Non Negative Matrix Factorization for Blind Source Separation

Aoulass Nabila[1], 2nd Chakkor Otman,
{ nabilaaoulass@gmail.com[1], o.chakkor@gmail.com[2] }

signal and image processing, university Abdlmalek Esaadi ENSAT laboratory NTT[1], telecommunication systems and networks engineering, university Abdlmalek Esaadi ENSAT[2]

Abstract. Non negative Matrix Factorization (NMF) has been a popular representation method for pattern classification problems. It tries to decompose a non negative matrix of data samples as the product of a non negative basis matrix and a non negative coefficient matrix in NMF both supervised and unsupervised mode of operations is used. Among them supervised mode outperforms well due to the use of pre-learned basis vectors corresponding to each underlying sources. In this paper NMF algorithms such as method based in the Frobinuis norm, Kullback Leibler divergence , and an extension to NMF, by incorporating sparsity. Algorithms are used to evaluate the performance of BSS in which supervised mode is used. We further illustrate the effect of hyperparameter as the rank k let the metric chooses and the initialization of decomposition matrices, on the speed of convergence of NMF algorithm.

Keywords: Source Separation, Blind Source Separation, NMF, sparsity.

1 Introduction

The separation of sources is the operation which, from the observations, makes it possible to obtain a set of signals proportional to the sources and to identify the contribution of each sources within the observed mixture. Thus we distinguish two subproblems:

—(i) the identification of the mixture.

—(ii) the reconstruction of the sources.

This opposite problem is badly posed because without any information on the sources and on the mixture, an infinity of solutions would be admissible. It is then necessary to formulate additional hypotheses and to take into account additional information on mixing and sources.

The problem of separation sources can be approached from two points of view. The first is that the decomposition of observations on a basis of elementary signals to eliminate the redundancy of information between the different observations. So, the first methods were proposed by C. Jutten and J. Hérault [1] who realized a nonlinear (ACP) in which we can diagonalize the covariance matrix by the decomposition in eigenvalues (EVD). Due to the limitation of diagonalizable matrices, singular value decomposition (SVD) makes (PCA) always possible

ICCWCS 2019, April 24-25, Kenitra, Morocco

DOI 10.4108/eai.24-4-2019.2284117

based on the orthogonality constraint. It offers the least error (with respect to some measures) with the same reduced complexity, compared to other models. But it is not the only. The NMF is used in place of other low rank factorizations, such as the (SVD) [2,3], because of its two primary advantages: storage and interpretability. Due to the non negativity constraints, the NMF produces a so-called "additive parts-based" representation [2,4] of the data. One consequence of this that the factors of decomposition matrix are generally naturally sparse, there by saving a great idea of storage when compared with the (SVD)'s dense factors[5] . But is not for free. On the one hand, the decomposition of the SVD is known to have a polynomial complexity. On the other hand, it has been recently demonstrated that the factorization of NMF has a non-deterministic polynomial computation complexity (NP) [2] for which the existence of an optimal algorithm of a polynomial time is unknown. Moreover, non-orthogonal factors do not allow representation as in (PCA) but are used as a basis for unsupervised or prior modeling for supervised learning [6].

A second, more recent approach is that of the Independent Component Analysis (ICA), it will be necessary to wait for the work of P.Comon [7] to generalize this concept. The latter demonstrates, in the case of linear mixtures, that if the source signals are assumed to be mutually independent and non- Gaussian (except for at most one source), it is possible to separate these signals to a scale factor and a permutation by seeking to minimize the dependence measurements between the estimated signals at the output of the separation system. The implicit objective of the (ICA) is often to find physically significant components. However, in some field of environmental science, and using data that has the property of non-negativity, the solutions estimated by the methods based on the (ICA) lack of physical interpretability. In addition, the (ICA) can not determine the variances (energies) of the independent components as well as the order of the independent sources because the basic functions are classified by non-Gaussianities [3]. In NMF, the non-negativity constraint leads to the representation based on parts of the input mixture that helps to develop structural constraints on the source signals. NMF does not require independent evaluation and is not limited to the length of the data. It provides more important basic vectors for the reconstruction of the underlying signal than the activation vectors [3].

Among the difficulties of matrix factorization in the area of blind separation, the ratio between the number of observations and the number of sources is a problem of interest for a large number of applications and has allowed the taxonomy that we recall below. In most applications and relying on instantaneous linear mixtures, the number m of samples in X is much larger than the numbers n of observations and p of sources. We then separate the determined case p = min (n, m), over-determined $p < n$, finally underdetermined such that $p > min(n, m)$.

