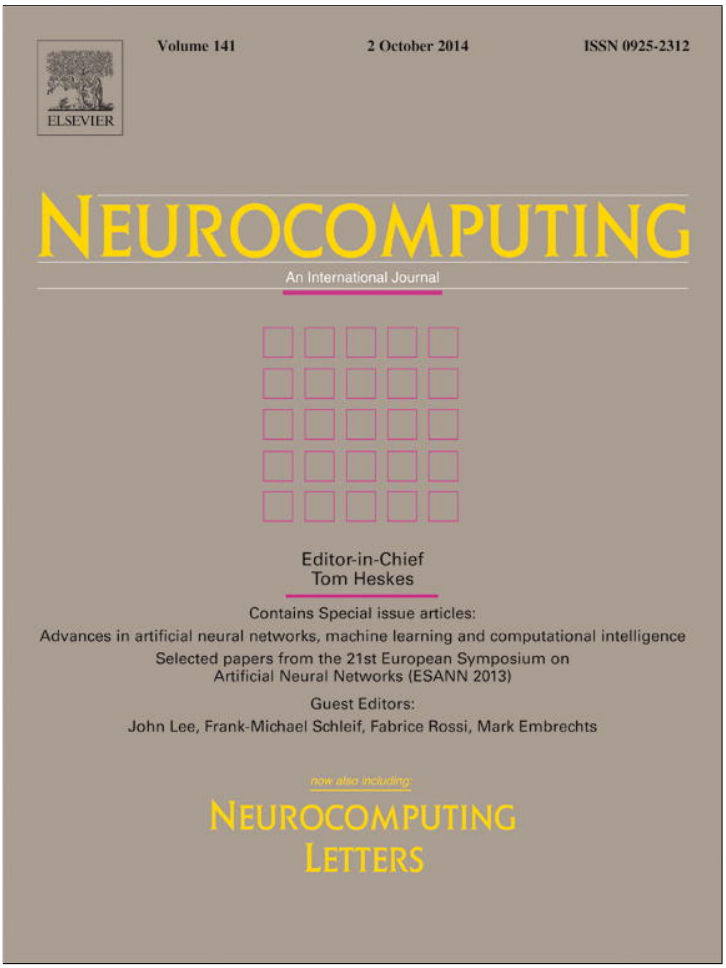

Neurocomputing 141 (2014) 15–25

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Two algorithms for orthogonal nonnegative matrix factorization with application to clustering ☆

Filippo Pompili [a], Nicolas Gillis [b,*], P.-A. Absil [c], François Glineur [c,d]

[a] Department of Electronic and Information Engineering, University of Perugia, Italy
[b] Department of Mathematics and Operational Research, Université de Mons, Rue de Houdain 9, 7000 Mons, Belgium
[c] Université catholique de Louvain, ICTEAM Institute, B-1348 Louvain-la-Neuve, Belgium
[d] Université catholique de Louvain, CORE, Voie du Roman Pays 34, B-1348 Louvain-la-Neuve, Belgium

ARTICLE INFO

Article history:
Received 21 June 2013
Received in revised form
24 January 2014
Accepted 3 February 2014
Available online 8 April 2014

Keywords:
Nonnegative matrix factorization
Orthogonality
Clustering
Document classification
Hyperspectral images

ABSTRACT

Approximate matrix factorization techniques with both nonnegativity and orthogonality constraints, referred to as orthogonal nonnegative matrix factorization (ONMF), have been recently introduced and shown to work remarkably well for clustering tasks such as document classification. In this paper, we introduce two new methods to solve ONMF. First, we show mathematical equivalence between ONMF and a weighted variant of spherical k-means, from which we derive our first method, a simple EM-like algorithm. This also allows us to determine when ONMF should be preferred to k-means and spherical k-means. Our second method is based on an augmented Lagrangian approach. Standard ONMF algorithms typically enforce nonnegativity for their iterates while trying to achieve orthogonality at the limit (e.g., using a proper penalization term or a suitably chosen search direction). Our method works the opposite way: orthogonality is strictly imposed at each step while nonnegativity is asymptotically obtained, using a quadratic penalty. Finally, we show that the two proposed approaches compare favorably with standard ONMF algorithms on synthetic, text and image data sets.

1. Introduction

We consider the orthogonal nonnegative matrix factorization (ONMF) problem, which can be formulated as follows. Given an m-by-n nonnegative matrix M and a factorization rank k (with $k<n$), solve

$$\min_{U\in\mathbb{R}^{m\times k},V\in\mathbb{R}^{k\times n}} \quad ||M-UV||_F^2 \tag{1a}$$

$$\text{subject to} \quad U\geq 0,\ V\geq 0, \tag{1b}$$

$$VV^T=I_k, \tag{1c}$$

where $\|\cdot\|_F$ denotes the Frobenius norm, (1b) means that the entries of matrices U and V are nonnegative, and I_k stands for the $k\times k$ identity matrix.

The ONMF problem (1) can be viewed as the well-known nonnegative matrix factorization (NMF) problem, (1a) and (1b), with an additional orthogonality constraint, (1c), that considerably modifies the nature of the problem. In particular, it is readily seen that constraints (1b) and (1c) imply that V has at most one nonzero entry in each column; we let i_j denote the index of the nonzero entry (if any) in column j of V. Therefore, any solution (U^*,V^*) of (1) has the following property: for $j=1,\ldots,n$, index i_j is such that column i_j of U^* achieves the smallest angle with column j of data matrix M, while $V^*(i_j,j)$ scales column i_j of U^* to make it as close as possible to column j of M (in the sense of the Euclidean norm). Hence it is clear that the ONMF problem relates to data clustering and, indeed, empirical evidence suggests that the additional orthogonality constraint (1c) can improve clustering performance compared to standard NMF or k-means [7,20].

Current approaches to ONMF problems are based on suitable modifications of the algorithms developed for the original NMF problem. They enforce nonnegativity of the iterates at each step, and strive to attain orthogonality at the limit (but never attain exactly orthogonal solutions). This can be done using a proper penalization term [10], a projection matrix formulation [20] or by choosing a suitable search direction [7]. Note that, for a given data matrix M, different methods may converge to different pairs (U,V), where the objective function (1a) may take different values. Furthermore, under random initialization, which is used by most

☆This work was carried out when NG was a Postdoctoral Researcher of the Fonds de la Recherche Scientifique (F.R.S.-FNRS). This paper presents research results of the Belgian Network DYSCO (Dynamical Systems, Control, and Optimization), funded by the Interuniversity Attraction Poles Programme initiated by the Belgian Science Policy Office.

* Corresponding author. Tel.: +32 65374680.
E-mail addresses: filippopompili.09@gmail.com (F. Pompili), nicolas.gillis@umons.ac.be (N. Gillis), pa.absil@uclouvain.be (P.-A. Absil), francois.glineur@uclouvain.be (F. Glineur).

http://dx.doi.org/10.1016/j.neucom.2014.02.018
0925-2312/

NMF algorithms [5], two runs of the same method may yield different results. This situation is due to the multimodal nature of the ONMF problem (1)—it may have multiple local minima—along with the inability of practical methods to guarantee more than convergence to local, possibly nonglobal, minimizers. Hence, ONMF methods not only differ in their computational cost, but also in the quality of the clustering encoded in the returned pair (U,V) for a given problem.

In this paper, we first show the equivalence of ONMF with a weighted variant of spherical k-means, which leads us to design an EM-like algorithm for ONMF. We also explain in which situations ONMF should be preferred to k-means and spherical k-means. Then, we propose a new ONMF method, dubbed ONP-MF, that relies on a strategy reversal: instead of enforcing nonnegativity of the iterates at each step and striving to attain orthogonality at the limit, ONP-MF enforces orthogonality of its iterates while obtaining nonnegativity at the limit. A resulting advantage of ONP-MF is that rows of factor V can be initialized directly with the right singular vectors of M (which is the optimal solution of the problem without the nonnegativity constraints), whereas the other methods require a prior alteration of the singular vectors to make them nonnegative [5]. We show that, on some clustering problems, the new algorithm outperforms other clustering methods, including ONMF-based methods, in terms of clustering quality.

The paper is organized as follows. In Section 2, we analyze the relationship between ONMF and clustering problems and show that it is closely related to spherical k-means. Based on this analysis, we develop an EM-like algorithm which features a rank-one NMF problem at its core. This also allows us to shed some light on the differences among k-means, spherical k-means and ONMF, which we illustrate on synthetic data sets. Section 3 introduces another algorithm to perform ONMF using an augmented Lagrangian and a projected gradient scheme, which enforce orthogonality at each step while obtaining nonnegativity at the limit. Finally, in Section 4, we experimentally show that our two new approaches perform competitively with standard ONMF algorithms on text data sets and on different image decomposition problems.

This paper is an extended version of the proceedings paper [18].

2. Equivalence of ONMF with a weighted variant of spherical *k*-means

In this section, we briefly recall how NMF with an additional constraint is equivalent to a fundamental clustering technique (see Eq. (c1)): Euclidean k-means [8,9]. We then observe that relaxing this constraint leads to (1c) and (1d), that is, ONMF, which is therefore not exactly equivalent to k-means but rather to another problem closely related to spherical k-means [2]. More precisely, ONMF is equivalent to weighted spherical k-means in a particular metric, see Theorem 1. Based on this analysis, we propose a new EM-like algorithm to solve ONMF problems, highlight the differences among k-means, spherical k-means and ONMF, and illustrate these results on synthetic data sets.

