Approximating sparse semi-nonnegative matrix factorization for X-Ray covid-19 Image Classification

Manel Sekma[1], Amel Mhamdi[2], and Wady Naanaa[3]

[1] Higher Institute of Computer Sciences and Mathematics of Monastir, University of Monastir, 5000 Monastir, Tunisia.
manel.sekma@isimm.rnu.tn
[2] Aivancity School for Technology, Business and Society, Paris-Cachan, France.
mhamdi@aivancity.ai
[3] University of Tunis El Manar, 1000 Tunis, Tunisia
wady.naanaa@enit.utm.tn

Abstract. Medical imaging has been intensively used to help the radiologists do the correct diagnosis for the COVID-19 disease. In particular, *chest X-ray imaging* is one of the prevalent information sources for COVID-19 diagnosis. The obtained images can be viewed as numerical data and processed by non-negative matrix factorization (NMF) algorithms, one of the available numerical data analysis tools.
In this work, we propose a new sparse semi-NMF algorithm that can classify the patients into COVID-19 and normal patients, based on chest X-ray images. We show that the huge volume of data resulting from X-ray images can be significantly reduced without significant loss of classification accuracy. Then, we evaluate our algorithm by carrying out an experiment on a publicly available dataset, having a known chest X-ray image bi-partition.
Experimental results demonstrate that the proposed sparse semi-NMF algorithm can predict COVID-19 patients with high accuracy, compared to state-of-the-art algorithms.

Keywords: Non-negative matrix factorization · sparse semi-NMF · classification · K-Means · medical images · COVID-19

1 Introduction

Non-negative matrix factorization (NMF) is one of the most effective unsupervised technique in the field of numerical data analysis. It consists in approximating a given non-negative matrix $\mathbf{M}$ by a matrix product $\mathbf{WS}^T$, where both $\mathbf{W}$ and $\mathbf{S}$ are required to be non-negative. The goal of such an approximation is to reduce the dimensionality of the data, since the number of components comprised in both of the two factors is typically much less than the number of components of the input matrix. In counter part, it is unlikely to obtain a perfect equality between the input matrix and its approximation. The quality

of the approximation, $\mathbf{M} \approx \mathbf{W}\mathbf{S}^T$, is therefore evaluated using some matrix norm or divergence [4]. In addition to providing low-rank approximation for non negative data, NMF algorithms have an inherent clustering property. Indeed, a r-clustering, where r is the factorization rank, of the columns of the input matrix can be easily deduced, once we have obtained the two factors [3].

Standard NMF can be formulated as follows:

$$\underset{\mathbf{W},\mathbf{S}}{\text{Minimize}} \ \|\mathbf{M} - \mathbf{W}\mathbf{S}^T\|_F^2 \text{ subject to } \mathbf{W} \succeq \mathbf{0}, \quad \mathbf{S} \succeq \mathbf{0} \tag{1}$$

where the matrix norm employed is given by $\|\mathbf{X}\|_F = \sqrt{\operatorname{tr} \mathbf{X}^T\mathbf{X}}$.

However, NMF has many variations for different contexts. In this paper, we focus on a variation that relaxes the non-negativity constraint on one of the factors, namely $\mathbf{W}$, resulting in the *Semi-NMF* variant. Moreover, in order to guide the factorization to the relevant factors, we resort to an optimisation criterion that favours sparse matrices for one of the factors, namely $\mathbf{S}$. The motivation behind favouring sparse factors has been widely discussed [6], and simplifying the data interpretation is one of the most mentioned arguments.

Nonetheless, even if we consider the more constrained settings of standard NMF, a unique identification of $\mathbf{W}$ and $\mathbf{S}$ is not possible, because any one of the two factors might be permuted and scaled provided that the other factor is transformed accordingly. Indeed, if $\mathbf{P}$ is a permutation matrix and $\mathbf{\Lambda}$ a non-singular positive diagonal matrix then we have $\mathbf{W}\mathbf{S}^T = (\mathbf{W}\mathbf{P}\mathbf{\Lambda})(\mathbf{\Lambda}^{-1}\mathbf{P}^{-1}\mathbf{S}^T)$. The matrix pairs $(\mathbf{W}, \mathbf{S})$ and $(\mathbf{W}\mathbf{P}\mathbf{\Lambda}, \mathbf{S}\mathbf{P}^{-T}\mathbf{\Lambda}^{-1})$ are, therefore, regarded as equivalent solutions in the context of NMF.

