)},\boldsymbol{H}^{(0)})}$, $\mathcal{L}^2_{(\boldsymbol{W}^{(0)},\boldsymbol{H}^{(0)})}$ and $\mathcal{L}^3_{(\boldsymbol{W}^{(0)},\boldsymbol{H}^{(0)})}$ as follows:

$$\begin{aligned}
\mathcal{L}^1_{(\boldsymbol{W}^{(0)},\boldsymbol{H}^{(0)})} &= \{(\boldsymbol{W},\boldsymbol{H}) \in \mathcal{F} \mid f(\boldsymbol{W},\boldsymbol{H}) \le f(\boldsymbol{W}^{(0)},\boldsymbol{H}^{(0)}),\\
&\qquad \|\boldsymbol{w}_k\|_2 \le \mu_k \text{ and } \|\boldsymbol{h}_k\|_2 \le \nu_k \text{ for all } k\}\\
\mathcal{L}^2_{(\boldsymbol{W}^{(0)},\boldsymbol{H}^{(0)})} &= \{(\boldsymbol{W},\boldsymbol{H}) \in \mathcal{F} \mid f(\boldsymbol{W},\boldsymbol{H}) \le f(\boldsymbol{W}^{(0)},\boldsymbol{H}^{(0)}),\\
&\qquad \|\boldsymbol{w}_k\|_2 \le \mu_k + \sigma_{\max}(\boldsymbol{X})\nu_k/\delta \text{ and } \|\boldsymbol{h}_k\|_2 \le \nu_k \text{ for all } k\},\\
\mathcal{L}^3_{(\boldsymbol{W}^{(0)},\boldsymbol{H}^{(0)})} &= \{(\boldsymbol{W},\boldsymbol{H}) \in \mathcal{F} \mid f(\boldsymbol{W},\boldsymbol{H}) \le f(\boldsymbol{W}^{(0)},\boldsymbol{H}^{(0)}),\\
&\qquad \|\boldsymbol{w}_k\|_2 \le \mu_k + \sigma_{\max}(\boldsymbol{X})\nu_k/\delta \text{ and}\\
&\qquad \|\boldsymbol{h}_k\|_2 \le \nu_k(\mu_k + \sigma_{\max}(\boldsymbol{X})\nu_k/\delta) \text{ for all } k\}
\end{aligned}$$

where

$$\begin{aligned}
\mu_k &= \max\{1, \|\boldsymbol{w}_k^{(0)}\|_2\},\\
\nu_k &= \max\{f(\boldsymbol{W}^{(0)},\boldsymbol{H}^{(0)}) + \mathbf{1}^{\mathrm{T}}_{N\times 1}\boldsymbol{X}^{\mathrm{T}}\mathbf{1}_{M\times 1} + N, \|\boldsymbol{h}_k^{(0)}\|_2\}
\end{aligned}$$

for $k = 1, 2, \ldots, K$ and $\sigma_{\max}(\boldsymbol{X})$ is the largest singular value of $\boldsymbol{X}$. It is clear that all of the three sets defined above are compact subsets of $\mathcal{F}$. It is also clear that $(\boldsymbol{W}^{(0)}, \boldsymbol{H}^{(0)}) \in \mathcal{L}^1_{(\boldsymbol{W}^{(0)},\boldsymbol{H}^{(0)})}$. Furthermore, the following lemma holds.

Lemma 8 *The following statements are true for $k = 1, 2, \ldots, K$.*

1. *If $(\boldsymbol{W},\boldsymbol{H}) \in \mathcal{L}^1_{(\boldsymbol{W}^{(0)},\boldsymbol{H}^{(0)})}$ then $D^{\mathrm{W}}_k(\boldsymbol{W},\boldsymbol{H}) \subseteq \mathcal{L}^2_{(\boldsymbol{W}^{(0)},\boldsymbol{H}^{(0)})}$.*
2. *If $(\boldsymbol{W},\boldsymbol{H}) \in \mathcal{L}^2_{(\boldsymbol{W}^{(0)},\boldsymbol{H}^{(0)})}$ then $S^{\mathrm{H}}_k(\boldsymbol{W},\boldsymbol{H}) \subseteq \mathcal{L}^3_{(\boldsymbol{W}^{(0)},\boldsymbol{H}^{(0)})}$.*
3. *If $(\boldsymbol{W},\boldsymbol{H}) \in \mathcal{L}^3_{(\boldsymbol{W}^{(0)},\boldsymbol{H}^{(0)})}$ then $S^{\mathrm{W}}_k(\boldsymbol{W},\boldsymbol{H}) \subseteq \mathcal{L}_{(\boldsymbol{W}^{(0)},\boldsymbol{H}^{(0)})}$.*
4. *If $(\boldsymbol{W},\boldsymbol{H}) \in \mathcal{L}_{(\boldsymbol{W}^{(0)},\boldsymbol{H}^{(0)})}$ then $D^{\mathrm{H}}_k(\boldsymbol{W},\boldsymbol{H}) \subseteq \mathcal{L}^1_{(\boldsymbol{W}^{(0)},\boldsymbol{H}^{(0)})}$.*

Proof We first prove the first statement. Suppose that $(\boldsymbol{W},\boldsymbol{H}) \in \mathcal{L}^1_{(\boldsymbol{W}^{(0)},\boldsymbol{H}^{(0)})}$. Then $\|\boldsymbol{w}_k\|_2 \le \mu_k$ and $\|\boldsymbol{h}_k\|_2 \le \nu_k$ hold. Using these inequalities, we have

$$\begin{aligned}
\left\| \frac{[\boldsymbol{R}_k\boldsymbol{h}_k + \delta\boldsymbol{w}_k]_+}{\|\boldsymbol{h}_k\|_2^2 + \delta} \right\|_2 &\le \left\| \frac{\boldsymbol{X}\boldsymbol{h}_k + \delta\boldsymbol{w}_k}{\delta} \right\|_2\\
&\le \frac{1}{\delta}\|\boldsymbol{X}\boldsymbol{h}_k\|_2 + \|\boldsymbol{w}_k\|_2\\
&\le \frac{\sigma_{\max}(\boldsymbol{X})}{\delta}\|\boldsymbol{h}_k\|_2 + \|\boldsymbol{w}_k\|_2\\
&\le \frac{\sigma_{\max}(\boldsymbol{X})}{\delta}\nu_k + \mu_k
\end{aligned}$$

which means that $D^{\mathrm{W}}_k(\boldsymbol{W},\boldsymbol{H}) \subseteq \mathcal{L}^2_{(\boldsymbol{W}^{(0)},\boldsymbol{H}^{(0)})}$.

We next prove the second statement. Suppose that $(\boldsymbol{W},\boldsymbol{H}) \in \mathcal{L}^2_{(\boldsymbol{W}^{(0)},\boldsymbol{H}^{(0)})}$. Then $\|\boldsymbol{w}_k\|_2 \le (\sigma_{\max}(\boldsymbol{X})\nu_k/\delta + \mu_k)$ and $\|\boldsymbol{h}_k\|_2 \le \nu_k$ hold. Using these inequalities, we have

$$\|\boldsymbol{h}_k\|_2\|\boldsymbol{w}_k\|_2 \le \nu_k\left(\frac{\sigma_{\max}(\boldsymbol{X})}{\delta}\nu_k + \mu_k\right)$$

which means that $S^{\mathrm{H}}_k(\boldsymbol{W},\boldsymbol{H}) \subseteq \mathcal{L}^3_{(\boldsymbol{W}^{(0)},\boldsymbol{H}^{(0)})}$.

