A Bregman-proximal point algorithm for robust non-negative matrix factorization with possible missing values and outliers - application to gene expression analysis

Stephane Chretien, Christophe Guyeux, Bastien Conesa, Régis Delage-Mouroux, Michèle Jouvenot, Philippe Huetz, Françoise Descotes

Chrétien *et al. BMC Bioinformatics* 2016, **17**(Suppl 8):284
DOI 10.1186/s12859-016-1120-8

BMC Bioinformatics

RESEARCH Open Access

A Bregman-proximal point algorithm for robust non-negative matrix factorization with possible missing values and outliers - application to gene expression analysis

Stéphane Chrétien[1*†], Christophe Guyeux[2†], Bastien Conesa[3], Régis Delage-Mouroux[4], Michèle Jouvenot[4], Philippe Huetz[5] and Françoise Descôtes[6]

From 11th International Symposium on Bioinformatics Research and Applications (ISBRA '15)
Norfolk, VA, USA. 7-10 June 2015

Abstract

Background: Non-Negative Matrix factorization has become an essential tool for feature extraction in a wide spectrum of applications. In the present work, our objective is to extend the applicability of the method to the case of missing and/or corrupted data due to outliers.

Results: An essential property for missing data imputation and detection of outliers is that the uncorrupted data matrix is low rank, i.e. has only a small number of degrees of freedom. We devise a new version of the Bregman proximal idea which preserves nonnegativity and mix it with the Augmented Lagrangian approach for simultaneous reconstruction of the features of interest and detection of the outliers using a sparsity promoting ℓ_1 penality.

Conclusions: An application to the analysis of gene expression data of patients with bladder cancer is finally proposed.

Keywords: Feature extraction, Non-negative matrix factorization, Gene expression analysis, Outliers and missing data

Background

Non-Negative Matrix Factorization (NMF) is a very efficient approach to feature extraction in machine learning when the data is naturaly non-negative. It has been applied to an extremely large range of situations such as clustering [1], email surveillance [2], hyperspectral image analysis [3], face recognition [4], blind source separation [5], etc. It has recently also been applied to microarray data analysis [6] and biomedicine [7]. Given a dataset consisting of n vectors $x_1, \ldots, x_n$ in $\mathbb{R}^d$, the NMF approach builds a matrix M whose columns are $x_1, \ldots, x_n$ and then factorizes this matrix as

$$M = UV^t + E,$$

where E is an error term, U and V are componentwise non-negative, and U has a small number of columns. The features are the columns of U. They are often interpretable and summarize the data in an efficient manner since each data then consists of a mixture of these columns. For many real datasets, the rank of the obtained matrix, i.e. the number of features extracted, is usually small and the NMF thus provides a compact representation of the data.

The method was first explored by Lee and Seung [8] in the late 90's and it then enjoyed a significant growth of interest in many application fields and especially machine learning. There exists a wide variety of methods for computing the NMF. One most employed strategy is the famous alternating minimization scheme, which consists in successively minimizing in U and then in V. Notice that minimization in U (resp. V) is a convex and easy optimization problem. Furthermore, it has been observed as quite

*Correspondence: stephane.chretien@npl.co.uk
†Equal Contributors
[1] National Physical Laboratory, Hampton Road, Teddington, Middlesex, UK
Full list of author information is available at the end of the article

efficient in practice. The main drawback of this approach however, is that no convergence garantee towards a global minimizer has been proved so far. Moreover, handling the nonnegativity constraints appears to be cumbersome is certain settings and the convergence speed of the method depends on the way these constraints are incorporated into the iterations. The work described in is [9] is a very interesting contribution to the study of potential convexifications for the NMF problem. It uses certain separability assumptions. Separability is the property that the features are some data vectors already belonging to the sample. Following shortly after, [10] proposed an efficient approach based on linear programming which also relied on separability. Recently, under similar assumptions, [11] proposed a very simple approach based on successive projections. When separability holds the above algorithms are the methods of choice for NMF. Unfortunately, separability does not hold in very important cases, and there is still a lot of work to do in order to understand the performance guarantees of the existing algorithms for NMF. Back to the not necessarily separable case, [12, 13] proposed Bregman divergence based iterative methods for NMF. Bregman-divergence based proximal approaches have been the subject of great interest recently due to good practical performances and connection with mirror descent type algorithms (see for instance the survey [14]).

In the present article we devise a Bregman-proximal method for NMF naturally extends to the case where some data may be missing and/or corrupted by the occurrence of outliers. Missing data and outliers are very frequent in gene expression data. Our approach also borrows ideas from robust PCA [15], where the matrix we want to approximate is assumed to be splittable into a low rank part and a sparse part:

$$M = L + S.$$

The outliers are represented by the matrix S. Notice that the noise was not taken into account in the original article [15], whereas gene expression dataset are often corrupted by very large noise. This is easily overcome by performing least squares penalized estimation as in e.g. the code GoDec [16]. In the present work, an efficient method is proposed that denoises the data, estimates the missing values, and identifies the outliers in M via nonnegative low-rank + sparse + noise matrix factorization. Our algorithm is inspired by the recent work [17], which presents a clear interpretation of the ADMM in terms of proximal method-type iterations. In our approach, a Bregman divergence is chosen for the proximal scheme which allows to easily take into account the nonnegativity constraints.