Since the work by Paatero and Tapper [10], many other were developed [7]. The block principle pivoting algorithm is a fairly simple NMF algorithm which has, however, proved to be very efficient [7]. Since then, it forms the core of many state-of-the-art algorithms. Yet another track explored for the purpose of performing NMF is the one relying on convex geometry. NMF algorithms adopting this approach are referred to as *geometric algorithms* [8]. One of these algorithms is EDA, for Extreme Direction Analysis, [9], which proceeds by identifying relevant facets of the *data cone*. These facets are identified by solving a number of linear programs using the simplex algorithm.

In the present paper, we focus on the sparse semi-NMF context, which can be met in various real world situations such as hyper-spectral pictures taken by a satellite [13].

The proposed sparse semi-NMF algorithm, REDA (for Rectangular EDA), is built up on EDA. Recall that this latter algorithm applies only when one of the factors is a square matrix. However, in the general context of sparse semi-NMF, each of the two factors may be a rectangular matrix. Hence, the main theoretical contribution of this article, which consists in transforming sparse semi-NMF into a particular sparse semi-NMF where one of the factors is a square matrix. REDA is, therefore, used to help in COVID-19 diagnosis via chest X-Ray images. We chosen to apply our algorithm to the analysis of these particular images because existing works concern mainly supervised algorithms [11, 12]. The paper

is structured as follows: In Section 2, we present the theoretical background of our algorithm. The proposed sparse semi-NMF algorithm, REDA, is detailed in Section 3. In Section 4, we report the details of an experiment that demonstrates the efficiency of REDA in classifying chest X-ray images into COVID-19 and non COVID-19 images.

2 Sparse semi-NMF as a linear optimisation problem

This work concerns sparse semi-NMF (SSNMF). This is about finding a pair of matrices $\mathbf{W}, \mathbf{S}$ whose product $\mathbf{W}\mathbf{S}^T$ is close to an input matrix $\mathbf{M}$, while also favouring a sparse matrix for $\mathbf{S}$. As in a lot of research, the quality of the approximation is measured by the Frobenius norm of the difference between the input matrix and its approximation. In turn, the sparseness of $\mathbf{S}$ may be evaluated via the multichannel sparseness criterion, which is defined as follows $\mu(\mathbf{S}) = (\sqrt{\det \mathbf{S}^T\mathbf{S}})/(\Pi_{j=1}^{n}\|\mathbf{s}_j\|_1)$, where $\mathbf{s}_j$ denotes the j^{th} column of $\mathbf{S}$. It has been shown, in [9], that $0 \leq \mu(\mathbf{S}) \leq 1$ and $\mu(\mathbf{S})$=1 if and only if $\mathbf{S}$ is a column orthonormal matrix. Moreover, since $\mathbf{S}$ is a non negative matrix, $\mu(\mathbf{S})$=1 implies that $\mathbf{S}$ contains, at least, $n(r-1)$ zeros, which corresponds to a rather sparse matrix.

In order to simplify the expression of $\mu(\mathbf{S})$, we enforce the columns of $\mathbf{S}$ to sum to 1. By the equivalence between NMF solutions, this additional constraint does not entail any loss of generality. Using matrix notation, the latter constraint can be written as $\mathbf{S}^T\mathbf{1} = \mathbf{1}$, where $\mathbf{1}$ denotes an all-one vector that has the appropriate size. It follows that $\mu(\mathbf{S})^2 = \det \mathbf{S}^T\mathbf{S}$.

The studied NMF variation is therefore specified as follows

$$\underset{\mathbf{W},\mathbf{S}}{\text{Minimize}} \ \|\mathbf{M} - \mathbf{W}\mathbf{S}^T\|_F^2 \ + \ \gamma \det \mathbf{S}^T\mathbf{S}, \ \text{subject to } \mathbf{S} \succeq \mathbf{0}, \ \mathbf{S}^T\mathbf{1} = \mathbf{1} \quad (2)$$

where $\gamma \leq 0$. As it can be seen from (2), the objective is a weighted sum of two terms: an *approximation* term and a *sparseness* term. We note that minimizing the sparseness term is a rather difficult task, because this is neither a concave nor a convex function of $\mathbf{S}$. The objective could, however, be approximated by considering the two terms of the objective separately. Consider therefore the following Semi-NMF sub-problem, which is obtained from (2) by setting $\gamma = 0$:

$$\underset{\mathbf{W},\mathbf{S}}{\text{Minimize}} \ \|\mathbf{M} - \mathbf{W}\mathbf{S}^T\|_F^2 \ \text{ subject to } \mathbf{S} \succeq \mathbf{0}, \quad \mathbf{S}^T\mathbf{1} = \mathbf{1} \quad (3)$$

The following theorem[4] is the main theoretical contribution of this paper.

