

# Csiszár's Divergences for Non-Negative Matrix Factorization: Family of New Algorithms

Andrzej CICHOCKI<sup>1\*</sup>, Rafal ZDUNEK<sup>1\*\*</sup> and Shun-ichi AMARI<sup>2</sup>

<sup>1</sup> Laboratory for Advanced Brain Signal Processing

<sup>2</sup> Laboratory for Mathematical Neuroscience,  
RIKEN BSI, Wako-shi JAPAN

**Abstract.** In this paper we discuss a wide class of loss (cost) functions for non-negative matrix factorization (NMF) and derive several novel algorithms with improved efficiency and robustness to noise and outliers. We review several approaches which allow us to obtain generalized forms of multiplicative NMF algorithms and unify some existing algorithms. We give also the flexible and relaxed form of the NMF algorithms to increase convergence speed and impose some desired constraints such as sparsity and smoothness of components. Moreover, the effects of various regularization terms and constraints are clearly shown. The scope of these results is vast since the proposed generalized divergence functions include quite a large number of useful loss functions such as the squared Euclidean distance, Kulback-Leibler divergence, Itakura-Saito, Hellinger, Pearson's chi-square, and Neyman's chi-square distances, etc. We have applied successfully the developed algorithms to blind (or semi blind) source separation (BSS) where sources can be generally statistically dependent, however they satisfy some other conditions or additional constraints such as nonnegativity, sparsity and/or smoothness.

## 1 Introduction and Problem Formulation

The non-negative matrix factorization (NMF) approach is promising in many applications from engineering to neuroscience since it is designed to capture alternative structures inherent in the data and, possibly to provide more biological insight [1–6]. Lee and Seung introduced NMF in its modern formulation as a method to decompose patterns or images [3, 7].

In this paper we impose nonnegativity constraints and other penalties such as sparseness and/or smoothness. The NMF decomposes the data matrix  $\mathbf{Y} = [\mathbf{y}(1), \mathbf{y}(2), \dots, \mathbf{y}(N)] \in \mathbb{R}^{m \times N}$  as a product of two matrices  $\mathbf{A} \in \mathbb{R}^{m \times n}$  and  $\mathbf{X} = [\mathbf{x}(1), \mathbf{x}(2), \dots, \mathbf{x}(N)] \in \mathbb{R}^{n \times N}$  having only non-negative elements. Although some decompositions or matrix factorizations provide an exact reconstruction of the data (i.e.,  $\mathbf{Y} = \mathbf{AX}$ ), we shall consider here decompositions which are approximative in nature, i.e.,

$$\mathbf{Y} = \mathbf{AX} + \mathbf{V}, \quad \mathbf{A} \geq 0, \quad \mathbf{X} \geq 0 \quad (1)$$

---

\* On leave from Warsaw University of Technology, Poland

\*\* On leave from Institute of Telecommunications, Teleinformatics and Acoustics, Wrocław University of Technology, Poland

or equivalently  $\mathbf{y}(k) = \mathbf{A}\mathbf{x}(k) + \mathbf{v}(k)$ ,  $k = 1, 2, \dots, N$  or in a scalar form as  $y_i(k) = \sum_{j=1}^n a_{ij}x_j(k) + \nu_i(k)$ ,  $i = 1, \dots, m$ , where  $\mathbf{V} \in \mathbb{R}^{m \times N}$  represents noise or error matrix,  $\mathbf{y}(k) = [y_1(k), \dots, y_m(k)]^T$  is a vector of the observed signals (typically nonnegative) at the discrete time instants  $k$  while  $\mathbf{x}(k) = [x_1(k), \dots, x_n(k)]^T$  is a vector of components or source signals at the same time instant [8]. Our objective is to estimate the mixing (basis) matrix  $\mathbf{A}$  and sources  $\mathbf{X}$  subject to nonnegativity constraints on all entries. Usually, in BSS applications it is assumed that  $N \gg m \geq n$  and  $n$  is known or can be relatively easily estimated using SVD or PCA. Throughout this paper, we use the following notations:  $x_j(k) = x_{jk}$ ,  $y_i(k) = y_{ik}$  and  $z_{ik} = [\mathbf{A}\mathbf{X}]_{ik}$  means  $ik$ -element of the matrix  $(\mathbf{A}\mathbf{X})$ , the  $ij$ -th element of the matrix  $\mathbf{A}$  is denoted by  $a_{ij}$ .

