



# Non-negative Matrix Factorization with Orthogonality Constraints and its Application to Raman Spectroscopy

HUALIANG LI, TÜLAY ADALI AND WEI WANG

*Department of CSEE, University of Maryland Baltimore County, Baltimore, MD, USA*

DARREN EMGE

*US Army Research, Aberdeen Proving Grounds, Aberdeen, MD 21010, USA*

ANDRZEJ CICHOCKI

*Laboratory for Advanced Brain Signal Processing, Brain Science Institute, Riken 2-1 Hirosawa, Wako-shi, Saitama, 351-0198, Japan*

ANDRZEJ CICHOCKI

*Warsaw University of Technology, Warsaw, Poland*

*Received: 05 May 2006; Revised: 25 August 2006; Accepted: 18 October 2006*

**Abstract.** We introduce non-negative matrix factorization with orthogonality constraints (NMFOC) for detection of a target spectrum in a given set of Raman spectra data. An orthogonality measure is defined and two different orthogonality constraints are imposed on the standard NMF to incorporate prior information into the estimation and hence to facilitate the subsequent detection procedure. Both multiplicative and gradient type update rules have been developed. Experimental results are presented to compare NMFOC with the basic NMF in detection, and to demonstrate its effectiveness in the chemical agent detection problem.

**Keywords:** Raman spectroscopy, non-negative matrix factorization, NMFOC

## 1. Introduction

Raman spectroscopy can be used for detecting a wide range of chemicals since Raman spectrum is unique to a given compound. It can also quantify chemicals on a surface, in a liquid, or in air. The Raman effect arises when the incident light excites molecules in the sample which subsequently scatter the light. While most of the scattered light is at the same wavelength as the incident light, some is

scattered at a different wavelength. This inelastically scattered light is called the Raman scatter. It results from the change of the molecular motions. The energy difference between the incident light and the Raman scattered light is called the Raman shift. A plot of Raman intensity vs. Raman shift constitutes a Raman spectrum.

Due to the presence of background fluorescence or measurement errors in the Raman spectra data, it can be assumed that the observations is a linear mixture of several underlying chemical spectra, i.e., given observed spectra  $\mathbf{X}$ , we have

This project is supported by Edgewood Chemical Biological Center, US Army RDECOM under contract no: W91ZLK-04-P-0950.

$$\mathbf{X} = \mathbf{AS} + \mathbf{N} \quad (1)$$

where the columns of  $\mathbf{A}$  represent the contribution of chemical spectra, rows of  $\mathbf{S}$  the chemical spectra, and  $\mathbf{N}$  is the noise.

One traditional way to detect the chemicals from Raman spectra is based on linear regression which can identify single or multiple chemicals by solving for the contribution of chemical spectra [8]. The decision-making procedure evaluates signal strength, a measurement probability resulting from a statistical  $F$ -test and the relative strength of each chemical identified as present in the measurement. This approach relies on the available library of Raman spectra of the chemicals of interest. The problem with this approach is that the library spectra are not always reliable, can not reflect the variation of spectral profile due to environmental changes.

Another way for solving this detection problem is to combine a blind source separation method with a detection scheme [20]. For example, independent component analysis (ICA) can be used to identify constituent chemical spectra directly from the observed data  $\mathbf{X}$ . If the largest absolute correlation between the ICA estimates and target chemical spectrum exceeds a certain threshold, then the target present hypothesis is accepted. Instead of yielding a decision for one measurement as in [8], this approach yields one detection result for multiple measurements. The problem with this approach is that the assumption of the independence among constituent spectra is not realistic, as in many cases, they can be highly correlated. If this is true, the performance of the ICA approach may be adversely affected [21].

Since we interpret  $\mathbf{A}$  as the contributions and  $\mathbf{S}$  as the constituent spectra in the linear mixture model described above, they are non-negative in nature. Hence, non-negative matrix factorization (NMF) [9, 10] can be used to find such a linear, non-negative data representation. In contrast to ICA, the non-negativity constraints make the representation purely additive which is more meaningful for the spectroscopy application [5, 6, 13]. A complete survey of the development and use of low-rank approximate NMF algorithms is given in [1] along with applications for feature extraction and identification in fields ranging from text mining to spectral data analysis.

Unlike ICA, NMF does not impose a strong condition such as the independence but seeks a factorization that minimizes a chosen metric. Due to the nonconvex nature of the NMF cost function, additional constraints on the factorization are usually

imposed to decrease the sensitivity of NMF solutions to different initializations [10]. For example, Li et al. [11] proposed a local non-negative matrix factorization (LNMF) algorithm for learning spatially localized, parts-based subspace representation of visual patterns. Hoyer showed how explicitly incorporating the sparseness constraint improved the final decomposition [7]. In [16], a constrained non-negative matrix factorization (cNMF) algorithm is used to recover constituent spectra by including a constraint on the minimum amplitude of the recovered spectra to deal with observations having negative values. In [15], a nonsmooth constraint is imposed to control the sparseness in both the basis and the encoding vectors.

Moreover, Donoho and Stodden developed a geometric view of the underlying NMF factorization and derived geometric conditions under which the factorization is essentially unique [3]. The distance measures considered in NMF are also extended to other measures, for instance, the Csiszár's divergences in [2], the Bregman divergence in [4].

In solving for the NMF problem, the update rules given in [10] take a multiplicative form and satisfy the non-negativity constraint implicitly and elegantly. The extension of the multiplicative updates of [10] can be found, for instance, in [17], where a multiplicative update rule is proposed to solve a non-negative quadratic programming problem. However, NMF with multiplicative updates exhibits some weaknesses as well. The most important of which is related to its slow convergence behavior. As a special case of bound optimization, this behavior is investigated in [18]. Recently some other forms of update rules are studied to improve the performance, especially to increase the speed of convergence. For instance, a quasi-Newton method for NMF is considered in [22], and a projected gradient method is proposed in [7] and [12].

In this paper, we first introduce non-negative matrix factorization with orthogonality constraints (NMFOC) for extracting a certain target chemical spectrum from the mixture. It incorporates a priori information of the target chemical spectrum as the reference and introduces a measure of closeness between the recovered spectra and the reference into the NMF contrast function, facilitating the subsequent detection procedure. Two methods with different orthogonality constraints, namely the augmented NMFOC and Lagrangian NMFOC, are used to solve the constrained optimization problem, each followed by a corresponding detection scheme. For the simpler

augmented NMFOC case, we impose the constraint using both multiplicative and projected gradient updates to optimize the desired cost function. The corresponding update rules are derived for each case with discussions on their convergence properties.

The rest of the paper is organized as follows: Section 2 introduces NMF. This is followed by the formulation of NMFOC which is introduced in Section 3. NMFOC learning procedures are presented and the convergence properties of NMFOC algorithms are shown. Section 4 presents experimental results illustrating properties of NMFOC and its performance in chemical detection compared to the basic NMF.

## 2. Non-negative Matrix Factorization

Given a non-negative  $n \times m$  observation matrix  $\mathbf{X}$ , NMF determines non-negative matrices  $\mathbf{A}$  and  $\mathbf{S}$  such that

$$\mathbf{X} \approx \mathbf{AS} \quad (2)$$

where  $\mathbf{A}$  is  $n \times k$  and  $\mathbf{S}$  is  $k \times m$ , and both  $\mathbf{A}$  and  $\mathbf{S}$  have all non-negative entries. Usually  $k$  is chosen to be smaller than  $n$  or  $m$ . Since NMF enforces the non-negativity constraints on  $\mathbf{A}$  and  $\mathbf{S}$ , the constituent chemical spectra in  $\mathbf{S}$  can be combined to form observations in an intuitive, additive fashion. Different from the application of NMF on the face-related image processing [9], here each row of  $\mathbf{S}$  represents one spectrum. Each row of  $\mathbf{X}$  is a linear combination of rows of  $\mathbf{S}$  through the mixing coefficients of  $\mathbf{A}$ .

Two measures, generalized Kullback–Leibler divergence and the Euclidean distance metric, have been considered in [10] to quantify the quality of the approximation.

### 2.1. KL Divergence

The generalized Kullback–Leibler divergence of  $\mathbf{X}$  from its approximation  $\mathbf{Y} = \mathbf{AS}$ , is defined as [10]

$$D_{KL}(\mathbf{X}||\mathbf{Y}) = \sum_{i,j} \left( x_{ij} \log \frac{x_{ij}}{y_{ij}} - x_{ij} + y_{ij} \right) \quad (3)$$

As shown in [16], the minimization of this cost is equivalent to the maximization of the likelihood of generating the observations  $\mathbf{X}$  from  $\mathbf{A}$  and  $\mathbf{S}$  when  $x_{i,j}$  is assumed to be Poisson distributed with mean  $y_{i,j}$ .

