

Multilayer nonnegative matrix factorization using projected gradient approaches

Andrzej Cichocki*, Rafal Zdunek**

Laboratory for Advanced Brain Signal Processing, Brain Science Institute, RIKEN, Wako-shi, Saitama 351-0198, Japan
e-mail: cia@brain.riken.jp

Received: date / Revised version: date

Abstract The most popular algorithms for Nonnegative Matrix Factorization (NMF) belong to the class of multiplicative Lee-Seung algorithms which have usually relative low complexity but are characterized by slow-convergence and risk to stuck in local minima. In this paper, we present and compare the performance of the additive algorithms based on three different variations of a projected gradient approach. Additionally, we shortly discuss a novel multi-layer approach to NMF algorithms combined together with multi-start initializations procedure, which in general, considerably improves the performance of all NMF algorithms. We demonstrate that this approach (the multi-layer system with projected gradient algorithms) can usually give much better performance than standard multiplicative algorithms, especially, if data are ill-conditioned, badly-scaled, and/or a number of observations is only slightly greater than a number of nonnegative hidden components. Our new implementations of NMF are demonstrated with the simulations performed for Blind Source Separation (BSS) data.

1 Introduction

NMF and its extended versions, nonnegative matrix deconvolution (NMD), and nonnegative tensor factorization (NTF) are relatively new and promising techniques with many potential scientific and engineering applications including: classification [1–5], clustering and segmentation of patterns [6–10], dimensionality reduction [11,12], face or object recognition [12–15], spectra recovering [16–18], language modeling, speech processing,

data mining and data analysis, e.g., text analysis [19] and music transcription [5,17,20].

NMF is often able to recover hidden structures in the data, and to provide biological insight. Depending on an application, the hidden structures may have different interpretation. For example, Lee and Seung in [6] introduced NMF as a method to decompose an image (face) into parts-based representations (parts reminiscent of features such as lips, eyes, nose, etc.). In blind source separation [21], the recovered components are unknown hidden (lateral) nonnegative components that cannot be observed directly. In many cases, NMF performs dimensionality reduction, and the retrieval components in a low-dimensional space have the similar interpretation (pattern analysis) as, e.g. the components obtained with PCA.

The simplest linear model used in NMF is of the form:

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

where $\mathbf{Y} = [y_{it}] \in \mathbb{R}^{m \times T}$ is a matrix of observations, $\mathbf{A} = [a_{ij}] \in \mathbb{R}^{m \times n}$ is an unknown basis or mixing matrix with nonnegative entries, $\mathbf{X} = [x_{jt}] \in \mathbb{R}^{n \times T}$ is a matrix of unknown hidden nonnegative components or sources, and $\mathbf{V} \in \mathbb{R}^{m \times T}$ is a matrix of additive noise; typically $T \gg m > n$. The objective is to estimate \mathbf{A} and \mathbf{X} knowing only \mathbf{Y} . To solve the problem, we usually define a suitable cost function and perform alternating minimization similar to Expectation Maximization (EM) approach [6].

There are many possibilities for defining the cost function $D(\mathbf{Y}||\mathbf{A}\mathbf{X})$, and many procedures for performing its alternating minimization, which leads to several kinds NMF algorithms: multiplicative, projected gradient, and fixed point [6,22,21,23–27].

The most known and used adaptive multiplicative algorithm for NMF is based Squared Euclidean distance

* On leave from Warsaw University of Technology, Poland

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

(expressed as squared Frobenius norm):

$$\begin{aligned} D_F(\mathbf{Y}||\mathbf{A}\mathbf{X}) &= \frac{1}{2} \|\mathbf{Y} - \mathbf{A}\mathbf{X}\|_F^2 \\ &= \frac{1}{2} \sum_{i=1}^m \sum_{t=1}^T |y_{it} - [\mathbf{A}\mathbf{X}]_{it}|^2 \\ &\text{s. t. } a_{ij} \geq 0, \quad x_j(t) = x_{jt} \geq 0 \quad \forall i, j, t, \end{aligned} \quad (2)$$

which is optimal for a Gaussian distributed noise). On basis of this cost function Lee and Seung proposed the following multiplicative algorithm [6]:

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

which is an extension of the well known ISRA (Image Space Reconstruction Algorithm) algorithm [28].

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 (generalized) Kullback-Leibler I-divergence [6,29]:

$$\begin{aligned} D_{KL}(\mathbf{Y}||[\mathbf{A}\mathbf{X}]) &= \sum_{it} \left(y_{it} \log \frac{y_{it}}{[\mathbf{A}\mathbf{X}]_{it}} + [\mathbf{A}\mathbf{X}]_{it} - y_{it} \right) \\ \text{s. t. } x_{jt} &\geq 0, \quad a_{ij} \geq 0, \quad \|\mathbf{a}_j\|_1 = \sum_{i=1}^m a_{ij} = 1. \end{aligned} \quad (4)$$

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

$$x_{jt} \leftarrow x_{jt} \frac{\sum_{i=1}^m a_{ij} (y_{it}/[\mathbf{A}\mathbf{X}]_{it})}{\sum_{q=1}^m a_{jq}}, \quad (5)$$

$$a_{ij} \leftarrow a_{ij} \frac{\sum_{t=1}^T x_{jt} (y_{it}/[\mathbf{A}\mathbf{X}]_{it})}{\sum_{t=1}^T x_{jt}}, \quad (6)$$

which are extensions (by alternating minimization) of the well known EMLL or Richardson-Lucy algorithm (RLA) [28].

However, the performance of the above multiplicative NMF algorithms may be quite poor, especially, when the unknown nonnegative components are badly scaled (ill-conditioned data), insufficiently sparse, and a number of observations is equal or only slightly greater than a number of latent (hidden) components.

In the paper, we focus on new projected gradient algorithms in the context of application to NMF. Additionally, we shortly present the multilayer extension to NMF that considerably improves the performance of all the existing NMF algorithms, especially the gradient projected algorithms.

2 Projected gradient techniques

In contrast to the multiplicative Lee-Seung NMF algorithms [6], this class of algorithms has additive updates. The algorithms discussed here mostly use the squared

Euclidean distance (2). The projected gradient method can be generally expressed by iterative updates:

$$\mathbf{X}^{(k+1)} = P_\Omega[\mathbf{X}^{(k)} + \eta_X^{(k)} \mathbf{P}_X], \quad (7)$$

$$\mathbf{A}^{(k+1)} = P_\Omega[\mathbf{A}^{(k)} + \eta_A^{(k)} \mathbf{P}_A], \quad (8)$$

where $P_\Omega[\xi]$ is a projection of ξ onto feasible set Ω , \mathbf{P}_X and \mathbf{P}_A are ascent directions for \mathbf{X} and \mathbf{A} , respectively. There are many rules for choosing the learning rates $\eta_X^{(k)}$ and $\eta_A^{(k)}$. In the presented methods, the learning rules are adjusted in this way to maintain nonnegativity constraints, which is equivalent to perform the projection $P_\Omega[\xi]$.