The third statement is clear from the definition of the point-to-set mapping S_k^{W}, and the fourth statement is clear from the proof of Lemma 5. □

From Lemma 8, we can restrict the domains of the point-to-set mappings D_k^{W}, S_k^{H}, S_k^{W}, D_k^{H} to $\mathcal{L}^1_{(\boldsymbol{W}^{(0)},\boldsymbol{H}^{(0)})}$, $\mathcal{L}^2_{(\boldsymbol{W}^{(0)},\boldsymbol{H}^{(0)})}$, $\mathcal{L}^3_{(\boldsymbol{W}^{(0)},\boldsymbol{H}^{(0)})}$ and $\mathcal{L}_{(\boldsymbol{W}^{(0)},\boldsymbol{H}^{(0)})}$, respectively. This means that we can restrict the domain of the point-to-set mapping A to $\mathcal{L}^1_{(\boldsymbol{W}^{(0)},\boldsymbol{H}^{(0)})}$. The next lemma shows the closedness of A restricted to $\mathcal{L}^1_{(\boldsymbol{W}^{(0)},\boldsymbol{H}^{(0)})}$.

Lemma 9 *For any $(\boldsymbol{W}^{(0)}, \boldsymbol{H}^{(0)}) \in \mathcal{F}$, the point-to-set mapping A restricted to $\mathcal{L}^1_{(\boldsymbol{W}^{(0)},\boldsymbol{H}^{(0)})}$ is closed on $\mathcal{L}^1_{(\boldsymbol{W}^{(0)},\boldsymbol{H}^{(0)})}$.*

Proof It is clear that the composite mapping $S_k^{\mathrm{H}} \circ D_k^{\mathrm{W}}$ from $\mathcal{L}^1_{(\boldsymbol{W}^{(0)},\boldsymbol{H}^{(0)})}$ to the subsets of $\mathcal{L}^3_{(\boldsymbol{W}^{(0)},\boldsymbol{H}^{(0)})}$ is closed on its domain for all k. Also, it follows from Lemma 7 and the continuity of D_k^{H} that the composite mapping $D_k^{\mathrm{H}} \circ S_k^{\mathrm{W}}$ from $\mathcal{L}^3_{(\boldsymbol{W}^{(0)},\boldsymbol{H}^{(0)})}$ to the subsets of $\mathcal{L}^1_{(\boldsymbol{W}^{(0)},\boldsymbol{H}^{(0)})}$ is closed on its domain for all k. Because $\mathcal{L}^3_{(\boldsymbol{W}^{(0)},\boldsymbol{H}^{(0)})}$ is a compact subset of $\mathcal{F}$, by [53, Corollary 4.2.1], the composite mapping $(D_k^{\mathrm{H}} \circ S_k^{\mathrm{W}}) \circ (S_k^{\mathrm{H}} \circ D_k^{\mathrm{W}})$ from $\mathcal{L}^1_{(\boldsymbol{W}^{(0)},\boldsymbol{H}^{(0)})}$ to the subsets of $\mathcal{L}^1_{(\boldsymbol{W}^{(0)},\boldsymbol{H}^{(0)})}$ is closed on its domain for all k. Furthermore, since $\mathcal{L}^1_{(\boldsymbol{W}^{(0)},\boldsymbol{H}^{(0)})}$ is a compact subset of $\mathcal{F}$, by [53, Corollary 4.2.1], we can conclude that A, which is a composition of the mappings $(D_k^{\mathrm{H}} \circ S_k^{\mathrm{W}}) \circ (S_k^{\mathrm{H}} \circ D_k^{\mathrm{W}})$, restricted to $\mathcal{L}^1_{(\boldsymbol{W}^{(0)},\boldsymbol{H}^{(0)})}$ is closed on its domain. □

We should note that even if $\boldsymbol{u}_k$ in Step 5 of Algorithm 1 is replaced with a constant nonnegative unit vector such as $(1/\sqrt{M})\mathbf{1}_{M\times 1}$ and $(1, 0, 0, \ldots, 0)^{\mathrm{T}}$ we can prove Theorem 5 without changing the definition of the mapping S_k^{W}.

5 Stopping conditions

We have proved that the HALS algorithm using the proposed update rule shown in Algorithm 1 is globally convergent to $\mathcal{S}$ in the sense of Definition 1. Therefore, combining this update rule with an appropriate stopping condition, we can design an algorithm that always stops in a finite number of iterations. In this section, we consider two approaches for deriving stopping conditions.

5.1 Relaxed KKT conditions

The first approach, which has already been used in the literature [29, 30, 38, 44, 46, 47], is to relax the KKT conditions (3) as follows:

$$\begin{cases} (\nabla_{\boldsymbol{W}} f(\boldsymbol{W}, \boldsymbol{H}))_{mk} \geq -\kappa_1, & \text{if } w_{mk} \leq \kappa_2, \\ \left|(\nabla_{\boldsymbol{W}} f(\boldsymbol{W}, \boldsymbol{H}))_{mk}\right| \leq \kappa_1, & \text{if } w_{mk} > \kappa_2, \end{cases} \quad m = 1, 2, \ldots, M,\ k = 1, 2, \ldots, K, \tag{21}$$

$$\begin{cases} (\nabla_{\boldsymbol{H}} f(\boldsymbol{W}, \boldsymbol{H}))_{nk} \geq -\kappa_1, & \text{if } h_{nk} \leq \kappa_2, \\ \left|(\nabla_{\boldsymbol{H}} f(\boldsymbol{W}, \boldsymbol{H}))_{nk}\right| \leq \kappa_1, & \text{if } h_{nk} > \kappa_2, \end{cases} \quad n = 1, 2, \ldots, N,\ k = 1, 2, \ldots, K \tag{22}$$

Algorithm 2 HALS Algorithm for NMF using Algorithm 1 and Relaxed KKT Conditions.

Input: $\boldsymbol{X} \in \mathbb{R}_+^{M \times N}$, $(\boldsymbol{W}^{(0)}, \boldsymbol{H}^{(0)}) \in \mathcal{F}$, $\delta, \kappa_1, \kappa_2 \in \mathbb{R}_{++}$
Output: $(\boldsymbol{W}, \boldsymbol{H}) \in \mathcal{F}$ satisfying (21) and (22)
1: Set $(\boldsymbol{W}, \boldsymbol{H}) \leftarrow (\boldsymbol{W}^{(0)}, \boldsymbol{H}^{(0)})$.
2: If (21) and (22) hold, return $(\boldsymbol{W}, \boldsymbol{H})$ and stop.
3: Update $(\boldsymbol{W}, \boldsymbol{H})$ using Algorithm 1, and go to Step 2.

where κ_1 and κ_2 are positive constants.

The HALS algorithm for NMF using Algorithm 1 and the stopping condition described by (21) and (22) is shown in Algorithm 2. For this algorithm, the following theorem holds. The proof is omitted because it is similar to that of Theorem 2 in [46].

Theorem 6 *Algorithm 2 stops in a finite number of iterations.*

5.2 Projected gradient norm

The second approach is to make use of the projected gradient [36]. To be more specific, the inequality

$$\psi_{\tau_2}(\boldsymbol{W}, \boldsymbol{H}) \leq \tau_1 \psi_{\tau_2}(\boldsymbol{W}^{(0)}, \boldsymbol{H}^{(0)}) \tag{23}$$

is used as the stopping condition, where τ_1 and τ_2 are positive constants, and $\psi_{\tau_2}(\boldsymbol{W}, \boldsymbol{H})$ is defined as

$$\psi_{\tau_2}(\boldsymbol{W}, \boldsymbol{H}) = \sqrt{\left\| \boldsymbol{G}^{\mathrm{W}}_{\tau_2}(\boldsymbol{W}, \boldsymbol{H}) \right\|_{\mathrm{F}}^2 + \left\| \boldsymbol{G}^{\mathrm{H}}_{\tau_2}(\boldsymbol{W}, \boldsymbol{H}) \right\|_{\mathrm{F}}^2}.$$