In the next section, the Bregman proximal scheme is presented and in the subsequent Section, a version taking into account potential outliers and/or missing data is described in full details. The choice of the relaxation parameter is also addressed. An application to the analysis of gene expression data of patients with bladder cancer is proposed in the last Section.

Method

A Bregman proximal scheme for non-negative matrix factorization

Let h be a strictly convex real valued function. Assume that h is continuously differentiable and defined on a closed convex set $\mathcal{C}$. Then, for all $x, y \in \mathcal{C}$, the Bregman divergence associated to h is given by

$$D_h(y,x) = h(y) - h(x) - \left\langle \nabla h(x), (y-x) \right\rangle. \quad (0.1)$$

The space alternating Bregman-proximal scheme

In this section, we will consider the following Bregman-proximal algorithm, which alternates minimization in the variable U and minimization in the variable V:

$$U^{(k+1)} \leftarrow \operatorname{argmin}_{U\in\mathbb{R}^{d\times n}} \left\| M - UV^{(k)^t} \right\|_F^2 + \rho D_h\left(U, U^{(k)}\right) \quad (0.2)$$

$$V^{(k+1)} \leftarrow \operatorname{argmin}_{V\in\mathbb{R}^{d\times n}} \left\| M - U^{(k+1)}V^t \right\|_F^2 + \rho D_h\left(V, V^{(k)}\right) \quad (0.3)$$

where $D_h(.,.)$ is the Bregman's divergence associated with $h(x) = x\ln(x)$, so we obtain

$$D_h(y,x) = x\ln\left(\frac{x}{y}\right) + y - x, \quad (0.4)$$

and ρ is a positive constant. Let us consider the problem

$$\operatorname{argmin}_{U\in\mathbb{R}^{d\times r}} \frac{1}{2}\left\| M - UV^{(k)^t t} \right\|_F^2 + \rho\left(U^{(k)}\ln\left(\frac{U}{U^{(k)}}\right) - \left(U - U^{(k)}\right)\right). \quad (0.5)$$

The gradient of $\phi(U)$ is given by

$$\nabla\phi(U) = -(M - UV^t)V.$$

Let us now compute the gradient of $\varphi(U)$ defined by $\varphi(U) = D_h(U, U^{(k)})$. A straightforward computation gives

$$\frac{\partial\varphi}{\partial U_{ij}}(U) = \ln\left(\frac{U_{ij}}{U_{ij}^k}\right).$$

Therefore, taking one step in our Bregman-penalized subpace method sums up to solving

$$(M - UV^t)V = \rho\ \ln\left(\frac{U_{ij}}{U_{ij}^k}\right).$$

Since no explicit solution to this decoupled system of equations, we will use a fixed point approach defined as follows.

1. Take $U^{(k+1,0)} = U^{(k)}$.
2. $\forall\, l \in \mathbb{N}^*$, define

$$U_{ij}^{(k+1,l+1)} = \exp\left(\frac{1}{\rho}\left[\left(M - U^{(k+1,l)}V^t\right)V\right]_{i,j} + \ln U_{ij}^k\right). \tag{0.6}$$

3. Stop when the difference between two successive iterates is sufficiently small, e.g., less that 1e-3. Denote by l^* the iteration number when this occurs and output $U^{(k+1)} = U^{(k+1,l^*)}$.

The iterate $V^{(k+1)}$ can be obtained from $V^{(k)}$ using the same approach. The corresponding optimization problem associated to step $k+1$ is

$$\operatorname{argmin}_{V\in\mathbb{R}^{n\times r}} \frac{1}{2}\left\|M^t - VU^{(k)^t}\right\|_F^2 + \rho\left(V^{(k)}\ln\left(\frac{V}{V^{(k)}}\right) - \left(V - V^{(k)}\right)\right).$$

A toy numerical experiment

We start with a simple random example programmed in Matlab. Let U_0 be a random matrix in $\mathbb{R}^{50\times 8}$ with i.i.d. components having the uniform distribution on $[0,1]$. Let V_0 be a random matrix in $\mathbb{R}^{70\times 8}$ with components having the same distribution. Take $M = U_0V_0^t$, $\rho = 100$, and random initial matrices. Figure 1 shows that the method converges to M in the sense that it produces a sequence of matrices $U^{(k)}$ and $V^{(k)}$ whose product $U^{(k)}V^{(k)^t}$ converges to M.