2.1 Projected gradient

We formulated the NMF problem as the following optimization problem:

$$\min_{\mathbf{X} \in \mathbb{R}^{n \times N}, \mathbf{A} \in \mathbb{R}^{m \times n}} D_F(\mathbf{Y}||\mathbf{A}\mathbf{X}), \quad \text{s.t. } a_{ij}, x_{jt} \geq 0, \quad (9)$$

which can be also solved with the following alternating Projected Gradient (PG) iterative updates

$$\mathbf{X}^{(k+1)} = P_\Omega[\mathbf{X}^{(k)} + \eta_X^{(k)} \mathbf{P}_X^{(k)}], \quad (10)$$

$$\mathbf{A}^{(k+1)} = P_\Omega[\mathbf{A}^{(k)} + \eta_A^{(k)} \mathbf{P}_A^{(k)}], \quad (11)$$

where

$$\mathbf{P}_X^{(k)} = -\nabla_X D_F(\mathbf{Y}||\mathbf{A}^{(k)}\mathbf{X})|_{\mathbf{X}=\mathbf{X}^{(k)}},$$

$$\mathbf{P}_A^{(k)} = -\nabla_A D_F(\mathbf{Y}||\mathbf{A}\mathbf{X}^{(k+1)})|_{\mathbf{A}=\mathbf{A}^{(k)}},$$

$$P_\Omega[\xi] = \begin{cases} \xi & \text{if } \xi \geq 0, \\ 0 & \text{otherwise} \end{cases}. \quad (12)$$

Several choices are available for selecting the optimal values of $\eta_X^{(k)}$, $\eta_A^{(k)}$ in each iteration k .

Recently, Chih-Jen Lin in [26] applied the Armijo rule to the squared Euclidean distance. For computation of \mathbf{X} , such a value of η_X is decided on which

$$\eta_X^{(k)} = \beta^{m_k}, \quad (13)$$

where m_k is the first non-negative integer m for which

$$\begin{aligned} D_F(\mathbf{Y}||\mathbf{A}\mathbf{X}^{(k+1)}) - D_F(\mathbf{Y}||\mathbf{A}\mathbf{X}^{(k)}) &\leq \\ &\leq \sigma \nabla_X D_F(\mathbf{Y}||\mathbf{A}\mathbf{X}^{(k)})^T (\mathbf{X}^{(k+1)} - \mathbf{X}^{(k)}). \end{aligned}$$

The similar rule is applied for computing \mathbf{A} . The parameters β and σ decide about a convergence speed. In this algorithm we set $\sigma = 0.01$, $\beta = 0.1$. We extend the Lin's algorithm to the multilayer system, which substantially improves the performance of the NMF. Moreover, it is possible to extend this approach to other divergences $D(\mathbf{Y}||\mathbf{A}\mathbf{X})$, such as the Kullback-Leibler, dual Kullback-Leibler or α -divergences [21].

2.2 Interior point gradient

The Interior Point Gradient (IPG) algorithm was proposed by Merritt and Zhang in [30] to a nonnegatively constrained least-squares problem. We extend this algorithm to NMF applications. This algorithm is based on the scaled gradient descent method, where the descent direction \mathbf{P}_X for \mathbf{X} is determined by a negative scaled gradient, i.e.

$$\mathbf{P}_X = -\mathbf{D} \odot \nabla_X D_F(\mathbf{Y}||\mathbf{A}\mathbf{X}), \quad (14)$$

with the scaling vector

$$\mathbf{D} = \mathbf{X} \odot (\mathbf{A}^T \mathbf{A}\mathbf{X}), \quad (15)$$

where \odot and \oslash mean component-wise multiplication and division, respectively. The cost function $D_F(\mathbf{Y}||\mathbf{A}\mathbf{X})$ was assumed to be the squared Euclidean distance (2). The IPG algorithm can be written in the compact form

$$\mathbf{A} \leftarrow \mathbf{A} + \eta_A [\mathbf{A} \oslash (\mathbf{A}\mathbf{X}\mathbf{X}^T)] \odot [(\mathbf{A}\mathbf{X} - \mathbf{Y})\mathbf{X}^T], \quad (16)$$

$$\mathbf{X} \leftarrow \mathbf{X} + \eta_X [\mathbf{X} \oslash (\mathbf{A}^T \mathbf{A}\mathbf{X})] \odot [\mathbf{A}^T (\mathbf{A}\mathbf{X} - \mathbf{Y})], \quad (17)$$

where $\eta_A > 0$ and $\eta_X > 0$ are suitably chosen learning rates.

In interior-point gradient methods, the learning rates are adjusted in each iteration to keep the iterates positive. In the IPG algorithm, the learning rates are chosen so that to be close to η_X^* and η_A^* which are the exact minimizers of $D_F(\mathbf{Y}||\mathbf{A}(\mathbf{X} + \eta_X \mathbf{P}_X))$ and $D_F(\mathbf{Y}||(\mathbf{A} + \eta_A \mathbf{P}_A)\mathbf{X})$, respectively, and on the other hand, to keep some distance to the boundary of the nonnegative orthant.

We used the following novel implementation of the IPG algorithm for NMF:

Algorithm 1 (IPG-NMF)

Set $\mathbf{A}, \mathbf{X}, \tau \in (0, 1)$, % Initialization
For $s = 0, 1, \dots$, % Alternating
 Step 1: Do **A-IPG** iterations with Algorithm 2,
 Step 2: Do **X-IPG** iterations with Algorithm 3,
End % Alternating

Algorithm 2 (A-IPG)

For $n = 0, 1, \dots$, % Inner loop for \mathbf{A}
 $\nabla_A D(\mathbf{Y}||\mathbf{A}\mathbf{X}) \leftarrow (\mathbf{A}\mathbf{X} - \mathbf{Y})\mathbf{X}^T$,
 $\mathbf{D} \leftarrow \mathbf{A} \oslash (\mathbf{A}\mathbf{X}\mathbf{X}^T)$,
 $\mathbf{P}_A \leftarrow -\mathbf{D} \odot \nabla_A D(\mathbf{Y}||\mathbf{A}\mathbf{X})$,
 $\eta_A^* = -\frac{(\text{vec}(\mathbf{P}_A)^T \text{vec}(\nabla_A D(\mathbf{Y}||\mathbf{A}\mathbf{X})))}{(\text{vec}(\mathbf{P}_A \mathbf{X})^T \text{vec}(\mathbf{P}_A \mathbf{X}))}$,
 $\hat{\eta}_A = \max\{\eta_A : \mathbf{A} + \eta_A \mathbf{P}_A \geq 0\}$,
 Set: $\hat{\tau} \in [\tau, 1)$, $\eta_A = \min(\hat{\tau} \hat{\eta}_A, \eta_A^*)$,
 $\mathbf{A} \leftarrow \mathbf{A} + \eta_A \mathbf{P}_A$,
End

Algorithm 3 (X-IPG)

