NEW ALGORITHMS FOR NON-NEGATIVE MATRIX FACTORIZATION IN APPLICATIONS TO BLIND SOURCE SEPARATION

Andrzej CICHOCKI[*]*, Rafal ZDUNEK*[†]*, and Shun-ichi AMARI*

RIKEN Brain Science Institute,
Wako-shi, JAPAN, a.cichocki@riken.jp

ABSTRACT

In this paper we develop several algorithms for non-negative matrix factorization (NMF) in applications to blind (or semi blind) source separation (BSS), when sources are generally statistically dependent under conditions that additional constraints are imposed such as nonnegativity, sparsity, smoothness, lower complexity or better predictability. We express the non-negativity constraints using a wide class of loss (cost) functions, which leads to an extended class of multiplicative algorithms with regularization. The proposed relaxed forms of the NMF algorithms have a higher convergence speed with the desired constraints. Moreover, the effects of various regularization and constraints are clearly shown. The scope of the results is vast since the discussed loss functions include quite a large number of useful cost functions such as weighted Euclidean distance, relative entropy, Kullback Leibler divergence, and generalized Hellinger, Pearson's, Neyman's distances, etc.

1. INTRODUCTION AND PROBLEM FORMULATION

Many problems in signal and image processing can be expressed in terms of matrix factorizations. Different cost functions and imposed constraints may lead to different types of matrix factorization. In this paper we impose nonnegativity constraints and other constraints such as sparseness and smoothness. Non-negative matrix factorization (NMF) decomposes the data matrix $\boldsymbol{Y} = [\boldsymbol{y}(1), \boldsymbol{y}(2), \ldots, \boldsymbol{y}(N)] \in \mathbb{R}^{m \times N}$ as a product of two matrices $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ and $\boldsymbol{X} = [\boldsymbol{x}(1), \boldsymbol{x}(2), \ldots, \boldsymbol{x}(N)] \in \mathbb{R}^{n \times N}$ having only non-negative elements. Although some decompositions or matrix factorizations provide an exact reconstruction data (i.e., $\boldsymbol{Y} = \boldsymbol{AX}$), we shall consider here the decompositions which are approximative in nature, i.e.,

$$\boldsymbol{Y} = \boldsymbol{AX} + \boldsymbol{V} \tag{1}$$

or equivalently in a scalar form as $y_i(k) = y_{ik} = \sum_{j=1}^{n} a_{ij} x_j(k) + \nu_i(k), \;\; i = 1, \ldots, m,$ where $\boldsymbol{V} \in \mathbb{R}^{m \times N}$ represents a noise or error matrix, $\boldsymbol{y}(k) = [y_1(k), \ldots, y_m(k)]^T$ is a vector of the observed signals at the discrete time instants[1] k, while $\boldsymbol{x}(k) = [x_1(k), \ldots, x_n(k)]^T$ is a vector of components or source signals at the same time instant [1, 2].

Our objective is to estimate the mixing (basis) matrix $\boldsymbol{A}$ and sources $\boldsymbol{X}$, subject to nonnegativity constraints. Usually, in BSS applications $N >> m \geq n$ and n is known or can be estimated using SVD. We use the following notations: $x_j(k) = x_{jk}$, $y_i(k) = y_{ik}$, $z_{ik} = [\boldsymbol{AX}]_{ik}$ means ik-element of the matrix $\boldsymbol{Z} = \boldsymbol{AX}$.

Unlike the other matrix factorizations, NMF permits the combination of multiple basis signals to represent original signals. But only additive combinations are allowed, because the nonzero elements of $\boldsymbol{A}$ and $\boldsymbol{X}$ are all positive. Thus in such decomposition no subtractions can occur. For these reasons, the non-negativity constraints are compatible with the intuitive notion of combining components to form a whole signal or image, which is how NMF learns a parts-based representation [3, 4, 5, 2].

The NMF (Non-negative Matrix Factorization) sometimes called also PMF (Positive Matrix Factorization)does not assume explicitly or implicitly sparseness or mutual statistical independence of components, however usually provides sparse decomposition [3]. The NMF found many applications in spectroscopy, chemometrics and environmental science where the matrices have clear physical meanings and some normalization are imposed to them (for example, the matrix $\boldsymbol{A}$ has columns normalized to unit length) [4, 6, 7]

The NMF method is designed to capture alternative structures inherent in the data, and possibly to provide more biological insight. Lee and Seung introduced NMF in its modern formulation as a method to decompose images [3]. For example, in this context, NMF yielded a decomposition of human faces into parts reminiscent of features such as lips, eyes, nose, etc.

