On the clustering aspect of various nonnegative matrix factorization objectives

Andri Mirzal and Masashi Furukawa
Graduate School of Information Science and Technology,
Hokkaido University, Kita 14 Nishi 9, Kita-Ku,
Sapporo 060-0814, Japan

Abstract: The clustering aspect of various nonnegative matrix factorization (NMF) objectives which include standard NMF, orthogonal NMF, sparse NMF, Semi-NMF, and Convex NMF have been reported in many papers. However, there is still no comprehensive study that provides a theoretical explanation on this aspect yet. In this work, we provide such explanation by showing that at the respective stationary points in nonnegative orthant of the feasible regions, which are the solutions pursued by NMF algorithms, the NMF objectives are equivalent to the graph clustering objective, therefore the clustering aspect of NMF has a solid justification.

Keywords: bound-constrained optimization, clustering method, nonnegative matrix factorization, Karush-Kuhn-Tucker conditions.

1 Introduction

NMF is a matrix approximation technique that factorizes a nonnegative matrix into a pair of other nonnegative matrices of much lower rank:

$$\mathbf{A} \approx \mathbf{BC}, \tag{1}$$

where $\mathbf{A} \in \mathbb{R}_+^{M\times N} = [\mathbf{a}_1, \ldots, \mathbf{a}_N]$ denotes the data matrix, $\mathbf{B} \in \mathbb{R}_+^{M\times K} = [\mathbf{b}_1, \ldots, \mathbf{b}_K]$ denotes the basis matrix, $\mathbf{C} \in \mathbb{R}_+^{K\times N} = [\mathbf{c}_1, \ldots, \mathbf{c}_N]$ denotes the coefficient matrix, and K denotes the number of factors which usually is chosen so that $K \ll \min(M, N)$. Note that the definitions of $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$ above are chosen to simplify the interpretations of NMF.

To compute $\mathbf{B}$ and $\mathbf{C}$, usually eq. 1 is rewritten into a minimization problem in Frobenius norm criterion.

$$\min_{\mathbf{B},\mathbf{C}} J(\mathbf{B}, \mathbf{C}) = \frac{1}{2}\|\mathbf{A} - \mathbf{BC}\|_F^2 \text{ s.t. } \mathbf{B} \geq \mathbf{0}, \mathbf{C} \geq \mathbf{0}. \tag{2}$$

In addition to the usual Frobenius norm, the family of Bregman divergences—which Frobenius norm and Kullback-Leibler divergence are part of it—can also be used as the affinity measures. Detailed discussion on the Bregman divergences for NMF can be found in [1].

1.1 Local and holistic interpretations

There are other methods to decompose a nonnegative rectangular matrix, e.g., singular value decomposition (SVD) and QR decomposition. But, NMF is particularly interesting because it allows each data vector to be represented as a linear combination of the basis vectors:

$$\mathbf{a}_n \approx c_{1n}\mathbf{b}_1 + \cdots + c_{Kn}\mathbf{b}_K, \ \forall n, \tag{3}$$

where c_{kn} is the k-th entry of $\mathbf{c}_n$. As shown in eq. 3, the basis vectors can be thought as either basic building components (every data vector constructed from $\mathbf{b}_k, \forall k$) or shared features (all data vectors constructed from the same set $\mathcal{B} \in \{\mathbf{b}_1, \ldots, \mathbf{b}_K\}$) of the data. The basic building components viewpoint leads to the part-based interpretation which is a popular NMF property due to the work of Lee and Seung [2] and then verified by others, e.g., [3, 4, 5, 6, 7, 8]. The shared features viewpoint leads to the holistic representation of the data which first studied by Li et al. [3] and then verified by Hoyer [4].

As reported in [3, 4], NMF can only produce either local representation (part-based) or holistic representation. In short, if the data is well-aligned then NMF will produce sparse basis matrix which associated with the part-based interpretation, and if the data is not well-aligned then NMF will produce dense basis matrix which associated with the holistic representation. The mechanisms for adjusting the sparseness of the basis matrix are then introduced to ensure NMF be able to give the part-based interpretation [3, 4, 8] which is an important application in image processing. And a theoretical work that gives necessary conditions for NMF to be able to correctly identify the parts of objects is provided by Donoho and Stodden [9].

1.2 Clustering interpretation

In addition to the part-based interpretation, Lee and Seung [2] also show that NMF has clustering interpretation by utilizing it to extract topics from a document corpus. The clustering aspect of NMF then is investigated further by Xu et al. [10] followed by others, e.g.,

[11, 12, 13, 14, 15, 16]. This aspect is intuitive since objective in eq. 2 can be rewritten into:

$$\min_{\mathbf{b}_k,\mathbf{c}_n} \sum_{n=1}^{N}\sum_{k=1}^{K} c_{kn}\|\mathbf{a}_n - \mathbf{b}_k\|^2, \tag{4}$$

which is the objective of K-means clustering with the coefficient c_{kn} denotes the degree of membership of the data vector $\mathbf{a}_n$ to the cluster centroid $\mathbf{b}_k$.

1.3 Research background

The part-based, holistic, and clustering interpretations are the direct results of NMF formulation which are not found in other matrix decomposition techniques (at least not directly). Different from the part-based interpretation in which NMF can fail to give meaningful results [3, 4, 17], the clustering aspect seems to be the most stable and powerful property as so far apparently there is no work that disputes it and there are numerous works that show NMF and its variants are superior methods compared to, e.g., spectral clustering [10] and K-means clustering [12, 13, 14, 15].

However, a comprehensive theoretical work for supporting the clustering aspect of NMF seems to be overlooked. Perhaps because unlike the part-based interpretation where a counterexample is immediately presented [3] (so that motivating researchers to build both theoretical conditions [9] and practical frameworks [3, 4] for NMF to be able to give such interpretation), no counterexample has been presented to disprove the clustering aspect of NMF yet. So far the best approaches to explain this aspect are by

1. showing the equivalence of standard NMF objective in eq. 2 to K-means clustering objective in eq. 4 [12, 13, 14, 15, 16], and

2. applying zero gradient conditions to some NMF objectives [13, 14, 15, 16] to show their equivalences to graph clustering objective, i.e., *ratio association*.

The problem with the first approach is there is no obvious way to incorporate the nonnegativity constraints (and other auxiliary constraints, e.g., orthogonality, sparsity, and convexity constraints) into the K-means objective. And the problem with the second approach is it discards the nonnegativity constraints, thus is equivalent to finding stationary points on the respective unbounded feasible regions. Thus, NMF which is a bound-constrained optimization turns into an unbounded optimization, and consequently there is no guarantee the stationary points that being utilized to prove the equivalences are located in the nonnegative orthant.

In this work, we attempt to provide a theoretical support for the clustering aspect of NMF—specifically for NMF objectives that are reported to have clustering capabilities which include standard NMF, orthogonal NMF, sparse NMF, Semi-NMF, and Convex NMF—by analyzing the objectives at the respective stationary points. The stationary points are important in proving the clustering aspect of NMF objectives because

1. local and global optima which are the solutions pursued by NMF algorithms must be stationary points,

2. NMF algorithms can only guarantee the stationary of the solutions, and

3. the strict Karush-Kuhn-Tucker (KKT) optimality conditions can be utilized to derive the objectives at the respective stationary points.

We will show that at the stationary points, those NMF objectives are equivalent to the relaxed *ratio association* objective (see [18] for details about various graph clustering objectives), therefore the clustering aspect of NMF has a solid justification. Note that in deriving the equivalences, unlike previous works [13, 14, 15, 16], we will not set the Lagrange multipliers to zeros. Thus, the stationary points under investigation are guaranteed to be located in nonnegative orthant of the corresponding feasible regions.

2 Limit points of the sequences generated by NMF algorithms

All NMF algorithms are formulated in alternating fashion, fixing one matrix while solving the other (the popular Lee and Seung multiplicative update algorithms [19] and their derivatives [3, 4, 10, 14, 15, 20] also use alternating strategy, but cannot be represented by generic algorithm below). This strategy is employed because NMF is nonconvex with respect to $\mathbf{B}$ and $\mathbf{C}$, but is convex with respect to $\mathbf{B}$ or $\mathbf{C}$ [21]. Thus, the alternating strategy transforms NMF problem into a convex optimization. The modification of a nonconvex problem into corresponding convex problem is a common practice in optimization researches because (1) convex optimization is more tractable, (2) usually convex methods are more efficient, (3) any local optimum is necessarily a global optimum, and (4) the algorithms are easy to initialize [22].

