Gibbs regularized nonnegative matrix factorization for blind separation of locally smooth signals

Article · January 2007

CITATIONS
5

READS
81

2 authors:

Rafal Zdunek
Wroclaw University of Science and Technology
138 PUBLICATIONS **4,957** CITATIONS
SEE PROFILE

Andrzej Cichocki
Skolkovo Institute of Science and Technology
1,040 PUBLICATIONS **39,957** CITATIONS
SEE PROFILE

Gibbs regularized nonnegative matrix factorization for blind separation of locally smooth signals

Rafal Zdunek*,† and Andrzej Cichocki*,‡
*Laboratory for Advanced Brain Signal Processing, Brain Science Institute, RIKEN, Wako-shi, Saitama, JAPAN
Emails: {zdunek, a.cichocki}@brain.riken.jp
† Institute of Telecommunications, Teleinformatics, and Acoustics, Wroclaw University of Technology, POLAND
‡ Warsaw University of Technology, POLAND

***Abstract*— Nonnegative Matrix Factorization (NMF) has already found many applications in signal processing and data analysis. One of them is blind separation of images or nonnegative signals. In the paper, we propose to improve the selected NMF algorithms to be more efficient for blind separation of locally smooth nonnegative signals. Our modifications are related with incorporation of the Gibbs prior, which is well-known in many tomographic image reconstruction applications, to a underlying blind source separation model. Our considerations are confirmed with the numerical tests.**

I. Introduction

Nonnegative Matrix Factorization (NMF) attempts to recover hidden nonnegative structures or patterns from usually redundant data. This technique has been successfully applied in many applications, e.g. in data analysis (pattern recognition, segmentation, clustering, dimensionality reduction) [1]–[12], signal and image processing (blind source separation, spectra recovering) [13], language modeling, text analysis [14], music transcription [3], [15], or neuro-biology (gene separation) [16], [17].

NMF decomposes the data matrix $\boldsymbol{Y} = [y_{ik}] \in \mathbb{R}^{I \times K}$ as a product of two nonnegative matrices $\boldsymbol{A} = [a_{ij}] \in \mathbb{R}^{I \times J}$ and $\boldsymbol{X} = [x_{jk}] \in \mathbb{R}^{J \times K}$, i.e.

$$\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{X}, \tag{1}$$

where $\forall i, j, k : a_{ij} \geq 0,\ x_{jk} \geq 0$.

Depending on an application, the hidden components may have different interpretation. For example, Lee and Seung in [4] 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 (BSS) [18], the matrix $\boldsymbol{Y}$ represents the observed mixed (superposed) signals or images, $\boldsymbol{A}$ is a mixing operator, and $\boldsymbol{X}$ is a matrix of true source signals or images. Each row of $\boldsymbol{Y}$ or $\boldsymbol{X}$ is a signal or 1D image representation, where I is a number of observed mixed signals and J is a number of hidden (source) components. The index k usually denotes a sample (discrete time instant), where K is a number of samples. In BSS, we usually have $K >> I \geq J$, and J is known or can be relatively easily estimated using SVD.

Our objective is to estimate the mixing matrix $\boldsymbol{A}$ and sources $\boldsymbol{X}$ subject to nonnegativity constraints of all the entries, given $\boldsymbol{Y}$ and possibly the prior knowledge on the nature of the true signals to be estimated or on a statistical distribution of noisy disturbances.

The basic approach to NMF is the alternating minimization of the specific cost function $D(\boldsymbol{Y}||\boldsymbol{A}\boldsymbol{X})$ that measures the distance between $\boldsymbol{Y}$ and $\boldsymbol{A}\boldsymbol{X}$. Lee and Seung [4] were the first who proposed two types of NMF algorithms. One minimizes the Euclidean distance, which is optimal for a Gaussian distributed additive noise, and the other for minimization of the Kullback-Leibler divergence, which is suitable for a Poisson distributed noise. The NMF algorithms that are optimal for many other distribution of additive noise can be found, e.g. in [18]–[20].