From the log inequality, we know that if  $z > 0$ , then  $\log(z) \leq z - 1$ , with equality if and only if  $z = 1$ . Hence, given any positive  $x_{i,j}$  and  $y_{i,j}$ , we have

$$\log \left( \frac{y_{i,j}}{x_{i,j}} \right) \leq \frac{y_{i,j}}{x_{i,j}} - 1$$

leading to  $x_{ij} \log \left( \frac{x_{ij}}{y_{ij}} \right) - x_{ij} + y_{ij} \geq 0$ . The equality holds if and only if  $x_{ij} = y_{ij}$ , which implies that the cost function defined in Eq. (3) is lower bounded by zero, and equals zero if and only if  $\mathbf{X} = \mathbf{Y}$ .

NMF factorization as defined in [10] is a solution to the following optimization problem:

$$\begin{aligned} \min_{\mathbf{A}, \mathbf{S}} \quad & D(\mathbf{X}||\mathbf{AS}) \\ \text{such that} \quad & \mathbf{A}, \mathbf{S} \geq 0, \quad \sum_i a_{i,j} = 1 \quad \forall j. \end{aligned} \quad (4)$$

The column-wise unit sum constraint is imposed to alleviate the scale ambiguity problem.

This optimization problem can be solved using the multiplicative update rules as [10]:

$$\begin{aligned} \mathbf{S} &= \mathbf{S} \odot (\mathbf{A}^T (\mathbf{X} \oslash \mathbf{AS})) \oslash (\mathbf{A}^T \mathbf{J}), \\ \mathbf{A} &= \mathbf{A} \odot ((\mathbf{X} \oslash \mathbf{AS}) \mathbf{S}^T) \oslash (\mathbf{J} \mathbf{S}^T), \end{aligned} \quad (5)$$

where  $\mathbf{J}$  is an  $n \times m$  matrix of 1s,  $\odot$  and  $\oslash$  denote elementwise multiplication and division, respectively.

### 2.2. Euclidean Distance

The Euclidean distance metric

$$D_{EU}(\mathbf{X}||\mathbf{Y}) = \sum_{i,j} (x_{ij} - y_{ij})^2 \quad (6)$$

is also lower bounded by zero and vanishes if and only if  $\mathbf{X} = \mathbf{Y}$ . The corresponding NMF multiplicative update rules for minimizing the Euclidean distance are given as [10]:

$$\begin{aligned} \mathbf{S} &= \mathbf{S} \odot (\mathbf{A}^T \mathbf{X}) \oslash (\mathbf{A}^T \mathbf{AS}), \\ \mathbf{A} &= \mathbf{A} \odot (\mathbf{XS}^T) \oslash (\mathbf{ASS}^T). \end{aligned} \quad (7)$$

### 2.3. Projected Gradient

The multiplicative update rules given in Eqs. (5) and (7) satisfy the nonnegativity constraint naturally

but may require large number of iterations to reach a local minimum. However, if the stepsize is chosen properly, with additive projected gradient updates the same value can be achieved much faster. In [7], a combination of multiplicative and projected gradient method is shown to achieve better performance than NMF multiplicative updates on a face image database factorization problem. In [12], the same problem is investigated by using bound-constrained optimization theory. The only difference between the projected gradient and the regular gradient method is the use of an additional projection procedure. For instance, the gradient of KL divergence with respect to  $\mathbf{S}$  and  $\mathbf{A}$  can be calculated as

$$\begin{aligned}\nabla_{\mathbf{S}}(D_{KL}) &= \mathbf{A}^T \mathbf{J} - \mathbf{A}^T (\mathbf{X} \oslash \mathbf{A} \mathbf{S}), \\ \nabla_{\mathbf{A}}(D_{KL}) &= \mathbf{J} \mathbf{S}^T - (\mathbf{X} \oslash \mathbf{A} \mathbf{S}) \mathbf{S}^T,\end{aligned}\quad (8)$$

and the gradient of Euclidean distance with respect to  $\mathbf{S}$  and  $\mathbf{A}$  are

$$\begin{aligned}\nabla_{\mathbf{S}}(D_{EU}) &= \mathbf{A}^T \mathbf{A} \mathbf{S} - \mathbf{A}^T \mathbf{X}, \\ \nabla_{\mathbf{A}}(D_{EU}) &= \mathbf{A} \mathbf{S} \mathbf{S}^T - \mathbf{X} \mathbf{S}^T.\end{aligned}\quad (9)$$

If regular gradient updates are used, there is no guarantee that all entries of  $\mathbf{A}$  and  $\mathbf{S}$  will be nonnegative in all iterations. When it is negative, we may project the updated  $\mathbf{A}$  and  $\mathbf{S}$  to be nonnegative. Such a method is a special case of the application of bound-constrained optimization theory.

Observing the update rules given in Eqs. (5) and (7), if  $\mathbf{A}$  and  $\mathbf{S}$  are invariant under these update rules, it can easily be shown that this implies a zero-gradient condition [10, 12].

In this paper, we impose an orthogonality constraint using multiplicative and projected gradient separately, and derive the corresponding learning rules.

### 3. Non-negative Matrix Factorization with Orthogonality Constraints

As we noted earlier, NMF does not impose a strong condition such as independence in ICA, and hence multiple solutions that are significantly different from each other are likely. This immediately suggests that a more direct control over the estimated factorization is needed. In NMFOC, we incorporate

the available information of the target spectrum through orthogonality constraints to guide the factorization. Since we are interested in detecting a single target, it is sufficient to estimate just two sources. Thus we can specify the source matrix in Eq. (2) as  $\mathbf{S} = [\mathbf{s}_1 \ \mathbf{s}_2]^T$ .

We define an orthogonality measure  $\psi(\cdot)$  for two given vectors  $\mathbf{u}$  and  $\mathbf{v}$  as

$$\psi(\mathbf{u}, \mathbf{v}) = \frac{(\mathbf{u}^T \mathbf{v})^2}{\|\mathbf{u}\|^2 \|\mathbf{v}\|^2} \quad (10)$$

such that  $\psi(\mathbf{u}, \mathbf{v})$  is zero when  $\mathbf{u}$  and  $\mathbf{v}$  are orthogonal. By Cauchy–Schwarz inequality,  $\psi(\mathbf{u}, \mathbf{v}) \leq 1$  with equality if and only if  $\mathbf{u}$  and  $\mathbf{v}$  are linearly dependent, i.e.,  $\mathbf{u} = a\mathbf{v}$  for some scalar  $a$ .

We consider two formulations of NMFOC that impose two different orthogonality constraints separately on the recovered NMF spectra, such that the optimization problem given in Eq. (4) is augmented by the additional constraint that for a given target chemical spectrum,  $\mathbf{r}$ , we have  $\|\mathbf{r}\| = 1$ , and

Case 1:  $\psi(\mathbf{s}_1, \mathbf{r}) = 0$ .

Case 2:  $\psi(\mathbf{s}_1, \mathbf{r}) \geq \eta$ , where  $\eta \in (0, 1)$  and  $\mathbf{r}$  is centralized.

The basic idea for the first case is that we constrain one of the recovered spectra to be orthogonal to  $\mathbf{r}$ , which will force the other estimate to be in the same direction as  $\mathbf{r}$  when the constituent spectra are mutually orthogonal. If the correlation between the unconstrained estimate  $\mathbf{s}_2$  and  $\mathbf{r}$  exceeds a certain threshold, then the target present hypothesis is accepted.

For the second case, we have

$$\rho^2(\mathbf{s}_1, \mathbf{r}) = \frac{(\mathbf{s}_1^T \mathbf{r})^2}{\|\mathbf{s}_1 - E\{\mathbf{s}_1\}\|^2} > \psi(\mathbf{s}_1, \mathbf{r}) \quad (11)$$

where  $\rho(\cdot)$  is the correlation measure. By setting a proper threshold  $\eta$ , any solution of the second problem yields one constrained estimate  $\mathbf{s}_1$  close to the target as given by the orthogonality measure, and

also implies a high correlation with the target. For example, if  $\psi(\mathbf{s}_1, \mathbf{r}) \geq \eta = 0.45$ , the absolute correlation between  $\mathbf{s}_1$  and  $\mathbf{r}$  is greater than  $\sqrt{\eta} = 0.67$ . As shown in Section 3, the detection scheme in this case can be made based on the divergence of the Lagrangian multiplier.

For comparison, we use an augmented penalty method to solve Case 1 and Lagrangian method to solve Case 2.

### 3.1. NMFOC Based on KL Divergence with Multiplicative Updates (NMFOC-KLM)

We define the following measure as the new objective function for Case 1 given in Section 3:

$$D_a(\mathbf{X}||\mathbf{Y}) = \sum_{ij} \left( x_{ij} \log \frac{x_{ij}}{y_{ij}} - x_{ij} + y_{ij} \right) + \alpha \psi(\mathbf{s}_1, \mathbf{r}) \quad (12)$$

where  $\mathbf{Y} = \mathbf{A}\mathbf{S}$  and  $\alpha$  is a positive number, which determines the weight given to the orthogonality measure. Since the first term of Eq. (12) is close to zero after convergence,  $\alpha = 1$  provides satisfactory performance for most cases.

