Nonnegative matrix factorization (NMF) for determined and underdetermined BSS problems

Ivica Kopriva

Ruđer Bošković Institute

e-mail: ikopriva@irb.hr ikopriva@gmail.com
Web: http://www.lair.irb.hr/ikopriva/

Course outline

- Motivation with illustration of applications (lecture I)
- Mathematical preliminaries with principal component analysis (PCA)? (lecture II)
- Independent component analysis (ICA) for linear static problems: information-theoretic approaches (lecture III)
- ICA for linear static problems: algebraic approaches (lecture IV)
- ICA for linear static problems with noise (lecture V)
- Dependent component analysis (DCA) (lecture VI)

Course outline

- Underdetermined blind source separation (BSS) and sparse component analysis (SCA) (lecture VII/VIII)
- Nonnegative matrix factorization (NMF) for determined and underdetermined BSS problems (lecture VIII/IX)
- BSS from linear convolutive (dynamic) mixtures (lecture X/XI)
- Nonlinear BSS (lecture XI/XII)
- Tensor factorization (TF): BSS of multidimensional sources and feature extraction (lecture XIII/XIV)

Seminar problems

1. Blind separation of two uniformly distributed signals with maximum likelihood (ML) and AMUSE/SOBI independent component analysis (ICA) algorithm. Blind separation of two speech signals with ML and AMUSE/SOBI ICA algorithm. **Theory, MATLAB demonstration and comments of the results.**

2. Blind decomposition/segmentation of multispectral (RGB) image using ICA, dependent component analysis (DCA) and nonnegative matrix factorization (NMF) algorithms. **Theory, MATLAB demonstration and comments of the results.**

3. Blind separation of acoustic (speech) signals from convolutive dynamic mixture. **Theory, MATLAB demonstration and comments of the results.**

Seminar problems

4. Blind separation of images of human faces using ICA and DCA algorithms (innovation transform and ICA, wavelet packets and ICA) **Theory, MATLAB demonstration and comments of the results.**

5. Blind decomposition of multispectral (RGB) image using sparse component analysis (SCA): clustering + L_p norm ($0<p\leq 1$) minimization. **Theory, MATLAB demonstration and comments of the results.**

6. Blind separation of four sinusoidal signals from two static mixtures (a computer generated example) using sparse component analysis (SCA): clustering + L_p norm ($0<p\leq 1$) minimization in frequency (Fourier) domain. **Theory, MATLAB demonstration and comments of the results.**

Seminar problems

7. Blind separation of three acoustic signals from two static mixtures (a computer generated example) using sparse component analysis (SCA): clustering + L_p norm ($0<p\leq 1$) minimization in time-frequency (short-time Fourier) domain. **Theory, MATLAB demonstration and comments of the results.**

8. Blind extraction of five pure components from mass spectra of two static mixtures of chemical compounds using sparse component analysis (SCA): clustering a set of single component points + L_p norm ($0<p\leq 1$) minimization in m/z domain. **Theory, MATLAB demonstration and comments of the results.**

9. Feature extraction from protein (mass) spectra by tensor factorization of disease and control samples in joint bases. Prediction of prostate/ovarian cancer. **Theory, MATLAB demonstration and comments of the results.**

Blind source separation

A theory for multichannel blind signal recovery requiring minimum of *a priori* information.

Problem:

$$\mathbf{X}=\mathbf{AS} \quad \mathbf{X}\in R^{NxT},\ \mathbf{A}\in R^{NxM},\ \mathbf{S}\in R^{MxT}$$

Goal: find **A** and **S** based on **X** only.

Solution $\mathbf{X}=\mathbf{A}\mathbf{T}^{-1}\mathbf{T}\mathbf{S}$ must be characterized with $\mathbf{T}=\mathbf{P}\Lambda$ where **P** is permutation and Λ is diagonal matrix i.e.: $\mathbf{Y}\cong\mathbf{P}\Lambda\mathbf{S}$

A. Cichocki, S. Amari, "Adaptive Blind Signal and Image Processing," John Wiley, 2002.

Independent component analysis

- Number of mixtures *N* must be greater than or equal to *M*.

- source signals $s_i(t)$ must be statistically independent.

$$\mathrm{p}(\mathbf{s}) = \prod_{m=1}^{M} p_m(s_m)$$

- source signals $s_m(t)$, except one, must be non-Gaussian.

$$\{C_n(s_m) \neq 0\}_{m=1}^{M} \qquad \forall n > 2$$

- mixing matrix **A** must be nonsingular.

$$\mathbf{W} \cong \mathbf{A}^{-1}$$

Nonnegative matrix factorization (NMF)

Many BSS problems arising in imaging, chemo- and/or bioinformatics are described by superposition of non-negative latent variables (sources):

$$\mathbf{X} = \mathbf{AS} \quad \mathbf{X} \in \mathbb{R}_{0+}^{N \times T},\ \mathbf{A} \in \mathbb{R}_{0+}^{N \times M} \ \textit{and} \ \mathbf{S} \in \mathbb{R}_{0+}^{M \times T}$$

where *N* represents number of sensors, *M* represents number of sources and *T* represents number of samples.

Thus, solution of related decomposition problem can be obtained by imposing non-negativity constraints on **A** and **S**, to narrow down number of possible decomposition of **X**. This leads to NMF algorithms.

Due to non-negativity constraints some other constraints (statistical independence) can be relaxed/replaced in applications where they are not fulfilled. (Non-negative sources are partially dependent).

Nonnegative matrix factorization

Modern approaches to NMF problems have been initiated by Lee-Seung' Nature paper, Ref. [1], where it is proposed to estimate **A** and **S** through alternative minimization procedure of the possibly two different cost functions:

Set Randomly initialize: $\mathbf{A}^{(0)}$, $\mathbf{S}^{(0)}$,

For *k*=1,2,…, *until convergence* do

Step 1: $$\mathbf{S}^{(k+1)} = \arg\min_{s_{mt}\geq 0} D_{\mathbf{s}}\left(\mathbf{X}\middle\|\mathbf{A}^{(k)}\mathbf{S}\right)_{\mathbf{S}^{(k)}}$$

Step 2: $$\mathbf{A}^{(k+1)} = \arg\min_{a_{nm}\geq 0} D_{\mathbf{A}}\left(\mathbf{X}\middle\|\mathbf{A}\mathbf{S}^{(k+1)}\right)_{\mathbf{A}^{(k)}}$$

If both cost functions represent squared Euclidean distance (Froebenius norm) we obtain alternating least square (ALS) approach to NMF.

1. D D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," Nature **401** (6755), 788-791 (1999).

Nonnegative matrix factorization

ALS-based NMF:

$$\left(\mathbf{A}^*, \mathbf{S}^*\right) = \arg\min_{\mathbf{A},\mathbf{S}} D\left(\mathbf{X}\middle\|\mathbf{AS}\right) = \frac{1}{2}\left\|\mathbf{X}-\mathbf{AS}\right\|_2^2 \quad s.t.\ \mathbf{A}\geq\mathbf{0}, \mathbf{S}\geq\mathbf{0}$$

There are two problems with above factorization:

1) Minimization of the square of Euclidean norm of approximation error **E=X-AS** is from the maximum likelihood viewpoint justified only if error distribution is Gaussian:

$$p\left(\mathbf{X}\middle|\mathbf{A},\mathbf{S}\right) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{\left\|\mathbf{X}-\mathbf{AS}\right\|_2^2}{2\sigma^2}\right)$$

2) In many instances non-negativity constraints imposed on **A** and **S** do not suffice to obtain solution that is unique up to standard BSS indeterminacies: permutation and scaling.

Nonnegative matrix factorization

In relation to original Lee-Seung NMF algorithm, [1], additional constraints are necessary to obtain factorization unique up to permutation and scaling. Generalization that involves constraints is given in [2]:

$$D\left(\mathbf{X}\|\mathbf{AS}\right)=\frac{1}{2}\left\|\mathbf{X}-\mathbf{AS}\right\|_2^2+\alpha_{\mathbf{S}}J_{\mathbf{S}}(\mathbf{S})+\alpha_{\mathbf{A}}J_{\mathbf{A}}(\mathbf{A})$$

where $J_{\mathbf{S}}(\mathbf{S})=\sum_{mt}s_{mt}$ and $J_{\mathbf{A}}(\mathbf{A})=\sum_{nm}a_{nm}$ are sparseness constraints that correspond with L_1-norm of **S** and **A** respectively. $\alpha_{\mathbf{S}}$ and $\alpha_{\mathbf{A}}$ are regularization constants. Gradient components in matrix form are:

$$\frac{\partial D\left(\mathbf{A},\mathbf{S}\right)}{\partial a_{nm}}=\left[-\mathbf{X}\mathbf{S}^{\mathrm{T}}+\mathbf{A}\mathbf{S}\mathbf{S}^{\mathrm{T}}\right]_{nm}+\alpha_{\mathbf{A}}\frac{\partial J_{\mathbf{A}}(\mathbf{A})}{\partial a_{nm}}$$

$$\frac{\partial D\left(\mathbf{A},\mathbf{S}\right)}{\partial s_{mt}}=\left[-\mathbf{A}^{\mathrm{T}}\mathbf{X}+\mathbf{A}^{\mathrm{T}}\mathbf{A}\mathbf{S}\right]_{mt}+\alpha_{\mathbf{S}}\frac{\partial J_{\mathbf{S}}(\mathbf{S})}{\partial s_{mt}}$$

2. A. Cichocki, R. Zdunek, and S. Amari, "Csiszár's Divergences for Non-negative Matrix Factorization: Family of New Algorithms," LNCS **3889,** 32-39 (2006).

