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

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

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)

$$\min_{U\ge 0,V\ge 0}\|M-UV\|_F^2\quad\text{s.t. } V\text{ satisfies (c1)},\tag{2}$$

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):

$$VV^T=I_k\quad\text{and}\quad V\ge 0.\tag{c2}$$

It can be easily checked that (c1) ⇒ (c2) while (c2)⇏(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:

$$\min_{U\ge 0,V\ge 0}\|M-UV\|_F^2\quad\text{such that } VV^T=I_k.\tag{3}$$

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.