A local solution to the above minimization problem can be found by using the following update rules: Given  $\mathbf{X}$  and  $\mathbf{r}$ , initialize  $\mathbf{A}$  and  $\mathbf{S}$  by a random matrix with non-negative entries, then update the entries of  $\mathbf{A}$  and  $\mathbf{S}$  such that:

$$\mathbf{S} = \mathbf{S} \odot (\mathbf{A}^T (\mathbf{X} \oslash \mathbf{A}\mathbf{S})) \oslash (\mathbf{A}^T \mathbf{J} + \mathbf{E}), \quad (13)$$

$$\mathbf{s}_1 = \mathbf{s}_1 / \|\mathbf{s}_1\|, \quad (14)$$

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

$$\mathbf{A} = \mathbf{A} \oslash (\mathbf{P}\mathbf{A}), \quad (16)$$

where  $\mathbf{P}$  is an  $n \times n$  matrix of 1s,  $\mathbf{E}$  is an  $k \times m$  matrix with entries  $e_{ij} = 2\alpha(\mathbf{s}_1^T \mathbf{r})r_{j1}\delta(i-1)$ . To derive the update rules given in Eqs. (13), (14), (15), and (16), we use auxiliary functions as in [10].

If  $G(\mathbf{S})$  is an auxiliary function of  $F(\mathbf{S})$ , then  $F(\mathbf{S})$  is nonincreasing under the update

$$\mathbf{S}^{t+1} = \arg \min_{\mathbf{S}} G(\mathbf{S}, \mathbf{S}^t). \quad (17)$$

**Update of  $\mathbf{S}$ :**  $\mathbf{S}$  is updated by minimizing  $F(\mathbf{S}) = D_a(\mathbf{X}||\mathbf{A}\mathbf{S})$  with  $\mathbf{A}$  fixed. An auxiliary function is constructed for  $F(\mathbf{S})$  as:

$$G(\mathbf{S}, \mathbf{S}^t) = G'(\mathbf{S}, \mathbf{S}^t) + \alpha \psi(\mathbf{s}_1, \mathbf{r}) \quad (18)$$

where

$$G'(\mathbf{S}, \mathbf{S}^t) = \sum_{ij} (x_{ij} \log x_{ij} - x_{ij}) + \sum_{ij} y_{ij} - \sum_{i,j,k} x_{ij} \frac{a_{ik}s_{kj}^t}{\sum_b a_{ib}s_{bj}^t} \times \left( \log(a_{ik}s_{kj}) - \log \frac{a_{ik}s_{kj}^t}{\sum_b a_{ib}s_{bj}^t} \right)$$

is an auxiliary function defined for Eq. (3). As in [10], we can verify that  $G(\mathbf{S}, \mathbf{S}) = F(\mathbf{S})$  and  $G(\mathbf{S}, \mathbf{S}^t) \geq F(\mathbf{S})$ .

To minimize  $F(\mathbf{S})$  with respect to  $\mathbf{S}$ , we can thus update  $\mathbf{S}$  using Eq. (17). Such a matrix  $\mathbf{S}$  can be found by setting the gradient to zero for all  $v$  and  $u$ , i.e., by writing

$$\frac{\partial G(\mathbf{S}, \mathbf{S}^t)}{\partial s_{vu}} = - \sum_i x_{iu} \frac{a_{iv}s_{vu}^t}{\sum_b a_{ib}s_{bu}^t} \frac{1}{s_{vu}} + \sum_i a_{iv} + 2\alpha(\mathbf{s}_1^T \mathbf{r})r_{u1}\delta(v-1) = 0 \quad (19)$$

where  $r_{u1}$  is the  $u$ th entry in the reference spectrum vector  $\mathbf{r}$ . Thus, the update rule for Eq. (17) takes the form

$$s_{vu}^{t+1} = s_{vu}^t \frac{\sum_i a_{iv}x_{iu}/y_{iu}}{\sum_i a_{iv} + 2\alpha(\mathbf{s}_1^T \mathbf{r})r_{u1}\delta(v-1)}. \quad (20)$$

which is Eq. (13) if written in a compact matrix form. To obtain Eq. (19), we assume that  $\mathbf{s}_1$  has unit energy which is guaranteed by Eq. (14). Since there

is no other constraint on  $\mathbf{A}$  except non-negativity, we obtain the same update rule as in [10] for  $\mathbf{A}$ , which is given in Eq. (15).

From the above analysis, we conclude that the updates given by Eqs. (13), (14), (15), and (16) result in a sequence of non-increasing values of  $D_a(\mathbf{X}||\mathbf{A}\mathbf{S})$ , and hence converges to a local minimum. After convergence, the correlation value between  $\mathbf{s}_2$  and  $\mathbf{r}$  is calculated. If it exceeds a certain threshold, the hypothesis that the target is present is accepted.

### 3.2. NMFOC Based on KL Divergence with Additive Projected Gradient Updates (NMFOC-KLA)

The gradient of cost defined in Eq. (12) with respect to  $\mathbf{S}$  is calculated as

$$\nabla_{\mathbf{S}}(D_a) = \mathbf{A}^T \mathbf{J} + \mathbf{E} - \mathbf{A}^T (\mathbf{X} \odot \mathbf{A}\mathbf{S}). \quad (21)$$

It is noted that the orthogonality constraint is also lower bounded by zero, as is the KL divergence, and is a quadratic function of  $\mathbf{S}$ . Based on Eq. (21) and using Eqs. (5) and (8), we can easily obtain Eq. (13) without the need to use an auxiliary function. This argument provides a promising way to heuristically obtain a fixed-point update rule for any problem similar to NMF. Furthermore, if we exchange the position of the numerator and the denominator in Eq. (13), the KL divergence will keep increasing to infinity since it is not upper bounded.

For the implementation of projected gradient method in this case, we adopt the scheme, given in [12]:

Given  $0 < \beta < 1$ ,  $0 < \sigma < 1$ , initialize  $\mathbf{S}$  to any feasible value

**for**  $k = 1, 2, \dots, \mathcal{K}$   
 $\mathbf{S}^{k+1} = \max(\mathbf{S}^k - \mu \nabla_{\mathbf{S}}(D_a), 0)$ ,

where  $\mu = \beta^t$ , and  $t$  is the first non-negative integer for which

$D(\mathbf{S}^{k+1}, \mathbf{A}) - D(\mathbf{S}^k, \mathbf{A}) \leq \sigma \nabla_{\mathbf{S}}(D_a)(\mathbf{S}^{k+1} - \mathbf{S}^k)$   
**end(for)**

The scheme above is used for updating  $\mathbf{S}$  by treating  $\mathbf{A}$  as constant. Repeating the same procedure,  $\mathbf{A}$  can be updated similarly, this time by treating  $\mathbf{S}$  as constant. Thus, the update procedure alternates between the updates of  $\mathbf{A}$  and  $\mathbf{S}$ .

Compared with the multiplicative updates, the main shortcoming of the projected gradient method is that it requires the selection of an optimal stepsize. However, if we compare the number of total iterations to achieve a local minima, multiplicative updates may be not as efficient as projected gradient updates as observed in the simulations.

### 3.3. NMFOC Based on Euclidean Distance

If the Euclidean distance metric is used instead of using generalized KL divergence, the cost function with an orthogonality constraint is given by

$$D_a(\mathbf{X}||\mathbf{Y}) = \sum_{i,j} (x_{ij} - y_{ij})^2 + \alpha \psi(\mathbf{s}_1, \mathbf{r}). \quad (22)$$

Its gradient with respect to  $\mathbf{S}$  and  $\mathbf{A}$  can be calculated as

$$\begin{aligned} \nabla_{\mathbf{S}}(D_a) &= \mathbf{A}^T \mathbf{A}\mathbf{S} + \mathbf{E} - \mathbf{A}^T \mathbf{X}, \\ \nabla_{\mathbf{A}}(D_a) &= \mathbf{A}\mathbf{S}\mathbf{S}^T - \mathbf{X}\mathbf{S}^T. \end{aligned} \quad (23)$$

As shown in the previous section, based on this gradient, we can easily write the multiplicative update rules as:

$$\begin{aligned} \mathbf{S} &= \mathbf{S} \odot (\mathbf{A}^T \mathbf{X}) \odot (\mathbf{A}^T \mathbf{A}\mathbf{S} + \mathbf{E}), \\ \mathbf{A} &= \mathbf{A} \odot (\mathbf{X}\mathbf{S}^T) \odot (\mathbf{A}\mathbf{S}\mathbf{S}^T), \end{aligned} \quad (24)$$

without constructing any auxiliary functions for NMFOC-EUM (NMFOC based on Euclidean distance with multiplicative updates). NMFOC based on Euclidean distance with additive projected gradient updates (NMFOC-EUA) can be derived by following the scheme presented in Section (3.2) by using the gradient equations given in Eq. (23).