Maximum a posteriori probability BSS/NMF

Maximization of a-posterior probability (MAP) $P(\mathbf{A},\mathbf{S}|\mathbf{X})$ yields minimum L_1-norm as the solution:

$$\left(\mathbf{A}^*,\mathbf{S}^*\right)=\max_{\mathbf{AS}=\mathbf{X}} P\left(\mathbf{A},\mathbf{S}\middle|\mathbf{X}\right)\propto\max_{\mathbf{AS}=\mathbf{X}} P\left(\mathbf{X}\middle|\mathbf{A},\mathbf{S}\right)P(\mathbf{A})P\left(\mathbf{S}\right)\ \ s.t.\ \mathbf{A}\geq\mathbf{0},\mathbf{S}\geq\mathbf{0}$$

Above formulation is equivalent to maximizing likelihood probability $P(\mathbf{X}|\mathbf{A},\mathbf{S})$ and maximizing prior probabilities $P(\mathbf{A})$ and $P(\mathbf{S})$. Assuming normal distribution of approximation error $\mathbf{E}=\mathbf{X}-\mathbf{AS}$ yields:

$$\left(\mathbf{A}^*,\mathbf{S}^*\right)=\arg\min_{(\mathbf{A},\mathbf{S})}\frac{1}{2}\left\|\mathbf{X}-\mathbf{AS}\right\|_2^2+\alpha_{\mathbf{S}}J_{\mathbf{S}}(\mathbf{S})+\alpha_{\mathbf{A}}J_{\mathbf{A}}(\mathbf{A})\ \ s.t.\ \mathbf{A}\geq\mathbf{0},\mathbf{S}\geq\mathbf{0}.$$

Maximum a posteriori probability BSS/NMF

Above formulation is equivalent to maximizing likelihood probability *P*(**X**|**A**,**S**) and maximizing prior probabilities *P*(**A**) and *P*(**S**). Assuming normal distribution of approximation error **E**=**X**-**AS**, non-informative prior on **A**: *P*(**A**)=const and *Laplacian* (sparse) prior on $P(\mathbf{S}) = \exp -\left(|\mathbf{s}_1| + \ldots + |\mathbf{s}_M|\right)$ yields

$$\left(\mathbf{A}^*, \mathbf{S}^*\right) = \arg\min_{(\mathbf{A},\mathbf{S})} \frac{1}{2}\|\mathbf{X} - \mathbf{AS}\|_2^2 + \alpha_S \|\mathbf{S}\|_1 \quad s.t.\, \mathbf{A} \geq \mathbf{0}, \mathbf{S} \geq \mathbf{0}.$$

It is possible to select other than Laplacian prior for *P*(**S**) that leads to sparseness constraint different than L_1-norm of **S**:

$$\left(\mathbf{A}^*, \mathbf{S}^*\right) = \arg\min_{(\mathbf{A},\mathbf{S})} \frac{1}{2}\|\mathbf{X} - \mathbf{AS}\|_2^2 + \alpha_S \|\mathbf{S}\|_p \quad s.t.\, 0 < p \leq 1, \mathbf{A} \geq \mathbf{0}, \mathbf{S} \geq \mathbf{0}.$$

Nonnegative matrix factorization

Since NMF problem deals with non-negative variables the idea is to ensure non-negativity of **A** and **S** through learning automatically. That is achieved by multiplicative learning equations:

$$\mathbf{A} \leftarrow \mathbf{A} \otimes \frac{\nabla_{\mathbf{A}}^{-} D(\mathbf{A},\mathbf{S})}{\nabla_{\mathbf{A}}^{+} D(\mathbf{A},\mathbf{S})} \qquad \mathbf{S} \leftarrow \mathbf{S} \otimes \frac{\nabla_{\mathbf{S}}^{-} D(\mathbf{A},\mathbf{S})}{\nabla_{\mathbf{S}}^{+} D(\mathbf{A},\mathbf{S})}$$

where $\otimes$ denotes component (entry) wise multiplication, $\nabla_{\mathbf{A}}^{-} D(\mathbf{A},\mathbf{S})$ and $\nabla_{\mathbf{A}}^{+} D(\mathbf{A},\mathbf{S})$ denote respectively negative and positive part of the gradient $\nabla_{\mathbf{A}} D(\mathbf{A},\mathbf{S})$. Likewise, $\nabla_{\mathbf{S}}^{-} D(\mathbf{A},\mathbf{S})$ and $\nabla_{\mathbf{S}}^{+} D(\mathbf{A},\mathbf{S})$ are negative and positive part of the gradient $\nabla_{\mathbf{S}} D(\mathbf{A},\mathbf{S})$.

When gradients converge to zero corrective terms converge to one. Since learning equations include multiplications and divisions of non-negative terms, non-negativity is ensured automatically.

Nonnegative matrix factorization

Multiplicative learning rules for NMF based on regularized squared L_2-norm of the approximation are obtained as:

$$\mathbf{A} \leftarrow \mathbf{A} \otimes \frac{\left[\mathbf{X}\mathbf{S}^{\mathrm{T}} - \alpha_{\mathbf{A}} \dfrac{\partial J_{\mathbf{A}}(\mathbf{A})}{\partial \mathbf{A}}\right]_{+}}{\mathbf{A}\mathbf{S}\mathbf{S}^{\mathrm{T}} + \varepsilon \mathbf{1}_{NM}} \qquad \mathbf{S} \leftarrow \mathbf{S} \otimes \frac{\left[\mathbf{A}^{\mathrm{T}}\mathbf{X} - \alpha_{\mathbf{S}} \dfrac{\partial J_{\mathbf{S}}(\mathbf{S})}{\partial \mathbf{S}}\right]_{+}}{\mathbf{A}^{\mathrm{T}}\mathbf{A}\mathbf{S} + \varepsilon \mathbf{1}_{MT}}$$

where $[x]_+=\max\{\varepsilon,x\}$ with small ε. For L_1-norm based regularization, derivatives of sparseness constraints in above expressions are equal to 1, i.e.:

$$\mathbf{A} \leftarrow \mathbf{A} \otimes \frac{\left[\mathbf{X}\mathbf{S}^{\mathrm{T}} - \alpha_{\mathbf{A}} \mathbf{1}_{NM}\right]_{+}}{\mathbf{A}\mathbf{S}\mathbf{S}^{\mathrm{T}} + \varepsilon \mathbf{1}_{NM}} \qquad \mathbf{S} \leftarrow \mathbf{S} \otimes \frac{\left[\mathbf{A}^{\mathrm{T}}\mathbf{X} - \alpha_{\mathbf{S}} \mathbf{1}_{MT}\right]_{+}}{\mathbf{A}^{\mathrm{T}}\mathbf{A}\mathbf{S} + \varepsilon \mathbf{1}_{MT}}$$

Nonnegative matrix factorization

NMF through minimization of Froebenius norm is optimal when data are corrupted by additive Gaussian noise. Another cost function that is used most often for NMF is Kullback-Leibler divergence, also called I-divergence:

$$D\left(\mathbf{X}\middle\|\mathbf{AS}\right)=\sum_{nt}\left(x_{nt}\ln\frac{x_{nt}}{\left[\mathbf{AS}\right]_{nt}}-x_{nt}+\left[\mathbf{AS}\right]_{nt}\right)$$

It can be shown that minimization of Kullback-Leibler divergence is equivalent to the maximization of the Poisson likelihood

$$L\left(\mathbf{X}\middle|\mathbf{A},\mathbf{S}\right)=\prod_{nt}\left(\frac{\left[\mathbf{AS}\right]_{nt}}{x_{nt}!}\exp\left(-\left[\mathbf{AS}\right]_{nt}\right)\right)$$

2. A. Cichocki, R. Zdunek, and S. Amari, "Csiszár's Divergences for Non-negative Matrix Factorization: Family of New Algorithms," LNCS **3889,** 32-39 (2006).

Nonnegative matrix factorization

Calculating gradients of I-divergence cost function w.r.t. a_{nm} and s_{mt} the following learning rules in MATLAB notation are obtained

$$\mathbf{S}^{(k+1)} = \left(\mathbf{S}^{(k)} \otimes \left(\mathbf{A}^{\mathrm{T}} \left(\mathbf{X} \oslash \left(\mathbf{A}\mathbf{S}^{(k)} \right) \right) \right)^{.[\omega]} \right)^{.[1+\alpha_{\mathbf{S}}]}$$

$$\mathbf{A}^{(k+1)} = \left(\mathbf{A}^{(k)} \otimes \left(\left(\mathbf{X} \oslash \left(\mathbf{A}^{(k)} \mathbf{S} \right) \right) \mathbf{S}^{\mathrm{T}} \right)^{.[\omega]} \right)^{.[1+\alpha_{\mathbf{A}}]}$$

where $\otimes$ denotes component-wise multiplication, and $\oslash$ denotes component-wise division. Relaxation parameter $\omega \in (0,2]$ provides improvement of the convergence, while $\alpha_{\mathbf{S}} \geq 0$ and $\alpha_{\mathbf{A}} \geq 0$ are sparseness constraints that are typically confined in the interval [0.001, 0.005].

