ORTHOGONAL NONNEGATIVE MATRIX FACTORIZATION VIA MINIMIZATION OVER THE NULL SPACE

XIAOJUN CHEN*, WEN LI†, AND QILUN LUO‡

Abstract. This paper gives a necessary and sufficient condition for a nonnegative matrix that has an orthogonal nonnegative matrix factorization (ONMF) via characterization of the null space. We propose an optimization model to minimize the Frobenius norm of the product of a given nonnegative matrix and a variable matrix subject to the constraints defined by the necessary and sufficient condition. Moreover, we present an augmented Lagrangian algorithm for solving this minimization model, and prove the global convergence to a stationary point. Two factor matrices for the ONMF of the given matrix can be easily obtained by the outputs of the algorithm. Preliminary numerical results using synthetic and real-world data with applications in clustering show that our approach outperforms some existing ONMF methods regarding accuracy and robustness.

Key words. Orthogonal nonnegative matrix factorization, null space, augmented Lagrangian method, clustering.

MSC codes. 90C26, 90C30, 65F30

1. Introduction. Given a nonnegative matrix $A \in \mathbb{R}^{m\times n}_+$ with $m \geq n$ and a positive integer $r < n$, the minimization problem

$$\min \|A - BC^\top\|_F^2, \qquad \text{subject to} \quad B \in \mathbb{R}^{m\times r}_+, C \in \mathbb{R}^{n\times r}_+, \, C^\top C = I \tag{1.1}$$

is a popular mathematical model for Orthogonal Nonnegative Matrix Factorization (ONMF), which aims to approximate A as the product of two matrices $B, C^\top$ in the feasible set of (1.1). ONMF has been successfully applied in various areas including clustering, image science, hyperspectral unmixing [12, 14, 19, 20]. However, solving (1.1) is generally NP-hard due to the combinatorial features. In particular, for any feasible point, each row of C has at most one positive element and each column of C takes the unit norm. Many methods have been developed for solving (1.1) including multiplicative update rules [7, 17], hierarchical alternating least squares (HALS) [13, 15, 21], and subspace exploration techniques [1]. Moreover, problem (1.1) can be considered as a special case of optimization problems with nonnegative orthogonal constraints [6, 12]. In this paper, we focus on problem (1.1) where the objective function is a continuously differentiable bilinear composite function.

We say that a matrix $A \in \mathbb{R}^{m\times n}_+$ has an ONMF if the optimal value of (1.1) is 0, that is, there exists a feasible point of (1.1) such that $A = BC^\top$. If A has an ONMF, then $AC = B$ and thus $A = ACC^\top$. In [19], Pan and Ng proposed the following minimization model

$$\min \|A - ACC^\top\|_F^2, \qquad \text{subject to} \quad C \in \mathbb{O}^{n\times r}_+, \tag{1.2}$$

*Department of Applied Mathematics, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong. Email: maxjchen@polyu.edu.hk. The research of this author is supported by the National Key R & D Program of China under grant 2023YFA1009303 and Hong Kong Research Grant Council project PolyU15300123.

†School of Mathematics and Statistics, Hanshan Normal University, Chaozhou, Guangdong, China. School of Mathematical Sciences, South China Normal University, Guangzhou, Guangdong, China. Email: liwen@scnu.edu.cn. The research is supported in part by the National Natural Science Foundation of China under grant 12471351.

‡School of Mathematical Sciences, South China Normal University, Guangzhou, Guangdong, China. Email: qilunluo92@gmail.com. The research is supported in part by the National Natural Science Foundation of China under grant 12301483, and the Natural Science Foundation of Guangdong Province of China under grants 2024A1515011527 and 2025A1515011611. Corresponding author.

where $\mathbb{O}_+^{n\times r} = \{C \in \mathbb{R}_+^{n\times r} \,|\, C^\top C = I\}$. Let $K = CC^\top$. Then $K = K^\top, K^2 = K, 0 \le K_{i,j} \le 1, \|K\|^2 = r$ [19, Theorem 3]. Using these properties of K, Pan and Ng proposed an algorithm to solve the following convex optimization model

$$\min \|A - AK\|_F^2 + \theta\|K\|_1 + \beta\|K\|_*, \qquad \text{subject to} \quad K = K^\top, K \in \mathbb{R}_+^{n\times n} \tag{1.3}$$

and set eigenvectors corresponding to the r largest eigenvalues of the solution K^* of (1.3) to be C. In (1.3), $\|K\|_*$ is the nuclear norm of K, $\|K\|_1 = \sum_{i,j=1}^n |K_{i,j}|$, and $\theta > 0, \beta > 0$. Under the assumption that A has an ONMF, some properties of the solution of (1.3) are established in [19]. Numerical results show Pan-Ng approach is better than some existing methods. Problem (1.3) is a nice convex approach for ONMF, but the solution of (1.3) may fail to satisfy $K^* = CC^\top$ and the eigenvectors corresponding to the r largest eigenvalues of K^* may have negative elements. Moreover, such approach arises an interesting question when A has an ONMF. To address this question and overcome difficulties in handling the orthogonal nonnegative constraints in ONMF, this paper presents new theoretical results in matrix analysis and an optimization algorithm for ONMF.

The main contributions of this paper are summarized as follows.

- We prove that a matrix $A \in \mathbb{R}_+^{m\times n}$ has an ONMF if and only if there is a matrix $Z \in \mathbb{R}^{n\times(n-r)}$ with $0 < r < n$ such that $AZ = 0$, $Z^\top Z = I$, $ZZ^\top \le I$.
- Based on the necessary and sufficient condition for A to have an ONMF, we propose an optimization model

$$\begin{aligned} &\min \ \frac{1}{2}\|AZ\|_F^2, \\ &\text{subject to } \ Z \in \Omega := \{Z \in \mathbb{R}^{n\times(n-r)} \,|\, Z^\top Z = I,\ ZZ^\top \le I\}. \end{aligned} \tag{1.4}$$

  Problem (1.4) has a nonempty and bounded solution set, and its optimal value is 0 if and only if A has an ONMF.
- We propose an augmented Lagrangian algorithm for solving (1.4) and prove the convergence of the algorithm to a stationary point. Moreover, we describe a simple method to obtain C and B from outputs of the algorithm by spectral decomposition. Preliminary numerical results are presented using synthetic and real-world datasets, which show our approach substantially improves some existing approaches in clustering.

The rest of the paper is organized as follows. In Section 2, we derive the sufficient and necessary conditions for a nonnegative matrix A to admit an ONMF and establish the error bound. Section 3 proposes an augmented Lagrangian (AL) algorithm to solve the model (1.4) and prove its convergence. In Section 4, we present numerical results to demonstrate the effectiveness of our approach and compare with other approaches.

Notation Let I represent the identity matrix of appropriate dimensions and $\mathbb{O}^{n\times r} = \{C \in \mathbb{R}^{n\times r} \,|\, C^\top C = I\}$. The null space of A is denoted by Null(A). We use $\|\cdot\|$ to denote the Frobenius norm $\|\cdot\|_F$ for matrices.

2. Characterization of ONMF.

This section delves into the theoretical framework for matrices that admit ONMF. We first present Lemma 2.1, which serves as a key tool in the proof of Theorem 2.2, where we provide the necessary and sufficient conditions for a nonnegative matrix to admit an ONMF. At the end of this section, we give a perturbation error bound. These results provide a solid foundation for the development of algorithms, which are further explored in the next section.

LEMMA 2.1. *Let $Q \in \mathbb{O}^{n\times r}$. If $QQ^\top \in \mathbb{R}_+^{n\times n}$, then there exists $C \in \mathbb{O}_+^{n\times r}$ such that*

$$CC^\top = QQ^\top. \tag{2.1}$$

Proof. We prove this lemma by mathematical induction on n.

If $n = 1$, then $QQ^\top = 1$ and $C = 1$.

Assume that this lemma holds for each $n \in \{1, \ldots, k\}$. We will prove it for $n = k + 1$. Let $S = QQ^\top \in \mathbb{R}^{(k+1)\times(k+1)}$. The eigenvalues of S are either 1 or 0.

Case 1: If S is irreducible, the Perron-Frobenius theorem [2] implies that 1 is a simple eigenvalue of S with a positive eigenvector v, while all other eigenvalues are 0. Note that $\operatorname{rank}(S) = 1$, which implies $r = 1$. Let $C = \frac{v}{\|v\|_2} \in \mathbb{O}_+^{(k+1)\times 1}$ be the normalized eigenvector. Then, S can be expressed as:

$$S = \begin{bmatrix} C & U \end{bmatrix} \operatorname{diag}(1, 0, \ldots, 0) \begin{bmatrix} C & U \end{bmatrix}^\top = CC^\top,$$

where $U \in \mathbb{O}^{(k+1)\times k}$ is the orthogonal complement of C.

Case 2: Suppose that S is reducible. By the definition of the reducible matrix, there is a permutation matrix P such that

$$P^\top SP = \begin{bmatrix} K_1 & K_{12} \\ 0 & K_2 \end{bmatrix}, \tag{2.2}$$

where K_i is an square matrix of order s_i and $s_1 + s_2 = k + 1$. Notice that S is symmetric, so we have $K_{12} = 0$. Observe that S is idempotent and nonnegative, it follows that K_1 and K_2 are also symmetric, idempotent and nonnegative, and thus the eigenvalues of each K_1 and K_2 are 1 or 0. Denote $r_1 = \operatorname{rank}(K_1)$ and $r_2 = \operatorname{rank}(K_2)$ with $r = r_1 + r_2$. By the spectral decomposition, it is evident that there exist $Q_1 \in \mathbb{O}^{s_1\times r_1}, Q_2 \in \mathbb{O}^{s_2\times r_2}$, such that $K_1 = Q_1Q_1^\top$ and $K_2 = Q_2Q_2^\top$.

If $r_1 r_2 > 0$, then the induction hypothesis ensures the existence of $C_1 \in \mathbb{O}_+^{s_1\times r_1}$ and $C_2 \in \mathbb{O}_+^{s_2\times r_2}$ such that

$$C_1C_1^\top = K_1, \quad C_2C_2^\top = K_2.$$

Hence this lemma holds with

$$C = P \begin{bmatrix} C_1 & 0 \\ 0 & C_2 \end{bmatrix} \in \mathbb{O}_+^{(k+1)\times r}.$$

If $r_1 = r$ and $r_2 = 0$, then we can justify this lemma by taking

$$C = P \begin{bmatrix} C_1 \\ 0 \end{bmatrix} \in \mathbb{O}_+^{(k+1)\times r}.$$

The similar result holds for $r_1 = 0$ and $r_2 = r$. The proof is complete. ∎

THEOREM 2.2. *A matrix $A \in \mathbb{R}_+^{m\times n}$ has an ONMF if and only if there exists $Z \in \mathbb{R}^{n\times(n-r)}$ with $0 < r < n$ such that*

$$AZ = 0, \quad Z^\top Z = I_{n-r}, \quad ZZ^\top \le I. \tag{2.3}$$

Proof. $\Leftarrow$: Assume there exists $Z \in \mathbb{R}^{n\times(n-r)}$ such that (2.3) holds. Let $Q \in \mathbb{O}^{n\times r}$ be the orthogonal complement of Z, then it yields $QQ^\top = I - ZZ^\top$. Since $ZZ^\top \le I$, it follows that $QQ^\top \in \mathbb{R}_+^{n\times n}$.

Following from Lemma 2.1, there exists $C \in \mathbb{O}_+^{n\times r}$ such that $CC^\top = QQ^\top$. Let $B = AC$. Then $B \in \mathbb{R}_+^{m\times r}$. Since $Z \in \text{Null}(A)$, we have:

$$BC^\top = ACC^\top = AQQ^\top = A(I - ZZ^\top) = A,$$

which implies that A admits an ONMF.

