Hierarchical ALS Algorithms for Nonnegative Matrix and 3D Tensor Factorization

Andrzej Cichocki[1], Rafal Zdunek[2], and Shun-ichi Amari[3]

[1] Dept. of EE, Warsaw University of Technology, and IBS PAN Warsaw, Poland
[2] Institute of Telecommunications, Teleinformatics and Acoustics, Wroclaw University of Technology, Poland
[3] RIKEN Brain Science Institute, Wako-shi, Saitama, Japan
{cia,zdunek,amari}@brain.riken.jp

Abstract. In the paper we present new Alternating Least Squares (ALS) algorithms for Nonnegative Matrix Factorization (NMF) and their extensions to 3D Nonnegative Tensor Factorization (NTF) that are robust in the presence of noise and have many potential applications, including multi-way Blind Source Separation (BSS), multi-sensory or multi-dimensional data analysis, and nonnegative neural sparse coding. We propose to use local cost functions whose simultaneous or sequential (one by one) minimization leads to a very simple ALS algorithm which works under some sparsity constraints both for an under-determined (a system which has less sensors than sources) and over-determined model. The extensive experimental results confirm the validity and high performance of the developed algorithms, especially with usage of the multi-layer hierarchical NMF. Extension of the proposed algorithm to multidimensional Sparse Component Analysis and Smooth Component Analysis is also proposed.

1 Introduction - Problem Formulation

Nonnegative Matrix Factorization (NMF) and its multi-way extensions: Nonnegative Tensor Factorization (NTF) and Parallel Factor analysis (PARAFAC) models with sparsity and/or non-negativity constraints have been recently proposed as promising and quite efficient tools for processing sparse signals, images, or general data [1,2,3,4,5,6,7,8]. From a viewpoint of data analysis, NMF/NTF provides nonnegative and usually sparse common factors or hidden (latent) components with physiological meaning and interpretation [6,9]. NMF, NTF, and Sparse Component Analysis (SCA) are used in a variety of applications, ranging from neuroscience and psychometrics to chemometrics [10,1,6,7,9,11,12].

In this paper, we propose new Hierarchical Alternating Least Squares (HALS) algorithms for NMF/NTF. By incorporating the regularization and penalty terms into the local squared Euclidean norms, we are able to achieve sparse and local representations of the desired solution, and to alleviate the problem of getting stuck in local minima.

M.E. Davies et al. (Eds.): ICA 2007, LNCS 4666, pp. 169–176, 2007.

We impose nonnegativity and sparsity constraints to the following NTF (i.e., standard PARAFAC with nonnegativity constraints) model [3]:

$$\boldsymbol{X}_q = \boldsymbol{A}\boldsymbol{D}_q\tilde{\boldsymbol{S}} + \boldsymbol{E}_q, \qquad (q = 1, 2, \ldots, Q) \tag{1}$$

where $\boldsymbol{X}_q \in \mathbb{R}_+^{I\times T}$ are frontal slices (matrices) of the observed 3D tensor data or signals $\underline{\boldsymbol{X}} \in \mathbb{R}^{I\times T\times Q}$, $\boldsymbol{D}_q \in \mathbb{R}_+^{J\times J}$ are diagonal scaling matrices, $\boldsymbol{A} = [\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_J] \in \mathbb{R}_+^{I\times J}$ is a mixing or basis matrix, $\tilde{\boldsymbol{S}} \in \mathbb{R}_+^{J\times T}$ represents unknown sources or hidden (nonnegative and sparse) components, and $\boldsymbol{E}_q \in \mathbb{R}^{I\times T}$ represents the q-th frontal slice of the tensor $\underline{\boldsymbol{E}} \in \mathbb{R}^{I\times T\times Q}$ representing a noise or error. In the special case for $Q = 1$, the model simplifies to the standard NMF model. The objective is to estimate the set of all nonnegative matrices: $\boldsymbol{A}$, $\{\boldsymbol{D}_q\}$, $\tilde{\boldsymbol{S}}$[1]. The problem can be converted to a tri-NMF model by applying averaging of frontal slices: In this section, we develop the alternative algorithm which converts the problem to a simple tri-NMF model (under condition that all frontal slices $\boldsymbol{X}_q$ have the same dimension):