The most of known and used adaptive multiplicative algorithms for NMF are based on two cost functions: Square Euclidean distance expressed by Frobenius norm $D_F(\boldsymbol{A}, \boldsymbol{X}) = \frac{1}{2}\|\boldsymbol{Y} - \boldsymbol{AX}\|_F^2$ which is optimal for Gaussian distributed

[*] On leave from Warsaw University of Technology, Poland

[†] On leave from Institute of Telecommunications, Teleinformatics and Acoustics, Wroclaw University of Technology, Poland

[1] The data are often represented not in the time domain but in a transform domain such as the time frequency domain, so index k may have different meaning.

noise, and the generalized Kullback-Leibler divegence

$$D_{KL}(\boldsymbol{A},\boldsymbol{X}) = \sum_{i,k}(y_{ik}\log(y_{ik}/[\boldsymbol{A}\boldsymbol{X}]_{ik}) + [\boldsymbol{A}\boldsymbol{X}]_{ik} - y_{ik})$$

which is related the Poisson likelihood. The most existing NMF algorithms perform blind source separation rather very poorly due to non-uniqueness of solution and/or lack of additional constraints which should be satisfied. The main objective of this contribution is to propose a flexible NMF approach and generalize or combine several different criteria in order to extract physically meaningful sources from their linear mixtures and noise. Whereas most applications of NMF focused on grouping elements of images into parts (using the matrix $\boldsymbol{A}$), we take the dual viewpoint by focusing primarily on grouping samples into components representing by the matrix $\boldsymbol{X}$ of source signals.

2. MULTIPLICATIVE NMF ALGORITHMS WITH REGULARIZATION AND SPARSITY CONSTRAINTS

Although standard NMF (without any auxiliary constraints) provides sparseness of its component, we can achieve some control of this sparsity as well as smoothness of components by imposing additional constraints to natural non-negativity constraints. In fact, we can incorporate smoothness or sparsity constraints in several ways [5]. One of the simples approach is to implement in each iteration step a nonlinear projection which can increase sparseness and/or smoothness of the estimated components. An alternative approach is to add to the loss function suitable regularization or penalty terms. Let us consider the following constrained optimization problem:
Minimize:

$$D_{F\alpha}(\boldsymbol{A},\boldsymbol{X}) = \frac{1}{2}\|\boldsymbol{Y} - \boldsymbol{A}\boldsymbol{X}\|_F^2 + \alpha_A J_A(\boldsymbol{A}) + \alpha_X J_X(\boldsymbol{X})$$

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

where α_A and $\alpha_X \geq 0$ are nonnegative regularization parameters and terms $J_X(\boldsymbol{X})$ and $J_A(\boldsymbol{A})$ are used to enforce a certain application-dependent characteristic of the solution. As special practical case, we have $J_X(\boldsymbol{X}) = \sum_{jk} f_X(x_{jk}) = \sum_{jk} x_{jk}$ where $f(\cdot)$ are suitably chosen functions which are measure of smoothness or sparsity. In order to achieve sparse representation we usually chose $f(x_j) = |x_j|$ or simply $f(x_j)= x_j$ or alternatively $f(x_j) = x_j \log(x_j)$ with constraints $x_j \geq 0$. Similar regularization terms can be implemented also for the matrix $\boldsymbol{A}$. Note, that we treat both matrices $\boldsymbol{A}$ and $\boldsymbol{X}$ in a symmetric way. Applying the standard gradient descent approach, we have

$$\begin{aligned} a_{ij} &\leftarrow a_{ij} - \eta_{ij}\frac{\partial J_{F\alpha}(\boldsymbol{A},\boldsymbol{X})}{\partial a_{ij}}, \\ x_{jk} &\leftarrow x_{jk} - \eta_{jk}\frac{\partial J_{F\alpha}(\boldsymbol{A},\boldsymbol{X})}{\partial x_{jk}}, \end{aligned}$$

