NEW ALGORITHM FOR NONNEGATIVE MATRIX FACTORIZATION USING GIVENS PARAMETERIZATION

El Mostafa FADAILI and Antoine SOULOUMIAC

CEA, LIST, Laboratoire des Processus Stochastiques et Spectres, F-91191 Gif-Sur-Yvette, France.
phone: + 33 (0)1.69.08.84.91, fax: + 33(0)1.69.08.78.19
email: fadaili.el-mostafa@cea.fr, antoine.souloumiac@cea.fr

ABSTRACT

In this paper, the problem of nonnegative matrix factorization (NMF) is considered. It is formulated as the optimization of a criterion with bound constraints. We propose an approach based on Givens parameterization of some positive vector, and criterion minimization is achieved using Levenberg-Marquardt algorithm. The performance of the developed NMF method is illustrated for the separation of a linear mixture of images. [1]

1. INTRODUCTION

Nonnegative matrix factorization (NMF) was first introduced by Paatero in [14] and reformulated by Lee in [7]. NMF decomposes the data matrix as a product of two matrices that are constrained by having nonnegative values.
Given an $m \times n$ ($m \leq n$) matrix $\mathbf{V}$ with $V_{ij} \geq 0, \forall i, j$ and $r \in \mathbb{N}^* < min(m,n)$, NMF finds two nonnegative matrices $\mathbf{W}$ of size $m \times r$ and $\mathbf{H}$ of size $r \times n$ such that they minimize the functional

$$C(\mathbf{W},\mathbf{H}) = \|\mathbf{V} - \mathbf{W}\mathbf{H}\|_F^2, \;\; \mathbf{W},\mathbf{H} \geq 0, \tag{1}$$

where $\|.\|$ denotes the Frobenius norm.

NMF has been used in several research areas such as biomedical imaging [8, 15], spectroscopy [3, 13], polyphonic music transcription [16], among others. The growing interest for this decomposition technique in addition to the dimensionality reduction is due to the fact that in many applications where observed data in matrix $\mathbf{V}$ can only take nonnegative values (for example pixel in imagery data or molecular mass intensity in mass spectroscopy), the solutions of NMF are nonnegative and can be physically meaningful. Notice that in the general case, the nonnegativity constraint is not guaranteed by using PCA (Principal Components analysis) or ICA (Independent components analysis) techniques [2].

One of the most popular algorithm to solve (1) was proposed by Lee and Seung in [7]. The approach is based in multiplicative update rules for both $\mathbf{W}$ and $\mathbf{H}$: after initialization by two nonnegative matrices, the elements of $\mathbf{W}$ and $\mathbf{H}$ are multiplied by certain factor at each iteration until convergence or after *kmax* iterations. The method is very simple to implement and preserves nonnegativity at each iteration. Unfortunately, this algorithm is known to be notoriously slow to converge and still lacks convergence results.
In order to overcome some of these shortcomings, many modifications were proposed later to accelerate the convergence [4] or to guarantee the convergence to a stationary point using projected gradient methods [10]. However, the results of these methods depend widely on the initialization matrix and still present some convergence problems.

We propose in this communication a new approach to optimize criterion (1) which exhibits interesting optimization properties and convergence rate when compared to the previous methods. This new approach is based in a parameterization with Givens rotation and an optimization with the Levenberg-Marquardt algorithm which is known to be more robust than the gradient descent one.

2. ALGORITHM

2.1 Givens parameterization of positive unit norm vector

A parameterization of any real positive unit-norm vector $\mathbf{w}$ with r entries can be provided by a sequence of Givens rotations:

$$\mathbf{w} = \mathbf{R}_1(\alpha_1)\mathbf{R}_2(\alpha_2)\dots\mathbf{R}_{r-1}(\alpha_{r-1})\mathbf{e}_1 \tag{2}$$

where

$$\mathbf{R}_i(\alpha_i) = \left(\begin{array}{ccccc} c_i & \dots & -s_i & \dots & 0 \\ & \mathbf{I}_{i-1} & & \dots & 0 \\ & & & & \vdots \\ s_i & \dots & c_i & \dots & \\ 0 & \dots & & \mathbf{I}_{r-i-1} & \end{array} \right), \tag{3}$$