For $n = 0, 1, \dots$, % Inner loop for \mathbf{X}
 $\nabla_X D(\mathbf{Y}||\mathbf{A}\mathbf{X}) \leftarrow \mathbf{A}^T (\mathbf{A}\mathbf{X} - \mathbf{Y})$,
 $\mathbf{D} \leftarrow \mathbf{X} \oslash (\mathbf{A}^T \mathbf{A}\mathbf{X})$,
 $\mathbf{P}_X \leftarrow -\mathbf{D} \odot \nabla_X D(\mathbf{Y}||\mathbf{A}\mathbf{X})$,
 $\eta_X^* = -\frac{(\text{vec}(\mathbf{P}_X)^T \text{vec}(\nabla_X D(\mathbf{Y}||\mathbf{A}\mathbf{X})))}{(\text{vec}(\mathbf{A}\mathbf{P}_X)^T \text{vec}(\mathbf{A}\mathbf{P}_X))}$,
 $\hat{\eta}_X = \max\{\eta_X : \mathbf{X} + \eta_X \mathbf{P}_X \geq 0\}$,
 Set: $\hat{\tau} \in [\tau, 1)$, $\eta_X = \min(\hat{\tau} \hat{\eta}_X, \eta_X^*)$,
 $\mathbf{X} \leftarrow \mathbf{X} + \eta_X \mathbf{P}_X$,
End

where \mathbf{P}_A and \mathbf{P}_X are descent directions, $\hat{\eta}_A$ and $\hat{\eta}_X$ are step lengths towards boundary of the nonnegative orthant, and η_A and η_X are current step lengths, respectively.

2.3 Regularized minimal residual norm steepest descent algorithm

The Minimal Residual Norm Steepest Descent (MRNSD) algorithm, which has been proposed by Nagy and Strakos [31] to image restoration problems, has been found to be the same as the EMLS algorithm developed by Kaufman [32]. The original MRNSD solves the following problem (assuming that the basis matrix \mathbf{A} is known)

$$\Phi(\mathbf{x}(t)) = \frac{1}{2} \|\mathbf{y}(t) - \mathbf{A}\mathbf{x}(t)\|_2^2, \quad (18)$$

subject to $\mathbf{x}(t) \geq 0$, $t = 1, \dots, T$.

The nonnegativity constraints are achieved by assuming the nonlinear transformation $\mathbf{x}(t) = \exp\{\mathbf{z}(t)\}$, and then

$$\begin{aligned} \nabla_{\mathbf{z}(t)} \Phi(\mathbf{x}(t)) &= \text{diag}(\mathbf{x}(t)) \nabla_{\mathbf{x}(t)} \Phi(\mathbf{x}(t)) \\ &= \text{diag}(\mathbf{x}(t)) \mathbf{A}^T (\mathbf{A}\mathbf{x}(t) - \mathbf{y}(t)) = 0 \end{aligned} \quad (19)$$

satisfies the KKT conditions. The solution is updated by the following rules:

$$\begin{aligned} \mathbf{p}(t) &\leftarrow \text{diag}(\mathbf{x}(t)) \mathbf{A}^T (\mathbf{A}\mathbf{x}(t) - \mathbf{y}(t)), \\ \mathbf{x}(t) &\leftarrow \mathbf{x}(t) + \eta \mathbf{p}(t). \end{aligned}$$

We applied the MRNSD approach to the regularized cost function:

$$D_F^{(\alpha_X)}(\mathbf{Y}||\mathbf{A}\mathbf{X}) = \frac{1}{2} \|\mathbf{Y} - \mathbf{A}\mathbf{X}\|_F^2 + \alpha_X J_X(\mathbf{X}), \quad (20)$$

where both matrix \mathbf{A} and \mathbf{X} are unknown and regularization term of the form $J_X(\mathbf{X}) = \sum_{j=1}^r \sum_{t=1}^T x_{jt}$ has been introduced in the order to provide the sparseness of the matrix \mathbf{X} .

We have implemented the following MRNSD algorithm for NMF:

Algorithm 4 (MRNSD-NMF)

Set $\mathbf{A}, \mathbf{X}, \mathbf{1}_m = [1, \dots, 1]^T \in \mathbb{R}^m$,
 $\mathbf{1}_r = [1, \dots, 1]^T \in \mathbb{R}^r$, $\mathbf{1}_T = [1, \dots, 1]^T \in \mathbb{R}^T$,
For $s = 0, 1, \dots$, % Alternating
Step 1: Do **A-MRNSD** iterations with
Algorithm 5,
Step 2: Do **X-MRNSD** iterations
with Algorithm 6,
End % Alternating

Algorithm 5 (A-MRNSD)

$\mathbf{G} = \nabla_{\mathbf{A}} D(\mathbf{Y} || \mathbf{A}\mathbf{X})$
 $= (\mathbf{A}\mathbf{X} - \mathbf{Y})\mathbf{X}^T$,
For $n = 0, 1, \dots$, % Inner loop for \mathbf{A}
 $\gamma = \mathbf{1}_m^T [\mathbf{G} \odot \mathbf{X} \odot \mathbf{G}] \mathbf{1}_r^T$,
 $\mathbf{P}_A = -\mathbf{A} \odot \mathbf{G}$, $\mathbf{U} = \mathbf{P}_A \mathbf{X}$,
 $\eta^* = \gamma (\mathbf{1}_m^T [\mathbf{U} \odot \mathbf{U}] \mathbf{1}_T)^{-1}$,
 $\eta = \min \{ \eta^*, \min_{p_{ij} < 0} (-\mathbf{A} \odot \mathbf{P}_A) \}$,
 $\mathbf{A} \leftarrow \mathbf{A} + \eta \mathbf{P}_A$, $\mathbf{Z} = \mathbf{U}\mathbf{X}^T$, $\mathbf{G} \leftarrow \mathbf{G} + \eta \mathbf{Z}$,
End % Alternating

Algorithm 6 (X-MRNSD)

$\mathbf{G} = \nabla_{\mathbf{X}} D(\mathbf{Y} || \mathbf{A}\mathbf{X})$
 $= \mathbf{A}^T (\mathbf{A}\mathbf{X} - \mathbf{Y}) + \eta_X$,
For $n = 0, 1, \dots$, % Inner loop for \mathbf{X}
 $\gamma = \mathbf{1}_r^T [\mathbf{G} \odot \mathbf{X} \odot \mathbf{G}] \mathbf{1}_T^T$,
 $\mathbf{P}_X = -\mathbf{X} \odot \mathbf{G}$, $\mathbf{U} = \mathbf{A}\mathbf{P}_X$,
 $\eta^* = \gamma (\mathbf{1}_m^T [\mathbf{U} \odot \mathbf{U}] \mathbf{1}_T)^{-1}$,
 $\eta = \min \{ \eta^*, \min_{p_{jt} < 0} (-\mathbf{X} \odot \mathbf{P}_X) \}$,
 $\mathbf{X} \leftarrow \mathbf{X} + \eta \mathbf{P}_X$, $\mathbf{Z} = \mathbf{A}^T \mathbf{U}$, $\mathbf{G} \leftarrow \mathbf{G} + \eta \mathbf{Z}$,
End % Alternating