### 3.4. NMFOC using Augmented Lagrangian (NMFOC-L)

In this case, the constrained optimization problem is defined as in Eq. (4) by adding the constraint  $c(\mathbf{S}) =$

$\psi(\mathbf{s}_1, \mathbf{r}) - \eta \geq 0$ , where  $c(\mathbf{S})$  represents the constraint imposed on  $\mathbf{S}$  and  $\eta$  is a given threshold. Here we only consider the generalized KL divergence as an example of using the Augmented Lagrangian approach. The extension to the Euclidean distance is straightforward. Also the projected gradient method is not considered here.

Using Lagrangian multiplier method, the augmented Lagrangian function is defined as [14]:

$$\mathcal{L}(\mathbf{A}, \mathbf{S}, \lambda, \mu) = D(\mathbf{X}|\mathbf{A}\mathbf{S}) \quad (25)$$

$$+ \begin{cases} -\lambda c(\mathbf{S}) + \frac{1}{2\mu} c^2(\mathbf{S}) & \text{if } c(\mathbf{S}) - \mu\lambda \leq 0 \\ -\frac{\mu}{2} \lambda^2 & \text{otherwise.} \end{cases}$$

and the following procedure is used to find the solution [14]:

Given  $\mu_0 > 0$ , starting points  $(\mathbf{A}_0^s, \mathbf{S}_0^s)$  and  $\lambda_0$ ;  
**for**  $k = 0, 1, 2, \dots, \mathcal{K}$

    Find an approximate minimizer  $(\mathbf{A}_k, \mathbf{S}_k)$  of  
      $\mathcal{L}(\mathbf{A}, \mathbf{S}, \lambda_k, \mu_k)$ , starting at  $(\mathbf{A}_k^s, \mathbf{S}_k^s)$

    If final convergence test is satisfied

**STOP** with approximate solution  $(\mathbf{A}_k, \mathbf{S}_k)$ ;

    Update Lagrangian multipliers using \indent  
     \indent  $\lambda(k+1) = \max(\lambda(k) - \frac{c(\mathbf{S}_k)}{\mu_k}, 0)$ ;

    Choose new penalty parameter  $\mu_{k+1} \in (0, \mu_k)$ ;

    Set starting point for the next iteration as

$(\mathbf{A}_{k+1}^s, \mathbf{S}_{k+1}^s) = (\mathbf{A}_k, \mathbf{S}_k)$ ;

**end(for)**

Generally the approximate minimum of Eq. (25) is searched by using a gradient type algorithm. Here we propose an update rule that guarantees non-negativity as in [10] and can obtain the minimum of Eq. (25) for a given  $\lambda$  and  $\mu$ .

### 3.5. Convergence Properties of NMFOL-L

In the above algorithmic framework, within each iteration  $k$ ,  $\mathbf{S}$  is updated by minimizing the augmented Lagrangian function  $F(\mathbf{S}) = \mathcal{L}(\mathbf{A}, \mathbf{S}, \lambda, \mu)$  while  $\mathbf{A}$  is fixed.

An auxiliary function is constructed for  $F(\mathbf{S})$  as

$$G(\mathbf{S}, \mathbf{S}^t) = G^t(\mathbf{S}, \mathbf{S}^t)$$

$$+ \begin{cases} -\lambda c(\mathbf{S}) + \frac{1}{2\mu} c^2(\mathbf{S}) & \text{if } c(\mathbf{S}) \leq \mu\lambda \\ -\frac{\mu}{2} \lambda^2 & \text{otherwise} \end{cases}$$

The inequality  $c(\mathbf{S}) \leq \mu\lambda$  indicates that the constraint imposed on  $\mathbf{S}$  is active. As in NMFOL, it is easy to verify that  $G(\mathbf{S}, \mathbf{S}) = F(\mathbf{S})$  and  $G(\mathbf{S}, \mathbf{S}^t) \geq F(\mathbf{S})$ .

To minimize  $F(\mathbf{S})$ , we can update  $\mathbf{S}$  using Eq. (17). Such a matrix  $\mathbf{S}$  can be found by setting  $\frac{\partial G(\mathbf{S}, \mathbf{S}^t)}{\partial s_{vu}} = 0$  for all  $v$  and  $u$ . When  $c(\mathbf{S}) - \mu\lambda > 0$ , we obtain the same update rule for  $\mathbf{S}$  as in [10]. If not, we obtain:

$$\frac{\partial G(\mathbf{S}, \mathbf{S}^t)}{\partial s_{vu}} = - \sum_i x_{iu} \frac{a_{iv} s_{vu}^t}{\sum_b a_{ib} s_{bu}^t} \frac{1}{s_{vu}} + \sum_i a_{iv}$$

$$- \lambda \frac{\partial c(\mathbf{S})}{\partial s_{vu}} + \frac{c(\mathbf{S})}{\mu} \frac{\partial c(\mathbf{S})}{\partial s_{vu}} = 0 \quad (26)$$

where

$$\frac{\partial c(\mathbf{S})}{\partial s_{vu}} = \left( \frac{2(\mathbf{s}_1^T \mathbf{r}) r_{u1}}{\|\mathbf{s}_1\|^2} - \frac{2(\mathbf{s}_1^T \mathbf{r})^2}{\|\mathbf{s}_1\|^4} s_{vu} \right) \delta(v-1). \quad (27)$$

Assuming that the effects of  $s_{vu}$  on  $\mathbf{s}_1^T \mathbf{r}$  and  $\|\mathbf{s}_1\|$  are not significant, we approximate Eq. (27) by treating  $\mathbf{s}_1^T \mathbf{r}$  and  $\|\mathbf{s}_1\|$  as constant for current estimate  $s_{vu}$ . After some straightforward algebraic manipulations, for the first row of  $\mathbf{S}$  ( $v=1$ ), on which the constraint is applied to, we can rewrite Eq. (26) as

$$\theta_1 s_{vu}^2 + \theta_2 s_{vu} + \theta_3 = 0 \quad (28)$$

where

$$\theta_1 \equiv \left( \lambda\mu - \frac{(\mathbf{s}_1^T \mathbf{r})^2}{\|\mathbf{s}_1\|^2} + \eta \right) \frac{2(\mathbf{s}_1^T \mathbf{r})^2}{\mu \|\mathbf{s}_1\|^4},$$

$$\theta_2 \equiv \sum_i a_{iv} - \left( \lambda\mu - \frac{(\mathbf{s}_1^T \mathbf{r})^2}{\|\mathbf{s}_1\|^2} + \eta \right) \frac{2(\mathbf{s}_1^T \mathbf{r}) r_{u1}}{\mu \|\mathbf{s}_1\|^2}, \text{ and}$$

$$\theta_3 \equiv - \sum_i x_{iu} \frac{a_{iv} s_{vu}^t}{\sum_b a_{ib} s_{bu}^t}.$$

Thus, the update rule for  $\mathbf{S}$  takes the form

$$s_{vu}^{t+1} = \frac{-\theta_2 + \sqrt{\theta_2^2 - 4\theta_1\theta_3}}{2\theta_1}. \quad (29)$$

Since  $\theta_1 \geq 0$  and  $\theta_3 \leq 0$ ,  $s_{vu}^{t+1}$  will always be nonnegative. If  $v \neq 1$ , the update rule for  $\mathbf{S}$  reduces to

$$s_{vu}^{t+1} = \sum_i x_{iu} \frac{a_{iv}s_{vu}^t}{\sum_b a_{ib}s_{bu}^t} \quad (30)$$

when  $\sum_i a_{iv} = 1$ .

Since there is no other constraint on  $\mathbf{A}$  except nonnegativity, we obtain the same update rule for  $\mathbf{A}$  (given in Eq. (15)). From the above analysis, we conclude that these update rules, Eqs. (15) and (16) for  $\mathbf{A}$ , and Eqs. (29) and (30) for  $\mathbf{S}$ , result in a sequence of non-increasing values of  $\mathcal{L}(\mathbf{A}, \mathbf{S}, \lambda, \mu)$ , and hence converges to a local minimum.

Test simulations indicate that a solution can be found to satisfy the constraint even when the target is not present, which means further analysis is needed for the subsequent detection stage. By constraining all the rows of  $\mathbf{S}$  to satisfy  $\mathbf{s}_i^T \mathbf{r} \geq 0, \forall i$ , it is shown next that the final divergence  $D(\mathbf{X}||\mathbf{AS})$  will be lower bounded by some non-zero value if the target is absent.

