

NON NEGATIVE MATRIX FACTORIZATION CLUSTERING CAPABILITIES; APPLICATION ON MULTIVARIATE IMAGE SEGMENTATION

Cosmin Lazar
CReSTIC
University of Reims
Reims, France
cosmic.lazar@univ-reims.fr

Andrei Doncescu
LAAS-CNRS
University of Toulouse
Toulouse FRANCE
adoncesc@laas.fr

Abstract

The clustering capabilities of the Non Negative Matrix Factorization algorithm is studied. The basis images are considered like the membership degree of the data to a particular class. A hard clustering algorithm is easily derived based on these images. This algorithm is applied on a multivariate image to perform image segmentation. The results are compared with those obtained by Fuzzy K-means algorithm and better clustering performances are found for NMF based clustering. We also show that NMF performs well when we deal with uncorrelated clusters but it cannot distinguish correlated clusters. This is an important drawback when we try to use NMF to perform data clustering.

1. Introduction

Recently, NMF received an increasing attention in the data analyzing community due to its visual clustering capabilities. There are some applications which use NMF to perform different clustering tasks [1], [2]. In our paper we investigate the NMF clustering capabilities and we try to find out its meanings. We also developed a NMF based clustering algorithm and we applied it to an application of multivariate image segmentation.

Above all, NMF is a linear blind source separation method which learns base data representations [3]. But as all BSS algorithms, NMF finds a new basis in the original data space where the dataset is projected. Then the new dataset is nothing but the linear projection of the original data onto the new basis directions. The clustering capabilities of the linear projection methods derive from how well the new basis directions follow the clouds of the data (indicating clusters) in the original data space. In real data applications, clusters are distributed according to irregular directions, so any unrealistic constraint (such as the orthogonality) imposed on the matrix factorization will have as result a significant loss of information which may worsen the clustering results.

Different matrix factorization algorithms (PCA or Independent Component Analysis [4]) incorporate such constraints (orthogonality or independence) on one of the factor matrices. More than that, the coefficients are of both sign which makes that in applications such as image processing, one cannot give them a realistic interpretation.

The only constraint that NMF imposes on the factor matrices is that of non negativity which makes NMF to be the best candidate for finding the real directions among whom a data set is distributed into a multidimensional space. Now, if a single cluster is distributed among a single direction in a data space, its clustering capabilities are obvious.

The paper is organized as follows: in the next section we review the basis of the linear matrix factorization algorithms. Then we analyze NMF as a spectral clustering algorithm and we outline its principal drawbacks. A clustering algorithm is derived and it is tested in an application of multivariate image segmentation. The results are then compared with those obtained by classical Fuzzy K-means clustering. Finally we conclude with some discussions.

2. Overview of the matrix factorization algorithms

Let the input matrix $X = (x_1 \dots x_n)$ be a set of n data columns and d lines, then a matrix factorization algorithm tries to find a subspace in which the majority of the data lies. In general, the input matrix X is factorized into two matrices:

$$X = FG^T \quad (1)$$

where $X \in \mathbb{R}^{d \times n}$, $F \in \mathbb{R}^{d \times k}$, $G \in \mathbb{R}^{n \times k}$, generally, $k \ll \min(d, n)$. These algorithms are also called linear projection methods because the new data is nothing but the projections of the original data onto the new subspace axes defined by the column vectors of F .

The new data can be written as:

$$G^T = HX \quad (2)$$

where $H \in \mathbb{R}^{k \times d}$ is the inverse matrix of F if F is quadratic, $k = \min(d, n)$ or the pseudo-inverse of F if F is non quadratic, $k \ll \min(d, n)$.