The following generic algorithm describes the alternating fashion for solving NMF which will generate a solution sequence $\{\mathbf{B}^{(l)}, \mathbf{C}^{(l)}\}_{l=0}^{L}$:

$$\mathbf{B}^{(l+1)} = \arg\min_{\mathbf{B}\geq\mathbf{0}} \frac{1}{2}\|\mathbf{A} - \mathbf{B}\mathbf{C}^{(l)}\|_F^2 \tag{5}$$

$$\mathbf{C}^{(l+1)} = \arg\min_{\mathbf{C}\geq\mathbf{0}} \frac{1}{2}\|\mathbf{A} - \mathbf{B}^{(l+1)}\mathbf{C}\|_F^2, \tag{6}$$

where in eq. 5, $\mathbf{C}$ is kept constant while solving for $\mathbf{B}$, and in eq. 6, $\mathbf{B}$ is kept constant while solving for $\mathbf{C}$. This generic algorithm is known as alternating nonnegativity-constrained least square (ANLS) algorithm, and usually are solved by decomposing each subproblem into nonnegativity-constrained least square (NNLS) problems where there are many algorithms that guarantee the global-optimality of the NNLS problem.

$$\hat{\mathbf{b}}_m^{T(l+1)} = \underset{\hat{\mathbf{b}}_m^T \geq \mathbf{0}}{\arg\min} \frac{1}{2}\|\hat{\mathbf{a}}_m^T - \mathbf{C}^{T(l)}\hat{\mathbf{b}}_m^T\|_F^2, \quad \forall m \quad (7)$$

$$\mathbf{c}_n^{(l+1)} = \underset{\mathbf{c}_n \geq \mathbf{0}}{\arg\min} \frac{1}{2}\|\mathbf{a}_n - \mathbf{B}^{(l+1)}\mathbf{c}_n\|_F^2, \quad \forall n, \quad (8)$$

where $\hat{\mathbf{x}}_i$ is the i-th row of matrix $\mathbf{X}$.

According to Grippo and Sciandrone [23] any limit point of the sequence $\{\mathbf{B}^{(l)}, \mathbf{C}^{(l)}\}_{l=0}^{L}$ generated by ANLS algorithms that optimally solve the convex subproblem eq. 5 and eq. 6 is a stationary point. And such ANLS based NMF algorithms exist, e.g., [6, 7, 20, 21, 24, 25], therefore there is guarantee that we can reach the stationary points on the feasible region for NMF problem in eq. 2. And as NNLS is the building block for ANLS, any NNLS algorithm that guarantees to find optimal solutions of eq. 7 and eq. 8, e.g., [26, 27, 28] can also be employed to search the stationary points for NMF problem in eq. 2.

And as will be shown in section 4, NMF objectives implicitly put upper bounds on the feasible regions (the lower bounds are explicit: the nonnegativity constraints), thus NMF is a bound-constrained optimization problem, consequently $\{\mathbf{B}^{(l)}, \mathbf{C}^{(l)}\}_{l=0}^{L}$ has at least one limit point [21]. This completes the conditions for any NMF algorithm that optimally solves subproblem eq. 5 and eq. 6 to be able to find a stationary point in the nonnegative orthant of the feasible region which is the necessary condition for our proofs on the clustering aspect of NMF.

3 Some issues in solving NMF objectives using NMF algorithms

The discussion on stationary of the limit points in section 2 is only for the standard NMF objective in eq. 2. Because we aim to explain the clustering aspect of the standard NMF, sparse NMF, orthogonal NMF, Semi-NMF, and Convex NMF, there is a need to verify that for each NMF objective at least one algorithm exists to guarantee the stationary of the limit points.

For the standard NMF, as stated previously, there are many algorithms that guarantee the convergence, e.g., [6, 7, 20, 21, 24, 25]. For sparse NMF, there also exists such algorithms, e.g., [24, 25] (there also exists sparse NMF algorithms that do not guarantee the convergence, e.g., sparse NMF by Hoyer [4], local NMF [3], ALS [5], CNMF [8], GD-CLS [11], and ACLS/AHCLS [29]).

But unfortunately, so far there is no algorithm for orthogonal NMF, Semi-NMF, and Convex NMF that guarantee to reach the stationary points. The reason is because algorithms for these objectives are all based on multiplicative update rules which as shown numerically by Gonzales and Zhang [30], proved by Lin [20], and explained qualitatively by Berry et al. [5], multiplicative update based algorithms can only guarantee the stationary of limit points in the interior of the corresponding feasible regions, and when the limit points lie on the boundary of the feasible regions, their stationary can not be determined.

Thus, in appendix we propose algorithms for uni-orthogonal NMF, bi-orthogonal NMF, Semi-NMF, and Convex NMF which based on additive update rules that has been proven by Lin [20] to have convergence property. Note that these additive update based algorithms have more works per iteration than their multiplicative counterparts which are known to have slow convergence. Hence, whenever possible, the same framework as shown in [24, 25] should be used to recast the auxiliary constraints into the ANLS framework for allowing more efficient (and converged) NMF algorithms be employed.

4 Clustering aspect of NMF

In this section a theoretical support for various NMF objectives that are reported to have clustering aspect is provided. We utilize the strict KKT optimality conditions to investigate the objectives at the corresponding stationary points in nonnegative orthant of the feasible regions. Unlike previous approaches where Lagrange multipliers are set to nulls [13, 14, 15, 16], we make no assumption about Lagrange multipliers, thus the stationary points are guaranteed to be in nonnegative orthant where the solutions of NMF problems should be located.

Further we also show that the upper bounds, which is a necessary condition for guaranteeing the existence of limit point of sequence $\{\mathbf{B}^{(l)}, \mathbf{C}^{(l)}\}_{l=0}^{L}$, is implicitly imposed by each NMF objective.

And for interpretability reason, $\mathbf{A}$ is considered as feature-by-item data matrix for the rest of this paper, where feature and item correspond to row and column respectively.

4.1 Standard NMF

In this subsection we prove that applying the standard NMF to $\mathbf{A}$ leads to the clustering of similar items and related features as reported in, e.g., [2, 7, 10, 12, 13, 14, 29, 31].

Theorem 1. *Minimizing the following objective*

$$\min_{\mathbf{B},\mathbf{C}} J_a(\mathbf{B},\mathbf{C}) = \frac{1}{2}\|\mathbf{A}-\mathbf{B}\mathbf{C}\|_F^2 \quad (9)$$
$$\text{s.t. } \mathbf{B}\geq\mathbf{0}, \mathbf{C}\geq\mathbf{0},$$

leads to the feature clustering indicator matrix $\mathbf{B}$ *and the item clustering indicator matrix* $\mathbf{C}$.

Proof.

$$\|\mathbf{A}-\mathbf{B}\mathbf{C}\|_F^2 = \text{tr}\,(\mathbf{A}^T\mathbf{A} - 2\mathbf{C}\mathbf{A}^T\mathbf{B} + \mathbf{B}^T\mathbf{B}\mathbf{C}\mathbf{C}^T).$$

Thus, minimizing J_a is equivalent to simultaneously optimizing:

$$\max_{\mathbf{B},\mathbf{C}} \text{tr}\,(\mathbf{C}\mathbf{A}^T\mathbf{B}) \quad (10)$$
$$\min_{\mathbf{B},\mathbf{C}} \text{tr}\,(\mathbf{B}^T\mathbf{B}\mathbf{C}\mathbf{C}^T). \quad (11)$$

Note that because $\text{tr}\,(\mathbf{X}\mathbf{Y}) \leq \text{tr}\,(\mathbf{X})\text{tr}\,(\mathbf{Y})$, minimizing Eq. 11 is equivalent to:

$$\min_{\mathbf{B}} \text{tr}\,(\mathbf{B}^T\mathbf{B}) \text{ and} \quad (12)$$
$$\min_{\mathbf{C}} \text{tr}\,(\mathbf{C}\mathbf{C}^T). \quad (13)$$

The Lagrangian of objective in eq. 9 is:

$$L_a(\mathbf{B},\mathbf{C}) = J_a(\mathbf{B},\mathbf{C}) - \text{tr}\,(\mathbf{\Gamma_B}\mathbf{B}^T) - \text{tr}\,(\mathbf{\Gamma_C}\mathbf{C}), \quad (14)$$

where $\mathbf{\Gamma_B} \in \mathbb{R}_+^{M\times K}$ and $\mathbf{\Gamma_C} \in \mathbb{R}_+^{N\times K}$ are the Lagrange multipliers. By applying the KKT conditions to L_a we get:

$$\nabla_{\mathbf{B}} L_a = \mathbf{B}\mathbf{C}\mathbf{C}^T - \mathbf{A}\mathbf{C}^T - \mathbf{\Gamma_B} = \mathbf{0} \quad (15)$$
$$\nabla_{\mathbf{C}} L_a = \mathbf{B}^T\mathbf{B}\mathbf{C} - \mathbf{B}^T\mathbf{A} - \mathbf{\Gamma_C}^T = \mathbf{0}, \quad (16)$$

with complementary slackness:

$$\mathbf{\Gamma_B} \otimes \mathbf{B} = \mathbf{0}, \text{ and } \mathbf{\Gamma_C}^T \otimes \mathbf{C} = \mathbf{0},$$

where $\otimes$ denotes component-wise multiplications. Eq. 15 and eq. 16 lead to:

$$\mathbf{B} = (\mathbf{A}\mathbf{C}^T + \mathbf{\Gamma_B})(\mathbf{C}\mathbf{C}^T)^{-1} \quad (17)$$
$$\mathbf{C} = (\mathbf{B}^T\mathbf{B})^{-1}(\mathbf{B}^T\mathbf{A} + \mathbf{\Gamma_C}^T). \quad (18)$$

Substituting eq. 18 to eq. 10 leads to:

$$\max_{\mathbf{B}} \text{tr}\,\big((\mathbf{B}^T\mathbf{B})^{-1}(\mathbf{B}^T\mathbf{A}\mathbf{A}^T\mathbf{B} + \mathbf{\Gamma_C}^T\mathbf{A}^T\mathbf{B})\big), \quad (19)$$

which is equivalent to simultaneously optimizing:

$$\max_{\mathbf{B}} \text{tr}\,(\mathbf{B}^T\mathbf{A}\mathbf{A}^T\mathbf{B}) \quad (20)$$
$$\max_{\mathbf{B}} \text{tr}\,(\mathbf{\Gamma_C}^T\mathbf{A}^T\mathbf{B}) \quad (21)$$
$$\min_{\mathbf{B}} \text{tr}\,(\mathbf{B}^T\mathbf{B}). \quad (22)$$