The notations $\boldsymbol{G}^{\mathrm{W}}_{\tau_2}(\boldsymbol{W}, \boldsymbol{H})$ and $\boldsymbol{G}^{\mathrm{H}}_{\tau_2}(\boldsymbol{W}, \boldsymbol{H})$ denote a modified projected gradients with respect to $\boldsymbol{W}$ and $\boldsymbol{H}$, respectively, which are defined by

$$(\boldsymbol{G}^{\mathrm{W}}_{\tau_2}(\boldsymbol{W}, \boldsymbol{H}))_{mk} = \begin{cases} \min\{0, (\nabla_{\boldsymbol{W}} f(\boldsymbol{W}, \boldsymbol{H}))_{mk}\}, & \text{if } w_{mk} \leq \tau_2, \\ (\nabla_{\boldsymbol{W}} f(\boldsymbol{W}, \boldsymbol{H}))_{mk}, & \text{if } w_{mk} > \tau_2 \end{cases}$$

and

$$(\boldsymbol{G}^{\mathrm{H}}_{\tau_2}(\boldsymbol{W}, \boldsymbol{H}))_{nk} = \begin{cases} \min\{0, (\nabla_{\boldsymbol{H}} f(\boldsymbol{W}, \boldsymbol{H}))_{nk}\}, & \text{if } h_{nk} \leq \tau_2, \\ (\nabla_{\boldsymbol{H}} f(\boldsymbol{W}, \boldsymbol{H}))_{nk}, & \text{if } h_{nk} > \tau_2. \end{cases}$$

Note that our definition of the projected gradient is slightly different from the one used in the literature [20, 26, 27, 36], which corresponds to the case where $\tau_2 = 0$. It is clear that if $(\boldsymbol{W}, \boldsymbol{H})$ is a stationary point of (1) then (23) is satisfied because $\psi_{\tau_2}(\boldsymbol{W}, \boldsymbol{H}) = 0$ holds. Therefore, (23) is considered as relaxed KKT conditions.

The proposed HALS algorithm for NMF using Algorithm 1 and the stopping condition (23) is shown in Algorithm 3. For this algorithm, the following theorem holds.

Theorem 7 *Algorithm 3 stops in a finite number of iterations.*

Proof The proof is done by contradiction. We first assume that Algorithm 3 does not stop for some $(\boldsymbol{W}^{(0)}, \boldsymbol{H}^{(0)}) \in \mathcal{F}$. Let $\{(\boldsymbol{W}^{(t)}, \boldsymbol{H}^{(t)})\}_{t=0}^{\infty}$ be an infinite sequence generated by Algorithm 3. Then, we see from Step 1 that $\psi_{\tau_2}(\boldsymbol{W}^{(0)}, \boldsymbol{H}^{(0)})$ must be positive. Also, by Theorem 5, this sequence has at least one subsequence that converges to a stationary point

of (1). Let $\{(\boldsymbol{W}^{(t_i)}, \boldsymbol{H}^{(t_i)})\}_{i=0}^{\infty}$ be one of such subsequences and $(\boldsymbol{W}^{(\infty)}, \boldsymbol{H}^{(\infty)}) \in \mathcal{F}$ be its limit. Because the limit is a stationary point of (1), it satisfies

$$\left(\nabla_{\boldsymbol{W}} f(\boldsymbol{W}^{(\infty)}, \boldsymbol{H}^{(\infty)})\right)_{mk} \begin{cases} \geq 0, & \text{if } w_{mk}^{(\infty)} = 0, \\ = 0, & \text{if } w_{mk}^{(\infty)} > 0, \end{cases}$$
$$m = 1, 2, \ldots, M, \ k = 1, 2, \ldots, K,$$
$$\left(\nabla_{\boldsymbol{H}} f(\boldsymbol{W}^{(\infty)}, \boldsymbol{H}^{(\infty)})\right)_{nk} \begin{cases} \geq 0, & \text{if } h_{nk}^{(\infty)} = 0, \\ = 0, & \text{if } h_{nk}^{(\infty)} > 0, \end{cases}$$
$$n = 1, 2, \ldots, N, \ k = 1, 2, \ldots, K.$$

Let us define a positive constant μ as

$$\mu = \frac{1}{\sqrt{MK + NK}} \tau_1 \psi_{\tau_2}(\boldsymbol{W}^{(0)}, \boldsymbol{H}^{(0)}).$$

Because $\nabla_{\boldsymbol{W}} f(\boldsymbol{W}, \boldsymbol{H})$ and $\nabla_{\boldsymbol{H}} f(\boldsymbol{W}, \boldsymbol{H})$ are continuous on $\mathcal{F}$, the following statements hold true.

1. For any (m, k) such that $(\nabla_{\boldsymbol{W}} f(\boldsymbol{W}^{(\infty)}, \boldsymbol{H}^{(\infty)}))_{mk} = 0$, there exists a positive integer I_{mk}^{W} such that
$$\left|(\nabla_{\boldsymbol{W}} f(\boldsymbol{W}^{(t_i)}, \boldsymbol{H}^{(t_i)}))_{mk}\right| \leq \mu$$
for all $i \geq I_{mk}^{\mathrm{W}}$.
2. For any (m, k) such that $(\nabla_{\boldsymbol{W}} f(\boldsymbol{W}^{(\infty)}, \boldsymbol{H}^{(\infty)}))_{mk} > 0$, there exists a positive integer I_{mk}^{W} such that
$$(\nabla_{\boldsymbol{W}} f(\boldsymbol{W}^{(t_i)}, \boldsymbol{H}^{(t_i)}))_{mk} > 0, \quad w_{mk}^{(t_i)} \leq \tau_2$$
for all $i \geq I_{mk}^{\mathrm{W}}$.
3. For any (n, k) such that $(\nabla_{\boldsymbol{H}} f(\boldsymbol{W}^{(\infty)}, \boldsymbol{H}^{(\infty)}))_{nk} = 0$, there exists a positive integer I_{nk}^{H} such that
$$\left|(\nabla_{\boldsymbol{H}} f(\boldsymbol{W}^{(t_i)}, \boldsymbol{H}^{(t_i)}))_{nk}\right| \leq \mu$$
for all $i \geq I_{nk}^{\mathrm{H}}$.
4. For any (n, k) such that $(\nabla_{\boldsymbol{H}} f(\boldsymbol{W}^{(\infty)}, \boldsymbol{H}^{(\infty)}))_{nk} > 0$, there exists a positive integer I_{nk}^{H} such that
$$(\nabla_{\boldsymbol{H}} f(\boldsymbol{W}^{(t_i)}, \boldsymbol{H}^{(t_i)}))_{nk} > 0, \quad h_{nk}^{(t_i)} \leq \tau_2$$
for all $i \geq I_{nk}^{\mathrm{H}}$.

From these statements, we see that

$$\left|(\boldsymbol{G}_{\tau_2}^{\mathrm{W}}(\boldsymbol{W}^{(t_i)}, \boldsymbol{H}^{(t_i)}))_{mk}\right| \leq \mu, \quad m = 1, 2, \ldots, M, \quad k = 1, 2, \ldots, K,$$
$$\left|(\boldsymbol{G}_{\tau_2}^{\mathrm{H}}(\boldsymbol{W}^{(t_i)}, \boldsymbol{H}^{(t_i)}))_{nk}\right| \leq \mu, \quad n = 1, 2, \ldots, N, \quad k = 1, 2, \ldots, K$$

for all $i \geq I = \max\{I_{11}^{\mathrm{W}}, \ldots, I_{MK}^{\mathrm{W}}, I_{11}^{\mathrm{H}}, \ldots, I_{NK}^{\mathrm{H}}\}$. Therefore, the inequality

$$\psi_{\tau_2}(\boldsymbol{W}^{(t_i)}, \boldsymbol{H}^{(t_i)}) = \sqrt{\|\boldsymbol{G}_{\tau_2}^{\mathrm{W}}(\boldsymbol{W}^{(t_i)}, \boldsymbol{H}^{(t_i)})\|_{\mathrm{F}}^2 + \|\boldsymbol{G}_{\tau_2}^{\mathrm{H}}(\boldsymbol{W}^{(t_i)}, \boldsymbol{H}^{(t_i)})\|_{\mathrm{F}}^2}$$

Algorithm 3 HALS Algorithm for NMF using Algorithm 1 and Projected Gradient-Based Stopping Condition

Input: $\boldsymbol{X} \in \mathbb{R}_+^{M\times N}$, $(\boldsymbol{W}^{(0)}, \boldsymbol{H}^{(0)}) \in \mathcal{F}$, $\delta, \tau_1, \tau_2 \in \mathbb{R}_{++}$
Output: $(\boldsymbol{W}, \boldsymbol{H}) \in \mathcal{F}$ satisfying (23)
1: If $\psi_{\tau_2}(\boldsymbol{W}^{(0)}, \boldsymbol{H}^{(0)}) = 0$, return $(\boldsymbol{W}^{(0)}, \boldsymbol{H}^{(0)})$ and stop. Otherwise set $(\boldsymbol{W}, \boldsymbol{H}) \leftarrow (\boldsymbol{W}^{(0)}, \boldsymbol{H}^{(0)})$ and $\boldsymbol{E} \leftarrow \boldsymbol{X} - \boldsymbol{W}\boldsymbol{H}^{\mathrm{T}}$.
2: Update $(\boldsymbol{W}, \boldsymbol{H})$ and $\boldsymbol{E}$ using Algorithm 1.
3: If (23) holds, return $(\boldsymbol{W}, \boldsymbol{H})$ and stop. Otherwise go to Step 2.