Considering that all row vectors of H have unit length and since they are orthogonal to the column vectors of F then we can write:

$$G^T = (F^T F)^{-1} F^T X \quad (3)$$

The factor $F^T F$ changes the projection basis F only with a scaling factor, so we can write:

$$G^T = F^T X \quad (4)$$

By developing , eq. 4 we write:

$$G^T = (f_1 \dots f_k)^T (x_1 \dots x_n)^T \quad (5)$$

$$G^T = \begin{pmatrix} f_1 \cdot x_1 & f_1 \cdot x_2 & \dots & f_1 \cdot x_n \\ f_2 \cdot x_1 & f_2 \cdot x_2 & \dots & f_2 \cdot x_n \\ \vdots & \vdots & \ddots & \vdots \\ f_k \cdot x_1 & f_k \cdot x_2 & \dots & f_k \cdot x_n \end{pmatrix} \quad (6)$$

The dot product $f_i \cdot x_j$ can be written as:

$$f_i \cdot x_j = |f_i| \cdot |x_j| \cdot \cos(f_i, x_k) \quad (7)$$

We impose that, $|f_i| = 1$ and then $f_i \cdot x_j = |x_j| \cdot \cos(f_i, x_k)$.

This shows that all linear matrix factorization algorithms find a new subspace in the original data space where the new data are the projections of the original data onto the new subspace axis; the new data depends only on the original data norm and the spectral angle between them and the new subspace axis. The better the new subspace axis (feature vectors of F) follows the clouds or clusters of data in the original multidimensional space, the better the new data is discriminated and appropriated for clustering.

In the following we overview three matrix factorization methods.

- 1) Principal Component Analysis is the most popular matrix factorization method. It can be written as:

$$X = UV^T \quad (8)$$

where U is the matrix of the eigenvectors of the covariance matrix of X and V is the new data in the eigenvectors feature space. Here, X is the mean adjusted raw data.

- 2) Independent Component Analysis is an other well known matrix factorization technique. It can be expressed as PCA:

$$x = AS^T \quad (9)$$

where A is the mixing matrix and S is the sources matrix. A and S are obtained under the independence constraint imposed on the rows of matrix S .

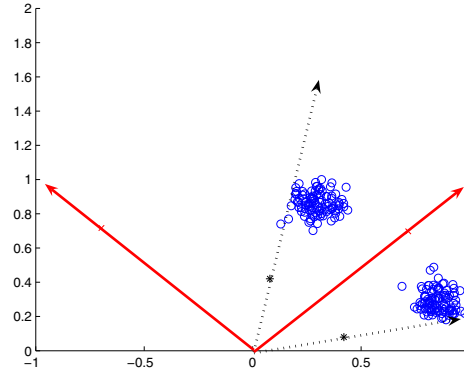


Figure 1. The difference between the PCA and NMF projection directions; solid line - PCA, dotted line - NMF

- 3) Non Negative Matrix Factorization is a well suited matrix factorization when the input data is non negative. It is expressed as:

$$X_+ = W_+ H_+^T \quad (10)$$

Matrices W and H are estimated bases only on the non-negativity assumption; no further assumptions on the statistical dependencies of the factor matrices are made.

The difference between PCA and NMF projection basis is illustrated on a toy example in figure 1. While the PCA basis vectors are orthogonal (one basis vector follows the direction indicating the maximum variance of the data), each NMF basis vector follows a single cluster meaning that NMF finds the most discriminating subspace where the data lies.

3. NMF and Fuzzy K-means

Previous works show that NMF has obvious clustering effects [3], [5] and it has been recently used in different clustering applications [1], [2]. There are some theoretical results which show the relation between NMF and the classical K-means clustering [6], [7]. Previous results are reviewed and we try to go further, providing a NMF based clustering algorithm and a useful discussion based on a toy example.

