Electronics Letters, Vol. 42, No. 16 (2006), pp. 947-948.

Multilayer Nonnegative Matrix Factorization

A. Cichocki and R. Zdunek

Abstract: A multilayer approach to Nonnegative Matrix Factorization (NMF) algorithms is proposed. It considerably improves their performance; especially if a problem is ill-conditioned, or data are badly-scaled, and projected gradient algorithms are used. This is fully confirmed by our extensive simulations with diverse types of data in application to Blind Source Separation (BSS).
Indexing terms: Nonnegative Matrix Factorization (NMF), Blind Source Separation (BSS), Data Mining and Analysis

Introduction: NMF and its extended version, nonnegative matrix deconvolution (NMD), are relatively new and promising techniques with many potential scientific and engineering applications including: classification, clustering and segmentation of patterns, dimensionality reduction, face/image recognition, language modeling, speech processing, data mining and data analysis, e.g., text analysis and music transcription [1-4].

The simplest linear model used in NMF is of the form: $\mathbf{X} = \mathbf{AS} + \mathbf{V}$, where $\mathbf{X} \in \Re^{m \times T}$ is a matrix of observations, $\mathbf{A} \in \Re^{m \times n}$ is an unknown basis matrix with nonnegative entries, $\mathbf{S} \in \Re^{n \times T}$ is a matrix of unknown hidden nonnegative components, and $\mathbf{V} \in \Re^{m \times T}$ is a matrix of additive noise; typically $T >> m > n$.

There are many possibilities for defining the cost function $D(\mathbf{X} \| \mathbf{AS})$, and many procedures for performing its alternating minimization, which leads to several NMF algorithms; most of them are multiplicative or projected gradient [1, 3]. However, the performance of many existing 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.

New Results: 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: In the first step, we perform the basic decomposition $\mathbf{X} = \mathbf{A}_1\mathbf{S}_1$ using any available NMF algorithm. In the second stage, the results obtained from the first stage are used to perform the similar decomposition: $\mathbf{S}_1 = \mathbf{A}_2\mathbf{S}_2$ 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: $\mathbf{X} = \mathbf{A}_1\mathbf{A}_2 \ldots \mathbf{A}_L\mathbf{S}_L$, with the basis matrix defined as $\mathbf{A} = \mathbf{A}_1\mathbf{A}_2 \ldots \mathbf{A}_L$. Physically, this means that we build up a system that has many layers or cascade connection of L mixing subsystems. The key point in our novel approach is that the learning

(update) process to find parameters of sub-matrices $\mathbf{A}_l$ and $\mathbf{S}_l$ is performed sequentially, i.e., layer by layer and in each layer we use multi-start initialization. We select such random initialization which provides fastest decrease of a specific cost function (typically generalized KL divergence) [6].

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 the following algorithm:

Set: $\mathbf{X}_0 = \mathbf{X}$, Initialize randomly basis matrix $\mathbf{A}_1^{(0)}$ and/or $\mathbf{S}_1^{(0)}$:

For $l = 1, 2, \ldots, L$ do:

For $t = 0, 1, \ldots, T_{\max}$ do:

$$\mathbf{S}_l^{(t+1)} = \arg\min_{\mathbf{S} \geq 0} D_l(\mathbf{X}_l \,||\, \mathbf{A}_l^{(t)} \mathbf{S}) \Big|_{\mathbf{S} = \mathbf{S}_l^{(t)}} \tag{1}$$

$$\mathbf{A}_l^{(t+1)} = \arg\min_{\mathbf{A} \geq 0} \tilde{D}_l(\mathbf{X}_l \,||\, \mathbf{A} \mathbf{S}_l^{(t+1)}) \Big|_{\mathbf{A} = \mathbf{A}_l^{(t)}}, \quad \mathbf{A}_l^{(t+1)} \leftarrow \left[a_{ij} / \sum_{i=1}^{m} a_{ij} \right]_l^{(t+1)}, \tag{2}$$

End (for *t*)

$$\mathbf{X}_{l+1} = \mathbf{S}_l^{(T_{\max}+1)}$$

End (for *l*)

In the above algorithm, the cost functions $D(\mathbf{X} \,||\, \mathbf{AS})$ and $\tilde{D}(\mathbf{X} \,||\, \mathbf{AS})$ can take various forms, e.g.: the Amari alpha divergence, Bregman divergence, Csiszar divergence, beta divergence, Euclidean distance [4, 5].