Unfortunately, the alternating minimization does not provide a unique solution, and often some additional constraints must be imposed to select a solution that is close to the true one. For example, finding such $\boldsymbol{P} > 0$ for which $\boldsymbol{P}^{-1} > 0$, we have: $\boldsymbol{A}\boldsymbol{X} = (\boldsymbol{A}\boldsymbol{P}^{-1})(\boldsymbol{P}\boldsymbol{X}) = \tilde{\boldsymbol{A}}\tilde{\boldsymbol{X}} = \boldsymbol{Y}$, where $\boldsymbol{A} \neq \tilde{\boldsymbol{A}}$ and $\boldsymbol{X} \neq \tilde{\boldsymbol{X}}$. Obviously, $\boldsymbol{P}$ could be any permutation matrix. Also, the alternating minimization is not convex with respect to both sets of the arguments $\{\boldsymbol{A}, \boldsymbol{X}\}$, even though the cost function is expressed by a quadratic function. To relax the ambiguity and non-convexity effects, the common approach is to incorporate some penalty terms to the cost function, which adequately regularizes the solution or restricts a set of all admissible solutions. Such regularization has been widely discussed in the literature with respect to various criteria for selection of the desired solution. The penalty terms can enforce sparsity, smoothness, continuity, closure, unimodality, orthogonality, or local rank-selectivity. A widely-used approach in many NMF applications is to apply sparsity constraints [19], [21]–[24].

In the paper, we apply the penalty term that enforces local smoothness in time-series representations of the estimated signals. This case may take place in many BSS applications with continuous and slowly-varying underlying processes, e.g. half-rectified harmonic signals.

The penalty term, which we use in the paper, is motivated by the Markov Random Field (MRF) models that are widely applied in image reconstruction. Such models, which are often expressed by the Gibbs prior, determine local roughness (smoothness) in the analyzed image with consideration of pair-wise interactions among adjacent pixels in a given neighborhood of a singe pixel. By analogy, the MRF model for time-series provides the information on pair-wise interactions

between adjacent samples. Thus, a total smoothness in time-series representations can be expressed by a joint Gibbs distribution with a nonlinear energy function. In our approach, we use the Green's function for measuring strength of the pair-wise sample interactions. Using a Bayesian framework, we get the Gibbs regularized Euclidean cost function that is minimized with a gradient descent alternating minimization technique subject to nonnegativity constrains that can be imposed in many ways. One of them is achieved with standard multiplicative updates that were used, e.g. by Lee and Seung [4]. Another approach is to apply the projected Alternating Least Squares (ALS) algorithms [24], which are generally more efficient to NMF problems than standard multiplicative algorithms.

II. Gibbs Regularized Algorithms

Since in practice a Gaussian noise occurs the most often in BSS applications, we restrict our considerations only to the following joint multivariate normal likelihood model:

$$p(\boldsymbol{Y}|\boldsymbol{X}) \propto \exp\left\{-\frac{1}{2}\,\mathrm{tr}\{(\boldsymbol{Y}-\boldsymbol{A}\boldsymbol{X})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{Y}-\boldsymbol{A}\boldsymbol{X})\}\right\}, \quad (2)$$

where each sample $\boldsymbol{y}_k$ from $\boldsymbol{Y}=[\boldsymbol{y}_1,\ldots,\boldsymbol{y}_K]$ is assumed to follow the same statistics with the covariance matrix $\boldsymbol{\Sigma}$.

Let us assume the prior information on total smoothness of the estimated signals is given by the following Gibbs distribution

$$p(\boldsymbol{X}) = \frac{1}{Z}\exp\left\{-\alpha U(\boldsymbol{X})\right\}, \quad (3)$$

where Z is a partition function, α is a regularization parameter, and $U(\boldsymbol{X})$ is a total energy function that measures the total roughness in the object of interest. The function $U(\boldsymbol{X})$ is often formulated with respect to the Markov Random Field (MRF) model that is often used in image reconstruction to enforce local smoothing.

The prior can be incorporated to the likelihood function with a Bayesian framework:

$$p(\boldsymbol{X}|\boldsymbol{Y}) = \frac{p(\boldsymbol{Y}|\boldsymbol{X})p(\boldsymbol{X})}{p(\boldsymbol{Y})}, \quad (4)$$

where $p(\boldsymbol{Y})$ is a marginal likelihood function. Thus the Gibbs regularized Euclidean cost function can be expressed in the form:

$$\Psi = -2\ln p(\boldsymbol{X}|\boldsymbol{Y}) = ||\boldsymbol{Y}-\boldsymbol{A}\boldsymbol{X}||_F^2 + 2\alpha U(\boldsymbol{X}) + c, \quad (5)$$

where c is a constant.