In the case of two sources, we assume that the observation matrix  $\mathbf{X} = [\mathbf{x}_1 \mathbf{x}_2]^T$  is generated by  $\mathbf{X} = \mathbf{A}^* \mathbf{S}^*$ , where

$$\mathbf{A}^* = \begin{bmatrix} \alpha & 1 - \alpha \\ \beta & 1 - \beta \end{bmatrix} \text{ and } \mathbf{S}^* = \begin{bmatrix} s_1^{*T} \\ s_2^{*T} \end{bmatrix}.$$

Here we assume that the total contribution of all spectra are constant, and is equal to one. NMFOC-L approximates  $\mathbf{X}$  by  $\mathbf{Y} = [\mathbf{y}_1 \mathbf{y}_2]^T = \mathbf{AS}$ , where the sum of each column of  $\mathbf{A}$  is constrained to be unity. If  $D(\mathbf{X}||\mathbf{Y}) = 0$  and  $\mathbf{X} = \mathbf{Y}$ , we have  $\mathbf{x}_1 + \mathbf{x}_2 = \mathbf{y}_1 + \mathbf{y}_2$  and thus obtain

$$\tau \mathbf{s}_1^* + (2 - \tau) \mathbf{s}_2^* = \mathbf{s}_1 + \mathbf{s}_2 \quad (31)$$

where  $\tau = \alpha + \beta$ . Assuming  $\mathbf{s}_i^{*T} \mathbf{r} \geq 0$  and  $\|\mathbf{s}_i^*\|^2 = 1$  for all  $i$ , we define  $\psi_i \equiv (\mathbf{s}_i^T \mathbf{r})^2$  and write

$$\begin{aligned} \psi(\mathbf{s}_1, \mathbf{r}) &= \frac{(\mathbf{s}_1^T \mathbf{r})^2}{\|\mathbf{s}_1\|^2} \\ &\leq \frac{(\tau\sqrt{\psi_1} + (2 - \tau)\sqrt{\psi_2})^2}{\|\mathbf{s}_1\|^2} \\ &\leq \frac{4 \max(\psi_1, \psi_2)}{\|\mathbf{s}_1\|^2} = \gamma \end{aligned} \quad (32)$$

by using Eq. (31) and assuming  $\mathbf{s}_i^T \mathbf{r} \geq 0$  for all  $i$ . To simplify the subsequent discussions, we assume that  $\|\mathbf{s}_1\|^2 = 1$ . This can be achieved by simply normalizing  $\mathbf{s}_1$  after iterations in Eq. (29).

Since  $\mathbf{r}$  is centralized, a non-negative vector  $\mathbf{s}$  can not have same direction as  $\mathbf{r}$ , thus  $\psi(\mathbf{s}, \mathbf{r})$  is strictly less than 1. Under additional assumption that all constituent spectra have same variance, it can be shown that the target chemical spectrum achieves the maximum of  $\psi(\mathbf{s}, \mathbf{r})$ . For subsequent discussions, we let  $\gamma_0$  or  $\gamma_1$  denote the upper bound  $\gamma$  corresponding to the case that the target is absent or present, respectively. Now assuming target is absent, we know that  $D(\mathbf{X}||\mathbf{AS}) = 0$  implies  $\psi(\mathbf{s}_1, \mathbf{r}) \leq \gamma_0$ . In this case, we set a threshold  $\eta$  with  $\gamma_1 > \eta > \gamma_0$ . If NMFOC-L can still find a pair  $(\mathbf{A}, \mathbf{S})$  satisfying  $\psi(\mathbf{s}_1, \mathbf{r}) \geq \eta$ , this implies  $D(\mathbf{X}||\mathbf{Y}) > 0$  as the upper bound is violated. Also since  $\psi(\mathbf{s}_1, \mathbf{r}) \geq \eta > \gamma_0$ , we have  $\mathbf{s}_1^T \mathbf{r} > \tau\sqrt{\psi_1} + (2 - \tau)\sqrt{\psi_2}$  and  $(\mathbf{y}_1 + \mathbf{y}_2 - \mathbf{x}_1 - \mathbf{x}_2)^T \mathbf{r} \geq \sqrt{\eta} - \sqrt{\gamma_0}$ . By Cauchy-Schwarz inequality, we also have  $\|\mathbf{y}_1 + \mathbf{y}_2 - \mathbf{x}_1 - \mathbf{x}_2\| \geq (\mathbf{y}_1 + \mathbf{y}_2 - \mathbf{x}_1 - \mathbf{x}_2)^T \mathbf{r}$ . Finally we obtain

$$\|\mathbf{y}_1 + \mathbf{y}_2 - \mathbf{x}_1 - \mathbf{x}_2\| \geq \sqrt{\eta} - \sqrt{\gamma_0} \quad (33)$$

Hence, given that the observation matrix  $\mathbf{X}$  is generated by two sources, we can have either of the following two conclusions for the detection problem:

- (1) If the target is absent, the orthogonality constraint can not be satisfied or it can be satisfied but with a non-zero  $D(\mathbf{X}||\mathbf{AS})$  which implies Eq. (33). Thus the Lagrangian multiplier diverges when the threshold can not be satisfied.

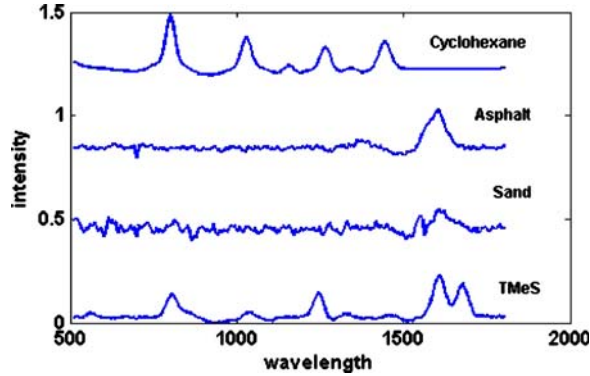


Figure 1. Raman spectra used in the simulation.

Table 1. Correlations and orthogonality measure values for the spectra of four chemicals and the target spectrum.

Chemicals	r (TMeS)	
	Correlation, $\rho$	Orthogonality, $\psi$
TMeS	1	0.47
Sand	0.52	0.038
Asphalt	0.6	0.11
Cyclohexane	0.23	0.027

- (2) If the chemical is present, the orthogonality constraint can be satisfied with a very small  $D(\mathbf{X}||\mathbf{AS})$  which can be approximately zero since  $\eta < \gamma_1$  and the upper bound in Eq. (32) is not violated.

## 4. Experiments

### 4.1. Augmented NMFOC Algorithms on Multiple Sources Mixture

To demonstrate feasibility and performance of the four augmented NMFOC algorithms (NMFOC-

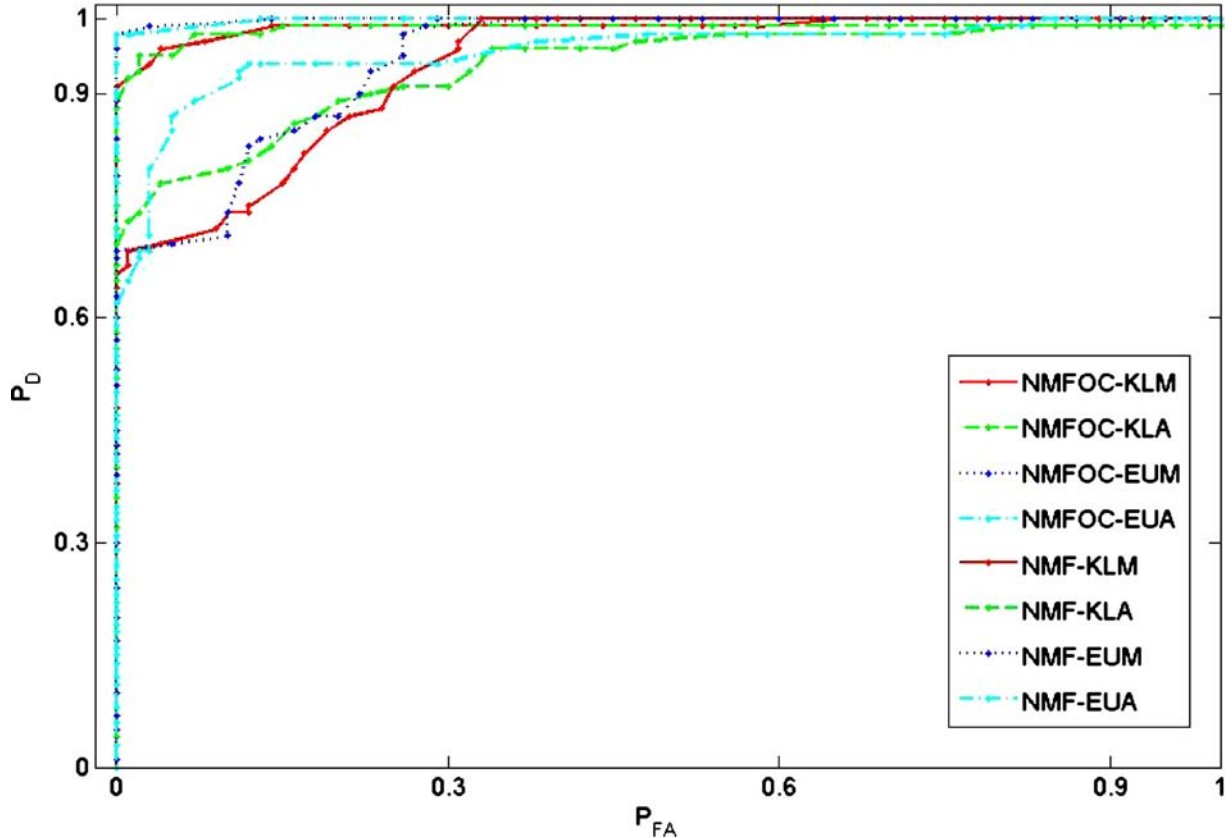


Figure 2. ROCs of four augmented NMFOC and NMF counterparts for 10 dB.