and

$$c_i = \cos(\alpha_i), s_i = \sin(\alpha_i), 0 \leq \alpha_i < \pi/2, \tag{4}$$

where the $r-1$ constraints $0 \leq \alpha_i < \pi/2$ are indispensable to keep $\mathbf{w}$ nonnegative and $\mathbf{e_1} = [1, 0, \dots, 0]^T$ of size $(r \times 1)$. The parameter vector associated to $\mathbf{w}(\alpha)$ is $\alpha = [\alpha_1, \dots, \alpha_{m-1}]^T$.
For nonnegative matrix $\mathbf{W}$ with unit-norm column vectors $\mathbf{w}_i(\alpha_i)$, with $\alpha_i = [\alpha_{1,i}, \dots, \alpha_{r-1,i}]^T$, $i \in \{1, \dots, r\}$, it can be parametrized by a vector $\theta = \text{vec}([\alpha_1, \dots, \alpha_r])$, where the vec$\{.\}$ operator vectorizes a matrix by stacking its columns (it is convention that column rather than row stacking is used).

[1] This work has been supported by CEA/DRT (Grant ACAV 'Imagerie Simultanée Multitraceurs').

2.2 Levenberg-Marquardt optimization

We assume here that matrix $\mathbf{W}$ is parametrized by unique parameter $\mathbf{W} = \mathbf{W}(\theta)$. Let us consider the criterion:

$$\begin{aligned} C(\theta) &= \|\mathbf{V} - \mathbf{W}(\theta)\mathbf{H}\|_F^2 \\ &= \|\mathsf{vec}(\mathbf{V} - \mathbf{W}(\theta)\mathbf{H})\|^2 \\ &= \|\mathbf{f}(\theta)\|^2 \end{aligned}$$

where $\mathbf{f}(\theta) = \mathsf{vec}(\mathbf{V} - \mathbf{W}(\theta)\mathbf{H})$.

In order to minimize this quadratic criterion, we propose to use a Levenberg-Marquardt (LM) optimization scheme [9, 11] which has become a suitable technique for non-linear least-squares problems.
We define the Jacobian $\mathbf{J}(\theta)$:

$$\mathbf{J}(\theta) = [\frac{\partial \mathbf{f}(\theta)}{\partial \theta_1}, \frac{\partial \mathbf{f}(\theta)}{\partial \theta_2}, \ldots], \tag{5}$$

and the gradient $\mathbf{g}(\theta)$ of $C(\theta)$:

$$\mathbf{g}(\theta) = 2\{\mathbf{J}(\theta)^T \mathbf{f}(\theta)\}.$$

Finally, the Hessian can be approximated by:

$$\mathbf{Hs}(\theta) \approx 2\mathbf{J}(\theta)^T \mathbf{J}(\theta).$$

The LM update step is given by:

$$\theta^{k+1} = \theta^k - (\mu_k \mathbf{I} + \mathsf{diag}(\mathbf{Hs}(\theta_k)))^{-1} \mathbf{g}(\theta^k) \tag{6}$$

where the operator $\mathsf{diag}\{.\}$ denotes the diagonal matrix constructed from diagonal elements of its argument, μ_k is nonnegative scalar and $\mathbf{I}$ is the identity matrix. We will give here the explicit form of matrix $\mathbf{J}(\theta)$. For a given column vector $\mathbf{w}_l(\alpha_l)$, $1 \le l \le r$, we consider the vector of derivatives

$$\begin{aligned} \mathbf{F}(\alpha_{i,l}) &= \frac{\partial \mathbf{f}(\alpha_{i,l})}{\partial \alpha_{i,l}} \\ &= \mathsf{vec}(\frac{\partial (\mathbf{V} - \mathbf{W}(\alpha)\mathbf{H})}{\partial \alpha_{i,l}}) \\ &= -\mathsf{vec}(\frac{\partial \mathbf{w}(\alpha_l)}{\partial \alpha_{i,l}} \mathbf{H}_l), \end{aligned}$$

