Neurocomputing 138 (2014) 142–156

Contents lists available at ScienceDirect

Neurocomputing

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

A new Bayesian approach to nonnegative matrix factorization: Uniqueness and model order selection

R. Schachtner [a,d], G. Pöppel [a], A.M. Tomé [b], C.G. Puntonet [c], E.W. Lang [d,*]

[a] Infineon Technologies AG, 93049 Regensburg, Germany
[b] IEETA, DETI, Universidade de Aveiro, P-3810-193 Aveiro, Portugal
[c] DATC/ESTII, Universidad de Granada, E-18071 Granada, Spain
[d] CIML Group, Biophysics Department, University of Regensburg, D-93040 Regensburg, Germany

ARTICLE INFO

Article history:
Received 11 December 2012
Received in revised form
29 January 2014
Accepted 3 February 2014
Communicated by A. Prieto
Available online 8 April 2014

Keywords:
Bayes NMF
Variational Bayes
Bayesian optimality criterion
Generalized Lee–Seung update rules

ABSTRACT

NMF is a blind source separation technique decomposing multivariate non-negative data sets into meaningful non-negative basis components and non-negative weights. There are still open problems to be solved: uniqueness and model order selection as well as developing efficient NMF algorithms for large scale problems. Addressing uniqueness issues, we propose a Bayesian optimality criterion (BOC) for NMF solutions which can be derived in the absence of prior knowledge. Furthermore, we present a new Variational Bayes NMF algorithm *VBNMF* which is a straight forward generalization of the canonical Lee–Seung method for the Euclidean NMF problem and demonstrate its ability to automatically detect the actual number of components in non-negative data.

1. Introduction

Decomposing any given data set into nonnegative factors has received wide-spread interest within the machine learning community in recent years [16]. The idea first appeared as positive matrix factorization in [50,49] and became popular as nonnegative matrix factorization (NMF) through the seminal paper by Lee and Seung [40]. Similar objectives were pursued also by nonnegative independent component analysis [51,52] or rectified factor analysis [29]. Originally, NMF was introduced as an unsupervised, parts-based learning paradigm involving the approximative decomposition of a nonnegative matrix $\mathbf{X}$ into a product of two nonnegative matrices, $\mathbf{W}$ and $\mathbf{H}$, via a multiplicative update algorithm. All these data analysis tools account for the fact that many physical measurements yield results with exclusively non-negative quantities which can be approximated by a non-subtractive superposition of exclusively non-negative underlying features which explain the systematic structure of the data set. Also for physical reasons the superimposed components cannot partially compensate each other. Hence decomposing any measurements into a parts-based representation seems more natural than other constraints often invoked in exploratory matrix factorization procedures which would allow for partial compensation of the components upon their superposition.

NMF has seen numerous applications in recent years [16,8] where it has been primarily applied in an unsupervised setting in image and natural language processing [42,58,13] and sparse coding [33,70]. Further applications include text mining [7] and music transcription [68]. More recently, it has been successfully utilized also in a variety of applications in computational biology [17], especially in molecular pattern discovery [56,57,59].

Basically, NMF can be formulated as a minimization problem. A suitable cost function such as the quadratic error function D_E

$$D_E(\mathbf{X},\mathbf{WH}) = \frac{1}{2}\sum_i\sum_j (X_{ij} - [WH]_{ij})^2 \qquad (1)$$

comes to be minimized subject to the constraint that either factor matrix has only non-negative entries, i.e. $\mathbf{W} \geq 0$, $\mathbf{H} \geq 0$. The cost function D_E quantifies the reconstruction error $\mathbf{E}$ between an $(N \times M)$-dimensional non-negative data matrix $\mathbf{X}$ and the product of an $(N \times K)$-dimensional weight matrix $\mathbf{W}$ and a $(K \times M)$-dimensional matrix $\mathbf{H}$ of hidden features or sources according to the data model:

$$\mathbf{X} = \mathbf{WH} + \mathbf{E} \quad \text{subject to } \mathbf{W} \geq 0,\ \mathbf{H} \geq 0. \qquad (2)$$