As example, we present a very efficient and simple algorithm that uses the regularized Euclidean distance, i.e. $D(\mathbf{X}\,\|\,\mathbf{AS}) = \|\,\mathbf{X}-\mathbf{AS}\,\|_F^2 + \alpha_S\Omega_S(\mathbf{S}) + \alpha_A\Omega_A(\mathbf{A})$, where additional regularization terms: $\Omega_S(\mathbf{S}) = tr\{\mathbf{S}^T\mathbf{E}\,\mathbf{S}\}$, and $\Omega_A(\mathbf{A}) = tr\{\mathbf{A}\,\mathbf{E}\,\mathbf{A}^T\}$ (*tr* means trace of a matrix and $\mathbf{E}\in\mathbb{R}^{R\times R}$ is a matrix with all ones) are added to enforce smoothness of solution and to avoid local minima.

Assuming $\nabla_{\mathbf{S}} D(\mathbf{X}\,\|\,\mathbf{AS}) = \mathbf{0}$, $\nabla_{\mathbf{A}} D(\mathbf{X}\,\|\,\mathbf{AS}) = \mathbf{0}$ for positive entries in **A** and **S**, which occurs when a stationary point is reached, we have:

$$\mathbf{S}_l^{(t+1)} = \max\left\{\varepsilon, \left(\mathbf{A}^T\mathbf{A} + \alpha_S^{(t)}\mathbf{E}\right)^+ \mathbf{A}^T\mathbf{X}_l\right\}\Big|_{\mathbf{A}=\mathbf{A}_l^{(t)}}, \tag{3}$$

$$\mathbf{A}_l^{(t+1)} = \max\left\{\varepsilon, \mathbf{X}_l\,\mathbf{S}^T\left(\mathbf{S}\,\mathbf{S}^T + \alpha_A^{(t)}\mathbf{E}\right)^+\right\}\Big|_{\mathbf{S}=\mathbf{S}_l^{(t+1)}}, \tag{4}$$

where $\mathbf{B}^+$ is a Moore-Penrose inverse of **B**, ε is a small constant (10^{-9}) to enforce positive entries. To avoid local minima, we assume $\alpha_A^{(t)} = \alpha_S^{(t)} = \alpha_0 \exp\{-t/\tau\}$, which is motivated by a temperature schedule in the simulated annealing technique, where α_0 and τ are some constants. Alternatively, as a cost function, we can use Amari alpha divergence [5]: $D_A(x_{ik}\,\|\,z_{ik}) = \sum_{ik} x_{ik}\dfrac{x_{ik}^{\alpha-1} - z_{ik}^{\alpha-1}}{\alpha(\alpha-1)z_{ik}^{\alpha-1}} + \dfrac{z_{ik}-x_{ik}}{\alpha}$, where $x_{ik} = [\mathbf{X}]_{ik}$, $z_{ik} = [\mathbf{AS}]_{ik}$. Using (1)-(2) we derived a new family of NMF algorithms for $\alpha \neq 0$:

$$a_{ij} \leftarrow a_{ij}\left(\sum_k s_{jk}\left(\frac{x_{ik}}{[\mathbf{AS}]_{ik}+\varepsilon}\right)^{\alpha}\right)^{\frac{1}{\alpha}}, \quad a_{ij} = \frac{a_{ij}}{\sum_i a_{ij}}, \qquad \mathbf{S} \leftarrow \max\left\{\varepsilon, \mathbf{A}^+\mathbf{X}\right\}.$$

Experiments: To show robustness of our technique, we selected a difficult case in which four nonnegative badly scaled sources [Fig.1(a)] were mixed by the Hilbert matrix ($\mathbf{A} \in \Re^{5\times 4}$: $a_{ij} = (i + j - 1)^{-1}$) with condition number $\kappa = 8956$ [Fig.1(b)]. We applied the new algorithm (3) – (4) with 10 layers, and 1000 iterations in each layer, starting with different random initial conditions. Note that the performance of 10 layers system with 1000 iterations is substantially better than for 10000 iterations in a single layer (see Fig. 2.), but computational costs are the same. All the columns of the mixing matrix and all the sources were estimated with mean-SIR (Signal-to-Interference Ratio) larger than 120dB. We tested many existing NMF algorithms and found that they failed to estimate the sources in such a difficult scenario.