The basic approach to NMF is alternating minimization or alternating projection: the specified loss function is alternately minimized with respect to two sets of parameters  $\{x_{jk}\}$  and  $\{a_{ij}\}$ , each time optimizing one set of arguments while keeping the other one fixed [2, 3, 8].

The most popular adaptive multiplicative algorithms for NMF are based on two loss functions: 1. square Euclidean distance expressed by the Frobenius norm:

$$D_F(\mathbf{A}, \mathbf{X}) = \frac{1}{2} \|\mathbf{Y} - \mathbf{A}\mathbf{X}\|_F^2 = \frac{1}{2} \sum_{i=1}^m \sum_{k=1}^N |y_{ik} - [\mathbf{A}\mathbf{X}]_{ik}|^2$$

$$\text{s. t. } a_{ij} \geq 0, \quad x_j(k) = x_{jk} \geq 0 \quad \forall i, j, k, \quad (2)$$

which is optimal for a Gaussian distributed noise). Based on of this cost function Lee and Seung proposed the following multiplicative algorithm:

$$a_{ij} \leftarrow a_{ij} \frac{[\mathbf{Y}\mathbf{X}^T]_{ij}}{[\mathbf{A}\mathbf{X}\mathbf{X}^T]_{ij}}, \quad x_{jk} \leftarrow x_{jk} \frac{[\mathbf{A}^T\mathbf{Y}]_{jk}}{[\mathbf{A}^T\mathbf{A}\mathbf{X}]_{jk}}. \quad (3)$$

which is an extension of the well known ISRA (Image Space Reconstruction Algorithm) algorithm [9]. Alternative mostly used loss function that intrinsically ensures non-negativity constraints and it is related to the Poisson likelihood is a functional based on the Kullback-Leibler divergence [3, 5]:

$$D_{KL}(\mathbf{Y} \parallel [\mathbf{A}\mathbf{X}]) = \sum_{ik} \left( y_{ik} \log \frac{y_{ik}}{[\mathbf{A}\mathbf{X}]_{ik}} + [\mathbf{A}\mathbf{X}]_{ik} - y_{ik} \right) \quad (4)$$

$$\text{s. t. } x_{jk} \geq 0, \quad a_{ij} \geq 0, \quad \|\mathbf{a}_j\|_1 = \sum_{i=1}^m a_{ij} = 1.$$

Using the alternating minimization approach, Lee and Seung derived the following multiplicative learning rules:

$$x_{jk} \leftarrow x_{jk} \frac{\sum_{i=1}^m a_{ij} (y_{ik}/[\mathbf{A}\mathbf{X}]_{ik})}{\sum_{q=1}^m a_{jq}}, \quad a_{ij} \leftarrow a_{ij} \frac{\sum_{k=1}^N x_{jk} (y_{ik}/[\mathbf{A}\mathbf{X}]_{ik})}{\sum_{p=1}^N x_{jp}}, \quad (5)$$

which are extensions (by alternating minimization) of the well known EMLL or Richardson-Lucy algorithm (RLA) [9].

It should be noted that the most existing NMF algorithms perform blind source separation rather very poorly due to the non-uniqueness of solution and/or the lack of additional constraints which should be satisfied. The main objective of this contribution is to propose flexible and improved NMF algorithms that generalize or combine several different criteria in order to extract physically meaningful sources, especially for biomedical signal applications such as EEG and MEG.