Table 1 Statistics of the datasets used in the experiments

	Features (M)	Instances (N)	Classes (K)
Olivetti	4096	400	40
CLUTO (tr41)	7454	878	10

$$\leq \mu\sqrt{MK + NK}$$
$$= \tau_1 \psi_{\tau_2}(\boldsymbol{W}^{(0)}, \boldsymbol{H}^{(0)})$$

holds for all $i \geq I$. This means that the stopping condition (23) holds in a finite number of iterations. However, this contradicts the assumption that Algorithm 3 does not stop. □

6 Numerical experiments

In order to examine the practical performance of the proposed update rule, the authors conducted numerical experiments using the real-world datasets: Olivetti[1] and CLUTO[2] (tr41). The former is a dataset of face images, and the latter is that of documents. The statistics of these two datasets is shown in Table 1. In the experiments, two global-convergence-guaranteed update rules were applied to the nonnegative matrices obtained from the datasets. One is Algorithm 1 (denoted as 'proposed') and the other is the update rule described by (10) and (11) (denoted as 'positive'). These two update rules are compared in terms of the evolution of the objective function value and the number of unsatisfied inequalities in the relaxed KKT conditions, and the characteristics of the obtained factor matrices.

Experimental setup is shown in Table 2. The value of δ in the proposed update rule is set to 10^{-8} in all experiments, while the value of ϵ in the positive one is set to 10^{-4} or 10^{-8} depending on the experiment. The iteration is terminated when the stopping condition described by (21) and (22) is satisfied or the number of iterations reaches 500. The values of κ_1 and κ_2 in the stopping condition are set to 1.0 and 2ϵ, respectively, in all experiments. Note that the finite termination of the positive update rule is guaranteed if κ_2 is greater than ϵ. This can be proved in the same way as Theorem 7 (see [29] for details). Three different initial solutions are generated for each dataset in such a way that each element is drawn from independent uniform distributions on the intervals $[0, 1]$, $[0, 0.5]$ and $[0, 0.25]$ which are called the 'large', 'medium' and 'small' initial solutions, respectively.

Results of Experiment 1 are summarized in Fig. 2 and Table 3. Figure 2 shows the evolution of the objective function value and the number of unsatisfied inequalities in (21) and (22).

[1] https://scikit-learn.org/0.19/datasets/olivetti_faces.html.
[2] http://glaros.dtc.umn.edu/gkhome/views/cluto.

Table 2 Experimental setup

	Dataset	δ	ϵ	κ_1	κ_2
Experiment 1	Olivetti	10^{-8}	10^{-4}	1.0	2×10^{-4}
Experiment 2	Olivetti	10^{-8}	10^{-8}	1.0	2×10^{-8}
Experiment 3	CLUTO (tr41)	10^{-8}	10^{-4}	1.0	2×10^{-4}
Experiment 4	CLUTO (tr41)	10^{-8}	10^{-8}	1.0	2×10^{-8}

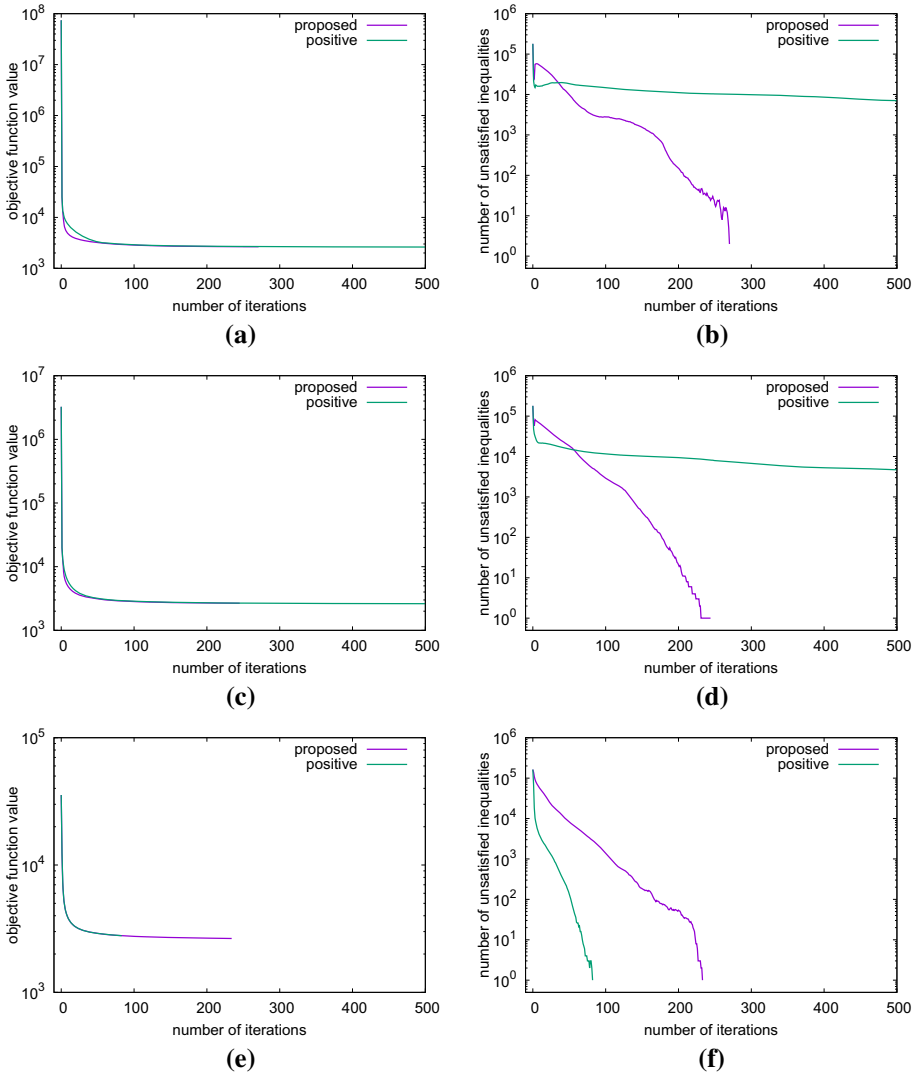

Fig. 2 The evolution of the objective function value (left column) and the number of unsatisfied inequalities in (21) and (22) (right column) in Experiment 1. The first, second and third rows show the results for the large, medium and small initial solutions, respectively

Table 3 Characteristics of the solutions obtained by the proposed and positive update rules in Experiment 1. The notation 'positive+replacement' means that all ϵ in the factor matrices obtained by the positive update rule are replaced with zero

Initial solution	Large	Medium	Small
Iterations (proposed)	271	245	234
Iterations (positive)	500	500	83
Objective function value (proposed)	2651.977	2665.153	2650.139
Objective function value (positive)	2619.553	2622.726	2786.625
Variables at the lower bound (proposed)	40017	40850	39429
Variables at the lower bound (positive)	41630	40340	42450
Unsatisfied inequalities (proposed)	0	0	0
Unsatisfied inequalities (positive)	7059	4729	0
Unsatisfied inequalities (positive+replacement)	40639	28311	4

We easily see from the figure that the two update rules decrease the objective function value in a similar way until one of them satisfies the stopping condition. In contrast, the behavior of these update rules with respect to the number of unsatisfied inequalities is quite different. The proposed update rule decreases the number at a similar rate for all the initial solutions, and satisfies the stopping condition between 200 and 300 iterations. This is because the normalization process is included in the proposed update rule. The positive update rule decreases the number very slowly, and cannot satisfy the stopping condition in 500 iterations for the large and medium initial solutions, while for the small initial solution it decreases the number very fast and satisfies the stopping condition in less than 100 iterations.

