REGULARIZED SPLIT GRADIENT METHOD FOR NONNEGATIVE MATRIX FACTORIZATION

Henri Lantéri, Céline Theys, Cédric Richard, David Mary

Laboratoire Fizeau, Université de Nice Sophia-Antipolis
Observatoire de la Côte d'Azur, UMR CNRS 6525, Parc Valrose 06108 Nice France
E-mail: Henri.Lanteri, Celine.Theys, Cedric.Richard, David.Mary@unice.fr

ABSTRACT

This article deals with a regularized version of the split gradient method (SGM), leading to multiplicative algorithms. The proposed algorithm is available for the optimization of any divergence depending on two data fields under positivity constraint. The SGM-based algorithm is derived to solve the nonnegative matrix factorization (NMF) problem. An example with a Frobenius norm on both the data consistency and the penalty term is developped and applied to hyperspectral data unmixing.

Index Terms— SGM, NMF, regularization

1. INTRODUCTION

In a recent paper, [1], we proposed the Split Gradient Method (SGM) allowing to obtain the multiplicative algorithms dedicated to NMF for any convex functional expressing the data consistency, that is the discrepancy between data and model.

Extending a previous paper dedicated to the deconvolution problem, [2], we propose here a method to regularize the NMF problem in the SGM context. For such a problem, the regularization functional acts on the columns of the unknown matrices.

The method is founded on the separation of the negative gradient, not only on the discrepancy function but also on the penalty term, this gives clearly gradient descent algorithms which convergence is ensured with a search on the step size; this is not the case in [3] and references therein. We show as an example the forms of the algorithms in the case of a Tikhonov regularization for smoothness constraints. Obviously, others penalty functions can be used as well.

In the present communication we use the Frobenius norm to express both the data consistency and the regularization term. Clearly any other convex divergence can be used for both terms.

In section 2 we specify the problem at hand and the notations for NMF while in section 3 we briefly recall the main points of the SGM. In section 4, we develop our argumentation to justify the decomposition of the gradient of the penalty term, the application of the regularization to hyperspectral imagery is given in section 5. A numerical example is finally shown in section 6.

2. NONNEGATIVE MATRIX FACTORIZATION

We consider here the problem of nonnegative matrix factorization (NMF), which is now a popular dimension reduction technique, employed for non-subtractive, part-based representation of nonnegative data. Given a data matrix $\mathbf{V}$ of dimension $F \times N$ with nonnegative entries, the NMF consists of seeking a factorization of the form

$$\mathbf{V} \approx \mathbf{WH} \tag{1}$$

where $\mathbf{W}$ and $\mathbf{H}$ are matrices, with non negative entries, of dimensions $F \times K$ and $K \times N$, respectively. Dimension K is usually chosen such that $FK + KN \ll FN$, hence reducing the data dimensionality.

The factorization (1) is usually sought through the minimization problem

$$\min_{\mathbf{W},\mathbf{H}} \mathcal{J}(\mathbf{V},\mathbf{W},\mathbf{H}) \quad \text{s.t.} \quad [\mathbf{W}]_{ij} \geq 0, [\mathbf{H}]_{ij} \geq 0 \tag{2}$$

To solve (2), one can use a minimization method of the SGM-type, alternatively on $\mathbf{W}$ and $\mathbf{H}$.

3. SPLIT GRADIENT METHOD

We briefly recall the main points of the SGM in the NMF context, [4, 5, 1]. The SGM is based on the Karush-Kuhn-Tucker conditions at the optimum $\mathbf{W}^*$ and $\mathbf{H}^*$. We only develop the minimization w.r.t $\mathbf{W}$ (minimization w.r.t $\mathbf{H}$ follows exactly the same scheme), this leads to:

$$\begin{aligned}[\mathbf{W}^*]_{ij}[\nabla_W \mathcal{J}(\mathbf{V},\mathbf{W}^*,\mathbf{H})]_{ij} = 0 \\ \Leftrightarrow [\mathbf{W}^*]_{ij}[-\nabla_W \mathcal{J}(\mathbf{V},\mathbf{W}^*,\mathbf{H})]_{ij} = 0\end{aligned} \tag{3}$$