2.1. Equivalence with Euclidean k-means

Let $M=(m_1,\dots,m_n)\in\mathbb{R}^{m\times n}_+$ be a nonnegative data matrix whose columns represent a set of n points $\{m_j\}_{j=1}^n\in\mathbb{R}^m_+$. Solving the clustering problem means finding a set $\{\pi_i\}_{i=1}^k$ of k disjoint clusters:

$$\pi_i\subseteq\{1,2,\dots,n\}\quad\forall i,\quad\bigcup_{1\le i\le k}\pi_i=\{1,2,\dots,n\},$$

and

$$\pi_i\cap\pi_j=\emptyset,\quad\forall i\neq j,$$

such that each cluster π_i contains objects as similar as possible to each other according to some quantitative criterion. When choosing the Euclidean distance, we obtain the k-means problem, which can be formulated as follows [8]:

$$\min_{\{\pi_i\}_{i=1}^k}\sum_{i=1}^k\sum_{j\in\pi_i}\|m_j-c_i\|^2,$$

where $c_i=\sum_{j\in\pi_i}m_j/|\pi_i|$ are the cluster centroids. Equivalently, we can define a binary cluster indicator matrix $B\in\{0,1\}^{k\times n}$ as follows:

$$B=\{b_{ij}\}_{k\times n}\quad\text{where } b_{ij}=1\iff j\in\pi_i.$$

Disjointness of clusters π_i means that rows of B are orthogonal, i.e., BB^T is diagonal. Therefore we can normalize them to obtain an orthogonal matrix $V=\{v_{ij}\}_{k\times n}=(BB^T)^{-1/2}B$ (a weighted cluster indicator matrix) which satisfies the following condition: There exists a set of clusters $\{\pi_i\}_{i=1}^k$ such that

$$v_{ij}=\begin{cases}\dfrac{1}{\sqrt{|\pi_i|}} & \text{if } j\in\pi_i,\\ 0 & \text{otherwise.}\end{cases}\tag{c1}$$

It has been shown in [9] that the NMF problem with matrix V satisfying condition (c1)

$$\min_{U\ge0,V\ge0}\|M-UV\|_F^2\quad\text{s.t. } V\text{ satisfies (c1)},\tag{2}$$

is equivalent to k-means. In fact, since V in problem (2) is a normalized indicator matrix which satisfies $v_{ij}=|\pi_i|^{-1/2}\iff j\in\pi_i$, we have

$$\begin{aligned}\|M-UV\|_F^2&=\sum_{j=1}^n\left\|m_j-\sum_{i=1}^k u_iv_{ij}\right\|^2\\&=\sum_{i=1}^k\sum_{j\in\pi_i}\|m_j-u_iv_{ij}\|^2\\&=\sum_{i=1}^k\sum_{j\in\pi_i}\left\|m_j-u_i\frac{1}{\sqrt{|\pi_i|}}\right\|^2,\end{aligned}$$

which implies that, at optimality, each column u_i of U must correspond (up to a multiplicative factor) to a cluster centroid with $u_i=\sqrt{|\pi_i|}c_i=\sum_{j\in\pi_i}m_j/\sqrt{|\pi_i|}\ \ \forall i=1,\dots,k$.

2.2. ONMF and a weighted variant of spherical k-means

Let us now define a condition weaker than (c1):

$$VV^T=I_k\quad\text{and}\quad V\ge0.\tag{c2}$$

It can be easily checked that (c1) $\Rightarrow$ (c2) while (c2)$\not\Rightarrow$(c1). The difference between conditions (c1) and (c2) is that condition (c2) does not require the rows of V to have their nonzero entries equal to each other. Now, if we only impose the weaker condition (c2) on NMF, we obtain a relaxed version of (2) which, by definition, corresponds to orthogonal NMF:

$$\min_{U\ge0,V\ge0}\|M-UV\|_F^2\quad\text{such that } VV^T=I_k.\tag{3}$$

In the following, we show the equivalence of problem (3) with a particular weighted variant of the spherical k-means problem:

Theorem 1. *For a nonnegative data matrix $M\in\mathbb{R}^{m\times n}_+$, the ONMF problem (3) is equivalent to the following weighted variant of spherical k-means*

$$\max_{\{\pi_i,u_i\in\mathbb{R}^m_+,\|u_i\|_2=1\}_{i=1}^k}\sum_{i=1}^k\sum_{j\in\pi_i}\|m_j\|^2\left(\frac{m_j^T}{\|m_j\|}u_i\right)^2,\tag{4}$$

where $\{\pi_i\}_{i=1}^k$ is a set of disjoint clusters.

Proof. The claim is that (3) and (4) are *equivalent*, i.e., a solution of (3) is obtained from a solution of (4) by means of elementary arithmetic operations, and vice-versa.

First, without loss of generality, we assume that k is sufficiently small so that the solutions U of (3) do not have vanishing columns.

We then redefine "$U \geq 0$" (resp. "$V \geq 0$") to mean that U (resp. V) is nonnegative without vanishing columns (resp. rows). This redefinition does not alter the solutions of (3).

Observe that (3) is equivalent to the following problem obtained by imposing the unit-norm constraint on the columns $\{u_i\}_{i=1}^k$ of U instead of the rows of V:

$$\min_{U \geq 0, V \geq 0} \|M - UV\|_F^2 \quad \text{s.t. } (VV^T)_{ij} = 0, \ \forall i \neq j \text{ and } \|u_i\| = 1, \ \forall i. \tag{5}$$

Indeed, since the function $\psi : (U,V) \mapsto (UD^{-1}, DV)$ with $D = \text{diag}(\|u_1\|, \ldots, \|u_k\|)$ is a homeomorphism from the feasible set of (3) onto the feasible of (5) that does not modify the objective value, it is readily seen that if (U,V) is a solution of (3), then $\psi(U,V)$ is a solution of (5); proving the reverse direction is equally straightforward.

It remains to show equivalence between (5) and (4). We say that a partition $\pi = \{\pi_i\}_{i=1}^k$ and a matrix V (with k rows) are *compatible*, which we write $V \sim \pi$, if the inclusion $j \in \pi_i$ holds whenever $V_{ij} \neq 0$. Using this notion, we first notice the crucial fact that $V \sim \pi$ for some π if and only if each column of V has at most one nonzero element (at position (i,j) where i is determined by $\pi_i \ni j$). Defining $\mathcal{V}$ to be the feasible set for V in (5), it is now clear that $V \sim \pi$ and $V \geq 0$ imply that $V \in \mathcal{V}$ and that, in the reverse direction, one can easily check that $V \in \mathcal{V}$ implies the existence of a partition π such that $V \sim \pi$.

We can now show that the following four propositions are equivalent, from which the main claim follows:

1. U and V minimize $\|M - UV\|_F^2$ subject to $U \geq 0$, $\|u_i\| = 1 \ \forall i$, $V \geq 0$, $(VV^T)_{ij} = 0 \ \forall i \neq j$, and π is obtained by elementary operations to satisfy $V \sim \pi$.
2. U, V, and π minimize $\|M - UV\|_F^2$ subject to $U \geq 0$, $\|u_i\| = 1 \ \forall i$, $V \geq 0$, $V \sim \pi$.
3. U, V, and π minimize $\sum_{i=1}^k \sum_{j \in \pi_i} \|m_j - u_i v_{ij}\|^2$ subject to $U \geq 0$, $\|u_i\| = 1 \ \forall i$, $V \geq 0$, $V \sim \pi$.
4. U and π maximize $\sum_{i=1}^k \sum_{j \in \pi_i} (m_j^T u_i)^2$ subject to $U \geq 0$, $\|u_i\| = 1 \ \forall i$, and V is obtained by the elementary operations
$$\begin{cases} v_{ij} = 0 & \text{if } j \notin \pi_i \\ v_{ij} = m_j^T u_i & \text{if } j \in \pi_i. \end{cases}$$

The equivalence between 1 and 2 follows from the discussion in the previous paragraph. The equivalence between 2 and 3 follows from a rewriting of the objective function made possible by the constraints.