where η_{ij} and η_{jk} are positive learning rates. The gradient components can be expressed in compact matrix forms as:

$$\begin{aligned} \frac{\partial D_{F\alpha}(\boldsymbol{A},\boldsymbol{X})}{\partial a_{ij}} &= [-\boldsymbol{Y}\boldsymbol{X}^T + \boldsymbol{A}\boldsymbol{X}\boldsymbol{X}^T]_{ij} + \alpha_A\frac{\partial J_A(\boldsymbol{A})}{\partial a_{ij}} \\ \frac{\partial D_{F\alpha}(\boldsymbol{A},\boldsymbol{X})}{\partial x_{jk}} &= [-\boldsymbol{A}^T\boldsymbol{Y} + \boldsymbol{A}^T\boldsymbol{A}\boldsymbol{X}]_{jk} + \alpha_X\frac{\partial J_X(\boldsymbol{X})}{\partial x_{jk}} \end{aligned}$$

Here, we follow the Lee and Seung's proposition to choose specific learning rates [3, 6]

$$\eta_{ij} = \frac{a_{ij}}{[\boldsymbol{A}\boldsymbol{X}\boldsymbol{X}^T]_{ij}}, \quad \eta_{jk} = \frac{x_{jk}}{[\boldsymbol{A}^T\boldsymbol{A}\,\boldsymbol{X}]_{jk}}, \tag{3}$$

which leads to a generalized robust multiplicative update rules:

$$a_{ij} \leftarrow a_{ij}\,\frac{\Big[\boldsymbol{Y}\,\boldsymbol{X}^T]_{ij} - \alpha_A\varphi_A(a_{ij})\Big]_\varepsilon}{[\boldsymbol{A}\,\boldsymbol{X}\,\boldsymbol{X}^T]_{ij} + \varepsilon} \tag{4}$$

$$x_{jk} \leftarrow x_{jk}\,\frac{\Big[[\boldsymbol{A}^T\,\boldsymbol{Y}]_{jk} - \alpha_X\,\varphi_X(x_{jk})\Big]_\varepsilon}{[\boldsymbol{A}^T\boldsymbol{A}\,\boldsymbol{X}]_{jk} + \varepsilon}, \tag{5}$$

where the nonlinear operator is defined as $[x]_\varepsilon = \max\{\varepsilon, x\}$ with small ε, and the functions $\varphi_A(a_{ij})$ and $\varphi_X(x_{jk})$ are defined as

$$\varphi_A(a_{ij}) = \frac{\partial J_A(\boldsymbol{A})}{\partial a_{ij}}, \quad \varphi_X(x_{jk}) = \frac{\partial J_X(\boldsymbol{X})}{\partial x_{jk}}. \tag{6}$$

Typically, $\varepsilon = 10^{-9}$ is introduced in order to ensure non-negativity constraints and to avoid possible division by zero. In the special case, by using the l_1 norm regularization terms $f(\boldsymbol{x}) = \|\boldsymbol{x}\|_1$ for both matrices $\boldsymbol{X}$ and $\boldsymbol{A}$ the above multiplicative learning rules can be simplified as follows:

$$a_{ij} \leftarrow a_{ij}\,\frac{\Big[[\boldsymbol{Y}\,\boldsymbol{X}^T]_{ij} - \alpha_A\Big]_\varepsilon}{[\boldsymbol{A}\,\boldsymbol{X}\,\boldsymbol{X}^T]_{ij} + \varepsilon} \tag{7}$$

$$x_{jk} \leftarrow x_{jk}\,\frac{\Big[[\boldsymbol{A}^T\,\boldsymbol{Y}]_{jk} - \alpha_X\Big]_\varepsilon}{[\boldsymbol{A}^T\boldsymbol{A}\,\boldsymbol{X}]_{jk} + \varepsilon} \tag{8}$$

with a normalization of columns of the matrix $\boldsymbol{A}$ in each iteration as $a_{ij} \leftarrow a_{ij}/\sum_i a_{ij}$.

The above algorithm provides a sparse representation of the estimated matrices and the sparseness measure increases with increasing values of regularization coefficients, typically $\alpha_X = 0.01 - 0.5$.