## 2 Generalized Divergences for NMF

There are three large classes of generalized divergences which can be potentially useful for developing new flexible algorithms for NMF: the Bregman divergences, Amari's alpha divergence [1] and the Csiszár's  $\varphi$ -divergences [10]. In this contribution we limit our discussion to the Csiszár's divergences and as the special case the alpha divergence. The Csiszár's  $\varphi$ -divergence is defined as

$$D_C(\mathbf{z}||\mathbf{y}) = \sum_{k=1}^N z_k \varphi\left(\frac{y_k}{z_k}\right) \quad (6)$$

where  $y_k \geq 0, z_k \geq 0$  and  $\varphi : [0, \infty) \rightarrow (-\infty, \infty)$  is a function which is convex on  $(0, \infty)$  and continuous at zero. Depending on the application, we can impose different restrictions on  $\varphi$ . In order to use the Csiszár's divergence as a distance measure, we assume that  $\varphi(1) = 0$  and that it is strictly convex at 1.

Several basic examples include ( $u_{ik} = y_{ik}/z_{ik}$ ):

1. If  $\varphi(u) = (\sqrt{u} - 1)^2$ , then  $D_{C-H} = \sum_{ik} (\sqrt{y_{ik}} - \sqrt{z_{ik}})^2$  -Hellinger distance;
2. If  $\varphi(u) = (u - 1)^2$ , then  $D_{C-P} = \sum_{ik} (y_{ik} - z_{ik})^2 / z_{ik}$  -Pearson's distance;
3. For  $\varphi(u) = u(u^{\beta-1} - 1) / (\beta^2 - \beta) + (1 - u) / \beta$  we have a family of Amari's alpha divergences:

$$D_A^{(\beta)}(\mathbf{A}\mathbf{X}||\mathbf{Y}) = \sum_{ik} y_{ik} \frac{(y_{ik}/z_{ik})^{\beta-1} - 1}{\beta(\beta-1)} + \frac{z_{ik} - y_{ik}}{\beta}, \quad z_{ik} = [\mathbf{A}\mathbf{X}]_{ik}, \quad (7)$$

where  $\beta = (1 + \alpha)/2$  [1] (see also Ali-Silvey, Liese & Vajda, Cressie-Read disparity, Eguchi beta divergence, Kompass) [11, 12]. It is interesting to note that in the special cases for  $\beta = 2, 0.5, -1$ , we obtain Pearson's, Hellinger and Neyman's chi-square distances, respectively (while for the cases  $\beta = 1$  and  $\beta = 0$  the divergences have to be defined as limiting cases as  $\beta \rightarrow 1$  and  $\beta \rightarrow 0$ , respectively). When these limits are evaluated one gets for  $\beta \rightarrow 1$  the generalized Kullback-Leibler divergence (called I-divergence) defined by equations (4) and for  $\beta \rightarrow 0$  the dual generalized KL divergence:

$$D_{KL}(\mathbf{A}\mathbf{X}||\mathbf{Y}) = \sum_{ik} \left( [\mathbf{A}\mathbf{X}]_{ik} \log \frac{[\mathbf{A}\mathbf{X}]_{ik}}{y_{ik}} - [\mathbf{A}\mathbf{X}]_{ik} + y_{ik} \right) \quad (8)$$

As an illustrative example, let us derive a new multiplicative learning rule for the loss function (8). By applying multiplicative exponentiated gradient (EG) descent updates:

$$x_{jk} \leftarrow x_{jk} \exp\left(-\eta_j \frac{\partial D_{KL}}{\partial x_{jk}}\right), \quad a_{ij} \leftarrow a_{ij} \exp\left(-\tilde{\eta}_j \frac{\partial D_{KL}}{\partial a_{ij}}\right), \quad (9)$$

we obtain new simple multiplicative learning rules for NMF

$$x_{jk} \leftarrow x_{jk} \exp\left(\sum_{i=1}^m \eta_j a_{ij} \log\left(\frac{y_{ik}}{[\mathbf{A}\mathbf{X}]_{ik}}\right)\right) = x_{jk} \prod_{i=1}^m \left(\frac{y_{ik}}{[\mathbf{A}\mathbf{X}]_{ik}}\right)^{\eta_j a_{ij}}, \quad (10)$$

$$a_{ij} \leftarrow a_{ij} \exp\left(\sum_{k=1}^N \tilde{\eta}_j x_{jk} \log\left(\frac{y_{ik}}{[\mathbf{A}\mathbf{X}]_{ik}}\right)\right) = a_{ij} \prod_{k=1}^N \left(\frac{y_{ik}}{[\mathbf{A}\mathbf{X}]_{ik}}\right)^{\tilde{\eta}_j x_{jk}}, \quad (11)$$