$\Rightarrow$: If A has an ONMF, then there exist a matrix $B \in \mathbb{R}_+^{m\times r}$ and a matrix $C \in \mathbb{O}_+^{n\times r}$ such that $A = BC^\top$. Take $Z \in \mathbb{R}^{n\times(n-r)}$ such that $[C, Z]$ is orthogonal. Then $AZ = BC^\top Z = 0$, $Z^\top Z = I_{n-r}$ and $ZZ^\top = I - CC^\top \leq I$. ∎

Remark 2.3. In Theorem 2.2, the matrix Z satisfying (2.3) may not be unique, but $ZZ^\top$ is unique if $\text{rank}(A) = r$, as it is the orthogonal projection onto the null space of A.

If $AZ = 0$ holds only for the zero matrix Z, then $A \in \mathbb{R}_+^{m\times n}$ has full column rank. It is straightforward to verify that A admits an ONMF of the form $A = BC^\top$, where $B = A$ and $C = I$. For the remainder of this paper, we only consider $r < n$ unless explicitly stated otherwise.

Example 2.1 Consider a matrix $A = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}$, which admits an ONMF with $B = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix}$ and $C^\top = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 \\ 0 & 0 & 1 \end{bmatrix}$. Both vectors $Z_1 = \begin{bmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & 0 \end{bmatrix}^\top$ and $Z_2 = \begin{bmatrix} -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 \end{bmatrix}^\top$ satisfy (2.3) in Theorem 2.2. Although $Z_1 \neq Z_2$, we have $Z_1 Z_1^\top = Z_2 Z_2^\top$.

Suppose A is the ground truth matrix, and $\tilde{A}$ is an observed matrix from the data with noise. Let Z and $\tilde{Z}$ be solutions of problem (1.4) with A and $\tilde{A}$, respectively. From Theorem 2.2, $AZ = 0$ and $\tilde{A}\tilde{Z} = 0$ if and only if A and $\tilde{A}$ have ONMF. The following theorem gives the error metric $\Delta M = \tilde{Z}\tilde{Z}^\top - ZZ^\top$. This error metric enables us to assess the accuracy of $\tilde{Z}$ in approximating Z, thereby providing a rigorous means to evaluate the performance of algorithms for ONMF in preserving the structural properties of the original matrix A.

THEOREM 2.4. *Let $\tilde{Z}$ be a solution of problem (1.4) with $\tilde{A} \in \mathbb{R}_+^{m\times n}$. Assume that A has an ONMF with $A = BC^\top$, $B \in \mathbb{R}_+^{n\times r}$ and $C \in \mathbb{O}_+^{n\times r}$. Set $Z \in \mathbb{R}^{n\times(n-r)}$ such that $[C, Z]$ is orthogonal, and let $M = ZZ^\top$, $\tilde{M} = \tilde{Z}\tilde{Z}^\top$. Define the difference $\Delta A = \tilde{A} - A$ and $\Delta M = \tilde{M} - M$. Then, the following bound holds:*

$$\|\Delta M\| \leq 2\sqrt{2}\|C^\top A^\dagger\|_2\|\Delta A\|. \tag{2.4}$$

In addition, if $\tilde{A}$ has an ONMF, then

$$\|\Delta M\| \leq \sqrt{2}\|C^\top A^\dagger \Delta A \tilde{M}\|. \tag{2.5}$$

Proof. Since $\tilde{Z}$ is in the feasible set Ω of problem (1.4), there is $\tilde{C} \in \mathbb{O}_+^{n\times r}$ such that $\tilde{M} = I - \tilde{C}\tilde{C}^\top$ by Lemma 2.1. Let $U = [C, C']$ and $V = [\tilde{C}, \tilde{C}']$ be $n \times n$ orthogonal matrices. Noting $M = I - CC^\top$. It is easy to see that $\|\Delta M\|^2 = \|\Delta MU\|^2 = \|\Delta MC\|^2 + \|\Delta MC'\|^2$ and $\|\Delta MC'\| = \|\tilde{C}^\top C'\|$. By CS Decomposition Theorem [18], we have $\|\tilde{C}'^\top C\| = \|\tilde{C}^\top C'\|$. Note that

$$\tilde{C}'^\top \Delta M C = \tilde{C}'^\top (CC^\top - \tilde{C}\tilde{C}^\top)C = \tilde{C}'^\top C,$$

we have

$$\|\tilde{C}'^{\top}C\| = \|\tilde{C}'^{\top}\Delta MC\|.$$

Therefore,

$$\|\Delta M\|^2 = \|\Delta MC\|^2 + \|\tilde{C}'^{\top}\Delta MC\|^2 \leq 2\|\Delta MC\|^2,$$

which implies that

$$\|\Delta M\| \leq \sqrt{2}\|\Delta MC\| = \sqrt{2}\|C^{\top}\Delta M\|.$$

Since A has an ONMF, we have $AM = AZZ^{\top} = 0$. Using $\tilde{A}(I - \tilde{C}\tilde{C}^{\top}) = \tilde{A}\tilde{Z}\tilde{Z}^{\top}$, we have

$$A\Delta M = A(\tilde{M} - M) = A\tilde{M}.$$

Hence, we obtain

$$C^{\top}\Delta M = B^{\dagger}A\Delta M = B^{\dagger}A\tilde{M},$$

which yields

$$\|\Delta M\| \leq \sqrt{2}\|B^{\dagger}A\tilde{M}\|. \tag{2.6}$$

Note that C is column orthonormal, by the properties of Moore-Penrose inverse, we have $C^{\dagger} = C^{\top}$, and thus $A^{\dagger} = (C^{\dagger})^{\top}B^{\dagger} = CB^{\dagger}$, which leads to $B^{\dagger} = C^{\top}A^{\dagger}$. In view of this and equation (2.6), we obtain

$$\begin{aligned}
\|\Delta M\| \leq \sqrt{2}\|C^{\top}A^{\dagger}A\tilde{M}\| &\leq \sqrt{2}\|C^{\top}A^{\dagger}\|_2(\|\Delta A\tilde{M}\| + \|\tilde{A}\tilde{M}\|) \\
&\leq \sqrt{2}\|C^{\top}A^{\dagger}\|_2(\|\Delta A\| + \|\tilde{A}\tilde{Z}\|) \\
&\leq \sqrt{2}\|C^{\top}A^{\dagger}\|_2(\|\Delta A\| + \|\tilde{A}Z\|) \\
&= \sqrt{2}\|C^{\top}A^{\dagger}\|_2(\|\Delta A\| + \|\Delta AZ\|) \\
&= \sqrt{2}\|C^{\top}A^{\dagger}\|_2(\|\Delta A\| + \|\Delta A\|),
\end{aligned}$$

where the fourth inequality uses that $\tilde{Z}$ is a solution of problem (1.4) with $\tilde{A} \in \mathbb{R}_{+}^{m\times n}$. Hence the bound (2.4) is obtained.

In addition, if $\tilde{A}$ has an ONMF, then $\tilde{A}\tilde{M} = 0$, which implies $A\tilde{M} = (\tilde{A} - \Delta A)\tilde{M} = -\Delta A\tilde{M}$, thus it obtains the bound (2.5). This completes the proof of the theorem. ∎

3. An augmented Lagrangian (AL) algorithm. In contrast to previous algorithms [8,20] to solve (1.1) or solve a convex approximation problem (1.3) as in [19], our approach is to solve (1.4) grounded in Theorem 2.2. The feasible region Ω of (1.4) guarantees that $Z \in \mathbb{O}^{n\times(n-r)}$ and $ZZ^{\top} \leq I$, as required by the conditions in Theorem 2.2. Moreover, it is straightforward to verify that Ω is non-empty, since $I_{n\times(n-r)}$ is a valid point, where $I_{n\times(n-r)}$ denotes the matrix formed by the first $(n-r)$ columns of the identity matrix I.

Since the feasible region is compact and the objective function is continuous, problem (1.4) admits a global minimum by the Weierstrass extreme value theorem. To facilitate the application of the AL algorithm, we reformulate problem (1.4) by using auxiliary variables. Let

$$\mathcal{W} = (X, Z) \in \mathbb{R}^{n\times n} \times \mathbb{R}^{n\times(n-r)}, \quad \mathbb{F} := \left\{\mathcal{W} \,|\, X \geq 0, Z^{\top}Z = I\right\}$$

and

$$c(\mathcal{W}) = X - I + ZZ^{\top}. \tag{3.1}$$

Problem (1.4) can be reformulated as:

$$\min_{\mathcal{W}\in\mathbb{F}} \frac{1}{2}\|AZ\|^2, \qquad \text{subject to } c(\mathcal{W}) = 0. \tag{3.2}$$

For any given penalty parameter $\rho > 0$ and Lagrange multiplier $\Lambda \in \mathbb{R}^{n\times n}$, the AL function for problem (3.2) is given by

$$\mathcal{L}(\mathcal{W}, \Lambda, \rho) := \frac{1}{2}\|AZ\|^2 + \frac{1}{2\rho}(\|\Lambda + \rho c(\mathcal{W})\|^2 - \|\Lambda\|^2). \tag{3.3}$$

At each outer iteration, for fixed Λ, ρ, we solve the following AL subproblem:

$$\min_{\mathcal{W}\in\mathbb{F}} \mathcal{L}(\mathcal{W}, \Lambda, \rho). \tag{3.4}$$

It can be shown that the following condition is a necessary condition [10, 23] for a local minimizer of problem (3.4).

$$H(\mathcal{W}, \Lambda, \rho) := \begin{pmatrix} (I - ZZ^\top)\left[A^\top A - 2\rho(I - X - \frac{1}{\rho}\Lambda)\right] Z \\ Z^\top Z - I \\ X - \left(I - ZZ^\top - \frac{1}{\rho}\Lambda\right)_+ \end{pmatrix} = 0, \tag{3.5}$$

where $(\cdot)_+ = \max\{\cdot, 0\}$ is the projection onto the nonnegative space. Let $\mathcal{W}^{\mathbf{feas}}$ be an feasible point of problem (3.2). For instance, $\mathcal{W}^{\mathbf{feas}} = (X^{\mathbf{feas}}, Z^{\mathbf{feas}})$, where $Z^{\mathbf{feas}} = I_{n\times(n-r)}$ is the first $n - r$ columns of the $n \times n$ identity matrix I, and $X^{\mathbf{feas}} = I - Z^{\mathbf{feas}}(Z^{\mathbf{feas}})^\top$.

Following the approach in [5], we present the AL algorithm for solving problem (3.2), as detailed in Algorithm 3.1.

3.1. Convergence Analysis. To establish the convergence properties of Algorithm 3.1, we first define a stationary point for problem (3.2). The convergence results for Algorithm 3.1 are then presented in Theorem 3.2.

DEFINITION 3.1 (Stationary point). *Let $\Omega_w = \{\mathcal{W} \in \mathbb{F} \mid c(\mathcal{W}) = 0\}$ be the feasible region of problem (3.2), we say $\mathcal{W}^* \in \Omega_w$ is a stationary point of (3.2) if there exists $\Lambda^* \in \mathbb{R}^{n\times n}$ such that*

$$(I - Z^*(Z^*)^\top)\left[\frac{1}{2}\nabla\|AZ\|^2\big|_{Z^*} + \nabla_Z\langle c(\mathcal{W}), \Lambda^*\rangle\big|_{Z^*}\right] = 0. \tag{3.11}$$

If $\mathcal{W}^*$ is a local minimizer of (3.2), then it is a stationary point of (3.2). See [3,10].

THEOREM 3.2. *Let $\{\mathcal{W}^k\}$ be the sequence generated by Algorithm 3.1 and assume that $\{\mathcal{W}^k\}$ is bounded* [1]. *Then, the following statements hold:*

(i) $c(\mathcal{W}^k) \to 0$ as $k \to \infty$, where $c(\cdot)$ is defined in (3.1).

(ii) Any accumulation point $\mathcal{W}^*$ of $\{\mathcal{W}^k\}$ is a stationary point of (3.2).

Proof. To prove statement (i), we consider the following two separate cases.

(i) Case (a): $\{\rho_k\}$ is bounded.

[1] The boundness of $\{\mathcal{W}^k\}$ is proved in Theorem 3.4

Algorithm 3.1 Augmented Lagrangian Method for Model (3.2)

Input: Observed data $A \in \mathbb{R}_+^{m\times n}$, rank r, $\rho_{\max}$, tolerance ε, max iteration it_{max}.

