Applied Numerical Mathematics 85 (2014) 1–15

Contents lists available at ScienceDirect

Applied Numerical Mathematics

www.elsevier.com/locate/apnum

Non-negative Matrix Factorization under equality constraints—a study of industrial source identification

A. Limem [a,b], G. Delmaire [a,b,*], M. Puigt [a,b], G. Roussel [a,b], D. Courcot [a,c]

[a] Université Lille Nord de France, 59655 Villeneuve d'Ascq, France

[b] Laboratoire d'Informatique Signal Image de la Côte d'Opale, EA 4491, Université du Littoral Côte d'Opale, Maison de la Recherche Blaise Pascal, 50 rue Ferdinand Buisson, BP 699, 62228 Calais Cedex, France

[c] Unité de Chimie Environnementale et Interactions sur le Vivant, EA 4492, Université du Littoral Côte d'Opale, Maison de la Recherche en Environnement Industriel de Dunkerque, 145 Avenue Maurice Schumann, 59140 Dunkerque, France

ARTICLE INFO

Article history:
Received 5 March 2014
Received in revised form 30 May 2014
Accepted 31 May 2014
Available online 16 June 2014

Keywords:
Non-negative Matrix Factorization
Quadratic Optimization
Air quality

ABSTRACT

This work is devoted to the factorization of an observation matrix into additive factors, respectively a contribution matrix G and a profile matrix F which enable to identify many pollution sources. The search for G and F is achieved through Non-negative Matrix Factorization techniques which alternatively look for the best updates on G and F.

These methods are sensitive to noise and initialization, and—as for any blind source separation method—give results up to a scaling factor and a permutation. A Weighted Non-negative Matrix Factorization extension has also been proposed in the literature, so that different standard deviations of the data matrix components are taken into account. However, some estimated profile components may be inconsistent with practical experience. To prevent this issue, we propose an informed Non-negative Matrix Factorization, where some components of the profile matrix are set to zero or to a constant positive value. A special parametrization of the profile matrix is developed in order to freeze some profile components and to let free the other ones.

The problem amounts to solve a family of quadratic sub-problems. A Maximization Minimization strategy leads to some global analytical expressions of both factors.

These techniques are used to estimate source contributions of airborne particles from both industrial and natural influences. The relevance of the proposed approach is shown on a real dataset.

1. Introduction

Approximate matrix factorization techniques are of great importance in several areas, such as pattern recognition, antenna array processing, or environmental data processing. Non-negative Matrix Factorization (NMF) aims to factorize a non-negative matrix as the product between two matrices, such that each entry of the latter matrices is null or positive. It emerged under the name Positive Matrix Factorization (PMF) with Paatero and Tapper's contributions [29,30]. It then appeared in the field of signal and image processing with the work by Lee and Seung [24]. Some of the most well known

* Corresponding author at: Laboratoire d'Informatique Signal Image de la Côte d'Opale, EA 4491, Université du Littoral Côte d'Opale, Maison de la Recherche Blaise Pascal, 50 rue Ferdinand Buisson, BP 699, 62228 Calais Cedex, France.

E-mail addresses: abdelhakim.limem@lisic.univ-littoral.fr (A. Limem), gilles.delmaire@lisic.univ-littoral.fr (G. Delmaire), matthieu.puigt@lisic.univ-littoral.fr (M. Puigt), gilles.roussel@lisic.univ-littoral.fr (G. Roussel), dominique.courcot@univ-littoral.fr (D. Courcot).

http://dx.doi.org/10.1016/j.apnum.2014.05.009
0168-9274/

approaches have been proposed by Cichocki and Zdunek [7] and Lin [26], and are based on a projected gradient method. Recently, some weighted versions of NMF were investigated and took into account specific standard deviations associated to each data point. Inconsistencies in practical results lead to investigate additional assumptions such as sparseness constraints [20,2], fixed row and column sums constraints [18], or orthogonality constraints [35,5].

Over the last decade, matrix factorization has been extensively investigated in the field of atmospheric sciences, and particularly for receptor models with the aim to re-construct the impacts of emissions from different pollutant sources, e.g., airborne particles [19]. Such source apportionment studies enable to identify the main emission sources and their relative contribution at different monitoring sites. However, Viana et al. [33] reported that, depending on the applied NMF method, these techniques may provide highly different solutions. Moreover—and to the best of our knowledge—NMF tools fail in discriminating some natural sources (i.e., crustal matter from local or regional re-suspension and road dust), as well as sources containing sulfate and nitrate components. Moreover, the separation of mineral particles emitted from both industrial dust emissions and natural sources appears as a difficult task [9]. In this paper, we propose to take into account some partial *a priori* knowledge on some source profiles in order to improve the estimation to the unknown components.

The remainder of the paper is organized as follows. In Section 2, we present the concepts of some state-of-the-art NMF methods. Section 3 introduces our proposed approach while its performance is investigated in Section 4. We conclude and discuss future directions of this work in Section 5.

2. A short review of Non-negative Matrix Factorization

NMF aims to estimate two non-negative matrices whose product approximates the observed one. The non-negativity constraints make the decomposition purely additive whereas other ones like Principal Components Analysis (PCA) and Independent Components Analysis (ICA) can provide negative components. Informed NMF is located between NMF methods—which only assume non-negativity of the matrices—on one hand, and on the other hand, regression models which assume that one matrix is exactly known. In this paper, we introduce an informed NMF method, since matrices may be partially known.

2.1. Receptor modelling

This model is very generic since it is often used in signal and image processing. In environmental field, the factorization model (1) below is best known as receptor modelling and links the data matrix and the active sources. The factorization enables to approximate the $n \times m$ data matrix X by the product of two matrices,

$$X \approx G \cdot F, \tag{1}$$

where

- X is the $n \times m$ data matrix, where n is the number of samples and m is the number of species. In environmetrics, x_{ij}—the (i, j)th element of X—accounts for the concentration, expressed in ng/m^3, of the jth chemical species coming from the ith sample. A row gathers concentrations of all registered species for the current sample.
- G stands for the $n \times p$ contribution matrix, where p is the pre specified number of sources and n is the number of samples. In environmetrics and chemometrics, the (i, k)th element of G—denoted g_{ik}—is referred to as the massic contribution from Source k to Sample i, expressed in µg/m^3.
- F is a $p \times m$ matrix of profiles. Its (k, j)th term f_{kj} stands for a mass ratio (in ng/µg) corresponding to a percentage of the jth species with respect to the whole mass of the Source k.

The number p of sources is usually chosen such that $np + pm \ll nm$. The *a priori* choice for the appropriate value of p mostly depends on the nature of the investigated data [16] and on the expert's knowledge of the number of sources. In the absence of such information, p can be estimated in a pre-processing stage by applying some information theory techniques [11] such as AIC or BIC or extended versions.

2.2. Non-negative matrix factorization

Non-negative matrix factorization has received a lot of attention since Lee and Seung [24] published their popular multiplicative algorithm. In fact, NMF looks for an approximate factorization of the data matrix according to receptor modelling (1) under the additional non-negativity constraint. The basic NMF solves

$$\min_{G,F \geq 0} \sum_{i=1}^{n} \sum_{j=1}^{m} \big((x_{ij} - (GF)_{ij})\big)^2 = \min_{G,F \geq 0} \|X - G \cdot F\|_{\mathrm{f}}^2, \tag{2}$$