Nonnegative matrix factorization

In order to obtain NMF algorithms optimal for different statistics of data and noise the α-divergence cost function can be used

$$D\left(\mathbf{X}\middle\|\mathbf{AS}\right)=\frac{1}{\alpha(\alpha-1)}\sum_{nt}\left(x_{nt}^{\alpha}\left[\mathbf{AS}\right]_{nt}^{1-\alpha}-\alpha x_{nt}+(\alpha-1)\left[\mathbf{AS}\right]_{nt}\right)$$

I-divergence is obtained in the limit when $\alpha\to1$, and dual Kullback-Leibler divergence when $\alpha\to0$. Using MATLAB notation the following update rules are obtained for $\alpha\neq0,1$.

$$\mathbf{S}\leftarrow\left(\mathbf{S}.*\left(\mathbf{A}^{\mathrm{T}}*\left(\mathbf{X}./\left[\mathbf{AS}+\varepsilon\mathbf{1}_{NT}\right]_{+}\right)_{.}^{\alpha}\right)_{.}^{\omega/\alpha}\right)_{.}^{1+\alpha_{\mathbf{S}}}$$

$$\mathbf{A}\leftarrow\left(\mathbf{A}.*\left(\left(\mathbf{X}./\left[\mathbf{AS}+\varepsilon\mathbf{1}_{NT}\right]_{+}\right)_{.}^{\alpha}\mathbf{S}^{\mathrm{T}}\right)_{.}^{\omega/\alpha}\right)_{.}^{1+\alpha_{\mathbf{A}}}$$

$$\mathbf{A}\leftarrow\mathbf{A}*diag\left(1./sum(\mathbf{A},1)\right)$$

Hierarchical ALS NMF

Local or hierarchical ALS NMF algorithms were recently derived in [3-5]. They are biologically plausible and employ minimization of the global cost function to learn the mixing matrix and minimization of set of local cost functions to learn the sources. Global cost function can for example be squared Euclidean norm:

$$D\left(\mathbf{X}\middle\|\mathbf{AS}\right)=\frac{1}{2}\left\|\mathbf{X}-\mathbf{AS}\right\|_2^2+\alpha_{\mathbf{S}}J_{\mathbf{S}}(\mathbf{S})+\alpha_{\mathbf{A}}J_{\mathbf{A}}(\mathbf{A})$$

Local cost functions can be also squared Euclidean norms

$$D^{(m)}\left(\mathbf{X}^{(m)}\middle\|\mathbf{a}_m\underline{\mathbf{s}}_m\right)=\frac{1}{2}\left\|\mathbf{X}^{(m)}-\mathbf{a}_m\underline{\mathbf{s}}_m\right\|_2^2+\alpha_{\mathbf{s}}^{(m)}J_{\mathbf{S}}(\underline{\mathbf{s}}_m)+\alpha_{\mathbf{a}}^{(m)}J_{\mathbf{a}}(\mathbf{a}_m)\quad m=1,\ldots,M$$

$$\mathbf{X}^{(m)}=\mathbf{X}-\sum_{j\neq m}\mathbf{a}_j\underline{\mathbf{s}}_j$$

3. A. Cichocki, R. Zdunek, S.I. Amari, Hierarchical ALS Algorithms for Nonnegative Matrix Factorization and 3D Tensor Factorization, LNCS **4666** (2007) 169-176

4. A. Cichocki, A-H. Phan, R. Zdunek, and L.-Q. Zhang, "Flexible component analysis for sparse, smooth, nonnegative coding or representation," LNCS **4984**, 811-820 (2008).

5. A. Cichocki, R. Zdunek, S. Amari, Nonnegative Matrix and Tensor Factorization, IEEE Sig. Proc. Mag. **25** (2008) 142-145.

Hierarchical ALS NMF

Minimization of above cost functions in ALS manner with sparseness constraints imposed on **A** and/or **S** yields

$$\left\{ \underline{\mathbf{s}}_m \leftarrow \left[\mathbf{a}_m^{\mathrm{T}} \mathbf{X}^{(m)} - \alpha_{\mathbf{s}}^{(m)} \mathbf{1}_{1\times T} \right]_+ \right\}_{m=1}^{M}$$

$$\mathbf{A} \leftarrow \left[\left(\mathbf{X}\mathbf{S}^{\mathrm{T}} - \alpha_{\mathbf{A}} \mathbf{1}_{N\times M} \right) \left(\mathbf{S}\mathbf{S}^{\mathrm{T}} + \lambda \mathbf{I}_M \right)^{-1} \right]_+$$

$$\left\{ \mathbf{a}_m \leftarrow \mathbf{a}_m \Big/ \left\| \mathbf{a}_m \right\|_2 \right\}_{m=1}^{M}$$

where $\mathbf{I}_{1\times T}$ is an $M\times M$ identity matrix, $\mathbf{1}_{1\times T}$ and $\mathbf{1}_{N\times M}$ are row vector and matrix with all entries equal to one and $[\xi]_+ = \max\{\varepsilon, \xi\}$ (e.g., $\varepsilon = 10^{-16}$).

Regularization constant λ changes as a function of the iteration index as $\lambda_k = \lambda_0 \exp(-k/\tau)$ (with λ_0 = 100 and τ = 0.02 in the experiments).

Multilayer NMF

Great improvement in the performance of the NMF algorithms is obtained when they are applied in the multilayer mode [89,90], whereas sequential decomposition of the nonnegative matrices is performed as follows.

In the first layer, the basic approximation decomposition is performed:

$$\mathbf{X} \cong \mathbf{A}^{(1)}\mathbf{S}^{(1)} \in \mathbb{R}_{0+}^{N\times T}$$

In the second layer result from the first layer is used to build up new input data matrix for the second layer $\mathbf{X} \leftarrow \mathbf{S}^{(1)} \in \mathbb{R}_{0+}^{M\times T}$. This yields $\mathbf{X}^{(1)} \cong \mathbf{A}^{(2)}\mathbf{S}^{(2)} \in \mathbb{R}_{0+}^{M\times T}$.

After *L* layers data decomposes as follows $\mathbf{X} \cong \mathbf{A}^{(1)}\mathbf{A}^{(2)}\cdots\mathbf{A}^{(L)}\mathbf{S}^{(L)}$

6. A. Cichocki, and R. Zdunek, "Multilayer Nonnegative Matrix Factorization," El. Letters **42**, 947-948 (2006).
7. A. Cichocki, R. Zdunek, A. H. Phan, S. Amari, Nonnegative Matrix and Tensor Factorizations-Applications to Exploratory Multi-way Data Analysis and Blind Source Separation, John Wiley, 2009.

Multi-start initialization for NMF algorithms

Combined optimization of the cost function $D(\mathbf{X}||\mathbf{AS})$ with respect to **A** and **S** is nonconvex optimization problem. Hence, some strategy is necessary to decrease probability that optimization process will get stuck in some local minima. Such procedure is outlined with the following pseudo code: Select R-number of restarts, K_i number of alternating steps, K_f number of final alternating steps.

for $r = 1,\ldots,R$ **do**

Initialize randomly $\mathbf{A}^{(0)}$ and $\mathbf{S}^{(0)}$

$\{\mathbf{A}^{(r)},\mathbf{S}^{(r)}\} \leftarrow$ nmf_algorithm($\mathbf{X},\mathbf{A}^{(0)},\mathbf{S}^{(0)},K_i$);

compute $d=D(\mathbf{X}||\mathbf{A}^{(r)}\mathbf{S}^{(r)})$;

end

$r_{min}=\text{argmin}_{1\le n\le R} d_r$;

$\{\mathbf{A},\mathbf{S}\} \leftarrow$nmf_algorithm($\mathbf{X},\mathbf{A}^{(rmin)},\mathbf{S}^{(rmin)},K_f$);

$L_{1/2}$ –sparsity constrained NMF

Very recently it has been proven in [8] $L_{1/2}$-regularizer is the most sparse and robust among L_p regularizers when $1/2 \le p < 1$, and when $0 < p < 1/2$, the L_p regularizers have similar properties as the $L_{1/2}$ regularizer. In [9] $L_{1/2}$-NMF algorithm has been derived for unmixing hyperspectral image. The algorithm is derived as solution of:

$$\left(\mathbf{A}^*, \mathbf{S}^*\right) = \arg\min_{(\mathbf{A},\mathbf{S})} \frac{1}{2}\left\|\mathbf{X} - \mathbf{AS}\right\|_2^2 + \alpha_S \left\|\mathbf{S}\right\|_{1/2} \; s.t. \; \mathbf{A} \ge \mathbf{0}, \mathbf{S} \ge \mathbf{0}.$$

Multiplicative update rules are:

$$\mathbf{A} \leftarrow \mathbf{A} \otimes \frac{\mathbf{X}\mathbf{S}^{\mathrm{T}}}{\mathbf{A}\mathbf{S}\mathbf{S}^{\mathrm{T}} + \varepsilon \mathbf{1}_{NM}} \quad \mathbf{S} \leftarrow \mathbf{S} \otimes \frac{\mathbf{A}^{\mathrm{T}}\mathbf{X}}{\mathbf{A}^{\mathrm{T}}\mathbf{A}\mathbf{S} + \frac{\lambda}{2}\left[\mathbf{S} + \varepsilon \mathbf{1}_{MT}\right]_+^{-1/2}}$$

8. X. ZongBen, Z. Hai, W. Yao, C. XiangYu, L. Yong, L_{1-2} regularization, Science China, series F, 53 (2010) 1159-1169.
9. Y. Qian, S. Jia, J. Zhou, A. Robles-Kelly, "Hyperspectral unmixing via $L_{1/2}$ Sparsity-Constrained Nonnegative Matrix Factorization," IEEE Transactions on Geoscience abd Remote Sensing, vol. 49, No. 11, 4282-4297, 2011.