Table 3 shows the characteristics of the solutions obtained by the proposed and positive update rules. Some important facts are observed in this table. The first one is that a small objective function value does not necessarily mean that the number of unsatisfied inequalities is small. In fact, the solution obtained by the positive update rule for the large initial solution gives the smallest objective function value and the largest number of unsatisfied inequalities. Also, the solution obtained by the positive update rule for the small initial solution gives the largest objective function value but satisfies all the inequalities. The second fact is that about a quarter of the variables are at the lower bound in all cases. Hence the solutions obtained by the proposed update rule are sparse because the lower bound is zero. In contrast, the solutions obtained by the positive update rule are dense because the lower bound is a positive constant ϵ. The third fact is that the replacement of all ϵ with zero in each solution obtained by the positive update rule increases the number of unsatisfied inequalities. In particular, the replacement changes a solution that satisfies the stopping condition to another one that does not. It is thus not always possible to find a sparse solution that satisfies the relaxed KKT conditions using the positive update rule, while we can always do it using the proposed update rule. This is an advantage of the proposed update rule against the positive one.

Results of Experiment 2 are summarized in Fig. 3 and Table 4 just like Experiment 1. The evolution of the objective function value and the number of unsatisfied inequalities in (21) and (22) shown in Fig. 3 are similar to those in Experiment 1 (see Fig. 2), though the values of ϵ and κ_2 are quite different. The characteristics shown in Table 4 are similar to those in Table 3 but there is one important difference. The number of unsatisfied inequalities is zero before and after the replacement of all ϵ with zero in the solution obtained by the positive update rule for the small initial solution. This indicates that we can find a sparse solution that

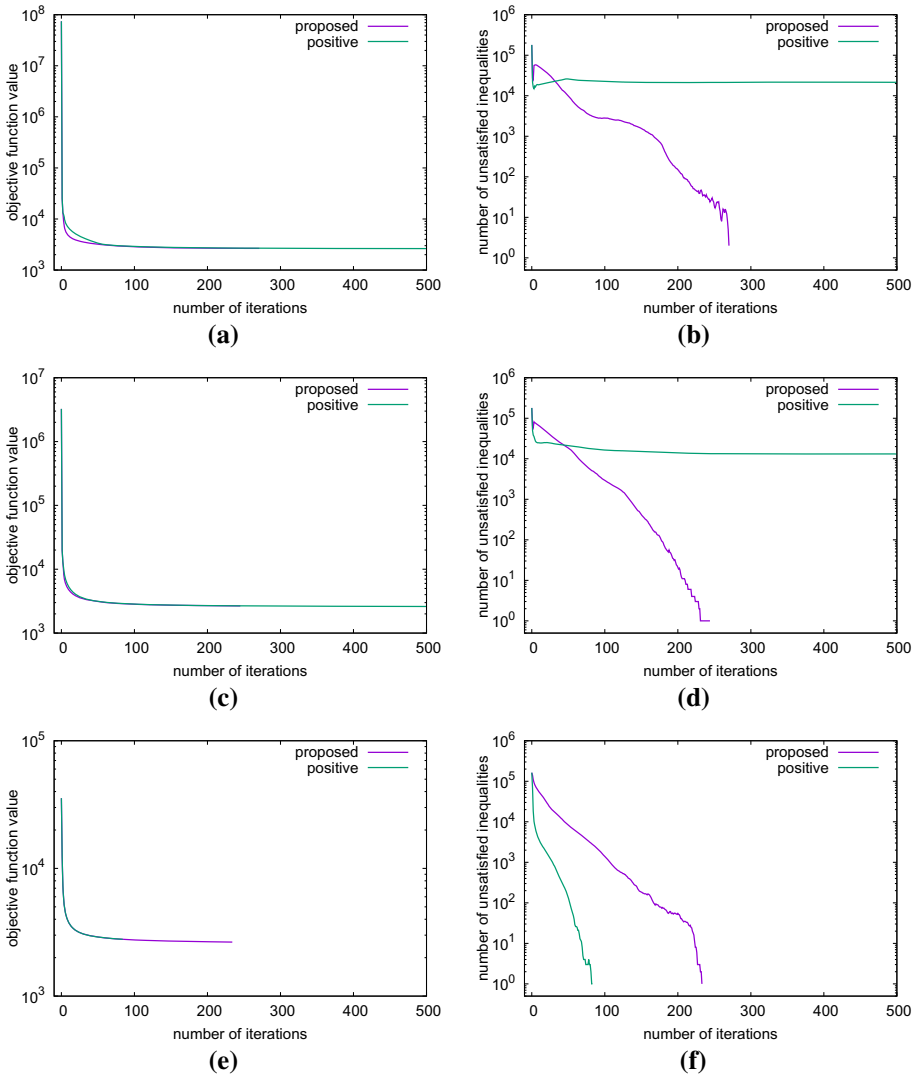

Fig. 3 The evolution of the objective function value (left column) and the number of unsatisfied inequalities in (21) and (22) (right column) in Experiment 2. The first, second and third rows show the results for the large, medium and small initial solutions, respectively

satisfies the relaxed KKT conditions using the positive update rule if the magnitude of the initial solution and the value of ϵ are sufficiently small. However, it is difficult in general to know in advance how small these values should be.

Results of Experiment 3 are summarized in Fig. 4 and Table 5. The evolution of the objective function value and the number of unsatisfied inequalities in (21) and (22) shown in Fig. 4 are similar to those in Experiment 1 (see Fig. 2), though the dataset is different. The characteristics shown in Table 5 are also similar to those in Table 3 but there are two main differences. One is that a solution with a smaller objective function value satisfies more inequalities in (21) and (22). The other is that the number of variables at the lower bound in

Table 4 Characteristics of the solutions obtained by the proposed and positive update rules in Experiment 2

Initial solution	Large	Medium	Small
Iterations (proposed)	271	245	234
Iterations (positive)	500	500	84
Objective function value (proposed)	2651.977	2665.153	2650.139
Objective function value (positive)	2632.812	2612.250	2783.772
Variables at the lower bound (proposed)	40017	40850	39429
Variables at the lower bound (positive)	41912	40942	42430
Unsatisfied inequalities (proposed)	0	0	0
Unsatisfied inequalities (positive)	21481	13133	0
Unsatisfied inequalities (positive+replacement)	45555	34177	0

Table 5 Characteristics of the solutions obtained by the proposed and positive update rules in Experiment 3

Initial solution	Large	Medium	Small
Iterations (proposed)	286	328	271
Iterations (positive)	500	500	182
Objective function value (proposed)	504367.3	504367.3	504367.3
Objective function value (positive)	504440.2	504402.8	504386.0
Variables at the lower bound (proposed)	42377	42377	42378
Variables at the lower bound (positive)	49595	47846	46570
Unsatisfied inequalities (proposed)	0	0	0
Unsatisfied inequalities (positive)	71	2	0
Unsatisfied inequalities (positive+replacement)	5556	1041	49

Table 6 Characteristics of the solutions obtained by the proposed and positive update rules in Experiment 4

Initial solution	Large	Medium	Small
Iterations (proposed)	286	328	271
Iterations (positive)	209	245	182
Objective function value (proposed)	504367.3	504367.3	504367.3
Objective function value (positive)	504367.3	504367.3	504367.3
Variables at the lower bound (proposed)	42377	42377	42378
Variables at the lower bound (positive)	42384	42377	42385
Unsatisfied inequalities (proposed)	0	0	0
Unsatisfied inequalities (positive)	0	0	0
Unsatisfied inequalities (positive+replacement)	0	0	0

each solution obtained by the positive update rule is higher than that in the corresponding solution obtained by the proposed update rule.