where $\frac{\partial \mathbf{w}(\alpha_l)}{\partial \alpha_{i,l}}$ expression is given in appendix and $\mathbf{H}_l$ is a row vector of matrix $\mathbf{H}$.
Then, the Jacobian $\mathbf{J}(\theta)$ reads:

$$\mathbf{J}(\theta) = [\mathbf{F}(\alpha_{1,1}), \mathbf{F}(\alpha_{2,1}), \ldots, \mathbf{F}(\alpha_{m-1,r})] \tag{7}$$

To minimize criterion (1) for both $\mathbf{H}$ and θ, we propose to use alternate multiplicative update rule for $\mathbf{H}$ proposed in [7] and LM update equation (6) for $\mathbf{W}$. In fact, the same parameterization can be chosen for the row vectors of $\mathbf{H}$, and because of high computational cost of this operation (in practice, the matrix $\mathbf{H}$ is very large), we consider a simple multiplicative rule for estimation of $\mathbf{H}$. Then, the scheme of our proposed method reads:

1. Initialization by θ^0 and $\mathbf{H}^0$.

2. Calculate the Jacobian from (7)

3. Update $\mathbf{H}$ and θ:

$$\mathbf{H}_{ij}^{k+1} = \mathbf{H}_{ij}^k (\mathbf{W}(\theta^k)^T \mathbf{V})_{ij} / (\mathbf{W}(\theta^k)^T \mathbf{W}(\theta^k) \mathbf{H}^k)_{ij}$$

$$\theta^{k+1} \equiv (\theta^k - (\mu_k \mathbf{I} + \mathsf{diag}(\mathbf{Hs}(\theta^k)))^{-1} \mathbf{g}(\theta^k))[\pi/2]$$

4. Update μ_k taking into account the error $\Delta C(\theta, \mathbf{H})$

5. Repeat from 2 until $\|\Delta C(\theta, \mathbf{H})\| < \varepsilon$ or $k = maxiter$

where *maxiter* is the maximum number of iterations and $\equiv [.]$ denotes the congruent operator. The use of congruent operator $\equiv [\pi/2]$ for updating θ enforces the nonnegative values for the matrix $\mathbf{W}(\theta)$.

Concerning the initialization of the algorithm, most of the NMF algorithms use simple random initialization, *i.e.*, $\mathbf{W}$ and $\mathbf{H}$ are initialized as matrices of random numbers between 0 and 1. It is known that this kind of initialization does not generally provide a good first estimate of NMF algorithm. Another alternatives to this initialization: *centroid initialization* [17] or nonnegative SVD decomposition [1] can also be used.
As it is classical in LM implementation, the values of μ_k during the iterative process are chosen in the following way: at the beginning of the iterations, μ_0 is set to a large value, and in each iteration, if $\Delta C(\theta, \mathbf{H}) < 0$, decrease μ_k by certain amount (divided by 10 for example) to speed up the convergence; otherwise, increase μ_k value to enlarge the searching area (trust-region).

However, regarding the monotony of the algorithm, Lee and Seung have proved that the objective function decreases; *i.e.*, for two successive iterations k and $k+1$, we have $\|\mathbf{V} - \mathbf{W}(\theta^k)\mathbf{H}^{k+1}\|_F^2 \le \|\mathbf{V} - \mathbf{W}(\theta^k)\mathbf{H}^k\|_F^2$. However, the LM algorithm for updating θ ensure decreasing (with constant $\mathbf{H}$), so $\|\mathbf{V} - \mathbf{W}(\theta^{k+1})\mathbf{H}^{k+1}\|_F^2 \le \|\mathbf{V} - \mathbf{W}(\theta^k)\mathbf{H}^{k+1}\|_F^2$. We conclude that $\|\mathbf{V} - \mathbf{W}(\theta^{k+1})\mathbf{H}^{k+1}\|_F^2 \le \|\mathbf{V} - \mathbf{W}(\theta^k)\mathbf{H}^k\|_F^2$ and so the decreasing of objective function. Nevertheless, our algorithm contains a modulo $[\frac{\pi}{2}]$ step, thus general convergence results of the LM algorithm are not valid here, but seems to not affect the convergence of the method.
Finally, the complexity of the proposed algorithm is higher than [7], due in particular to the Hessian inversion operation.