Similarly, substituting eq. 17 to eq. 10 leads to:

$$\max_{\mathbf{C}} \text{tr}\,\big((\mathbf{C}\mathbf{A}^T\mathbf{A}\mathbf{C}^T + \mathbf{C}\mathbf{A}^T\mathbf{\Gamma_B})(\mathbf{C}\mathbf{C}^T)^{-1}\big), \quad (23)$$

which is equivalent to simultaneously optimizing:

$$\max_{\mathbf{C}} \text{tr}\,(\mathbf{C}\mathbf{A}^T\mathbf{A}\mathbf{C}^T) \quad (24)$$
$$\max_{\mathbf{C}} \text{tr}\,(\mathbf{C}\mathbf{A}^T\mathbf{\Gamma_B}) \quad (25)$$
$$\min_{\mathbf{C}} \text{tr}\,(\mathbf{C}\mathbf{C}^T). \quad (26)$$

As shown, eq. 22 and eq. 26 recover eq. 12 and eq. 13 respectively, so there is no need to substituting eq. 17 and eq. 18 into eq. 11.

Now we concentrate on the basis matrix $\mathbf{B}$ first. Eq. 20 – 22 give the conditions that must be satisfied by $\mathbf{B}$ at the stationary point. Note that if we consider $\mathbf{A}$ to be affinity matrix induced from bipartite graph $\mathcal{G}(\mathbf{A})$ (which is a reasonable thought since any feature-by-item matrix can be modeled by a bipartite graph), then $\mathcal{G}(\mathbf{A}\mathbf{A}^T)$ is the feature graph where edge weights describe the similarity between corresponding vertex pairs. So, eq. 20 looks like *ratio association* objective applied to $\mathcal{G}(\mathbf{A}\mathbf{A}^T)$. But without orthogonality constraint $\mathbf{B}^T\mathbf{B} = \mathbf{I}$ (part of *ratio association* objective), one can optimize eq. 20 by setting $\mathbf{B}$ to be an infinity matrix. However, this violates eq. 22 which favours small $\mathbf{B}$. Similarly, one can optimize eq. 22 by setting $\mathbf{B}$ to be a zero matrix. But again, this violates eq. 20. Thus, eq. 20 and eq. 22 create implicit lower and upper bound constraints on $\mathbf{B}$: $\mathbf{0} \leq \mathbf{B} \leq \mathbf{\Upsilon_B}$.

For convenience, objective in eq. 22 can be restated as:

$$\min_{\mathbf{B}} \text{tr}\,(\mathbf{B}^T\mathbf{B}) \equiv \min_{\mathbf{B}} \text{tr}\,(\mathbf{B}^T\mathbf{B}\mathbf{B}^T\mathbf{B}). \quad (27)$$

By using the fact $\text{tr}\,(\mathbf{X}^T\mathbf{X}) = \|\mathbf{X}\|_F^2$, eq. 27 can be rewritten into:

$$\min_{\mathbf{B}} \Big(\|\mathbf{B}^T\mathbf{B}\|_F^2 = \sum_i \big(\mathbf{b}_i^T\mathbf{b}_i\big)^2 + \sum_{i\neq j} \big(\mathbf{b}_i^T\mathbf{b}_j\big)^2 \Big), \quad (28)$$

Therefore, the objectives in eq. 20 – 22 can be restated into:

$$\max_{\mathbf{B}} \text{tr}\,(\mathbf{B}^T\mathbf{A}\mathbf{A}^T\mathbf{B}) \quad (29)$$
$$\max_{\mathbf{B}} \text{tr}\,(\mathbf{\Gamma_C}^T\mathbf{A}^T\mathbf{B}) \quad (30)$$
$$\min_{\mathbf{b}} \Big(\underbrace{\sum_i \big(\mathbf{b}_i^T\mathbf{b}_i\big)^2}_{j_{b1}} + \underbrace{\sum_{i\neq j} \big(\mathbf{b}_i^T\mathbf{b}_j\big)^2}_{j_{b2}} \Big) \quad (31)$$
$$\text{s.t. } \mathbf{0} \leq \mathbf{B} \leq \mathbf{\Upsilon_B}.$$

Now the feasible region is bounded, thus guaranteeing the existence of at least one limit point of the sequence.

However, even though the objectives are now transformed into bound constrained optimization problem, since there is no column-orthogonality constraint,

maximizing eq. 29 can be easily done by setting each entry of $\mathbf{B}$ to the corresponding largest possible value (in graph term this means to only create one partition on $\mathcal{G}(\mathbf{A}\mathbf{A}^T)$). But this scenario results in a large value of eq. 31, which violates the objective. Similarly, minimizing eq. 31 to the smallest possible value violates eq. 29. Since minimizing j_{b1} implies minimizing j_{b2}, but not vice versa, simultaneous optimizing eq. 29 and eq. 31 can be done by setting j_{b2} as small as possible and balancing j_{b1} with eq. 29. This scenario is the relaxed ratio association applied to $\mathcal{G}(\mathbf{A}\mathbf{A}^T)$, and as long as vertices of $\mathcal{G}(\mathbf{A}\mathbf{A}^T)$ are clustered, this leads to the feature clustering indicator matrix $\mathbf{B}$.

The remaining problem is objective in eq. 30. Since we know nothing about $\mathbf{\Gamma_C}$, the best bet will be making $\mathbf{A}^T\mathbf{B}$ as dense as possible. This can be done by setting $\mathbf{B}$ to largest possible values, but this scenario violates objective in eq. 31. So, the most reasonable scenario will be making $\mathbf{A}^T\mathbf{B}$ denser at the entries near diagonal region to guarantee the objective near optimal. This can be achieved by using $\mathbf{B}$ from previous discussion. As $\mathbf{B}$ is the feature clustering indicator matrix, multiplying $\mathbf{A}^T$ with $\mathbf{B}$ will result in a matrix that has denser entries near diagonal region, therefore it can be expected that eq. 30 will have good optimality. Thus simultaneously optimizing eq. 29 – 31 leads to the feature clustering indicator matrix $\mathbf{B}$.

By applying the similar approach to the coefficient matrix $\mathbf{C}$, optimizing eq. 24 – 26 is equivalent to optimizing:

$$\max_{\mathbf{C}} \operatorname{tr}\left(\mathbf{C}\mathbf{A}^T\mathbf{A}\mathbf{C}^T\right) \tag{32}$$

$$\max_{\mathbf{C}} \operatorname{tr}\left(\mathbf{C}\mathbf{A}^T\mathbf{\Gamma_B}\right) \tag{33}$$

$$\min_{\hat{\mathbf{c}}} \left(\sum_i \left(\hat{\mathbf{c}}_i\hat{\mathbf{c}}_i^T\right)^2 + \sum_{i\neq j}\left(\hat{\mathbf{c}}_i\hat{\mathbf{c}}_j^T\right)^2\right) \tag{34}$$

$$\text{s.t. } \mathbf{0} \leq \mathbf{C} \leq \mathbf{\Upsilon_C},$$

where $\hat{\mathbf{c}}_i$ denotes i-th row of $\mathbf{C}$. And by following the previous discussion, it can be shown that simultaneously optimizing eq. 32 – 34 leads to the item clustering indicator matrix $\mathbf{C}$. □

4.2 Uni-orthogonal NMF

Uni-orthogonal NMF objective is introduced by Ding et al. [15] for improving clustering capabilities by imposing auxiliary orthogonality constraint either on the basis matrix ($\mathbf{B}^T\mathbf{B} = \mathbf{I}$) or the coefficient matrix ($\mathbf{C}\mathbf{C}^T = \mathbf{I}$). The clustering aspect of this objective is reported in [13, 15, 32]. Because $(\mathbf{A} - \mathbf{B}\mathbf{C})^T = \mathbf{A}^T - \mathbf{C}^T\mathbf{B}^T$, it is sufficient to discuss the orthogonality constraint either on $\mathbf{B}$ or $\mathbf{C}$. Here we choose $\mathbf{C}\mathbf{C}^T = \mathbf{I}$ as the auxiliary constraint.

Theorem 2. *Minimizing the following objective*

$$\min_{\mathbf{B},\mathbf{C}} J_b(\mathbf{B},\mathbf{C}) = \frac{1}{2}\|\mathbf{A} - \mathbf{B}\mathbf{C}\|_F^2 \tag{35}$$
$$\text{s.t. } \mathbf{B} \geq \mathbf{0}, \mathbf{C} \geq \mathbf{0}, \mathbf{C}\mathbf{C}^T = \mathbf{I}$$

leads to the feature clustering indicator matrix $\mathbf{B}$ *and the item clustering indicator matrix* $\mathbf{C}$.