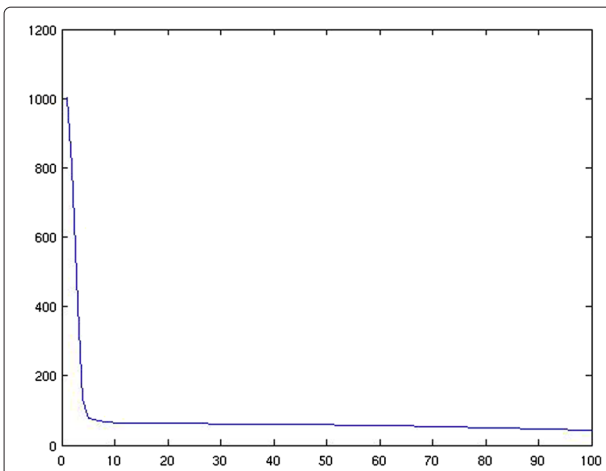

Fig. 1 Evolution of the error $M - U^{(k)}V^{(k)^t}$ as k goes from 1 to 100 on a random example

At this point, we can see on a toy example that the method converges, but it is obviously not proven in the general case. However, the present state of knowledge in the analysis of numerical algorithms for NMF is still lacunary and the case of outliers is therefore even more out of reach for the moment. The only case where a method has been proposed that finds an optimal solution in polynomial time is the case of separable data (see [18]). We were not able to prove that the assumptions are satisfied by our data set. Proving convergence of the method to a stationary point is not convincing either for practical purposes and is out of scope for the present study. To our opinion, and in view of the state of the art, showing a nice behavior of the algorithm in a toy example can be a sufficient practical evidence that the method has a stable behavior, which is what the practitioners want to know before going into more details. The convergence analysis of the algorithm will be studied in a later project and we hope to obtain a more precise theoretical understanding in a near future.

The case of outliers and missing data

Let Ω denotes the set of couples (i,j) for which an observation of $M_{i,j}$ is available. The matrix factorization problem can be addressed by considering the following optimization problem

$$\min_{Y,S,U,V\geq 0} \frac{1}{2}\sum_{(i,j)\in\Omega}(Y_{i,j} - (UV^t)_{i,j})^2 + \lambda\,\|S\|_1 \tag{0.7}$$

subject to

$$M = Y + S, \tag{0.8}$$

with $\|S\|_1 = \sum_{i,j}\left|S_{ij}\right|$, and λ is a relaxation parameter whose value is discussed in Section 6.

The augmented Lagrange function

The Lagrange function $L(Y,S,U,V,\Lambda)$ for our problem is equal to:

$$\frac{1}{2}\sum_{(i,j)\in\Omega}(Y_{i,j} - (UV^t)_{i,j})^2 + \lambda\,\|S\|_1 + \sum_{(i,j)\in\Omega}\Lambda_{ij}(M_{ij} - Y_{ij} - S_{ij}). \tag{0.9}$$

In order to enforce the constraint $M_{ij} = Y_{ij} - S_{ij}$ for all $(i,j)\in\Omega$, we introduce the following augmented Lagrange function

$$L^{aug}(Y,S,U,V,\Lambda) = L(Y,S,U,V) + \rho\,\frac{1}{2}\sum_{(i,j)\in\Omega}(M_{ij} - Y_{ij} - S_{ij})^2. \tag{0.10}$$

We now introduce an Alternating Direction Method of Multipliers. This method consists of solving iteratively in all the variables one after the other, and then updating the dual variable. For this purpose, we compute in the next

subsections the optimum value for the problem of minimizing the augmented Lagrange function with respect to each variable.

Individual minimization subproblems in *Y*, *S*, *U*, and *V*

Minimization with respect to Y.
The problem reduces to minimizing the function of the variable Y given by

$$\frac{1}{2}\sum_{i,j}(Y_{ij}-(UV^t)_{ij})^2+\sum_{(i,j)\in\Omega}\Lambda_{ij}(M_{ij}-Y_{ij}-S_{ij})+\rho\frac{1}{2}\sum_{(i,j)\in\Omega}(M_{ij}-Y_{ij}-S_{ij})^2$$

Let us denote by Y^* a solution to this problem. We have to consider two cases separately: either $(i,j)\in\Omega$, or $(i,j)\notin\Omega$. The case $(i,j)\notin\Omega$ is obvious, since it is straightforward to check that $Y^*_{ij}=(UV^t)_{ij}$ is a solution. Setting the partial derivative to zero gives the result of the case $(i,j)\in\Omega$. To summarize, we obtain

$$Y^*_{ij}=\begin{cases}\frac{1}{1+\rho}((UV^t)_{ij}+\Lambda_{ij}+\rho(M_{ij}-S_{ij})), \text{ if}(i,j)\in\Omega,\\ (UV^t)_{ij}\text{otherwise.}\end{cases}\tag{0.11}$$

Minimization with respect to S.
We have to minimize the function of S given by

$$\lambda\,\|S\|_1+\sum_{(i,j)\in\Omega}\Lambda_{ij}(M_{ij}-Y_{ij}-S_{ij})+\rho\frac{1}{2}\sum_{(i,j)\in\Omega}(M_{ij}-Y_{ij}-S_{ij})^2.\tag{0.12}$$

This can be performed by optimizing each component of S independently of the other. As for the case of minimizing with respect to Y, we distinguish between two cases, while if $(i,j)\notin\Omega$ one easily checks that $S^*_{ij}=0$. If $(i,j)\in\Omega$, we will use the following result.