3 Multilayer approach

In order to improve performance of the NMF, especially for ill-conditioned and badly scaled data and also to reduce risk of getting stuck in local minima of a cost function, we have developed a simple hierarchical and multi-stage procedure in which we perform a sequential decomposition (factorization) of nonnegative matrices as follows:

Algorithm 7 (Multilayer Algorithm)

Set: $\mathbf{X}_0 = \mathbf{Y}$,
For $l = 1, 2, \dots, L$, **do** :
Initialize randomly $\mathbf{A}_l^{(0)}$ and/or $\mathbf{X}_l^{(0)}$,
For $k = 1, 2, \dots, K$, **do** :
 $\mathbf{X}_l^{(k)} = \arg \min_{\mathbf{X}_l \geq 0} \left\{ D(\mathbf{X}_{l-1} || \mathbf{A}_l^{(k-1)} \mathbf{X}_l) \right\}$,
 $\mathbf{A}_l^{(k)} = \arg \min_{\mathbf{A}_l \geq 0} \left\{ \tilde{D}(\mathbf{X}_{l-1} || \mathbf{A}_l \mathbf{X}_l^{(k)}) \right\}$,
 $[\mathbf{A}_l^{(k)}]_{ij} \leftarrow \left[\frac{a_{ij}}{\sum_{i=1}^m a_{ij}} \right]_l^{(k)}$,
End
 $\mathbf{X}_l = \mathbf{X}_l^{(K)}$, $\mathbf{A}_l = \mathbf{A}_l^{(K)}$,
End

End

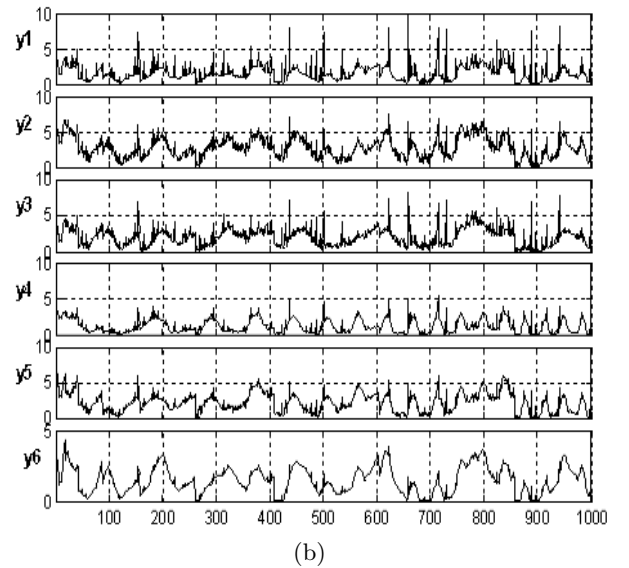
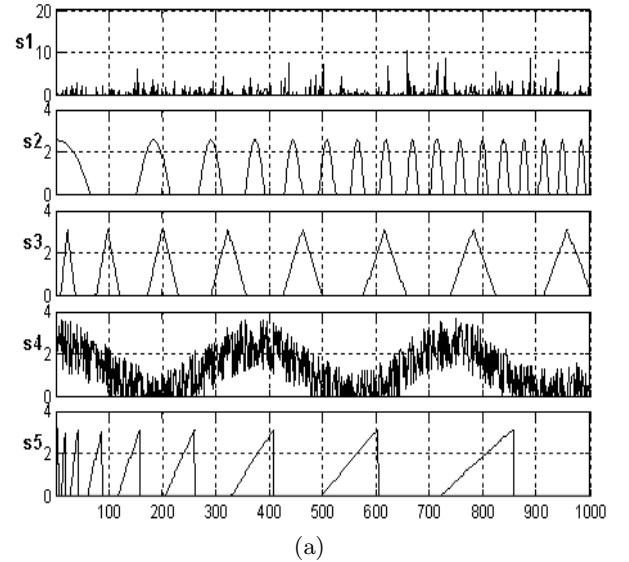
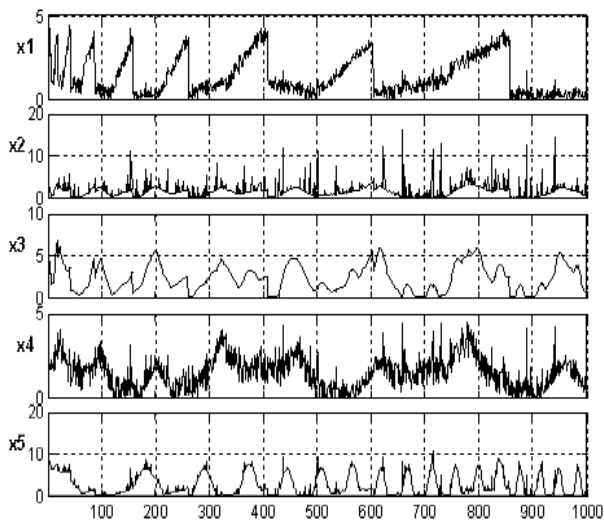
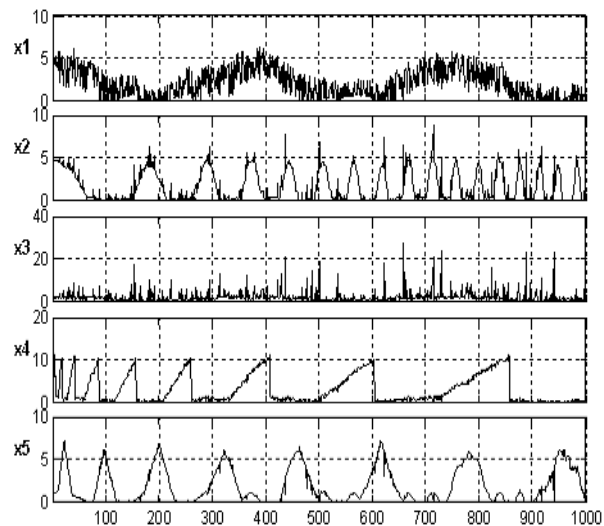


Fig. 1 (a) Original 5 source signals; (b) observed 6 mixed signals mixed synthetically by random generated ill-conditioned matrix.