In [6], NMF is motivated by K-means clustering [8]. Let $W = (w_1 \dots w_k)$ be the cluster centroids obtained via K-means clustering. Let H be the cluster indicators: $h_{k,i} = 1$ if x_i belongs to the cluster c_k and $h_{k,i} = 0$ otherwise. The K-means objective can be written as:

$$J_{KM} = \sum_{i=1}^n \sum_{k=1}^k h_{i,k} \|x_i - w_k\|^2 = \|X - WH^T\|^2 \quad (11)$$

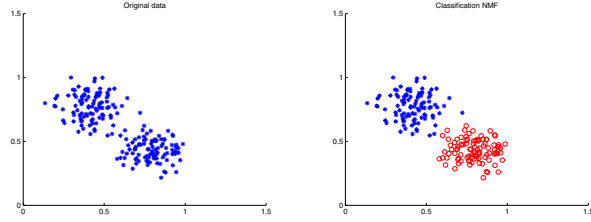


Figure 2. Left: two gaussian clusters in a two dimensional space; right: the classification results after NMF clustering

Its fuzzy extension, Fuzzy K-means allows one datum to belong to one or more clusters in the same time. The objective function is:

$$J_{FKM} = \sum_{i=1}^n \sum_{k=1}^k h_{i,k}^m \|x_i - w_k\|^2 = \|X - W(H^T)^m\|^2 \quad (12)$$

where $h_{i,k}$ is the membership degree of a data pattern x_i to the cluster c_k and m is a real number greater than 1 which controls the fuzziness. For $m = 1$ the hard K-means clustering is obtained.

NMF can also be performed by minimizing the objective function from equation 11 and so NMF can be thought as a soft clustering by relaxing the values of $h_{i,k}$ from binary to continuous non-negative values [6].

In the following, we consider the lines of the matrix H as the membership degree of the data to a particular class. But the question we are asking further is: does NMF always perform fuzzy clustering? The answer is no and we try to argue by some simple examples.

As all linear projection methods, NMF estimates some directions in the original data space (as we can see in figure 1). The new data is nothing but the projection of the data on these directions. So data which lies along some direction estimated by NMF has significant values in the corresponding line of the matrix H (because of the small angle between data vectors and the new basis vector w_k), while the rest of the data has less significant values in the corresponding lines of the matrix H .

Now if one and only one cluster is distributed along a single direction we can associate this cluster with one direction and in this case, NMF performs fuzzy clustering and the data is well clustered as the figure 2 shows.

If two clusters are distributed along a single direction (correlated clusters), NMF cannot be used for data clustering unless data decorrelation has been performed before. In this case, we cannot associate to a direction a single cluster but two or more correlated clusters and so, NMF cannot perform an effective clustering as it can be seen in figure 3. Data

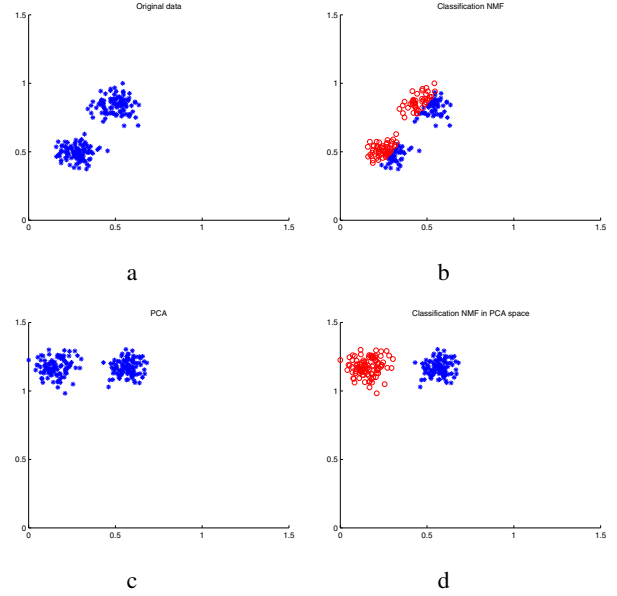


Figure 3. a - original dataset - two correlated clusters; b - clustering results after NMF clustering; c - data after decorrelation; d - clustering results after NMF clustering in decorrelated dataset

decorrelation is done by a simple rotation in the data space (generally PCA is used for data decorrelation).