Finally, the equivalence between 3 and 4 is established as follows. Referring to 3, given a feasible U and π, we have for each term $\|m_j - u_i v_{ij}\|^2$ that the optimal v_{ij}^* is given by

$$\begin{aligned} v_{ij}^* &= \operatorname*{argmin}_{x \geq 0} \|m_j - u_i x\|^2 \\ &= \operatorname*{argmin}_{x \geq 0} (m_j^T m_j - 2x m_j^T u_i + x^2) \\ &= m_j^T u_i, \quad 1 \leq i \leq k, \ j \in \pi_i. \end{aligned} \tag{6}$$

(Observe that $m_j^T u_i \geq 0$ in view of the nonnegativity of M and U.) Backsubstituting the optimal coefficients (6) in 3, we have that U and π of 3 minimize the function

$$\begin{aligned} \sum_{i=1}^k \sum_{j \in \pi_i} \|m_j - (m_j^T u_i) u_i\|^2 &= \sum_{i=1}^k \sum_{j \in \pi_i} (m_j^T m_j - 2(m_j^T u_i)^2 + (m_j^T u_i)^2) \\ &= \sum_{i=1}^k \sum_{j \in \pi_i} -(m_j^T u_i)^2 + \text{cst}. \end{aligned}$$

Hence they maximize the function

$$\sum_{i=1}^k \sum_{j \in \pi_i} (m_j^T u_i)^2. \tag{7}$$

This shows that 3 implies 4, and the converse is readily established by contradiction. □

It is insightful to compare formulation (4) of ONMF with the spherical k-means problem [2], which is a variant of k-means where both data points and centroids are constrained to have unit norm:

$$\begin{aligned} \min_{\{\pi_i, u_i\}_{i=1}^k} \sum_{i=1}^k \sum_{j \in \pi_i} \left\| \frac{m_j}{||m_j||} - u_i \right\|^2 \quad \text{s.t. } ||u_i|| = 1, \\ \equiv \max_{\{\pi_i, u_i\}_{i=1}^k} \sum_{i=1}^k \sum_{j \in \pi_i} \frac{m_j^T}{||m_j||} u_i \quad \text{s.t. } ||u_i|| = 1. \end{aligned} \tag{8}$$

Note that both problems (4) and (8) relate to maximizing the cosines of the angles between u_i and the data points from the corresponding cluster. However, we observe that

- Because of coefficients $||m_i||^2$, problem (4) is sensitive to the norm of the data points, as opposed to spherical k-means (8) which only depends on their direction.
- Even for normalized data points (i.e., $||m_i|| = 1 \ \forall i$), problem (4) is similar but not equivalent to spherical k-means (8) because it tries to maximize *the sum of squares* of the cosines (instead of their sum).
- In contrast to problem (4), spherical k-means (8) does not require nonnegativity of u_i's, although it will clearly hold at optimality when data points m_j are nonnegative.

2.3. Which model should be used: k-means, spherical k-means or ONMF?

Based on the analysis of ONMF from Section 2 (in particular, Theorem 1), we explain in this section in which situations ONMF should be preferred to k-means and spherical k-means. This issue can be settled by addressing the following two questions:

1. *Should scaling of the data points influence the cluster assignment?* Given the cluster centroids, spherical k-means and ONMF are invariant to scaling in the sense that, for any $\alpha > 0$, a data point x and its scaling αx will be assigned to the same cluster (the one minimizing the angle between x and the cluster centroid; see Section 2.4). On the contrary, k-means is very sensitive to scaling as it assigns data points to clusters based on distances; see Fig. 1 for an illustration. In practice, there are many situations where cluster assignment should be independent of scaling so that spherical k-means and ONMF should be preferred to k-means. For example, in document classification, two documents discussing the same topic will roughly be multiple of one another (scaling depending then on the relative lengths of the documents), and, in hyperspectral imaging, pixels containing the same material will have their spectral signatures multiple of one another (scaling depending on the relative illumination conditions; see Section 4 for numerical experiments).
2. *Does the noise added to each data point depend on its norm?* Spherical k-means is invariant under normalization of the data points (see Eq. (8)) while ONMF gives more importance to data points with larger norm. For example, if the noise added to each data point is independent of its norm, ONMF should be preferred. In fact, in that situation, data points with larger norm are relatively contaminated with less noise hence should be given more importance. Another example is when data points with larger norm are statistically more significant. This is usually the case for example in document classification; assuming that each document discusses only one topic and that each topic is a distribution over the words, longer

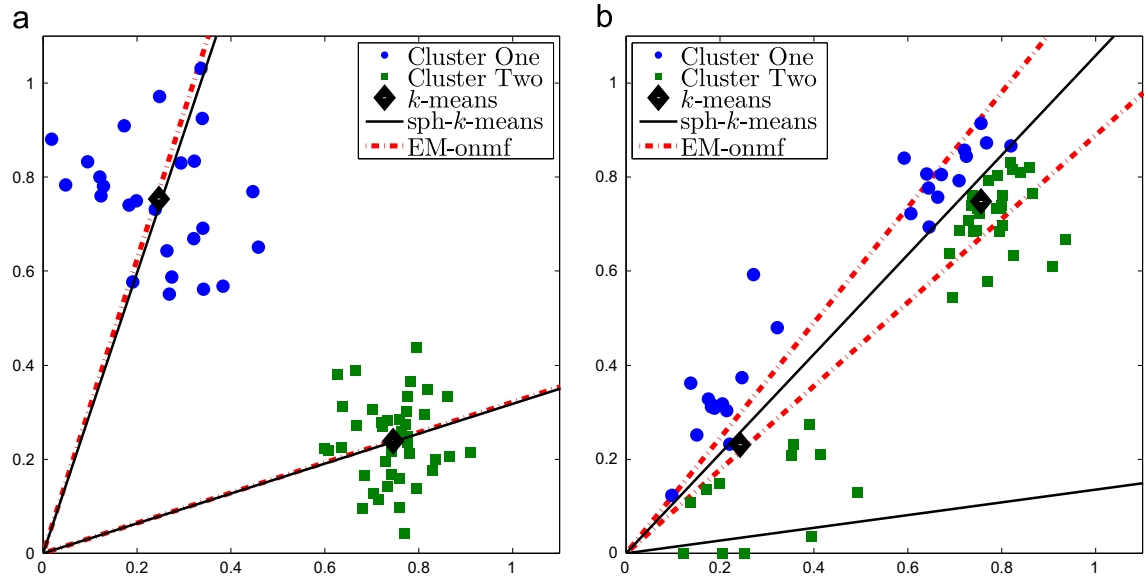

Fig. 1. Comparison of k-means, standard spherical k-means and ONMF. Diamonds are cluster centroids found by k-means, continuous lines are spherical k-means centroid directions while dashed lines are ONMF centroid directions. Circles and squares are data points as clustered by ONMF. As expected, k-means is not sensitive to the alignment of the clusters as opposed to spherical k-means and ONMF. In (a), the clusters are well separated and the three techniques perform similarly. In (b),the directional effect is clearly visible for both ONMF and spherical k-means. However, there is an important difference between the two: ONMF is more sensitive to the data points with larger norm,while spherical k-means treats all the points the same way (including the ones from the lower left cluster with smaller norm but wider angular distribution) and its centroids are therefore further apart from each other.

documents represent a larger sample of the corresponding topic distribution and should be given more importance. (In fact, most document classification software typically discards very short documents: ONMF implicitly takes care of this issue by giving less importance to shorter documents.) In hyperspectral imaging, background pixels contain mostly noise and should not be given too much importance; hence ONMF should also be preferred in this situation.

Fig. 1 displays a comparison among k-means, spherical k-means and ONMF on two simple examples.

To conclude, ONMF should be preferred to both k-means and spherical k-means when

(1) scaling should not affect cluster assignment, and
(2) data points with larger norm are more reliable and should be given more importance.

To illustrate this, we generate several synthetic data sets as follows. Each data set has six clusters $\{\pi_i\}_{i=1}^6$, each containing $100-(i-1)10$ data points for a total of 450 data points. Each cluster centroid $u_i \in \mathbb{R}^{10} 1 \le i \le 6$ is generated uniformly at random in the unit cube $[0,1]^6$. Each data point m_j $1 \le j \le 450$ is a multiple of its corresponding cluster centroid: $m_j = \alpha u_k$ where $\alpha > 0$ is picked uniformly at random in the interval $[0.1, 1]$. Hence, because of this scaling, k-means is not appropriate and will perform poorly. Each data point is then perturbed by some additive noise with fixed magnitude (i.e. independent of the norm of the data point, which should lead to ONMF performing better than spherical k-means). Concretely, each noise entry is drawn from a normal distribution with zero mean and fixed standard deviation ϵ. Finally, the negative entries of each data point are set to zero to obtain a nonnegative input matrix (note that this can only reduce the noise). Letting $\{\pi'_i\}_{i=1}^6$ be the clusters extracted by an algorithm and $\{\pi_i\}_{i=1}^6$ be the true clusters, the accuracy is defined as

$$\text{Accuracy} = \max_{P \in [1,2,\ldots,k]} \frac{1}{450}\left(\sum_{i=1}^{6} |\pi_i \cap \pi'_{P(i)}|\right) \in [0,1], \tag{9}$$

where $[1,2,\ldots,k]$ is the set of permutations of $\{1,2,\ldots,k\}$. For each noise level $\epsilon \in [0,1]$, we generate ten synthetic data sets as described above, and Fig. 2 reports the average accuracy of each algorithm using ten random initializations (except for ONMF which was solved using ONP-MF, which does not use random initialization, see Section 3).

As expected, we observe that ONMF outperforms k-means and spherical k-means. We will use the same synthetic data sets to compare the different ONMF algorithms in Section 4.