When a single-channel source separation problem is considered under-determined, it can not usually be solved without prior knowledge of the sources in the mixture. For this reason, the problem of estimating multiple overlapping

sources from an input mixture is unclear and complex in the (BSS) environment. But (NMF) provides a solution to this single-channel source separation problem by using its non-negativity constraint as well as a supervised mode of operation for source separation [3]. So, NMF is defined as:

$$X \approx F.G \ . \tag{1}$$

Where $X \in R_+^{n \times m}$ is the spectrogram, $F \in R_+^{n \times p}$ matrix of basis vectors (columns), $G \in R_+^{n \times m}$ is the matrix of activations (rows) of the input mixture. In NMF when the spectrogram of mixture X is given, the matrices G and F can be computed via an optimization problem by:

$$\min_{F,G} D(X||F.G) \ . \tag{2}$$

where D denotes the divergence. the reduced dimension p depends on the application and is imposed by the problem that we seek to solve. It is the same for the content of the product matrices that vary also depending on the application and the processed data and can have different physical meanings.

2 Algorithms for solving the NMF problem

2.1 NMF method based on the Frobenius norm :

The multiplicative methods can be obtained in two different ways, either by a heuristic approach, or by a Maximization-Minimization (MM) approach.
Heuristic approach : Multiplicative methods based on the Frobenius norm solve the problem (1) by rewriting it as a matrix trace, i.e.

$$J(G,F) = Tr((X - G.F)^T(X - G.F)) \ . \tag{3}$$

By developing this expression, we can show 3 functions in the expression of J(G, F), that is,

$$\begin{aligned} J(G,F) &= Tr(X^T.X) - Tr(F^T.G^T.X) - Tr(X^T.G.F) + Tr(F^T.G^T.G) \\ &= J_1 - J_2 + J_3 \ . \end{aligned} \tag{4}$$

The calculation of the gradient of each of these three functions is carried out here with respect to a matrix F, while noting that the calculation of that with respect to G is similar.

$$\frac{\partial J}{\partial F} = 2G^T.(G.F - X) \ . \tag{5}$$

we can identify the two non-negative functions appearing in the writing of the gradient,

$$\nabla_F^+ = 2GT.G.F \ . \quad (6)$$
$$\nabla_F^- = 2G^T.X \quad . \quad (7)$$

The heuristic approach consists in using these two non-negative functions in the update rules of F, and transpose the results for the update of G :

$$F \to F \circ \frac{\nabla_F^- J(G,F)}{\nabla_F^+ J(G,F)} \quad . \quad (8)$$

$$G \to G \circ \frac{\nabla_G^- J(G,F)}{\nabla_G^+ J(G,F)} \quad . \quad (9)$$

the multiplicative update rules for the Frobenius NMF are available in :

$$F \to F \circ \frac{(G^T .X)}{(G^T .G.F)} \ . \quad (10)$$
$$G \to G \circ \frac{(X.F^T)}{(G.F.F^T)} \quad . \quad (11)$$

The Maximization-Minimization (MM) approach :is based on two steps to obtain the update rules. The principle of this method consists of : first look for increasing the cost function by a function called auxiliary function. then in a second step, to perform the minimization of the auxiliary function.
In the problematic of the classical NMF, we can take the function J (G, F) reduced to its vector formulation and rewrite it in the form of a Taylor development in the second order, considering that , the column current of the matrix F, is the only variable of the problem

$$J(\theta) = J(\theta^k) + \nabla J(\theta^k)(\theta - \theta^k) + \frac{1}{2}(\theta - \theta^k)^T.G^T - G(\theta - \theta^k). \quad (12)$$

Where

$$J(\theta^k) = -G^T(X - G.\theta^k) \quad . \quad (13)$$

the quadratic general form of the auxiliary function H(,) is given as a function of a positive A matrix, Lee and Seung [] propose to choose the matrix A in the form:

$$\theta^{k+1} = \frac{\theta^k o\, G^T x}{G^T.G.\theta^k} \ . \quad (14)$$

The matrix formulation can be obtained by collecting the columns of the matrix F, which gives rise to the multiplicative updating rules for the Frobenius NMF problem :

$$\theta^{k+1} = \frac{\theta^k o\, G^T x}{G^T.G.\theta^k} \quad . \quad (15)$$
$$G \to G \circ \frac{(X.F^T)}{(G.F.F^T)} \quad . \quad (16)$$