The stationary points of Ψ can be derived from the gradients of Ψ with respect to $\boldsymbol{X}$ and $\boldsymbol{A}$. Thus:

$$\nabla_X\Psi = 2\boldsymbol{A}^T(\boldsymbol{A}\boldsymbol{X}-\boldsymbol{Y}) + 2\alpha\nabla_X U(\boldsymbol{X}) \equiv 0, \quad (6)$$

$$\nabla_A\Psi = (\boldsymbol{A}\boldsymbol{X}-\boldsymbol{Y})\boldsymbol{X}^T \equiv 0. \quad (7)$$

1) NMF Algorithms: From (6)–(7), we have:

$$\frac{[\boldsymbol{A}^T\boldsymbol{Y}-\alpha\nabla_X U(\boldsymbol{X})]_{jk}}{[\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{X}]_{jk}} = 1, \quad \frac{[\boldsymbol{Y}\boldsymbol{X}^T]_{ij}}{[\boldsymbol{A}\boldsymbol{X}\boldsymbol{X}^T]_{ij}} = 1. \quad (8)$$

Using multiplicative updates, we get the Gibbs regularized multiplicative NMF algorithm:

$$x_{jk} \leftarrow x_{jk}\,\frac{\left[[\boldsymbol{A}^T\boldsymbol{Y}]_{jk}-\alpha[\nabla_X U(\boldsymbol{X})]_{jk}\right]_\varepsilon}{[\boldsymbol{A}^T\boldsymbol{A}\,\boldsymbol{X}]_{jk}}, \quad (9)$$

$$a_{ij} \leftarrow a_{ij}\,\frac{[\boldsymbol{Y}\,\boldsymbol{X}^T]_{ij}}{[\boldsymbol{A}\,\boldsymbol{X}\,\boldsymbol{X}^T]_{ij}}, \quad (10)$$

$$a_{ij} \leftarrow \frac{a_{ij}}{\sum_{j=1}^J a_{ij}}, \quad (11)$$

where

$$[x]_\varepsilon = \max\{\varepsilon, x\} \quad (12)$$

is a nonlinear operator of the projection onto a positive orthant (subspace $\mathbb{R}_+$) with small ε (*eps*) Typically, $\varepsilon = 10^{-16}$. The normalization (11) additionally constrains the basis vectors to a unit l_1-norm, which relaxes the intrinsic scaling ambiguity in NMF.

It is easy to notice that for $\alpha = 0$ in (9), the updating rules (9)–(10) simplify to the standard Lee-Seung algorithm that minimizes the Euclidean distance (Frobenius norm).

The algorithm (9)–(11) can be also improved by replacing the step (10) with a more exact updating rule. It is well-known that multiplicative algorithms are slowly-convergent, and the system of linear equations to be solved in the step (10) is highly over-determined. Hence, the update (10) can be successfully replaced with the projected Moore-Penrose pseudo-inverse [24] or the quasi-Newton approach [23]. For simplicity, we consider only the former approach, thus from (7) we have

$$\boldsymbol{A} \leftarrow \left[\boldsymbol{Y}\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{X}^T)\right]_\varepsilon. \quad (13)$$

2) Markov Random Field Model: MRF models have been widely applied in many image reconstruction applications, especially in tomographic imaging. In our application, MRF models motives the definition of the total energy function in the Gibbs prior (3). Thus

$$U(\boldsymbol{X}) = \sum_{j=1}^{J}\sum_{k=1}^{K}\sum_{l\in S_k} w_{kl}\,\psi\left(x_{jk}-x_{jl},\delta\right), \quad (14)$$

where S_k is a set of pixels in the neighborhood of sample k, w_{kl} is a weighting factor, δ is a scaling factor, and $\psi\left(\xi,\delta\right)$ is some potential function of ξ, which can take different forms. For example, in our application we can use the functions proposed by the following scientists: Besag (Laplacian) [25], Hebert and Leahy [26], Geman and McClure [27], Geman and Reynolds [28], Stevenson and Delp (Hubert) [29], or Green [30]. Also, a typical Gaussian function may be also useful.