2.4. EM-like algorithm for ONMF

We present here a simple EM-like alternating algorithm designed to tackle the ONMF problem (3) based on its equivalence with the weighted variant of spherical k-means (4). It is very similar to the standard spherical k-means algorithm [2], except for the computation of cluster centroids. Specifically, it starts with an initial set of centroids, either randomly chosen or supplied as initial values. It then alternates between two steps:

1. Given cluster centroids $\{u_i\}_{i=1}^k$, choose $\{\pi_i\}_{i=1}^k$ assigning each point to its closest cluster:

$$j \in \pi_i \Rightarrow i \in \operatorname*{argmax}_{1 \le \ell \le k} (m_j^T u_\ell)^2 = \operatorname*{argmax}_{1 \le \ell \le k} (m_j^T u_\ell).$$

Notice that this step is exactly equivalent to the one of standard spherical k-means [2].

2. Given the clustering $\{\pi_i\}_{i=1}^k$, compute the new optimal cluster centroids $\{u_i\}_{i=1}^k$ as follows. Define matrix $M_i \in \mathbb{R}^{m \times |\pi_i|}$ as the submatrix of M containing the columns belonging to cluster π_i. We have to solve problem (7) with respect to the u_i's:

$$\max_{\{u_i \ge 0, ||u_i|| = 1\}_{i=1}^k} \sum_{i=1}^{k} \sum_{j \in \pi_i} (m_j^T u_i)^2 = \sum_{i=1}^{k} ||M_i^T u_i||_2^2.$$

There are k independent problems: each u_i must maximize the term $||M_i^T u_i||_2^2$. The optimal solution u_i^* is given by the dominant left singular vector of M_i associated with $\sigma_1(M_i)$, the largest

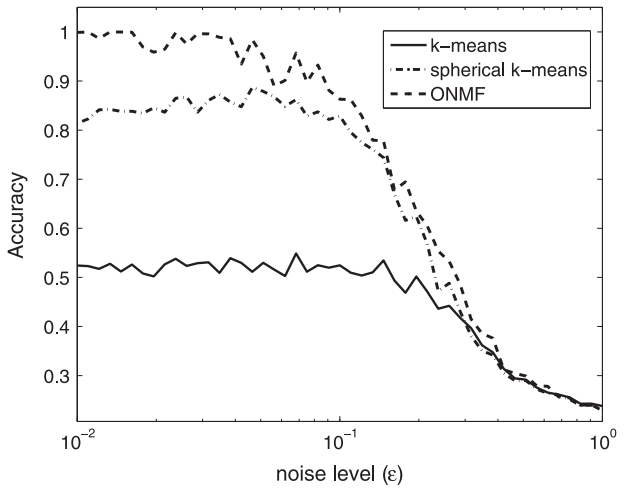

Fig. 2. Comparison of k-means, standard spherical k-means and ONMF on synthetic data sets.

singular value of M_i:

$$u_i^* = \underset{||u||_2=1}{\operatorname{argmax}} ||M_i^T u||_2^2 = \underset{||u||_2=1}{\operatorname{argmax}} u^T M_i M_i^T u,$$

for which we have $||M_i^T u_i^*||_2 = \sigma_1(M_i) = ||M_i||_2$. Moreover, since $M_i \geq 0$, the Perron–Frobenius theorem guarantees that u_i^* can always be chosen to be nonnegative.

Algorithm 1, referred to as EM-ONMF, implements this procedure. We will see in the last section that, despite its simplicity, it works well for text clustering tasks.

Algorithm 1. EM-like algorithm for ONMF (EM-ONMF).

input: Nonnegative data matrix M, and initial centroids $\{u_i\}_{i=1}^k$.
output: Clustering of the points $\{\pi_i\}_{i=1}^k$, with the corresponding centroid directions $\{u_i\}_{i=1}^k$.
while *not converged* **do**
  $\{\pi_i\}_{i=1}^k \leftarrow \emptyset$;
  for $j \leftarrow 1$ **to** n **do**
    find $i \in \operatorname{argmax}_{1 \leq \ell \leq k}(m_j^T u_\ell)$ and update cluster $\pi_i = \pi_i \cup \{j\}$.;
  end
  if $\pi_i = \emptyset$ for some i **then** randomly transfer a point to cluster π_i **end**;
  for $i \leftarrow 1$ **to** k **do**
    $u_i \leftarrow$ (any) nonnegative dominant singular vector of the data submatrix $M_i = M(:, \pi_i)$.;
  end
end

Note that Algorithm 1 does not explicitly provide a solution to ONMF. However, a candidate solution of (5) can be obtained by taking $U = [u_1 \ \ldots \ u_k]$ and selecting V according to (6), from which a candidate solution of ONMF (3) is readily obtained as $(UD, D^{-1}V)$ where $D = \operatorname{diag}(\|V(1,:)\|_2, \ldots, \|V(k,:)\|_2)$.

It is interesting to relate this with the original ONMF problem (5): given a partitioning $\{\pi_i\}_{i=1}^k$, let us denote $w_i = (v_{ij})_{j \in \pi_i}$ the subvector containing only the positive entries of the ith row of V. Then,

$$\|M - UV\|_F^2 = \sum_{i=1}^k \sum_{j \in \pi_i} \|m_j - u_i v_{ij}\|^2 = \sum_{i=1}^k ||M_i - u_i w_i^T||_F^2,$$

so that the optimal (u_i, w_i) must be an optimal solution of

$$\min_{||u_i||=1, u_i \geq 0, w_i \geq 0} ||M_i - u_i w_i^T||_F^2. \tag{10}$$

Each of these problems looks for the best nonnegative rank-one approximation of a nonnegative matrix (i.e., a rank-one NMF problem). This in turn can be solved by combining the Eckart–Young and Perron–Frobenius theorems: taking the first rank-one factor generated by the singular value decomposition (SVD) (making sure it is nonnegative in case of non-uniqueness) leads to a minimum value for (10) equal to $||M_i||_F^2 - \sigma_1^2(M_i)$. Therefore, solving ONMF amounts to finding a partitioning $\{\pi_i\}_{i=1}^k$ such that the sum of squares of the first singular values of submatrices M_i's is maximized, that is, *ONMF problem (3) is equivalent to* $\max_{\{\pi_i\}_{i=1}^k} \sum_{i=1}^k \sigma_1^2(M_i)$.

3. Augmented Lagrangian method for ONMF

In this section, we present an alternative approach to solve ONMF problems.[1] Typically, ONMF algorithms strictly enforce nonnegativity for each iterate while trying to achieve orthogonality at the limit. This can be done using a proper penalization term [10], a projection matrix formulation [20] or by choosing a suitable search direction [7]. We propose here a method working the opposite way: at each iteration, a (continuous) projected gradient scheme is used to ensure that the V iterates are orthogonal (but not necessarily nonnegative).

Nonnegativity constraints in the ONMF formulation (3) will be handled using the following augmented Lagrangian, defined for a matrix of Lagrange multipliers $\Lambda \in \mathbb{R}_+^{k \times n}$ associated to the nonnegativity constraints:

$$L_\rho(U, V, \Lambda) = \frac{1}{2}\left|\left|M - UV\right|\right|_F^2 + \langle \Lambda, -V \rangle + \frac{\rho}{2}||\min(V, 0)||_F^2, \tag{11}$$

where ρ is the quadratic penalty parameter. Ideally, we would like to solve the Lagrangian dual

$$\max_{\Lambda \geq 0} f(\Lambda) \quad \text{where } f(\Lambda) = \min_{U \geq 0, VV^T = I_k} L_\rho(U, V, \Lambda).$$

Observe that, regardless of the value of ρ, the solutions (U,V) of the ONMF problem (1) are the solutions of

$$\min_{U \geq 0, VV^T = I_k} \max_{\Lambda \geq 0} L_\rho(U, V, \Lambda).$$

We propose here a simple alternating scheme to update variables U, V, Λ while, as announced, explicitly enforcing $U \geq 0$ and $VV^T = I_k$:

1. For V and Λ fixed, the optimal U can be computed by solving a nonnegative least squares problem $U \leftarrow \operatorname{argmin}_{X \in \mathbb{R}_+^{m \times k}} \|M - XV\|_F^2$. We use the efficient active-set method proposed in[2] [14].
2. For U and Λ fixed, we update matrix V by means of a projected gradient scheme. Computing the projection of a matrix $\hat{V}$ onto the feasible set of orthogonal matrices, known as the Stiefel manifold,[3] amounts to solving the following problem:

$$\operatorname{Proj}_{St}(\hat{V}) = \underset{X}{\operatorname{argmin}} ||\hat{V} - X||_F^2 \quad \text{such that } XX^T = I_k,$$

whose optimal solution X^* can be computed in closed form from the unitary factor of a polar decomposition of $\hat{V}$, see, e.g.,

[1] Our code for the proposed algorithm is available at https://bitbucket.org/filp/onmf/src.
[2] Available at http://www.cc.gatech.edu/~hpark/.
[3] The Stiefel manifold is the set of all $n \times k$ orthogonal matrices, i.e., $\mathrm{St}(k,n) = \{X \in \mathbb{R}^{n \times k} : X^T X = I_k\}$.

[13,1]. Our projected gradient scheme then reads:

$V \leftarrow \mathrm{Proj}_{St}(V - \beta \nabla_V L_\rho(U, V, \Lambda))$,

where the step length β is chosen with a backtracking line search similar to that in [15] (step length is increased as long as there is a decrease in the objective function, and decreased otherwise).