Non-negative matrix under-approximation (NMU)

A sequential approach to NMF has been recently proposed in [10] by estimating rank-1 one factors $\mathbf{a}_m\mathbf{s}_m$ one at a time. Each time $\mathbf{a}_m\mathbf{s}_m$ is estimated it is removed from $\mathbf{X} \rightarrow \mathbf{X}-\mathbf{a}_m\mathbf{s}_m$. To prevent subtraction from being negative the under-approximation constraint is imposed on $\mathbf{a}_m\mathbf{s}_m$: $\mathbf{a}_m\mathbf{s}_m \leq \mathbf{X}$.

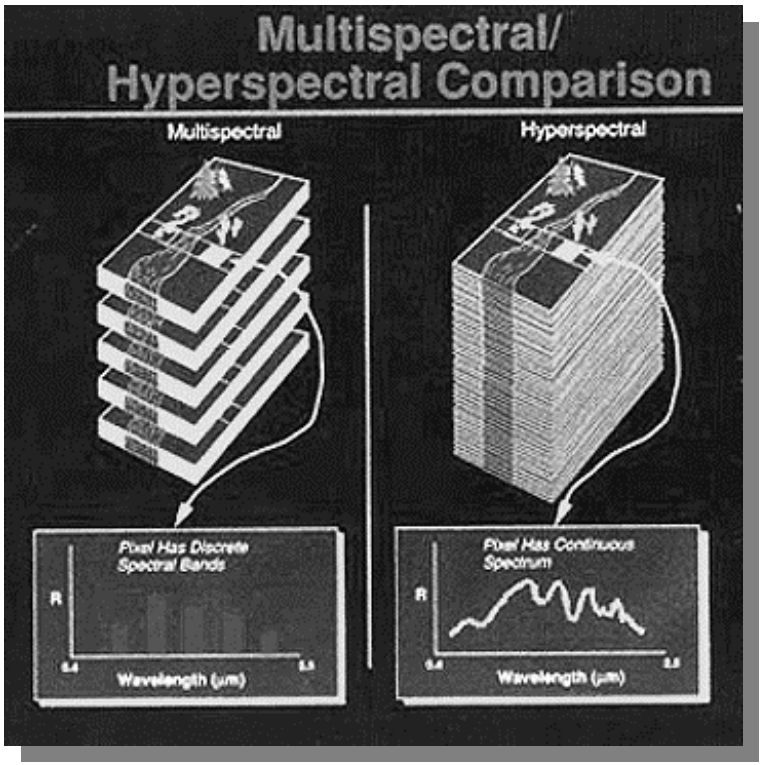
Hence, the NMU algorithm is obtained as a solution of:

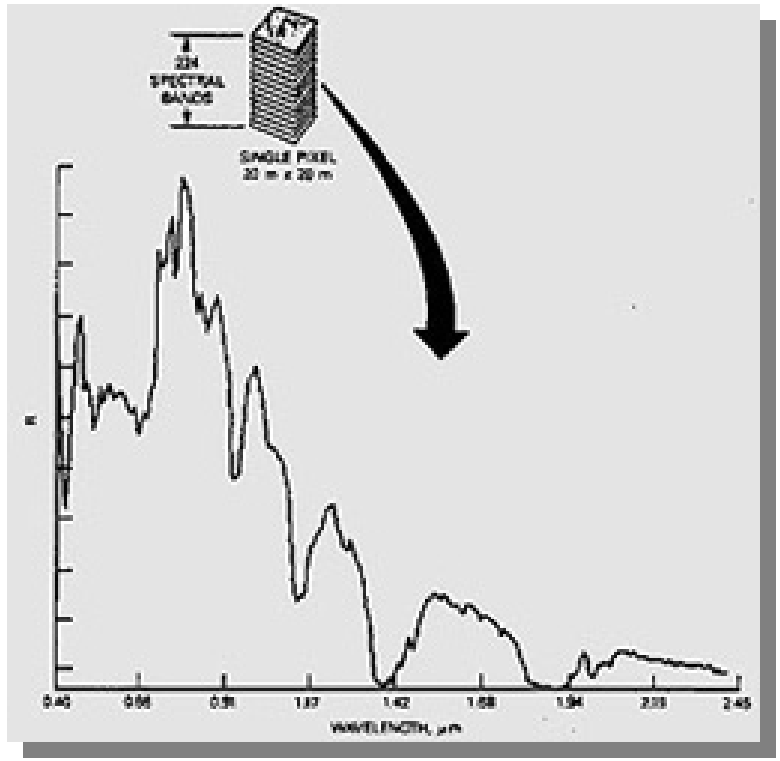
$$\left(\mathbf{A}^*, \mathbf{S}^*\right) = \arg\min_{(\mathbf{A},\mathbf{S})} \frac{1}{2}\left\|\mathbf{X}-\mathbf{AS}\right\|_2^2 \; s.t. \; \mathbf{A} \geq \mathbf{0}, \mathbf{S} \geq \mathbf{0}, \; \mathbf{AS} \leq \mathbf{X}.$$

The underapproximation constraint ensures sparse (parts based) factorization of **X**. Since no explicit regularization is used there are no difficulties associated with selecting values of regularization constants. MATLAB code for NMU algorithm is available at: *http://www.core.ucl.ac.be/~ngillis/papers/recursiveNMU.m*

10. N. Gillis, and F. Glineur, "Using underapproximations for sparse nonnegative matrix factorization," Patt. Recog., vol. 43, pp. 1676-1687, 2010.

Unsupervised segmentation of multispectral images





- SPOT- 4 bands, LANDSAT -7 bands, AVIRIS-224 bands (0.38μ-2.4μ);
- Objects with very similar reflectance spectra are *difficult to discriminate*.

Unsupervised segmentation of multispectral images

Hyperspectral/multispectral image and static linear mixture model. For image consisting of N bands and M materials linear data model is assumed:

$$\mathbf{X} = \mathbf{AS} = \sum_{m=1}^{M} \mathbf{a}_m \mathbf{s}_m$$

$$\left[\mathbf{a}_1 \quad \mathbf{a}_2 \quad \ldots \quad \mathbf{a}_M\right] \equiv \mathbf{A}$$

$$\left[\mathbf{s}_1 \quad \mathbf{s}_2 \quad \ldots \quad \mathbf{s}_M\right]^T \equiv \mathbf{S}$$

X - measured data intensity matrix, $\mathbf{X} \in \mathbb{R}_{0+}^{N \times T}$

S - unknown class matrix, $\mathbf{S} \in \mathbb{R}_{0+}^{M \times T}$

A – unknown spectral reflectance matrix. $\mathbf{A} \in \mathbb{R}_{0+}^{N \times M}$

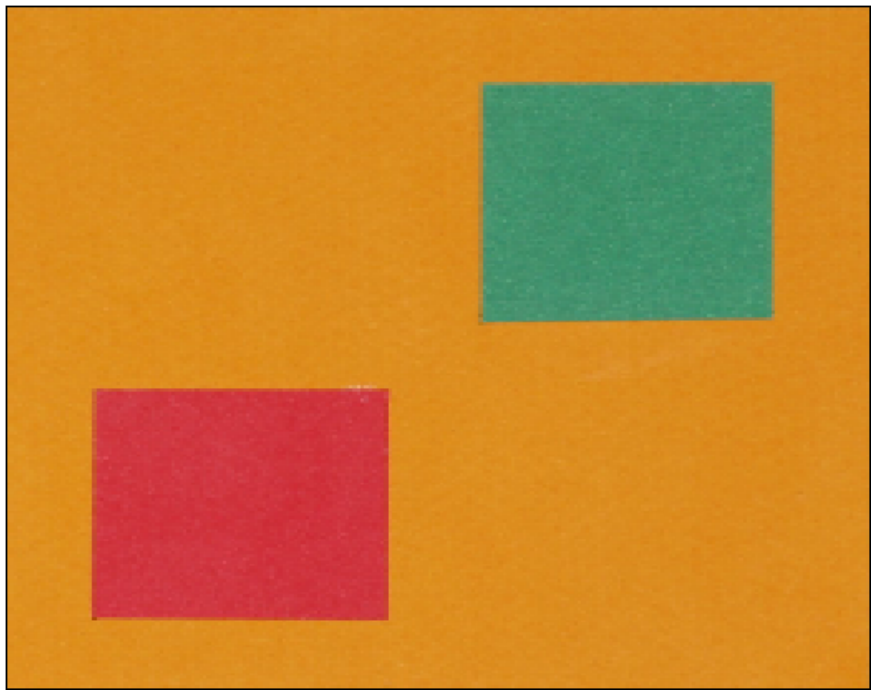
Unsupervised segmentation of multispectral images

Spectral similarity between the sources s_m and s_n implies that corresponding column vectors are close to collinear i.e. $\mathbf{a}_m \cong c\mathbf{a}_n$.