1: Choose $\Lambda^0 = \mathbf{0}_{n\times n} \in \mathbb{R}^{n\times n}$, $\rho_0 > 0, \gamma \in (1,\infty), \tau \in (0,1), \eta \in (0,1)$, a positive sequence $\{\epsilon_k\}$ with $\lim_{k\to\infty} \epsilon_k = 0$, and a constant

$$\Gamma \geq \max\{\|AZ^{\mathbf{feas}}\|, \mathcal{L}(\mathcal{W}^{\mathbf{feas}}, \Lambda^0, \rho_0)\}. \tag{3.6}$$

2: **for** $k = 0, 1, \cdots, \mathrm{it}_{max}$ **do**

3: Solve problem (3.4) with $\Lambda = \Lambda^k$ and $\rho = \rho_k$ to find an approximate stationary point $\mathcal{W}^k \in \mathbb{F}$ of (3.4) such that

$$\|H(\mathcal{W}^k, \Lambda^k, \rho_k)\| \leq \epsilon_k, \quad \mathcal{L}(\mathcal{W}^k, \Lambda^k, \rho_k) \leq \Gamma. \tag{3.7}$$

4: Set

$$\Lambda^{k+1} = \Lambda^k + \rho_k c(\mathcal{W}^k). \tag{3.8}$$

5: **if** $k > 0$ and

$$\|c(\mathcal{W}^k)\| \leq \eta \|c(\mathcal{W}^{k-1})\| \tag{3.9}$$

then

6: Set $\rho_{k+1} = \rho_k$

7: **else**

8: Set

$$\rho_{k+1} = \max\{\gamma\rho_k, \|\Lambda^{k+1}\|^{1+\tau}\}. \tag{3.10}$$

9: **end if**

10: **if** $\|c(\mathcal{W}^k)\| < \varepsilon$ or $\rho_{k+1} > \rho_{\max}$ **then**

11: **break**

12: **end if**

13: **end for**

Output: $\mathcal{W}^k$

From the updating scheme in Algorithm 3.1 and $\gamma > 1$, we can see that $\{\rho_k\}$ is updated by (3.10) only for finite times. It thus implies that there is $k_0 > 0$ such that (3.9) holds for all $k \geq k_0$, that is,

$$\|c(\mathcal{W}^k)\| \leq \eta \|c(\mathcal{W}^{k-1})\|, \qquad \forall k \geq k_0.$$

Together with $\eta \in (0,1)$, we obtain

$$\lim_{k\to\infty} \|c(\mathcal{W}^k)\| = 0, \tag{3.12}$$

which clearly implies that statement *(i)* holds.

Case (b): $\{\rho_k\}$ is unbounded.

By the updating scheme in Algorithm 3.1, we can observe that $\{\rho_k\}$ must be updated by (3.10) for infinite times. Let $\{\rho_{j_1}, \rho_{j_2}, \cdots\}$ denote all elements in $\{\rho_k\}$ that are updated by (3.10) and $\mathbb{J} = \{j_1, j_2, \ldots\}$ is arranged in increasing order. It then follows that $\rho_{j_\ell} \to \infty$ as $\ell \to \infty$ and

$$\rho_i = \rho_{j_\ell}, \quad j_\ell \leq i < j_{\ell+1}, \ \ell \geq 1, \tag{3.13}$$

$$\rho_{j_\ell} = \max\{\gamma\rho_{j_\ell - 1}, \|\Lambda^{j_\ell}\|^{1+\tau}\}, \quad \ell \geq 1. \tag{3.14}$$

Let $\underline{j}(k) := \max\{j \in \mathbb{J} \,|\, k \geq j\}$ for every $k \geq j_1$.

Next we claim that for every $k \geq j_1$,

$$\frac{\|\Lambda^k\|}{\rho_k} \leq \frac{\|\Lambda^{\underline{j}(k)}\|}{\rho_{\underline{j}(k)}} + \sum_{i=1}^{k-\underline{j}(k)} \|c(\mathcal{W}^{k-i})\|. \tag{3.15}$$

In fact, $k \geq \underline{j}(k)$ by the definition of $\underline{j}(k)$. Clearly, (3.15) holds with equality when $k = \underline{j}(k)$. Now suppose $k > \underline{j}(k)$, we have $\rho_{k-i} = \rho_k$ for $1 \leq i \leq k - \underline{j}(k)$. In view of this and (3.8), it follows

$$\begin{aligned} \frac{\|\Lambda^k\|}{\rho_k} = \frac{\|\Lambda^k\|}{\rho_{k-1}} &\leq \frac{\|\Lambda^{k-1}\|}{\rho_{k-1}} + \|c(\mathcal{W}^{k-1})\| \\ &\leq \cdots \leq \frac{\|\Lambda^{\underline{j}(k)}\|}{\rho_{\underline{j}(k)}} + \sum_{i=1}^{k-\underline{j}(k)} \|c(\mathcal{W}^{k-i})\|, \end{aligned} \tag{3.16}$$

which proves (3.15).

We then show that for every $k \geq j_1$,

$$\frac{\|\Lambda^k\|}{\rho_k} \leq \frac{\|\Lambda^{\underline{j}(k)}\|}{\rho_{\underline{j}(k)}} + \frac{1}{1-\eta}\|c(\mathcal{W}^{\underline{j}(k)})\|. \tag{3.17}$$

Indeed, it follows from the definition of $\underline{j}(k)$ and the updating scheme of $\{\rho_\ell\}$ that

$$\|c(\mathcal{W}^{k-i})\| \leq \eta\|c(\mathcal{W}^{k-i-1})\|, \quad \text{for } 1 \leq i < k - \underline{j}(k),$$

which leads to

$$\|c(\mathcal{W}^{k-i})\| \leq \eta^{k-\underline{j}(k)-i}\|c(\mathcal{W}^{\underline{j}(k)})\|, \quad \text{for } 1 \leq i \leq k - \underline{j}(k). \tag{3.18}$$

Then inequality (3.17) follows from (3.15) and the fact that

$$\sum_{\ell=0}^{k-\underline{j}(k)-1} \eta^\ell \leq \sum_{\ell=0}^{\infty} \eta^\ell = \frac{1}{1-\eta}.$$

In the following, we will show that

$$\lim_{k\to\infty} \|\Lambda^{\underline{j}(k)}\|/\rho_{\underline{j}(k)} = 0, \tag{3.19}$$

$$\lim_{k\to\infty} \|c(\mathcal{W}^{\underline{j}(k)})\| = 0. \tag{3.20}$$

Indeed, recall that $\rho_{j_\ell} \to \infty$ as $\ell \to \infty$. It thus follows that $\rho_{\underline{j}(k)} \to \infty$ as $k \to \infty$. By (3.10) and the definition of $\underline{j}(k)$, one has

$$\rho_{\underline{j}(k)} = \max\left\{\gamma\rho_{\underline{j}(k)-1}, \|\Lambda^{\underline{j}(k)}\|^{1+\tau}\right\}, \qquad k \geq j_1. \tag{3.21}$$

This yields

$$\|\Lambda^{\underline{j}(k)}\| \leq \rho_{\underline{j}(k)}^{\frac{1}{1+\tau}},$$

which implies that

$$\|\Lambda^{\underline{j}(k)}\|/\rho_{\underline{j}(k)} \leq (\rho_{\underline{j}(k)})^{-\frac{\tau}{1+\tau}}, \qquad k \geq j_1.$$

The relations (3.19) then follows from this and $\rho_{\underline{j}(k)} \to \infty$ as $k \to \infty$. Further, by the second inequality in (3.7) and the definition of the AL function (3.3), one has

$$\frac{1}{2}\|AZ^{\underline{j}(k)}\|^2 + \frac{\|\Lambda^{\underline{j}(k)} + \rho_{\underline{j}(k)} c(\mathcal{W}^{\underline{j}(k)})\|^2 - \|\Lambda^{\underline{j}(k)}\|^2}{2\rho_{\underline{j}(k)}} \leq \Gamma,$$

which leads to

$$\left\| c(\mathcal{W}^{\underline{j}(k)}) + \frac{\Lambda^{\underline{j}(k)}}{\rho_{\underline{j}(k)}} \right\|^2 \leq \frac{1}{\rho_{\underline{j}(k)}}[2\Gamma - \|AZ^{\underline{j}(k)}\|^2] + \frac{\|\Lambda^{\underline{j}(k)}\|^2}{\rho^2_{\underline{j}(k)}}. \tag{3.22}$$

Using this, (3.19), the boundedness of $\{\mathcal{W}^k\}$ and $\rho_{\underline{j}(k)} \to \infty$ as $k \to \infty$, one can see that

$$\lim_{k\to\infty} \left\| c(\mathcal{W}^{\underline{j}(k)}) + \frac{\Lambda^{\underline{j}(k)}}{\rho_{\underline{j}(k)}} \right\| = 0, \tag{3.23}$$

which together with (3.19) implies that the relation (3.20) holds. In view of (3.17), (3.19) and (3.20), we conclude that

$$\lim_{k\to\infty} \|\Lambda^k\|/\rho_k = 0. \tag{3.24}$$

By the same argument used to establish (3.22) for any k, we obtain

$$\left\| c(\mathcal{W}^k) + \frac{\Lambda^k}{\rho_k} \right\|^2 \leq \frac{1}{\rho_k}[2\Gamma - \|AZ^k\|^2] + \frac{\|\Lambda^k\|^2}{\rho_k^2}. \tag{3.25}$$

Together with (3.24), the boundedness of $\{\mathcal{W}^k\}$, and $\rho_k \to \infty$ as $k \to \infty$, we conclude that $c(\mathcal{W}^k) \to 0$ as $k \to \infty$. This completes the proof of statement *(i)*.

(ii) Since $\mathcal{W}^*$ is an accumulation point and by the continuity of $c(\cdot)$, we have $c(\mathcal{W}^*) = 0$. Moreover, given that $\{\mathcal{W}^k\} \subseteq \mathbb{F}$ and is bounded, the closedness of $\mathbb{F}$ ensures that $\mathcal{W}^* \in \mathbb{F}$, which implies that $\mathcal{W}^*$ is a feasible point.

Let $\{\mathcal{W}^k\}_{\mathcal{K}} \to \mathcal{W}^*$ for some subsequence $\mathcal{K}$. To show that $\mathcal{W}^*$ is a stationary point, we first establish that

$$\|[I - Z^k(Z^k)^\top](A^\top A + 2\Lambda^{k+1})Z^k\| \to 0 \quad \text{as} \quad k \to \infty,\ k \in \mathcal{K}. \tag{3.26}$$

Recall that the first relation in (3.7) holds for any positive sequence $\{\epsilon_k\}$. Letting $\epsilon_k \to 0$ as $k \to \infty$, it follows that $\{\epsilon_k\}_{\mathcal{K}} \to 0$ as well. Since $\{\mathcal{W}^k\} \subset \mathbb{F}$, one has $(I - Z^k(Z^k)^\top)Z^k = 0$. From (3.8), we have

$$\begin{aligned}
&[I - Z^k(Z^k)^\top][A^\top A - 2\rho_k(I - X^k - \frac{\Lambda^k}{\rho_k})]Z^k \\
=&[I - Z^k(Z^k)^\top][A^\top A + 2\rho_k(\frac{\Lambda^k}{\rho_k} + X^k - I + Z^k(Z^k)^\top)]Z^k \\
=&[I - Z^k(Z^k)^\top]\{A^\top A + 2[\Lambda^k + \rho_k c(\mathcal{W}^k)]\}Z^k \\
=&[I - Z^k(Z^k)^\top](A^\top A + 2\Lambda^{k+1})Z^k.
\end{aligned} \tag{3.27}$$

Then, we obtain (3.26) by the first inequality in (3.7).

Equation (3.26) implies that

$$[I - Z^k(Z^k)^\top](A^\top A + 2\Lambda^{k+1})Z^k \to 0, \quad \text{as } k \in \mathcal{K} \to \infty. \tag{3.28}$$

Note that $\operatorname{rank}(Z^k) = n - r$. Let $Q^k \in \mathbb{O}^{n\times r}$ be the orthogonal complement of Z, then $I - Z^k(Z^k)^\top = Q^k(Q^k)^\top$. Substituting this into equation (3.28), we obtain

$$Q^k(Q^k)^\top(A^\top A + 2\Lambda^{k+1})Z^k = Q^k(Q^k)^\top A^\top AZ^k + 2Q^k\bar{\Lambda}^k \to 0, \tag{3.29}$$

as $k \to \infty, k \in \mathcal{K}$, where $\bar{\Lambda}^k = (Q^k)^\top \Lambda^{k+1} Z^k$.