The second formulation of eq.(3), while trivial, makes appear the descent direction. To solve this equation iteratively, three points have to be noticed. The first one is that $\mathbf{M} \cdot (-\nabla_W \mathcal{J})$ is a gradient related descent direction of $\mathcal{J}$ if $\mathbf{M}$ is a matrix with positive entries, where $\cdot$ denotes the Hadamard product. The second one is that $[-\nabla_W \mathcal{J}]_{ij}$ can always be decomposed as $[\mathcal{P}]_{ij} - [\mathcal{Q}]_{ij}$, where $[\mathcal{P}]_{ij}$ and $[\mathcal{Q}]_{ij}$ are positive entries. Last but not least, the third one is that equations of the form $\varphi(\mathbf{W}) = 0$ can be solved with a fixed-point algorithm, under some conditions on the function φ, by considering the problem $\mathbf{W} = \mathbf{W} + \varphi(\mathbf{W})$. Implementing this fixed-point strategy with (3) and using

$$[\mathbf{M}]_{ij} = \frac{1}{[\mathcal{Q}]_{ij}} \tag{4}$$

we obtain the gradient related descent algorithm

$$[\mathbf{W}^{k+1}]_{ij} = [\mathbf{W}^k]_{ij} + \alpha_{ij}^k \frac{[\mathbf{W}^k]_{ij}}{[\mathcal{Q}^k]_{ij}} \left([\mathcal{P}^k]_{ij} - [\mathcal{Q}^k]_{ij}\right) \tag{5}$$

with α_{ij}^k a positive step size that allows to control convergence of the algorithm. We take the same step size for all indices. This step size must be searched by simplified search methods (Armijo rule for example, [6]) in the range $[0, \alpha_{\max}^k]$ where $\alpha_{\max}^k$ is explicitely computed from (5) to ensure the non-negativity of the components of $\mathbf{W}^k$. It can be easily shown that $\alpha_{\max}^k$ is always greater than one and for $\alpha^k = 1$, we get the very simple and well-known multiplicative form:

$$[\mathbf{W}^{k+1}]_{ij} = [\mathbf{W}^k]_{ij} \frac{[\mathcal{P}^k]_{ij}}{[\mathcal{Q}^k]_{ij}} \tag{6}$$

Positiveness is satisfied if $[\mathbf{W}^0]_{ij} > 0$, but convergence of the algorithm (6) is not guaranteed.

4. REGULARIZATION

In this section, we propose to introduce the principles of regularization in the context of SGM, with the objective to give multiplicative algorithms for NMF. The regularization penalty term is applied separately on the columns of $\mathbf{W}$ and $\mathbf{H}$ and is added to the data consistency $\mathcal{D}(\mathbf{V}, \mathbf{WH})$. Then the problem (2) is expressed as follows:

$$\min_{\mathbf{W},\mathbf{H}} \mathcal{J}(\mathbf{V}, \mathbf{W}, \mathbf{H}) \quad [\mathbf{W}]_{ij} \geq 0, [\mathbf{H}]_{ij} \geq 0 \tag{7}$$

$$\mathcal{J}(\mathbf{V}, \mathbf{W}, \mathbf{H}) = \mathcal{D}(\mathbf{V}, \mathbf{WH}) + \gamma \mathcal{F}_1(\mathbf{W}) + \mu \mathcal{F}_2(\mathbf{H}) \tag{8}$$

The terms $\mathcal{F}_1(\mathbf{W})$ and $\mathcal{F}_2(\mathbf{H})$ are the penalty functions, γ and μ are the regularization factors.

4.1. Regularized SGM

The general rules given for SGM remain true for the regularized versions of the algorithms. For the minimization with respect to $\mathbf{W}$:

$$-\nabla_W \mathcal{J} = -\nabla_W \mathcal{D} - \gamma \nabla_W \mathcal{F}_1 \tag{9}$$

If we denote the decomposition of the negative gradients as

$$[-\nabla_W \mathcal{D}]_{ij} = [\mathbf{P}]_{ij} - [\mathbf{Q}]_{ij} \tag{10}$$

$$[-\nabla_W \mathcal{F}_1]_{ij} = [\mathbf{P}_R]_{ij} - [\mathbf{Q}_R]_{ij} \tag{11}$$

We have

$$[-\nabla_W \mathcal{J}]_{ij} = [\mathcal{P}]_{ij} - [\mathcal{Q}]_{ij} \tag{12}$$

with

$$[\mathcal{P}]_{ij} = [\mathbf{P}]_{ij} + \gamma [\mathbf{P}_R]_{ij}, \qquad [\mathcal{Q}]_{ij} = [\mathbf{Q}]_{ij} + \gamma [\mathbf{Q}_R]_{ij} \tag{13}$$

The same type of decomposition is available for the gradient with respect to $\mathbf{H}$.