The nonnegative learning rates  $\eta_j, \tilde{\eta}_j$  can take different forms. Typically, in order to guarantee stability of the algorithm we assume that  $\eta_j = \omega (\sum_{i=1}^m a_{ij})^{-1}$ ,  $\tilde{\eta}_j = \omega (\sum_{k=1}^N x_{jk})^{-1}$ , where  $\omega \in (0, 2)$  is an over-relaxation parameter. The above algorithm can be considered as an alternating minimization/projection extension of the well known SMART (Simultaneous Multiplicative Algebraic Reconstruction Technique) [9].

Similarly, for  $\beta \neq 0$  we have developed the following new algorithm (the proof is omitted due to the lack of space)

$$x_{jk} \leftarrow x_{jk} \left(\sum_{i=1}^m a_{ij} (y_{ik}/[\mathbf{A}\mathbf{X}]_{ik})^\beta\right)^{1/\beta}, \quad a_{ij} \leftarrow a_{ij} \left(\sum_{k=1}^N (y_{ik}/[\mathbf{A}\mathbf{X}]_{ik})^\beta x_{jk}\right)^{1/\beta}$$

with normalization of columns of  $\mathbf{A}$  in each iteration to unit length:  $a_{ij} \leftarrow a_{ij} / \sum_p a_{pj}$ . The algorithm can be written in a compact matrix form using some MATLAB notations:

$$\begin{aligned} \mathbf{X} &\leftarrow \mathbf{X} .* \left(\mathbf{A}^T ((\mathbf{Y} + \varepsilon) ./ (\mathbf{A}\mathbf{X} + \varepsilon)).^\beta\right)^{1/\beta} \\ \mathbf{A} &\leftarrow \mathbf{A} .* \left(((\mathbf{Y} + \varepsilon) ./ (\mathbf{A}\mathbf{X} + \varepsilon)).^\beta \mathbf{X}^T\right)^{1/\beta}, \quad \mathbf{A} \leftarrow \mathbf{A} \text{diag}\{1./\text{sum}(\mathbf{A}, 1)\}, \end{aligned} \quad (12)$$

where in practice a small constant  $\varepsilon = 10^{-9}$  is introduced in order to ensure non-negativity constraints and avoid possible division by zero.

### 3 Modified Multiplicative NMF Algorithms with Regularization, Sparsity and/or Smoothing

Although the standard NMF (without any auxiliary constraints) provides sparseness of its components, we can achieve some control of this sparsity by imposing additional constraints in addition to non-negativity constraints. In fact, we can incorporate smoothness or sparsity constraints in several ways. As an illustrative

example, let us consider a modified alpha divergence with regularization terms (which is an extension of the generalized divergence proposed recently by Raul Kompass [12]):

$$D_{Ko}(\mathbf{Y}||\mathbf{AX}) = \sum_{ik} \left( y_{ik} \frac{y_{ik}^{\beta-1} - [\mathbf{AX}]_{ik}^{\beta-1}}{\beta(\beta-1)} + [\mathbf{AX}]_{ik}^{\beta-1} \frac{[\mathbf{AX}]_{ik} - y_{ik}}{\beta} \right) + \alpha_X f_X(\mathbf{X}) + \alpha_A f_A(\mathbf{A}), \quad (13)$$

where regularization parameters and terms  $f_A(\mathbf{A})$  and  $f_X(\mathbf{X})$  are introduced to enforce a certain application-dependent characteristic of solutions such as smoothness or sparsity. For example, in order to achieve sparse representation we usually choose  $f_X(x_j) = x_j$  with constraints  $x_j \geq 0$ .