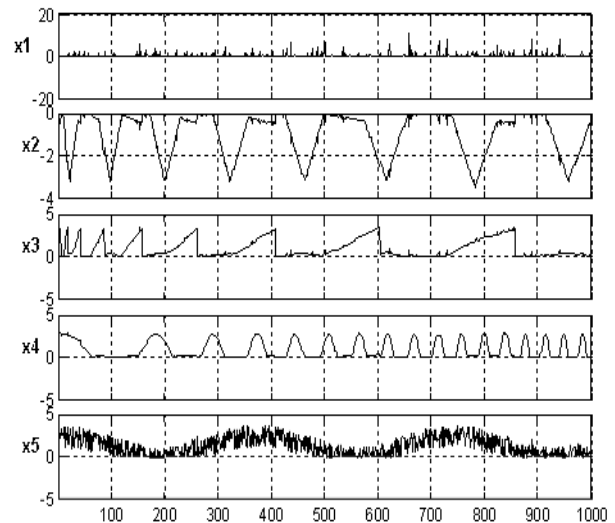
(a)



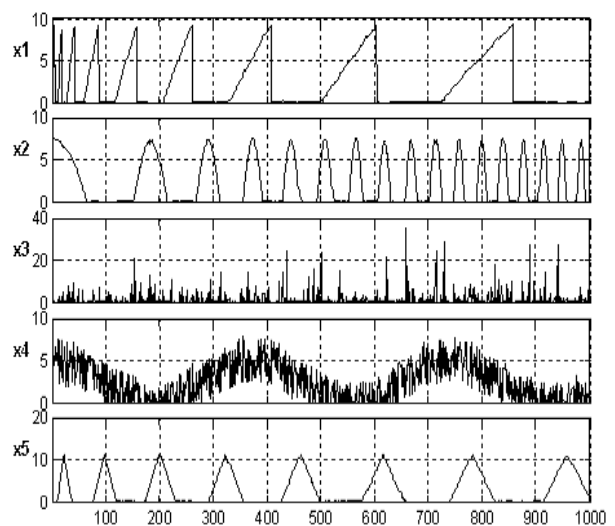
(b)

Fig. 2 Estimated sources with: (a) the Lee-Seung algorithm (EMML) with one single layer; (b) the standard Lee-Seung algorithm (EMML) with three layers (slight to moderate improvements are observed).

In the first step, we perform the basic decomposition $\mathbf{Y} = \mathbf{A}_1 \mathbf{X}_1$ using any available NMF algorithm, where $\mathbf{A}_1 \in \mathbb{R}^{m \times n}$ and $\mathbf{X}_1 \in \mathbb{R}^{n \times T}$ with $m \geq n$. In the second stage, the results obtained from the first stage are used to perform the similar decomposition: $\mathbf{X}_1 = \mathbf{A}_2 \mathbf{X}_2$, where $\mathbf{A}_2 \in \mathbb{R}^{n \times n}$ and $\mathbf{X}_2 \in \mathbb{R}^{n \times T}$, using the same or different update rules, and so on. We continue our decomposition taking into account only the last achieved components. The process can be repeated arbitrary many times until some stopping criteria are satisfied. In each step, we usually obtain gradual improvements of the performance. Thus, our model has the form: $\mathbf{Y} = \mathbf{A}_1 \mathbf{A}_2 \dots \mathbf{A}_L \mathbf{X}_L$, with the basis matrix defined as $\mathbf{A} = \mathbf{A}_1 \mathbf{A}_2 \dots \mathbf{A}_L \in \mathbb{R}^{m \times T}$. Physically, this means that we build up a system that has many layers or cascade connection of L



(a)

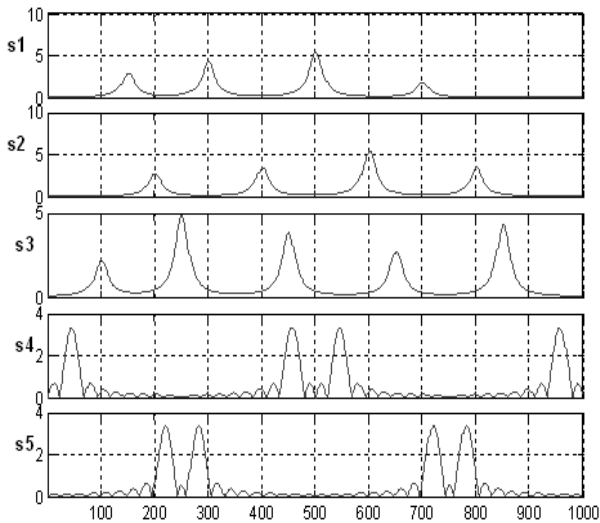


(b)

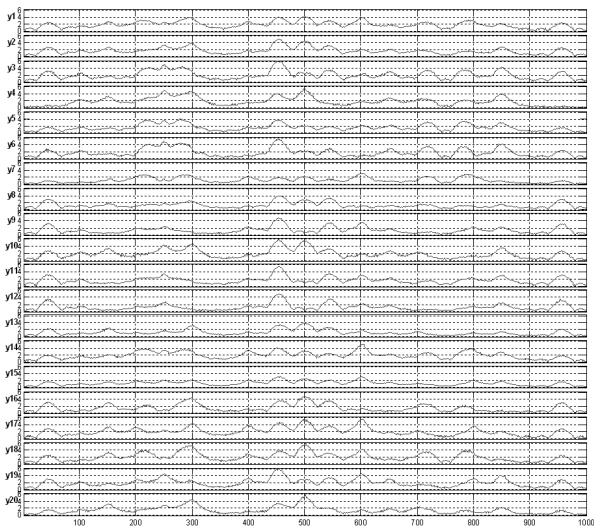
Fig. 3 Estimated sources with: (a) ThinICA algorithm with one layer; (b) IPG algorithm with three layers (in this case, we are able to reconstruct hidden components almost perfectly).

mixing subsystems. The key point in our novel approach is that the learning (update) process to find parameters of matrices \mathbf{A}_l and \mathbf{X}_l is performed sequentially, i.e., layer by layer. In each step or each layer, we can use the same cost (loss) functions, and consequently, the same learning (minimization) rules, or completely different cost functions and/or corresponding update rules. Thus, our approach can be described by Algorithm 7. The cost functions $D(\mathbf{Y}||\mathbf{A}\mathbf{X})$ and $\tilde{D}(\mathbf{Y}||\mathbf{A}\mathbf{X})$ can take various forms, e.g.: the alpha divergence, Bregman divergence, Csiszar divergence, beta divergence, Euclidean distance [21,27].

An open theoretical issue is to prove mathematically or explain more rigorously why the multilayer distributed NMF system results in considerable improve-



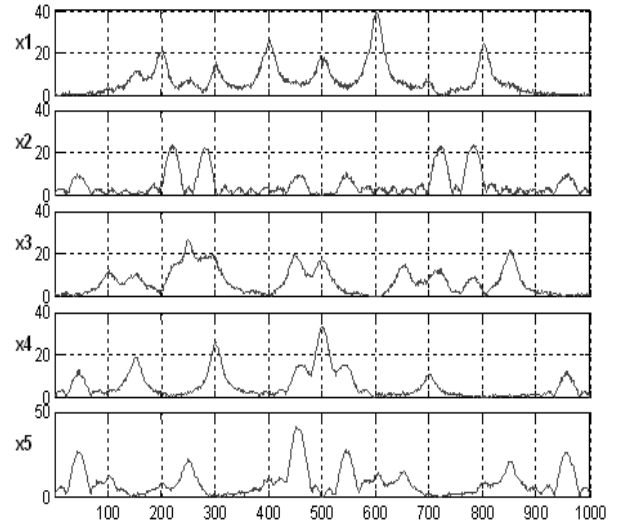
(a)