NMF can be seen as a spectral clustering approach, meaning that all data in a NMF cluster will present strong spectral similarities. So, a NMF cluster can contain one or more clusters whose data are similar from a spectral point of view. The only difference between data belonging to a NMF cluster regards their magnitude. So correlated clusters can be found by a histogram based method. Now for each NMF cluster the histogram of the data norm is computed and sub-clusters are determined by fixing a threshold in the local minima of each histogram. The number of sub-clusters equals the number of modes in the histogram. This criterium is often used to estimate the number of classes in algorithms based on the estimation of the probability density function [9], [10].

Under these considerations, our approach can be resumed in the following algorithm:

NMF based clustering algorithm

- 1) Estimate the number of principal directions in the original data space and perform data decorrelation by PCA.
- 2) Compute NMF on the new dataset (the number of factors equals the number of the most significant principal directions).
- 3) The rows of the matrix H are considered to be the data

membership degree to a NMF cluster (each cluster is associated to a single direction). By normalizing the column vectors of H to have unit length, we can provide a fuzzy clustering result close to that obtained by Fuzzy K-means clustering.

- 4) A hard clustering can be obtained by assigning to each data the label corresponding to the line where the maximum value of column vectors lies.
- 5) Optional (re-cluster the NMF clusters by a histogram based method. For each cluster, the number of sub clusters equals the number of modes in the histogram).

Fuzzy K-means and NMF based clustering have been investigated and results have been compared in an application of multivariate image segmentation using two relative validity clustering indices: Davies-Bouldin (DB) and compacity-separability (CS).

The choice of the cluster number is a mandatory step in data clustering. It can be done either by external or internal indices [11]. In this paper, internal indices were used to estimate the optimal number of classes. For Fuzzy K-means clustering, DB and CS indices were computed for different values of the cluster number parameter.

The Davies Bouldin index [?] is based on similarity measure of clusters (R_{ij}) which depends on the dispersion measure of a cluster s_i and the cluster dissimilarity measure d_{ij} . The similarity measure of clusters (R_{ij}) can be defined as:

$$R_{ij} = \frac{s_i - s_j}{d_{ij}} \quad (13)$$

where $d_{ij} = d(v_i, v_j)$ and $s_i = \frac{1}{nc_i} \sum_{x \in c_i} d(x, v_i)$.

Here, d indicates the euclidian distance between two vectors, v_i is the center of the i^{th} cluster, nc_i is the number of data in cluster i , c_i denotes the i^{th} cluster and x a data vector.

Then, Davies-Bouldin index can be written as:

$$DB = \frac{1}{n_c} \sum_{i=1}^{n_c} R_i \quad (14)$$

where

$$R_i = \max_{j=1 \dots n_c, i \neq j} R_{ij} \quad (15)$$

The compacity-separability index is computed as:

$$CS = \frac{c_0}{s} \quad (16)$$

where c_0 measures the intra-cluster compacity:

$$c_0 = \frac{1}{n} \sum_{i=1}^c \sum_{k=1}^n \|x_k - v_i\|^2 \quad (17)$$

and s measures clusters separability (the inter-class distance).

$$s = \left[\min_{i=1 \dots c; j=1 \dots c; j \neq i} \|v_i - v_j\| \right]^2 \quad (18)$$

For both indices, the optimal clustering solution is given by the value where the global minimum is obtained.

For NMF based clustering, the cluster number is considered to be the number of the most important principal directions indicated by PCA. This is not always an unrealistic supposition because clusters can be formed by means of local structures in a data space such as high density area [11].

4. Experimental results

A multispectral image of a cross section of a barley grain, acquired in microspectrofluorometry is analyzed to identify external tissues of the barley grain. The grains were furnished by INRA de Clermont Ferrand and the images were recorded by INRA Nantes thanks to M.-F. Devaux. The data set contains 19 images each one of 512x512 pixels, figure 4.