It is interesting to note that such defined divergence for  $\alpha_X = \alpha_A = 0$  and  $\beta = 2$  simplifies to the Frobenius norm (2); for  $\beta \rightarrow 0$  it tends to Itakura -Saito distance, and for  $\beta \rightarrow 1$  reduces to the Kullback-Leibler divergence (4).

Applying the standard gradient descent to (13) we have

$$x_{jk} \leftarrow x_{jk} - \eta_{jk} \left( \sum_{i=1}^m a_{ij} \left( [\mathbf{AX}]_{ik}^{\beta-1} - y_{ik}/[\mathbf{AX}]_{ik}^{2-\beta} \right) - \alpha_X \psi_X(\mathbf{X}) \right) \quad (14)$$

$$a_{ij} \leftarrow a_{ij} - \eta_{ij} \left( \sum_{k=1}^N \left( [\mathbf{AX}]_{ik}^{\beta-1} - y_{ik}/[\mathbf{AX}]_{ik}^{2-\beta} \right) x_{jk} - \alpha_A \psi_A(\mathbf{A}) \right), \quad (15)$$

where the functions  $\psi_A(\mathbf{A})$  and  $\psi_X(\mathbf{X})$  are defined as

$$\psi_A(\mathbf{A}) = \frac{\partial f_A(\mathbf{A})}{\partial a_{ij}}, \quad \psi_X(\mathbf{X}) = \frac{\partial f_X(\mathbf{X})}{\partial x_{jk}}. \quad (16)$$

Similar to the Lee and Seung approach, by choosing suitable learning rates:

$$\eta_{jk} = \frac{x_{jk}}{\sum_{i=1}^m a_{ij} [\mathbf{AX}]_{ik}^{\beta-1}}, \quad \eta_{ij} = \frac{a_{ij}}{\sum_{k=1}^N [\mathbf{AX}]_{ik}^{\beta-1} x_{jk}}, \quad (17)$$

we obtain multiplicative update rules:

$$x_{jk} \leftarrow x_{jk} \frac{[\sum_{i=1}^m a_{ij} (y_{ik}/[\mathbf{AX}]_{ik}^{2-\beta}) - \alpha_X \psi_X(\mathbf{X})]_{\varepsilon}}{\sum_{i=1}^m a_{ij} [\mathbf{AX}]_{ik}^{\beta-1}}, \quad (18)$$

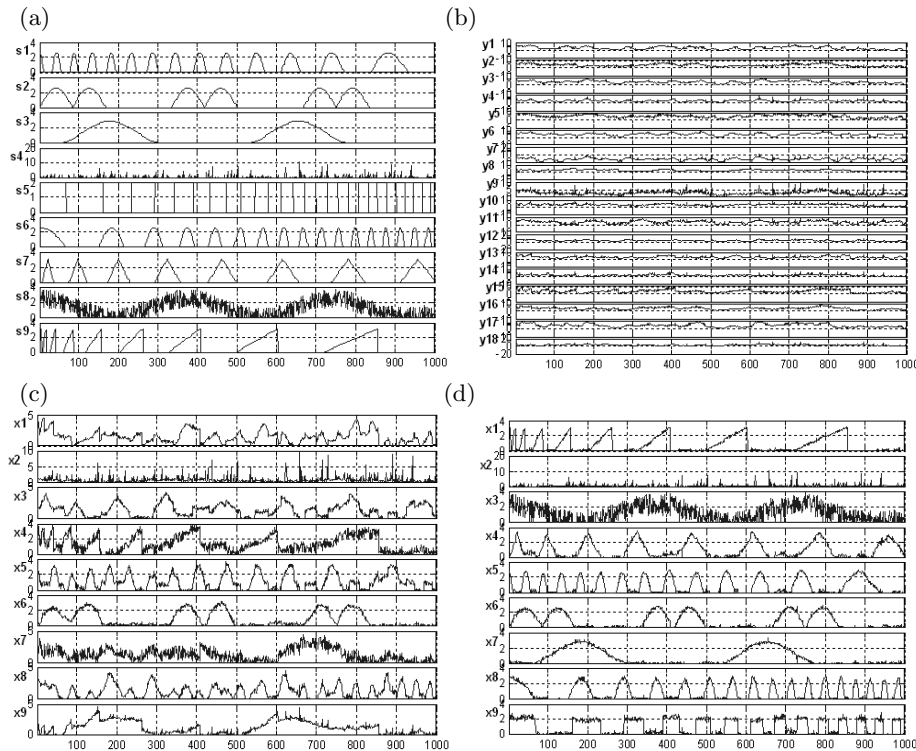
$$a_{ij} \leftarrow a_{ij} \frac{[\sum_{k=1}^N (y_{ik}/[\mathbf{AX}]_{ik}^{2-\beta}) x_{jk} - \alpha_A \psi_A(\mathbf{A})]_{\varepsilon}}{\sum_{k=1}^N [\mathbf{AX}]_{ik}^{\beta-1} x_{jk}}, \quad (19)$$