4.2. Tikhonov regularization

We develop the argumentation in the simple case of the Tikhonov regularization, expressing some smoothness property of the solution:

$$\mathcal{F}(\mathbf{X}) = \sum_{ij} ([\mathbf{X}]_{ij} - c)^2 \tag{14}$$

possibly with $c = 0$ or:

$$\mathcal{F}(\mathbf{X}) = \sum_{ij} [DX]_{ij}^2 \tag{15}$$

where D is the first derivative or the second derivative operator. In all these penalties, we always express the discrepancy between the current solution and the reference solution by means of the squared Frobenius norm. More generally, the penalty functions used in the literature express a discrepancy between two functions of the solution. Note that the Tikhonov regularization used in this paper is just an example.

5. APPLICATION TO HYPERSPECTRAL IMAGERY

Hyperspectral imaging has received considerable attention in the last few years. See for instance [7], [8] and references therein. It consists of data acquisition with high sensitivity and resolution in hundreds contiguous spectral bands, geo-referenced within the same coordinate system. With its ability to provide extremely detailed data regarding the spatial and spectral characteristics of a scene, this technology offers immense new possibilities

in collecting and managing information for civilian and military application areas.

Each vector pixel of an hyperspectral image characterizes a local spectral signature. Usually, one consider that each vector pixel can be modeled accurately as a linear mixture of different pure spectral components, called endmembers. Referring to our notations, each column of $\mathbf{V}$ can thus be interpreted as a spectral signature obtained by linear mixing of the spectra of endmembers, i.e., the columns of $\mathbf{W}$. The problem is then to estimate the endmember spectra $\mathbf{W}$ and the abundance coefficients $\mathbf{H}$ from the spectral signatures $\mathbf{V}$.

We propose to regularize the columns of $\mathbf{W}$, the endmember spectra, as in terrestrial imagery.

5.1. Regularization on the columns of W

In what follows, we shall implement the first order regularization (15) as $\mathcal{F}(\mathbf{X}) = \sum_{ij}[DX]_{ij}^2$. Let us note that when we use the expression DX, we can write:

$$[\mathcal{D}X]_{ij} = [\mathbf{X}]_{ij} - [\mathbf{AX}]_{ij} \tag{16}$$

where, when we use the first derivative, $\mathbf{AX}$ corresponds to the convolution of each column of the current solution $\mathbf{X}$ by the mask $[1, 0, 0]$, (a shifted version of the solution) and when we use the second derivative, $\mathbf{AX}$ corresponds to the convolution of each column of $\mathbf{X}$ by the mask $[0.5, 0, 0.5]$, (a low pass version of the solution). We must first compute the gradient of the penalty term with respect to the elements of the columns of $\mathbf{W}$. If the reference solution depends on the current solution, we have:

$$\mathcal{F}_1(\mathbf{W}) = \frac{1}{2}\sum_j\sum_i([\mathbf{AW}]_{ij} - [\mathbf{W}]_{ij})^2 \tag{17}$$

then, the opposite of the gradient can be expressed in the matrix form

$$-\frac{\partial \mathcal{F}_1}{\partial[\mathbf{W}]_{ij}} = [\mathbf{A}^T\mathbf{W} - \mathbf{A}^T\mathbf{AW} + \mathbf{AW} - \mathbf{W}]_{ij} \tag{18}$$

The decomposition (11) that must be used in SGM is:

$$\mathbf{P}_R = \mathbf{A}^T\mathbf{W} + \mathbf{AW}, \qquad \mathbf{Q}_R = \mathbf{A}^T\mathbf{AW} + \mathbf{W} \tag{19}$$

If the reference solution is constant, the penalty term is:

$$\mathcal{F}_1(\mathbf{W}) = \frac{1}{2}\sum_j\sum_i(c - [\mathbf{W}]_{ij})^2 \tag{20}$$

then, the opposite of the gradient can be expressed in matrix form

$$-\frac{\partial \mathcal{F}_1}{\partial[\mathbf{W}]_{ij}} = c - [\mathbf{W}]_{ij} \tag{21}$$