Results of Experiment 4 are summarized in Fig. 5 and Table 6. As for the proposed update rule, the evolution of the objective function value and the number of unsatisfied inequalities in (21) and (22) are similar to those in Experiment 3 (see Fig. 4). In contrast,

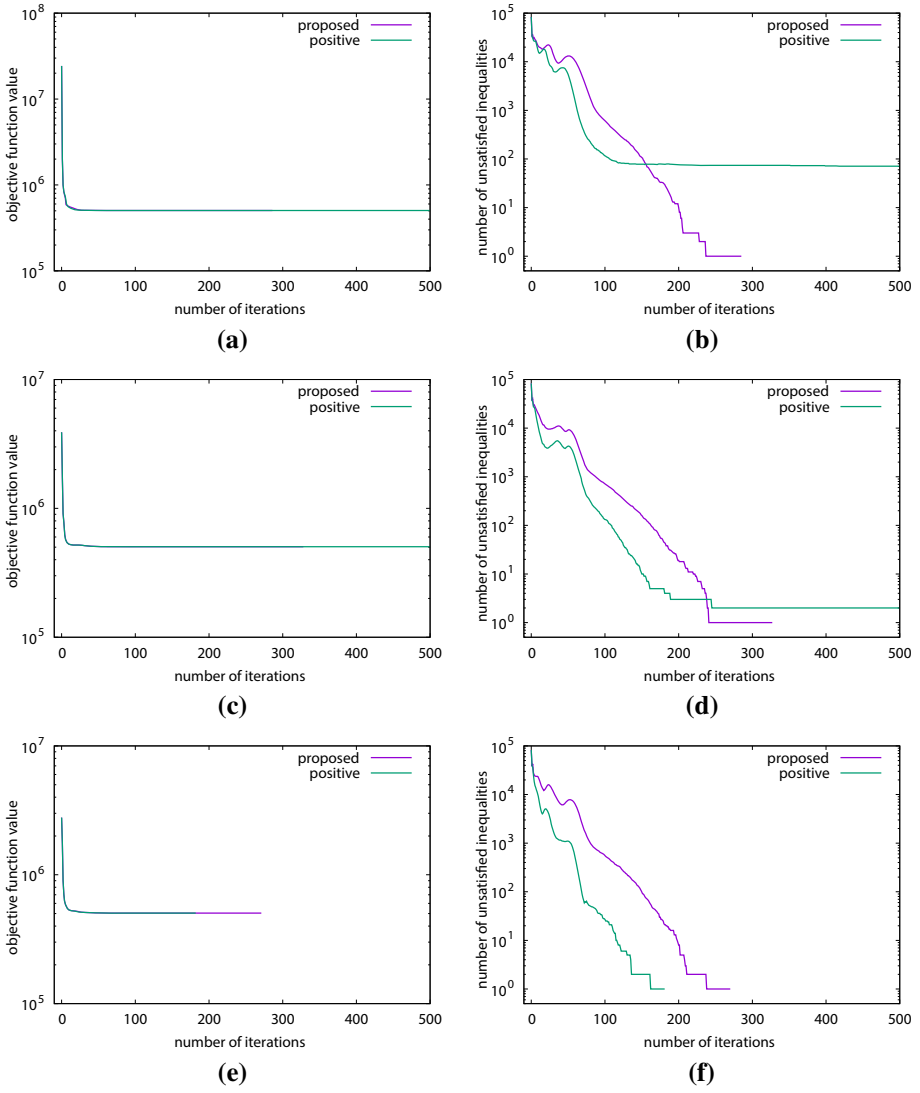

Fig. 4 The evolution of the objective function value (left column) and the number of unsatisfied inequalities in (21) and (22) (right column) in Experiment 3. The first, second and third rows show the results for the large, medium and small initial solutions, respectively

the behavior of the positive update rule is quite different from that in Experiment 3. The number of unsatisfied inequalities decreases faster than the proposed update rule for all the initial solutions, and reaches zero in less than 200 iterations. All the solutions obtained by the proposed and positive update rules have almost the same objective function value and very similar numbers of unsatisfied inequalities, as shown in Table 6. In addition, the number of unsatisfied inequalities is not affected by the replacement of all ϵ with zero for all the solutions obtained by the positive update rule. This indicates that we can find a sparse solution that

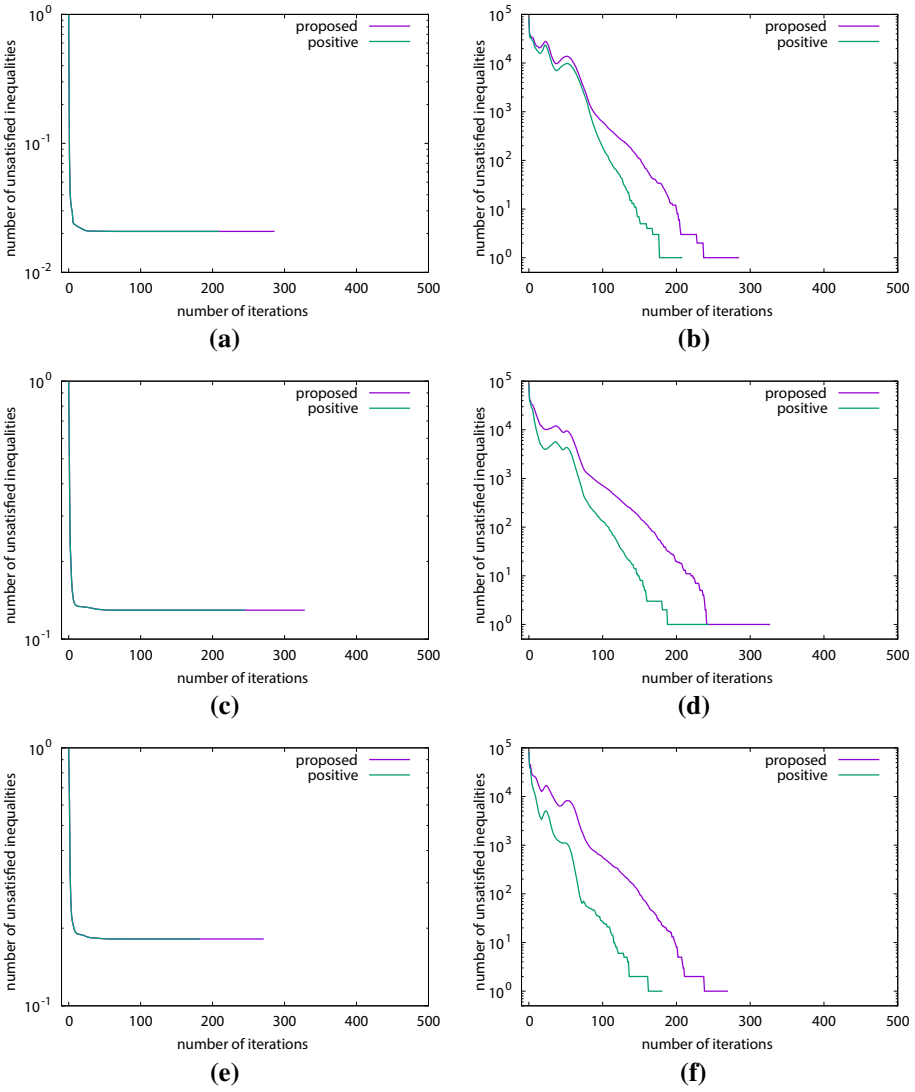

Fig. 5 The evolution of the objective function value (left column) and the number of unsatisfied inequalities in (21) and (22) (right column) in Experiment 4. The first, second and third rows show the results for the large, medium and small initial solutions, respectively

satisfies the relaxed KKT conditions using the positive update rule if the magnitude of the initial solution and the value of ϵ are properly selected.