where the additional nonlinear operator is introduced in practice defined as  $[x]_{\varepsilon} = \max\{\varepsilon, x\}$  with a small  $\varepsilon$  in order to avoid zero and negative values.

Another simple approach which can be used for control of sparsification of estimated variables is to apply nonlinear projections via suitable nonlinear monotonic functions which increase or decrease the sparseness. In this paper we have applied a very simple nonlinear transformation  $x_{jk} \leftarrow (x_{jk})^{1+\alpha_{sX}}$ ,  $\forall k$ , where  $\alpha_{sX}$  is a small coefficient typically, from 0.001 to 0.005 and it is positive if we want to increase sparseness of an estimated component and negative if we want to decrease the sparseness (see Table 1 for practical implementations).

**Table 1.** New Multiplicative NMF algorithms with regularization and/or sparsity constraints

Minimization of loss function subject to $a_{ij} \geq 0$ and $x_{jk} \geq 0$	Iterative Learning Algorithm Relaxation parameter $\omega \in (0, 2)$
Alpha divergence, $\beta \neq 0, \beta \neq 1$	$x_{jk} \leftarrow \left( x_{jk} \left( \sum_{i=1}^m a_{ij} \left( \frac{y_{ik}}{[\mathbf{A}\mathbf{X}]_{ik}} \right)^\beta \right)^{\omega/\beta} \right)^{1+\alpha_{sX}}$
$\sum_{ik} \left\{ \frac{y_{ik}^\beta z_{ik}^{1-\beta} - \beta y_{ik} + (\beta-1)z_{ik}}{\beta(\beta-1)} \right\}$	$a_{ij} \leftarrow \left( a_{ij} \left( \sum_{k=1}^N x_{jk} \left( \frac{y_{ik}}{[\mathbf{A}\mathbf{X}]_{ik}} \right)^\beta \right)^{\omega/\beta} \right)^{1+\alpha_{sA}}$
Pearson and Hellinger distances	$a_{ij} \leftarrow a_{ij} / \sum_p a_{pj},$
$\sum_{ik} \frac{(y_{ik} - [\mathbf{A}\mathbf{X}]_{ik})^2}{[\mathbf{A}\mathbf{X}]_{ik}},$	$(\beta = 2)$
$\sum_{ik} \left( \sqrt{[\mathbf{A}\mathbf{X}]_{ik}} - \sqrt{y_{ik}} \right)^2,$	$(\beta = 0.5)$
Kulback-Leibler divergence	$x_{jk} \leftarrow \left( x_{jk} \left( \sum_{i=1}^m a_{ij} \frac{y_{ik}}{[\mathbf{A}\mathbf{X}]_{ik}} \right)^\omega \right)^{1+\alpha_{sX}}$
$\sum_{ik} (y_{ik} \log \frac{y_{ik}}{[\mathbf{A}\mathbf{X}]_{ik}} - y_{ik} + [\mathbf{A}\mathbf{X}]_{ik})$	$a_{ij} \leftarrow \left( a_{ij} \left( \sum_{k=1}^N x_{jk} \frac{y_{ik}}{[\mathbf{A}\mathbf{X}]_{ik}} \right)^\omega \right)^{1+\alpha_{sA}}$