Proof.

$$\|\mathbf{A} - \mathbf{B}\mathbf{C}\|_F^2 = \operatorname{tr}\left(\mathbf{A}^T\mathbf{A} - 2\mathbf{C}\mathbf{A}^T\mathbf{B} + \mathbf{B}^T\mathbf{B}\right).$$

Thus, minimizing J_b is equivalent to simultaneously optimizing:

$$\max_{\mathbf{B},\mathbf{C}} \operatorname{tr}\left(\mathbf{C}\mathbf{A}^T\mathbf{B}\right) \tag{36}$$

$$\min_{\mathbf{B}} \operatorname{tr}\left(\mathbf{B}^T\mathbf{B}\right). \tag{37}$$

The Lagrangian function:

$$L_b(\mathbf{B},\mathbf{C}) = J_b(\mathbf{B},\mathbf{C}) - \operatorname{tr}\left(\mathbf{\Gamma_B}\mathbf{B}^T\right) - \operatorname{tr}\left(\mathbf{\Gamma_C}\mathbf{C}\right) + \operatorname{tr}\left(\mathbf{\Lambda_C}\left(\mathbf{C}\mathbf{C}^T - \mathbf{I}\right)\right), \tag{38}$$

where $\mathbf{\Gamma_B} \in \mathbb{R}_+^{M\times K}$, $\mathbf{\Gamma_C} \in \mathbb{R}_+^{N\times K}$, $\mathbf{\Lambda_B} \in \mathbb{R}_+^{K\times K}$, and $\mathbf{\Lambda_C} \in \mathbb{R}_+^{K\times K}$ are the Lagrange multipliers. By applying the KKT conditions we get:

$$\nabla_{\mathbf{B}} L_b = \mathbf{B}\mathbf{C}\mathbf{C}^T - \mathbf{A}\mathbf{C}^T - \mathbf{\Gamma_B} = \mathbf{0} \tag{39}$$

$$\nabla_{\mathbf{C}} L_b = \mathbf{B}^T\mathbf{B}\mathbf{C} - \mathbf{B}^T\mathbf{A} - \mathbf{\Gamma_C}^T + 2\mathbf{\Lambda_C}\mathbf{C} = \mathbf{0}. \tag{40}$$

Therefore,

$$\mathbf{B} = \left(\mathbf{A}\mathbf{C}^T + \mathbf{\Gamma_B}\right) \tag{41}$$

$$\mathbf{C} = \left(\mathbf{B}^T\mathbf{B} + 2\mathbf{\Lambda_C}\right)^{-1}\left(\mathbf{B}^T\mathbf{A} + \mathbf{\Gamma_C}^T\right). \tag{42}$$

Substituting eq. 42 to eq. 36 leads to:

$$\max_{\mathbf{B}} \operatorname{tr}\left(\left(\mathbf{B}^T\mathbf{B} + 2\mathbf{\Lambda_C}\right)^{-1}\left(\mathbf{B}^T\mathbf{A}\mathbf{A}^T\mathbf{B} + \mathbf{\Gamma_C}^T\mathbf{A}^T\mathbf{B}\right)\right), \tag{43}$$

which is equivalent to simultaneously optimizing:

$$\max_{\mathbf{B}} \operatorname{tr}\left(\mathbf{B}^T\mathbf{A}\mathbf{A}^T\mathbf{B}\right) \tag{44}$$

$$\max_{\mathbf{B}} \operatorname{tr}\left(\mathbf{\Gamma_C}^T\mathbf{A}^T\mathbf{B}\right) \tag{45}$$

$$\min_{\mathbf{B}} \operatorname{tr}\left(\mathbf{B}^T\mathbf{B} + 2\mathbf{\Lambda_C}\right) \equiv \min_{\mathbf{B}} \operatorname{tr}\left(\mathbf{B}^T\mathbf{B}\right). \tag{46}$$

The objectives in eq. 44 – 46 are equivalent to the objectives in eq. 20 – 22, and consequently lead to the feature clustering indicator matrix $\mathbf{B}$ which is bounded by: $\mathbf{0} \leq \mathbf{B} \leq \mathbf{\Upsilon_B}$.

Similarly, substituting eq. 41 to eq. 36 leads to:

$$\max_{\mathbf{C}} \operatorname{tr}\left(\mathbf{C}\mathbf{A}^T\mathbf{A}\mathbf{C}^T + \mathbf{C}\mathbf{A}^T\mathbf{\Gamma_B}\right), \tag{47}$$

which is equivalent to simultaneously optimizing:

$$\max_{\mathbf{C}} \text{tr}\,(\mathbf{C}\mathbf{A}^T\mathbf{A}\mathbf{C}^T) \quad (48)$$

$$\max_{\mathbf{C}} \text{tr}\,(\mathbf{C}\mathbf{A}^T\mathbf{\Gamma}_\mathbf{B}) \quad (49)$$

$$\text{s.t. } \mathbf{C}\mathbf{C}^T = \mathbf{I}.$$

Optimizing objective in eq. 48 with the orthogonality constraint $\mathbf{C}\mathbf{C}^T = \mathbf{I}$ is equivalent to applying *ratio association* to the item graph $\mathcal{G}(\mathbf{A}^T\mathbf{A})$, and hence leads to the clustering of similar items. And by following previous discussion, the item clustering indicator matrix $\mathbf{C}$ also leads to nearly optimal objective in eq. 49. □

4.3 Bi-orthogonal NMF

Bi-orthogonal NMF objective is introduced by Ding et al. [15] by imposing auxiliary orthogonality constraints on both the basis matrix and the coefficient matrix. Because both $\mathbf{B}$ and $\mathbf{C}$ are constrained to be orthogonal, the approximation of $\mathbf{A}$ by $\mathbf{BC}$ will lead to the poor result. To avoid this, Ding et al. [15] introduce matrix $\mathbf{S} \in \mathbb{R}_+^{K\times K}$ to absorb the different scales of $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$. The clustering aspect of this objective is reported in [13, 15, 33].

Theorem 3. *Minimizing the following objective*

$$\min_{\mathbf{B},\mathbf{C}} J_c(\mathbf{B},\mathbf{C}) = \frac{1}{2}\|\mathbf{A}-\mathbf{B}\mathbf{S}\mathbf{C}\|_F^2 \quad (50)$$

$$\text{s.t. } \mathbf{B}\geq\mathbf{0}, \mathbf{S}\geq\mathbf{0}, \mathbf{C}\geq\mathbf{0}, \mathbf{B}^T\mathbf{B}=\mathbf{I}, \mathbf{C}\mathbf{C}^T=\mathbf{I}$$

leads to the feature clustering indicator matrix $\mathbf{B}$ *and the item clustering indicator matrix* $\mathbf{C}$.

Proof. By absorbing $\mathbf{S}$ into $\mathbf{B}$, objective in eq. 50 is equivalent to eq. 35, and therefore leads to the item clustering indicator matrix $\mathbf{C}$. Similarly, by absorbing $\mathbf{S}$ into $\mathbf{C}$, objective in eq. 50 is also equivalent to eq. 35, thus leads to the feature clustering indicator matrix $\mathbf{B}$. □

4.4 Sparse NMF

There are many sparse NMF objectives available. Here we enlist some of them:

1. *Local NMF* [3]:

$$\min_{\mathbf{B},\mathbf{C}} J = \frac{1}{2}\|\mathbf{A}-\mathbf{B}\mathbf{C}\|_F^2+\alpha\|\mathbf{B}^T\mathbf{B}\|_F-\beta\text{tr}(\mathbf{C}\mathbf{C}^T), \quad (51)$$

with α and β are regularized parameters. Note that the original local NMF objective uses divergence instead of Frobenius norm.

2. *Hoyer's sparse NMF* [4]:

$$\min_{\mathbf{B},\mathbf{C}} J = \frac{1}{2}\|\mathbf{A}-\mathbf{B}\mathbf{C}\|_F^2 + \alpha\sum_{k=1}^{K} S(\mathbf{b}_k) + \beta\sum_{k=1}^{K} S(\hat{\mathbf{c}}_k), \quad (52)$$

where S is the Hoyer's sparseness function, and $\hat{\mathbf{c}}_k$ is k-th row of $\mathbf{C}$.

3. *Sparse NMF with L_1-norm constraint* [24]:

$$\min_{\mathbf{B},\mathbf{C}} J = \frac{1}{2}\|\mathbf{A}-\mathbf{B}\mathbf{C}\|_F^2 + \alpha\|\mathbf{B}\|_F^2 + \beta\sum_{n=1}^{N}\|\mathbf{c}_n\|_1^2. \quad (53)$$

4. *Constrained NMF* [8]:

$$\min_{\mathbf{B},\mathbf{C}} J = \frac{1}{2}\|\mathbf{A}-\mathbf{B}\mathbf{C}\|_F^2 + \alpha J_1(\mathbf{B}) + \beta J_2(\mathbf{C}), \quad (54)$$

where J_1 and J_2 are penalty terms used to enforce certain constraints on the solution.

To prove the clustering aspect of sparse NMF, we choose constrained NMF objective by setting $J_1(\mathbf{B}) = 1/2\|\mathbf{B}\|_F^2$ and $J_2(\mathbf{C}) = 1/2\|\mathbf{C}\|_F^2$. This is because constrained NMF is more general than any other sparse NMF objectives, and Frobenius norm seems to be the most widely used criterion to measure the sparseness of matrices in NMF problems [8, 12, 24, 25, 31]. The clustering aspect of other sparse NMF objectives can also be proven in similar fashion.