$$\boldsymbol{X} = \boldsymbol{A}\boldsymbol{D}\tilde{\boldsymbol{S}} + \boldsymbol{E} = \boldsymbol{A}\boldsymbol{S} + \boldsymbol{E}, \tag{2}$$

where $\boldsymbol{X} = \sum_{q=1}^{Q} \boldsymbol{X}_q$, $\boldsymbol{D} = \sum_{q=1}^{Q} \boldsymbol{D}_q = \mathrm{diag}\{d_{q1}, d_{q2}, \ldots, d_{qJ}\}$, $\boldsymbol{E} = \sum_{q=1}^{Q} \boldsymbol{E}_q$, and $\boldsymbol{S} = \boldsymbol{D}\tilde{\boldsymbol{S}}$ is a scaled matrix of sources. The above system of linear algebraic equations can be represented in an equivalent scalar form as follows $x_{it} = \sum_j a_{ij}s_{jt} + e_{it}$, or equivalently in the vector form: $\boldsymbol{X} = \sum_j \boldsymbol{a}_j\, \underline{\boldsymbol{s}}_j + \boldsymbol{E}$ where $\underline{\boldsymbol{s}}_j$ are rows of $\boldsymbol{S}$, and $\boldsymbol{a}_j$ are columns of $\boldsymbol{A}$ $(j = 1, 2, \ldots, J)$. Such a simple model provides improved performance if the noise (in the frontal slices) is not correlated.

The majority of NMF/NTF algorithms for BSS applications works only if the following assumption $T >> I \geq J$ is held, where J is known or can be estimated using SVD. In the paper, we propose the NMF algorithm that can work also for an under-determined case, i.e. $T >> J > I$, if signal representations are enough sparse. Our objective is to estimate the mixing (basis) matrix $\boldsymbol{A}$ and the sources $\boldsymbol{S}$, subject to nonnegativity and sparsity constraints.

2 Locally Regularized ALS Algorithm

The most of known and used adaptive algorithms for NMF are based on alternating minimization of the squared Euclidean distance expressed by the Frobenius norm:

$$D_F(\boldsymbol{X}||\boldsymbol{A}\boldsymbol{S}) = \frac{1}{2}\|\boldsymbol{X} - \boldsymbol{A}\boldsymbol{S}\|_F^2 + \alpha_A||\boldsymbol{A}||_1 + \alpha_S||\boldsymbol{S}||_1, \tag{3}$$

subject to nonnegativity constraints of all the elements in $\boldsymbol{A}$ and $\boldsymbol{S}$, where $||\boldsymbol{A}||_1 = \sum_{ir} a_{ir}$, $||\boldsymbol{S}||_1 = \sum_{jt} s_{jt}$, and α_A and α_S are nonnegative regularization coefficients controlling sparsity of the matrices [9].

[1] Usually, the common factors, i.e., matrices $\boldsymbol{A}$ and $\tilde{\boldsymbol{S}}$ are normalized to unit length column vectors and rows, respectively, and are forced to be as sparse as possible.

The basic approach to NMF is alternating minimization or alternating projection: the specified cost function is alternately minimized with respect to two sets of the parameters $\{s_{jt}\}$ and $\{a_{ij}\}$, each time optimizing one set of arguments while keeping the other one fixed [9,1].