Theorem 0.1. *The solution to*

$$\min_{x\in\mathbb{R}}\frac{1}{2}(y-x)^2+\lambda\,|x|\tag{0.13}$$

is given by

$$x^*=\begin{cases}y-\lambda & \text{if} \quad y>\lambda,\\ y+\lambda & \text{if} \quad y<\lambda,\\ 0 & \text{otherwise.}\end{cases}\tag{0.14}$$

Based on this result, we easily obtain

$$S^*_{ij}=\begin{cases}-M_{ij}+Y_{ij}+\Lambda_{ij}-\frac{\lambda}{\rho}, \text{if} -M_{ij}+Y_{ij}+\Lambda_{ij}>\frac{\lambda}{\rho},\\ -M_{ij}+Y_{ij}+\Lambda_{ij}+\frac{\lambda}{\rho}, \text{if} -M_{ij}+Y_{ij}+\Lambda_{ij}<\frac{\lambda}{\rho},\\ 0 \quad \text{otherwise.}\end{cases}\tag{0.15}$$

Minimization with respect to U and V.
We just have to use the fixed point subroutine given by (0.6).

Our Bregman proximal-type ADMM

We will choose the starting values as follows. Set $U^{(0)}$, $V^{(0)}$, $\Lambda^{(0)}$, $S^{(0)}=0$. Set

$$Y^0_{ij}=\begin{cases}M_{ij},\ \forall(i,j)\in\Omega,\\ \text{imputation by the mean over all other observed values row } i, \forall(i,j)\notin\Omega.\end{cases}$$

Note that mean imputation is a widely used approach for dealing with missing data. This is also the most basic one. One of the most efficient method for missing data is the proposal of [19]. However, this latter is based on standard multivariate analysis. It does not take into account the nonnegativity of the data, and it does not address the joint problem of extracting relevant features.

The Bregman-Proximal point ADMM is then given by

1. Set $S=S^{(k)}, U=U^{(k)}, V=V^{(k)}, \Lambda=\Lambda^{(k)}$ and obtain $Y^{(k+1)}=Y^*$ given by (0.11),
2. Set $Y=Y^{(k+1)}, U=U^{(k)}, V=V^{(k)}, \Lambda=\Lambda^{(k)}$ and obtain $S^{(k+1)}=S^*$ given by (0.15),
3. Set $Y=Y^{(k+1)}, S=S^{(k+1)}, U=U^{(k)}, \Lambda=\lambda^{(k)}$ and obtain $U^{(k+1)}$ using the fixed point method (0.6)
4. Set $Y=Y^{(k+1)}, S=S^{(k+1)}, U=U^{(k+1)}, \Lambda=\Lambda^{(k)}$ and obtain $U^{(k+1)}$ using the fixed point method (0.6).
5. Set $\Lambda^{(k+1)}=\Lambda^{(k)}+M-Y-S$

Choosing the value of λ

The choice of the parameter λ is crucial for the good performances of the proposed method. We performed a selection of λ using the following approach.

1. Propose an a priori range of values for λ such that its maximum values leads to S equals the null matrix at optimality.
2. For each value of λ, select a set $\mathcal{S}$ of s entries chosen uniformly at random in M and consider them as missing data temporarily.
   (a) Find the solution of (0.7), where the missing data incorporate the set of entries which were artificially declared as missing in the previous step.
   (b) Compute the average squared error on the data artificially declared as missing. Denote this quantity by err_λ.
3. Select λ in the prescribed range as the one which minimizes the average squared error err_λ.

Results

Description of the data

In this section, we apply our method to bladder cancer expression data. The number of new patients affected by bladder cancer in 2013 attained 10,000 in France, thus improving the diagnosis of bladder cancer is a Public Health priority. Determining the genes responsible for bladder cancer would undoubtedly permit to design an efficient and adapted set of medical treatments.

First of all the treatment should depend on the advancement of the tumor. For this purpose, researchers have gathered important gene expression data and the corresponding state of the malignant tumor in the bladder of 100 patients in the Lyon region (France), as described in [20]. The prospective multicentre study has been performed between September 2007 and May 2008, it formerly included 108 bladder tumours (45 pTa, 35 pT1 and 28>pT1). In this study, 34 genes have been selected from the lists provided by the Biometric Research Branch class comparison analyses. For this purpose, researchers have gathered important gene expression data and the corresponding state of the malignant tumor in the bladder of 100 patients in the Lyon region (France), as described in [20]. From the lists of genes provided by the Biometric Research Branch class comparison analyses [19], the microarray results of 34 selected differentially expressed genes were analyzed for validation using real-time quantitative PCR in other bladder tumour cohort.