Contribution at certain pixel location t is: $\mathbf{a}_m s_{mt} + \mathbf{a}_n s_{nt} \cong c\mathbf{a}_n s_{mt} + \mathbf{a}_n s_{nt}$. This implies that $\mathbf{s}_n$ and $c\mathbf{s}_m$ are indistinguishable i.e. they are statistically dependent.

Thus, spectral similarity between the sources causes ill-conditioning problems of the basis matrix as well as statistical dependence among the sources. **Both conditions imposed by ICA algorithm on SLMM are not satisfied.**

Unsupervised segmentation of RGB image with three materials: NMF with sparseness constrains, DCA, ICA.[11]



Original RGB image

11. I. Kopriva and A . Cichocki, "Sparse component analysis-based non-probabilistic blind decomposition of low-dimensional multi-spectral images," *Journal of Chemometrics*, vol. 23, Issue 11, pp. 590-597 (2009).

Unsupervised segmentation of multispectral images

Evidently degree of overlap between materials in spatial domain is very small i.e. $s_m(t)*s_n(t)\approx\delta_{nm.}$. Hence RGB image decomposition problem can be solved either with clustering and L_1-norm minimization or with HALS NMF algorithm with sparseness constraints.

For the L_1-norm minimization estimate of the mixing (spectral reflectance matrix) **A** and number of materials M is necessary. For HALS NMF only estimate of M is necessary. Both tasks can be accomplished by data clustering algorithm presented in ref.[12].

Because materials in principle do not overlap in spatial domain it applies $||\mathbf{s}(t)||_0\approx 1$

12. F. M. Naini, G.H. Mohimani, M. Babaie-Zadeh, Ch. Jutten, "Estimating the mixing matrix in Sparse Component Analysis (SCA) based on partial *k*-dimensional subspace clustering," Neurocomputing **71**, 2330-2343 (2008).

Unsupervised segmentation of multispectral images

Assuming unit L_2-norm of $\boldsymbol{a}_m$ we can parameterize column vectors in 3D space by means of azimuth and elevation angles

$$\mathbf{a}_m = [\cos(\varphi_m)\sin(\theta_m) \;\; \sin(\varphi_m)\sin(\theta_m) \;\; \cos(\theta_m)]^{\mathrm{T}}$$

Due to nonnegativity constraints both angles are confined in $[0,\pi/2]$. Now estimation of **A** and M is obtained by means of data clustering algorithm:

- We remove all data points close to the origin for which applies: $\left\{\left|\mathbf{x}(t)\right|_2 \le \varepsilon\right\}_{t=1}^{T}$ where ε represents some predefined threshold.

- Normalize to unit L_2-norm remaining data points $\mathbf{x}(t)$, i.e., $\left\{\mathbf{x}(t) \to \mathbf{x}(t)/\left|\mathbf{x}(t)\right|_2\right\}_{t=1}^{\bar{T}}$

Unsupervised segmentation of multispectral images

• Calculate function *f*(**a**):

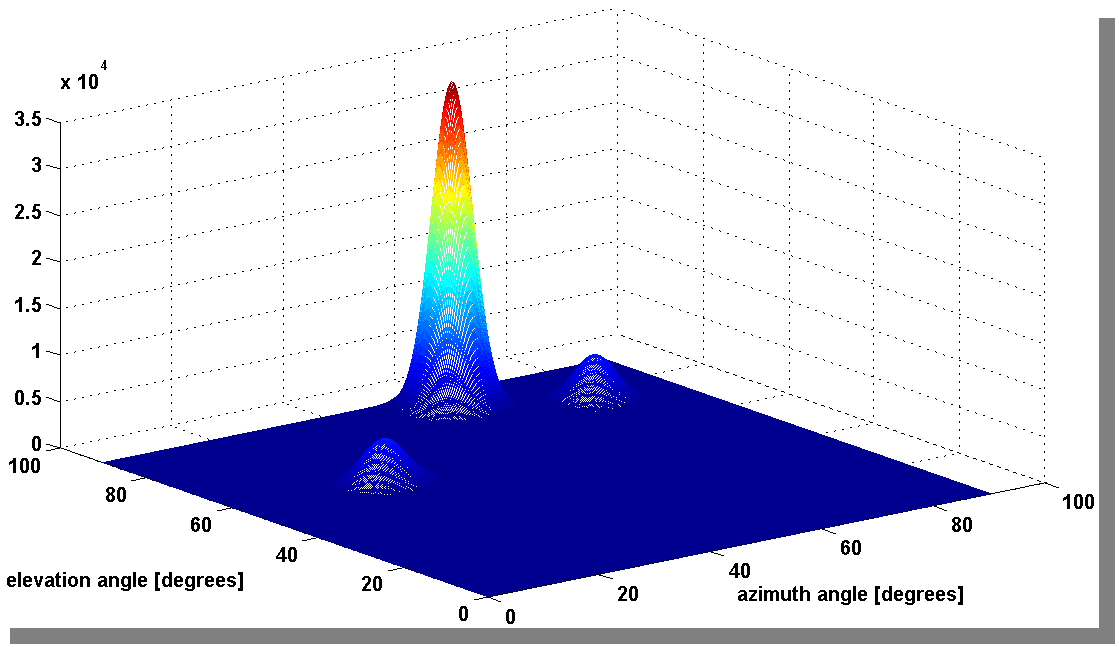
$$f(\mathbf{a}) = \sum_{t=1}^{\bar{T}} \exp\left(-\frac{d^2(\mathbf{x}(t), \mathbf{a})}{2\sigma^2} \right)$$

where $d(\mathbf{x}(t), \mathbf{a}) = \sqrt{1 - (\mathbf{x}(t) \cdot \mathbf{a})^2}$ and $(\mathbf{x}(t) \cdot \mathbf{a})$ denotes inner product. Parameter σ is called dispersion. If set to sufficiently small value, in our experiments this turned out to be $\sigma \approx 0.05$, the value of the function *f*(**a**) will approximately equal the number of data points close to **a**. Thus by varying mixing angles $0 \leq \varphi, \theta \leq \pi/2$ we effectively cluster data.

• Number of peaks of the function *f*(**a**) corresponds with the estimated number of materials *M*. Locations of the peaks correspond with the estimates of the mixing angles $\left\{ \left(\hat{\varphi}_m, \hat{\theta}_m \right) \right\}_{m=1}^{\hat{M}}$, i.e., mixing vectors $\{ \hat{\mathbf{a}}_m \}_{m=1}^{\hat{M}}$.

Unsupervised segmentation of RGB image with three materials: NMF with sparseness constrains, DCA, ICA.

Clustering algorithm is used to estimate number of materials M.



Thee peaks suggest existence of three materials in the RGB image **i.e. *M*=3**.

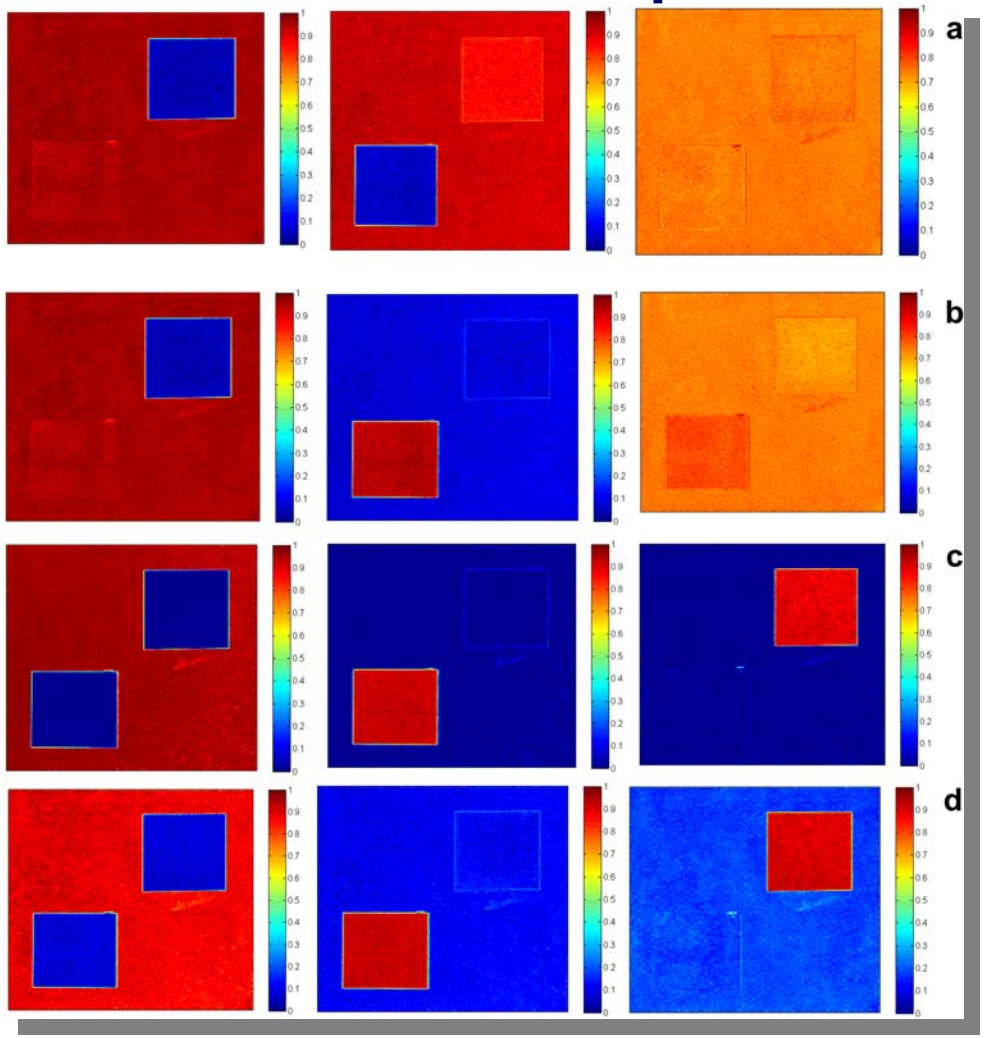
Unsupervised segmentation of RGB image with three materials: NMF with sparseness constrains, DCA, ICA.

