Journal of Computational and Applied Mathematics 260 (2014) 149–166

Contents lists available at ScienceDirect

Journal of Computational and Applied Mathematics

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

A convergent algorithm for orthogonal nonnegative matrix factorization

Andri Mirzal *

Faculty of Computing, Universiti Teknologi Malaysia, 81310 Johor Bahru, Malaysia

ARTICLE INFO

Article history:
Received 28 May 2012
Received in revised form 31 July 2013

MSC:
65F30
15A23

Keywords:
Nonnegative matrix factorization
Orthogonality constraint
Convergent algorithm
Clustering methods

ABSTRACT

This paper proposes a convergent algorithm for nonnegative matrix factorization (NMF) with orthogonality constraint on the factors. We design the algorithm based on the additive update rule algorithm for the standard NMF proposed by Lee and Seung, and derive the convergent version by generalizing the convergence proof of the algorithm developed by Lin. Further we use the proposed algorithms to improve clustering capability of the standard NMF using the Reuter document corpus, a standard dataset in clustering research.

1. Introduction

The nonnegative matrix factorization (NMF) is a technique that decomposes a nonnegative data matrix into a pair of other nonnegative matrices with lower ranks:

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

where $\mathbf{A} \in \mathbb{R}_+^{M\times N} = [\mathbf{a}_1, \ldots, \mathbf{a}_N]$ denotes the data matrix, $\mathbf{B} \in \mathbb{R}_+^{M\times R} = [\mathbf{b}_1, \ldots, \mathbf{b}_R]$ denotes the basis matrix, $\mathbf{C} \in \mathbb{R}_+^{R\times N} = [\mathbf{c}_1, \ldots, \mathbf{c}_N]$ denotes the coefficient matrix, R denotes the number of factors which usually is chosen so that $R \ll \min(M, N)$, and $\mathbb{R}_+^{M\times N}$ denotes M by N nonnegative real matrix. To compute $\mathbf{B}$ and $\mathbf{C}$, Eq. (1) is usually rewritten into a minimization problem in Frobenius norm criterion:

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

The NMF was first studied by Paatero et al. [1,2], and made popular by Lee and Seung in which the authors introduced a simple algorithm based on multiplicative update rules (MUR) and demonstrated its uses in two application domains: learning the parts of objects and discovering semantic features of text [3]. In the subsequent work [4], Lee and Seung gave an analysis of the algorithm in which they showed that the algorithm has nonincreasing property, a necessary but not sufficient condition for the convergence. The NMF has been successfully extended and used in many problem domains including clustering and classification [5–18], image analysis [19–23], spectral analysis [24–30], and blind source separation [31–35].

As shown in many works, the performance of the standard NMF in Eq. (2) (we will refer Eq. (2) as the standard NMF for the rest of this paper) often can be improved by appending auxiliary constraints such as sparseness [19,20,8,9],

* Tel.: +60 75532260.
E-mail addresses: andrimirzal@utm.my, andrimirzal@gmail.com.

0377-0427/$ – see front matter
http://dx.doi.org/10.1016/j.cam.2013.09.022

smoothness [5,24,26,33], and orthogonality [36–39,27] into its formulation. The auxiliary constraints are usually designed so that the computed solutions can be directed to have some characteristics (e.g., sparseness constraint is enforced when parts-based representations are desired [19], smoothness constraint is enforced to reduce influence of noises [25], and orthogonality constraint is enforced to improve clustering capability of the NMF [36]). For NMF objective functions that are differentiable, due to the convenience of deriving the algorithms directly from the corresponding objective functions, most works proposed MUR based algorithms. However because MUR based NMF algorithms do not have convergence guarantee [25,40] (see Section 2 for the detailed discussion), developing convergent algorithms for various NMF objective functions is still an open research problem.

Orthogonal NMFs were introduced by Ding et al. [36] to enforce orthogonality constraints on columns of $\mathbf{B}$ and/or rows of $\mathbf{C}$ in order to improve clustering capability of the standard NMF. Because clustering indicator matrices are orthogonal (in hard clustering case), imposing orthogonality on columns of $\mathbf{B}$ (rows of $\mathbf{C}$) can potentially produce a sharper row clustering indicator matrix (column clustering indicator matrix), and therefore it is expected that this mechanism can lead to better clustering methods. However, as the original orthogonal NMF algorithms [36] and the variants [37–39] are all based on the MUR, there is no convergence guarantee for these algorithms. And because the orthogonality constraints cannot be recast into alternating nonnegativity-constrained least square (ANLS) framework (see [9,11] for discussion on ANLS), convergent algorithms for the standard NMF, e.g., [11,40–44] cannot be utilized to solving the orthogonal NMF problems. Thus, there is still no convergent algorithm for orthogonal NMF.