2.1 NMF methods based on Kullback Leibler divergence :

The strategy is of type MM and the auxiliary function obtained is based on the concavity of the logarithmic function. The cost function to be minimized is expressed as,

$$D_{KL}(X||F.G) = \sum_{i,j}\left[X\log\frac{X}{G.F}\right] - X + G.F]_{i,j} \ . \tag{17}$$

We adopt the notation f, and x as the respective current column vectors of F and X,

$$J(f) = D_{KL}(X||F.G) = \sum_i x_i \ \log x_i \ - x_i \ + \sum_j G_{i,j} \ f_j \ - \ x_i \ \log G_{i,j} \ f_j \ \ . \tag{18}$$

Using the concavity of the logarithmic function and Jensen's inequality, the previous cost can be increased by the following auxiliary function

$$H(f, f^k) = \sum_i x_i \ \log x_i \ + \sum_j G_{i,j} \ f_j \ - \ x_i \ \ \sum_j \frac{G_{i,j} \ f_j^k}{\sum_l G_{i,j} \ f_l^k} \ (\log G_{i,j} \ f_j log \frac{G_{i,j} \ f_j^k}{\sum_l G_{i,j} \ f_l^k}) \ . \tag{19}$$

The Maximization-Minimization Theorem implies that :

$$J(f^k) > \min\left(H(f, f^k)\right) = H(f^{K+1}, f^k)) \geq J(f^{K+1}). \tag{20}$$

Minimize $H(.,f^k)$ from f:

$$\frac{\partial H}{\partial f_j} = \sum_i G_{i,j} - \frac{f_j^k}{f_j} \sum_i \ x_i \ \frac{G_{i,j} \ f_j^k}{\sum_l G_{i,j} \ f_l^k} = 0 \quad . \tag{21}$$

The minimum is then given by :

$$f^{k+1} = \frac{f^{k+1}}{G.1} \circ \left(G^T \ \frac{x^\circ 1}{G.f^k}\right) \ . \tag{22}$$

By grouping these vectors, we get the update expression of F :

$$f^{k+1} = \frac{f^{k+1}}{G.1} \circ (G^T \ \frac{x^\circ 1}{G.f^k}) \ . \tag{23}$$

By transposition, the rule for updating the matrix G is written :

$$F = \frac{F}{G^T.1}\left(G^T \frac{x^\circ 1}{G.F}\right) \ . \tag{24}$$

2.2 Sparse NMF :

The concept of sparse coding refers to a representational scheme where only a few units (out of a large population) are effectively used to represent typical data vectors [8]. In effect, this implies most units taking values close to zero while only few take significantly non-zero values. In this paper, we use a sparseness measure based on the relationship between the L_1 norm and the L_2 norm [9] :

$$sparseness(x) = \frac{\sqrt{n} - \Sigma|x_i| \Big/ \sqrt{\Sigma x_i^2}}{\sqrt{n} - 1} \quad . \tag{25}$$

where n is the dimensionality of x. Our aim is to constrain (NMF) to find solutions with desired degrees of sparseness. The first question to answer is then : what exactly should be sparse? The basis vectors F or the coefficients G? This is a question that cannot be given a general answer; it all depends on the specific application in question. Further, just transposing the data matrix gnone) must be made by the experimenter. The sparse NMF problem can be formulated as :

$$E(F, G) = \|X - F.G\|^2 \quad . \tag{26}$$

is minimized, under optional constraints :

$$\text{sparseness}(f_i) = S_f, \forall \text{ i} . \tag{27}$$
$$\text{sparseness}(g_i) = \text{Sg}, \forall \text{ i} . \tag{28}$$

Where f_i is the i-th column of F and g_i is the i-th row of G. Here, S_f and Sg are the desired sparsenesses of F and G (respectively). These parameters are set by the user.

3 uniqueness of NMF solution :

Given the formulation of the NMF problem, it is clearly to be feared the existence of invertible square matrices T such that the pairs (WT, $T^{-1}H$) are also solutions of the problem, since

$$F.G = (F.S)(S^{-1}.H) . \tag{29}$$

and that the cost of reconstruction depends only on the product F G. Given a solution (F^0,G^0), the pair(F^0.S,S^1.G) is the solution of the problem if and only if the two matrices F^0.S and (S^{-1}.G) have positive or zero coefficients. At least two types of cases can be exhibited where this is possible :