Theorem 4. *Minimizing the following objective*

$$\min_{\mathbf{B},\mathbf{C}} J_d(\mathbf{B},\mathbf{C}) = \frac{1}{2}\left\{\|\mathbf{A}-\mathbf{B}\mathbf{C}\|_F^2 + \alpha\|\mathbf{B}\|_F^2 + \beta\|\mathbf{C}\|_F^2\right\} \quad (55)$$

$$\text{s.t. } \mathbf{B}\geq\mathbf{0}, \mathbf{C}\geq\mathbf{0},$$

leads to the feature clustering indicator matrix $\mathbf{B}$ *and the item clustering indicator matrix* $\mathbf{C}$.

Proof.

$$\|\mathbf{A}-\mathbf{B}\mathbf{C}\|_F^2 + \alpha\|\mathbf{B}\|_F^2 + \beta\|\mathbf{C}\|_F^2 =$$
$$\text{tr}\,(\mathbf{A}^T\mathbf{A} - 2\mathbf{C}\mathbf{A}^T\mathbf{B} + \mathbf{B}^T\mathbf{B}\mathbf{C}\mathbf{C}^T + \alpha\mathbf{B}^T\mathbf{B} + \beta\mathbf{C}\mathbf{C}^T).$$

Following proof of theorem 1:

$$\min(\mathbf{B}^T\mathbf{B}\mathbf{C}\mathbf{C}^T) \equiv \min(\mathbf{B}^T\mathbf{B}) \text{ and } \min(\mathbf{C}\mathbf{C}^T),$$

thus minimizing J_d is equivalent to simultaneously optimizing:

$$\max_{\mathbf{B},\mathbf{C}} \text{tr}\,(\mathbf{C}\mathbf{A}^T\mathbf{B}) \quad (56)$$

$$\min_{\mathbf{B}} \text{tr}\,(\mathbf{B}^T\mathbf{B}) \quad (57)$$

$$\min_{\mathbf{C}} \text{tr}\,(\mathbf{C}\mathbf{C}^T). \quad (58)$$

The Lagrangian function:

$$L_d(\mathbf{B},\mathbf{C}) = J_d(\mathbf{B},\mathbf{C}) - \text{tr}\,\left(\mathbf{\Gamma}_\mathbf{B}\mathbf{B}^T\right) - \text{tr}\,\left(\mathbf{\Gamma}_\mathbf{C}\mathbf{C}\right). \quad (59)$$

By applying the KKT conditions we get:

$$\nabla_\mathbf{B} L_d = \mathbf{B}\mathbf{C}\mathbf{C}^T - \mathbf{A}\mathbf{C}^T + \alpha\mathbf{B} - \mathbf{\Gamma}_\mathbf{B} = \mathbf{0} \quad (60)$$

$$\nabla_\mathbf{C} L_d = \mathbf{B}^T\mathbf{B}\mathbf{C} - \mathbf{B}^T\mathbf{A} + \beta\mathbf{C} - \mathbf{\Gamma}_\mathbf{C}^T = \mathbf{0}. \quad (61)$$

Therefore,

$$\mathbf{B} = (\mathbf{A}\mathbf{C}^T + \mathbf{\Gamma_B})(\mathbf{C}\mathbf{C}^T + \alpha\mathbf{I})^{-1} \quad (62)$$

$$\mathbf{C} = (\mathbf{B}^T\mathbf{B} + \beta\mathbf{I})^{-1}(\mathbf{B}^T\mathbf{A} + \mathbf{\Gamma}_\mathbf{C}^T). \quad (63)$$

Substituting eq. 63 to eq. 56 leads to:

$$\max_{\mathbf{B}} \operatorname{tr}\left((\mathbf{B}^T\mathbf{B} + \beta\mathbf{I})^{-1}(\mathbf{B}^T\mathbf{A}\mathbf{A}^T\mathbf{B} + \mathbf{\Gamma}_\mathbf{C}^T\mathbf{A}^T\mathbf{B})\right), \quad (64)$$

which is equivalent to simultaneously optimizing:

$$\max_{\mathbf{B}} \operatorname{tr}(\mathbf{B}^T\mathbf{A}\mathbf{A}^T\mathbf{B}) \quad (65)$$

$$\max_{\mathbf{B}} \operatorname{tr}(\mathbf{\Gamma}_\mathbf{C}^T\mathbf{A}^T\mathbf{B}) \quad (66)$$

$$\min_{\mathbf{B}} \operatorname{tr}(\mathbf{B}^T\mathbf{B} + \beta\mathbf{I}) \equiv \min_{\mathbf{B}} \operatorname{tr}(\mathbf{B}^T\mathbf{B}). \quad (67)$$

The objectives in eq. 65 – 67 are equivalent to the objectives in eq. 20 – 22, and consequently lead to the feature clustering indicator matrix $\mathbf{B}$.

Similarly, substituting eq. 62 to eq. 56 leads to:

$$\max_{\mathbf{C}} \operatorname{tr}\left((\mathbf{C}\mathbf{A}^T\mathbf{A}\mathbf{C}^T + \mathbf{C}\mathbf{A}^T\mathbf{\Gamma_B})(\mathbf{C}\mathbf{C}^T + \alpha\mathbf{I})^{-1}\right), \quad (68)$$

which is equivalent to simultaneously optimizing:

$$\max_{\mathbf{C}} \operatorname{tr}(\mathbf{C}\mathbf{A}^T\mathbf{A}\mathbf{C}^T) \quad (69)$$

$$\max_{\mathbf{C}} \operatorname{tr}(\mathbf{C}\mathbf{A}^T\mathbf{\Gamma_B}) \quad (70)$$

$$\min_{\mathbf{C}} \operatorname{tr}(\mathbf{C}\mathbf{C}^T + \alpha\mathbf{I}) \equiv \min_{\mathbf{C}} \operatorname{tr}(\mathbf{C}\mathbf{C}^T). \quad (71)$$

The objectives in eq. 69 – 71 are equivalent to the objectives in eq. 24 – 26, and consequently lead to the item clustering indicator matrix $\mathbf{C}$. □

4.5 Semi-NMF

Semi-NMF is introduced by Ding et al. [14] to extend NMF for mixed signs data matrix $\mathbf{A} \in \mathbb{R}_{\pm}^{M\times N}$. The clustering aspect of Semi-NMF is reported in [14]. The factorization is done by releasing nonnegativity constraint on the basis matrix, while keeping the nonnegativity constraint on the coefficient matrix: $\mathbf{A}_\pm \approx \mathbf{B}_\pm\mathbf{C}_+$, thus unlike traditional NMF objectives, the feasible region is no longer located on the nonnegative orthant. Semi-NMF is motivated by K-means clustering which can be employed to find clustering for mixed signs data [14].

Because in Semi-NMF only $\mathbf{C}$ is used for clustering purpose, we prove the clustering aspect of Semi-NMF for item clustering only.

Theorem 5. *Minimizing the following objective*

$$\min_{\mathbf{B},\mathbf{C}} J_e(\mathbf{B},\mathbf{C}) = \frac{1}{2}\|\mathbf{A} - \mathbf{B}\mathbf{C}\|_F^2 \quad (72)$$
$$\text{s.t. } \mathbf{C} \geq \mathbf{0},$$

leads to the item clustering indicator matrix $\mathbf{C}$.

Proof.

$$\|\mathbf{A} - \mathbf{B}\mathbf{C}\|_F^2 = \operatorname{tr}(\mathbf{A}^T\mathbf{A} - 2\mathbf{C}\mathbf{A}^T\mathbf{B} + \mathbf{B}^T\mathbf{B}\mathbf{C}\mathbf{C}^T).$$

Following proof of theorem 1:

$$\min(\mathbf{B}^T\mathbf{B}\mathbf{C}\mathbf{C}^T) \equiv \min(\mathbf{B}^T\mathbf{B}) \text{ and } \min(\mathbf{C}\mathbf{C}^T),$$

thus minimizing J_e is equivalent to simultaneously optimizing:

$$\max_{\mathbf{B},\mathbf{C}} \operatorname{tr}(\mathbf{C}\mathbf{A}^T\mathbf{B}) \quad (73)$$

$$\min_{\mathbf{B}} \operatorname{tr}(\mathbf{B}^T\mathbf{B}) \quad (74)$$

$$\min_{\mathbf{C}} \operatorname{tr}(\mathbf{C}\mathbf{C}^T). \quad (75)$$

The Lagrangian function:

$$L_e(\mathbf{B},\mathbf{C}) = J_e(\mathbf{B},\mathbf{C}) - \operatorname{tr}(\mathbf{\Gamma_C}\mathbf{C}). \quad (76)$$

By applying the KKT conditions we get:

$$\nabla_\mathbf{B} L_e = \mathbf{B}\mathbf{C}\mathbf{C}^T - \mathbf{A}\mathbf{C}^T = \mathbf{0} \quad (77)$$

Therefore,

$$\mathbf{B} = (\mathbf{A}\mathbf{C}^T)(\mathbf{C}\mathbf{C}^T)^{-1} \quad (78)$$

Substituting eq. 78 to eq. 73 leads to:

$$\max_{\mathbf{C}} \operatorname{tr}\left((\mathbf{C}\mathbf{A}^T\mathbf{A}\mathbf{C}^T)(\mathbf{C}\mathbf{C}^T)^{-1}\right), \quad (79)$$

which is equivalent to simultaneously optimizing:

$$\max_{\mathbf{C}} \operatorname{tr}(\mathbf{C}\mathbf{A}^T\mathbf{A}\mathbf{C}^T) \quad (80)$$

$$\min_{\mathbf{C}} \operatorname{tr}(\mathbf{C}\mathbf{C}^T). \quad (81)$$

The objectives in eq. 80 and 81 are equivalent to the objectives in eq. 24 and 26, and consequently lead to the item clustering indicator matrix $\mathbf{C}$. □

Note that because Semi-NMF has a simple constraint (only nonnegativity constraint on $\mathbf{C}$), K-means clustering in eq. 4 can also be used to "weakly" prove the clustering aspect of Semi-NMF. This equivalence also is stated by the authors [14].