3. APPLICATION TO IMAGES SEPARATION

The effectiveness of the proposed method has been illustrated for blind separation of images. This is suitable for nonnegative matrix factorization since the pixels have nonnegative values.
The $n = 3$ gray images used in this simulation are shown in Figure 1. The images are 128×128 with integer pixel intensities in $[0, 255]$. Matrix $\mathbf{H}$ of size 3×128^2 is constructed from the *vectorization* of the previous images with unit norm row vectors. Figure 2 shows $m = 9$ mixtures of the 3 images with random matrix $\mathbf{W}$ distributed uniformly between 0 and 1.

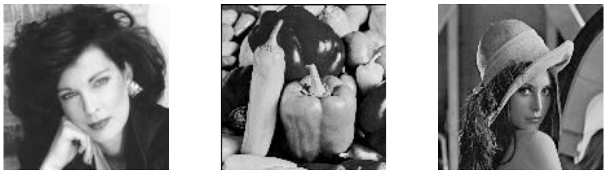

Figure 1: The $n=3$ gray source images of size 128×128.

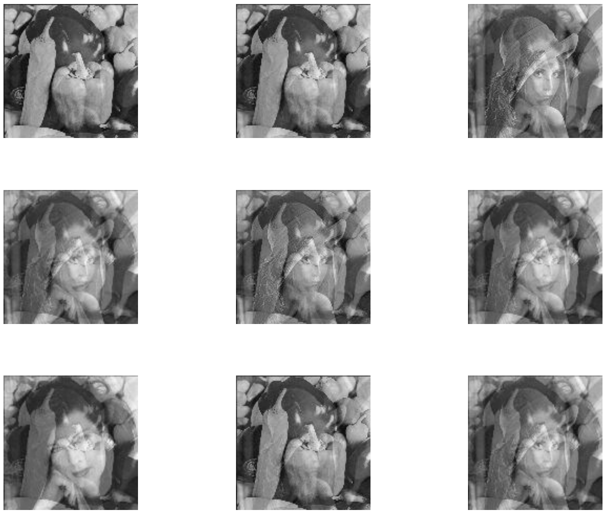

Figure 2: The $m=9$ mixed images of Figure 1 by random matrix $\mathbf{W}$ distributed uniformly between 0 and 1.

We denote by nmfgivens our proposed method and we will compare it with two other methods:

- nmfmult: nonnegative matrix factorization by multiplicative update rule [7];
- nmfpgrad: nonnegative matrix factorization using projected gradient method [10].

To measure the performance of the previous methods, we propose to use two different criteria. The first one is based on the quadratic reconstruction error between $\mathbf{V}$ and the estimated nonnegative matrices $\hat{\mathbf{W}}, \hat{\mathbf{H}}$:

$$\mathsf{err}_{\mathsf{re}} = \frac{\|\mathbf{V} - \hat{\mathbf{W}}\hat{\mathbf{H}}\|_F^2}{\|\mathbf{V}\|_F^2} \tag{8}$$

The second one is the following performance index which measures the separation quality [12]:

$$\begin{aligned} \mathsf{err}_{\mathsf{s}}(\mathbf{G}) \quad &= \quad \frac{1}{r(r-1)} \sum_{i=1}^{r} \left(\sum_{j=1}^{r} \frac{|(\mathbf{G})_{i,j}|^2}{\max\limits_{\ell} |(\mathbf{G})_{i,\ell}|^2} - 1 \right) \\ &+ \quad \frac{1}{r(r-1)} \sum_{j=1}^{r} \left(\sum_{i=1}^{r} \frac{|(\mathbf{G})_{i,j}|^2}{\max\limits_{\ell} |(\mathbf{G})_{\ell,j}|^2} - 1 \right), \end{aligned}$$

with r being the dimension of the square matrix $\mathbf{G} = \hat{\mathbf{W}}^{\dagger}\mathbf{W}$ and † denotes the pseudo-inverse operator.