Since the Green's function [30] satisfies all the properties mentioned in [31], i.e. it is nonnegative, even, 0 at $\xi = 0$, strictly increasing for $\xi > 0$, unbounded, convex, and has bounded first-derivative, we decided to select this function to our tests. Thus

$$\psi(\xi, \delta) = \delta \log[\cosh(\xi/\delta)], \tag{15}$$

which leads to

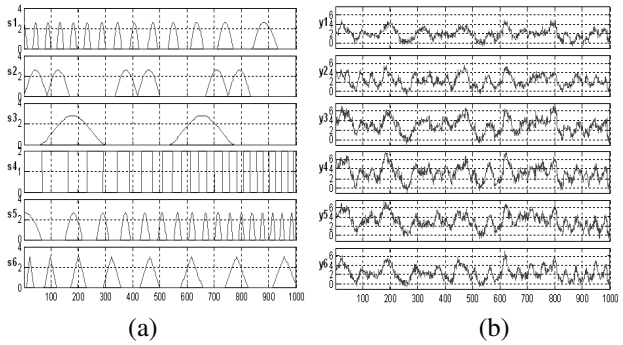
$$[\nabla_X U(\boldsymbol{X})]_{jk} = \sum_{l \in S_k} w_{kl} \tanh\left(\frac{x_{jk} - x_{jl}}{\delta}\right). \tag{16}$$

The set S_k and the associated weighting factors w_{kl} can be defined in many ways; usually, the nearest neighborhood is only taken into account, and w_{kl} are defined by the MRF model. Considering the nearest neighborhood in our application, we have $S_k = \{k-1, k+1\}$, and $w_{kl} = 1$. However, there is also possible to include another samples to enforce smoothness in a wider window. As example, we can take $S_k = \{k-2, k-1, k+1, k+2\}$ and $\forall l \in S_k : w_{kl} = 1$, or $w_{kl} = 1$ for $l = k-1$ and $l = k+1$, and $w_{kl} = 0.5$ for $l = k-2$ and $l = k+2$. The latest settings considerably favor the nearest neighborhood.

Usually, the potential functions in (14) are parameter-dependent. At least, one parameter (in our case, the parameter δ) must be set up in advance, or simultaneously with the estimation. Generally, this can be regarded as a hyperparameter, and consequently estimated with maximization of the marginal likelihood function $p(\boldsymbol{Y})$ in (4). However, a direct estimation of the parameter from the data usually involves a high computational complexity, and it is not absolutely needed if we operate on one class of data for which preliminary simulations can be performed. We notice that for our class of data, the parameter has a very slight impact on the estimation in quite a wide range of its values. Thus, we set $\delta = 10^{-3}$ in all the tests in the paper.

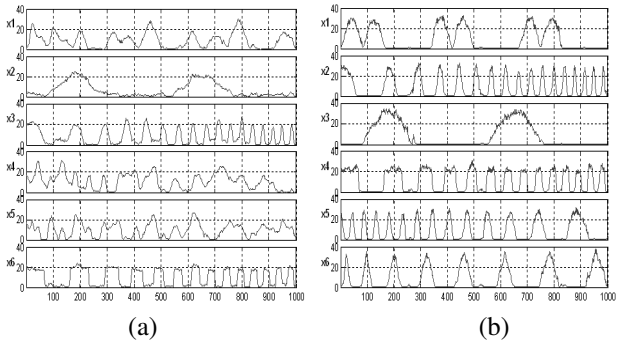
III. Numerical Tests

The proposed algorithms have been extensively tested for various sets of the parameters (α and δ), and the algorithms are compared with the standard NMF algorithm. For the numerical tests we used the benchmark of 6 smooth original signals (Fig. 1(a)) which were mixed with the dense random mixing matrix $\boldsymbol{A} \in \mathbb{R}^{30\times 6}$ with a uniform distribution. The mixtures were then corrupted with the noise of $SNR = 10$[dB]. Fig. 1(b) presents the selected mixed signals. The estimated signals with the standard Lee-Seung algorithm (the updates (9)–(11) at $\alpha = 0$) are shown in Fig. 2(a). The results obtained with the improved Gibbs regularized NMF algorithm: $\boldsymbol{X}$ updated with (9) and $\boldsymbol{A}$ updated with (13) with normalization (11) are illustrated in Fig. 2(b) for $\alpha = 0.28$. The estimations are also quantitatively assessed with the standard Signal-to-Interference Ratio (SNR). The same algorithms are also tested with the Monte Carlo (MC) analysis where for each run the initial conditions are randomly set. Fig. 3 presents the histograms obtained from 100 mean-SIR samples generated with the MC analysis for the above-mentioned NMF



(a) (b)

Fig. 1. (a) Original 6 smooth source signals; (b) Selected observations (totally 30 very noisy mixed signals with $SNR = 10$[dB]).



(a) (b)

Fig. 2. Estimated sources with: (a) standard Lee-Seung algorithm (9)–(11) at $\alpha = 0$ ($SIR_X = 1.1, 2.4, 11.8, 12.3, 10.3, 4.4$[dB], respectively); (b) Gibbs regularized algorithm: $\boldsymbol{X}$ with (9), $\boldsymbol{A}$ with (13), normalization (11), and parameters $\alpha = 0.28$ and $\delta = 10^{-3}$ ($SIR_X = 19, 20, 19.3, 14.2, 17.9, 18$[dB], respectively).