4.6 Convex NMF

Convex NMF is introduced by Ding et al. [14]. Both Semi-NMF and Convex NMF are extensions to the standard NMF to deal with mixed signs data matrix. However, unlike Semi-NMF where the basis matrix is nonnegativity-unconstrained, Convex NMF put nonnegative constraints on both the weight matrix $\mathbf{W}$ and the coefficient matrix $\mathbf{C}$ (see eq. 83), and thus like traditional NMF, the feasible regions are located in the nonnegative orthant. Compared to Semi-NMF, Convex NMF puts auxiliary constraint on the basis matrix

by restricting each basis vector $\mathbf{b}_k$ to be a convex combination of the data vectors $\mathbf{a}_n$:

$$\mathbf{b}_k = \sum_{n=1}^{N} w_{nk}\mathbf{a}_n, \ \forall k, \tag{82}$$

where w_{nk} is nonnegative weight. As shown in eq. 82, the basis matrix in Convex NMF captures the notion of clustering centroids much better than any other NMF objective discussed so far, and consequently more closely related to the K-means clustering. And Convex NMF can be written as:

$$\mathbf{A}_{\pm} \equiv \mathbf{A}_{\pm}\mathbf{W}_{+}\mathbf{C}_{+}, \tag{83}$$

where $\mathbf{W} \in \mathbb{R}_{+}^{N\times K}$ is the weight matrix.

Because in soft clustering (which is the clustering offered by NMF other than orthogonal NMF), usually each $\mathbf{a}_n$ in some degree belongs to small number of clusters, and each cluster comprises of a fraction of total number of the data vectors, $\mathbf{C}$ and $\mathbf{W}$ tend to be sparse. The experimental results that show the sparseness of $\mathbf{C}$ and $\mathbf{W}$ can be found in the original work [14].

Like Semi-NMF, in Convex NMF only $\mathbf{C}$ is used for clustering purpose, so we prove the clustering aspect of Convex NMF for item clustering only.

Theorem 6. *Minimizing the following objective*

$$\begin{aligned} &\min_{\mathbf{W},\mathbf{C}} J_f(\mathbf{W},\mathbf{C}) = \frac{1}{2}\|\mathbf{A}-\mathbf{AWC}\|_F^2 \\ &\text{s.t. } \mathbf{W}\geq\mathbf{0}, \mathbf{C}\geq\mathbf{0} \end{aligned} \tag{84}$$

leads to the item clustering indicator matrix $\mathbf{C}$.

Proof. By absorbing $\mathbf{W}$ into $\mathbf{A}$ to form mixed signs basis matrix $\mathbf{B} = \mathbf{AW}$, objective in eq. 84 is equivalent to eq. 72, and therefore leads to the item clustering indicator matrix $\mathbf{C}$. □

5 Related works

Some works show the equivalence between the standard NMF and K-means clustering [12, 13, 14, 15, 16], however as stated previously, there is no obvious way to incorporate the nonnegativity and other constraints into the K-means objective.

Ding et al. [15] provide a theoretical analysis on the equivalence between uni-orthogonal NMF and graph clustering, i.e., *ratio association*. However as their proof utilizes the zero gradient conditions, the hidden assumptions (setting the Lagrange multipliers to zeros) are not revealed there. And as stated previously, this approach is the KKT conditions applied to the nonnegativity-unconstrained version of eq. 35. Thus, there is no guarantee that the stationary points which being utilized to prove the equivalences are located in the nonnegative orthant. Then by using the same approach, Ding et al. [14] extend this effort to also include other objectives, i.e., Semi-NMF, Convex NMF, Cluster NMF, and Kernel NMF.

The first attempt of Ding et al. [16] to prove the clustering aspect of the standard NMF actually is better since there is no zero Lagrange multiplier assumption being made. However, the proof is only for symmetric matrices and due to the used approach, the theorem cannot be extended to rectangular matrices which so far are the usual form of the data (it seems that the practical applications of NMF are exclusively for rectangular matrices). Therefore, their result cannot be used to explain the abundant experimental results that show the power of the standard NMF in clustering, e.g., [2, 7, 10, 12, 13, 14, 29, 31]. Moreover, they made unnecessary step by proving the clustering indicator vectors are approximately orthogonal to each other, which is a little bit misleading since as shown by Xu et al. [10] the vectors point to cluster centroids in the nonnegative orthant. Therefore, when the centroids are close to each other, their proof will not be correct.

6 Conclusion and future works

By applying the strict KKT optimality conditions to the standard NMF, uni-orthogonal NMF, bi-orthogonal NMF, sparse NMF, Semi-NMF, and Convex NMF, the equivalences between these objectives to graph clustering objective, i.e., *ratio association* are obtained, thus giving a theoretical framework for supporting the clustering aspect of these objectives. There are highly possible that many other NMF objectives also have clustering capabilities. We believe that the same framework can also be utilized to derive the equivalences. However, the proofs presented can only explain the clustering aspect itself, without further explanation concerning the clustering quality differences among objectives. This issue is important since there are works that show sparse NMF tends to be better than the standard NMF [12, 31], and bi-orthogonal NMF is better than the standard NMF, Semi-NMF, and Convex NMF [13]. We will address this issue in future researches.

Some NMF objectives discussed, i.e., uni-orthogonal NMF, bi-orthogonal NMF, Semi-NMF, and Convex NMF have only multiplicative update based algorithms which have no convergence guarantee. Thus, in appendix we provide additive update based versions which for the standard NMF has been shown to have good convergence property [20]. And to anticipate other NMF objectives, we provide a more general form of the additive update algorithm in appendix E. However, as the convergence only being proven for the standard NMF, it is necessary to obtain formal convergence proofs for all algorithms in appendix. Further, it is also necessary to evaluate their performances

compared to the corresponding multiplicative update counterparts. We will address these problems in future researches.

References

[1] I.S. Dhillon and S. Sra, "Generalized nonnegative matrix approximation with Bregman divergences," UTCS Technical Reports, The University of Texas at Austin, 2005.

[2] D. Lee and H. Seung, "Learning the parts of objects by non-negative matrix factorization," Nature, 401(6755), pp. 788-91, 1999.

[3] S.Z. Li, X.W. Hou, H.J. Zhang, and Q.S. Cheng, "Learning spatially localized, parts-based representation," Proc. IEEE Comp. Soc. Conf. on Computer Vision and Pattern Recognition, pp. 207-12, 2001.

[4] P.O. Hoyer, "Non-negative matrix factorization with sparseness constraints," The Journal of Machine Learning Research, Vol. 5, pp. 1457-69, 2004.

[5] M. Berry, M. Brown, A. Langville, P. Pauca, and R.J. Plemmons, "Algorithms and applications for approximate nonnegative matrix factorization," Computational Statistics and Data Analysis, 2006. Preprint.

[6] D. Kim, S. Sra, and I.S. Dhillon, "Fast projection-based methods for the least squares nonnegative matrix approximation problem," Stat. Anal. Data Min., Vol. 1(1), pp. 38-51, 2008.

[7] D. Kim, S. Sra, I.S. Dhillon, "Fast newton-type methods for the least squares nonnegative matrix approximation problem," Proc. SIAM Conference on Data Mining, pp. 343-54, 2007.

[8] V.P. Pauca, J. Piper, and R.J. Plemmons, "Nonnegative matrix factorization for spectral data analysis," Linear Algebra and Its Applications, Vol. 416(1), pp. 29-47, 2006.

[9] D. Donoho and V. Stodden, "When does nonnegative matrix factorization give a correct decomposition into parts?," Proc. 16th Neural Information Processing Systems, pp. 1141-9, 2003.

[10] W. Xu, X. Liu and Y. Gong, "Document clustering based on non-negative matrix factorization," Proc. ACM SIGIR, pp. 267-73, 2003.

[11] F. Shahnaz, M.W. Berry, V. Pauca, and R.J. Plemmons, "Document clustering using nonnegative matrix factorization," Information Processing & Management, Vol. 42(2), pp. 373-86, 2006. Preprint.

[12] J. Kim and H. Park, "Sparse nonnegative matrix factorization for clustering," CSE Technical Reports, Georgia Institute of Technology, 2008.

[13] T. Li and C. Ding, "The relationships among various nonnegative matrix factorization methods for clustering," Proc. ACM 6th Int'l Conf. on Data Mining, pp. 362-71, 2006.

[14] C. Ding, T. Li, and M.I. Jordan, "Convex and semi-nonnegative matrix factorizations," IEEE Transactions on Pattern Analysis and Machine Intelligence, pp. 45-55, 2010.

[15] C. Ding, T. Li, W. Peng, and H. Park, "Orthogonal nonnegative matrix t-factorizations for clustering," Proc. 12th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining, pp. 126-35, 2006.

[16] C. Ding, X. He, and H.D. Simon, "On the equivalence of nonnegative matrix factorization and spectral clustering," Proc. SIAM Data Mining Conference, pp. 606-10, 2005.