KLM, NMFOC-KLA, NMFOC-EUM, NMFOC-EUA) and the Lagrangian NMFOC (NMFOC-L) algorithms, we conduct experiments using artificial mixtures of Raman spectral data. Given an  $n \times m$  observation matrix  $\mathbf{X}$ , where  $n$  is the number of spectra and  $m$  is the dimension of Raman spectrum, we choose between the two competing hypotheses:  $\mathcal{H}_0$  and  $\mathcal{H}_1$ , i.e., the target spectrum is absent or present. The target in our experiments is the Raman spectrum of the chemical, TMeS.

In the first simulation, the mixtures are generated by asphalt, sand, cyclohexane, and with or without TMeS. In the experiments, only the spectra with bandwidth of 509 to 1,805 Raman shift ( $\text{cm}^{-1}$ ) have been considered and are shown in Fig. 1. The similarity between the Raman spectra of the three non-target chemicals and the target spectrum is quantified by both correlation and the orthogonality measure in Table 1.

To generate the mixture under each hypothesis, we use random non-negative square mixing matrix  $\mathbf{A}$

and Gumbel distributed noise  $\mathbf{N}$  [19]. The signal to noise ratio  $SNR = 10 \log(E_S/E_N)$  is defined in terms of the energy of constituent chemical spectra and the energy of the noise.

After generating mixture  $\mathbf{X}$ , we estimate the latent Raman spectra using NMF and the four Augmented NMFOC algorithms. The dimension of  $\mathbf{X}$  we generated is  $3 \times m$  when the target is not present and  $4 \times m$  when the target is present. For all the algorithms, the number of maximum iterations is set to 80. For the additive versions, the related parameters are set as  $\sigma = 0.9$ ,  $\beta = 0.8$ , and  $\mathcal{K} = 100$ . The selection of these parameters may affect the speed of the convergence, however in our simulations we did not note significant sensitivity to different selections. For each method, if the largest absolute correlations between their estimates and the target spectrum exceed a certain threshold  $\eta \in (0, 1)$ , we accept  $\mathcal{H}_1$ . By adjusting the threshold, we obtain a sequence of detection rates for each method. The probability of false alarm ( $P_{FA}$ ) and the probability of detection

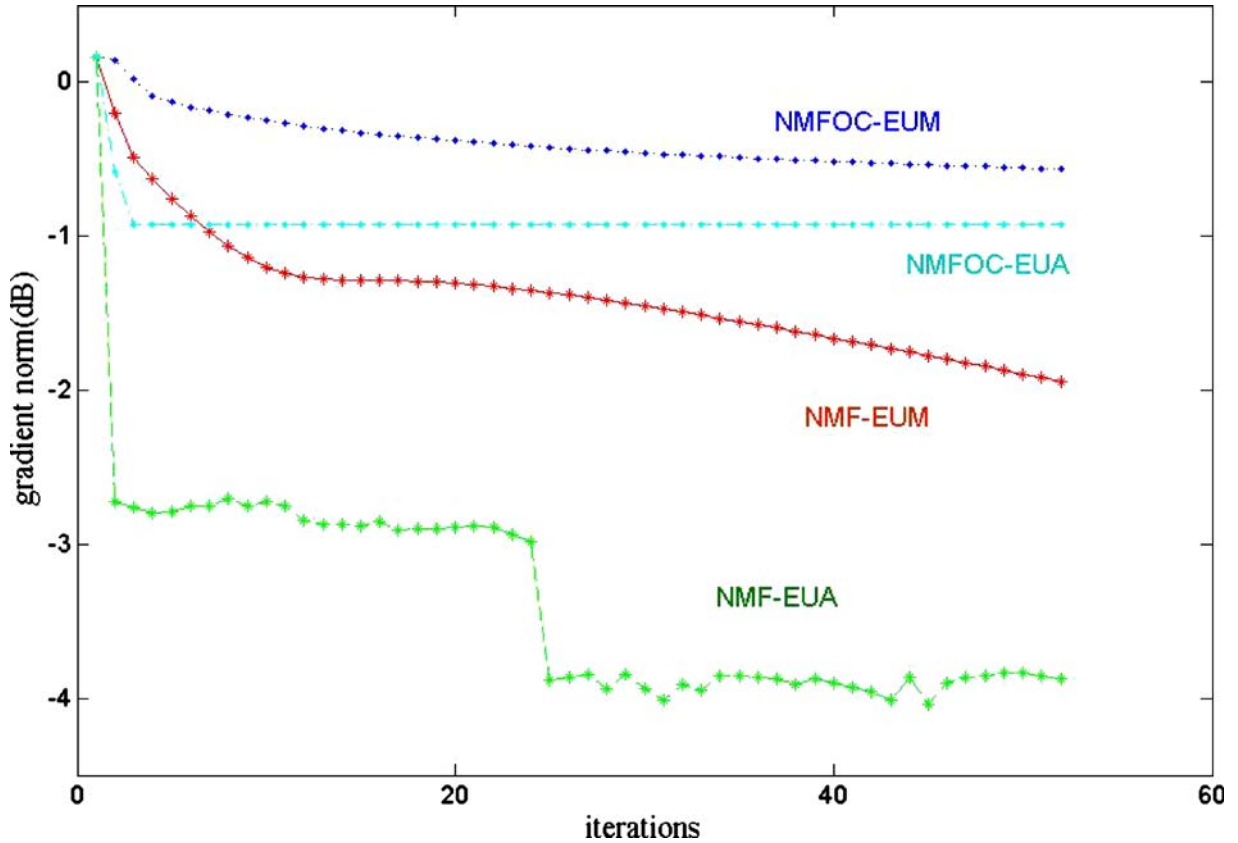


Figure 3. Convergence behavior of multiplicative updates and additive updates.

( $P_D$ ) are estimated as the ratio of detection rates and the number of trials. Here we provide the results based on 500 independent trials.

In Fig. 2 the Receiver Operating Characteristic (ROC) curves corresponding to different methods are plotted for  $SNR = 10$  dB case. In all the cases, the four Augmented NMFOC algorithms provide much better performance than the original NMF counterparts. Thus the NMFOC algorithms significantly improved the performance of NMF when used for detection.

The results also show that for the purpose of recovering the underlying spectra, the Euclidean distance may provide a better measure than the generalized KL divergence. Within each measure, there is no significant difference in performance between the multiplicative and additive methods. However, as shown in Fig. 3, the additive projected gradient updates require much smaller number of iterations to converge than the multi-

licative counterparts. It is observed that NMFOC-EUA typically converges within ten iterations, which is similar to what have been observed in [7]. One possible reason is that the factorization becomes more and more inflexible as the constraints imposed on  $\mathbf{A}$  and  $\mathbf{S}$  are increased.

#### 4.2. Application to Field Raman Data

In this case, the data are generated by the Laser Interrogation of Surface Agents (LISA) system [8]. In this real application, the target chemical TMeS was dropped on glass and the Raman spectra were collected using LISA detecting system. There are 100 measurements for each given drop size. Using the NMF and NMFOC-KLM algorithms, each time we analyze ten pulses and estimate two spectra. As shown in Fig. 4, the highest correlation between the estimated components and the target chemical increases as the drop size increases since the drop