In this paper we consider minimization of the set of local squared Euclidean cost functions:

$$D_F^{(j)}(\boldsymbol{X}^{(j)}||\boldsymbol{a}_j\underline{\boldsymbol{s}}_j) = \frac{1}{2}\|(\boldsymbol{X}^{(j)} - \boldsymbol{a}_j\underline{\boldsymbol{s}}_j)\|_F^2 + \alpha_A^{(j)}\ J_A(\boldsymbol{a}_j) + \alpha_S^{(j)}\ J_S(\underline{\boldsymbol{s}}_j), \quad (4)$$

for $j = 1, 2, \ldots, J$, subject to nonnegativity constraints for all elements: $a_{ij} \geq 0$ and $s_{jt} \geq 0$, where

$$\boldsymbol{X}^{(j)} = \boldsymbol{X} - \sum_{p \neq j} \boldsymbol{a}_p\underline{\boldsymbol{s}}_p, \quad (5)$$

$\boldsymbol{a}_j \in \mathbb{R}^{I\times 1}$ are columns of the basis mixing matrix $\boldsymbol{A}$, $\underline{\boldsymbol{s}}_j \in \mathbb{R}^{1\times T}$ are rows of $\boldsymbol{S}$, $\alpha_A^{(j)} \geq 0$ and $\alpha_S^{(j)} \geq 0$ are local parameters controlling a sparsity level of the individual vectors, and the penalty terms $J_A(\boldsymbol{a}_j) = \sum_i a_{ij}$ and $J_S(\underline{\boldsymbol{s}}_j) = \sum_t s_{jt}$ enforce sparsification of the columns in $\boldsymbol{A}$ and the rows in $\boldsymbol{S}$, respectively. The construction of such a set of local cost functions follows from the simple observation that the observed data can be decomposed approximately as follows $\boldsymbol{X} = \sum_{p=1}^{J} \boldsymbol{a}_p\underline{\boldsymbol{s}}_p + \boldsymbol{E}$ or more generally $\boldsymbol{X} = \sum_{p=1}^{J} \lambda_p\boldsymbol{a}_p\underline{\boldsymbol{s}}_p + \boldsymbol{E}$ with $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_J > 0$.

The gradients of the cost function (4) with respect to the unknown vectors $\boldsymbol{a}_j$ and $\underline{\boldsymbol{s}}_j$ are expressed by

$$\frac{\partial D_F^{(j)}(\boldsymbol{X}^{(j)}||\boldsymbol{a}_j\underline{\boldsymbol{s}}_j)}{\partial \underline{\boldsymbol{s}}_j} = \boldsymbol{a}_j^T\boldsymbol{a}_j\underline{\boldsymbol{s}}_j - \boldsymbol{a}_j^T\boldsymbol{X}^{(j)} + \alpha_S^{(j)}, \quad (6)$$

$$\frac{\partial D_F^{(j)}(\boldsymbol{X}^{(j)}||\boldsymbol{a}_j\underline{\boldsymbol{s}}_j)}{\partial \boldsymbol{a}_j} = \boldsymbol{a}_j\underline{\boldsymbol{s}}_j\underline{\boldsymbol{s}}_j^T - \boldsymbol{X}^{(j)}\underline{\boldsymbol{s}}_j^T + \alpha_A^{(j)}, \quad (7)$$

where the scalars $\alpha_S^{(j)}$ and $\alpha_A^{(j)}$ are added/substracted component-wise. By equating the gradient components to zero and assuming that we enforce the nonnegativity constraints with a simple "half-rectifying" nonlinear projection, we obtain a new set of sequential learning rules:

$$\underline{\boldsymbol{s}}_j \leftarrow \left[\frac{1}{\boldsymbol{a}_j^T\boldsymbol{a}_j}(\boldsymbol{a}_j^T\boldsymbol{X}^{(j)} - \alpha_S^{(j)})\right]_+ \qquad \boldsymbol{a}_j \leftarrow \left[\frac{1}{\underline{\boldsymbol{s}}_j\underline{\boldsymbol{s}}_j^T}(\boldsymbol{X}^{(j)}\underline{\boldsymbol{s}}_j^T - \alpha_A^{(j)})\right]_+, \quad (8)$$

for $j = 1, 2, \ldots, J$, where $[\xi]_+ = \max\{\epsilon, \xi\}$, and ϵ is a small constant to avoid numerical instabilities (usually $\epsilon = 10^{-16}$).