Then, the decomposition (11) is:

$$\mathbf{P}_R = c, \qquad \mathbf{Q}_R = \mathbf{W} \tag{22}$$

5.2. Complete form of the algorithm

We also use a Frobenius norm for data consistency:

$$\mathcal{D}(\mathbf{V}, \mathbf{WH}) = \frac{1}{2}\sum_j\sum_i([\mathbf{WH}]_{ij} - [\mathbf{V}]_{ij})^2 \tag{23}$$

with:

$$-\frac{\partial \mathcal{D}}{\partial[\mathbf{W}]_{ij}} = [\mathbf{VH}^T]_{ij} - [\mathbf{WHH}^T]_{ij} = [\mathbf{P}]_{ij} - [\mathbf{Q}]_{ij} \tag{24}$$

In this case, with equations (19), (24), the algorithm (5) writes:

$$[\mathbf{W}^{k+1}]_{ij} = [\mathbf{W}^k]_{ij} + \alpha^k[\mathbf{W}^k]_{ij}\left(\frac{[\mathbf{VH}^{Tk}]_{ij} + \gamma[(\mathbf{A}^T + \mathbf{A})\mathbf{W}^k]_{ij}}{[\mathbf{W}^k\mathbf{H}^k\mathbf{H}^{Tk}]_{ij} + \gamma[(\mathbf{A}^T\mathbf{A} + \mathbf{I})\mathbf{W}^k]_{ij}} - 1\right) \tag{25}$$

If the reference solution is constant, the previous expression of the algorithm becomes:

$$[\mathbf{W}^{k+1}]_{ij} = [\mathbf{W}^k]_{ij} + \alpha^k[\mathbf{W}^k]_{ij}\left(\frac{[\mathbf{VH}^{Tk}]_{ij} + \gamma c}{[\mathbf{W}^k\mathbf{H}^k\mathbf{H}^{Tk}]_{ij} + \gamma[\mathbf{W}^k]_{ij}} - 1\right) \tag{26}$$

For $\alpha^k = 1$, a multiplicative form for (25) and (26) are obtained. Computing the gradient over $\mathbf{H}$ of (23), and using the form (5) give the following expression for the actualization of $\mathbf{H}$ (here $\mathbf{H}$ is not regularized).

$$[\mathbf{H}^{k+1}]_{ij} = [\mathbf{H}^k]_{ij} + \beta^k[\mathbf{H}^k]_{ij}\left(\frac{[\mathbf{W}^{Tk+1}\mathbf{V}]_{ij}}{[\mathbf{W}^{Tk+1}\mathbf{W}^{k+1}\mathbf{H}^k]_{ij}} - 1\right) \tag{27}$$

where β^k is the step size.

6. SIMULATIONS RESULTS

Many simulations have been performed to validate the proposed algorithm, eqs. (25) and (27). Note that the different forms of the regularization term give approximatively the same practical results. The experiment presented in this paper corresponds to 10 linear mixtures of 3 endmembers, the length of each spectrum being 826. The three endmembers used in this example were extracted from the ENVI library [9] and correspond to the spectra of the construction concrete, green grass, and micaceous loam. A noise vector distributed according to a Gaussian distribution with zero-mean and covariance

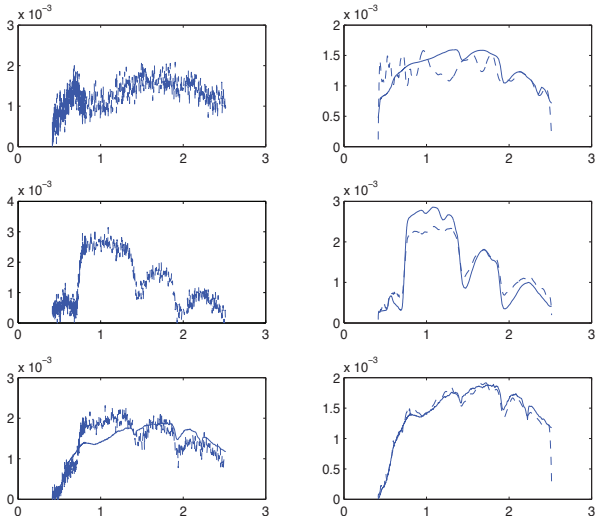

Fig. 1. Columns of $\mathbf{W}$. On each plot: solid line for true values, dashed line for estimated values. Left column: without regularization $\gamma = 0$. Right column with $\gamma = 0.1$