Theorem 1. *Let* $(\mathbf{M}, r)$ *be a Semi-NMF instance and let* $\mathbf{S}$ *be a non negative matrix whose columns sum to 1. Then* $\mathbf{S}$ *is an optimal solution for* $(\mathbf{M}, r)$ *if and only if it is a full column rank matrix whose columns are in the range of* $\mathbf{V}_{:,1:r}$*, where* $\mathbf{V}$ *is the right singular factor of* $\mathbf{M}$*.*

[4] The proof the this theorem is omitted in this short version of the paper.

From the above theorem, we can evaluate the quality of the proposed rank r approximation. beginAmel

Corollary 1. *The error of approximating an m-by-n matrix $\mathbf{M}$ by a rank r SNMF $\mathbf{M} \approx \mathbf{W}\mathbf{S}^T$ is given by $\|\mathbf{M} - \mathbf{W}\mathbf{S}^T\|_F^2 = \sum_{i=r+1}^{m} \sigma_{i,i}^2$, where $\sigma_{i,i}, i : r+1, \ldots, m$ are the smallest $m-r$ singular values of $\mathbf{M}$.*

In what follows $\bar{\mathbf{V}}$ will designate $\mathbf{V}_{:,1:r}$.

Let $\mathbf{c} = \bar{\mathbf{V}}^T\mathbf{1}$. Theorem 1 suggests that any n-by-r matrix $\mathbf{S}$ that verifies $\mathbf{S} = \bar{\mathbf{V}}\mathbf{X} \succeq 0$ and $\mathbf{S}^T\mathbf{1} = \mathbf{X}^T\mathbf{c} = \mathbf{1}$, for some non singular matrix $\mathbf{X}$, yields an optimal solution for Problem (3). Then solving Problem (3) reduces to solving the following problem:

$$\underset{\mathbf{X}}{\text{Minimize}}\ 1 \ \text{ subject to } \bar{\mathbf{V}}\mathbf{X} \succeq \mathbf{0},\ \mathbf{X}^T\mathbf{c} = \mathbf{1} \tag{4}$$

Problem (4) presents many advantages, with regard to Problem (3). First, the matrix to be computed, that is $\mathbf{X}$, is r-by-r, which is a relatively small matrix compared to $\mathbf{S}$ in a context where $r \ll n$. Moreover, Problem (4) is a satisfiability problem, that is, it has no objective. Thus, Problem (4) should be much more easier to solved. This simplified problem will be used as the starting point to cope with the main problem, namely Problem (2).

Thanks to Theorem 1, we obtain $\mathbf{S}^T\mathbf{S} = \mathbf{X}^T\mathbf{X}$. And since $\mathbf{X}$ is a square matrix, this implies that $\det \mathbf{S}^T\mathbf{S} = (\det \mathbf{X})^2$. Hence the following mathematical program

$$\underset{\mathbf{X}}{\text{Maximize}}\ |\det \mathbf{X}| \ \text{ subject to } \bar{\mathbf{V}}\mathbf{X} \succeq \mathbf{0},\ \mathbf{X}^T\mathbf{c} = \mathbf{1} \tag{5}$$

3 The algorithm

Thanks to Theorem 1, a suboptimal solution for Problem (2) can be obtained via EDA, the sparse semi-NMF algorithm described in [9]. Roughly speaking EDA proceeds by replacing the columns of matrix $\mathbf{X}$, one at a time, by other feasible vectors, in order to increase the objective $|\det \mathbf{X}|$. Thus, EDA can be used to find an approximate solution for Problem (2) at the expense of minor changes. These changes consist in the singular value decomposition and the choice of the r first right singular vectors of $\mathbf{M}$, performed at lines 2 and 3 of Algorithm 1. The resulting algorithm will be referred to by REDA (for rectangular EDA).