This fact motivates us to develop a convergent algorithm for orthogonal NMF. The proposed algorithm is designed by generalizing the work of Lin [40] in which he provided a convergence algorithm for the standard NMF based on a modified version of the additive update rules (AUR) framework introduced in Ref. [4]. The generalization presented in this paper is not trivial since the proof is developed in matrix form, thus providing a framework for developing convergence algorithms for other NMF objective functions that have matrix based auxiliary constraints with mutually interdependency between the columns and/or the rows (Lin used vector form for developing the proof, so the interdependency between the columns and/or the rows cannot be captured). Also, in the process of developing the proof, the objective function needs to be decomposed into the Taylor series. When the objectives have only up to second order derivatives, the nonincreasing properties can be proven by showing the positive-definiteness of the Hessian matrices of the objectives [4,40]. But in general cases, the objectives can have more than second order derivatives. And in particular, orthogonality constraint makes the objective has more than second order derivative. Thus, the same strategy can no longer be used for the general cases. Accordingly, we introduce a strategy to deal with this kind of objective functions. Note that the proof presented here is sufficiently general to be a framework for developing convergence algorithms for other NMF objectives with well-defined partial derivatives up to second order.

2. Multiplicative update algorithm

In [4], Lee and Seung introduced two MUR algorithms for the standard NMF using the Frobenius norm and the Kullback–Leibler divergence as the distance measures. They also showed how to modify the MUR algorithms into the corresponding AUR versions. However, due to numerical difficulties of the Kullback–Leibler divergence, and additional computational requirements of the AUR algorithms, only the Frobenius norm based MUR algorithms are being extensively studied. In this section, we review the Frobenius norm based MUR algorithm and discuss the reason why this kind of algorithms does not have convergence guarantee. As a tool, we utilize the Karush–Kuhn–Tucker (KKT) optimality conditions which must be satisfied by any stationary point.

The KKT function for the standard NMF objective is:

$$L(\mathbf{B}, \mathbf{C}) = J(\mathbf{B}, \mathbf{C}) - \operatorname{tr}\,(\mathbf{\Gamma_B} \mathbf{B}^T) - \operatorname{tr}\,(\mathbf{\Gamma_C} \mathbf{C}),$$

where $\mathbf{\Gamma_B} \in \mathbb{R}_+^{M\times R}$ and $\mathbf{\Gamma_C} \in \mathbb{R}_+^{N\times R}$ denote the KKT multipliers, and $\operatorname{tr}\,(\mathbf{X})$ denotes trace of matrix $\mathbf{X}$. Partial derivatives of L with respect to $\mathbf{B}$ and $\mathbf{C}$ can be written as:

$$\nabla_{\mathbf{B}} L(\mathbf{B}) = \nabla_{\mathbf{B}} J(\mathbf{B}) - \mathbf{\Gamma_B}, \quad \text{and}$$
$$\nabla_{\mathbf{C}} L(\mathbf{C}) = \nabla_{\mathbf{C}} J(\mathbf{C}) - \mathbf{\Gamma_C}^T,$$

with

$$\nabla_{\mathbf{B}} J(\mathbf{B}) = \mathbf{B}\mathbf{C}\mathbf{C}^T - \mathbf{A}\mathbf{C}^T, \quad \text{and}$$
$$\nabla_{\mathbf{C}} J(\mathbf{C}) = \mathbf{B}^T\mathbf{B}\mathbf{C} - \mathbf{B}^T\mathbf{A}.$$

By results from optimization studies, $(\mathbf{B}^*, \mathbf{C}^*)$ is a stationary point of Eq. (2) if it satisfies the KKT optimality conditions [45], i.e.:

$$\begin{aligned} &\mathbf{B}^* \geq \mathbf{0}, \qquad \mathbf{C}^* \geq \mathbf{0}, \\ &\nabla_{\mathbf{B}} J(\mathbf{B}^*) = \mathbf{\Gamma_B} \geq \mathbf{0}, \qquad \nabla_{\mathbf{C}} J(\mathbf{C}^*) = \mathbf{\Gamma_C}^T \geq \mathbf{0}, \\ &\nabla_{\mathbf{B}} J(\mathbf{B}^*) \odot \mathbf{B}^* = \mathbf{0}, \qquad \nabla_{\mathbf{C}} J(\mathbf{C}^*) \odot \mathbf{C}^* = \mathbf{0}, \end{aligned} \tag{3}$$