Figure 4. Multicomponent image of a cross section of a barley grain

Image segmentation has been performed by pixel classification. Fuzzy K-means and NMF based clustering have been investigated in this application and their results have

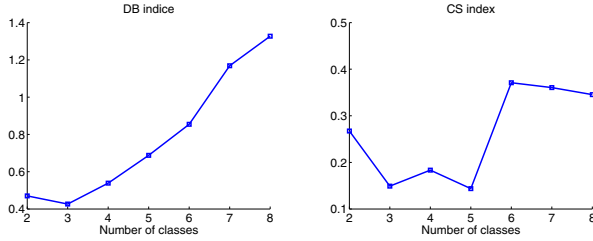


Figure 5. Davies-Bouldin and compactness-separability indices. Both indicate 3 clusters.

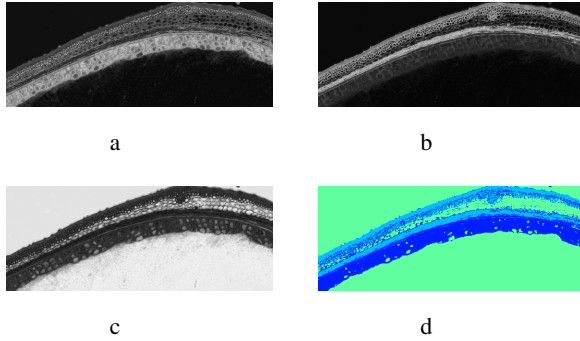


Figure 6. Fuzzy K-means results: a,b,c membership degrees; d - hard clustering

been compared using two relative validity clustering indices: compactness-separability (CS) and Davies-Bouldin (DB).

In previous works we used probability density based methods to perform pixel classification [13], [14]. A dimension reduction pre-processing step is mandatory for these approaches due to the computing time for density estimation which increases significant with data dimension. PCA and NMF were used to perform dimension reduction and then Parzen-Watershed algorithm [10], [15] clustered the data in the reduced feature space. This is a comun and useful manner to introduce matrix factorization algorithms in different clustering tasks especially when we deal with large sets of high dimensional data. In this case, the computing time for any clustering algorithm is an important drawback.

Fuzzy K-means clustering

Both Davies-Bouldin and compactness-separability indices indicate a local minimum for 3 clusters, figure 5. Clustering results are shown in figure 6. The simulations were performed with a fuzziness parameter $m = 3$ for 100 iterations.

NMF based clustering

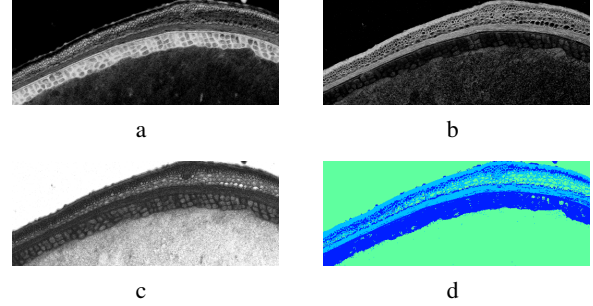


Figure 7. NMF based clustering: a, b, c - factor images; d - hard clustering

The results of the PCA indicate that the most part of the energy is distributed in the first three principal components, table 1. The number of spectral clusters is then fixed at 3. NMF is performed by chosing the number of factors equals 3. Factor images are presented in figure 7 a,b and c and hard clustering result is presented in figure 7 d.

PCA	73.0.69%	20.171%	3.056%	1.557%
-----	----------	---------	--------	--------

Table 1. Energy distribution

Clustering results comparison

Visualy, Fuzzy K-means and NMF clustering show very similar clustering results, but the performances of two clustering algorithms should be compared using some relative validity indices; any internal validity indice can be used as a relative index [11]. Davis-Bouldin and compactness-separability were used this time as relative indices and the results are shown in table 2. The results show that NMF based clustering performs better than Fuzzy K-means.