Figure 3 shows the performance index $\mathsf{err}_{\mathsf{re}}$ versus iteration number. All methods are initialized by the same uniform random positive matrices $\mathbf{W}^0$ and $\mathbf{H}^0$. From the graph it is seen that nmfgivens gives better performance than nmfmult in terms of objective function minimization, but still less good than nmfpgrad one because of the update of the matrix $\mathbf{H}$. In Figure 4, the performance of nmfgivens is well improved using projected gradient estimation of the matrix $\mathbf{H}$.

The values of the performance index $\mathsf{err}_{\mathsf{s}}$ are reported in Figure 5 for two different random initializations where the separation quality is better with nmfgivens. Figure 6 shows the reconstructed 3 sources from the mixture of the Figure 2.

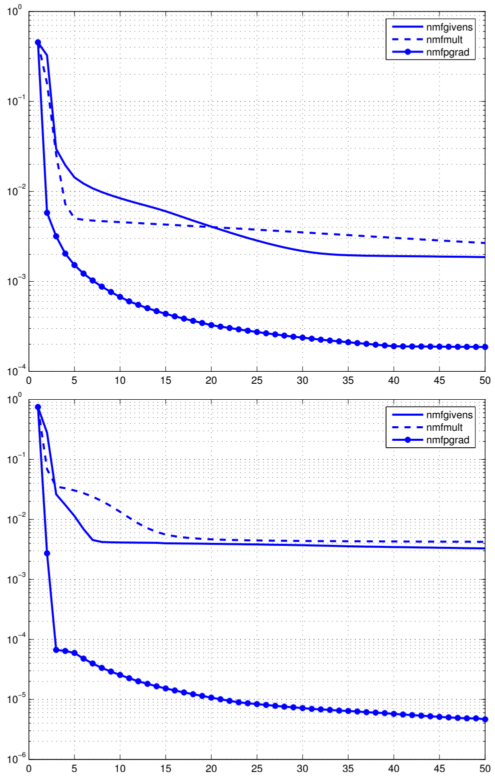

Figure 3: The performance index $\mathsf{err}_{\mathsf{re}}$ versus iteration number for nmfgivens, nmfmult, nmfpgrad.

4. CONCLUSION

In this paper, we have considered the problem of nonnegative matrix factorization and we have proposed a new approach based in a Givens parameterization of the mixing matrix and Levenberg-Marquardt optimization of some quadratic criterion.

Simulations on image data show that the proposed method gives much better results both in terms of criterion optimization and separation quality, when compared to classical algorithms for solving NMF.

As a future issue, it is interesting to investigate how to generalize the proposed method to the NMF problem taking into account the property of sparsity [5, 6]. The sparsity is achieved generally by adding some penalty term to the NMF objective function. Finally, the study of the behavior of the proposed method in noisy model can also be considered.