[17] M. Chu, F. Diele, R.J. Plemmons, and S. Ragni, "Optimality, computation, and interpretation of nonnegative matrix factorizations," Unpublished Report, 2004.

[18] I. S. Dhillon, Y. Guan, and B. Kulis, "Weighted graph cuts without eigenvectors: A multilevel approach," IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 29, No. 11, pp. 1944-57, 2007.

[19] D. Lee and H. Seung, "Algorithms for nonnegative matrix factorization," Proc. Advances in Neural Processing Information Systems, pp. 556-562, 2001.

[20] C.J. Lin, "On the convergence of multiplicative update algorithms for nonnegative matrix factorization," IEEE Transactions on Neural Networks, Vol. 18(6), 2007.

[21] C.J. Lin, "Projected gradient methods for nonnegative matrix factorization," Technical Report ISSTECH-95-013, Department of CS, National Taiwan University, 2005.

[22] H. Hindi, "A tutorial on convex optimization," Proc. American Control Conference, pp. 3252-65, 2004.

[23] L. Grippo and M. Sciandrone, "On the convergence of the block nonlinear Gauss-Seidel method under convex constraints," Operation Research Letters, Vol. 26, pp. 127-36, 2000.

[24] H. Kim and H. Park, "Nonnegative matrix factorization based on alternating nonnegativity constrained least squares and active set method," SIAM. J. Matrix Anal. & Appl., Vol. 30(2), pp. 713-30, 2008. Preprint.

[25] J. Kim and H. Park, "Toward faster nonnegative matrix factorization: A new algorithm and comparisons," Proc. 8th IEEE International Conference on Data Mining, pp. 353-62, 2008.

[26] M.H.V. Benthem and M.R. Keenan, "Fast algorithm for the solution of large-scale nonnegativity-constrained least squares problems," Journal of Chemometrics, Vol. 18, pp. 441-50, 2004.

[27] R. Bro and S.D. Jong, "A fast non-negativity-constrained least squares algorithm," Journal of Chemometrics, Vol. 11, pp. 393-401, 1997.

[28] C.L. Lawson and R.J. Hanson, "Solving least squares problems," SIAM Classic in Applied Mathematics, 1995.

[29] R. Albright, J. Cox, D. Duling, A. Langville, and C. Meyer, "Algorithms, Initializations, and Convergence for the Nonnegative Matrix Factorization," NCSU Technical Report Math 81706, North Carolina State University, 2006.

[30] E.F. Gonzales and Y. Zhang, "Accelerating the Lee-Seung algorithm for non-negative matrix factorization," Technical Report, Department of Computational and Applied Mathematics, Rice University, 2005.

[31] H. Kim and H. Park, "Sparse non-negative matrix factorizations via alternating non-negativity constrained least squares for microarray data analysis," Bioinformatics, Vol. 23(12), pp. 1495-502, 2007.

[32] J. Yoo and S. Choi, "Orthogonal nonnegative matrix factorization: Multiplicative updates on Stiefel manifolds," Proc. 9th Int'l Conf. Intelligent Data Engineering and Automated Learning, pp. 140-7, 2008.

[33] J. Yoo and S. Choi, "Orthogonal nonnegative matrix tri-factorization for co-clustering: Multiplicative updates on Stiefel manifolds," Information Processing & Management, Vol. 46(5), pp. 559-70, 2010.

Appendix

A Additive update algorithm for uni-orthogonal NMF

The following algorithm is uni-orthogonal NMF ($\mathbf{B}^T\mathbf{B} = \mathbf{I}$) algorithm proposed by Ding et al. [15] which is based on multiplicative update rules:

$$b_{mk} \longleftarrow b_{mk} \frac{(\mathbf{AC}^T)_{mk}}{(\mathbf{BB}^T\mathbf{AC}^T)_{mk}} \tag{85}$$

$$c_{kn} \longleftarrow c_{kn} \frac{(\mathbf{B}^T\mathbf{A})_{mk}}{(\mathbf{B}^T\mathbf{AC}^T\mathbf{C})_{kn}}. \tag{86}$$

The additive version can be written as:

$$b_{mk} \longleftarrow b_{mk} - \frac{b_{mk}}{(\mathbf{BB}^T\mathbf{AC}^T)_{mk}} (\underbrace{\mathbf{BB}^T\mathbf{AC}^T - \mathbf{AC}^T}_{\hat{\mathbf{B}}})_{mk} \tag{87}$$

$$c_{kn} \longleftarrow c_{kn} - \frac{c_{kn}}{(\mathbf{B}^T\mathbf{AC}^T\mathbf{C})_{kn}} (\underbrace{\mathbf{B}^T\mathbf{AC}^T\mathbf{C} - \mathbf{B}^T\mathbf{A}}_{\hat{\mathbf{C}}})_{mk}. \tag{88}$$

By inspection we can see that both algorithms are equivalent. To handle numerical difficulties and convergence issue, the following modifications are necessary [20]:

$$b_{mk} \longleftarrow b_{mk} - \frac{\hat{b}_{mk}}{(\mathbf{BB}^T\mathbf{AC}^T)_{mk} + \delta} \hat{\mathbf{B}}_{mk} \tag{89}$$

$$c_{kn} \longleftarrow c_{kn} - \frac{\hat{c}_{kn}}{(\mathbf{B}^T\mathbf{AC}^T\mathbf{C})_{kn} + \delta} \hat{\mathbf{C}}_{mk}, \tag{90}$$

where

$$\hat{b}_{mk} \equiv \begin{cases} b_{mk} & \text{if } \hat{\mathbf{B}}_{mk} \geq 0 \\ \max(b_{mk}, \sigma) & \text{if } \hat{\mathbf{B}}_{mk} < 0 \end{cases},$$

and

$$\hat{c}_{kn} \equiv \begin{cases} c_{kn} & \text{if } \hat{\mathbf{C}}_{kn} \geq 0 \\ \max(c_{kn}, \sigma) & \text{if } \hat{\mathbf{C}}_{kn} < 0 \end{cases},$$

with $\delta > 0$ and $\sigma > 0$ are very small adjustable constants (Lin [20] proposes setting $\delta = \sigma = 10^{-8}$). As stated in [20] this additive update algorithm is guaranteed to converge to a stationary point.

B Additive update algorithm for bi-orthogonal NMF

The multiplicative update rules based algorithm for bi-orthogonal NMF [15] can be written as follow:

$$b_{mk} \longleftarrow b_{mk} \frac{(\mathbf{AC}^T\mathbf{S}^T)_{mk}}{(\mathbf{BB}^T\mathbf{AC}^T\mathbf{S}^T)_{mk}} \tag{91}$$

$$c_{kn} \longleftarrow c_{kn} \frac{(\mathbf{S}^T\mathbf{B}^T\mathbf{A})_{kn}}{(\mathbf{S}^T\mathbf{B}^T\mathbf{AC}^T\mathbf{C})_{kn}} \tag{92}$$

$$s_{pq} \longleftarrow s_{pq} \frac{(\mathbf{B}^T\mathbf{AC}^T)_{pq}}{(\mathbf{B}^T\mathbf{BSCC}^T)_{pq}}. \tag{93}$$

And the additive version is:

$$b_{mk} \longleftarrow b_{mk} - \frac{b_{mk}}{(\mathbf{B}\mathbf{B}^T\mathbf{A}\mathbf{C}^T\mathbf{S}^T)_{mk}}\tilde{\mathbf{B}}_{mk} \tag{94}$$

$$c_{kn} \longleftarrow c_{kn} - \frac{c_{kn}}{(\mathbf{S}^T\mathbf{B}^T\mathbf{A}\mathbf{C}^T\mathbf{C})_{kn}}\tilde{\mathbf{C}}_{kn} \tag{95}$$

$$s_{pq} \longleftarrow s_{pq} - \frac{s_{pq}}{(\mathbf{B}^T\mathbf{B}\mathbf{S}\mathbf{C}\mathbf{C}^T)_{pq}}\tilde{\mathbf{S}}_{pq}, \tag{96}$$

where

$$\begin{aligned}\tilde{\mathbf{B}}_{mk} &= (\mathbf{B}\mathbf{B}^T\mathbf{A}\mathbf{C}^T\mathbf{S}^T - \mathbf{A}\mathbf{C}^T\mathbf{S}^T)_{mk}\\ \tilde{\mathbf{C}}_{kn} &= (\mathbf{S}^T\mathbf{B}^T\mathbf{A}\mathbf{C}^T\mathbf{C} - \mathbf{S}^T\mathbf{B}^T\mathbf{A})_{kn}\\ \tilde{\mathbf{S}}_{pq} &= (\mathbf{B}^T\mathbf{B}\mathbf{S}\mathbf{C}\mathbf{C}^T - \mathbf{B}^T\mathbf{A}\mathbf{C}^T)_{pq}.\end{aligned}$$

Then a similar modifications must be applied to deal with numerical difficulties and convergence issue as in uni-orthogonal case.