	Fuzzy K-means	NMF based clustering
DB indice	0.46	0.39
CS indice	0.149	0.117

Table 2. Clustering results comparison

5. Conclusion

Non Negative Matrix Factorization is investigated in this paper not from a blind source separation perspective but from the clustering point of view. We reinforce the idea that NMF can be considered as a fuzzy clustering approach which performs spectral clustering. We have shown that correlated clusters cannot be separated by simply applying

NMF on the data. We also developed a clustering algorithm and we applied it in a multivariate image segmentation application. Results are compared with those obtained by Fuzzy K-means clustering. We have shown that NMF based clustering algorithm offer better results than Fuzzy K-means. The choice of the cluster numbers is not imposed *a priori*: it is estimated by unsupervised approaches (PCA is used to estimate the principal directions in the data set which are associated to the number of spectral clusters). Spectral clusters can be re-clustered by a histogram based method where the number of sub-clusters are indicated by the number of modes in the histogram. As Fuzzy K-means algorithm, NMF based algorithm cannot deal with non-convex shape clusters. This drawback arrives from the fact that as all linear matrix factorization algorithms, NMF is only able to discover linear dependencies between data. If data are distributed along non-linear directions all these methods show their limits and so, NMF based clustering fail to discover the real clusters.

References

- [1] Farial Shahnaz , Michael W. Berry , V. Paul Pauca , Robert J. Plemmons, "Document clustering using nonnegative matrix factorization", *Information Processing and Management: an International Journal*, vol. 42 no. 2, pp. 373–386, March 2006.
- [2] Wei Xu , Xin Liu , Yihong Gong, "Document clustering based on non-negative matrix factorization," in *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, July 28-August 01, 2003, Toronto, Canada.
- [3] D.D.Lee and H.S.Seung, "Learning the parts of objects by non-negative matrix factorization", *Nature*, 401, pp. 788-791, 1999.
- [4] A. Hyvrinen and E. Oja. "Independent Component Analysis: Algorithms and Applications". *Neural Networks*, 13(4-5):411-430, 2000.
- [5] P. O. Hoyer , "Non-negative Matrix Factorization with sparseness constraints", *Journal of Machine Learning Research*, 5:1457-1469, 2004.
- [6] Li, T. and Ding, C., "The Relationships Among Various Nonnegative Matrix Factorization Methods for Clustering," in *In Proceedings of the Sixth international Conference on Data Mining*, december 18 - 22, 2006, Hong-Kong.
- [7] C. Ding, X. He, and H. D. Simon, "On the equivalence of nonnegative matrix factorization and spectral clustering," in *In Proc. SIAM Data Mining Conf*, April 21–23, Newport Beach, California, USA.
- [8] J. A. Hartigan and M. A. Wong (1979), "A K-Means Clustering Algorithm", *Applied Statistics*, vol. 28, no. 1, pp.100–108
- [9] M. Herbin, N. Bonnet, P. Vautrot, " Number of clusters and influence zones", *Pattern Recognition Letters* vol. 22, 1557–1568, 2001.
- [10] N. Bonnet, "Artificial intelligence and pattern recognition techniques in microscopic image processing and analysis", *Adv. Imag. Electron. Phys.* vol. 14, 1–77, 2000.
- [11] A. K. Jain and R. C. Dubes, "Algorithms for Clustering Data. Englewood Cliffs", NJ: *Prentice Hall, Inc.*, 1988.
- [12] D. L. Davies and D. W. Bouldin, "Cluster Separation Measure," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 1, pp. 95–104, 1979.
- [13] D. Nuzillard and C. Lazar, "Comparison of Two Unsupervised Methods of Classification for Segmenting Multi-spectral Images", *International Conference on Acoustics, Speech, and Signal Processing 2006*, Toulouse, France, may, 2006.
- [14] D. Nuzillard and C. Lazar, "Partitional Clustering Techniques for Multi-Spectral Image Segmentation", *Journal of Computers*, vol. 2, no. 10, pp. 1–8, december, 2007.
- [15] J. Cutrona, N. Bonnet, M. Herbin, F. Hofer, "Advances in the segmentation of the multi-component microanalytical images", *Ultramicroscopy*, vol.103, 141–152, 2005.