Spatial maps of the materials were extracted by NMF with 25 layers, linear programming, ICA and DCA methods.

Extracted spatial maps were rescaled to the interval [0,1] where 0 means full absence of the material and 1 means full presence of the material.

This enables visualization of the quality of decomposition process. Zero probability (absence of the material) is visualized with dark blue color and probability one (full presence of the material) is visualized with dark red color.

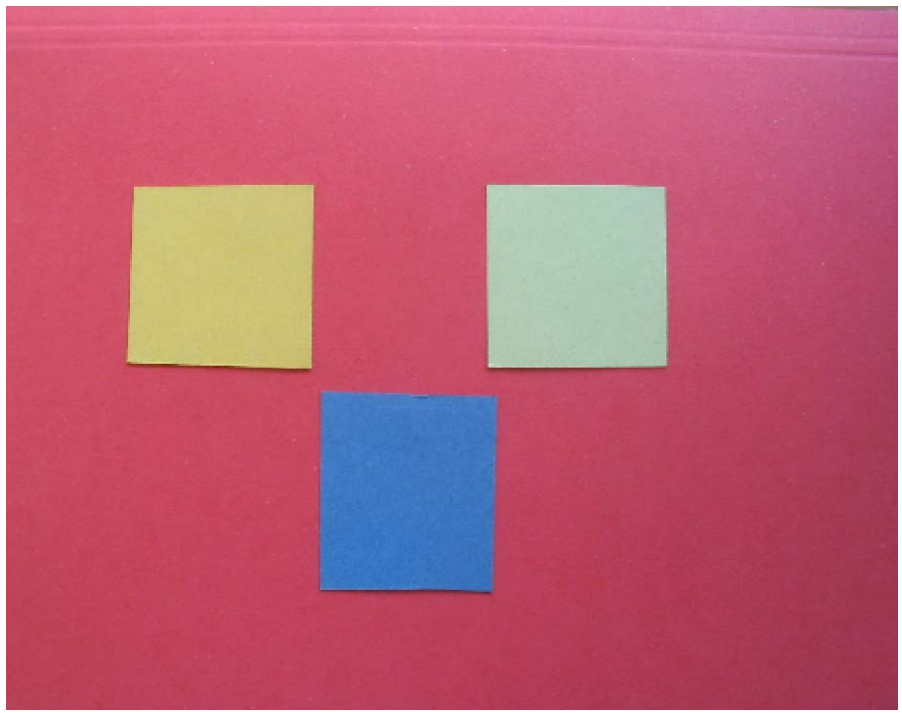
Unsupervised segmentation of RGB image with three materials: NMF with sparseness constrains, DCA, ICA.



a) DCA
b) ICA
c) NMF
d) Linear programming

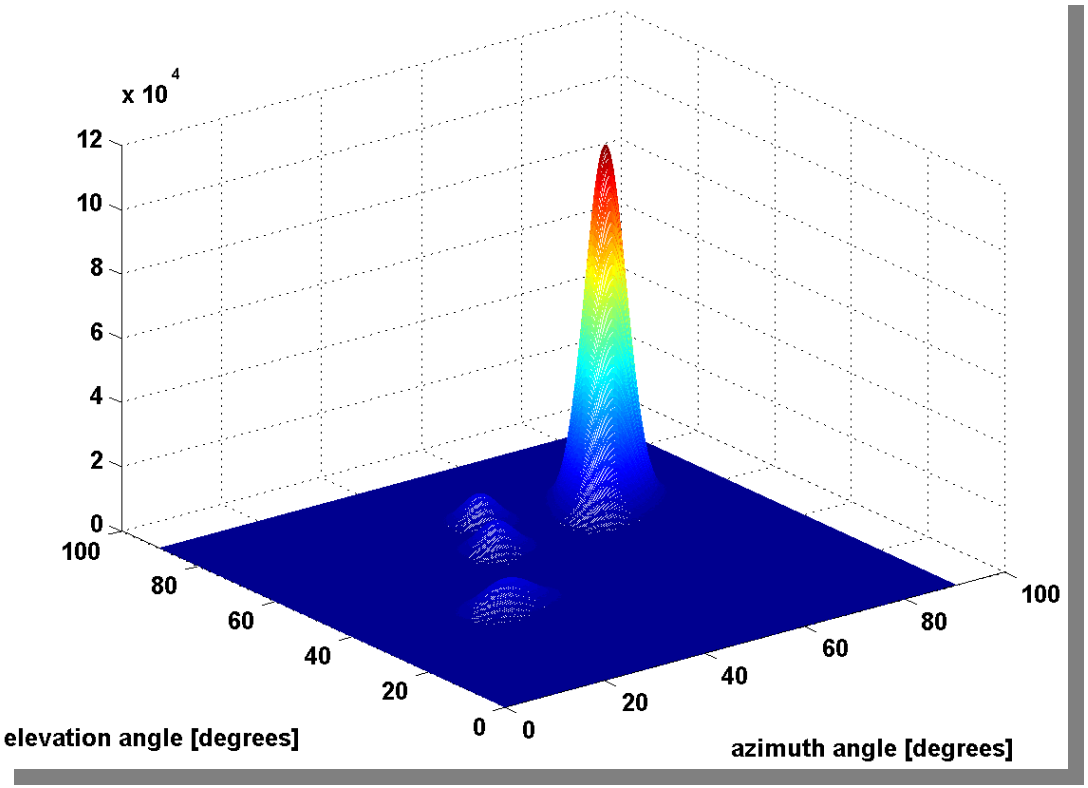
Unsupervised segmentation of multispectral images

Consider blind decomposition of the RGB image (*N*=3) composed of four materials *(M*=4*)*, ref.[11]:



Unsupervised segmentation of multispectral images

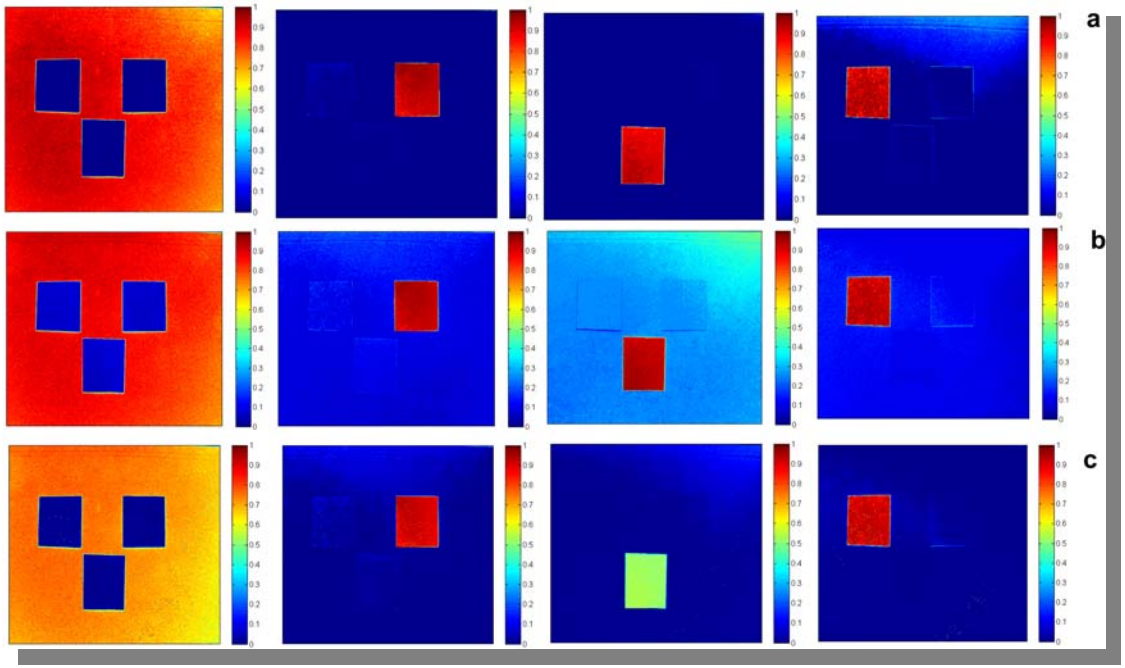
For shown experimental RGB image clustering function is obtained as:



Four peaks suggest existence of four materials in the RGB image **i.e. *M*=4**.