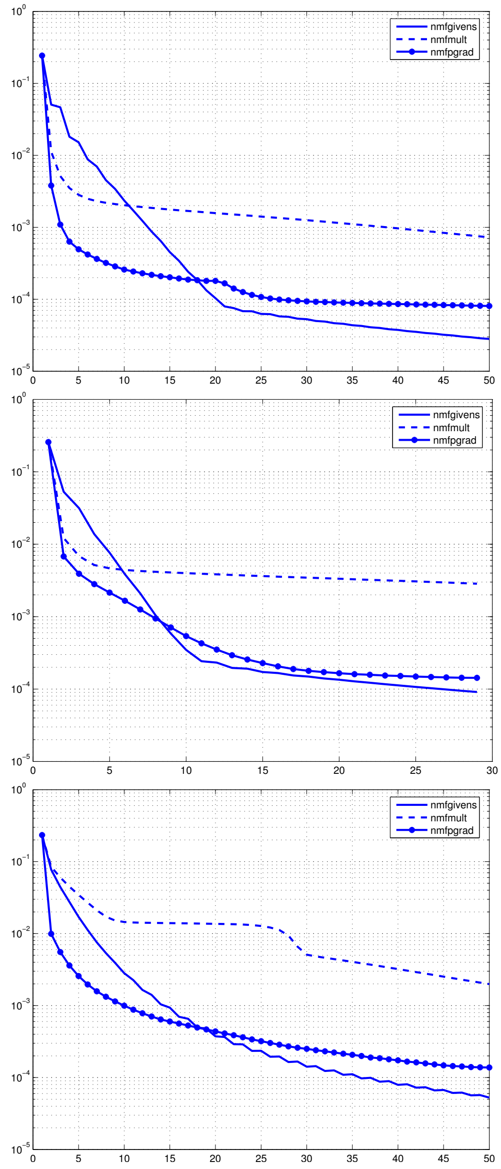

Figure 4: The performance index err_{re} versus iteration number for nmfgivens, nmfmult, nmfpgrad. The estimation of $\mathbf{H}$ in nmfgivens is done using projected gradient method.

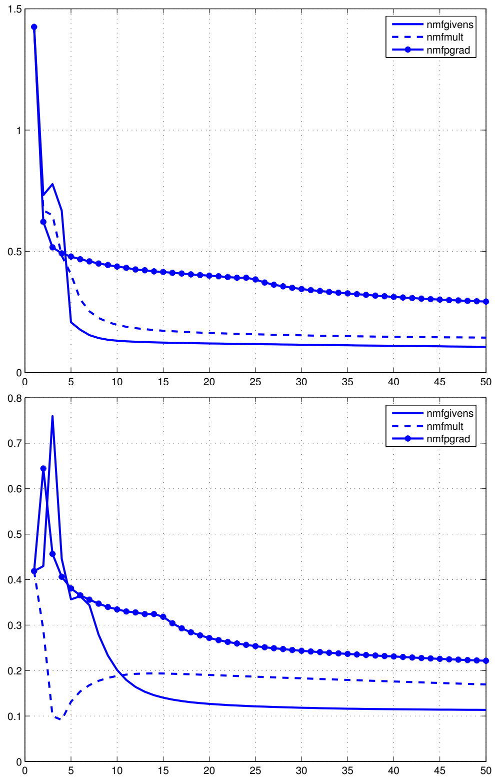

Figure 5: The performance index err_{s} versus iteration number for two different random initializations of θ between 0 and $\frac{\pi}{2}$ and $\mathbf{H}$ is initialized by [1] .

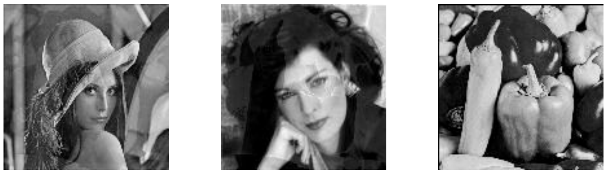

Figure 6: The 3 sources recovered using proposed nmfgivens algorithm.

APPENDIX

The derivative of vector $\mathbf{w}(\theta)$ of size $(N \times 1)$ parametrized by a vector $\theta = [\alpha_1, \ldots, \alpha_{N-1}]$ to each of the $N-1$ parameters reads:

$$\begin{aligned}\frac{\partial \mathbf{w}}{\partial \alpha_i} &= \mathbf{R}_1(\alpha_1)\mathbf{R}_2(\alpha_2)\ldots\mathbf{R}_{i-1}(\alpha_{i-1})\mathbf{R}'_i(\alpha_i)\mathbf{R}_{i+1}(\alpha_{i+1}) \\ &\quad \ldots\mathbf{R}_{N-1}(\alpha_{N-1})\mathbf{e_1}\end{aligned}$$

where

$$\mathbf{R}'_i(\alpha_i) = \begin{pmatrix} -s_i & & -c_i & \\ & \mathbf{0}_{i-1} & & \\ c_i & & -s_i & \\ & & & \mathbf{0}_{N-i-1} \end{pmatrix}$$

and $\mathbf{e_1} = [1, 0, \ldots, 0]^T$.