From the statistical perspective, the data can be analyzed using PCA, cluster analysis, and polytomic logistic regression. However, due to the noisy nature of the data together with the presence of outliers and missing data, such methods fails in producing interesting results. For instance, the usually very efficient sparse principal component analysis returns contradictory number of features when the alpha sparsity controlling parameter ranges from 0 to 0.5, see Fig. 2. Our approach in this section is to use Non-negative Matrix Factorization in order to jointly take into account the data's intrinsic non-negative nature and the necessity of clustering the data by performing efficient feature extraction. One of the main challenges in the study of such data sets is to take into account possible outliers. For the bladder cancer dataset, some outliers have been observed by using standard PCA visualization, thus enforcing the need to automatically detect such phenomena in order to avoid subsequent misinterpretations of the genes' respective influences on the tumor state.

The data array consists of one first column providing the tumor state. The next 34 columns provide the expression of 34 genes. The array has 100 lines which correspond to the number of patients. There are two principal classes of tumors:

- $TVNIM$: noninfiltrating tumors;
- $TVIM$: infiltrating tumors.

The tumor states have been classified into the following groups:

- Ta : noninfiltrating tumor in Urothelium;
- $T1a$: noninfiltrating tumor in Urothelium and parts of the chorion;
- $T1b$: noninfiltrating tumor in Urothelium and the full chorion;
- $> T1$: infiltrating tumor.

In the standard classification, the last group of the list incorporates states $T2$ to $T4b$.

Experimental results

The ADMM algorithm was run on the experimental data. Using such an elaborate feature extraction method can be justified by the fact that existing methods fail to

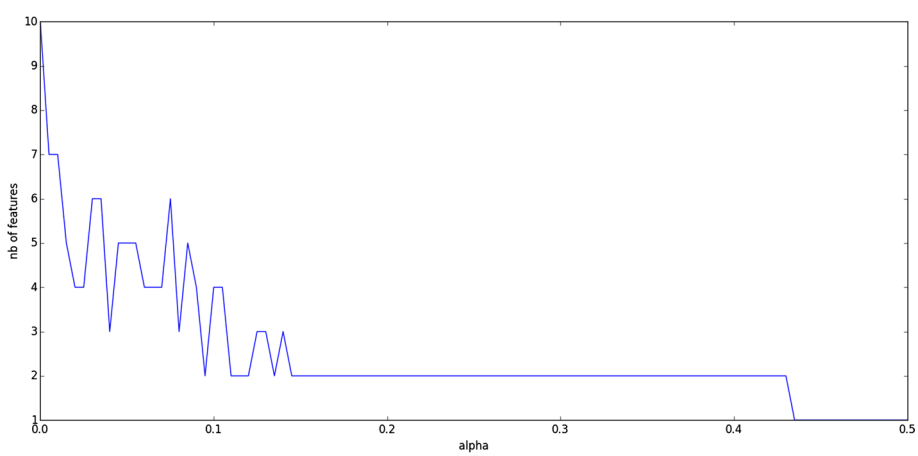

Fig. 2 Sparse PCA fails in finding relevant number of features

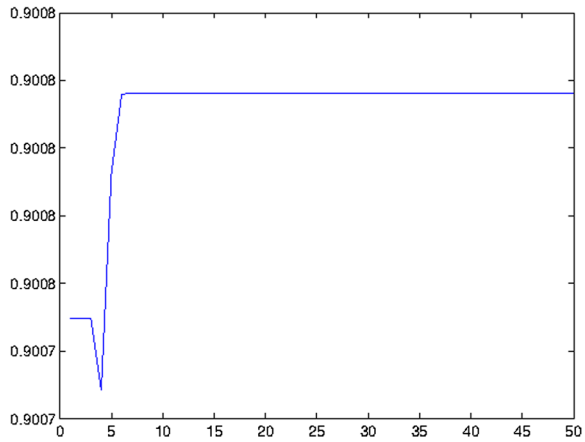

Fig. 3 The average squared prediction error on the artificially declared as missing entries as a function of $-1 - 2/5 * \log(\lambda/50)$

achieve this extraction. The choice of λ in our method was obtained using the strategy described in Section 6. The result obtained by this strategy is depicted in Fig. 3.

Based on the optimal choice of λ, the algorithm performances and the estimation results are depicted in Figs. 4 and 5.

The first subplot of Fig. 4 represents the matrix S after convergence, while the second subplot is the matrix V^t. The third subplot, for its part, represents the distance in Frobenius norm between two successive iterates of Λ. Finally, the fourth subplot represents the evolution of the relative error between M and its NMF $U^{(k)}V^{(k)^t}$.

Figure 5 shows the cluster index in each subgroup "pTa", "pT1a", "pT1b", and "more serious than pT1" (i.e., ">pT1"). We see a sort of continuous drift in these cluster indices from pTa to >pT1. Indeed,

- pTa mostly consists of 3 subgroups indexed by $\{1, 4, 6\}$;
- pT1a mostly consists of 3 subgroups indexed by $\{4, 5\}$;
- pT1b mostly consists of only one group, which is $\{5\}$;
- finally, >pT1 mostly consists of 5 subgroups indexed by $\{2, 3, 5, 7, 8\}$.

The intersection of the index subsets between two adjacent states is always a singleton, up to a discarded minority of individuals. The cluster indexed by 5 appears at medium to serious levels. The lowest level is characterized by cluster 6 while the most serious level is characterized by the more significant appearance of clusters 2,3,7 and 8.