Remark 1. In practice, it is necessary to normalize in each iteration step the column vectors $\boldsymbol{a}_j$ and the row vectors $\underline{\boldsymbol{s}}_j$ to unit length vectors (in the sense of norm l_p norm ($p = 1, 2, ..., \infty$)). In the special case of l_2 norms the above

algorithms can be further simplified by neglecting the denominator in (8). After estimating the normalized matrices $\boldsymbol{A}$ and $\widetilde{\boldsymbol{S}}$, we estimate the diagonal matrices as follows:

$$\boldsymbol{D}_q = \left[\mathrm{diag}\{\boldsymbol{A}^+ \boldsymbol{X}_q \tilde{\boldsymbol{S}}^+\}\right]_+, \qquad (q = 1, 2, \ldots, Q) \tag{9}$$

Remark 2. In this paper we have applied a simple nonlinear half-wave rectifying projection $[s_{jt}]_+ = \max\{\epsilon, s_{jt}\}$, $\forall t$ (element-wise) in order to impose nonnegativity constraints. However, other nonlinear projections or filtering can be applied to extract sources (not necessary nonnegative) with specific properties. First of all, the proposed method can be easily extended for semi-NMF and semi-NTF, where nonnegativity constraints are imposed only for some preselected sources, i.e, rows of the matrix $\boldsymbol{S}$ and/or some selected columns of the matrix $\boldsymbol{A}$ if some *a priori* information is available. Furthermore, instead of using the simple nonlinear half-rectifying projection, we can apply more complex nonlinear projections and filtering to estimate bipolar sources which have some specific properties, for example, sources can be bounded, sparse or smooth. In order to estimate the sparse bipolar sources, we can apply well-known adaptive (soft or hard) shrinking nonlinear transformations (e.g, the nonlinear projection can be defined as: $P_{sr}(s_{jt}) = s_{jt}$ for $|s_{jt}| > \delta$ and $P_{sr}(s_{jt}) = 0$ otherwise, with the adaptive threshold $\delta > 0$). Alternatively, we may apply a power nonlinear element-wise transformation: $P_{sp}(s_{jt}) = \mathrm{sign}(s_{jt})|s_{jt}|^{1+\gamma_s}$, $\forall t$, where γ_s is a small coefficient which controls a sparsity/density level of individual sources [11]. In order to achieve smoothness of the estimated sources, we may apply a local averaging operator (such as MA or ARMA models) or low pass filtering which gradually enforces some level of smoothness during an iterative process.

3 Possible Extensions and Improvements

To deal with the factorization problem (1) efficiently, we adopt several approaches from constrained optimization and multi-criteria optimization, where we minimize simultaneously several cost functions using alternating switching between the sets of parameters: $\{\boldsymbol{A}\}$, $\{\boldsymbol{S}\}$.

The above simple algorithm can be further extended or improved (in respect of convergence rate and performance). First of all, different cost functions can be used for estimation of the rows in the matrix $\boldsymbol{S}$ and the columns in the matrix $\boldsymbol{A}$. Furthermore, the columns of $\boldsymbol{A}$ can be estimated simultaneously, instead one by one. For example, by minimizing the set of cost functions in (4) with respect to $\underline{\boldsymbol{s}}_j$, and simultaneously the cost function (3) with normalization of the columns $\boldsymbol{a}_j$ to unit l_2-norm, we obtain the new ALS learning algorithm in which the rows of $\boldsymbol{S}$ are updated locally (row by row) and the matrix $\boldsymbol{A}$ is updated globally (all columns $\boldsymbol{a}_j$ simultaneously):

$$\underline{\boldsymbol{s}}_j \leftarrow \left[\boldsymbol{a}_j^T \boldsymbol{X}^{(j)} - \alpha_S^{(j)}\right]_+, \quad (j = 1, \ldots, J), \quad \boldsymbol{A} \leftarrow \left[(\boldsymbol{X}\boldsymbol{S}^T - \alpha_A)(\boldsymbol{S}\boldsymbol{S}^T)^{-1}\right]_+ \tag{10}$$

with normalization (scaling) of the columns in $\boldsymbol{A}$ to the unit length.