The sparsity and smoothness constraints can be extended to NMF algorithms corresponding to other loss functions and generalized divergences by adding suitable regularization terms and/or projecting data by suitable nonlinear functions. For example, for the generalized regularized Kullback-Leibler divergence (called also I-divergence):

$$\begin{aligned} D_{KL}(\boldsymbol{Y}||\boldsymbol{A}\boldsymbol{X}) &= \sum_{ik}\left(y_{ik}\,log\frac{y_{ik}}{[\boldsymbol{A}\boldsymbol{X}]_{ik}} + y_{ik} - [\boldsymbol{A}\boldsymbol{X}]_{ik}\right) \\ &\quad +\alpha_X J(\boldsymbol{X}) + \alpha_A J(\boldsymbol{A}) \end{aligned} \tag{9}$$

we developed a modified learning rule:

$$x_{jk} \leftarrow \left(x_{jk} \frac{\sum_{i=1}^{m} a_{ij} \, (y_{ik}/[\boldsymbol{A}\,\boldsymbol{X}]_{ik})}{[\sum_{q=1}^{m} a_{qj} + \alpha_X \varphi_X(\boldsymbol{X})]_\varepsilon} \right)^{1+\alpha_{sX}} \quad (10)$$

$$a_{ij} \leftarrow \left(a_{ij} \frac{\sum_{k=1}^{N} x_{jk} \, (y_{ik}/[\boldsymbol{A}\,\boldsymbol{X}]_{ik})}{[\sum_{p=1}^{N} x_{jp} + \alpha_A \varphi_A(\boldsymbol{A})]_\varepsilon} \right)^{1+\alpha_{sA}}, \quad (11)$$

where additional small positive regularization terms $\alpha_{sX} \geq 0$ and $\alpha_{sX} \geq 0$ are introduced in order to enforce sparseness of the solution, if necessary. Typical values of $\alpha_{sX} = \alpha_{sA} = 0.001 - 0.005$.

Alternative loss function which provides a wide class of flexible distance measures is the Amari's alpha divergence (see also Cressie-Read disparity family, Liese & Vajda, 1987) [1, 8] defined as

$$D_A(\boldsymbol{Y}||\boldsymbol{A}\boldsymbol{X}) = \sum_{ik} y_{ik} \frac{\left(\frac{y_{ik}}{[\boldsymbol{A}\boldsymbol{X}_{ik}]}\right)^{\beta-1} - 1}{(\beta-1)\beta} + \frac{[\boldsymbol{A}\boldsymbol{X}]_{ik} - y_{ik}}{\beta}, \quad (12)$$

which for $\beta = 0.5$ simplifies to Hellinger distance, for $\beta = 1$ converges to Kullback-Leibler distance, for $\beta = 2$ simplifies to Pearson's chi-square distance, and Neyman's chi-square distance is for $\beta = -1$. The minimization of the above loss function subject to nonnegativity constraints $x_{jk} \geq 0$ and $a_{ij} \geq 0$ using majorizing (auxiliary) cost function and alternating minimization/projection approach leads to a new generalized and universal algorithm:

$$x_{jk} \leftarrow \left(x_{jk} \left(\sum_{i=1}^{m} a_{ij} \, (y_{ik}/[\boldsymbol{A}\,\boldsymbol{X}]_{ik})^{\beta} \right)^{\omega/\beta} \right)^{1+\alpha_{sX}} \quad (13)$$

$$a_{ij} \leftarrow \left(a_{ij} \left(\sum_{k=1}^{N} x_{jk} \, (y_{ik}/[\boldsymbol{A}\,\boldsymbol{X}]_{ik})^{\beta} \right)^{\omega/\beta} \right)^{1+\alpha_{sA}} \quad (14)$$

$$a_{ij} \leftarrow a_{ij} / \sum_i a_{ij}, \quad \beta \neq 0. \quad (15)$$

The above algorithm can be described in the matrix form using MATLAB notations:

$$\boldsymbol{X} \leftarrow \left(\boldsymbol{X} .* \left(\boldsymbol{A}' * ([\boldsymbol{Y}]_\varepsilon ./ [\boldsymbol{A} * \boldsymbol{X}]_\varepsilon) .^{\beta} \right) .^{\omega/\beta} \right) .^{1+\alpha_{sX}}$$