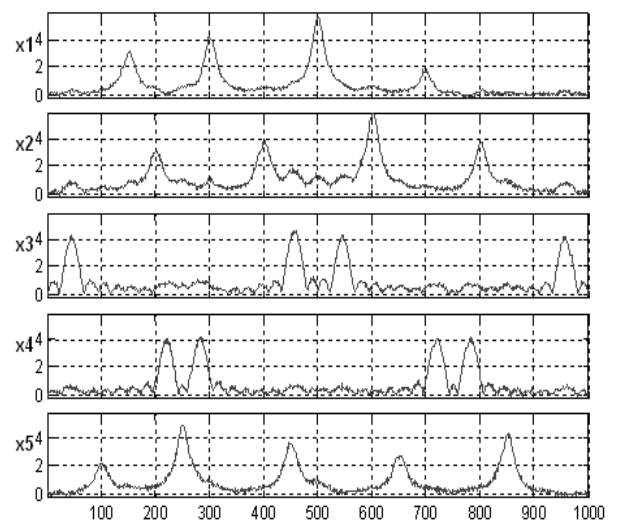
(b)

Fig. 4 (a) Original 5 source spectra; (b) observed 20 noisy mixed spectra (SNR = 20 dB).

ment in performance and reduces the risk of getting stuck at local minima. An intuitive explanation is as follows: the multilayer system provides a sparse representation of basis matrices \mathbf{A}_l , so even a true basis matrix \mathbf{A} is not sparse it can be represented by a product of sparse factors. In each layer we enforce (or encourage) a sparse representation. We found by extensive experiments that if the basis matrix is very sparse, most NMF algorithms have improved performance. However, not all data is a sufficiently sparse representation, so the main idea is to model any data by cascade connections of sparse sub-systems. On the other hand, such multilayer systems are biologically motivated and plausible.



(a)



(b)

Fig. 5 Estimated sources with: (a) the Lee-Seung algorithm (EMML) with one layer (SIR = 7.4, 6.9, 4.6, 3.9, 7.7 [dB]); (b) ThinICA algorithm with one layer (SIR = 16.7, 8.2, 17.2, 10.6, 14.9 [dB]).

4 Experiments

The proposed NMF algorithms have been extensively tested for many difficult benchmarks for signals and images with various statistical distributions. The simulation results confirmed that the developed algorithms together with the multilayer strategy are efficient and stable for a wide set of parameters. We show here three illustrative examples.

The five statistically dependent nonnegative signals shown in Fig. 1(a) have been mixed by randomly generated uniformly distributed nonnegative matrix $\mathbf{A} \in \mathbb{R}^{6 \times 5}$. The mixing signals are shown in Fig. 1(b). Using the standard multiplicative NMF algorithms we failed to estimate the original sources. Fig. 2(a) illustrates the re-

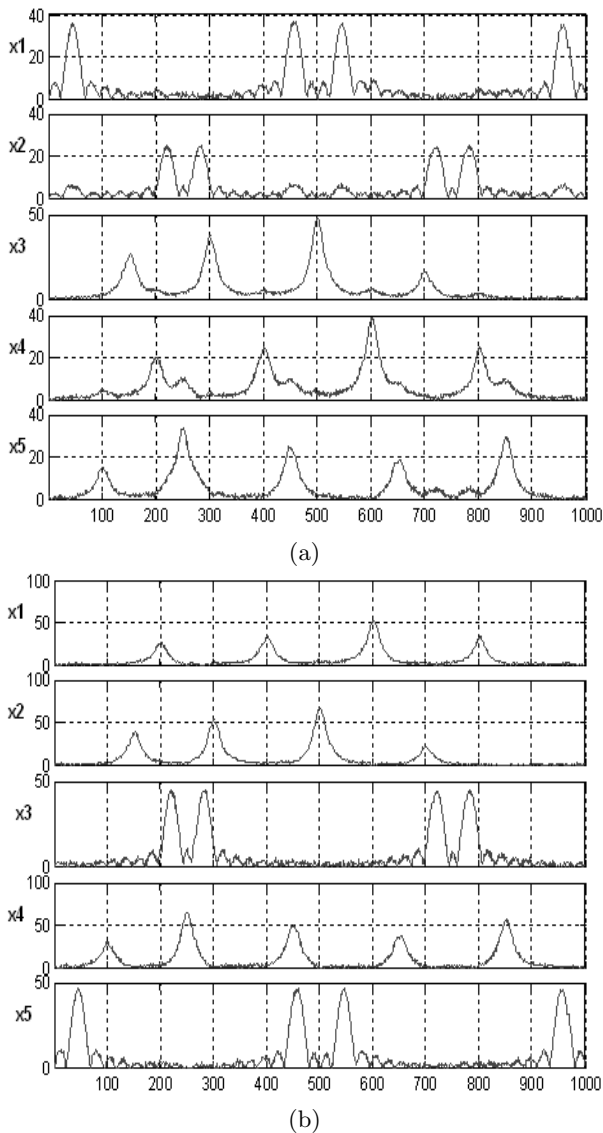


Fig. 6 Estimated sources with: (a) IPG algorithm with one layer (SIR = 19.4, 10.8, 17.8, 18.3, 11.6 [dB]); (b) IPG algorithm with three layers (SIR = 22.2, 21.1, 17.1, 22.9, 22 [dB]).

sults obtained with the standard Lee-Seung algorithm, referred here to as EMLM (Expectation Maximization Maximum Likelihood) based on alternating minimization of the Kullback-Leibler I-divergence. However, the same algorithm with only three layers of the multilayer technique gives much better results – see Fig. 2(b) and also compare the SIRs in Table 1 for this case. Slightly better performance for the multilayer system provides the second Lee-Seung algorithm referred to as ISRA (Image Space Reconstruction Algorithm). However, we have obtained the best performance for the projected gradient methods, especially for the IPG method – see Fig. 3(b) and Table 1. We also tried to apply the ICA algorithms to solve the problem but due to partial dependence of the sources the performance is rather poor,

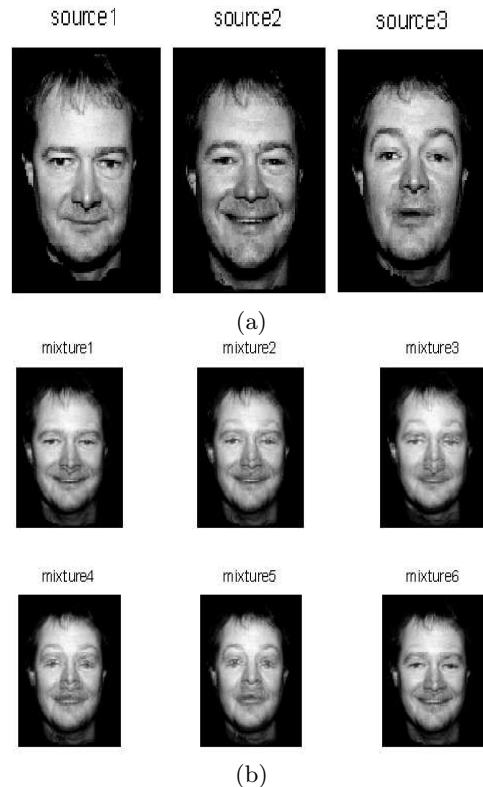


Fig. 7 (a) Original 3 source images; (b) observed 6 mixed images.

which is illustrated in Fig 3(a) for the powerful and flexible ThinICA algorithm [33]. The multilayer approach for the ThinICA algorithm does not improve the results. The most important feature of our approach consists in applying multi-layer technique that reduces the risk of getting stuck in local minima, and hence, considerably improves the performance of NMF.