Secondly, instead of the standard gradient descent approach we can apply the Quasi-Newton method [13,14] for estimation of matrix $\boldsymbol{A}$. Since the Hessian $\nabla_A^2(D_F) = \boldsymbol{I}_I \otimes \boldsymbol{S}\boldsymbol{S}^T \in \mathbb{R}^{IJ \times IJ}$ of $D_F(\boldsymbol{X}||\boldsymbol{A}\boldsymbol{S})$ has the diagonal block structure with the same blocks, we can simplify the update of $\boldsymbol{A}$ with the Newton method to the very simple form:

$$\boldsymbol{A} \leftarrow \left[\boldsymbol{A} - \nabla_A(D_F(\boldsymbol{X}||\boldsymbol{A}\boldsymbol{S}))\boldsymbol{H}_A^{-1}\right]_+, \tag{11}$$

where $\nabla_A D_F(\boldsymbol{X}||\boldsymbol{A}\boldsymbol{S}) = (\boldsymbol{A}\boldsymbol{S} - \boldsymbol{X})\boldsymbol{S}^T \in \mathbb{R}^{I \times J}$, and $\boldsymbol{H}_A = \boldsymbol{S}\boldsymbol{S}^T \in \mathbb{R}^{J \times J}$. The matrix $\boldsymbol{H}_A$ may be ill-conditioned, especially if $\boldsymbol{S}$ is sparse, and due to this the Levenberg-Marquardt approach is used to control ill-conditioning of the Hessian. Thus we have developed the following NMF/NTF algorithm:

$$\underline{\boldsymbol{s}}_j \leftarrow \left[\boldsymbol{a}_j^T \boldsymbol{X}^{(j)} - \alpha_S^{(j)}\right]_+, \quad \boldsymbol{A} \leftarrow \left[\boldsymbol{A} - (\boldsymbol{A}\boldsymbol{S} - \boldsymbol{X})\boldsymbol{S}^T(\boldsymbol{S}\boldsymbol{S}^T + \lambda \boldsymbol{I}_J)^{-1}\right]_+, \tag{12}$$

for $j = 1, \ldots, J$, where $\lambda \leftarrow \lambda_0 \exp\{-\tau k\}$, k is an index of a current alternating step, and $\boldsymbol{I}_J \in \mathbb{R}^{J \times J}$ is an identity matrix.

Since the alternating minimization technique in NMF is not convex, the selection of initial conditions is very important. Our algorithms are initialized with random uniform matrices. Thus, to minimize the risk of getting trapped in local minima of the cost functions, we use some steering technique that comes from a simulated annealing approach. The solution is triggered with the exponential rule. For our problems, we set heuristically $\lambda_0 = 100$ and $\tau = 0.02$.

3.1 Multi-layer NMF/NTF

In order to improve the performance of the NTF algorithms proposed in this paper, especially for ill-conditioned and badly scaled data and also to reduce risk of getting stuck in local minima in non-convex alternating minimization computations, we have developed a simple hierarchical multi-stage procedure [15] combined together with multi-start initializations, in which we perform a sequential decomposition of nonnegative matrices as follows. In the first step, we perform the basic decomposition (factorization) $\boldsymbol{X}_q \approx \boldsymbol{A}^{(1)}\boldsymbol{D}_q^{(1)}\boldsymbol{S}^{(1)}$ using any available NTF algorithm. In the second stage, the results obtained from the first stage are used to build up a new tensor $\underline{\widehat{\boldsymbol{S}}_1}$ from the estimated frontal slices defined as $\widehat{\boldsymbol{X}}_q^{(1)} = \boldsymbol{S}_q^{(1)} = \boldsymbol{D}_q^{(1)}\boldsymbol{S}^{(1)}$, $(q = 1, 2, \ldots, Q)$ and in the next step we perform the similar decomposition for the new available frontal slices: $\widehat{\boldsymbol{X}}_q^{(1)} = \boldsymbol{S}_q^{(1)} \approx \boldsymbol{A}^{(2)}\boldsymbol{D}_q^{(2)}\boldsymbol{S}^{(2)}$, using the same or different update rules. We continue our decomposition taking into account only the last achieved components. The process can be repeated arbitrarily many times until some stopping criteria are satisfied. In each step, we usually obtain gradual improvements of the performance. Thus, our NTF model has the form:

$$\boldsymbol{X}_q \approx \boldsymbol{A}^{(1)}\boldsymbol{A}^{(2)} \cdots \boldsymbol{A}^{(L)}\boldsymbol{D}_q^{(L)}\boldsymbol{S}^{(L)}, \qquad (q = 1, 2, \ldots, Q), \tag{13}$$

with final results $\boldsymbol{A} = \boldsymbol{A}^{(1)}\boldsymbol{A}^{(2)} \cdots \boldsymbol{A}^{(L)}$, $\boldsymbol{S} = \boldsymbol{S}^{(L)}$ and $\boldsymbol{D}_q = \boldsymbol{D}_q^{(L)}$.

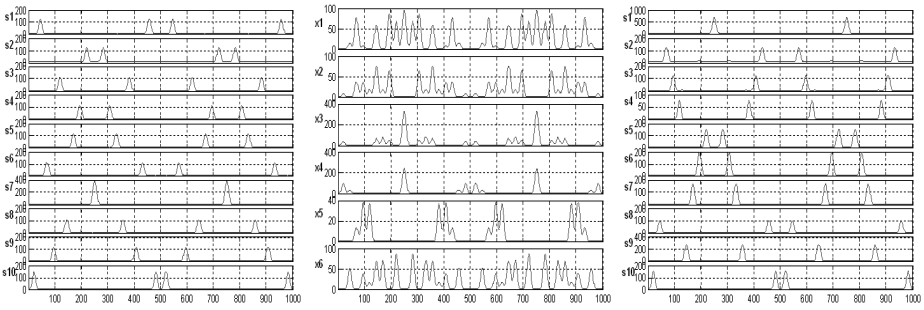

Fig. 1. (left) Original 10 sparse source signals ; (middle) observed 6 mixed signals with randomly generated mixing matrix $\boldsymbol{A} \in \mathbb{R}^{6\times 10}$ (under-determined case); (right) estimated 10 source signals using our new algorithm (12); For 10 layers we achieved the following performance: SIRs for $\boldsymbol{A}$ and $\boldsymbol{S}$ are as follows: $SIR_A = 38.1, 37.0, 35.9, 32.4, 28.2, 33.1, 34.5, 41.2, 25.1, 25.1$[dB] and $SIR_S = 23.1, 32.7, 35.2, 26.2, 29.8, 22.5, 41.8, 29.3, 30.2, 32.5$[dB], respectively

Physically, this means that we build up a system that has many layers or cascade connections of L mixing subsystems. The key point in our approach is that the learning (update) process to find parameters of matrices $\boldsymbol{S}^{(l)}$ and $\boldsymbol{A}^{(l)}$ is performed sequentially, i.e. layer by layer. In fact, we found that the hierarchical multi-layer approach is necessary to apply in order to achieve high performance for all the proposed algorithms.

4 Simulation Results

All the NTF algorithms presented in this paper have been extensively tested for many difficult benchmarks for signals and images with various statistical distributions and additive noise, and also for preliminary tests with real EEG data. Due to space limitations we present here only the selected simulations results in Figs.1–2. The synthetic benchmark illustrated in Fig.1(left) contains sparse non-negative and weakly statistically dependent 10 source components. The sources have been mixed by the randomly generated full rank matrix $\boldsymbol{A} \in \mathbb{R}_+^{6\times 10}$. The typical mixed signals are shown in Fig.1(middle). The results obtained with the new algorithm (12) with $\alpha_S^{(j)} = 0.05$ are illustrated in Fig.1(right) with average Signal-to-Interference (SIR) level greater than 25 [dB].