7 Applicability of proposed update rule to variants of HALS algorithm

In this section, we introduce some variants of the HALS algorithm to which our update rule can be applied in order to guarantee the well-definedness and/or the global convergence. First one is the accelerated HALS algorithm [17]. The idea behind this algorithm is very simple. In each iteration, $\boldsymbol{W}$ is updated several times while $\boldsymbol{H}$ is fixed, and then $\boldsymbol{H}$ is updated several times while $\boldsymbol{W}$ is fixed. It was shown through experiments using image and text datasets that this algorithm significantly outperforms the original HALS algorithm [17]. Now we claim that the global convergence of this algorithm is guaranteed if Algorithm 1 is incorporated into it. In each iteration, the algorithm updates all columns of $\boldsymbol{W}$ several times using the update rule in Step 3, next updates all columns of $\boldsymbol{W}$ and $\boldsymbol{H}$ using Steps 4 and 5 once, and then updates all columns of $\boldsymbol{H}$ several times using the update rule in Step 6.

The second one is the fast coordinate descent algorithm with variable selection [26]. In each iteration, M rows of $\boldsymbol{W}$ are updated one by one and then N rows of $\boldsymbol{H}$ are updated one by one. Each row of $\boldsymbol{W}$ or $\boldsymbol{H}$ is updated by repeating the following two steps until some condition is satisfied: i) selection of one element based on the potential decrease in the objective function value, and ii) update of the selected element. It was shown through experiments using synthetic and real-world datasets that this algorithm is considerably faster than conventional algorithms [26]. Again, we claim that the global convergence of this algorithm is guaranteed if Algorithm 1 is incorporated into it. To be more specific, when each row of $\boldsymbol{W}$ or $\boldsymbol{H}$ is updated, the update rule in Step 3 or Step 6 of Algorithm 1 can be used for both the computation of the potential decrease in the objective function value and the update of the selected element. One important point is that the normalization procedure in Steps 4 and 5 should be done between the update of M rows of $\boldsymbol{W}$ and the update of N rows of $\boldsymbol{H}$.

The third one is the randomized HALS algorithm [13] which is based on the probabilistic framework for low-rank approximations [21]. In the first step, this algorithm constructs a surrogate matrix $\boldsymbol{B} \in \mathbb{R}^{L\times N}$ with $K < L \ll M$ as follows. First, $\boldsymbol{X}$ is multiplied by a random matrix $\boldsymbol{\Omega} \in \mathbb{R}^{N\times L}$ to get $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\Omega}$. Next, a matrix $\boldsymbol{Q} \in \mathbb{R}^{M\times L}$ with orthogonal columns is obtained by performing the QR-decomposition of $\boldsymbol{Y}$. Finally, the surrogate matrix is obtained by $\boldsymbol{B} = \boldsymbol{Q}^{\mathrm{T}}\boldsymbol{X}$. The surrogate matrix $\boldsymbol{B}$ obtained like this is expected to capture the essential information of $\boldsymbol{X}$. In the next step, this algorithm solves the optimization problem:

$$\begin{aligned} &\text{minimize } \frac{1}{2}\left\|\boldsymbol{B} - \tilde{\boldsymbol{W}}\boldsymbol{H}^{\mathrm{T}}\right\|_{\mathrm{F}}^2 \\ &\text{subject to } \boldsymbol{Q}\tilde{\boldsymbol{W}} \geq \boldsymbol{0}_{M\times K},\ \boldsymbol{H} \geq \boldsymbol{0}_{N\times K} \end{aligned}$$

by an iterative algorithm very similar to the HALS algorithm. It was shown through experiments using hand-written digits and face image datasets that the randomized HALS algorithm has a substantially lower computational cost than the deterministic one, and attains almost the same reconstruction error as the deterministic one [13]. The technique used in our update rule can be easily applied to this algorithm in order to ensure that it is well-defined.

In addition to these three, there are many other algorithms to which our update rule can be applied. One example is the distributed HALS algorithm for multiagent networks [10]. This algorithm is based on the update rule given by (10) and (11) to guarantee the global convergence. By using one of our update rules, this algorithm can find a stationary point of the original optimization problem (1).

8 Conclusions

In this paper, we have proposed a novel update rule of the HALS algorithm for NMF, and proved its global convergence using Zangwill's global convergence theorem. The proposed update rule has the same computational complexity per iteration as the update rule in the original HALS algorithm. In addition, unlike the global-convergence-guaranteed update rules in the literature [29, 30], the proposed update rule does not restrict the range of each variable to a subset of $\mathbb{R}_{++}$. This allows us to obtain sparse factor matrices. We have also given two types of stopping conditions and proved the finite termination of the proposed update rule combined with these stopping conditions.

One future direction of this work is to extend our results to Nonnegative Tensor Factorization (NTF) [7, 55] which is expected to be used in various applications such as recommender systems [56], while the global convergence property has not yet been analyzed in depth.

Acknowledgements The authors would like to thank anonymous reviewers for their valuable comments and suggestions to improve the quality of the paper.

References

1. Arora, S., Ge, R., Kannan, R., Moitra, A.: Computing a nonnegative matrix factorization–provably. SIAM J. Comput. **45**(4), 1582–1611 (2016)
2. Berry, M.W., Browne, M., Langville, A.N., Pauca, V.P., Plemmons, R.J.: Algorithms and applications for approximate nonnegative matrix factorization. Comput. Stat. Data Anal. **52**(1), 155–173 (2007)
3. Boyd, S., Boyd, S.P., Vandenberghe, L.: Convex Optimization. Cambridge University Press, Cambridge (2004)
4. Cai, D., He, X., Han, J., Huang, T.S.: Graph regularized nonnegative matrix factorization for data representation. IEEE Trans. Pattern Anal. Mach. Intell. **33**(8), 1548–1560 (2010)
5. Cichocki, A., Phan, A.H.: Fast local algorithms for large scale nonnegative matrix and tensor factorizations. IEICE Trans. Fundam. Electron. Commun. Comput. Sci. **92**(3), 708–721 (2009)
6. Cichocki, A., Zdunek, R., Amari, S.I.: Hierarchical ALS algorithms for nonnegative matrix and 3D tensor factorization. In: proceedings of the 2017 International conference on independent component analysis and signal separation, pp. 169–176 (2007)
7. Cichocki, A., Zdunek, R., Phan, A.H., Amari, S.I.: Nonnegative Matrix and Tensor Factorizations: Applications to Exploratory Multi-Way Data Analysis and Blind Source Separation. John Wiley & Sons, Hoboken (2009)
8. Cunningham, J.P., Ghahramani, Z.: Linear dimensionality reduction: survey, insights, and generalizations. J. Mach. Learn. Res. **16**(1), 2859–2900 (2015)
9. Ding, C., Li, T., Peng, W., Park, H.: Orthogonal nonnegative matrix tri-factorizations for clustering. In: proceedings of the 12th ACM SIGKDD International conference on knowledge discovery and data mining, pp. 126–135 (2006)
10. Domen, Y., Migita, T., Takahashi, N.: A distributed HALS algorithm for Euclidean distance-based nonnegative matrix factorization. In: proceedings of the 2019 IEEE symposium series on computational intelligence, pp. 1332–1337 (2019)
11. Donoho, D., Stodden, V.: When does non-negative matrix factorization give a correct decomposition into parts? Adv. Neural Inf. Process. Syst. **16**, 1141–1148 (2003)