Example 2 shows the separation results of 5 more realistic signals (see Fig. 4 (a)) that represent spectra. The separation is performed from 20 noisy observations ($\mathbf{A} \in \mathbb{R}^{20 \times 5}$ is a uniformly distributed random matrix), where the Gaussian noise with SNR = 20dB was added. The observations are shown in Fig. 4 (a). The simulation results are illustrated in Figs. 5 and 6. The IPG algorithm with 3 layers still gives the best performance.

Finally, in Example 3 we have used the 3 correlated and statistically dependent images: faces of the same person with different facial expression (see Fig. 7 (a)) that have been mixed with uniformly distributed random matrix $\mathbf{A} \in \mathbb{R}^{6 \times 3}$. The mixtures are shown in Fig. 7 (b). Figs. 8 and 9 present the separation results. Again, we have confirmed that the multilayer system considerably improves the performance of all the tested algorithms, and the best results are obtained with the IPG algorithm.

Table 1 SIRs for estimation of sources and columns of mixing matrix [dB] from noise-free mixtures of signals in Fig. 1. The number in parenthesis close to the algorithm name denotes the number of layers.

Sources:	1	2	3	4	5	Mean SIRs
EMML(1)	2.77	14.48	1.66	4.17	10	6.62
EMML(3)	8.45	11.35	13.68	16.49	17.19	13.43
ISRA(1)	6.42	3.18	2.05	19.36	4.27	7.06
ISRA(3)	20.05	9.48	16.35	15.08	27.54	17.7
PG(1)	2.2	12.6	3.3	3.85	7.24	5.84
PG(3)	23.02	43.75	31.33	20.95	25.23	28.86
IPG(1)	8.47	20.54	18.93	6.77	10.02	12.95
IPG(3)	48.53	38.36	39.11	24.37	44.69	39.78
MRNSD(1)	6.95	12.54	6.25	3.33	7.94	7.4
MRNSD(3)	26.42	30.39	24.67	23.53	20.04	29.49
ThinICA	19.62	21.12	15.11	26.42	15.16	19.49
Columns:	1	2	3	4	5	Mean SIRs
EMML(1)	7.02	8.39	2.84	12.22	6.59	7.41
EMML(3)	18	18.9	22.84	12.11	24.9	19.35
ISRA(1)	12.02	7.47	6.98	10.95	7.38	8.96
ISRA(3)	21.45	19.24	18.03	14.51	29.71	20.59
PG(1)	4.91	9.72	5.3	7.6	2.87	6.08
PG(3)	35.64	30.61	28.44	34.52	27.44	31.33
IPG(1)	20	14.79	17.59	20.74	6.02	16.83
IPG(3)	41.92	30.34	37.87	45.89	42.89	39.78
MRNSD(1)	12.63	7.83	2.67	7.77	3.48	6.87
MRNSD(3)	31.98	36.83	20.72	28.52	29.39	25.01
ThinICA	22	33.79	20.93	26	20.32	24.52

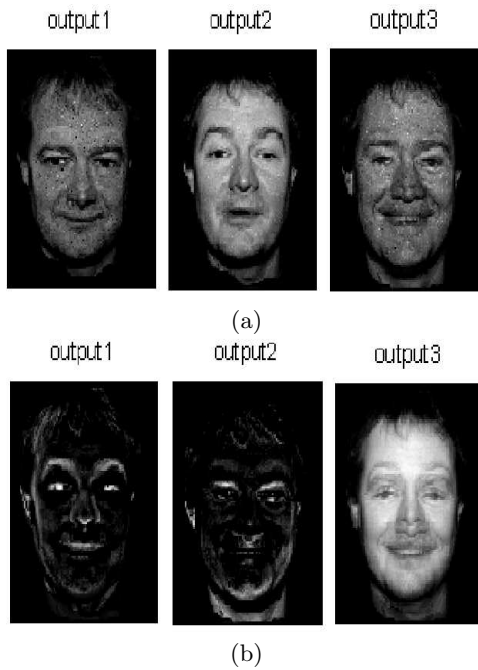


Fig. 8 Estimated images with: (a) the Lee-Seung algorithm (EMML) with one layer (SIR = 8.4, 8.6, 16.3 [dB]); (b) ThinICA algorithm with one layer (SIR = 12.8, 13.7, 12.9 [dB]).

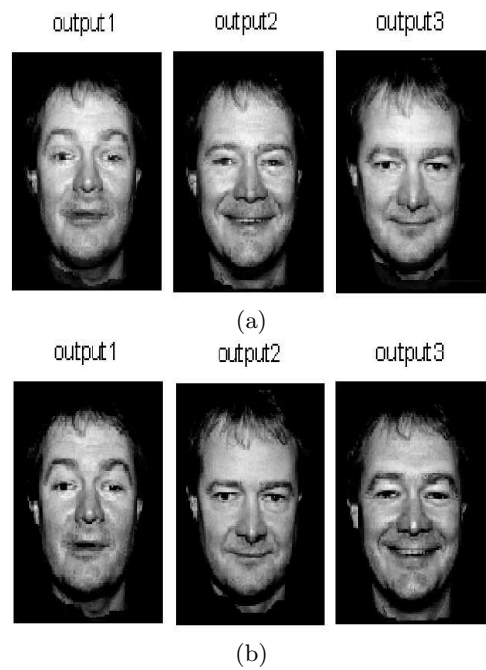


Fig. 9 Estimated images with: (a) IPG algorithm with one layer (SIR = 25.1, 22.4, 21 [dB]); (b) IPG algorithm with three layers (SIR = 44.5, 28.2, 45.7 [dB]).

5 Conclusions

In this paper we discuss three nonconventional implementations of NMF algorithms with additive updates based on projected gradient, interior-point gradient, and regularized minimal residual norm steepest descent techniques. We implemented them in MATLAB (see NMFLAB for Signal Processing – Matlab toolbox accessible in <http://www.bsp.brain.riken.jp>) using the novel multi-layer structure and found by extensive simulations their superior performance in comparison to the standard multiplicative algorithms, especially when the basis matrix is not sufficiently sparse and the number of observations is only slightly greater than the number of sources.