Next, we show how REDA can be employed to perform a classification tack. More precisely, we assume that the goal is classify the columns of the data matrix, $\mathbf{M}$, into r classes. To this end, we apply REDA to the SSNMF instance defined by $(\mathbf{M}, r)$ Once we have obtained an approximate solution, say $(\mathbf{W}, \mathbf{S})$, this latter is used to dispatch the columns of $\mathbf{M}$ into the r classes as follows. The class of each column is determined by measuring the similarity between each column of the approximate matrix $\mathbf{W}\mathbf{S}^T$, on the one hand, and each column of $\mathbf{W}$ on the other hand. The similarity between each pair of column vectors is evaluated by the cosine of the angle formed by the two vectors. The resulting cosine matrix,

which has size n-by-r, is used to determine the class of each of the n columns of $\mathbf{M}$. This is done by simply determining the position of the maximum in each row of the cosine matrix.

Algorithm 1: REDA

Data: $\mathbf{M}, r$
Result: $\mathbf{W}, \mathbf{S}$
// r must not exceed the rank of $\mathbf{M}$
$r \leftarrow \min(r, \text{rank}(\mathbf{M}))$
$[\mathbf{U}, \mathbf{\Sigma}, \mathbf{V}] \leftarrow \text{svd}(\mathbf{M})$
// $\bar{\mathbf{V}}$ is composed of the first r right singular vectors of $\mathbf{M}$
$\bar{\mathbf{V}} \leftarrow \mathbf{V}(:, 1:r)$
// Initializing matrix $\mathbf{X}$ to a non singular matrix
$\mathbf{f} \leftarrow \mathbf{1}$
for j $\leftarrow 1$ **to** r **do**
  $\mathbf{X}(:, j) \leftarrow \arg\max_{\mathbf{x}} |\mathbf{f}^T\mathbf{x}|$ subject to $\bar{\mathbf{V}}\mathbf{x} \succeq 0,\ \mathbf{c}^T\mathbf{x} = 1$
  $\mathbf{N} \leftarrow \texttt{nullspace}(\mathbf{X})$
  $\mathbf{f} \leftarrow \mathbf{N}^T\mathbf{1}$
end
stop $\leftarrow$ `false`
while not stop **do**
  stop $\leftarrow$ `true`
  max $\leftarrow 1$
  $\mathbf{Y} \leftarrow \mathbf{X}^{-1}$
  for j $\leftarrow 1$ **to** r **do**
    $\mathbf{f} \leftarrow \mathbf{Y}(:, j)$
    $\mathbf{x}^* \leftarrow \arg\max_{\mathbf{x}} |\mathbf{f}^T\mathbf{x}|$ subject to $\bar{\mathbf{V}}\mathbf{x} \succeq 0, \quad \mathbf{c}^T\mathbf{x} = 1$
    if $(|\mathbf{f}^T\mathbf{x}^*| \geq max)$ **then**
      $max \leftarrow |\mathbf{f}^T\mathbf{x}^*|$
      $j_{max} \leftarrow j$
      $\mathbf{x}_{max} \leftarrow \mathbf{x}^*$
      stop $\leftarrow$ `false`
    end
  end
  if not stop **then** $\mathbf{X}(:, j_{max}) \leftarrow \mathbf{x}_{max}$
end
$\mathbf{S} \leftarrow \bar{\mathbf{V}}\mathbf{X}$
$\mathbf{W} \leftarrow \mathbf{M}\mathbf{S}^{T\dagger}$

4 Experimental results

We experimentally evaluate the performances of our algorithm by attempting to partition a dataset of X-ray images into a COVID-19 and a non COVID-19 subsets. The dataset used for the experiments is obtained from the Kaggle

Table 1. Comparison between clustering algorithms. Performance values are reported in the form of accuracy percentages.

Classes	Kmeans	NMF	REDA
NORMAL	94.69	93.24	96,34
COVID-19	84.89	92.18	87,15

Table 2. Average classification accuracy. (Performance values are reported in the form of Accuracy percentages

Methods	Classification Accuracy
K-means	92.08
NMF	92.96
REDA	93,88

repository COVID-19 Radiography Database [2]. It contains 576 X-ray images of COVID-19 patients and 1583 images for normal people. All the results quoted in this work were performed using the original dataset.

In a first step, we applied an image preprocessing, which consists in resizing the X-ray radio images in order to obtain 180-by-150 images. This substantially reduce the input matrix size and boosts computational speed. The obtained images were subsequently converted into grayscale images, since the luminance is more important than the colours, for this kind of image.