Trivial invariances : If we insist that S and its inverse S^{-1} have positive or zero coefficients, the F.S and (S^{-1}.G) products are also positive and we are in the presence of a new solution to the initial problem. In this case, it is easy to prove that such a matrix S is necessarily the product of a permutation matrix and a diagonal matrix with positive coefficients [11]. The permutation matrix introduces K degrees of invariance, but does not really change the solution; moreover, the uniqueness of a quantity is often defined at a close permutation. With regard to scale factors (diagonal matrix), the question can be solved by imposing a standardization on one of the factors F or G (in practice, one will often choose to standardize the columns of F in norm L^2). These

invariances are therefore not a real obstacle to a possible uniqueness of the solution, which will be defined by a permutation and a change of scale.
Local invariances : The product ($F^0.G^0$) can also remain invariant on points in the vicinity of the couple that makes it.Suppose that F0 and G0 have strictly positive coefficients. Given a square matrix U, not necessarily with positive coefficients, we can find ε sufficiently small such that (I + εU) is invertible. One can make the limited development : (I + εU) ' (I −εU). The matrices (I +εU) and (I +εU)−1 perform local transformations around ($F^0.G^0$); provided that this point is not situated on the edges of the positive quadrant, and that ε is chosen sufficiently small, the points (F^0(I +εU)) and ((I +εU)−1)G^0 remain in this quadrant. Thus, we obtain a new solution to the problem of NMF, the pair (F^0(I +εU), (I +εU)$^{-1}$.G^0).

4 Non-uniqueness of the general case :

We have considered only pairs of solutions expressing themselves relative to each other via a linear transformation : (F, G) and (G.S,S−1.F). In reality, there is nothing to require that two solutions producing the same product F.G be connected in this way.

$$X=\begin{pmatrix}1&1&0&0\\1&0&1&0\\0&1&0&1\\0&0&1&1\end{pmatrix}=X.I=I.X \quad . \qquad (29)$$

where I denotes the identity matrix. The matrix V is of rank 3.We can choose as factorizations (F^0 = X,G^0 = I) or (F^1 = I,G^1 = X) There is no invertible matrix S such that $F^0S = F^1$, for reasons of rank. However, if we choose K = rg (X), we can show that such counterexamples are impossible. In this case, all solutions are connected to each other by linear transformations [11].

5 The disadvantages of NMF :

Geometrically, the (NMF) consists in finding a cone belonging to the positive orthant which includes the components of the vectors of the observed data [10]. From this point of view, the cone is not always unique without additional constraints. From this geometrical interpretation, on the one hand, it appears that the (NMF) is not unique ,which poses a problem for the BSS. On the other hand, in (NMF), the criteria can be convex only according to one of the two matrices produced but not for both. The algorithms therefore only allow to converge towards a local minimum. Therefore, the convergence result strongly depends on the initialization of the algorithm. practically, it is not guaranteed that the decomposition obtained an important interpretation. To avoid this problem, it is necessary to exploit certain preliminary information or to impose certain constraints on the decomposition. For example, information from locations or special signals is used in a so-called supervised (NMF) in [3]. This method improves the accuracy of automatic transcription, but requires well-organized advance information. Another strategy is to rely on specific constraints from the characteristics of the processed signals. However, This conditions are very difficult to satisfy in the case of real data.

6 The advantages of NMF :

The NMF is used in place of other low rank factorizations, because of its two primary advantages: storage and interpretability. Due to the nonnegativity constraints, the NMF produces a so-called “additive parts-based” representation of the data. One consequence of this is that the factors F and G are generally naturally sparse, there by saving a great deal of storage. The NMF also has impressive benefits in terms of interpretation of its factors. So, the basis vectors naturally correspond to conceptual properties of the data.

7 Application :

In this section, we present the empirical evidences that support NMF as a successful document clustering and topic modeling method.consider a text processing application that requires the factorization of a term-by-document matrix X. In this case, k can be considered the number of (hidden) topics present in the document collection. In this case, F becomes a term by-topic matrix whose columns are the NMF basis vectors. The non zero elements of column 1 of F, which is sparse and nonnegative, correspond to particular terms. So we compare the clustering quality between multiplicative update (MU) NMF based on Kullback Leibler divergence and other based on the Frobenius norm , Within the sparse NMF algorithms we compare the multiplicative updating (MU), Coordinate Descent(CD) wtih defferent initializations nndsvd, nndsvdar, random.

7-1 Data Sets ”20 Newsgroups”:

This data set consists of 20000 messages taken from 20 newsgroups. One thousand Usenet articles were taken from each of the following 20 newsgroups. Approximately 4 % of the articles are crossposted. The articles are typical postings and thus have headers including subject lines, signature files, and quoted portions of other articles.