$(\beta = 1)$	$a_{ij} \leftarrow a_{ij} / \left( \sum_p a_{pj} \right)$
K-L divergence (dual)	$x_{jk} \leftarrow \left( x_{jk} \prod_{i=1}^m \left( \frac{y_{ik}}{[\mathbf{A}\mathbf{X}]_{ik}} \right)^{\omega a_{ij}} \right)^{1+\alpha_{sX}}$
$\sum_{ik} ([\mathbf{A}\mathbf{X}]_{ik} \log \frac{[\mathbf{A}\mathbf{X}]_{ik}}{y_{ik}} + y_{ik} - [\mathbf{A}\mathbf{X}]_{ik})$	$a_{ij} \leftarrow \left( a_{ij} \prod_{k=1}^N \left( \frac{y_{ik}}{[\mathbf{A}\mathbf{X}]_{ik}} \right)^{\tilde{\eta}_j x_{jk}} \right)^{1+\alpha_{sA}}$
$(\beta = 0)$	$a_{ij} \leftarrow a_{ij} / \left( \sum_p a_{pj} \right), \tilde{\eta}_j = \omega \left( \sum_k x_{jk} \right)^{-1}$
Euclidean distance	$x_{jk} \leftarrow x_{jk} \frac{[[\mathbf{A}^T \mathbf{Y}]_{ik} - \alpha_X \psi_X(\mathbf{X})]_\varepsilon}{[\mathbf{A}^T \mathbf{A}\mathbf{X}]_{ik} + \varepsilon}$
$\ \mathbf{Y} - [\mathbf{A}\mathbf{X}]\ _F^2 + \alpha_X f_X(\mathbf{X}) + \alpha_A f_A(\mathbf{A})$	$a_{ij} \leftarrow a_{ij} \frac{[[\mathbf{Y}\mathbf{X}^T]_{ij} - \alpha_A \psi_A(\mathbf{A})]_\varepsilon}{[\mathbf{A}\mathbf{X}\mathbf{X}^T]_{ij} + \varepsilon}$
Kompass generalized divergence	$x_{jk} \leftarrow x_{jk} \frac{[\sum_{i=1}^m a_{ij} (y_{ik}/[\mathbf{A}\mathbf{X}]_{ik}^{2-\beta}) - \alpha_X \psi_X(\mathbf{X})]_\varepsilon}{\sum_{i=1}^m a_{ij} [\mathbf{A}\mathbf{X}]_{ik}^{\beta-1} + \varepsilon}$
$\sum_{ik} (y_{ik} \frac{y_{ik}^{\beta-1} - [\mathbf{A}\mathbf{X}]_{ik}^{\beta-1}}{\beta(\beta-1)} +$	$a_{ij} \leftarrow \left( a_{ij} \frac{[\sum_{k=1}^N x_{jk} (y_{ik}/[\mathbf{A}\mathbf{X}]_{ik}^{2-\beta})]_\varepsilon}{\sum_{k=1}^N x_{jk} [\mathbf{A}\mathbf{X}]_{ik}^{\beta-1} + \varepsilon} \right)^{1+\alpha_{sA}}$
$+ [\mathbf{A}\mathbf{X}]_{ik}^{\beta-1} \frac{[\mathbf{A}\mathbf{X}]_{ik} - y_{ik}}{\beta}) + \alpha_X f_X(\mathbf{X})$	$a_{ij} \leftarrow a_{ij} / \left( \sum_p a_{pj} \right), \beta \in [0, 2]$



**Fig. 1.** Example 1: (a) The original 9 source signals, (b) observed 18 mixed signals, (c) Estimated sources using standard Lee-Seung algorithm (5) (d) Estimated source signals using the new algorithm (12) for  $\beta = 2$  with nonlinear projection  $\alpha_{sX} = \alpha_{sA} = 0.002$  with SIR=32dB, 20dB, 19dB, 18dB, 23dB, 25dB, 27dB, 26dB, 19dB, for individual sources respectively.