Comparison using a Gaussian mixture model selection

The problem of choosing the number of clusters K a priori is a difficult one. This is usually done by comparing the penalized maximum likelihood values for different values of K and choosing the maximum one. Model selection can be performed too using the Bayesian Information Criterion (BIC). This criterion is the opposite of the maximum likelihood value penalized with $\log(n)\times$ the number of real parameters to estimate.

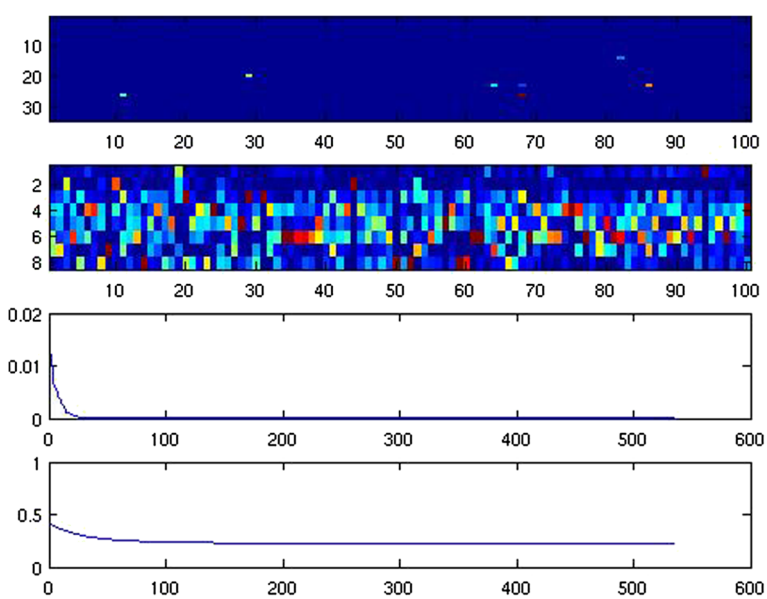

Fig. 4 The factorization and convergence curves. The first subplot is S after convergence. The second subplot is V^t. The third subplot, shows the distance between two successive iterates of Λ. The fourth subplot shows the relative error between M and its NMF $U^{(k)}V^{(k)^t}$ as a function of iteration number

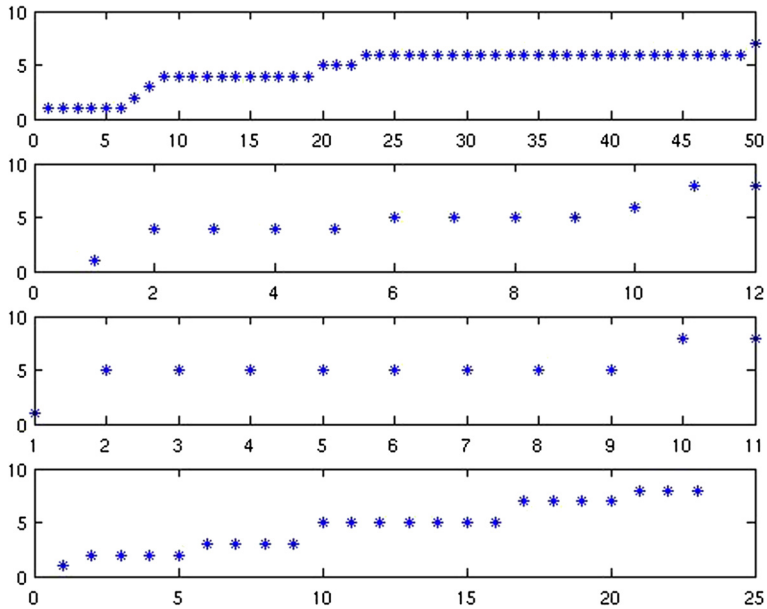

Fig. 5 Cluster index for each group of patients. Subplot 1 corresponds to pTa, subplot 2 to pT1a, subplot 3 to pT1b, and subplot 4 to > pT1

The first attempt on raw data failed to provide any useful information, due to outliers and missing data. This criterion has then been applied on the gene expression part of our denoised array, to determine the best way to cluster the set of genes. The number of mixture components has ranged from 1 to 29, and at each time the Bayesian information criterion for the current model fit has been computed (more precisely, for pretty prints, the logarithm of $x - minBIC$, where $minBIC$ is the smallest obtained BIC). As can be seen in obtained plot depicted in Fig. 6, the criterion has not provide any obvious result when considering the whole data. However, applying it on the 3 principal components of the denoised data emphasizes that the optimal number of clusters is 4, as previously. Such a result were encouraging, as we have 4 tumor states in the array. We then have performed a PCA on the raw data while colorizing each of the 4 clusters provided by the Gaussian mixture model. Obtained results are depicted in Fig. 7, they are coherent with the tumor state of each patient, and with results obtained in the previous section.

Some more precise investigations should now be performed in order to understand the biological meaning of these clusters, i.e., to understand the factors of gravity in this cancer.