Given that Q^k and Z^k are column orthonormal, it follows that Q^k and Z^k are bounded. Multiplying $(Q^k)^\top$ to the left of (3.29) shows the boundedness of $\bar{\Lambda}^k$. Thus there exists $\mathcal{K}_1 \subset \mathcal{K}$ such that $\{Q^k\}_{\mathcal{K}_1} \to Q^*, \{Z^k\}_{\mathcal{K}_1} \to Z^*$ and $\{\bar{\Lambda}^k\}_{\mathcal{K}_1} \to \bar{\Lambda}^*$, where Q^*, Z^* are column orthonormal. Observe that

$$Q^*(Q^*)^\top = \lim_{k\in\mathcal{K}_1\to\infty} Q^k(Q^k)^\top = \lim_{k\in\mathcal{K}_1\to\infty} \left(I - Z^k(Z^k)^\top\right) = I - Z^*(Z^*)^\top.$$

It follows that

$$\begin{aligned}
&\lim_{k\in\mathcal{K}_1\to\infty} Q^k(Q^k)^\top A^\top AZ^k + 2Q^k\bar{\Lambda}^k \\
&= Q^*(Q^*)^\top A^\top AZ^* + 2Q^*\bar{\Lambda}^* \\
&= Q^*(Q^*)^\top A^\top AZ^* + 2Q^*(Q^*)^\top Q^*\bar{\Lambda}^*(Z^*)^\top Z^* \\
&= Q^*(Q^*)^\top (A^\top A + 2\Lambda^*)Z^* \\
&= \left(I - Z^*(Z^*)^\top\right)(A^\top A + 2\Lambda^*)Z^*,
\end{aligned} \tag{3.30}$$

where $\Lambda^* = Q^*\bar{\Lambda}^*(Z^*)^\top$. Hence $[\nabla_Z \langle c(\mathcal{W}), \Lambda^*\rangle]_{Z^*} = 2\Lambda^* Z^*$. We prove that $\mathcal{W}^*$ is a stationary point. ∎

3.2. A block coordinate descent algorithm for subproblem (3.4). In this subsection, we propose a block coordinate descent (BCD) algorithm for solving the subproblem (3.4) with fixed $\Lambda = \Lambda^k$ and $\rho = \rho_k$. The goal is to find an approximate stationary point $\mathcal{W}^k \in \mathbb{F}$ that satisfies the conditions in (3.7).

For simplicity, we use $\mathcal{W}_t$ for the iterations $\mathcal{W}_t^k$ of the BCD algorithm. To guarantee the validity of the second relation in (3.7), one can select the initial point of the BCD method as follows:

$$\mathcal{W}_0 = \begin{cases} \mathcal{W}^{\mathbf{feas}} & \text{if } k = 0 \text{ or } \mathcal{L}(\mathcal{W}^{k-1}, \Lambda, \rho) > \Gamma \\ \mathcal{W}^{k-1} & \text{otherwise,} \end{cases} \tag{3.31}$$

where $\mathcal{W}^{\mathbf{feas}}$ denotes a feasible point for problem (3.2). For instance, let $Z = I_{n\times(n-r)}$ represent the first $n - r$ columns of the identity matrix I. Then, set $X = I - ZZ^\top$.

The subproblems for Z and X are solved alternately by fixing the other variable at each step. Specifically, during the t-th iteration of the BCD algorithm, the updates for Z and X are carried out as follows:

$$\begin{aligned}
Z_{t+1} &\in \operatorname*{argmin}_{Z^\top Z = I} \mathcal{L}(X_t, Z, \Lambda, \rho) \\
&= \operatorname*{argmin}_{Z^\top Z = I} \operatorname{tr}\left\{ Z^\top \left[A^\top A - 2\rho(I - X_t - \frac{\Lambda}{\rho}) \right] Z \right\},
\end{aligned} \tag{3.32}$$

$$\begin{aligned}
X_{t+1} &= \operatorname*{argmin}_{X \geq 0} \mathcal{L}(X, Z_{t+1}, \Lambda, \rho) \\
&= \operatorname*{argmin}_{X \geq 0} \frac{1}{2}\|X - I + Z_{t+1}Z_{t+1}^\top + \frac{\Lambda}{\rho}\|^2.
\end{aligned} \tag{3.33}$$

To find a solution of problem (3.32), we introduce the following lemma.

LEMMA 3.3 (Corollary 4.3.39 in [11]). *Let $S \in \mathbb{R}^{n\times n}$ be symmetric, and suppose $1 \le \ell \le n$. Then*

$$\lambda_1(S) + \cdots + \lambda_\ell(S) = \min_{Z\in\mathbb{R}^{n\times\ell}, Z^\top Z=I} \mathrm{tr}(Z^\top S Z), \tag{3.34}$$

where $\lambda_1(S) \le \cdots \lambda_\ell(S) \le \cdots \le \lambda_n(S)$ are eigenvalues of S, and the minimum is achieved at a matrix Z whose columns are orthonormal eigenvectors associated with $\lambda_1, \cdots, \lambda_\ell$.

Lemma 3.3 gives a closed form solution of (3.32) by setting $S = A^\top A - 2\rho(I - X_t - \frac{\Lambda}{\rho})$. That is, Z_{t+1} can be constructed by the orthonormal eigenvectors associated with the $n - r$ smallest eigenvalues of S.

The subproblem (3.33) also has a closed-form solution, which are given by:

$$X_{t+1} = \max\{I - Z_{t+1}Z_{t+1}^\top - \frac{\Lambda}{\rho}, 0\}. \tag{3.35}$$

The BCD algorithm for subproblem (3.4) is summarized in Algorithm 3.2.

Algorithm 3.2 BCD algorithm to solve subproblem (3.4)

Input: Result from the $(k-1)_{th}$ step of Algorithm 3.1 $\mathcal{W}^{k-1}$, $\Lambda = \Lambda_k$, $\rho = \rho_k$

1: Set tolerance ϵ_k, max iteration it_{max}.
2: Choose the initial point $\mathcal{W}_0$ by (3.31).
3: **for** $t = 0, 1, \cdots, \mathrm{it}_{max}$ **do**
4: Update Z_{t+1} by Lemma 3.3 with $S = A^\top A - 2\rho(I - X_t - \frac{\Lambda}{\rho})$.
5: Update X_{t+1} by using (3.35).
6: **if** $\|H(\mathcal{W}_{t+1}, \Lambda, \rho)\| < \epsilon_k$ **then**
7: **break**
8: **end if**
9: **end for**

Output: $\mathcal{W}^k = (X_{t+1}, Z_{t+1})$.

To ensure the boundedness of $\mathcal{W}^k$ and that the conditions in (3.7) are satisfied within Algorithm 3.1, we demonstrate that these claims are achieved through the BCD algorithm outlined in Algorithm 3.2, as formalized in Theorem 3.4.

THEOREM 3.4. *Let $\{\mathcal{W}_t\}$ be a sequence generated by Algorithm 3.2. Then, the following statements hold:*

(i) The sequence $\{\mathcal{W}_t\}$ is bounded.

(ii) For any $\epsilon_k > 0$, there is a positive integer $\bar{t}$, such that for any $t \ge \bar{t}$, we have

$$\|H(\mathcal{W}_t, \Lambda, \rho)\| \le \epsilon_k, \quad \mathcal{L}(\mathcal{W}_t, \Lambda, \rho) \le \Gamma, \tag{3.36}$$

where Γ is defined in (3.6) within Algorithm 3.1.

Proof. (i) We first show $\Lambda = \Lambda^k \ge 0$, for any k. By using the update scheme of

Λ^k in (3.8) and X_{t+1} in (3.35), respectively, one has:

$$\begin{aligned}\Lambda^{k+1} =&\rho_k\left(X^k - I + Z^k(Z^k)^\top + \frac{\Lambda^k}{\rho_k}\right)\\ =&\rho_k\left(\max\{I - Z^k(Z^k)^\top - \frac{\Lambda^k}{\rho_k}, 0\} - I + Z^k(Z^k)^\top + \frac{\Lambda^k}{\rho_k}\right)\\ =&\rho_k \max\{-I + Z^k(Z^k)^\top + \frac{\Lambda^k}{\rho_k}, 0\} \geq 0,\end{aligned} \tag{3.37}$$

where the third equality follows $\max(a,0) - a = \max(-a,0)$. Thus $\Lambda^k \geq 0$, for any k. Notice that Z_{t+1} is updated over the feasible set in (3.32), i.e., $Z_{t+1}^\top Z_{t+1} = I$, which implies $\{Z_{t+1}\}$ is bounded.

Since $\frac{\Lambda}{\rho} \geq 0$, one can obtain

$$0 \leq X_{t+1} = \max\{I - Z_{t+1}Z_{t+1}^\top - \frac{\Lambda}{\rho}, 0\} \leq \max\{I - Z_{t+1}Z_{t+1}^\top, 0\}, \tag{3.38}$$

which implies $\{X_t\}$ is bounded. Consequently, we conclude that $\{\mathcal{W}_t\}$ is bounded.

Furthermore, since $\|\mathcal{W}_t\|_F^2 = \|Z_t\|_F^2 + \|X_t\|_F^2$, we obtain

$$\|\mathcal{W}_t\|_F^2 \leq (n-r) + 2(2n-r)^2,$$

where the bound is independent of both t and k. Noting that $\mathcal{W}^k = \lim_{t\to\infty} \mathcal{W}_t^k$, it follows that the sequence $\{\mathcal{W}^k\}$ generated by Algorithm 3.1 is uniformly bounded.

(ii) From the scheme of Algorithm 3.2, it is easy to verify that for any $t \geq 0$,

$$-\frac{\|\Lambda\|^2}{2\rho} \leq \mathcal{L}(\mathcal{W}_{t+1}, \Lambda, \rho) \leq \mathcal{L}(Z_{t+1}, X_t, \Lambda, \rho) \leq \mathcal{L}(\mathcal{W}_t, \Lambda, \rho) \leq \mathcal{L}(\mathcal{W}_0, \Lambda, \rho) \leq \Gamma. \tag{3.39}$$

This shows that $\mathcal{L}$ is bounded with respect to $\mathcal{W}_t$.

Now we prove the first inequality in (3.36). From the structure of Algorithm 3.2, we have

$$\begin{pmatrix} (I - Z_{t+1}Z_{t+1}^\top)\left[A^\top A - 2\rho(I - X_t - \frac{\Lambda}{\rho})\right] Z_{t+1} \\ Z_{t+1}^\top Z_{t+1} - I \\ X_{t+1} - \left(I - Z_{t+1}Z_{t+1}^\top - \frac{\Lambda}{\rho}\right)_+ \end{pmatrix} = 0. \tag{3.40}$$

Hence, we have

$$\|H(\mathcal{W}_{t+1}, \Lambda, \rho)\| \leq 2\rho\|X_t - X_{t+1}\|.$$

Since $\mathcal{L}$ is strongly convex with respect to X, there is a constant α such that

$$\begin{aligned}\|X_t - X_{t+1}\| &\leq \alpha\left[\mathcal{L}(Z_{t+1}, X_t, \Lambda, \rho) - \mathcal{L}(\mathcal{W}_{t+1}, \Lambda, \rho)\right]\\ &\leq \alpha\left[\mathcal{L}(\mathcal{W}_t, \Lambda, \rho) - \mathcal{L}(\mathcal{W}_{t+1}, \Lambda, \rho)\right].\end{aligned}$$

Since $\mathcal{L}(\mathcal{W}_t, \Lambda, \rho)$ is decreasing and bounded below, one has $\mathcal{L}(\mathcal{W}_t, \Lambda, \rho)$ converges. Thus $\mathcal{L}(\mathcal{W}_t, \Lambda, \rho) - \mathcal{L}(\mathcal{W}_{t+1}, \Lambda, \rho) \downarrow 0$, we can claim that there is $\bar{t} > 0$, such that for $t > \bar{t}$, the first inequality in (3.36) holds. ∎

Now, one can see $\mathcal{L}(\mathcal{W}^k, \Lambda, \rho) = \lim_{t\to\infty} \mathcal{L}(\mathcal{W}_t^{k-1}, \Lambda, \rho) \leq \Gamma$ which ensures that the second condition in (3.7) holds at every iteration.