3. Finally, Lagrange multipliers are updated in order to penalize the negative values of V:

$\Lambda \leftarrow \max(0, \Lambda - \alpha V)$.

where α is the step length. As $-V$ is the gradient of function $\Lambda \mapsto L_\rho(U, V, \Lambda)$, this update is a (projected) gradient step with step length α. We choose a predefined step length sequence $\alpha = \alpha_0/t$, where t is the iteration counter and $\alpha_0 > 0$ is a constant parameter, that satisfies the usual "square summable but not summable" condition of online gradient methods [4, (5.1)].

To initialize the algorithm, we set Λ to zero and choose for the columns of V the first k right singular vectors of the data matrix M (which can be obtained with SVD).[4] Quadratic penalty parameter ρ is initially fixed to a small value ρ_0 and then increased geometrically after each iteration. Algorithm 2 implements this procedure, which we refer to as Orthogonal Nonnegatively Penalized Matrix Factorization (ONP-MF).

Algorithm 2. Orthogonal nonnegatively penalized matrix factorization (ONP-MF).

input: A nonnegative data matrix M, the number of clusters k, $\alpha_0 > 0$, $\rho_0 > 0$ and $C > 1$.
output: The centroid matrix U, and the cluster assignment matrix V.
Initialize $\Lambda^{(0)} = 0$, the rows of $V^{(0)}$ with the first k right singular vectors of M, and $\rho = \rho_0$.;
for $t = 1, 2, \ldots$ **do**
  Update $U^{(t)}$ with the optimal solution $U^* = \mathrm{argmin}_{U \geq 0} ||M - UV^{(t-1)}||_F^2$.;
  Update $V^{(t)}$ with projected gradient and a line search for step $\beta^{(t)}$:
  $V^{(t)} \leftarrow \mathrm{Proj}_{St}[V^{(t-1)} - \beta^{(t)} \nabla_V L_\rho(U^{(t)}, V^{(t-1)}, \Lambda^{(t-1)})]$.
  Update Lagrange multipliers (using an approximate subgradient):
  $\Lambda^{(t)} \leftarrow \max\left(0, \Lambda^{(t-1)} - \frac{\alpha_0}{t} V^{(t)}\right)$.
  Update $\rho \leftarrow C\rho$.
end

We observed that the term $||\min(V, 0)||_F$ decreases linearly to zero (as augmented Lagrangian methods are expected to, see [16, Theorem 17.2]) while $||M - UV||_F$ converges to a fixed value, see Fig. 3 for an example on the Hubble data set (cf. Section 4.3). A rigorous convergence proof is a topic for further research. In fact, it is difficult to analyze an augmented Lagrangian approach when the subproblems are not solved exactly (in our case, using a single loop of a block coordinate descent method) and, as far as we know, such a proof does not exist in the literature (see, e.g., [16]) although it is a very popular method.

[4] To overcome the sign ambiguity of each row of $V^{(0)}$, we flip its sign if the ℓ_2-norm of its negative entries is larger than the ℓ_2-norm of its positive entries.

4. Numerical experiments

In this section, we report some preliminary numerical experiments showing that ONP-MF (Algorithm 2) and EM-ONMF (Algorithm 1) perform competitively with two recently proposed methods for ONMF: CHNMF from Choi [7] and O-PNMF from Yang and Oja [20] (Euclidean variant). It should be noted that because ONP-MF is initialized with SVD, its results are deterministic and obtained with just one execution of the algorithm. However, it could be argued that the comparison with the other algorithms is not completely fair as CHNMF and O-PNMF are initialized with randomly generated factors. In order to perform a fairer comparison, we also initialize CHNMF and O-PNMF with an SVD-based initialization [5] (SVD cannot be used directly because its factors are not necessarily nonnegative), which will be denoted as CH(SVD) and O-P(SVD) respectively. Finally, we also report results from two standard EM clustering algorithms, namely k-means and spherical k-means (SKM) (see, e.g., [2]). We will see that EM-ONMF is quite efficient for text clustering tasks (see Section 4.2) while ONP-MF gives very good results for unsupervised image classification tasks (see Sections 4.3 and 4.4).

Parameters for ONP-MF are chosen as follows: $\alpha_0 = 100$, $\rho_0 = 0.01$ and $C = 1.01$ for all data sets. ONMF algorithms are run until a stopping condition is met (see below), or a maximum of 5000 iterations in case of random initializations (for CHNMF and O-PNMF) and 20 000 iterations for the SVD-based initialization (as done in [5]) was reached. The following stopping condition for CHNMF seems to work well in practice[5]:

$$\frac{|\,||M - U^{(t+1)}V^{(t+1)}||_F - ||M - U^{(t)}V^{(t)}||_F\,|}{||M||_F} < 10^{-7},$$

where t is the iteration counter. For O-PNMF, we use the stopping criterion suggested by its authors[6]:

$$\frac{||V^{(t-1)} - V^{(t)}||_F}{||V^{(t-1)}||_F} < 10^{-5}.$$

For ONP-MF, we check whether the current iterate is 'sufficiently' nonnegative, using

$$\frac{||\min(V, 0)||_F}{||V||_F} < 10^{-3}.$$

All EM-like algorithms, EM-ONMF included, were run until cluster assignment did not change for two consecutive iterations. The initial centroids were randomly selected among the data points. For each experiment, a number of 30 repetitions were executed in random conditions for both ONMF- and EM-like algorithms (except for the synthetic data sets where we only performed 10 as in Section 2.3). In the image experiments, we will display the best solution obtained, i.e, with the lowest error. All experiments were run on an Intel® Core™ i7-2630QM quad core CPU @2.00 GHz with 8 GB of RAM.

4.1. Synthetic data sets

In this section, we perform experiments on synthetic data sets as in Section 2.3 in order to compare the different ONMF algorithms. Fig. 4 reports the average accuracy of each algorithm.

We observe that

- ONP-MF or O-P(SVD) perform the best among all ONMF algorithms. In particular, they are the only algorithms able to perfectly identify all clusters for small noise levels ($\epsilon = 0.01$).

[5] It seems that 10^{-7} is a good trade-off: for example, using 10^{-8} instead leads to much larger computational times without significant improvements in clustering accuracy.

[6] Code available at http://users.ics.tkk.fi/rozyang/pnmf/index.html.

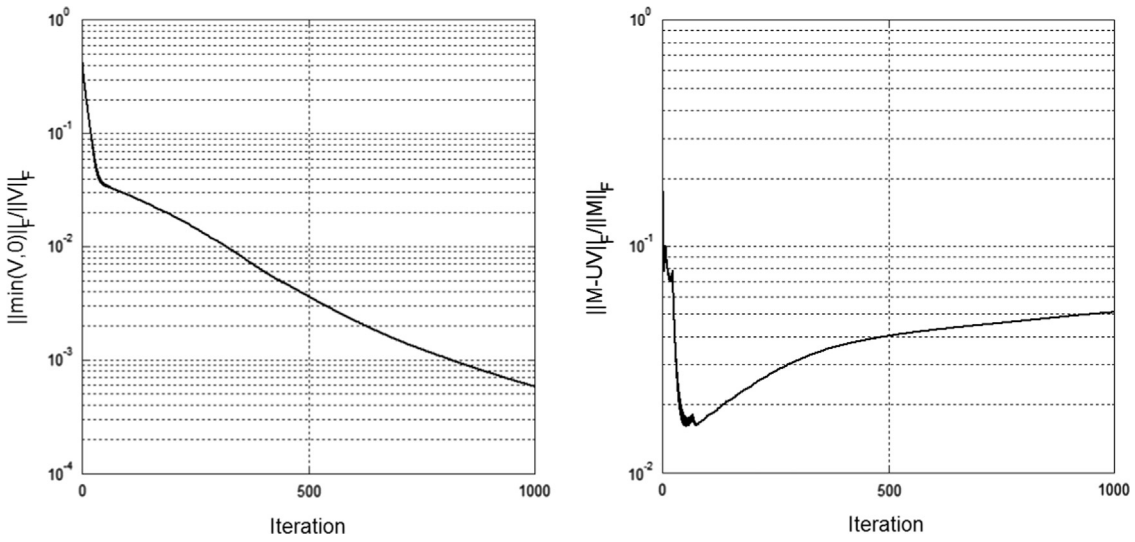

Fig. 3. Convergence of Algorithm 2 for the Hubble data set (left: constraint residual, right: approximation error).

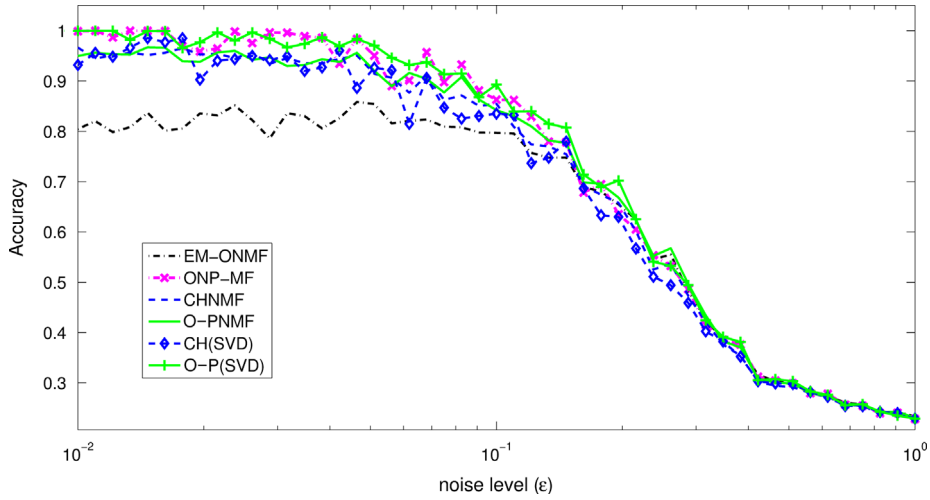

Fig. 4. Comparison of the different ONMF algorithms on synthetic data sets.