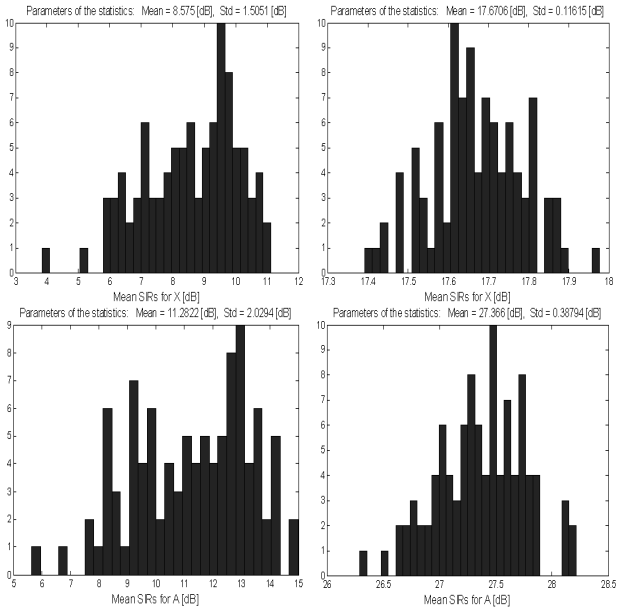


Fig. 3. Histograms from 100 mean-SIR samples generated with the following algorithms: (left) standard Lee-Seung algorithm; (right) Gibbs regularized algorithm; (top) estimation of $\boldsymbol{X}$ (sources); (bottom) estimation of columns in mixing matrix $\boldsymbol{A}$.

algorithms: unregularized version (left) and Gibbs regularized version (right).

IV. Conclusions

In the paper, we derived the new algorithm for NMF, which may be useful for estimation of locally smooth nonnegative signals in BSS applications. The algorithm exploits the information on pair-wise interactions between adjacent samples, which is motivated by MRF models in tomographic image reconstruction. Incorporating such a prior information to the NMF updating rules (especially for $\boldsymbol{X}$) is also very profitable for relaxing NMF ambiguity and non-convexity effects. The numerical results demonstrate the robustness of the proposed algorithm, especially for highly noisy data. The algorithm is much less sensitive to initialization in comparison to the standard NMF algorithms. This is confirmed with the MC simulations shown in Fig. 3. The proposed approach can be further extended with additional constraints or different updating rules. Also, another extension may concern the application of data-driven hyperparameter estimation techniques, especially for the regularization parameter.

References

[1] D. Guillamet, J. Vitrià, and B. Schiele, "Introducing a weighted nonnegative matrix factorization for image classification," *Pattern Recognition Letters*, vol. 24, no. 14, pp. 2447–2454, 2003.

[2] J.-H. Ahn, S. Kim, J.-H. Oh, and S. Choi, "Multiple nonnegative-matrix factorization of dynamic PET images," in *ACCV*, 2004.

[3] J. S. Lee, D. D. Lee, S. Choi, and D. S. Lee, "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, December 2001, pp. 556–562.

[4] D. D. Lee and H. S. Seung, "Learning the parts of objects by nonnegative matrix factorization," *Nature*, vol. 401, pp. 788–791, October 1999.

[5] H. Li, T. Adali, and D. E. W. Wang, "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.

[6] P. Carmona-Saez, R. D. Pascual-Marqui, F. Tirado, J. M. Carazo, and A. Pascual-Montano, "Biclustering of gene expression data by non-smooth non-negative matrix factorization," *BMC Bioinformatics*, vol. 7, no. 78, 2006.

[7] A. Pascual-Montano, J. M. Carazo, K. Kochi, D. Lehmean, and R. Pacual-Marqui, "Nonsmooth nonnegative matrix factorization (nsNMF)," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 28, no. 3, pp. 403–415, 2006.

[8] F. Shahnaz, M. Berry, P. Pauca, and R. Plemmons, "Document clustering using non-negative matrix factorization," *Journal on Information Processing and Management*, vol. 42, pp. 373–386, 2006.

[9] O. Okun and H. Priisalu, "Fast nonnegative matrix factorization and its application for protein fold recognition," *EURASIP Journal on Applied Signal Processing*, vol. 2006, pp. Article ID 71 817, 8 pages, 2006.