3.3. Deriving B and C from output Z. From Theorem 3.2, an output Z of Algorithm 3.1 is in $\Omega = \{Z \in \mathbb{R}^{n\times(n-r)} \mid Z^\top Z = I, ZZ^\top \leq I\}$. We can use $Z \in \Omega$ to obtain $B \in \mathbb{R}_+^{m\times r}$ and $C \in \mathbb{O}_+^{n\times r}$ for ONMF of A. This can be achieved using the adjacency graph $\mathcal{G}$ that constructed from $K = I - ZZ^\top$ with r strongly connected components. The complete procedure is detailed in Algorithm 3.3, and its main ideas are outlined below:

(a) Since the adjacency graph $\mathcal{G}$ is sensitive to small perturbations in $K = I - ZZ^\top$, even negligible entries can connect otherwise disjoint subgraphs. To eliminate these artifacts, we use the bisection method [22] to threshold K by setting entries with absolute values below δ to zero, ensuring that the resulting graph $\mathcal{G}$ consists of exactly r strongly connected components, as described in lines 3–15 of Algorithm 3.3.

(b) Identify the r strongly connected components $\mathcal{G}_1, \ldots, \mathcal{G}_r$ of $\mathcal{G}$ [9], with the corresponding adjacency matrices by $K_1, \ldots, K_r$, where $K_i \in \mathbb{R}^{s_i\times s_i}$ and $\sum_{i=1}^r s_i = n$. These components can be computed using the `conncomp` function in MATLAB, see line 16 of Algorithm 3.3.

(c) Since $\mathcal{G}_i$ is strongly connected, its adjacency matrix K_i is irreducible [4]. Combined with the symmetry and idempotence of K_i, there exists a positive vector u_i such that $K_i = u_i u_i^\top$, $i = 1, \cdots, r$.

(d) Let π_i be the vertex set of $\mathcal{G}_i$, we have $K_i = K(\pi_i, \pi_i)$, where $\pi_i \subseteq \{1, \cdots, n\}$, $\cup_{i=1}^r \pi_i = \{1, \cdots, n\}$, and $\pi_i \cap \pi_j = \emptyset$ for $i \neq j$. The i-th column of C is obtained by normalizing the first column of K_i: $C(\pi_i, i) = \frac{u_i}{\|u_i\|_2}$, and setting $C(\pi_i^C, i) = 0$, where π_i^C is the complement of π_i [19]. This ensures that $C \in \mathbb{O}_+^{n\times r}$. Set $B = AC \in \mathbb{R}_+^{m\times r}$, we then obtain both factors B and C, as implemented in lines 17–20 of Algorithm 3.3.

4. Numerical Experiments. In this section, we evaluate the performance of our approach on both synthetic and real-world datasets. We denote the proposed model in Algorithm 3.1 as NS-ONMF. To ensure completeness and reproducibility, we specify the parameter initialization strategy for NS-ONMF as follows. In particular, the factor Z is set as $Z = I_{n\times(n-r)}$, i.e., the first $(n-r)$ columns of the identity matrix I_n. The penalty parameter ρ_0 is selected from $\{10^{-10}, 10^{-6}, 10^{-4}\}$ depending on the dataset type, where smaller values are used for near-ONMF synthetic data to allow gradual constraint enforcement, while larger values are used for real data to promote faster convergence under noise. The decay factor η, which controls the growth of the penalty, is chosen from $\{0.99, 0.995, 0.999\}$ to ensure stable updates. The update factor γ is chosen from the interval $(1, 2)$, with a typical value of 1.2; the parameter τ is taken from $(0, 1)$ with $\tau = 0.1$. In Algorithm 3.2, the maximum number of inner iterations is set to 10 for synthetic data to allow more accurate subproblem solutions, while for large-scale or noisy real datasets, a single inexact update is used to reduce computational cost without compromising performance. We further let $\Gamma = \max\{\|AZ^{\mathbf{feas}}\|, \mathcal{L}(\mathcal{W}^{\mathbf{feas}}, \Lambda^0, \rho_0)\}$ and $\epsilon_k = 1/k$. All experiments were conducted in MATLAB 2023b on an Intel(R) Xeon(R) Gold 5320 CPU @ 2.20GHz with 512GB of RAM, and the complete source code is publicly available on GitHub at https://github.com/Qilun-Luo/ONMF.

4.1. Synthetic Data. We will first test our approach with synthetic data.

4.1.1. A has an ONMF. Given a matrix A, our approach aims to minimize $\|AZ\|^2$ over the set Ω. We compare the performance with other ONMF methods, including BiOR-NM3F [8] (Bi-Orthogonal 3-Factor NMF), MU-ONMF [7] (Multi-

Algorithm 3.3 ONMF of A for deriving $B \in \mathbb{R}_+^{m\times r}$ and $C \in \mathbb{O}_+^{n\times r}$

Input: Null space $Z \in \mathbb{R}^{n\times(n-r)}$ obtained from Algorithm 3.1, factor r.

Initialize $\delta_{min} = 10^{-10}$, $\delta_{max} = 10^{-1}$, $\text{it}_{max} = 100$.
Set $K = I - ZZ^\top$, $\delta = (\delta_{min} + \delta_{max})/2$.
for $k = 1, \cdots, \text{it}_{max}$ **do**
  Set $\tilde{K} = K$,and let $\tilde{K}(\text{abs}(\tilde{K}) < \delta) = 0$.
  Set $\tilde{G} = \tilde{K}$, and let $\tilde{G}(\tilde{G} \neq 0) = 1$.
  Construct the graph $\mathcal{G}$ from $\tilde{G}$, and determine its number of strongly connected components n_c using the MATLAB function '`conncomp`'.
  if $n_c == r$ **then**
    break
  else if $n_c > r$ **then**
    $\delta_{max} = \delta$
  else
    $\delta_{min} = \delta$
  end if
  Compute $\delta = (\delta_{min} + \delta_{max})/2$.
end for
Find the irreducible components $K_1, \ldots, K_r$ and the corresponding vertex sets $\pi_1, \ldots, \pi_r$ of the graph $\mathcal{G}$ using the '`conncomp`' function.
for $i = 1, \cdots, r$ **do**
  Set $C(\pi_i, i) = \frac{u_i}{\|u_i\|_2}$ and $C(\pi_i^C, i) = 0$, where u_i is the first column of K_i.
end for
Compute $B = AC$.

Output: Factor matrices $B \in \mathbb{R}_+^{m\times r}$, $C \in \mathbb{O}_+^{n\times r}$.

plicative Updates on Stiefel Manifolds), EM-ONMF [20] (EM-like alternating algorithm), ONPMF [20] (Orthogonal Nonnegatively Penalized Matrix Factorization), SN-ONMF [19] (Sparsity and Nuclear norm minimization for ONMF). To ensure a fair comparison, the parameters for competing methods are set following their original recommendations. Particularly, the maximum number of iterations of the proposed method is 1000, and the stopping criteria are as follows: for BiOR-NM3F, MU-ONMF, HALS, and EM-ONMF, $\|C^k - C^{k-1}\| < \epsilon$, where $\epsilon = 10^{-8}$; for ONPMF, $\|\min(C^k, 0)\|/\|C^k\| < 10^{-3}$; for SN-ONMF, $\max([\|K^k - K^{k-1}\|, \|X^k - K^k\|, \|Z^k - K^k\|]) < \epsilon$, where K, X, Z are defined in model (13) of [19]; and for NS-ONMF, $\|c(\mathcal{W}^k)\| < \epsilon$ or $\rho_{k+1} > \rho_{\max}$.

We generate a matrix $A \in \mathbb{R}_+^{300\times 100}$ for three distinct cases. In Case 1, A is constructed as $A = BC^\top$, where $C \in \mathbb{O}_+^{100\times 10}$ has a block diagonal structure, and $B \in \mathbb{R}_+^{300\times 10}$ is a random matrix. In Case 2, A is generated similarly to Case 1, but the orthogonal factor C does not possess the block diagonal structure. In Case 3, $A = BC^\top + R$, where B and C are generated in the same manner as in Case 1, and the noise $R = 0.01 \times \text{rand}(300, 100)$.

To assess the performance of the comparison methods, we use the following evaluation metrics:

$$\text{(4.1)}\quad \text{err}(\tilde{C}) = \|\tilde{C}\tilde{C}^\top - CC^\top\|, \text{orth}(\tilde{C}) = \|\tilde{C}^\top\tilde{C} - I\|^2, \text{res}(\tilde{B}, \tilde{C}) = \frac{\|A_{gt} - \tilde{B}\tilde{C}^\top\|}{\|A_{gt}\|},$$

where $\tilde{B}, \tilde{C}$ are the factors computed by the algorithms and $A_{gt} = BC^\top$ denotes the

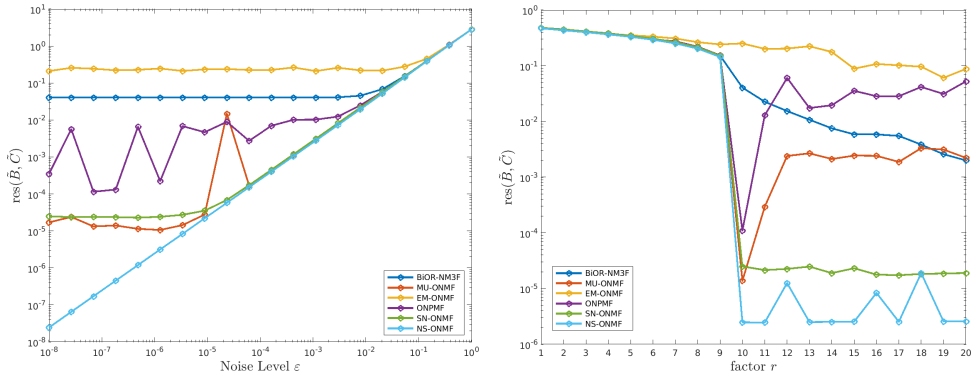
ground-truth orthogonal nonnegative decomposable matrix. All methods are run 10 times, and the average results are reported in Table 1.