Table 1
Average computational time in seconds for the synthetic data sets.

CHNMF	CH(SVD)	O-PNMF	O-P(SVD)	EM-ONMF	ONP-MF
2.19	3.25	2.38	2.73	0.41	3.80

- CHNMF, O-PNMF and CH(SVD) perform similarly, their accuracies being in most cases lower than those of ONP-MF and O-P(SVD).
- EM-ONMF performs rather poorly and is in general not able to perform a good clustering (although it is much faster than the other algorithms, see Table 1).

Table 1 gives for each algorithm the average computational time for one execution on a synthetic data set. ONP-MF is slightly slower but typically obtains one of the best factorizations using only a single deterministic initialization.

4.2. Text clustering

We selected 12 well-known preprocessed document databases described in [21]. Each data set is represented by a term-by-document matrix of varying characteristics, see Table 2. As a performance indicator, we use the *accuracy*; see Eq. (9). We report the average value of the obtained accuracy along with the standard deviations in Table 3. For more than half of the data sets, the average best result was

Table 2
Text mining data sets [21].

Data	m	n	r	#
classic	7094	41 681	4	223 839
ohscal	11162	11 465	10	674 365
hitech	2301	10 080	6	331 373
reviews	4069	18 483	5	758 635
sports	8580	14 870	7	1 091 723
la1	3204	31 472	6	484 024
la2	3075	31 472	6	455 383
k1b	2340	21 839	6	302 992
tr11	414	6429	9	116 613
tr23	204	5832	6	78 609
tr41	878	7454	10	171 509
tr45	690	8261	10	193 605

achieved by our algorithms, either EM-ONMF or ONP-MF. Moreover, our algorithms obtain the best performance among ONMF algorithms in 10 out of the 12 data sets (being close for the two remaining data sets). While EM-ONMF is very fast with a low number of iterations (as the other EM-like algorithms), ONP-MF is in general slower than the other ONMF algorithms (especially CHNMF), and typically requires a larger number of iterations to converge. Note that, in this paper, our focus is on the comparison of ONMF algorithms based on the

Table 3
Average accuracy and standard deviation (if applicable) in percent obtained by the different algorithms (in bold, best average performance; underlined, second best).

Data set	*k*-means	SKM	CHNMF	CH(SVD)	O-PNMF	O-P(SVD)	EM-ONMF	ONP-MF
classic	**58.9** ± 6.8	56.8 ± 4.7	55.0 ± 1.8	55.9	50.8 ± 1.9	53.9	58.8 ± 6.9	53.8
ohscal	28.8 ± 3.2	**42.8** ± 2.9	33.7 ± 2.6	34.0	35.0 ± 1.1	33.8	39.2 ± 3.3	34.0
hitech	32.2 ± 1.6	**48.7** ± 3.6	42.0 ± 4.2	41.5	47.0 ± 1.0	47.7	**48.7** ± 5.6	47.0
reviews	43.6 ± 5.5	**68.7** ± 6.5	52.6 ± 9.1	49.3	55.6 ± 6.4	52.8	63.7 ± 10.3	51.0
sports	38.8 ± 2.4	45.1 ± 4.1	42.3 ± 2.9	49.5	44.3 ± 3.8	49.0	**50.0** ± 6.6	**50.0**
la1	35.0 ± 1.9	48.0 ± 4.7	53.0 ± 5.4	44.3	55.4 ± 5.8	60.9	50.2 ± 7.3	**65.8**
la2	33.8 ± 1.7	46.6 ± 4.5	41.0 ± 2.7	42.1	48.0 ± 4.6	52.7	47.1 ± 6.1	**52.8**
k1b	66.8 ± 10.1	64.9 ± 8.2	74.9 ± 2.9	76.7	72.7 ± 5.7	76.4	75.3 ± 6.9	**79.0**
tr11	31.6 ± 1.7	**53.0** ± 5.2	47.1 ± 3.3	50.2	31.5 ± 8.5	33.8	42.4 ± 6.3	46.1
tr23	40.8 ± 2.5	**42.5** ± 5.2	37.0 ± 3.6	32.8	39.2 ± 2.0	41.2	40.7 ± 4.4	40.7
tr41	41.8 ± 6.6	**53.2** ± 5.8	46.5 ± 5.5	42.6	35.5 ± 7.8	43.1	**53.2** ± 7.4	43.1
tr45	27.9 ± 4.2	**54.2** ± 6.2	39.2 ± 1.7	39.3	35.7 ± 4.7	35.1	41.4 ± 6.6	35.9

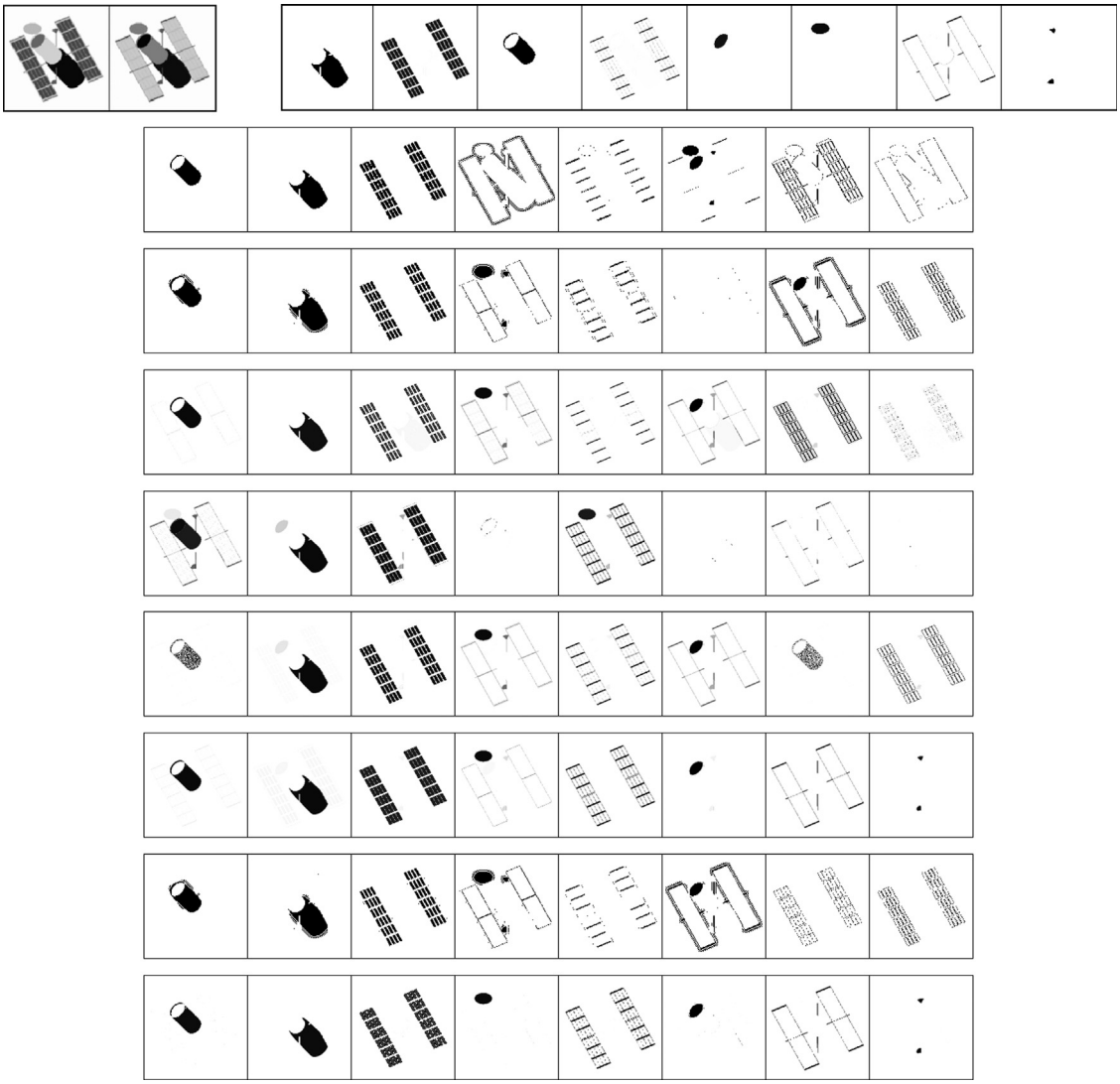

Fig. 5. Hubble data set decomposition. From top to bottom: sample images at different wavelengths along with the true constituent materials, *k*-means, spherical *k*-means, CHNMF, CH(SVD), O-PNMF, O-P(SVD), EM-ONMF and ONP-MF.