Since the proposed algorithms (alternating techniques) perform a non-convex optimization, the estimated components are initial condition dependent. To estimate the performance in a statistical sense, we present the histograms of 100 mean-SIR samples for estimation of $\boldsymbol{S}$ (Fig.2). We tested the two different algorithms (combination of the algorithms) – algorithm (10): ALS for $\boldsymbol{A}$ and HALS for $\boldsymbol{X}$ ($\alpha_A = 0,\ \alpha_S^{(j)} = 0.05$), and algorithm (12): quasi-Newton for $\boldsymbol{A}$ and HALS for $\boldsymbol{S}$.

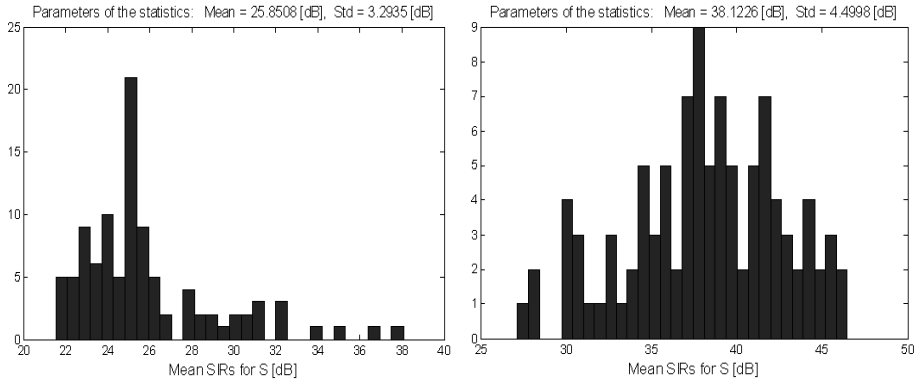

Fig. 2. Histograms of 100 mean-SIR samples for estimating $\boldsymbol{S}$ from Monte Carlo analysis performed using the following algorithms with 10 layers: (left) ALS for $\boldsymbol{A}$ and HALS for $\boldsymbol{S}$ (10); (right) quasi-Newton for $\boldsymbol{A}$ and HALS for $\boldsymbol{S}$ (12)

5 Conclusions and Discussion

The main objective and motivation of this paper is to derive simple algorithms which are suitable both for under-determined and over-determined cases. We have proposed the generalized and flexible cost function (controlled by sparsity penalty) that allows us to derive a family of robust and efficient alternating least squares algorithms for NMF and NTF. Exploiting gradient and Hessian properties, we have derived a family of efficient algorithms for estimating nonnegative sources even if the number of sensors is smaller than the number of hidden nonnegative sources under assumption that the sources are sufficiently sparse and not strongly overlapped. This is the unique modification of the standard ALS algorithm, and to the authors' best knowledge, the first time such a cost function and algorithms have been applied to NMF and NTF. The proposed algorithm gives also better performance (SIRs ans speed) than the ordinary ALS algorithm for NMF, and also some applications of the FOCUSS algorithm [16,17]. We implemented the discussed algorithms in our NMFLAB/NTFLAB MATLAB Toolboxes [18]. The algorithms may be also promising for other applications, such as Sparse Component Analysis, Smooth Component Analysis and EM Factor Analysis because they relax the problems of getting stuck to in local minima much better than the standard ALS algorithm.

We have motivated the use of the proposed models in three areas of data analysis (especially, EEG and fMRI) and signal/image processing: (i) multi-way blind source separation, (ii) model reduction and selection, and (iii) sparse image coding. Our preliminary experiments are promising. The models can be further extended by imposing additional, natural constraints such as smoothness, continuity, closure, unimodality, local rank - selectivity, and/or by taking into account a prior knowledge about specific 3D, or more generally, multi-way data.