TABLE 1
Results of different methods for finding C

Case #	Method	CPU	$\text{err}(\tilde{C})$	$\text{orth}(\tilde{C})$	$\text{res}(\tilde{B},\tilde{C})$
Case 1	BiOR-NM3F [8]	3.3850	9.6405×10^{-1}	9.6405×10^{-1}	7.2555×10^{-2}
	MU-ONMF [7]	0.6624	4.6774×10^{-4}	4.6074×10^{-4}	1.2026×10^{-5}
	EM-ONMF [20]	0.2079	4.2426×10^{-1}	4.5142×10^{-16}	4.3893×10^{-2}
	ONPMF [20]	0.9867	1.5283×10^{-3}	5.9523×10^{-15}	1.8239×10^{-4}
	SN-ONMF [19]	0.3765	2.2225×10^{-4}	3.6828×10^{-15}	2.6307×10^{-5}
	NS-ONMF	**0.0144**	$\mathbf{2.0297 \times 10^{-12}}$	$\mathbf{3.5108 \times 10^{-16}}$	$\mathbf{4.6018 \times 10^{-13}}$
Case 2	BiOR-NM3F [8]	3.8633	9.6695×10^{-1}	9.6695×10^{-1}	7.3526×10^{-2}
	MU-ONMF [7]	0.6414	4.6894×10^{-4}	4.6123×10^{-4}	1.5922×10^{-5}
	EM-ONMF [20]	0.0947	1.4142×10^{-1}	$\mathbf{3.7551 \times 10^{-16}}$	1.5074×10^{-2}
	ONPMF [20]	0.9535	1.5661×10^{-3}	4.6853×10^{-15}	1.9442×10^{-4}
	SN-ONMF [19]	0.3541	2.5372×10^{-5}	5.5830×10^{-15}	2.9005×10^{-6}
	NS-ONMF	**0.0049**	$\mathbf{4.5946 \times 10^{-12}}$	4.0030×10^{-16}	$\mathbf{1.0189 \times 10^{-12}}$
Case 3	BiOR-NM3F [8]	3.4089	9.8347×10^{-1}	9.8256×10^{-1}	8.0480×10^{-2}
	MU-ONMF [7]	0.6570	2.6336×10^{-2}	6.3920×10^{-3}	2.7990×10^{-2}
	EM-ONMF [20]	**0.0126**	2.3252×10^{0}	4.1919×10^{-16}	2.5274×10^{-1}
	ONPMF [20]	0.9625	3.0425×10^{-2}	6.3940×10^{-15}	2.7678×10^{-2}
	SN-ONMF [19]	0.6528	2.3917×10^{-2}	4.3275×10^{-15}	2.8026×10^{-2}
	NS-ONMF	2.3704	$\mathbf{1.9531 \times 10^{-2}}$	$\mathbf{2.2204 \times 10^{-16}}$	$\mathbf{2.5906 \times 10^{-2}}$

Table 1 reports the results of different methods for recovering the factor C across three synthetic cases. Both NS-ONMF and EM-ONMF achieve orthogonality up to machine precision (on the order of 10^{-16}), confirming the correctness of their formulations. Beyond orthogonality, NS-ONMF consistently attains the best reconstruction error and relative residual across all cases. In the noise-free settings (Cases 1 and 2), NS-ONMF is also the fastest method, converging within milliseconds while simultaneously achieving superior accuracy. This efficiency arises because the output produced by Algorithm 3.2 already yields factors close to the optimal solution of model (1.4), allowing Algorithm 3.1 to converge within only a few iterations. Moreover, Algorithm 3.2 employs closed-form updates for both subproblems and typically converges within ten iterations, further reducing runtime in clean data regimes. In the noisy setting (Case 3), NS-ONMF requires more runtime than some competitors, yet it still produces the lowest reconstruction error and residual, demonstrating greater robustness to noise. By contrast, methods such as EM-ONMF converge faster in noisy cases but at the expense of substantially larger reconstruction errors and residuals. These results highlight a clear trade-off: while several algorithms enforce orthogonality effectively, NS-ONMF achieves the most accurate and stable overall decomposition, offering both efficiency in noise-free regimes and robustness under noise.

In the following, we analyze the relationship between the noise level ε and the relative residual $\text{res}(\tilde{B},\tilde{C})$. Similar to Case 3, where $A = BC^\top + R$, we set $R = \varepsilon \times \text{rand}(300, 100)$, with $\varepsilon = \text{logspace}(-8, 0, 20)$, meaning ε is generated between 10^{-8} and 1 on a logarithmic scale with 20 points. Figure 1(a) shows that the proposed method outperforms the competitors, especially for small noise levels. This also explains why, in Case 1 and Case 2 (noise-free cases), the method NS-ONMF requires less CPU time, as lower noise leads to higher accuracy and fewer iterations.



(a) Residuals vs. noise level ε (b) Residuals vs. factor r

FIG. 1. *Residual* res($\tilde{C}$) *comparisons across noise levels (a) and factor values (b).*

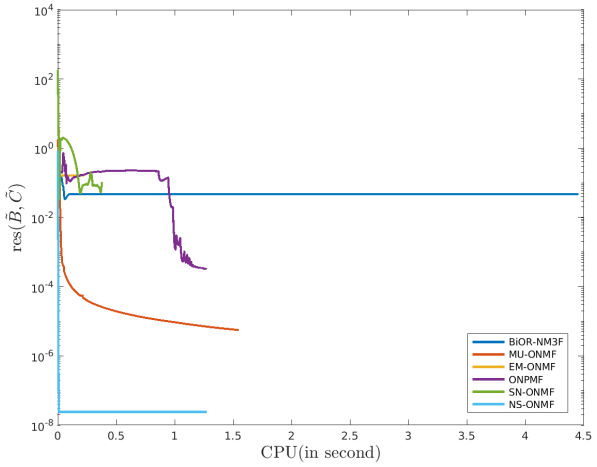


FIG. 2. *Relative residual error vs. CPU time*

Next, we examine the impact of the factor r on the residual res($\tilde{B},\tilde{C}$). Let $A \in \mathbb{R}_+^{300\times 100}$ be generated as in Case 3, where the true rank of A is 10. However, we do not assume the rank of A is known and instead vary the factor r from 1 to 20 to evaluate how the methods handle different values of r. From Figure 1(b), we observe that for $r < \text{rank}(A)$, all methods perform similarly, with the residual near 1. When $r = \text{rank}(A) = 10$, most methods improve their performance, with the proposed NS-ONMF achieving the best outcome, reaching a residual of 10^{-5}. For $r > \text{rank}(A)$, the performance of the methods slightly deteriorates, but NS-ONMF still maintains a residual of 10^{-5}. This result further demonstrates that our model does not require the rank of A to be known, in accordance with Theorem 2.2.

In this part, we examine the efficiency and convergence behavior of the compared methods, we construct the synthetic matrix $A = BC^\top + R$ using Case 3 with noise level 10^{-8}. Figure 2 plots the relative residual error res($\tilde{B},\tilde{C}$) against CPU time for all algorithms. This runtime-based comparison provides a more direct illustration of convergence efficiency. The results show that NS-ONMF rapidly decreases the residual

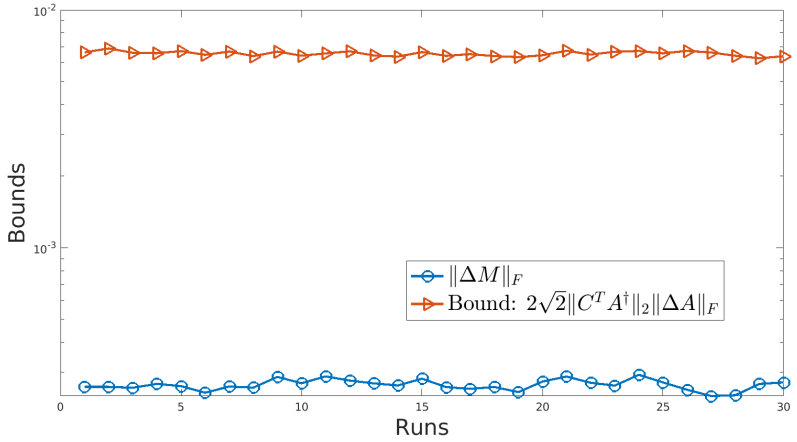


FIG. 3. *Error bounds computed by Theorem 2.4*

error, reaching nearly 10^{-8} within a fraction of a second. Although MU-ONMF also decreases steadily, it requires more runtime and does not achieve the same accuracy. ONPMF and SN-ONMF stagnate at higher error levels, while BiOR-NM3F maintains a relatively large error throughout. EM-ONMF is the fastest method though it attains a higher residual compared with NS-ONMF. These observations confirm that NS-ONMF is both effective and efficient: it converges quickly in terms of CPU time and consistently attains the smallest residual error among all tested methods.

To further validate Theorem 2.4, we construct a ground-truth orthogonal nonnegative matrix decomposition $A = BC^\top$, where $C \in \mathbb{O}_+^{100\times 10}$ and $B \in \mathbb{R}_+^{300\times 10}$. We then perturb A by defining $\tilde{A} = A + R$, with $R = 10^{-4} \times \text{rand}(300, 100)$. Applying the proposed method to $\tilde{A}$ yields the factors $\tilde{B}, \tilde{C}, \tilde{Z}$. We evaluate the deviation

$$\Delta M = ZZ^\top - \tilde{Z}\tilde{Z}^\top = CC^\top - \tilde{C}\tilde{C}^\top,$$

and compare the theoretical bound in (2.4) with the empirical value $\|\Delta M\|_F$. This experiment is repeated 30 times to ensure robustness. The results, shown in Figure 3, confirm that the empirical deviations are consistent with the theoretical bounds established in Theorem 2.4.

4.1.2. Synthetic clustering data. In this part, we generate the synthetic clustering data similar to [20]. Each dataset consists of six clusters, denoted as $\{\pi_i\}_{i=1}^6$, with each cluster containing $100 - (i-1) \times 10$ data points, for a total of 450 data points. Each cluster centroid $u_i \in \mathbb{R}^{1000}$ for $1 \le i \le 6$ is generated uniformly at random within the unit cube $[0, 1]^{1000}$. Each data point m_j for $1 \le j \le 450$ is a scalar multiple of its corresponding cluster centroid, given by $m_j = \alpha u_k$, where α is chosen uniformly at random from the interval $[0.2, 1]$. The data points are then perturbed by adding noise drawn from a normal distribution with mean zero and standard deviation ε (negative entries are set to zero). Finally, A is constructed by collecting these data points into a matrix of size 1000×450.

To obtain the clusters using ONMF methods, we utilize the nonnegative orthonormal factor C to determine the cluster assignments. Each row of C serves as a feature vector for a sample, and the column index of the positive entry in each row indicates the cluster to which the sample belongs. However, since some methods may yield multiple positive entries in a row, we assign the cluster based on the index of the

maximum entry. To evaluate the clustering results of the compared algorithms, we use accuracy as the performance metric. We examine the clustering performance of the proposed NS-ONMF method against six existing ONMF methods: BiOR-NM3F [8], MU-ONMF [7], HALS [21], EM-ONMF [20], ONPMF [20], and SN-ONMF [19].

The noise level ε is selected from the interval $[10^{-2}, 1]$ with a total of 50 points. Figure 4 presents the average accuracy of each method across different noise levels over 10 runs. The results indicate that the proposed NS-ONMF method and ONPMF consistently achieve the best performance. Both methods successfully identify the true clusters across a wide range of noise levels, demonstrating their robustness and effectiveness in clustering tasks.

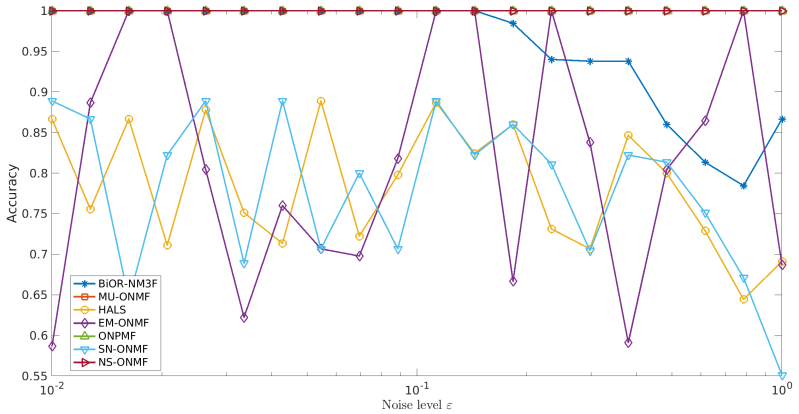
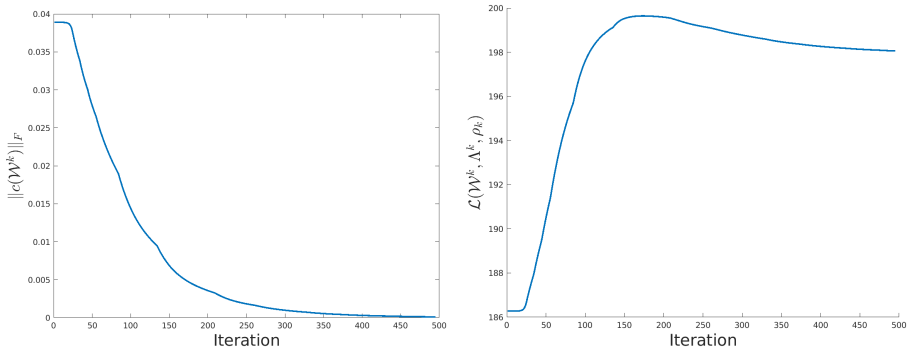


FIG. 4. *Comparison of ONMF methods on synthetic clustering data with varying noise levels.*