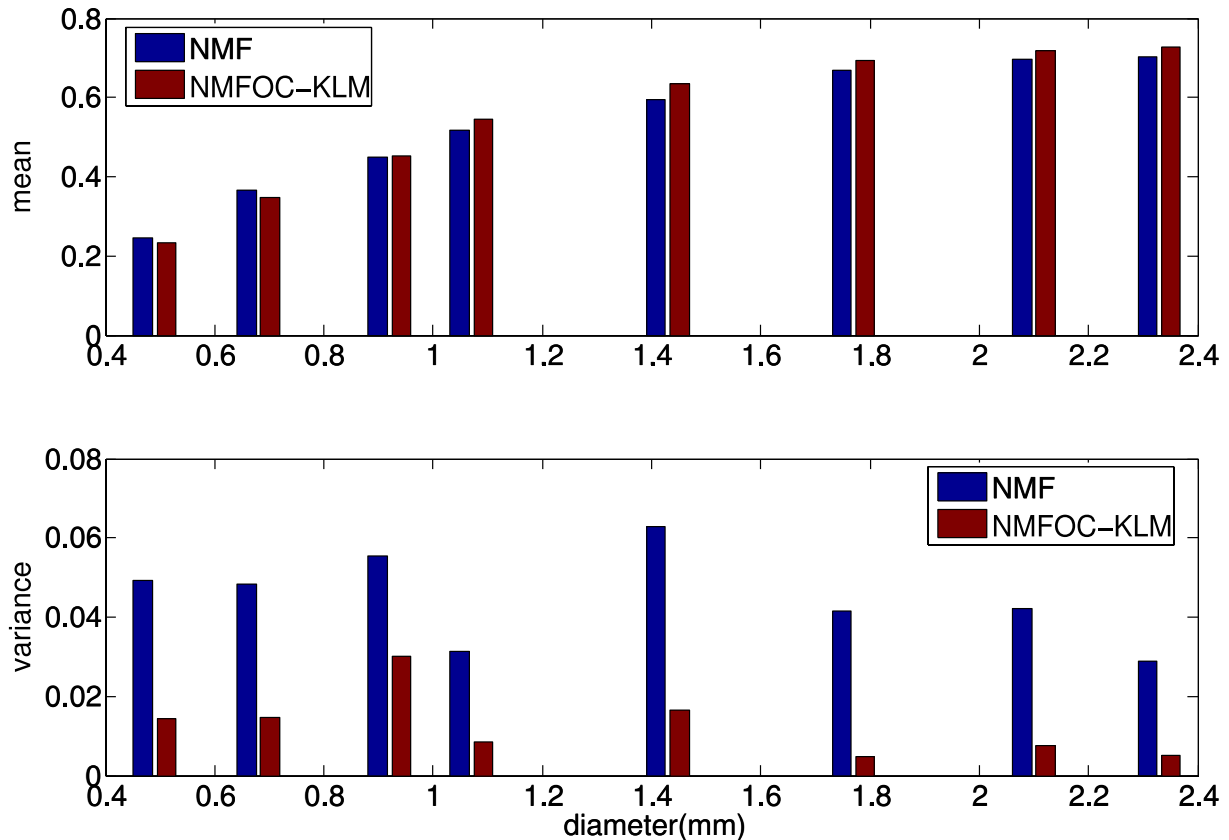


Figure 4. The real application of NMFOC algorithms.

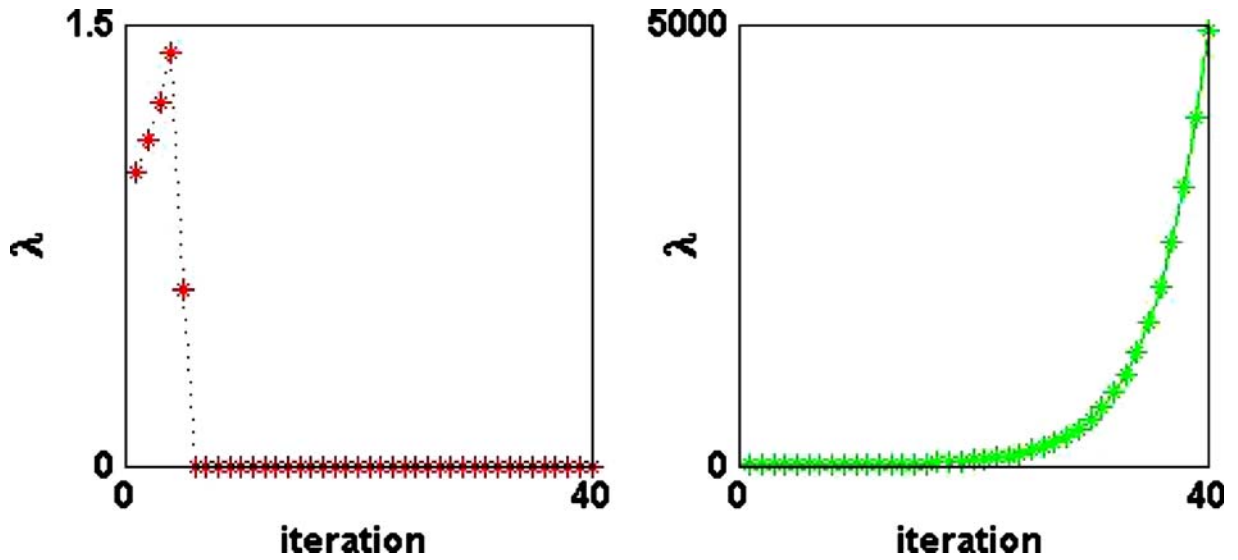


Figure 5. Examples for convergent and divergent behavior of  $\lambda$ .

size indicates the amount of signal contribution contributed by the given chemical. The NMFOC-KLM algorithm shows smaller variation than the standard NMF algorithm, most likely due to the decrease in the possibility of converging to undesirable local minima. Obviously, this is a very desirable property for a data analysis algorithm that needs to be employed in real-time as in this application.

#### 4.3. Lagrangian NMFOC Algorithm with a Two-Source Mixture

The detection scheme for NMFOC-L discussed at the end part of Section 3.5 can be simplified to just observing the behavior of the Lagrangian multiplier  $\lambda$ . The dimension of  $\mathbf{X}$  is  $2 \times m$  in this case. The related parameters are set as  $\mathcal{K} = 10, \lambda_0 = 1, \mu_0 =$

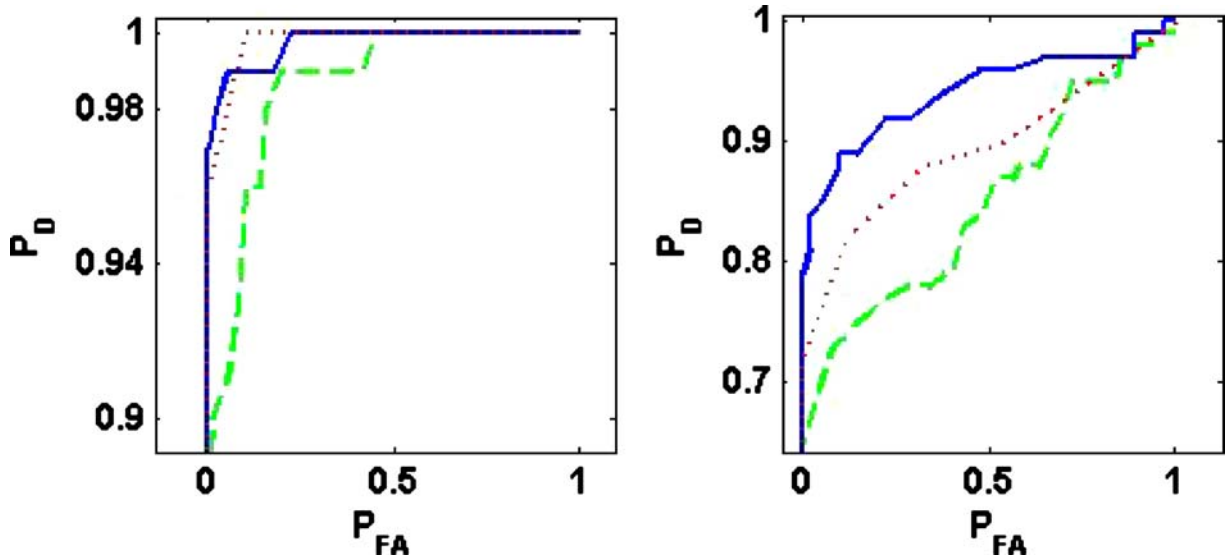


Figure 6. ROCs of NMF-KLM (dashed), NMFOC-KLM (solid), and NMFOC-L (dotted) for 10 dB (left) and 5 dB (right).

0.7, and  $\mu_{i+1} = 0.8\mu_i$ . Compared with the augmented NMFOC algorithms,  $\mathcal{K}$  is much smaller because the computational cost of Lagrangian NMFOC shows a significant increase within each iteration compared to NMFOC algorithms. When the threshold can be satisfied,  $\lambda$  will converge to zero. If not,  $\lambda$  will diverge. In the second example, we generated observations with the sand and asphalt spectra for  $\mathcal{H}_0$ , and TMeS and asphalt spectra for  $\mathcal{H}_1$  and used the initial values for the variables as  $\lambda_0 = 1$  and  $\mu_0 = 0.8$ . A typical behavior of  $\lambda$  for this example is shown in Fig. 5. Similarly as in the multiple sources case, by adjusting the threshold  $\eta$ , we obtain a sequence of detection rates and estimate the  $P_D$  and  $P_{FA}$  based on 100 independent trials.

As shown in Fig. 6, NMFOC-L is almost as good as NMFOC-KLM for  $SNR = 10$  dB case. As noise increases, the performance of NMFOC-L degrades, but it still outperforms NMF-KLM.

## 5. Discussion

In this paper, we propose five variations of the NMF algorithm, NMFOC-KLM, NMFOC-KLA, NMFOC-EUM, NMFOC-EUA and NMFOC-L, to solve a detection problem. An important element of our approach is the explicit use of a priori information in a generative model framework. It is incorporated through an orthogonality constraint to guide the matrix factorization. NMFOC-L can use the change in  $\Delta\lambda$  as a clear detection index at the expense of increased computational cost. In contrast, the other four NMFOC algorithms we derived using two distance metrics-KL divergence and Euclidean distance-and two types of update rules-multiplicative and additive-provide good detection performance without sacrificing the speed of NMF. Also important to note is that additive updates provide faster convergence provided that an appropriate stepsize is chosen.