7-2 Determining a Suitable Metric When using Non-negative Matrix Factorization :

In this paper we describes algorithms for nonnegative matrix factorization (NMF) with two cost functions of divergence.This cost functions parametrized by a single shape parameter beta that takes the Euclidean distance, the Kullback Leibler divergence and the Itakura-Saito divergence (the Itakura-Saito theory is established in [2]) . So to examine the effect of different divergences on convergence speed, we test them on multiplicative update.

Table 1. the speed of convergence as a function of defferences cost function taken by fixing the number of components at 10.

Methode NMF	Frobenius	KL	Itakura-Saito
time (s)	1.509	0.277	5

7-3 The effect of inisialization NMF with Sparseness contraint :

The NMF must be initialized and the initialization selected is crucial to getting good solutions. It is well-known that good initializations can improve the speed and accuracy of the solutions of many NMF algorithms. Add to this the fact that many NMF algorithms are sensitive with respect to the initialization of one or both NMF factors, and the impact of initializations becomes very important. In this section, we compare the results of three initialization procedures for sparse NMF :Nonnegative Double Singular Value Decomposition (NNDSVD), NNDSVD with zeros filled with small random values (NNDSVDar) and random initialization :

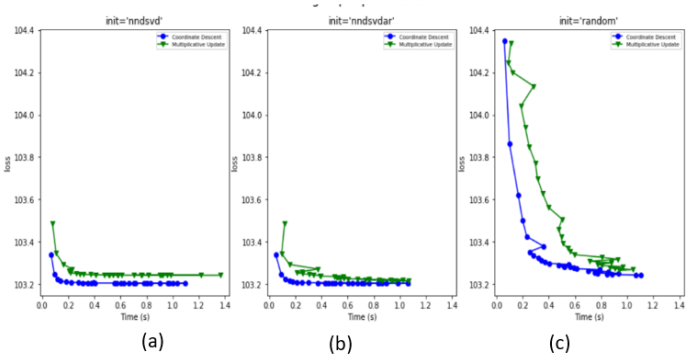

Fig. 1 graphs of different solvers taken with different inisializations: NNSDVD in (a), NNSDar in (b) and random inisialization in (c) .

.

Coordenate desent is established as the best algorithm for sparse NMF, and NN SDVD, NNSDar inisialization (graph (a), resp (b)) plays an important role to minimize convergence time and loss value. In all cases, a good initialization can improve the speed and accuracy of the algorithms, as it can produce faster convergence.

8 Conclusion :

Coordenate Non-negative matrix factorization (NMF) has proven itself a useful tool in the analysis of a diverse range of data. One of its most useful properties is that the resulting decompositions are often intuitive and easy to interpret because they are sparse. However, the sparseness achieved by NMF is not enough; in such situations like in extarction topic it might be useful to control some hyperparametres such as the inisialization of decompositions matrix that the best results are obtained by NNSDVD, NNSDar inisialisations for coordinate Descent

(CD) and suitable metric has also are markable effect on the convergence speed of the algorithm. So, we see that the best results for multiplicative update are obtained from Frobenius norm.

References

[1] Jutten and J. Herault, independent component analysis,1991.

[2] Nonnegative matrixfactorisation alorithms and applications, NGOC-DIEP HO,2008.

[3] An Experimental Survey on Non-Negative Matrix Factorization for Single Channel Blind Source Separation,Mona Nandakumar M and Edet Bijoy K,2014.

[4] Traitement des signaux parcimonieux et applications,Mohamed Aziz Sbai,2012.

[5] Algorithms, Initializations, and Convergence for the Nonnegative Matrix Factorization,Amy N. Langville ,Carl D. Meyer ,Russell Albright ,James Cox ,and David Duling,2014.

[6] NMF Factorisation par matrices non négatives,wikisat.

[7] Méthodes informées de factorisation matricielle non négative, Abdelhakim Limem,2017.

[8] O. Hoyer. Modeling receptive fields with non-negative sparse coding,2003.

[9] Non-negative Matrix Factorization with Sparseness Constraints,Patrik O. Hoyer,2004.

[10] Méthodes de séparation aveugle de sources et application à la télédétection spatiale ,Moussa Sofiane Karoui,2012 .

[11] Les factorisations en matrices non-négatives. Approches contraintes et probabilistes, application à la transcription automatique de musique polyphonique,Nancy Bertin,2010.