Conclusion

In this article, a new way to find a relevant dictionary for extracting the relevant features in a given dataset has been presented, in an original context of missing values and outliers. The well-known Non-negative Matrix Factorization (NMF) method has been extended on denoised data, where missing values have been guessed and outliers have been detected, leading to a mixture of Bregman proximal methods and the of Augmented Lagrangian scheme.

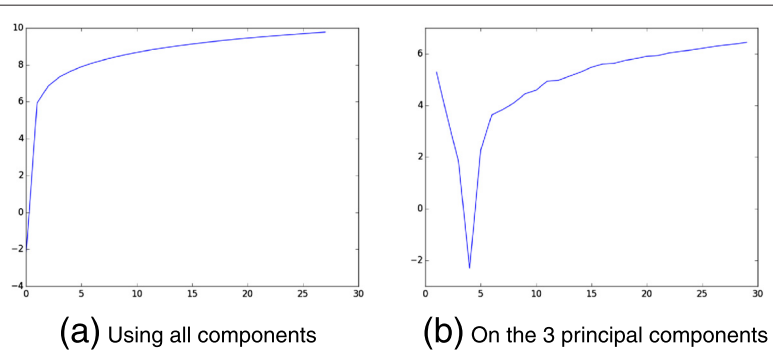

Fig. 6 Determination of the optimal number of clusters in denoized data: number of clusters for easting values and (log of) Bayesian Information Criterion (BIC) for northing ones

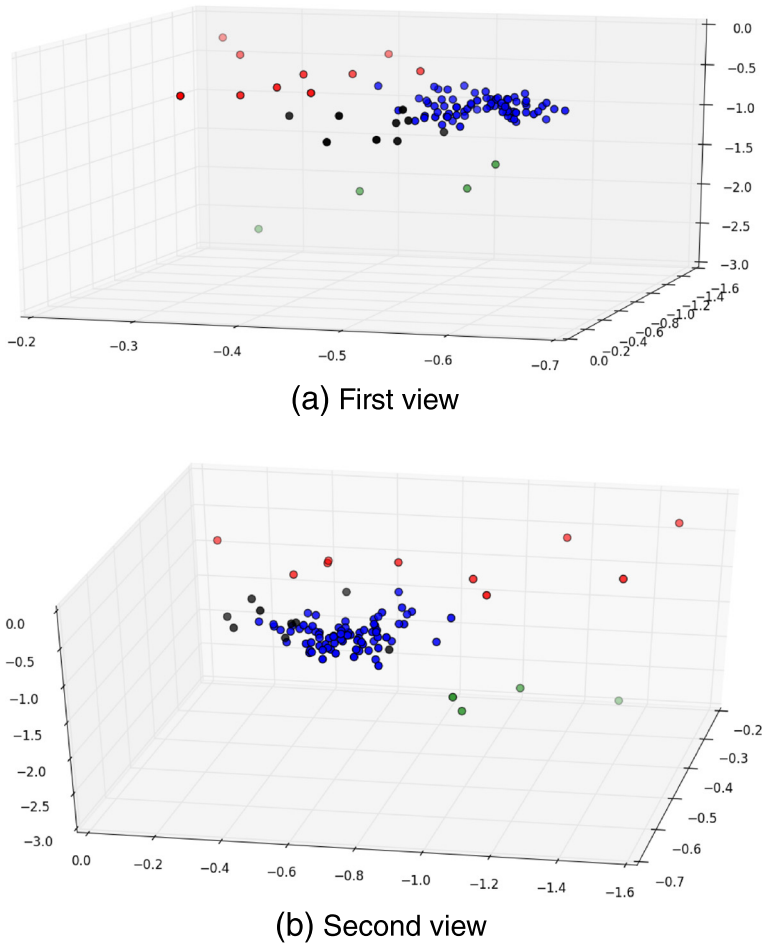

Fig. 7 PCA on raw data, colorized according to their cluster provided by the GMM

Finally, an application to the analysis of gene expression data of patients with bladder cancer has been provided for illustration purpose.

Acknowledgements
The first and second authors were funded by the grant Eléments Transposables from the Région Franche-Comté. The first author would like to acknowledge the help of Haokun Li for the preparation of the manuscript. The work of Regis Delage Mouroux and Michele Jouvenot was supported by the Région Franche-Comté and by the Ligue Contre le Cancer (CCIR-GE). P. Huetz acknowledges the Montbéliard and Besançon Leagues against Cancer for financial support.

Declarations
This article has been published as part of BMC Bioinformatics Volume 17 Supplement 8, 2016. Selected articles from the 11th International Symposium on Bioinformatics Research and Applications (ISBRA '15): bioinformatics. The full contents of the supplement are available online https://bmcbioinformatics.biomedcentral.com/articles/supplements/volume-17-supplement-8.

Funding
Publication costs for this work were funded by FEMTO-ST, Département DISC and the AND team, the Laboratoire de Mathematiques de Besancon, the National Physical Laboratory, the Région Franche-Comtéx and by the Ligue Contre le Cancer (CCIR-GE).