Frobenius norm. Comparison with ONMF algorithms using other measures, such as the Kullback–Leibler divergence, and comparison with other topic models (such as LDA [3]) is a topic for further research.

4.3. Hyperspectral unmixing

A hyperspectral image is a set of images of the same object or scene taken at different wavelengths. Each image is acquired by measuring the reflectance (i.e., the fraction of incident electromagnetic power reflected) of each individual pixel at a given wavelength. The aim is to classify the pixels in different clusters, each representing a different material. We want to cluster the columns of a wavelength-by-pixel reflectance matrix so that each cluster (a set of pixels) corresponds to a particular type of material.

4.3.1. Hubble telescope

We first use a synthetic data set from [17], see Fig. 5 (top row), in clean conditions (i.e., without noise or blur). It represents the Hubble telescope and is made up of 8 different materials, each having a specific spectral signature. Fig. 5 displays the clustering

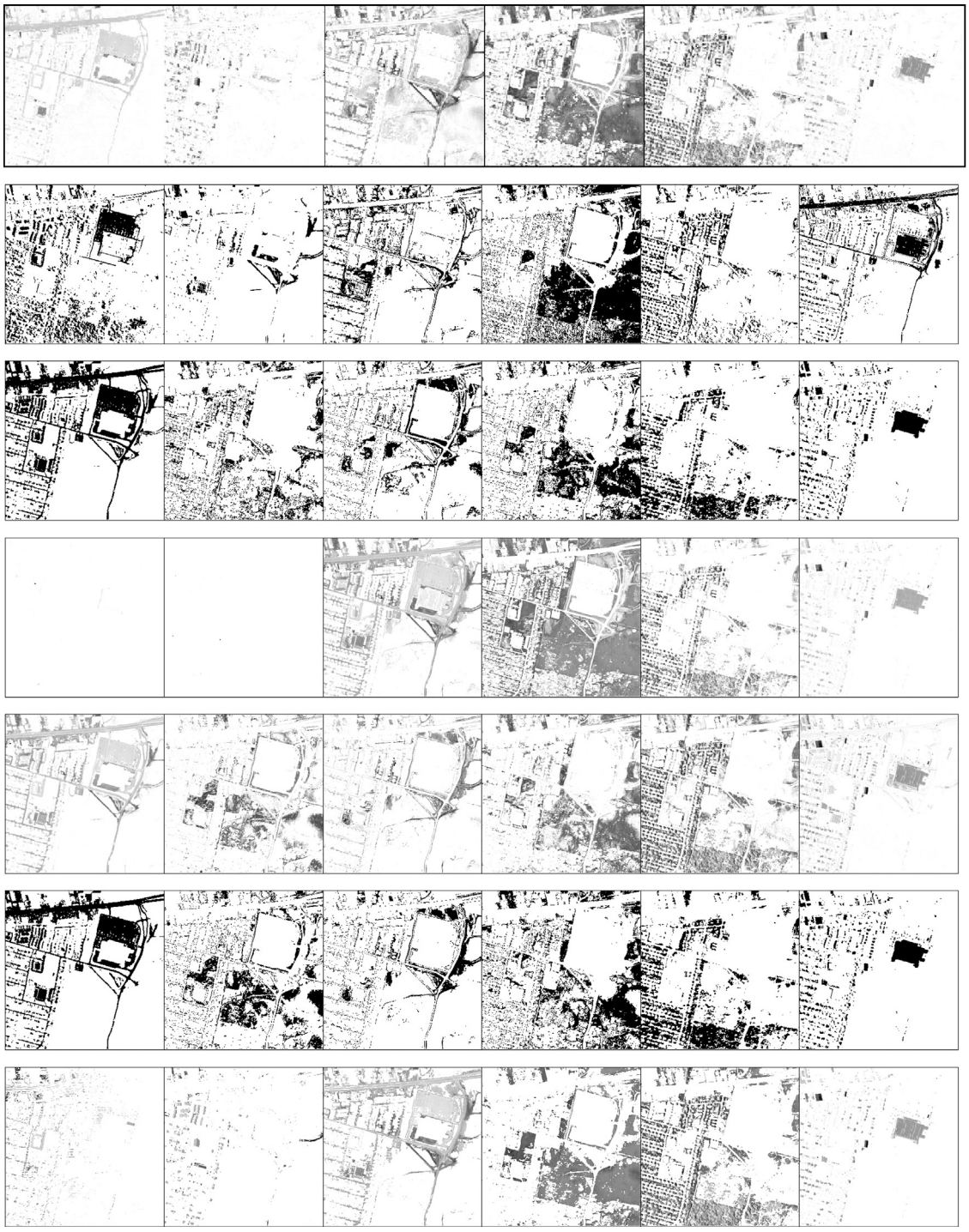

Fig. 6. Urban data set decomposition. From top to bottom: 'true' materials, *k*-means, spherical *k*-means, CHNMF, O-PNMF, EM-ONMF and ONP-MF.

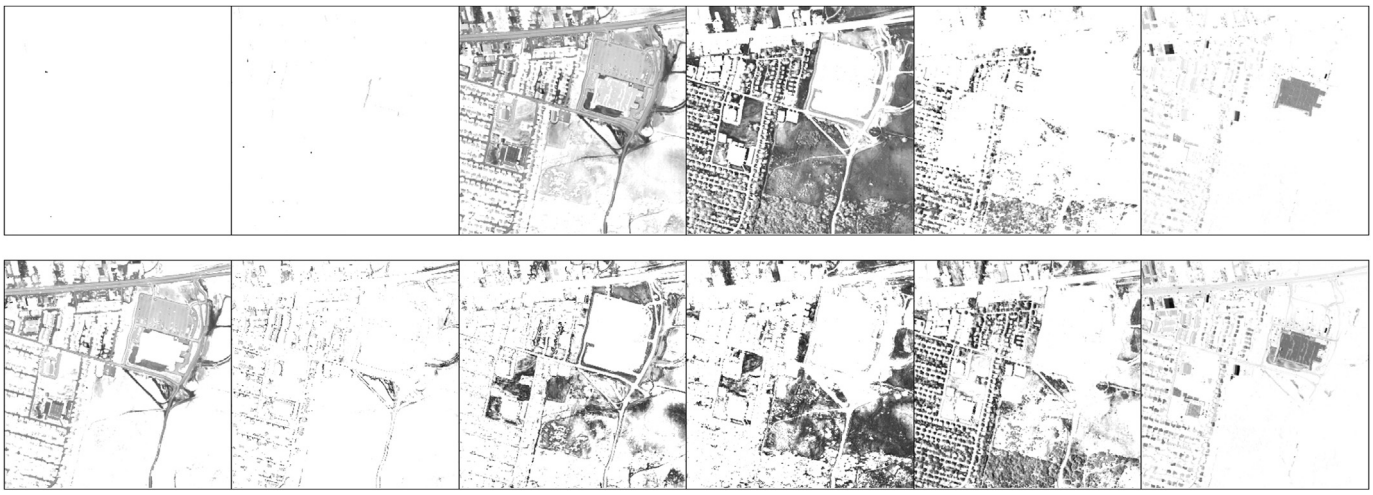

Fig. 7. Urban data set decomposition. From top to bottom: CH(SVD) and O-P(SVD).

obtained by the different algorithms[7] and we observe that only ONP-MF is able to successfully recover all eight materials without any mixing.

Even with the SVD-based initialization, CHNMF and 0-PNMF (i.e., CH(SVD) and O-P(SVD)) are not able to separate all materials properly; ONP-MF is the only algorithm able to perform this task almost perfectly.

4.3.2. Urban data set

The Urban hyperspectral image is taken from HYper-spectral Digital Imagery Collection Experiment (HYDICE) air-borne sensors. It contains 162 clean bands, and 307×307 pixels for each spectral image; it is mainly composed of 6 types of materials: road, dirt, trees, roofs, grass and metal (mostly metallic rooftops) as reported in [12,11]. The first row of Fig. 6 displays a very good clustering obtained using N-FINDR5 [19] plus manual adjustment from [12], along with the clusterings obtained with the different algorithms. The road and dirt are difficult to extract because their spectral signatures are similar (up to a multiplicative factor), and none of the algorithms is able to separate them perfectly. ONP-MF successfully extracts the grass, trees, and roofs and is the only algorithm able to extract the metal (second basis element), while only mixing the road and dirt together. Spherical k-means, CHNMF, O-P(SVD) (Fig. 7) and EM-ONMF also perform relatively well, being able to extract the road (mixed with dirt or metal), trees, grass (as two separate basis elements) and roofs. CH(SVD) and k-means perform relatively poorly: they are not able to separate as many materials properly.

4.4. Image segmentation: swimmer data set

The swimmer image data set consists of 256 binary images of a body with four limbs which can be positioned in four different ways each. The goal is to find a part-based decomposition of these images, i.e., isolate the different constitutive parts of the images (the body and the limbs, 17 in total). Moreover, these parts are not overlapping, and therefore no rows of V can share nonzero entries in the same column, and ONMF is an appropriate model. Fig. 8 displays the basis elements obtained with the different ONMF algorithms. It can be observed that, in this case, the SVD-based initialization is of no benefit, neither for CHNMF nor for O-PNMF. All algorithms are able to successfully find the correct parts except O-PNMF and O-P(SVD).