$$\boldsymbol{A} \leftarrow \left(\boldsymbol{A} .* \left(([\boldsymbol{Y}]_\varepsilon ./ [\boldsymbol{A} * \boldsymbol{X}]_\varepsilon) .^{\beta} * \boldsymbol{X}' \right) .^{\omega/\beta} \right) .^{1+\alpha_{sA}}$$

$$\boldsymbol{A} \leftarrow \boldsymbol{A} * \mathrm{diag}(1./sum(\boldsymbol{A}, 1)). \quad (16)$$

where $\omega \in (0, 2)$ is the relaxation parameter and the operators $.*$, $./$ and $.^{\beta}$ mean componentwise multiplication, division and rising to the power β each element of a corresponding matrix, respectively. For $\beta = 0$ the NMF algorithm takes the following form:

$$x_{jk} \leftarrow x_{jk} \prod_{i=1}^{m} \left(\frac{y_{ik}}{[\boldsymbol{A}\boldsymbol{X}]_{ik}} \right)^{\omega a_{ij}}, \quad (17)$$

$$a_{ij} \leftarrow a_{ij} \prod_{k=1}^{N} \left(\frac{y_{ik}}{[\boldsymbol{A}\boldsymbol{X}]_{ik}} \right)^{\tilde{\eta}_j x_{jk}}, \quad (18)$$

where $\tilde{\eta}_j = \omega(\sum_k x_{jk})^{-1}$ and a_{ij} is normalized in each step as: $a_{ij} \leftarrow a_{ij} / \sum_p a_{pj}$.

3. SIMULATION RESULTS

The proposed NMF algorithms have been extensively tested for many difficult benchmarks for signals and images with various statistical distributions. The simulations results confirmed that the developed algorithms are efficient and stable for a wide set of parameters. However, if a number of observations is very close to a number of unknown sources the algorithms do not guarantee estimation of all the sources and sometimes stuck in local minima, especially if regularization terms are absent or they are not suitably chosen. If a number of observations is much larger than a number of sources the proposed algorithms give consistent and satisfactory results, especially when sources or mixing matrices are sparse. We found that the regularization and nonlinear projections terms play a key role in improving of performance of blind source separation. Due to limit of space we give here only one illustrative example. Example 1: Five statistically dependent nonnegative signals shown in Fig.1 (top) have been mixed by randomly generated nonnegative matrix $\boldsymbol{A} \in \mathbb{R}^{9\times 5}$. To mixture we added Gaussian noise with SNR=20 dB. The mixing signals are shown in Fig.1 (bottom). Using the known standard NMF algorithms we failed to estimate original sources. Fig. 2 (top) illustrates the results applying algorithm (13)-(15) without any regularization and nonlinear projection with $\beta = \omega = 1$ and $\alpha_{sX} = \alpha_{sXA} = 0$. Applying the regularization and nonlinear projection techniques for the same algorithm with and $\alpha_{sX} = \alpha_{sXA} = 0.005$, $\beta = 2$ and $\omega = 1.9$ we reconstructed successfully all the sources with signal to interference ratios (SIR) as follows SIR=$24, 33, 34, 19, 23$ dB as shown in Fig.2 (bottom). Similar or even slightly better performance we achieved by applying three other developed algorithms. We initialized our simulations from arbitrary nonnegative matrices (for example, elements of matrices can be uniformly distributed from 0 to 1). Iterations should be continued until the error change will be negligible small (say, less than 0.01%). Since all the presented algorithms are based on a gradient descent approach they guarantee to achieve only local minima. To address this limitation, we can repeat the procedure several times starting from different initial matrices. The global convergence is still an open issue for NMF.