12. Dorffer, C., Puigt, M., Delmaire, G., Roussel, G.: Informed nonnegative matrix factorization methods for mobile sensor network calibration. IEEE Trans. Signal Inf. Process. Netw. **4**(4), 667–682 (2018)
13. Erichson, N.B., Mendible, A., Wihlborn, S., Kutz, J.N.: Randomized nonnegative matrix factorization. Pattern Recognit. Lett. **104**, 1–7 (2018)
14. Févotte, C., Bertin, N., Durrieu, J.L.: Nonnegative matrix factorization with the Itakura–Saito divergence: with application to music analysis. Neural Comput. **21**(3), 793–830 (2009)
15. Gillis, N.: Nonnegative Matrix Factorization. SIAM (2020)
16. Gillis, N., Glineur, F.: Nonnegative factorization and the maximum edge biclique problem. arXiv e-prints (2008)
17. Gillis, N., Glineur, F.: Accelerated multiplicative updates and hierarchical ALS algorithms for nonnegative matrix factorization. Neural Comput. **24**(4), 1085–1105 (2012)
18. Gligorijević, V., Panagakis, Y., Zafeiriou, S.: Non-negative matrix factorizations for multiplex network analysis. IEEE Trans. Pattern Anal. Mach. Intell. **41**(4), 928–940 (2018)
19. Gonzalez, E.F., Zhang, Y.: Accelerating the Lee-Seung algorithm for nonnegative matrix factorization. Tech. rep. (2005)
20. Guan, N., Tao, D., Luo, Z., Yuan, B.: NeNMF: an optimal gradient method for nonnegative matrix factorization. IEEE Trans. Signal Process. **60**(6), 2882–2898 (2012)
21. Halko, N., Martinsson, P.G., Tropp, J.A.: Finding structure with randomness: probabilistic algorithms for constructing approximate matrix decompositions. SIAM Rev. **53**(2), 217–288 (2011)
22. Hamon, R., Borgnat, P., Flandrin, P., Robardet, C.: Extraction of temporal network structures from graph-based signals. IEEE Trans. Signal Inf. Process. Netw. **2**(2), 215–226 (2016)
23. Ho, N.D.: Nonnegative matrix factorization algorithms and applications. Ph.D. thesis, Université catholique de Louvain (2008)
24. Hoyer, P.O.: Non-negative sparse coding. In: proceedings of the 12th IEEE Workshop on neural networks for signal processing, pp. 557–565 (2002)
25. Hoyer, P.O.: Non-negative matrix factorization with sparseness constraints. J. Mach. Learn. Res. **5**, 1457–1469 (2004)
26. Hsieh, C.J., Dhillon, I.S.: Fast coordinate descent methods with variable selection for non-negative matrix factorization. In: proceedings of the 17th ACM SIGKDD International conference on knowledge discovery and data mining, pp. 1064–1072 (2011)
27. Kim, J., He, Y., Park, H.: Algorithms for nonnegative matrix and tensor factorizations: a unified view based on block coordinate descent framework. J. Glob. Optim. **58**(2), 285–319 (2014)
28. Kim, J., Park, H.: Fast nonnegative matrix factorization: an active-set-like method and comparisons. SIAM J. Sci. Comput. **33**(6), 3261–3281 (2011)
29. Kimura, T., Takahashi, N.: Global convergence of a modified HALS algorithm for nonnegative matrix factorization. In: proceedings of the 2015 IEEE 6th International Workshop on computational advances in multi-sensor adaptive processing, pp. 21–24 (2015)
30. Kimura, T., Takahashi, N.: Gauss-Seidel HALS algorithm for nonnegative matrix factorization with sparseness and smoothness constraints. IEICE Trans. Fundam. Electron. Commun. Comput. Sci. **100**(12), 2925–2935 (2017)
31. Kitamura, D., Ono, N., Sawada, H., Kameoka, H., Saruwatari, H.: Determined blind source separation unifying independent vector analysis and nonnegative matrix factorization. IEEE/ACM Trans. Audio Speech Lang. Process. **24**(9), 1626–1641 (2016)
32. Lee, D.D., Seung, H.S.: Learning the parts of objects by non-negative matrix factorization. Nature **401**(6755), 788–791 (1999)
33. Lee, D.D., Seung, H.S.: Algorithms for non-negative matrix factorization. In: advances in Neural Information Processing Systems, pp. 556–562 (2001)
34. Li, Z., Tang, J., He, X.: Robust structured nonnegative matrix factorization for image representation. IEEE Trans. Neural Netw. Learn. Syst. **29**(5), 1947–1960 (2017)
35. Lin, C.J.: On the convergence of multiplicative update algorithms for nonnegative matrix factorization. IEEE Trans. Neural Netw. **18**(6), 1589–1596 (2007)
36. Lin, C.J.: Projected gradient methods for nonnegative matrix factorization. Neural Comput. **19**(10), 2756–2779 (2007)
37. Lu, S., Hong, M., Wang, Z.: A nonconvex splitting method for symmetric nonnegative matrix factorization: convergence analysis and optimality. IEEE Trans. Signal Process. **65**(12), 3120–3135 (2017)
38. Nakatsu, S., Takahashi, N.: A novel Newton-type algorithm for nonnegative matrix factorization with alpha-divergence. In: proceedings of the 2017 International conference on neural information processing, pp. 335–344. Springer (2017)
39. Paatero, P., Tapper, U.: Positive matrix factorization: a non-negative factor model with optimal utilization of error estimates of data values. Environmetrics **5**(2), 111–126 (1994)

40. Pauca, V.P., Piper, J., Plemmons, R.J.: Nonnegative matrix factorization for spectral data analysis. Linear Algebra Appl. **416**(1), 29–47 (2006)
41. Pauca, V.P., Shahnaz, F., Berry, M.W., Plemmons, R.J.: Text mining using non-negative matrix factorizations. In: proceedings of the 2004 SIAM International conference on data mining, pp. 452–456 (2004)
42. Razaviyayn, M., Hong, M., Luo, Z.Q.: A unified convergence analysis of block successive minimization methods for nonsmooth optimization. SIAM J. Optim. **23**(2), 1126–1153 (2013)
43. Recht, B., Re, C., Tropp, J., Bittorf, V.: Factoring nonnegative matrices with linear programs. Adv. Neural Inf. Process. Syst. **25**, 1214–1222 (2012)
44. Sano, T., Migita, T., Takahashi, N.: A damped Newton algorithm for nonnegative matrix factorization based on alpha-divergence. In: proceedings of the 2019 6th International conference on systems and informatics, pp. 463–468. IEEE (2019)
45. Sriperumbudur, B.K., Lanckriet, G.R.: On the convergence of the concave-convex procedure. In: proceedings of the 22nd International conference on neural information processing systems, pp. 1759–1767 (2009)
46. Takahashi, N., Hibi, R.: Global convergence of modified multiplicative updates for nonnegative matrix factorization. Comput. Optim. Appl. **57**(2), 417–440 (2014)
47. Takahashi, N., Katayama, J., Seki, M., Takeuchi, J.: A unified global convergence analysis of multiplicative update rules for nonnegative matrix factorization. Comput. Optim. Appl. **71**(1), 221–250 (2018)
48. Takahashi, N., Nishi, T.: Global convergence of decomposition learning methods for support vector machines. IEEE Trans. Neural Netw. **17**(6), 1362–1369 (2006)
49. Vavasis, S.A.: On the complexity of nonnegative matrix factorization. SIAM J. Optim. **20**(3), 1364–1377 (2010)
50. Wang, F., Li, T., Wang, X., Zhu, S., Ding, C.: Community discovery using nonnegative matrix factorization. Data Min. Knowl. Discov. **22**(3), 493–521 (2011)
51. Wold, S., Esbensen, K., Geladi, P.: Principal component analysis. Chemom. Intell. Lab. Syst. **2**(1–3), 37–52 (1987)
52. Yang, Z., Oja, E.: Unified development of multiplicative algorithms for linear and quadratic nonnegative matrix factorization. IEEE Trans. Neural Netw. **22**(12), 1878–1891 (2011)
53. Zangwill, W.I.: Nonlinear Programming: A Unified Approach. Prentice-Hall, Englewood Cliffs, New Jersey (1969)
54. Zdunek, R., Cichocki, A.: Non-negative matrix factorization with quasi-Newton optimization. In: International conference on artificial intelligence and soft computing, pp. 870–879 (2006)
55. Zdunek, R., Fonal, K.: Randomized nonnegative tensor factorization for feature extraction from high-dimensional signals. In: 2018 25th International conference on systems, signals and image processing, pp. 1–5 (2018)
56. Zhang, W., Sun, H., Liu, X., Guo, X.: Temporal QoS-aware web service recommendation via non-negative tensor factorization. In: proceedings of the 23rd International conference on World Wide Web, pp. 585–596 (2014)