To examine the convergence behavior of Algorithm 3.1, we set the noise level $\varepsilon = 0.1$ and generate synthetic clustering data as previously described. We plot $c(\mathcal{W}^k)$ and $\mathcal{L}(\mathcal{W}^k, \Lambda^k, \rho_k)$ in Figure 5. The first plot, Figure 5(a), shows that $c(\mathcal{W}^k)$ steadily decreases and converges to zero, consistent with the results presented in Theorem 3.2. This convergence indicates that the algorithm is progressively satisfying the constraints. The second plot, Figure 5(b), shows that $\mathcal{L}(\mathcal{W}^k, \Lambda^k, \rho_k)$ remains bounded throughout the process. We note that in this experiment the noise level is relatively large ($\varepsilon = 0.1$), which explains the higher iteration counts compared with the small-noise cases (e.g., $\varepsilon = 10^{-8}$). For small-noise settings, the Figure 2 demonstrates that Algorithm 3.1 converges within only a few iterations and rapidly reduces the residual error to the order of 10^{-7}. By contrast, Figure 5 highlights that even under larger noise Algorithm 3.1 still converges in the sense that $c(\mathcal{W}^k) \to 0$, confirming robustness. We further emphasize that the augmented Lagrangian does not exhibit a monotone convergence trend because the penalty parameter ρ^k is adaptively increased (see Algorithm 3.1, line 8 and Eq. (3.10)), which may cause $\mathcal{L}(\mathcal{W}^k, \Lambda^k, \rho^k)$ to increase even when the iterates approach a stationary point. Nevertheless, the stopping criterion $\|c(\mathcal{W}^k)\| < \epsilon$ is theoretically justified, since $\|c(\mathcal{W}^k)\| \to 0$ implies that $\mathcal{W}^k$ converges to a stationary point of problem (3.2) by Theorem 3.2. This bounded behavior confirms the algorithm's stability under the given conditions, and the overall results support the effectiveness of Algorithm 3.1 in converging to a solution that satisfies the imposed constraints while maintaining stability in the objective function.



(a) $c(\mathcal{W}^k)$ (b) $\mathcal{L}(\mathcal{W}^k, \Lambda^k, \rho_k)$

FIG. 5. *Convergence analysis of Algorithm 3.1*

4.2. Real-world Data. In this section, we will apply the proposed NS-ONMF method to real-world datasets, focusing on clustering. We compare the performance of NS-ONMF with six other ONMF methods: BiOR-NM3F [8], MU-ONMF [7], HALS [21], EM-ONMF [20], ONPMF [20], and SN-ONMF [19]. To evaluate the effectiveness of each method, we also use four metrics [16]: Purity, NMI, MIhat, and Accuracy. These metrics provide a comprehensive assessment of the clustering quality. To ensure comparisons focus on problem (1.1), clustering results are derived from the orthogonal nonnegative factor $\tilde{C}$, the computed solution of (1.1). Hard clustering is applied across all methods, assigning each sample to the cluster with the highest probability given by the factor matrix $\tilde{C}$.

We will evaluate the ONMF methods on real-world data clustering tasks, which include face clustering, document clustering, and object clustering. Each category comprises three well-known datasets, resulting in a total of nine datasets. The summary of these datasets is provided in Table 2.

TABLE 2
Summary of clustering datasets

Dataset	Type	Features (m)	Samples (n)	Classes (K)
PIE_pose27	Face	1024	2856	68
Yale	Face	1024	165	15
ORL	Face	1024	400	40
bbcsport	Document	4613	737	5
tr11	Document	6429	414	9
tr12	Document	5804	313	8
seeds	Object	7	210	3
wine	Object	13	173	3
glass	Object	9	214	6

Table 3 reports the performance of several ONMF methods on the *PIE_pose27*, *Yale*, and *ORL* face-clustering datasets. Across all three datasets, NS-ONMF consistently achieves the best clustering quality, attaining the highest Purity, NMI, MIhat, and Accuracy. In particular, on the challenging *PIE_pose27* dataset, NS-ONMF reaches a Purity of 0.8207 and an Accuracy of 0.7990, outperforming all competing

methods by a clear margin. Similar improvements are observed on *Yale* and *ORL*, where NS-ONMF delivers better NMI and MIhat values, indicating more coherent and informative cluster structures. Figure 6, which displays the basis images for the *PIE_pose27* dataset, further supports these findings. The basis images generated by NS-ONMF are significantly clearer and more distinguishable than those produced by other methods, indicating NS-ONMF's superior ability to capture essential clustering features. Although NS-ONMF incurs a higher computational cost, this is offset by its substantial improvements in clustering accuracy and the quality of the basis images.

TABLE 3
Results of different methods for face clustering

Dataset	Method	CPU	Purity	NMI	MIhat	Accuracy
PIE_pose27	BiOR-NM3F [8]	58.6085	0.3442	0.5843	0.5580	0.2773
	MU-ONMF [7]	15.9989	0.2868	0.4616	0.3919	0.2202
	HALS [21]	15.8625	0.1299	0.3239	0.2617	0.1278
	EM-ONMF [20]	**2.6473**	0.2721	0.5401	0.5322	0.2465
	ONPMF [20]	57.2348	0.1635	0.3228	0.2979	0.1488
	SN-ONMF [19]	156.2487	0.7122	**0.8911**	0.8672	0.6835
	NS-ONMF	422.0823	**0.8207**	0.8785	**0.8704**	**0.7990**
Yale	BiOR-NM3F [8]	0.6950	0.3818	0.4291	0.4120	0.3697
	MU-ONMF [7]	**0.4034**	0.3576	0.4154	0.3844	0.3091
	HALS [21]	0.7649	0.2606	0.3069	0.3110	0.2364
	EM-ONMF [20]	0.1460	0.3455	0.4375	0.4195	0.3394
	ONPMF [20]	2.7489	0.3091	0.3386	0.2962	0.3030
	SN-ONMF [19]	0.6933	0.3455	0.4028	0.3950	0.3333
	NS-ONMF	2.4502	**0.4909**	**0.5305**	**0.5253**	**0.4848**
ORL	BiOR-NM3F [8]	1.8026	0.4775	0.6495	0.6058	0.4225
	MU-ONMF [7]	1.0387	0.5350	0.7188	0.6773	0.4725
	HALS [21]	3.8143	0.1475	0.3815	0.3030	0.1425
	EM-ONMF [20]	**0.4160**	0.5275	0.7040	0.6770	0.4675
	ONPMF [20]	8.3403	0.1375	0.3167	0.2240	0.1325
	SN-ONMF [19]	2.9851	0.5200	0.6841	0.6801	0.4800
	NS-ONMF	8.2375	**0.7000**	**0.8048**	**0.7878**	**0.6500**

Table 4 compares the performance of ONMF methods on document clustering across the *bbcsport*, *tr11*, and *tr12* datasets. In the *bbcsport* dataset, NS-ONMF achieves the highest Purity (0.9457), NMI (0.8485), MIhat (0.8473), and Accuracy (0.9457), slightly outperforming ONPMF, which also attains competitive clustering quality. Figure 7 illustrating the *bbcsport* clustering results reinforces these metrics, showing that NS-ONMF produces well-separated clusters that closely reflect the ground truth. NS-ONMF consistently delivers superior separation and clarity, particularly in more challenging datasets like *tr11* and *tr12*, making it the most robust method for document clustering among those evaluated.

Table 5 presents a comparison of various ONMF methods applied to object clustering tasks, specifically for the *seeds*, *wine*, and *glass* datasets. NS-ONMF emerges as the most effective method, delivering the highest values in Purity, NMI, MIhat, and Accuracy across all datasets. In the *seeds* dataset, for example, NS-ONMF achieves a standout Purity of 0.8810 and an NMI of 0.6518, demonstrating its robustness in differentiating between object classes. Figure 8 depicting the *seeds* dataset clustering further illustrates NS-ONMF's capability to delineate distinct and coherent clusters,

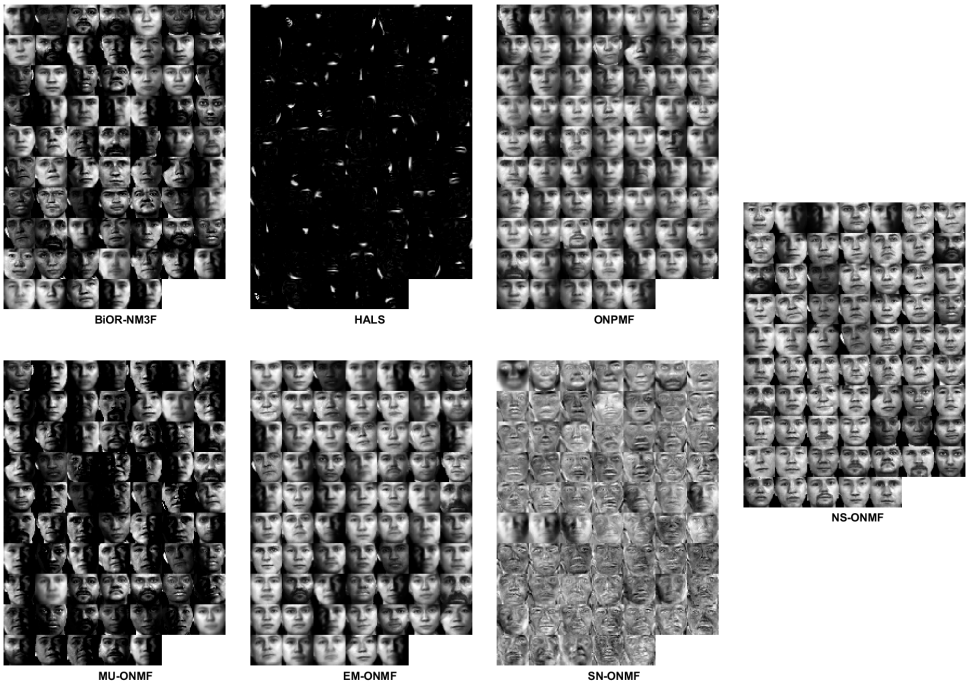


FIG. 6. *Basis images generated by different ONMF methods for the PIE_pose27 Dataset.*

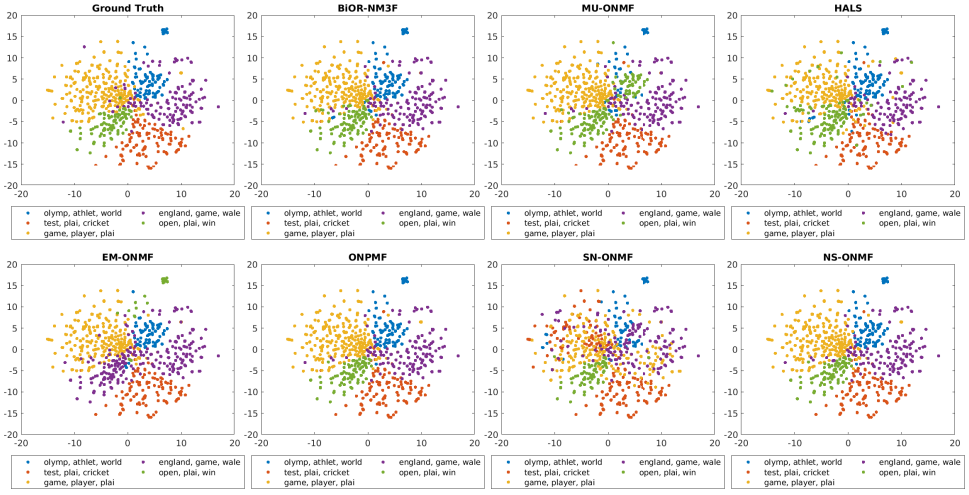


FIG. 7. *Document clustering by ONMF methods for the bbcsport Dataset.*

closely mirroring the ground truth distribution.