Vavasis [74] recently proofed that this problem is NP-hard. Without the nonnegativity constraint, singular value decomposition (SVD) provides an optimal solution to the factorization problem. Arora et al. recently discussed conditions when the NMF problem can be solved in polynomial time [4]. As posed above, NMF can be

[*] Corresponding author.
E-mail addresses: reinhard.schachtner@infineon.com (R. Schachtner), gerhard.poeppel@infineon.com (G. Pöppel), ana@ua.pt (A.M. Tomé), carlosgp@ugr.es (C.G. Puntonet), elmar.lang@biologie.ur.de (E.W. Lang).

http://dx.doi.org/10.1016/j.neucom.2014.02.021
0925-2312/© 2014 Elsevier B.V. All rights reserved.

considered a constrained, unsupervised feature extraction technique. However, usually K is chosen such that $K(N+M) \ll NM$ which brings clustering aspects into play as well but also raises the question of proper model order selection (MOS).

The cost function (1) is just the most popular choice for NMF as it is based on the Frobenius norm of the reconstruction error of the data matrix factorization as a proper distance measure. Other popular cost functions are based on information theoretic concepts like the generalized Kullback–Leibler divergence, the Itakura–Saito divergence or similar divergences [15,18] (see [16] for a recent review). Positivity constraints are, however, rarely sufficient to extract the underlying features uniquely. Hence, further constraints based on sparseness considerations [32,70,27] or minimum volume requirements [63,3] have been considered in addition. Such regularizing constraints have been added to enforce certain characteristics of the solutions, or to impose prior knowledge about the application considered. Technically, additional constraints can be realized via additional penalty terms in the NMF cost function, for example

$$D_{constr}(\mathbf{X}, \mathbf{WH}) = \frac{1}{2}\sum_i \sum_j (X_{ij} - [WH]_{ij})^2 + \lambda_W f(\mathbf{W}) + \lambda_H g(\mathbf{H}) \quad (3)$$

where the scalars λ_W, λ_H determine the balance between reconstruction accuracy and the desired properties of the factor matrices which are expressed by suitable functions f and g.

Also various optimization techniques and sampling procedures such as Expectation–Maximization (EM), Markov Chain Monte Carlo (MCMC) or variational Bayes (VB) [19], based on the Indian Buffet Process factor analysis model [37,28] or nonparametric Bayesian approaches [30] have been presented since. To further improve the performance of NMF algorithms and to reduce the risk of getting stuck in local minima, a multi-layer technique has been advocated recently [15,14]. This technique has been proposed earlier already in connection with hybridizing sparse NMF with a genetic algorithm to optimize the related discontinuous cost function [71].

Proofs of convergence of NMF algorithms are scarce (see the discussions in [69,8]) though for the popular class of multiplicative update algorithms convergence to a stationary point could be proven [43]. Also uniqueness of NMF-solutions is still an open issue despite some recent attempts to deal with the subject [20,33,73]. Necessary and sufficient conditions have recently been formulated for a given NMF solution to be unique [39]. In [63] a geometric approach has been taken considering the determinant of the span of basis vectors and optimizing the decomposition for a minimal determinant. Unique solutions, as the term will be used in the following, may still be subject to scaling and permutation indeterminacies which are ubiquitous in many blind source separation contexts. There are two popular routes to enforce uniqueness of the solutions. While in [32], arguments from sparse coding are invoked, the development of positive matrix factorization as surveyed in [31] rather proposes application-driven solutions requiring background knowledge. Both approaches are limited to special applications where specific information about the data is available or specific assumptions concerning the composition of data are necessary. The issue of model order selection (MOS) within NMF decompositions relates to the application of information theoretic criteria like Akaike's Information Criterion (AIC) [1,2], the Minimum Description Length (MDL) criterion [54] which is equivalent to the Bayes Information Criterion (BIC) [67], or the Risk Inflation Criterion (RIC) [22–24]. Recently, however, automatic relevance detection (ARD) schemes have been discussed in relation with Bayesian approaches to NMF [10,21,72] and showed promising results.