[10] Y. Wang, Y. Jia, C. Hu, and M. Turk, "Non-negative matrix factorization framework for face recognition," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 19, no. 4, pp. 495–511, 2005.

[11] W. Liu and N. Zheng, "Non-negative matrix factorization based methods for object recognition," *Pattern Recognition Letters*, vol. 25, no. 8, pp. 893–897, 2004.

[12] M. W. Spratling, "Learning image components for object recognition," *Journal of Machine Learning Research*, vol. 7, pp. 793–815, 2006.

[13] P. Sajda, S. Du, T. R. Brown, R. S. D. C. Shungu, X. Mao, and L. C. Parra, "Nonnegative matrix factorization for rapid recovery of constituent spectra in magnetic resonance chemical shift imaging of the brain," *IEEE Trans. Medical Imaging*, vol. 23, no. 12, pp. 1453–1465, 2004.

[14] I. S. Dhillon and D. M. Modha, "Concept decompositions for large sparse text data using clustering," *Machine Learning J.*, vol. 42, pp. 143–175, 2001.

[15] Y. C. Cho and S. Choi, "Nonnegative features of spectro-temporal sounds for classification," *Pattern Recognition Letters*, vol. 26, pp. 1327–1336, 2005.

[16] J.-P. Brunet, P. Tamayo, T. R. Golub, and J. P. Mesirov, "Metagenes and molecular pattern discovery using matrix factorization," vol. 101, no. 12. PNAS, 2000, pp. 4164–4169.

[17] N. Rao, S. J. Shepherd, and D. Yao, "Extracting characteristic patterns from genome – wide expression data by non-negative matrix factorization," in *Proc. of the 2004 IEEE Computational Systems Bioinformatics Conference (CSB 2004)*, Stanford, CA, August 2004.

[18] A. Cichocki, R. Zdunek, and S. Amari, "Csiszar's divergences for non-negative matrix factorization: Family of new algorithms," *Springer LNCS*, vol. 3889, pp. 32–39, 2006.

[19] I. Dhillon and S. Sra, "Generalized nonnegative matrix approximations with Bregman divergences," in *Neural Information Proc. Systems*, Vancouver, Canada, December 2005.

[20] R. Kompass, "A generalized divergence measure for nonnegative matrix factorization," *Neural Computation*, vol. 19, no. 3, pp. 780–791, 2006.

[21] P. Hoyer, "Non-negative matrix factorization with sparseness constraints," *Journal of Machine Learning Research*, vol. 5, pp. 1457–1469, 2004.

[22] K. Kreutz-Delgado, J. F. Murray, B. D. Rao, K. Engan, T.-W. Lee, and T. J. Sejnowski, "Dictionary learning algorithms for sparse representation," *Neural Computation*, vol. 15, no. 2, pp. 349–396, 2003.

[23] R. Zdunek and A. Cichocki, "Nonnegative matrix factorization with constrained second-order optimization," *Signal Processing*, 2007, in press.

[24] A. Cichocki and R. Zdunek, "Regularized alternating least squares algorithms for non-negative matrix/tensor factorization," *Springer LNCS*, vol. 4493, pp. 793–802, 2007.

[25] J. Besag, "Toward Bayesian image analysis," *J. Appl. Stat.*, vol. 16, pp. 395–407, 1989.

[26] T. Hebert and R. Leahy, "A generalized EM algorithm for 3-D Bayesian reconstruction form Poisson data using Gibbs priors," *IEEE Trans. Medical Imaging*, vol. 8, pp. 194–202, 1989.

[27] S. Geman and D. McClure, "Statistical methods for tomographic image reconstruction," *Bull. Int. Stat. Inst.*, vol. LII-4, pp. 5–21, 1987.

[28] S. Geman and G. Reynolds, "Constrained parameters and the recovery of discontinuities," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 14, pp. 367–383, 1992.

[29] R. Stevenson and E. Delp, "Fitting curves with discontinuities," in *Proc. 1-st Int. Workshop on Robust Computer Vision*, Seattle, Wash., USA, Oct. 1–3 1990.

[30] P. J. Green, "Bayesian reconstruction from emission tomography data using a modified EM algorithm," *IEEE Trans. Medical Imaging*, vol. 9, pp. 84–93, 1990.

[31] K. Lange and R. Carson, "EM reconstruction algorithms for emission and transmission tomography," *J. Comp. Assisted Tomo.*, vol. 8, no. 2, pp. 306–316, 1984.