For these real-world datasets, it is unlikely that the data exhibit a near-ONMF structure; instead, they are more accurately modeled as $A = BC^\top + R$, where R represents a substantial noise component. Under such conditions, NS-ONMF generally requires more iterations to converge, and the increased runtime should not be interpreted as the sole reason for its improved clustering performance. As illustrated in Figure 2, several competing methods (e.g., BiOR-NM3F, MU-ONMF, and ONPMF)

TABLE 4
Results of different methods for document clustering

Dataset	Method	CPU	Purity	NMI	MIhat	Accuracy
bbcsport	BiOR-NM3F [8]	6.7560	0.9199	0.7901	0.7847	0.9199
	MU-ONMF [7]	6.1448	0.8412	0.7034	0.6966	0.8412
	HALS [21]	3.8879	0.8602	0.6834	0.6771	0.8602
	EM-ONMF [20]	**0.9328**	0.8100	0.7226	0.6896	0.7788
	ONPMF [20]	12.6639	0.9444	0.8424	0.8421	0.9444
	SN-ONMF [19]	10.3098	0.6418	0.4667	0.4653	0.6418
	NS-ONMF	29.1887	**0.9457**	**0.8485**	**0.8473**	**0.9457**
tr11	BiOR-NM3F [8]	7.9396	0.7029	0.5655	0.5273	0.5097
	MU-ONMF [7]	9.6800	0.6957	0.5267	0.4946	0.4638
	HALS [21]	7.6255	0.6184	0.4841	0.4780	0.5072
	EM-ONMF [20]	**0.6722**	0.7343	0.5848	0.5430	0.4783
	ONPMF [20]	23.6050	0.7367	0.5762	0.5410	0.4734
	SN-ONMF [19]	3.8672	0.4324	0.2928	0.2783	0.3478
	NS-ONMF	9.1732	**0.7488**	**0.6173**	**0.5812**	**0.6256**
tr12	BiOR-NM3F [8]	5.2975	0.6869	0.5247	0.5106	0.5495
	MU-ONMF [7]	5.7893	0.7125	0.5576	0.5463	0.6070
	HALS [21]	5.1425	0.4984	0.3510	0.3268	0.4153
	EM-ONMF [20]	**1.3501**	0.5623	0.4556	0.4246	0.4505
	ONPMF [20]	13.7598	0.6454	0.5304	0.5219	0.5495
	SN-ONMF [19]	2.2722	0.3834	0.2040	0.2029	0.3259
	NS-ONMF	4.8724	**0.7157**	**0.6106**	**0.6006**	**0.6134**

consume comparable or even greater CPU times, yet their residual errors stagnate at much higher levels, resulting in weaker clustering outcomes. This confirms that clustering quality is determined by the quality of the feasible solution rather than by runtime alone. Moreover, subproblem (3.32) requires the computation of $n-r$ eigenpairs, which becomes computationally expensive for large n (as in document and face datasets). This further contributes to the higher CPU times observed in Tables 3-5. From another perspective, the computed matrix $Z \in \mathbb{R}^{n\times(n-r)}$ captures additional informative structure by including a larger set of basis directions. Furthermore, Algorithm 3.3 guarantees that C is both nonnegative and orthogonal, a property not strictly enforced in some competing methods (e.g., SN-ONMF). These structural advantages enhance the quality of both Z and C, which in turn explain the consistently superior clustering performance of NS-ONMF. The additional runtime observed in noisy or large-scale datasets should therefore be regarded as a reasonable trade-off for obtaining more accurate and reliable clustering results.

5. Conclusions. This paper advances the theoretical framework of ONMF by characterizing properties of the null space of a nonnegative matrix $A \in \mathbb{R}^{m\times n}$ for the existence of ONMF. We prove that A has an ONMF if and only if there is a matrix $Z \in \mathbb{R}^{n\times(n-r)}$ such that $AZ = 0$, $Z^\top Z = I$ and $ZZ^\top \le I$. The new necessary and sufficient condition provides solid mathematical foundation to study optimization model (1.4) for ONMF. We propose an augmented Lagrangian algorithm for solving (1.4) and prove the convergence of the algorithm. Numerical evaluations on both synthetic and real-world datasets demonstrate the efficacy of the proposed methods,

TABLE 5
Results of different methods for object clustering

Dataset	Method	CPU	Purity	NMI	MIhat	Accuracy
seeds	BiOR-NM3F [8]	0.1996	0.7429	0.4272	0.4252	0.7429
	MU-ONMF [7]	0.0445	0.6190	0.3396	0.3112	0.6190
	HALS [21]	0.0463	0.6667	0.5543	0.4420	0.6667
	EM-ONMF [20]	**0.0043**	0.8000	0.5095	0.5082	0.8000
	ONPMF [20]	0.1725	0.4571	0.0437	0.0430	0.4571
	SN-ONMF [19]	0.7741	0.5571	0.1499	0.1495	0.5571
	NS-ONMF	0.7995	**0.8810**	**0.6518**	**0.6508**	**0.8810**
wine	BiOR-NM3F [8]	0.1719	0.6348	0.3791	0.3764	0.5843
	MU-ONMF [7]	0.0518	0.6348	0.3711	0.3488	0.6348
	HALS [21]	0.0581	0.4045	0.0081	0.0142	0.4045
	EM-ONMF [20]	**0.0103**	0.6966	0.3823	0.3762	0.6966
	ONPMF [20]	0.2531	0.4607	0.0352	0.0340	0.4045
	SN-ONMF [19]	0.6889	0.4775	0.0771	0.0771	0.4775
	NS-ONMF	0.8507	**0.7022**	**0.3857**	**0.3803**	**0.7022**
glass	BiOR-NM3F [8]	0.2234	0.4907	0.1895	0.1805	0.3692
	MU-ONMF [7]	0.0683	0.5374	0.2804	0.2736	0.4252
	HALS [21]	0.0859	0.3598	0.0111	0.0293	0.3131
	EM-ONMF [20]	**0.0109**	0.5467	0.3096	0.3055	0.4673
	ONPMF [20]	0.2499	0.3598	0.0308	0.0245	0.3084
	SN-ONMF [19]	0.7417	0.4579	0.2652	0.2538	0.4252
	NS-ONMF	2.2538	**0.5701**	**0.3976**	**0.3615**	**0.5374**

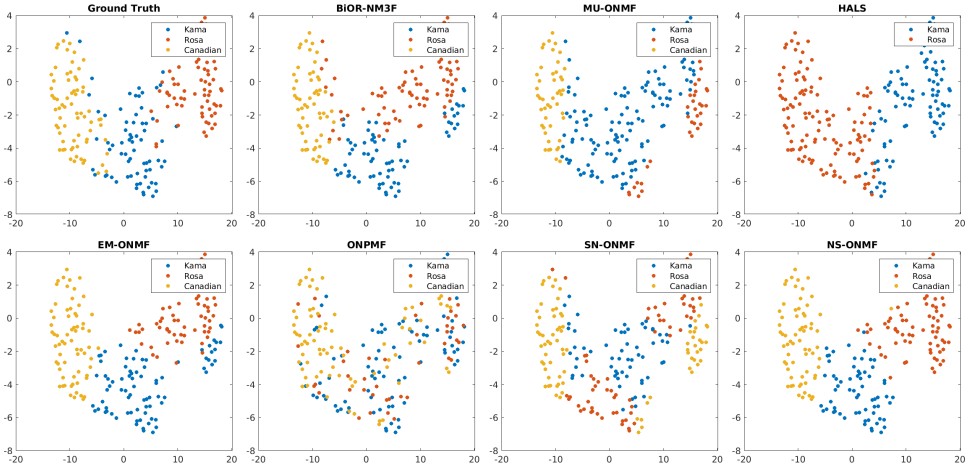


FIG. 8. *Object clustering by ONMF methods for the seeds Dataset.*

showcasing substantial improvements in clustering. These findings not only validate the robustness and accuracy of our approach but also underscore its utility in enhancing interpretability in practical applications.

Acknowledgement. The authors are grateful to the two anonymous reviewers for their valuable comments and for verifying the implementation of the code, and to

Prof. Delin Chu and Prof. Zaikun Zhang for their insightful suggestions.

REFERENCES

[1] M. Asteris, D. Papailiopoulos, and A. G. Dimakis, *Orthogonal NMF through subspace exploration*, Adv. Neural Inf. Process. Syst., 28 (2015).

[2] A. Berman and R. J. Plemmons, *Nonnegative Matrices in the Mathematical Sciences*, SIAM, 1994.

[3] E. G. Birgin and J. M. Martínez, *Practical Augmented Lagrangian Methods for Constrained Optimization*, SIAM, 2014.

[4] R. A. Brualdi, H. J. Ryser, et al., *Combinatorial matrix theory*, vol. 39, Springer, 1991.

[5] X. Chen, L. Guo, Z. Lu, and J. J. Ye, *An augmented Lagrangian method for non-Lipschitz nonconvex programming*, SIAM J. Numer. Anal., 55 (2017), pp. 168–193.

[6] X. Chen, Y. He, and Z. Zhang, *Tight error bounds for the sign-constrained Stiefel manifold*, SIAM J. Optim., 35 (2025), pp. 302–329.

[7] S. Choi, *Algorithms for orthogonal nonnegative matrix factorization*, in 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), 2008, pp. 1828–1832, https://doi.org/10.1109/IJCNN.2008.4634046.

[8] C. Ding, T. Li, W. Peng, and H. Park, *Orthogonal nonnegative matrix t-factorizations for clustering*, in Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, 2006, pp. 126–135.

[9] L. K. Fleischer, B. Hendrickson, and A. Pinar, *On identifying strongly connected components in parallel*, in Parallel and Distributed Processing: 15 IPDPS 2000 Workshops Cancun, Mexico, May 1–5, 2000 Proceedings 14, Springer, 2000, pp. 505–511.

[10] B. Gao, X. Liu, X. Chen, and Y.-X. Yuan, *A new first-order algorithmic framework for optimization problems with orthogonality constraints*, SIAM J. Optim., 28 (2018), pp. 302–332.

[11] R. A. Horn and C. R. Johnson, *Matrix Analysis*, Cambridge University Press, 2012.

[12] B. Jiang, X. Meng, Z. Wen, and X. Chen, *An exact penalty approach for optimization with nonnegative orthogonality constraints*, Math. Program., 198 (2023), pp. 855–897.

[13] K. Kimura, Y. Tanaka, and M. Kudo, *A fast hierarchical alternating least squares algorithm for orthogonal nonnegative matrix factorization*, in Asian Conference on Machine Learning, PMLR, 2015, pp. 129–141.

[14] B. Li, G. Zhou, and A. Cichocki, *Two efficient algorithms for approximately orthogonal nonnegative matrix factorization*, IEEE Signal Process. Lett., 22 (2014), pp. 843–846.

[15] W. Li, J. Li, X. Liu, and L. Dong, *Two fast vector-wise update algorithms for orthogonal nonnegative matrix factorization with sparsity constraint*, J. Comput. Appl. Math., 375 (2020), p. 112785.

[16] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*, vol. 39, Cambridge University Press, 2008.

[17] A. Mirzal, *A convergent algorithm for orthogonal nonnegative matrix factorization*, J. Comput. Appl. Math., 260 (2014), pp. 149–166.

[18] C. C. Paige and M. A. Saunders, *Towards a generalized singular value decomposition*, SIAM J. Numer. Anal., 18 (1981), pp. 398–405.

[19] J. Pan and M. Ng, *Orthogonal nonnegative matrix factorization by sparsity and nuclear norm optimization*, SIAM J. Matrix Anal. Appl., 39 (2018), pp. 856–875.

[20] F. Pompili, N. Gillis, P.-A. Absil, and F. Glineur, *Two algorithms for orthogonal nonnegative matrix factorization with application to clustering*, Neurocomputing, 141 (2014), pp. 15–25.

[21] M. Shiga, K. Tatsumi, S. Muto, K. Tsuda, Y. Yamamoto, T. Mori, and T. Tanji, *Sparse modeling of EELS and EDX spectral imaging data by nonnegative matrix factorization*, Ultramicroscopy, 170 (2016), pp. 43–59.

[22] K. Sikorski, *Bisection is optimal*, Numer. Math., 40 (1982), pp. 111–117.

[23] N. Xiao, X. Liu, and Y. Yuan, *Exact penalty function for 2,1 norm minimization over the stiefel manifold*, SIAM J. Optim., 31 (2021), pp. 3097–3126.