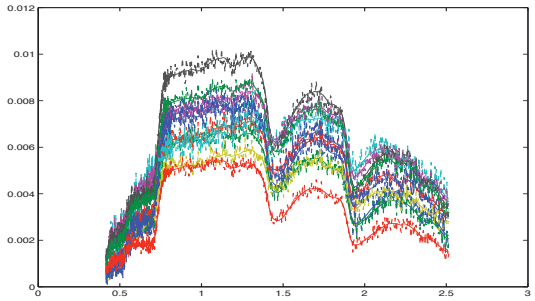

Fig. 2. Columns of $\mathbf{V}$, solid line for true values, dashed line for estimated values without regularization, $\gamma = 0$

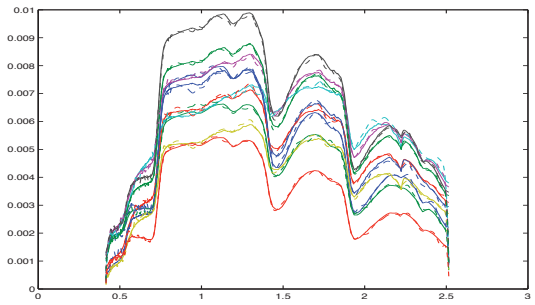

Fig. 3. Columns of $\mathbf{V}$, solid line for true values, dashed line for estimated values with $\gamma = 0.1$

matrix $\sigma^2\mathbf{I}_N$, where $\mathbf{I}_N$ is the $N \times N$ identity matrix has been added to each column of $\mathbf{V}$. Note that this statistical model assumes that the noise variances are the same in all bands. Results are given for a snr equal to $20dB$. Equations (25) and (27) have been implemented. Fig. 1 shows the estimated endmembers (columns of $\mathbf{W}$) after 12000 iterations, and compared with the true values with and without regularization. Fig. 2 and fig. 3 show the 10 reconstructed spectra in comparison with the true ones, respectively without and with regularization. We clearly see the interest of the regularization on the estimation.

7. CONCLUSION

A regularized version of the SGM has been developped for NMF, leading to the well-known multiplicative algorithms as a particular case. An application to hyperspectral imagery has been proposed and a simulation example shows clearly effectiveness of the algorithm. Let us note that the method can be applied for any convex criterion for both the data consistency and the penalty term.

8. REFERENCES

[1] H. Lantéri, C. Theys, C. Richard, and C. Févotte, "Split gradient method for nonnegative matrix factorization," in *EUSIPCO, Aalborg*, 2010.

[2] H. Lantéri, C. Aime, H. Beaumont, and P. Gaucherel, "Blind deconvolution using the Richardson-Lucy algorithm," in *European Symposium on Satellite and Remote Sensing*, 1994.

[3] A. Cichocki, R. Zdunek, and S. Amari, *Csiszár's Divergences for Non-negative Matrix Factorization: Family of New Algorithms*, vol. 3889 of *Lectures Notes in Computer Science*, Springer Berlin/ Heidelberg, 2006.

[4] H. Lantéri, M. Roche, O. Cuevas, and C. Aime, "A general method to devise maximum likelihood signal restoration multiplicative algorithms with non-negativity constraints," *Signal Processing*, vol. 54, no. 81, pp. 945–974, 2001.

[5] H. Lantéri, M. Roche, and C. Aime, "Penalized maximum likelihood image restoration with positivity constraints-multiplicative algorithms," *Inverse problems*, vol. 18, pp. 1397–1419, 2002.

[6] L. Armijo, "Minimization of functions having continuous derivatives," *Pacific Journal of Mathematics*, , no. 16, pp. 1–3, 1966.

[7] C. I. Chang, *Hyperspectral Imaging: Techniques for Spectral Detection and Classification*, Plenum Publishing Co., New York, 2003.

[8] D. A. Landgrebe, *Signal Theory Methods in Multispectral Remote Sensing*, Wiley, New York, 2003.

[9] RSI (Research Systems Inc.), *ENVI User's guide Version 4.0*, Boulder, CO 80301 USA, Sept. 2003.