Unsupervised segmentation of multispectral images

Spatial maps of the materials extracted by HALS NMF with 25 layers, linear programming and interior point method [13,14] are obtained as:



a) 25 layers HALS NMF; b) Interior point method, [74,90]; c) Linear programming.

13. S.J. Kim, K. Koh, M. Lustig, S. Boyd, D. Gorinevsky, "An Interior-Point Method for Large-Scale L_1 - Regularized Least Squares,"IEEE Journal of Selected Topics in Signal Processing **1**, 606-617 (2007).
14. http://www.stanford.edu/~boyd/l1_ls/.

Unsupervised segmentation of multispectral images

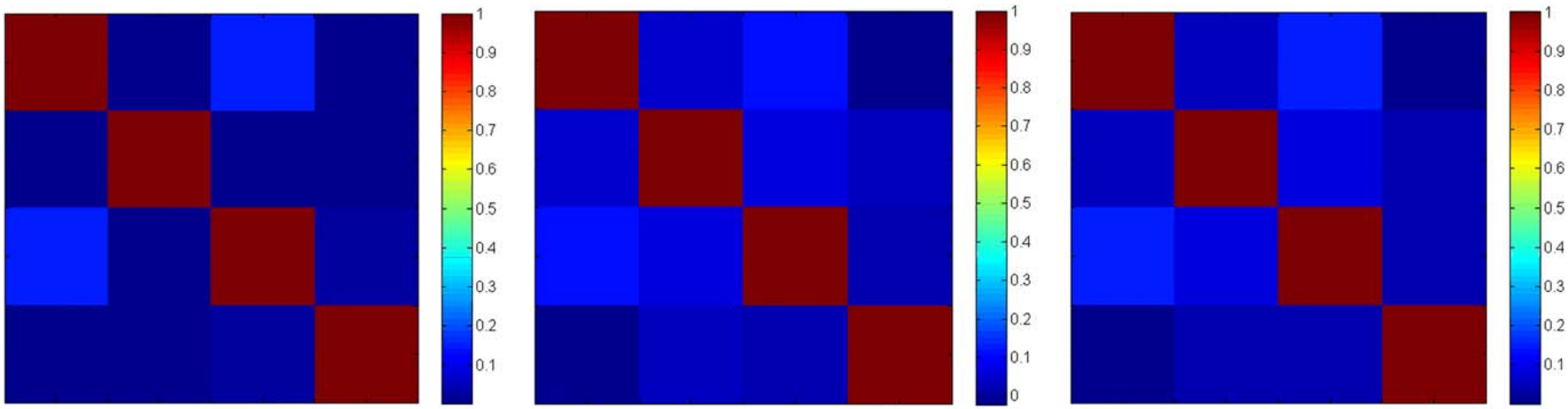
Because materials in the experimental RGB image are orthogonal (they do not overlap in spatial domain) we can evaluate performance of the employed blind image decomposition methods via the correlation matrix defined as $\mathbf{G=SS}^{\mathrm{T}}$. For perfect estimation the correlation matrix will be diagonal and performance is visualized as deviation from diagonal matrix. To quantify decomposition quality numerically we compute the correlation index in dB scale as

$$CR = -10\log_{10}\sum_{\substack{i,j=1\\ j\neq i}}^{M} g_{ij}^{2}$$

where before calculating correlation matrix **G** rows of **S** are normalized to unit L_2-norm.

Unsupervised segmentation of multispectral images

Correlation matrices



From left to right: 25 layers HALS NMF; Interior point method, [74,90]; c) Linear programming.

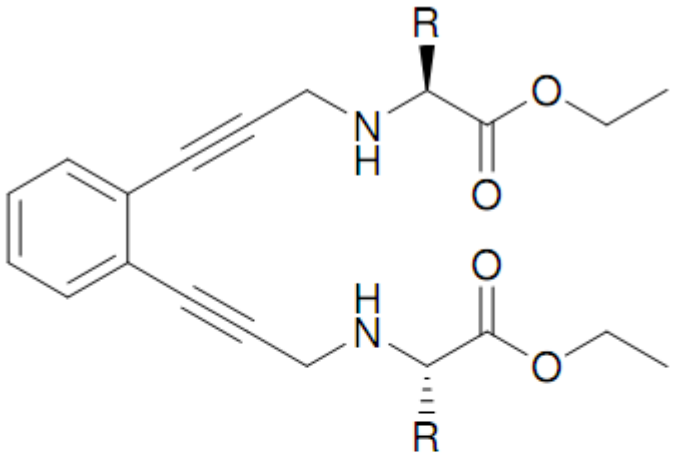
CR performance measure in dB

	Multilayer HALS NMF	**Interior-point method**	**Linear program**
CR [dB]	13.67	9.97	7.77
CPU time [s]*	3097	7751	3265

*MATLAB environment on 2.4 GHz Intel Core 2 Quad Processor Q6600 desktop computer with 4GB RAM.

Blind extraction of analytes (pure components) from mixtures of chemical compounds in mass spectrometry[15]

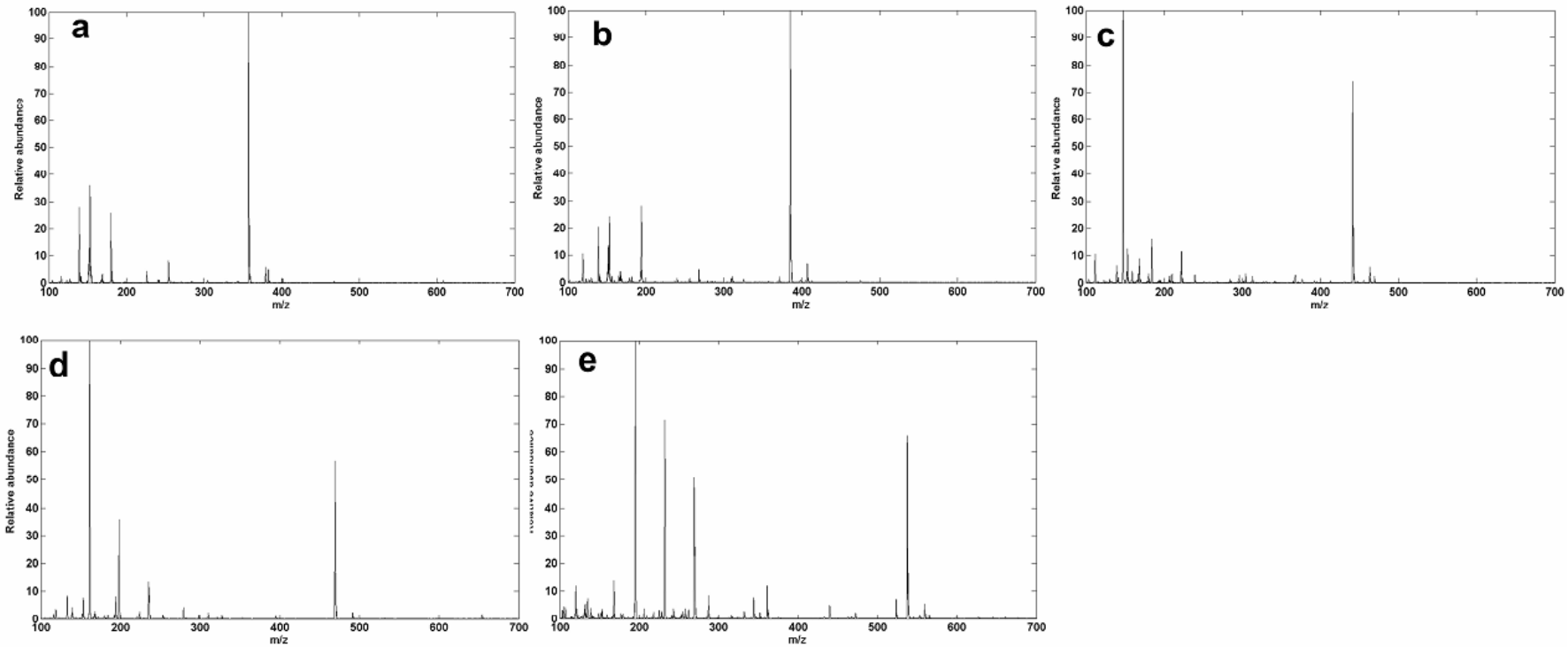
15. I. Kopriva, I. Jerić (2009). Multi-component Analysis: Blind Extraction of Pure Components Mass Spectra using Sparse Component Analysis, **Journal of Mass Spectrometry**, vol. 44, issue 9, pp. 1378-1388 (IF: 2.94).