Availability of data and materials
The code is available by email request to stephane.chretien@npl.co.uk. The data sets are available by email request to francoise.descotes@chu-lyon.fr.

Authors' contributions
SC proposed the Bregman proximal method, wrote the initial Matlab code and performed the preliminary experiments. CG did the experimental study, performed the comparison with Sparse PCA and wrote a large part of the paper. BC and PH perfomed large scale numerical experiments and verifications. RD-M and MJ contributed the biological interpretation. FD produced the dataset and supervised methodology for the biological part of the paper. All authors read and approved the final manuscript.

Competing interests
The authors declare that they have no competing interests.

Consent for publication
Not applicable.

Ethics approval and consent to participate
Not applicable.

Author details
[1] National Physical Laboratory, Hampton Road, Teddington, Middlesex, UK.
[2] FEMTO-ST Institute, UMR 6174 CNRS, DISC Computer Science Department, Université de Bourgogne Franche-Comté, 16, route de Gray, 25000 Besançon,

France. [3]ISIFC, Université de Bourgogne Franche-Comté, 23, rue Alain Savary, 25000 Besançon, France. [4]EA 3922/IFR133, UFR Sciences et Techniques, Université de Bourgogne Franche-Comté, 16, route de Gray, 25000 Besançon, France. [5]ABC&T, 33, rue Charles Nodier, 25000 Besançon, France. [6]Service de Biochimie et Biologie Moléculaire Sud, Pavillon 3D, Centre Hospitalier Lyon Sud, Pierre Bénite, Cedex 69495 Lyon, France.

Published: 31 August 2016

References

1. Xu W, Liu X, Gong Y. 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. ACM; 2003. p. 267–73.
2. Berry MW, Browne M. Email surveillance using non-negative matrix factorization. Comput Math Organ Theory. 2005;11(3):249–64.
3. Jia S, Qian Y. Constrained nonnegative matrix factorization for hyperspectral unmixing. Geosci Remote Sens IEEE Trans. 2009;47(1): 161–73.
4. Guillamet D, Vitria J. Non-negative matrix factorization for face recognition. In: Topics in Artificial Intelligence. Springer; 2002. p. 336–44.
5. Chan TH, Ma WK, Chi CY, Wang Y. A convex analysis framework for blind separation of non-negative sources. Signal Process IEEE Trans. 2008;56(10):5120–34.
6. Kim H, Park H. Sparse non-negative matrix factorizations via alternating non-negativity-constrained least squares for microarray data analysis. Bioinformatics. 2007;23(12):1495–502.
7. Li Y, Sima DM, Cauter SV, Sava C, Anca R, Himmelreich U, Pi Y, Van Huffel S. Hierarchical non-negative matrix factorization (hnmf): a tissue pattern differentiation method for glioblastoma multiforme diagnosis using mrsi. NMR Biomed. 2013;26(3):307–19.
8. Lee DD, Seung HS. Learning the parts of objects by non-negative matrix factorization. Nature. 1999;401(6755):788–91.
9. Esser E, Möller M, Osher S, Sapiro G, Xin J. A convex model for nonnegative matrix factorization and dimensionality reduction on physical space. Image Process IEEE Trans. 2012;21(7):3239–52.
10. Recht B, Re C, Tropp J, Bittorf V. Factoring nonnegative matrices with linear programs. In: Advances in Neural Information Processing Systems; 2012. p. 1214–22.
11. Gillis N, Vavasis SA. Fast and robust recursive algorithmsfor separable nonnegative matrix factorization. Pattern Analysis and Machine Intelligence, IEEE Transactions on. 2014;36(4):698–714.
12. Sra S, Dhillon IS. Generalized nonnegative matrix approximations with bregman divergences. In: Advances in Neural Information Processing Systems; 2005. p. 283–90.
13. Li L, Lebanon G, Park H. Fast bregman divergence nmf using taylor expansion and coordinate descent. In: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM; 2012. p. 307–15.
14. Bubeck S. Convex optimization: Algorithms and complexity. Foundations and Trends® in Machine Learning. 2015;8(3-4):231–357.
15. Candès EJ, Li X, Ma Y, Wright J. Robust principal component analysis? J ACM (JACM). 2011;58(3):11.
16. Zhou T, Tao D. Godec: Randomized low-rank & sparse matrix decomposition in noisy case. In: International Conference on Machine Learning. Omnipress; 2011.
17. Parikh N, Boyd SP. Proximal algorithms. Foundations Trends Optim. 2014;1(3):127–239.
18. Gillis N. The why and how of nonnegative matrix factorization. Regularization Optim Kernels Support Vector Mach. 2014;12:257.
19. Husson F, Josse J. missmda: Handling missing values with/in multivariate data analysis (principal component methods). R package version. 2010;1(2):.
20. Descotes F, Dessen P, Bringuier PP, Decaussin M, Martin PM, Adams M, Villers A, Lechevallier E, Rebillard X, Rodriguez-Lafrasse C, et al. Microarray gene expression profiling and analysis of bladder cancer supports the sub-classification of t1 tumours into t1a and t1b stages. BJU Intl. 2014;113(2):333–42.