The second step consists in building the input matrix. Then, the rows of each preprocessed image are concatenated to form a single column of the input matrix $\mathbf{M}$. Thus, the number of columns in $\mathbf{M}$ correspond to the number of images in the dataset, and the number of rows correspond to the number of pixels in each image.

REDA is compared with two existing non supervised clustering algorithms, namely the NMF algorithm described in [1], which was parametrized for clustering tasks, and the K-means algorithm [5]. This latter algorithm is a distance-based algorithm that assigns a points into clusters based on the notion of cluster centres, and it is known as one of the most competitive unsupervised clustering algorithms.

The results of our experiment is shown in Table 1. We used the COVID-19 label to designate the row containing the accuracy obtained for the class of COVID-19 infected patients, and the NORMAL label for the row containing the accuracy for the not infected patients.

As it can be noticed from Table 1, for all algorithms, the NORMAL class is identified with more accuracy than the COVID-19 class. And the highest accuracy is obtained by REDA (96,34%). In contrast, for the COVID-19 class, the state-of-the-art NMF obtained the highest accuracy. Finally, the overall accuracy (see Table 2) gave the advantage to REDA, which obtained highest global accuracy, with 93,88%.

5 Conclusion

In this paper, we have proposed a sparse semi-NMF algorithm, REDA, which has been built upon an existing geometric sparse semi-NMF algorithm. When applied to chest X-ray images, REDA significantly reduces the huge volume of data, issued from these images, without loss of classification accuracy.

The experimental results showed that the classification accuracy of the proposed algorithm is very competitive with dedicated state-of-the-art algorithms.

References

1. Cichocki, A., Phan, A.H.: Fast Local Algorithms for Large Scale Nonnegative Matrix and Tensor Factorizations. IEICE Transactions on Fundamentals of Electronics Communications and Computer Sciences **92**(3), 708–721 (Jan 2009)
2. Cohen, J.P., Morrison, P., Dao, L., Roth, K., Duong, T.Q., Ghassemi, M.: Covid-19 image data collection: Prospective predictions are the future. arXiv 2006.11988 (2020)
3. Ding, C., He, X., Simon, H.D.: On the equivalence of nonnegative matrix factorization and spectral clustering. In: in SIAM International Conference on Data Mining. pp. 606–610 (2005)
4. Gillis, N.: The why and how of nonnegative matrix factorization. CoRR **abs/1401.5226** (2014), http://arxiv.org/abs/1401.5226
5. Hartigan, J., Wong, M.: Algorithm AS 136: A K-means clustering algorithm. Applied Statistics pp. 100–108 (1979)
6. Kim, H., Park, H.: Sparse non-negative matrix factorizations via alternating non-negativity-constrained least squares for microarray data analysis. Bioinformatics **23**(12), 1495–1502 (05 2007)
7. Kim, H., Park, H.: Nonnegative matrix factorization based on alternating nonnegativity constrained least squares and active set method. SIAM J. Matrix Anal. Appl. **30**(2), 713–730 (2008)
8. Laurberg, H., Christensen, M.G., Plumbley, M.D., Hansen, L.K., Jensen, S.H.: Theorems on positive data: On the uniqueness of NMF. Comput. Intell. Neurosci. **2008** (2008)
9. Naanaa, W., Nuzillard, J.: Extreme direction analysis for blind separation of non-negative signals. Signal Process. **130**, 254–267 (2017)
10. Pentti, P., tapper Unto: Positive matrix factorization: a nonnegative factor model with optimal utilization of error estimates of data values. Environmetrics **5**(2), 111–126 (1994)
11. Ucara, F., Korkmaz, D.: Covidiagnosis-net: Deep bayes-squeezenet based diagnosis of the coronavirus disease 2019 (covid-19) from x-ray images. Med Hypotheses. **140**, 109761 (2020)
12. Waheed, A., Goyal, M., Gupta, D., Khanna, A., Al-Turjman, F., Pinheiro, P.: Covidgan: Data augmentation using auxiliary classifier gan for improved covid-19 detection. IEEE Access **PP**, 1–1 (05 2020)
13. Wang, Z., He, M., Wang, L., Xu, K., Xiao, J., Nian, Y.: Semi-nmf-based reconstruction for hyperspectral compressed sensing. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing **13**, 4352–4368 (2020). https://doi.org/10.1109/JSTARS.2020.3010332