## 4 Simulation Results

All the NMF algorithms discussed in this paper (see Table 1) have been extensively tested for many difficult benchmarks for sparse signals and images with various statistical distributions. The simulation results confirmed that the developed algorithms are stable, efficient and provide consistent results for a wide set of parameters. Due to the limit of space we give here only one illustrative example: Nine nonnegative sparse signals (some of them are statistically dependent) shown in Fig.1 (a) have been mixed by randomly generated nonnegative matrix  $A \in \mathbb{R}^{18 \times 9}$ . To the mixture we added a uniform distributed noise with SNR=20 dB. The mixed signals are shown in Fig.1 (b). Using the known standard NMF algorithm (5) we failed to estimate the original sources (see Fig.1 (c)). However, for the new algorithms we reconstructed successfully all the sources. Fig. 1 (d) illustrates the results obtained by using algorithm (12) with  $\beta = 2$  and the nonlinear projection with  $\alpha_{sX} = \alpha_{sA} = 0.002$  (see also Table 1). Similar

or even slightly better performance we achieved by applying the other proposed algorithms with regularization/projection (Table 1).

## 5 Conclusions and Discussion

In this paper we discuss loss functions which allow to derive us a very large class of flexible, robust and efficient NMF adaptive algorithms. The optimal choice of  $\beta$  depends on the distribution of data and *a priori* knowledge about noise. If such knowledge is not available, we may run NMF algorithms for various sets of parameters to find an optimal solution. For some tasks and distributions there are particular divergence measures that are uniquely suited. On the other hand, if the approximating model fits the true distribution well, then it does not matter which divergence measure is used, since all of them will give similar results. The discussed loss functions are jointly convex. This property is stronger than the individual convexity in  $\{y_{ik}\}$  and  $\{z_{ik}\}$ . However, it very difficult to prove the global convergence of the derived NMF algorithms. Our simulation experiments indicate that for  $m \gg n$ , typically  $m > 5n$  and  $N = 10^3 \sim 10^4$ , we usually avoid sticking in poor local minima. We found by extensive simulations that regularization/projections techniques play a key role in improving the performance of blind source separation by using the NMF approach.

## References

1. Amari, S.: Differential-Geometrical Methods in Statistics. Springer Verlag (1985)
2. Amari, S.: Information geometry of the EM and em algorithms for neural networks. *Neural Networks* **8** (1995) 1379–1408.
3. Lee, D.D., Seung, H.S.: Learning of the parts of objects by non-negative matrix factorization. *Nature* **401** (1999) 788–791.
4. Hoyer, P.: Non-negative matrix factorization with sparseness constraints. *Journal of Machine Learning Research* **5** (2004) 1457–1469.
5. Sajda, P., Du, S., Parra, L.: Recovery of constituent spectra using non-negative matrix factorization. In: *Proceedings of SPIE – Volume 5207, Wavelets: Applications in Signal and Image Processing* (2003) 321–331.
6. Cho, Y.C., Choi, S.: Nonnegative features of spectro-temporal sounds for classification. *Pattern Recognition Letters* **26** (2005) 1327–1336.
7. Lee, D.D., Seung, H.S.: Algorithms for nonnegative matrix factorization. Volume 13. *NIPS*, MIT Press (2001)
8. Cichocki, A., Amari, S.: *Adaptive Blind Signal And Image Processing* (New revised and improved edition). John Wiley, New York (2003)
9. Byrne, C.: Accelerating the EMM algorithm and related iterative algorithms by rescaled block-iterative (RBI) methods. *IEEE Transactions on Image Processing* **7** (1998) 100 – 109.
10. Csiszár, I.: Information measures: A critical survey. In: *Prague Conference on Information Theory*, Academia Prague. Volume A. (1974) 73–86.
11. Cressie, N.A., Read, T.: *Goodness-of-Fit Statistics for Discrete Multivariate Data*. Springer, New York (1988)
12. Kompass, R.: A generalized divergence measure for nonnegative matrix factorization, *Neuroinformatics Workshop*, Torun, Poland (September, 2005)