## References

1. M. Berry, M. Browne, A. Langville, P. Pauca, and R. Plemmons, "Algorithms and Applications for Approximate Nonnegative Matrix Factorization," *Comput. Stat. Data Anal.*, 2006 (in press).
2. A. Cichocki, R. Zdunek, and S. Amari, "Csiszár's Divergence for Non-negative Matrix Factorization: Family of New Algorithms," in *Proc. 6th Int. Conf. ICA and BSS*, Charleston SC, March 5–8, 2006, Springer LNCS, vol. 3889, pp. 32–39.
3. D. Donoho and V. Stodden, "When Does Non-negative Matrix Factorization Give a Correct Decomposition into Parts?" in *Proc. Neural Information Processing Systems*, vol. 16, 2003, pp. 1141–1149.
4. I. S. Dhillon and S. Sra, "Generalized Nonnegative Matrix Approximations with Bregman Divergences," in *Proc. NIPS*, Vancouver, BC, 2005.
5. C. Gobinet, E. Perrin, and R. Huez, "Application of Nonnegative Matrix Factorization to Fluorescence Spectroscopy," in *Proc. EUSIPCO 2004*, Vienna, Austria, Sept. 6–10, 2004.
6. F. Guimet, R. Boqué, and J. Ferré, "Application of Non-negative Matrix Factorization Combined with Fisher's Linear Discriminant Analysis for Classification of Olive Oil Excitation-emission Fluorescence Spectra," *Chemometr. Intell. Lab. Syst.*, vol. 81, 2006, pp. 94–106.
7. P. O. Hoyer, "Non-negative Matrix Factorization with Sparseness Constraints," *J. Mach. Learn. Res.*, vol. 5, 2004, pp. 1457–1469.
8. ITT Industries, Advanced Engineering and Sciences Division, "Tests of Laser Interrogation of Surface Agents System for On-the-move Standoff Sensing of Chemical Agents," in *Proc. Int. Symp. Spect. Sensing Research*, 2003.
9. D. D. Lee and H. S. Seung, "Learning the Parts of Objects by Non-negative Matrix Factorization," *Nature*, vol. 401, 1999, pp. 788–791.
10. D. D. Lee and H. S. Seung, "Algorithms for Non-negative Matrix Factorization," in *Proc. Neural Information Processing Systems*, vol. 13, 2000, pp. 556–562.
11. S. Z. Li, X. Hou, H. Zhang, and Q. Cheng, "Learning Spatially Localized, Parts-based Representation," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 1, 2001, pp. 207–212.
12. C.-J. Lin, "Projected Gradient Methods for Non-negative Matrix Factorization," Technical report, Department of Computer Science, National Taiwan University, 2005.
13. S. Moussaoui, D. Brie, C. Carteret, and A. Mohammad-Djafari, "Application of Bayesian Non-negative Source Separation to Mixture Analysis in Spectroscopy," in *Proc. 24th Int. Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering*, Max-Planck Institute, Garching, Munich, Germany, July 2004, pp. 25–30.
14. J. Nocedal and S. J. Wright, *Numerical Optimization*, Springer, 2000.
15. A. Pascual-Montano, J. M. Carazo, K. Kochi, D. Lehmann, and R. D. Pascual-Marqui, "Nonsmooth Nonnegative Matrix Factorization (nsNMF)," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 3, 2006, pp. 403–415, Mar.
16. P. Sajda, D. Shuyan, and L. Parra, "Recovery of Constituent Spectra Using Non-negative Matrix Factorization," *Proc. SPIE*, vol. 5207, 2003, pp. 321–331.
17. F. Sha, L. K. Saul, and D. D. Lee "Multiplicative Updates for Nonnegative Quadratic Programming in Support Vector Machines," in *Proc. Neural Information Processing Systems*, vol. 15, MIT Press, 2003.
18. R. Salakhutdinov, S. Roweis, and Z. Ghahramani, "On the Convergence of Bound Optimization Algorithms," in *Proc.*

*Li et al.*

- Conf. on Uncertainty in Artificial Intelligence*, vol. 19, 2003, pp. 509–516.
19. S. Sigurdsson, J. Larsen, P. Philipsen, M. Gniadecka, H. Wulf, and L. Hansen, “Estimating and Suppressing Background in Raman Spectra with an Artificial Neural Network,” *Informat-ics and Mathematical Modeling*, Technical Univ. Denmark, Tech. Rep. 2003–2020, 2003.
  20. W. Wang and T. Adalò, “Constrained ICA and its Application to Raman Spectroscopy,” in *AP-S/URSI Symposium 2005*, Washington, DC.
  21. W. Wang, T. Adalò, H. Li, and D. Emge, “Detection Using Correlation Bound and its Application to Raman Spectroscopy,” in *2005 IEEE Workshop on Machine Learning for Signal Processing*, September, 2005, pp. 259–264.
  22. R. Zdunek and A. Cichocki, “Non-negative Matrix Factorization with Quasi-Newton Optimization,” in *Proc. 8th Int. Conf. on Artificial Intelligence and Soft Computing*, ICAISC, Zakopane, Poland, 25–29 June, 2006, Springer Lectures Notes in Artificial Intelligence, vol. 4029, pp. 870–879.



**Hualiang Li** received the B.Eng. degree in Electrical Engineering from Zhejiang University, Hangzhou, China, in 1996. He is currently pursuing the Ph.D. degree in Electrical Engineering at the University of Maryland, Baltimore County. His research interests include optimization, complex-valued signal processing and the applications to independent component analysis, neural networks, biomedical imaging processing, and variations of non-negative matrix factorization algorithms.



**Tülay Adalı** received the Ph.D. degree in electrical engineering from North Carolina State University, Raleigh, in 1992 and joined the faculty at the University of Maryland Baltimore County (UMBC), Baltimore, the same year. She is currently a professor in the Department of Computer Science and Electrical Engineering at UMBC. She worked in the organization of a number of international conferences and workshops including the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), the IEEE International Workshop on Neural Networks for Signal Processing (NNSP), and the IEEE International Workshops on Machine Learning for Signal Processing (MLSP). She was the general co-chair for the NNSP workshops 2001–2003 and the technical chair of the MLSP workshops 2004–2006. She is the past chair and current member of the MLSP Technical Committee, and is serving on the IEEE publications board and the IEEE Signal Processing Society conference board. She is the associate editor of *IEEE Transactions on Signal Processing* and the *VLSI journal of signals and systems*. Her research interests are in the areas of statistical signal processing, machine learning for signal processing, biomedical data analysis (functional MRI, MRI, PET, CR, ECG, and EEG), bioinformatics, and signal processing for optical communications. Dr. Adalı is the recipient of a 1997 National Science Foundation (NSF) CAREER Award and the provost’s research faculty fellowship.



**Wei Wang** received the B.S. and M.S. degree in Engineering Mechanics from Huazhong University of Science and Tech-

## Non-negative Matrix Factorization with Orthogonality Constraints

nology, People's Republic of China, in 1992 and 1995, respectively. Currently he is a Ph.D. student in the Department of Computer Science and Electrical Engineering at the University of Maryland Baltimore County, USA. His primary research interests are in the area of statistical signal processing in target detection and optical communication systems.



**Mr. Darren Emge** earned a Bachelor's degree in Physics and a Master's degree in Electrical Engineering from the University of Maryland MD, USA, in 1992 and 2000, respectively. He then work for the Department of Neurology at the University of Maryland Medical System. During this time, he developed analysis algorithms for evoked response potentials, functional neuronal mapping and the acute pain center. In 2001, he joined the US Army SBCCOM passive detection team investigating advanced mathematical techniques for the detection of chemical vapors. Recently he has

transitioned to the active detection team where he is applying his knowledge to active detection systems.



**Dr. Andrzej Cichocki** received the M.Sc. (with honors), Ph.D. and Habilitate Doctorate (Dr.Sc.) degrees in Electrical Engineering, from the Warsaw University of Technology, Poland, in 1972, 1975, and 1982, respectively. He is the co-author of three international and successful books (two of them translated to Chinese): Adaptive Blind Signal and Image Processing, John Wiley 2002, MOS Switched-Capacitor and Continuous-Time Integrated Circuits and Systems (Springer-Verlag, 1989) and Neural Networks for Optimization and Signal Processing (J. Wiley and Teubner Verlag, 1993/1994) and author or co-author of more than 200 papers. He is Editor-in-Chief of Journal Computational Intelligence and Neuroscience and Associate Editor of IEEE Transactions on Neural Networks. Since 1997 he is the head of the laboratory for Advanced Brain Signal Processing in the Riken Brain Science Institute, Japan.