The efficiency of many NMF strategies is affected by the selection of the starting matrices. Poor initializations often result in slow convergence, and in certain instances may lead to an incorrect or irrelevant answer. It should be noted that in our approach a key role plays also a multi-start initialization for each layer. We usually start to run a specific algorithm for several (typically 10) randomly generated initial conditions (\mathbf{A} and \mathbf{X}) for a fixed but very small number of iterations (typically, 20 iterations) and select a such initial conditions which provide the lowest possible value of the Kullback-Leibler I-divergence after 20 iterations (see NMFLAB implementation).

References

- Guillamet, D., Vitrià, J., Schiele, B.: Introducing a weighted nonnegative matrix factorization for image classification. *Pattern Recognition Letters* **24** (2003) 2447–2454
- Guillamet, D., Schiele, B., Vitrià, J.: Analyzing nonnegative matrix factorization for image classification. In: 16th International Conference on Pattern Recognition (ICPR'02). Volume 2., Quebec City, Canada (2002) 116–119
- Guillamet, D., Vitrià, J.: Classifying faces with nonnegative matrix factorization. In: Proc. 5th Catalan Conference for Artificial Intelligence, Castello de la Plana, Spain (2002)
- Ahn, J.H., Kim, S., Oh, J.H., Choi, S.: Multiple nonnegative-matrix factorization of dynamic PET images. In: ACCV. (2004)
- Lee, J.S., Lee, D.D., Choi, S., Lee, D.S.: Application of nonnegative matrix factorization to dynamic positron emission tomography. In: 3rd International Conference on Independent Component Analysis and Blind Signal Separation, San Diego, CA (2001) 556–562
- Lee, D.D., Seung, H.S.: Learning the parts of objects by nonnegative matrix factorization. *Nature* **401** (1999) 788–791
- Li, H., Adali, T., Wang, D.E.: Non-negative matrix factorization with orthogonality constraints for chemical agent detection in raman spectra. In: IEEE Workshop on Machine Learning for Signal Processing, Mystic USA, (2005)
- Carmona-Saez, P., Pascual-Marqui, R.D., Tirado, F., Carazo, J.M., Pascual-Montano, A.: Biclustering of gene expression data by non-smooth non-negative matrix factorization. *BMC Bioinformatics* **7** (2006)
- Pascual-Montano, A., Carazo, J.M., Kochi, K., Lehman, D., Pascual-Marqui, R.: Nonsmooth nonnegative matrix factorization (nsNMF). *IEEE Trans. Pattern Analysis and Machine Intelligence* **28** (2006) 403–415
- Shahnaz, F., Berry, M., Pauca, P., Plemmons, R.: Document clustering using non-negative matrix factorization. *Journal on Information Processing and Management* **42** (2006) 373–386
- Brunet, J.P., Tamayo, P., Golub, T.R., Mesirov, J.P.: Metagenes and molecular pattern discovery using matrix factorization. Volume 101., PNAS (2000) 4164–4169
- Okun, O., Priisalu, H.: Fast nonnegative matrix factorization and its application for protein fold recognition. *EURASIP Journal on Applied Signal Processing* **2006** (2006) Article ID 71817, 8 pages
- Wang, Y., Jia, Y., Hu, C., Turk, M.: Non-negative matrix factorization framework for face recognition. *International Journal of Pattern Recognition and Artificial Intelligence* **19** (2005) 495–511
- Liu, W., Zheng, N.: Non-negative matrix factorization based methods for object recognition. *Pattern Recognition Letters* **25** (2004) 893–897
- Spratling, M.W.: Learning image components for object recognition. *Journal of Machine Learning Research* **7** (2006) 793–815
- Sajda, P., Du, S., Brown, T.R., Shungu, R.S.D.C., Mao, X., Parra, L.C.: Nonnegative matrix factorization for rapid recovery of constituent spectra in magnetic resonance chemical shift imaging of the brain. *IEEE Trans. Medical Imaging* **23** (2004) 1453–1465
- Sajda, P., Du, S., Brown, T., Parra, L., Stoyanova, R.: Recovery of constituent spectra in 3d chemical shift imaging using nonnegative matrix factorization. In: 4th International Symposium on Independent Component Analysis and Blind Signal Separation, Nara, Japan (2003) 71–76
- 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
- Dhillon, I.S., Modha, D.M.: Concept decompositions for large sparse text data using clustering. *Machine Learning J.* **42** (2001) 143–175
- Cho, Y.C., Choi, S.: Nonnegative features of spectrotemporal sounds for classification. *Pattern Recognition Letters* **26** (2005) 1327–1336
- Cichocki, A., Zdunek, R., Amari, S.: Csiszar's divergences for non-negative matrix factorization: Family of new algorithms. *LNCS* **3889** (2006) 32–39
- Chu, M., Plemmons, R.J.: Nonnegative matrix factorization and applications. *Bulletin of the International Linear Algebra Society* **34** (2005) 2–7
- Cho, H., Dhillon, I.S., Guan, Y., Sra, S.: Minimum sum squared residue based co-clustering of gene expression data. In: Proc. 4th SIAM International Conference on Data Mining (SDM), Florida (2004) 114–125
- Hoyer, P.: Non-negative matrix factorization with sparseness constraints. *Journal of Machine Learning Research* **5** (2004) 1457–1469

25. Finesso, L., Spreij, P.: Nonnegative matrix factorization and I-divergence alternating minimization. *Linear Algebra and its Applications* **416** (2006) 270–287
26. Lin, C.J.: Projected gradient methods for nonnegative matrix factorization. Technical report, Department of Computer Science, National Taiwan University, www.csie.ntu.edu.tw/~cjlin (2005)
27. Dhillon, I., Sra, S.: Generalized nonnegative matrix approximations with Bregman divergences. In: *Neural Information Proc. Systems, Vancouver, Canada* (2005) <http://www.cs.utexas.edu/ftp/pub/techreports/tr06-27.pdf>
28. Byrne, C.: Iterative projection onto convex sets using multiple bregman distances. In: *Inverse Problems*. Volume 15. (1999) 1295–1313
29. Lee, D.D., Seung, H.S.: Algorithms for nonnegative matrix factorization. In: *NIPS*. (2000) 556–562
30. Merritt, M., Zhang, Y.: An interior-point gradient method for large-scale totally nonnegative least squares problems. *J. Optimization Theory and Applications* **126** (2005) 191–202
31. Nagy, J.G., Strakos, Z.: Enforcing Nonnegativity in Image Reconstruction Algorithms, *Mathematical Modeling, Estimation, and Imaging*. Volume 4121. (2000) 182–190
32. Kaufman, L.: Maximum likelihood, least squares, and penalized least squares for PET. *IEEE Transactions on Medical Imaging* **12** (1993) 200–214
33. Cruces-Alvarez, S., Cichocki, A., Lathauwer, L.D.: Thin QR and SVD factorizations for simultaneous blind signal extraction. In: *Proc. of the European Signal Processing Conference (EUSIPCO), Vienna, Austria* (2004) 217–220