Obviously, there are many challenging open issues remaining, such as global convergence, an optimal choice of the associated parameters.

References

1. Cichocki, A., Amari, S.: Adaptive Blind Signal And Image Processing (New revised and improved edition). John Wiley, New York (2003)
2. Dhillon, I., Sra, S.: Generalized nonnegative matrix approximations with Bregman divergences. In: Neural Information Proc. Systems, Vancouver, Canada (2005)
3. Hazan, T., Polak, S., Shashua, A.: Sparse image coding using a 3D non-negative tensor factorization. In: International Conference of Computer Vision (ICCV), pp. 50–57 (2005)
4. Heiler, M., Schnoerr, C.: Controlling sparseness in non-negative tensor factorization. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006. LNCS, vol. 3951, pp. 56–67. Springer, Heidelberg (2006)
5. Hoyer, P.: Non-negative matrix factorization with sparseness constraints. Journal of Machine Learning Research 5, 1457–1469 (2004)
6. Morup, M., Hansen, L.K., Herrmann, C.S., Parnas, J., Arnfred, S.M.: Parallel factor analysis as an exploratory tool for wavelet transformed event-related EEG. NeuroImage 29, 938–947 (2006)
7. Smilde, A., Bro, R., Geladi, P.: Multi-way Analysis: Applications in the Chemical Sciences. John Wiley and Sons, New York (2004)
8. Oja, E., Plumbley, M.D.: Blind separation of positive sources by globally convergent gradient search. Neural Computation 16, 1811–1825 (2004)
9. Lee, D.D., Seung, H.S.: Learning the parts of objects by nonnegative matrix factorization. Nature 401, 788–791 (1999)
10. Berry, M., Browne, M., Langville, A., Pauca, P., Plemmons, R.: Algorithms and applications for approximate nonnegative matrix factorization. Computational Statistics and Data Analysis (in press, 2006)
11. Cichocki, A., Amari, S., Zdunek, R., Kompass, R., Hori, G., He, Z.: Extended SMART algorithms for non-negative matrix factorization. In: Rutkowski, L., Tadeusiewicz, R., Zadeh, L.A., Zurada, J.M. (eds.) ICAISC 2006. LNCS (LNAI), vol. 4029, pp. 548–562. Springer, Heidelberg (2006)
12. Kim, M., Choi, S.: Monaural music source separation: Nonnegativity, sparseness, and shift-invariance. In: Rosca, J., Erdogmus, D., Príncipe, J.C., Haykin, S. (eds.) ICA 2006. LNCS, vol. 3889, pp. 617–624. Springer, Heidelberg (2006)
13. Zdunek, R., Cichocki, A.: Non-negative matrix factorization with quasi-Newton optimization. In: Rutkowski, L., Tadeusiewicz, R., Zadeh, L.A., Zurada, J.M. (eds.) ICAISC 2006. LNCS (LNAI), vol. 4029, pp. 870–879. Springer, Heidelberg (2006)
14. Zdunek, R., Cichocki, A.: Nonnegative matrix factorization with constrained second-order optimization. Signal Processing 87, 1904–1916 (2007)
15. Cichocki, A., Zdunek, R.: Multilayer nonnegative matrix factorization. Electronics Letters 42, 947–948 (2006)
16. Murray, J.F., Kreutz-Delgado, K.: Learning sparse overcomplete codes for images. Journal of VLSI Signal Processing 45, 97–110 (2006)
17. Kreutz-Delgado, K., Murray, J.F., Rao, B.D., Engan, K., Lee, T.W., Sejnowski, T.J.: Dictionary learning algorithms for sparse representation. Neural Computation 15, 349–396 (2003)
18. Cichocki, A., Zdunek, R.: NTFLAB for Signal Processing. Technical report, Laboratory for Advanced Brain Signal Processing, BSI, RIKEN, Saitama, Japan (2006)