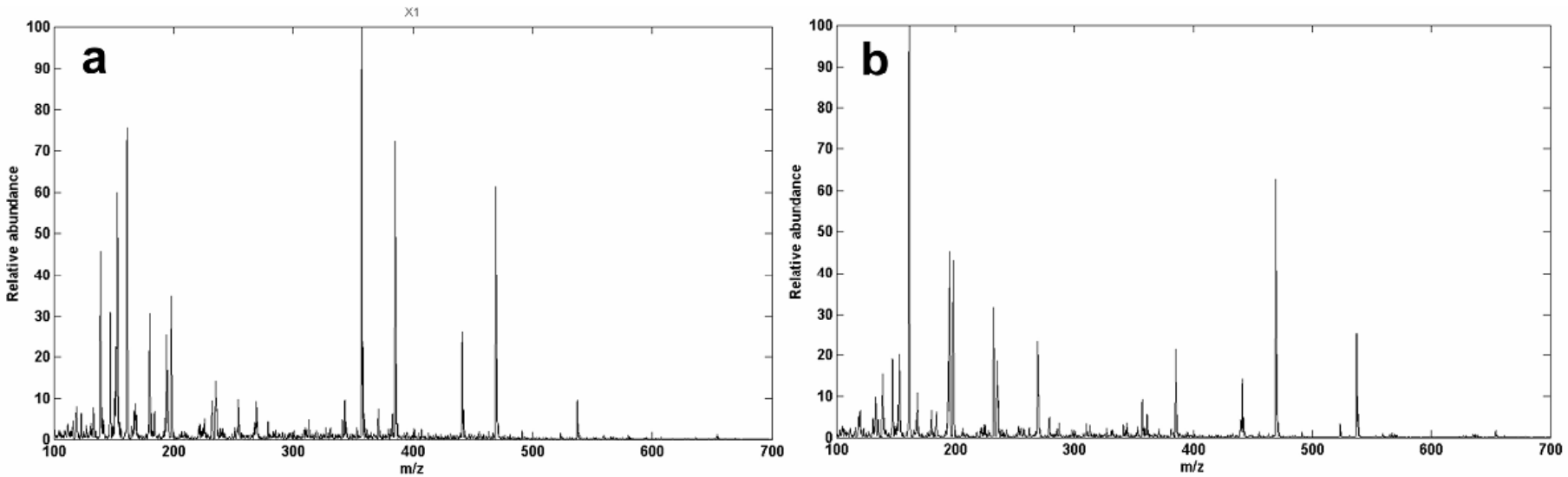
5 R=H
6 R=CH_3
7 R=$CH(CH_3)_2$
8 R=$CH_2CH(CH_3)_3$
9 R=$CH_2C_6H_5$

Figure S-1.

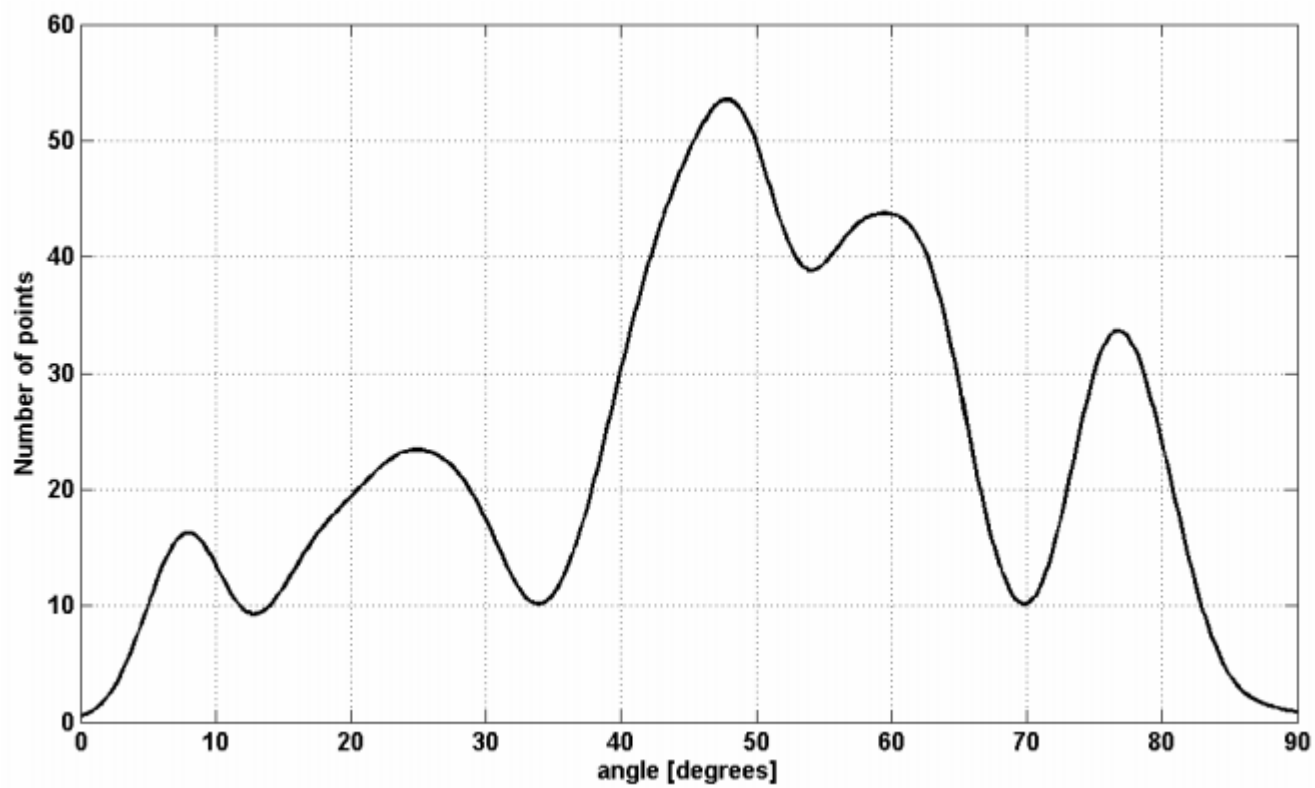
Chemical structure of five pure components.



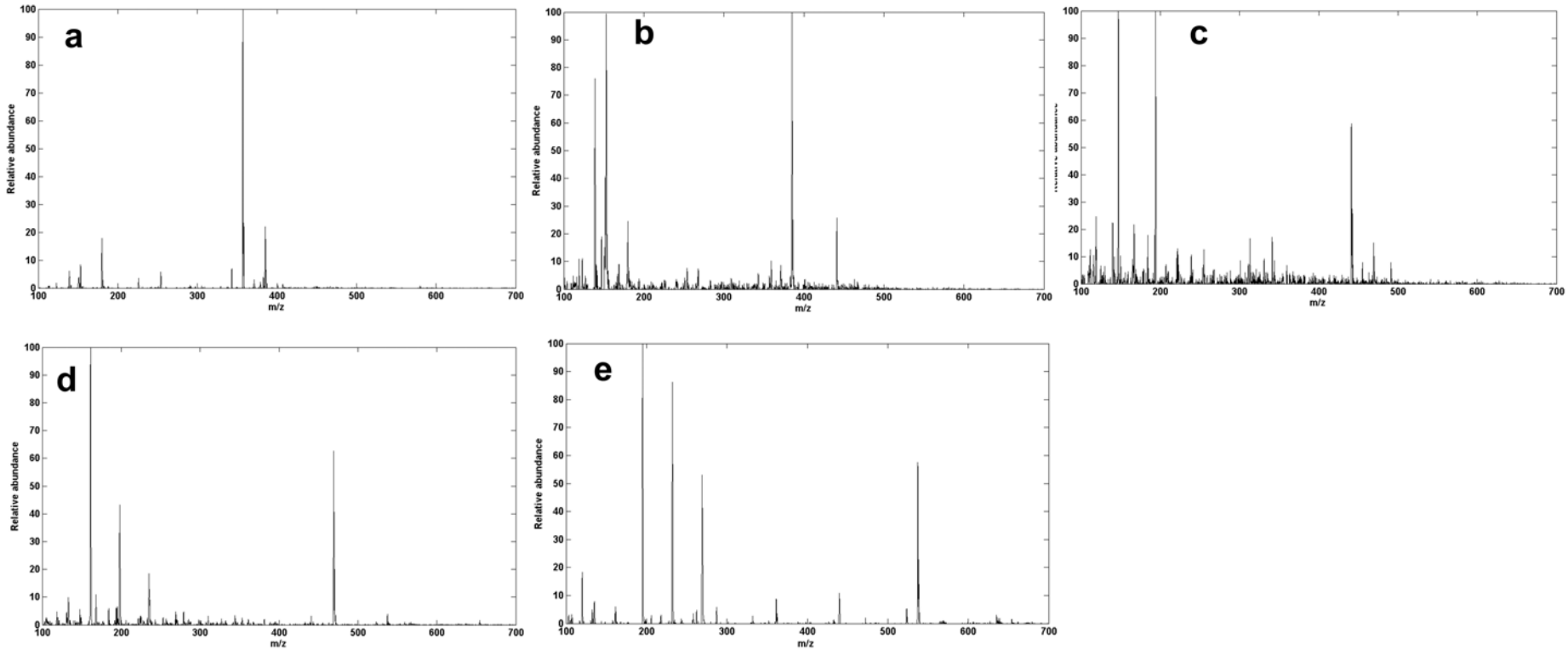
Mass spectra of five pure components.



Mass spectra of two mixtures



Dana clustering function in the mixing anagle domain. Five peaks indicate presence of five components in the mixtures spectra.



Mass spectra of five pure components estimated by multilayer (100 layers) HALS NMF algorithm with 500 iterations per layer and regularization constant α_S=0.5.

Normalized correlation coefficients with true pure components 1to 5: 0.9084, 0.7432, 0.7389, 0.9372, 0.9698.