REFERENCES

[1] C. Boutsidis, E. Gallopoulos, "On SVD-based initialization for nonnegative matrix factorization", *Technical Report* HPCLAB-SCG-6/08-05, University of Patras, Patras, Greece, 2005.

[2] J.-F. Cardoso and A. Souloumiac, "Blind Beamforming for non Gaussian Signals", *IEEE Proceedings-F*, Vol. 40, pp. 362-370, 1993.

[3] C. Gobinet, E. Perrin et R. Huez, "Application of nonnegative matrix factorization to fluorescence spectroscopy", *European Signal Processing Conference (EUSIPCO'2004)*, Vienna, Austria, septembre 2004.

[4] E. Gonzalez, Y. Zhang, "Accelerating the Lee-Seung algorithm for nonnegative matrix factorization", *Tech. Rep*. TR-05-02, Rice University, March 2005.

[5] P. O. Hoyer, "Non-negative sparse coding", *Neural Networks for Signal Processing XII (Proc. IEEE Workshop on Neural Networks for Signal Processing)*, pp. 557-565, Martigny, Switzerland, 2002.

[6] P.O. Hoyer, "Nonnegative Matrix Factorization with Sparseness Constraints", *J. Machine Learning Research*, vol. 5, pp. 1457-1469, 2004.

[7] D.D. Lee and H.S. Seung, "Learning the Parts of Objects by Nonnegative Matrix Factorization", *Nature*, vol. 401, pp. 788-791,1999.

[8] J. S. Lee, D. D. Lee, S. Choi, and D. S. Lee, "Application of non-negative matrix factorization to dynamic positron emission tomography", *in Proceedings of the International Conference on Independent Component Analysis and Signal Separation (ICA2001)*, San Diego, California, pp. 629-632, december 2001.

[9] K. Levenberg, "A Method for the Solution of Certain Non-linear Problems in Least Squares", *Quarterly of Applied Mathematics*, 2(2), pp. 164-168, July 1944.

[10] C.-J. Lin, "Projected gradient methods for non-negative matrix factorization", *Technical Report*. Information and Support Services Technical Report ISSTECH-95-013, Department of Computer Science, National Taiwan University.

[11] D.W. Marquardt, "An Algorithm for the Least-Squares Estimation of Nonlinear Parameters", *SIAM Journal of Applied Mathematics*, 11(2), pp. 431-441, Jun 1963.

[12] E. Moreau, "A generalization of joint-diagonalization criteria for source separation", *IEEE Trans. Signal Processing*, Vol. 49, No. 3, pp. 530-541, March 2001.

[13] S. Moussaoui, D. Brie, A. Mohammad-Djafari, C. Carteret, " Separation of Non-Negative Mixture of Non-Negative Sources Using a Bayesian Approach and MCMC Sampling", *IEEE Transactions on Signal Processing*, vol. 54(11), pp. 4133-4145, 2006.

[14] P. Paatero and U. Tapper, "Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values", *Environmetrics*, vol. 5, pp. 111-126, 1994.

[15] Paul Sajda, Shuyan Du, Truman R. Brown, Radka Stoyanova, Dikoma C. Shungu, Xiangling Mao, Lucas C. Parra, " Nonnegative matrix factorization for rapid recovery of constituent spectra in magnetic resonance chemical shift imaging of the brain", *IEEE Trans. Med. Imaging*, vol. 23(12), pp. 1453-1465, 2004.

[16] P. Smaragdis and J.C. Brown, "Non-Negative Matrix Factorization for Polyphonic Music Transcription", *in Proceedings of WASPAA'03*, New Paltz, NY, pp. 177-180, october 2003.

[17] S. Wild, J. Curry, and A. Dougherty, "Improving nonnegative matrix factorizations through structured initialization", *Journal of Pattern Recognition*, vol. 37(11), pp. 2217-2232, 2004.