Conclusions: In this letter, we proposed a novel approach for NMF, which considerably improves the accuracy and performance of the new and existing NMF algorithms. Furthermore, we proposed two new algorithms for NMF, which together with the multilayer procedure with multi-start initializations, give promising results. We implemented the proposed procedure and NMF algorithms in our NMFLAB toolbox [6] for MATLAB, and confirmed by extensive simulations their validity and usefulness for NMF problems.

References:

1. Lee D.D., and Seung H.S., 'Learning of the parts of objects by non-negative matrix factorization', Nature, 1999, 401, pp. 788–791

2. Cho Y.C., and Choi S., 'Nonnegative features of spectro-temporal sounds for classification', Pattern Recognition Letters, 2005, 26, pp.1327–1336.

3. Chu M., and Plemmons R.J., 'Nonnegative matrix factorization and applications', Bulletin of the International Linear Algebra Society, 2005, 34, pp.2–7

4. Dhillon I.S., and Sra S., 'Generalized nonnegative matrix approximations with Bregman divergences', Proc. NIPS, Vancouver, Canada, December 2005.

5. Cichocki A., Zdunek, R., and Amari, S. 'Csiszar's divergences for non-negative matrix factorization: Family of new multiplicative algorithms', (ICA06) Springer LNCS, 2006, 3889, pp. 32– 39

6. Cichocki A., and Zdunek R., NMFLAB for Signal Processing: Matlab Toolbox for NMF http://www.bsp.brain.riken.jp/index.php

Authors' affiliations:

A. Cichocki*, R. Zdunek**

Laboratory for Advanced Brain Signal Processing,

RIKEN BSI, Wako-shi, Saitama 351-0198, Japan

E-mail: cia@brain.riken.jp

* On leave from Warsaw University of Technology, Poland

** On leave from Wroclaw University of Technology, Poland

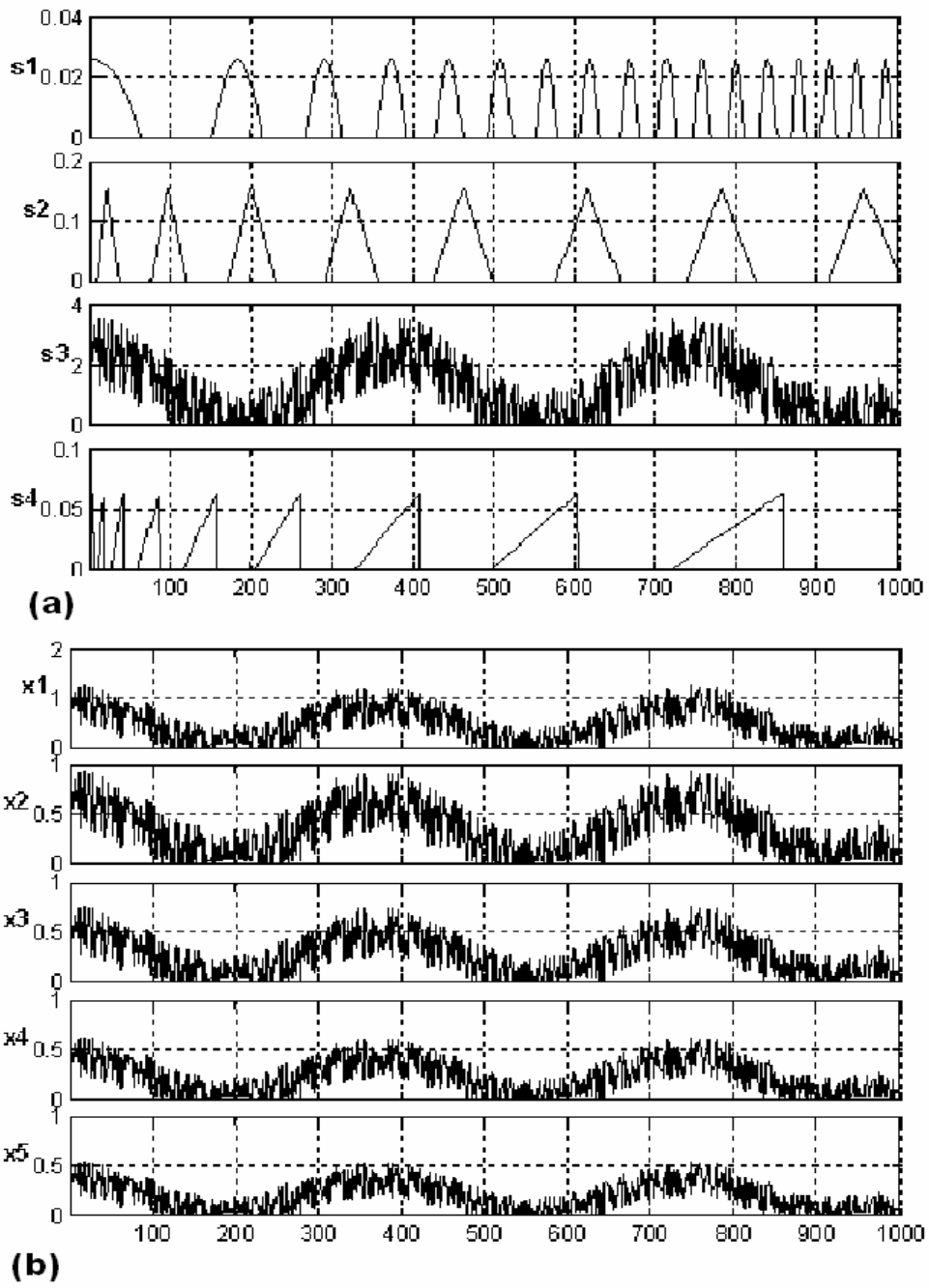

Fig.1. (a) Original badly-scaled sources, (b) Observed mixed signals with Hilbert mixing matrix $\mathbf{A} \in \Re^{5\times 4}$, (Estimated sources have SIRs above 120 dB, and neglecting scale and permutation, they are almost identical as original ones.)

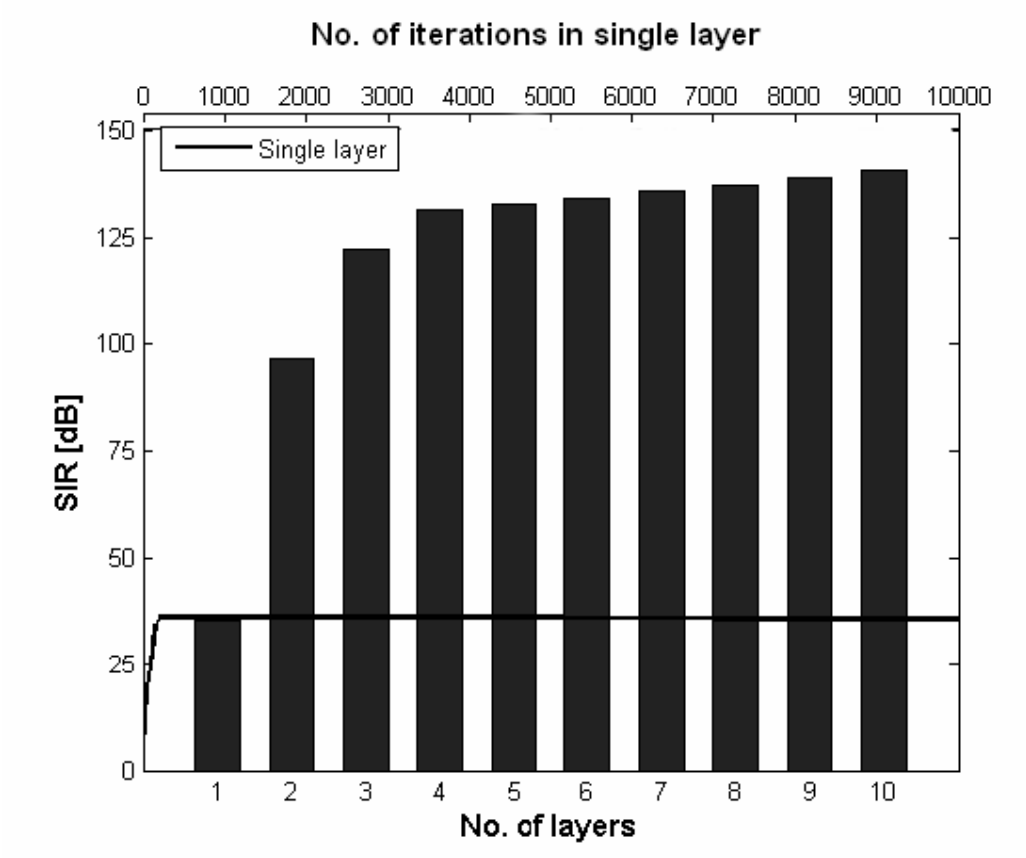

Fig.2. Performance of standard and multilayer NMF. Bars present SIRs of estimated sources in each layer. Solid line shows SIRs versus number of iterations for single layer.