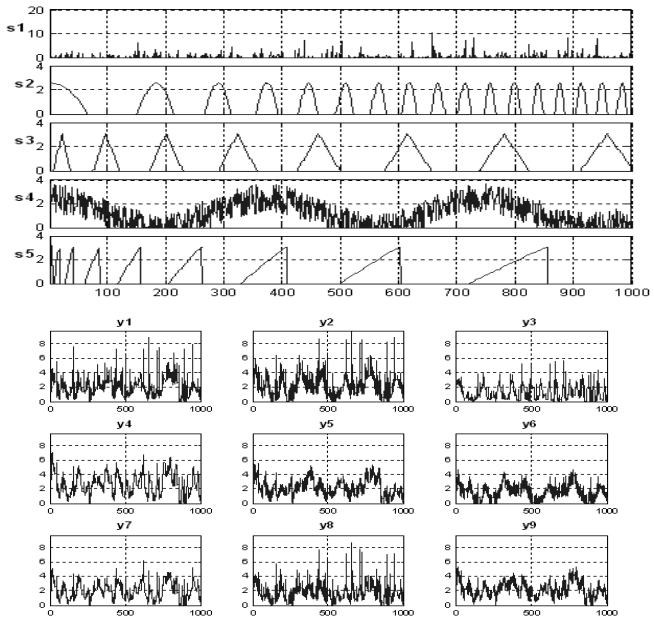

Fig. 1. Example 1, Top: original 5 source signals, and bottom: observed 9 mixed signals.

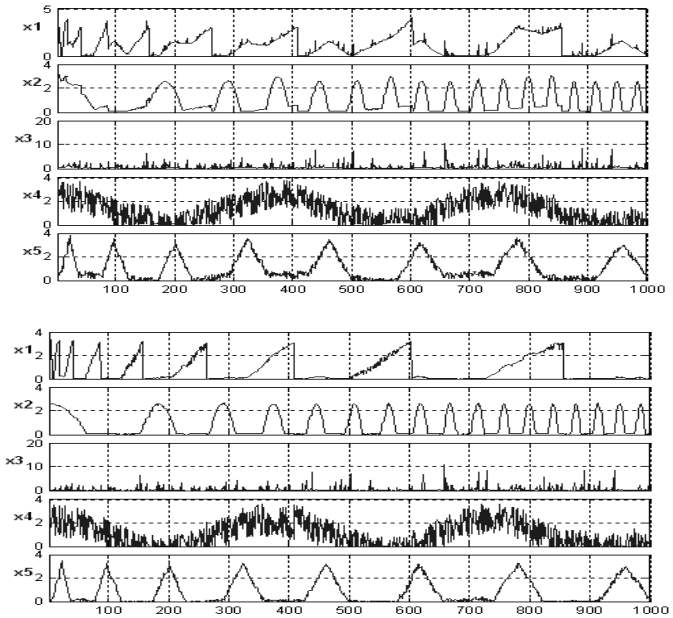

Fig. 2. Example 1, Top: Estimated sources using standard algorithm (13)-(15) for $\alpha_{sA} = \alpha_{sX} = 0$ and $\beta = \omega = 1$ with SIR= 6 dB, 17 dB, 9 dB, 24 dB, 13 dB; and bottom: Estimated source signals using the new algorithm (13)-(15) with regularization/projection $\alpha_{sX} = \alpha_{sA} = 0.005$ and $\beta = 2$, $\omega = 1.9$ with SIR=24 dB, 33 dB, 34 dB, 19 dB and 23 dB, respectively.

4. FINAL REMARKS AND CONCLUSIONS

The optimal choice of regularization parameters and β depends on the distribution of data and *a priori* knowledge about the hidden (latent) components. 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.

In summary, the proposed NMF multiplicative algorithms are efficient and robust for extracting and separation of statistically dependent sparse and/or smooth sources. However, the challenge that still remains is to prove the global convergence of such algorithms and to provide a meaningful physical interpretation to some of NMF discovered latent components or classes of components when the structures of the true sources are completely unknown.

5. REFERENCES

[1] S. Amari, *Differential-Geometrical Methods in Statistics*, Springer Verlag, 1985.

[2] A. Cichocki and S. Amari, *Adaptive Blind Signal And Image Processing (New revised and improved edition)*, John Wiley, New York, 2003.

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

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

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

[6] P. Sajda, S. Du, and L. Parra, "Recovery of constituent spectra using non-negative matrix factorization," in *Proceedings of SPIE – Volume 5207*. 2003, pp. 321–331., Wavelets: Applications in Signal and Image Processing.

[7] H. Li, T. Adali, and D. Emge 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)*. 2005, Mystic USA.

[8] N. A. Cressie and T.C.R. Read, *Goodness-of-Fit Statistics for Discrete Multivariate Data*, Springer, New York, 1988.