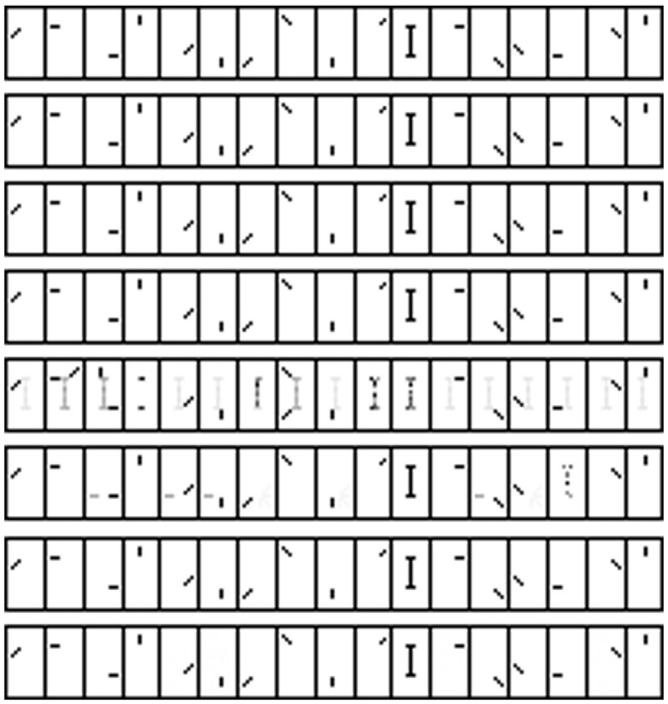

Fig. 8. Swimmer data set decomposition. From top to bottom: k-means, spherical k-means, CHNMF, CH(SVD), O-PNMF, O-P(SVD), EM-ONMF and ONP-MF.

5. Conclusion

In this paper, we have studied the ONMF problem and showed its equivalence with a weighted variant of spherical k-means (Theorem 1). This led us to design a new EM-like algorithm for solving ONMF problems (Algorithm 1, EM-ONMF). We have also proposed an alternative approach based on an augmented Lagrangian method imposing orthogonality at each step while relaxing the non-negativity constraint (Algorithm 2, ONP-MF).

[7] For EM-ONMF, k-means and SKM we preprocess the data by discarding pixels from the background (i.e., all columns of the input matrix with zero ℓ_2-norm). Recall that, for each algorithm, we keep the best solution (w.r.t. the error) among the 30 randomly generated initial matrices.

We then performed numerical experiments on some synthetic, text and image data sets. Note that Euclidean-based metric ONMF (3) is not particularly suited for document classification: First, the use of the Frobenius norm makes the implicit assumption that the noise is Gaussian, which is unrealistic for sparse data sets such as document data sets; see, e.g., the discussion in [6]. Second, ONMF assumes that each document is on a single topic, which is in general not true (there exist more general generative models such as LDA assuming that documents are mixtures of several topics). Text data sets are nevertheless a worthy benchmark on which to compare the effectiveness of various Euclidean-metric ONMF algorithms.

The experiments indicate that our ONP-MF algorithm is by far the most robust among existing algorithms for solving the Euclidean ONMF problem (1): it always gave very good results, the best in many cases, using only one initialization. In particular, we observed for all image experiments that a single (deterministic) run of ONP-MF worked better than all the other tested algorithms, despite the fact that those were allowed to keep the best solution obtained from 30 different (random) initializations. Since initialization is known to be an important component in the design of successful NMF methods [5], we believe that initializing the V factor with the unaltered right singular vectors of the data matrix, which is allowed by the workings of ONP-MF but impossible with other ONMF methods, plays an instrumental role in the clustering performance of ONP-MF observed in numerical experiments.

Acknowledgments

The authors would like to thank the guest editor and the reviewers for their feedback which helped improve the paper.

References

[1] P.-A. Absil, Jérôme Malick, Projection-like retractions on matrix manifolds, SIAM J. Optim. 22 (2012) 135–158.
[2] A. Banerjee, I. Dhillon, J. Ghosh, S. Sra,Generative model-based clustering of directional data, in: Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-03), ACM Press, 2003, pp. 19–28.
[3] D.M. Blei, A.Y. Ng, M.I. Jordan, Latent dirichlet allocation, J. Mach. Learn. Res. 3 (2003) 993–1022.
[4] L. Bottou, Online Algorithms and Stochastic Approximations, in: D. Saad (Ed.), Online learning and neural networks, Cambridge University Press, Cambridge, 1998.
[5] C. Boutsidis, E. Gallopoulos, Svd based initialization: a head start for nonnegative matrix factorization, Pattern Recognit. 41 (4) (2008) 1350–1362.
[6] E.C. Chi, T.G. Kolda, On tensors, sparsity, and nonnegative factorizations, SIAM J. Matrix Anal. Appl. 33 (2012) 1272–1299.
[7] S. Choi, Algorithms for orthogonal nonnegative matrix factorization, in: Proceedings of the International Joint Conference on Neural Networks, 2008, pp. 1828–1832.
[8] I.S. Dhillon, Y. Guan, B. Kulis, Weighted graph cuts without eigenvectors: a multilevel approach, IEEE Trans. Pattern Anal. Mach. Intell. 29 (11) (2007) 1944–1957.
[9] C. Ding, X. He, On the equivalence of nonnegative matrix factorization and spectral clustering, in: Proceedings of the Fifth SIAM Conference on Data Mining, 2005, pp. 606–610.
[10] C. Ding, T. Li, W. Peng, H. Park,Orthogonal nonnegative matrix tri-factorizations for clustering, in: Proceedings of the Twelfth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Philadelphia, PA, USA, August 20–23, 2006, 2006, pp. 126–135.
[11] N. Gillis, R.J. Plemmons, Dimensionality reduction, classification, and spectral mixture analysis using nonnegative underapproximation, Opt. Eng. 50 (2011) 027001.
[12] Z. Guo, T. Wittman, S. Osher, L1 unmixing and its application to hyperspectral image enhancement, in: Proceedings of SPIE Conference on Algorithms and Technologies for Multispectral, Hyperspectral, and Ultraspectral Imagery XV, 2009.
[13] R.A. Horn, C. Johnson, Matrix Analysis, Cambridge University Press, Cambridge, 1990.
[14] J. Kim, H. Park, Fast nonnegative matrix factorization: an active-set-like method and comparisons, SIAM J. Sci. Comput. 6 (2011) 3261–3281.
[15] C.-J. Lin, Projected gradient methods for nonnegative matrix factorization, Neural Comput. 19 (2007) 2756–2779.
[16] J. Nocedal, S.J. Wright, Numerical Optimization, Second Ed., Springer, New York, 2006.
[17] V.P. Pauca, J. Piper, R.J. Plemmons, Nonnegative matrix factorization for spectral data analysis, Linear Algebra Appl. 406 (1) (2006) 29–47.
[18] F. Pompili, N. gillis, P.-A. Absil, F. Glineur, Onp-mf: An orthogonal nonnegative matrix factorization algorithm with application to clustering, in: 21st European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, 2013.
[19] M. Winter, N-findr: an algorithm for fast autonomous spectral end-member determination in hyperspectral data, in: Proceedings of SPIE Conference on Imaging Spectrometry V, 1999.
[20] Z. Yang, E. Oja, Linear and nonlinear projective nonnegative matrix factorization, Trans. Neural Netw. 21 (2010) 734–749.
[21] S. Zhong, J. Ghosh, Generative model-based document clustering: a comparative study, Knowl. Inf. Syst. 8 (3) (2005) 374–384.

Filippo Pompili received a Master degree in engineering (2008) and a Ph.D. in computer science (2012) from Università degli Studi di Perugia, Italy. He has been working on medical image and data analysis problems and he is currently conducting industrial applied research for natural language processing and information retrieval. His research interests lie in machine learning, data mining, and numerical optimization.

Nicolas Gillis received a master degree and a Ph.D. degree in applied mathematics from Université catholique de Louvain (Belgium) in 2007 and 2011, respectively. He is currently an assitant professor at the Department of Mathematics and Operational Research, Faculté polytechnique, Université de Mons, Belgium. His research interests lie in optimization, numerical linear algebra, machine learning and data mining.

P.-A. Absil is professor of applied mathematics at Université catholique de Louvain in Louvain-la-Neuve, Belgium. He received an engineering degree (1998) and a PhD in applied sciences (2003) from the University of Liège, Belgium. His main research area is numerical optimization, with particular interests in numerics on manifolds, linear and nonlinear programming algorithms, and biomedical applications.

François Glineur received engineering degrees from Faculté Polytechnique de Mons (Belgium) and École supérieure délectricité (France) in 1997, and a Ph.D. in applied sciences from Faculté Polytechnique de Mons in 2001. He is currently professor of applied mathematics at the Engineering School of Université catholique de Louvain, member of the Center for Operations Research and Econometrics and the Institute of Information and Communication Technologies, Electronics and Applied Mathematics. His main research interests lie in optimization models and algorithms (with a focus on convex optimization) and their engineering applications, as well as in nonnegative matrix factorization.