C Additive update algorithm for Semi-NMF

Ding et al. [14] propose the following multiplicative update algorithm for Semi-NMF:

$$\mathbf{B} \longleftarrow \mathbf{A}\mathbf{C}^T(\mathbf{C}\mathbf{C}^T)^{-1} \tag{97}$$

$$c_{kn} \longleftarrow c_{kn}\sqrt{\frac{(\mathbf{B}^T\mathbf{A})^+_{kn} + \left[(\mathbf{B}^T\mathbf{B})^-\mathbf{C}\right]_{kn}}{(\mathbf{B}^T\mathbf{A})^-_{kn} + \left[(\mathbf{B}^T\mathbf{B})^+\mathbf{C}\right]_{kn}}}, \tag{98}$$

where $\mathbf{X}^+ = (|x|_{ij} + x_{ij})/2$ and $\mathbf{X}^- = (|x|_{ij} - x_{ij})/2$ and pseudo inverse is used if inverse cannot be calculated. The additive version can be written as:

$$\mathbf{B} \longleftarrow \mathbf{A}\mathbf{C}^T(\mathbf{C}\mathbf{C}^T)^{-1} \tag{99}$$

$$c_{kn} \longleftarrow c_{kn} - \frac{c_{kn}}{\sqrt{(\mathbf{B}^T\mathbf{A})^-_{kn} + \left[(\mathbf{B}^T\mathbf{B})^+\mathbf{C}\right]_{kn}}}\mathbf{S}_{kn}, \tag{100}$$

where

$$\mathbf{S}_{kn} = \sqrt{(\mathbf{B}^T\mathbf{A})^-_{kn} + \left[(\mathbf{B}^T\mathbf{B})^+\mathbf{C}\right]_{kn}} - \sqrt{(\mathbf{B}^T\mathbf{A})^+_{kn} + \left[(\mathbf{B}^T\mathbf{B})^-\mathbf{C}\right]_{kn}}\ .$$

Then a similar modifications must be applied to deal with numerical difficulties and convergence issue as in uni-orthogonal case.

D Additive update algorithm for Convex NMF

Convex NMF is introduced by Ding et al. [14] and they proposed the following multiplicative update algorithm to compute it:

$$w_{nk} = w_{nk}\sqrt{\frac{(\hat{\mathbf{A}}^+\mathbf{C}^T)_{nk} + (\hat{\mathbf{A}}^-\mathbf{W}\mathbf{C}\mathbf{C}^T)_{nk}}{(\hat{\mathbf{A}}^-\mathbf{C}^T)_{nk} + (\hat{\mathbf{A}}^+\mathbf{W}\mathbf{C}\mathbf{C}^T)_{nk}}} \tag{101}$$

$$c_{kn} = c_{kn}\sqrt{\frac{(\mathbf{W}^T\hat{\mathbf{A}}^+)_{kn} + (\mathbf{W}^T\hat{\mathbf{A}}^-\mathbf{W}\mathbf{C})_{kn}}{(\mathbf{W}^T\hat{\mathbf{A}}^-)_{kn} + (\mathbf{W}^T\hat{\mathbf{A}}^+\mathbf{W}\mathbf{C})_{kn}}}, \tag{102}$$

where $\hat{\mathbf{A}} = \mathbf{A}^T\mathbf{A}$, and $\mathbf{X}^+$ and $\mathbf{X}^-$ are defined similarly as in appendix C. The additive version can be written as:

$$w_{nk} = w_{nk} - \frac{w_{nk}}{\sqrt{(\hat{\mathbf{A}}^-\mathbf{C}^T)_{nk} + (\hat{\mathbf{A}}^+\mathbf{W}\mathbf{C}\mathbf{C}^T)_{nk}}}\breve{\mathbf{W}}_{nk} \tag{103}$$

$$c_{kn} = c_{kn} - \frac{c_{kn}}{\sqrt{(\mathbf{W}^T\hat{\mathbf{A}}^-)_{kn} + (\mathbf{W}^T\hat{\mathbf{A}}^+\mathbf{W}\mathbf{C})_{kn}}}\breve{\mathbf{C}}_{kn}, \tag{104}$$

where

$$\begin{aligned}\breve{\mathbf{W}}_{nk} =& \sqrt{(\hat{\mathbf{A}}^-\mathbf{C}^T)_{nk} + (\hat{\mathbf{A}}^+\mathbf{W}\mathbf{C}\mathbf{C}^T)_{nk}} -\\ & \sqrt{(\hat{\mathbf{A}}^+\mathbf{C}^T)_{nk} + (\hat{\mathbf{A}}^-\mathbf{W}\mathbf{C}\mathbf{C}^T)_{nk}}\\ \breve{\mathbf{C}}_{kn} =& \sqrt{(\mathbf{W}^T\hat{\mathbf{A}}^-)_{kn} + (\mathbf{W}^T\hat{\mathbf{A}}^+\mathbf{W}\mathbf{C})_{kn}} -\\ & \sqrt{(\mathbf{W}^T\hat{\mathbf{A}}^+)_{kn} + (\mathbf{W}^T\hat{\mathbf{A}}^-\mathbf{W}\mathbf{C})_{kn}}.\end{aligned}$$

Then a similar modifications must be applied to deal with numerical difficulties and convergence issue as in uni-orthogonal case.

E A more general form of additive update algorithm for NMF

A more general form of NMF objective formulation includes auxiliary constraints on $\mathbf{B}$ and/or $\mathbf{C}$ in addition to the nonnegativity constraints:

$$\min_{\mathbf{B},\mathbf{C}} J = \frac{1}{2}\|\mathbf{A}-\mathbf{B}\mathbf{C}\|_F^2 + \alpha J_1(\mathbf{B}) + \beta J_2(\mathbf{C}) + \gamma J_3(\mathbf{B},\mathbf{C}). \tag{105}$$

The Lagrangian:

$$L = J - \mathrm{tr}\,(\mathbf{\Gamma}_\mathbf{B}\mathbf{B}^T) - \mathrm{tr}\,(\mathbf{\Gamma}_\mathbf{C}\mathbf{C}), \tag{106}$$

where $\mathbf{\Gamma}_\mathbf{B} \in \mathbb{R}_+^{M\times K}$ and $\mathbf{\Gamma}_\mathbf{C} \in \mathbb{R}_+^{N\times K}$ are the Lagrange multipliers. By differentiating L with respect to $\mathbf{B}$ and $\mathbf{B}$ we get:

$$\mathbf{B}\mathbf{C}\mathbf{C}^T - \mathbf{A}\mathbf{C}^T + \alpha\nabla_\mathbf{B}J_1(\mathbf{B}) + \gamma\nabla_\mathbf{B}J_3(\mathbf{B},\mathbf{C}) = \mathbf{\Gamma}_\mathbf{B} \tag{107}$$

$$\mathbf{B}^T\mathbf{B}\mathbf{C} - \mathbf{B}^T\mathbf{A} + \beta\nabla_\mathbf{C}J_2(\mathbf{C}) + \gamma\nabla_\mathbf{C}J_3(\mathbf{B},\mathbf{C}) = \mathbf{\Gamma}_\mathbf{C}^T. \tag{108}$$

Then, by using the complementary slackness, the multiplicative update algorithm for NMF objective in

eq. 105 can be written as follow:

$$b_{mk} \longleftarrow b_{mk} \frac{(\mathbf{A}\mathbf{C}^T)_{mk}}{\left(\mathbf{B}\mathbf{C}\mathbf{C}^T + \alpha\nabla_{\mathbf{B}} J_1(\mathbf{B}) + \gamma\nabla_{\mathbf{B}} J_3(\mathbf{B},\mathbf{C})\right)_{mk}} \tag{109}$$

$$c_{kn} \longleftarrow c_{kn} \frac{(\mathbf{B}^T\mathbf{A})_{kn}}{\left(\mathbf{B}^T\mathbf{B}\mathbf{C} + \beta\nabla_{\mathbf{C}} J_2(\mathbf{C})\gamma\nabla_{\mathbf{C}} J_3(\mathbf{B},\mathbf{C})\right)_{kn}}. \tag{110}$$

And, the additive version can be written as:

$$b_{mk} \longleftarrow b_{mk} - \frac{b_{mk}}{\left(\mathbf{B}\mathbf{C}\mathbf{C}^T + \alpha\nabla_{\mathbf{B}} J_1(\mathbf{B}) + \gamma\nabla_{\mathbf{B}} J_3(\mathbf{B},\mathbf{C})\right)_{mk}} \bar{\mathbf{B}}_{mk} \tag{111}$$

$$c_{kn} \longleftarrow c_{kn} - \frac{c_{kn}}{\left(\mathbf{B}^T\mathbf{B}\mathbf{C} + \beta\nabla_{\mathbf{C}} J_2(\mathbf{C}) + \gamma\nabla_{\mathbf{C}} J_3(\mathbf{B},\mathbf{C})\right)_{kn}} \bar{\mathbf{C}}_{kn}, \tag{112}$$

where

$$\bar{\mathbf{B}}_{mk} = \left(\mathbf{B}\mathbf{C}\mathbf{C}^T + \alpha\nabla_{\mathbf{B}} J_1(\mathbf{B}) + \gamma\nabla_{\mathbf{B}} J_3(\mathbf{B},\mathbf{C}) - \mathbf{A}\mathbf{C}^T\right)_{mk}$$
$$\bar{\mathbf{C}}_{kn} = \left(\mathbf{B}^T\mathbf{B}\mathbf{C} + \beta\nabla_{\mathbf{C}} J_2(\mathbf{C}) + \gamma\nabla_{\mathbf{C}} J_3(\mathbf{B},\mathbf{C}) - \mathbf{B}^T\mathbf{A}\right)_{kn}.$$

Then a similar modifications must be applied to deal with numerical difficulties and convergence issue as in uni-orthogonal case.