The rest of this paper is organized as follows: in Section 2 we present a short overview of Bayesian approaches addressing issues of uniqueness and model order selection. Next we present in Section 3 a Bayesian Optimality Condition (BOC) for NMF with an Euclidean distance measure, and show that it leads to a minimum volume constraint for the optimization of an L_2-norm reconstruction error proposed earlier [63]. In Section 4 we finally present a new variational Bayesian NMF (VBNMF) approach to tackle the problem of model order selection and show in Section 5 with toy data sets that this approach implements an automatic relevance detection scheme. Finally, in 6, the potential of the VBNMF algorithm on binary toy data is explored. A preliminary version of this paper has been presented at a conference [60].

2. Bayesian approaches to NMF

While NMF was introduced in terms of optimizing a suitable cost function subject to non-negativity constraints, it is well-known that many popular NMF cost functions can be related to statistical models of the reconstruction error of the decomposition via Maximum Likelihood (ML) estimations. For example, the squared Euclidean distance measure is based on Gaussian error statistics, while KL- or IS-divergences as cost functions relate to alternative error statistics given by Poisson or Gamma distributed noise kernels. Hence, constrained optimization of proper cost functions can be interpreted within a statistical perspective as maximum likelihood estimation problems (see e.g. [55,15,64,21]). This opens the field to a conceptually more principled approach based on Bayesian probabilistic interpretations of NMF. It is not accidental that the Richardson–Lucy algorithm which was first presented in 1972 in a paper entitled *Bayesian-Based Iterative Method of Image Restoration* [53] is one of the forefathers of modern NMF algorithms [16].

Maximum Likelihood estimation is based on adequate reconstruction error statistics. For example, assuming the entries E_{ij} of the reconstruction error matrix $\mathbf{E}$ in Eq. (2) to be independently and identically distributed according to a Gaussian distribution with zero mean and variance σ_r^2, the joint distribution of all data items X_{ij} factorizes according to the following equation:

$$P(\mathbf{X}|\mathbf{W}, \mathbf{H}) = \prod_i \prod_j \frac{1}{\sqrt{2\pi}\sigma_r} \exp\left(-\frac{1}{2}\left(\frac{X_{ij} - [\mathbf{WH}]_{ij}}{\sigma_r}\right)^2\right) \quad (4)$$

Note that Eq. (2) implies the assumption of additive i.i.d. Gaussian noise. Hence, it is only an approximation, since the left hand side, i.e. the difference $X_{ij} - E_{ij}$ could turn negative, in principle. Since the right hand side of this equation, i.e. the term $[\mathbf{WH}]_{ij}$ is non-negative by definition, and the observed data is also non-negative, the noise cannot be independent from the data and small observations X_{ij} are related to low noise levels E_{ij} in practice. Since in the following we will consider situations only where the noise parameter σ_r is sufficiently small, this subtlety can be neglected here. In the following, we will refer to Eq. (4) as Gaussian likelihood for NMF.

For Gaussian noise kernels, maximizing the log-likelihood of the data corresponds to minimizing the quadratic error function D_E (see Eq. (1)). However, in addition to specifying proper cost functions according to data statistics, suitable prior distributions can be used to integrate existing knowledge about the data and enforce desired characteristics of the solutions. For example, non-negative sparse coding [32] actually is a *maximum a posteriori* (MAP) estimation, assuming independent exponential prior distributions of the weights W_{ij} and flat priors on the features H_{kj}. Several papers suggest Bayesian techniques to explicitly incorporate prior knowledge on the factor matrices in NMF, including independent Gamma priors [48], Gaussian process priors [64], or Gamma chain priors [75] for audio signal modeling. Also various volume priors have been discussed in the context of volume