Nonlinear sparse component analysis with applications in medical image analysis, bioinformatics and chemometrics

Ivica Kopriva

Ruđer Bošković Institute

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

Acknowledgment:

Croatian Science Foundation Grant 9.01/232 “Nonlinear component analysis with applications in chemometrics and pathology”**.**

Rudjer Boskovich Institute, Zagreb, Croatia

https://www.irb.hr/eng

The Ruđer Bošković Institute is regarded as Croatia’s leading scientific institute in the natural and biomedical sciences as well as marine and environmental research, owing to its size, scientific productivity, international reputation in research, and the quality of its scientific personnel and research facilities.

The Institute is the leading and internationally most competitive Croatian institute by virtue of its participation in international research projects, such as the IAEA and EC FP5-7 programs funded by the European Commission, NATO, NSF, SNSF, DAAD and other international scientific foundations.

Today, the Ruđer Bošković Institute has over 550 scientists and researchers in more than 80 laboratories pursuing research in theoretical and experimental physics, physics and materials chemistry, electronics, physical chemistry, organic chemistry and biochemistry, molecular biology and medicine, the sea and the environment, informational and computer sciences, laser and nuclear research and development.

Roger Joseph Boskovich

http://en.wikipedia.org/wiki/Roger_Joseph_Boscovich

Ruđer Bošković (18 May 1711 – 13 February 1787) was a physicist, astronomer, mathematician, philosopher, diplomat, poet, theologian, Jesuit priest, and a polymath from the city of Dubrovnik in the Republic of Ragusa (today Croatia), who studied and lived in Italy and France where he also published many of his works.

Among his many achievements he was the first to suggest least absolute deviation based regression (1757). That was studied by Laplace (1793) and predated the least square technique originally developed by Legendre (1805) and Gauss (1823):

P. Bloomfield and W. L. Steiger. *Least Absolute Deviations: Theory, Applications, and Algorithms.* Birkhauser, Boston, MA, 1983.

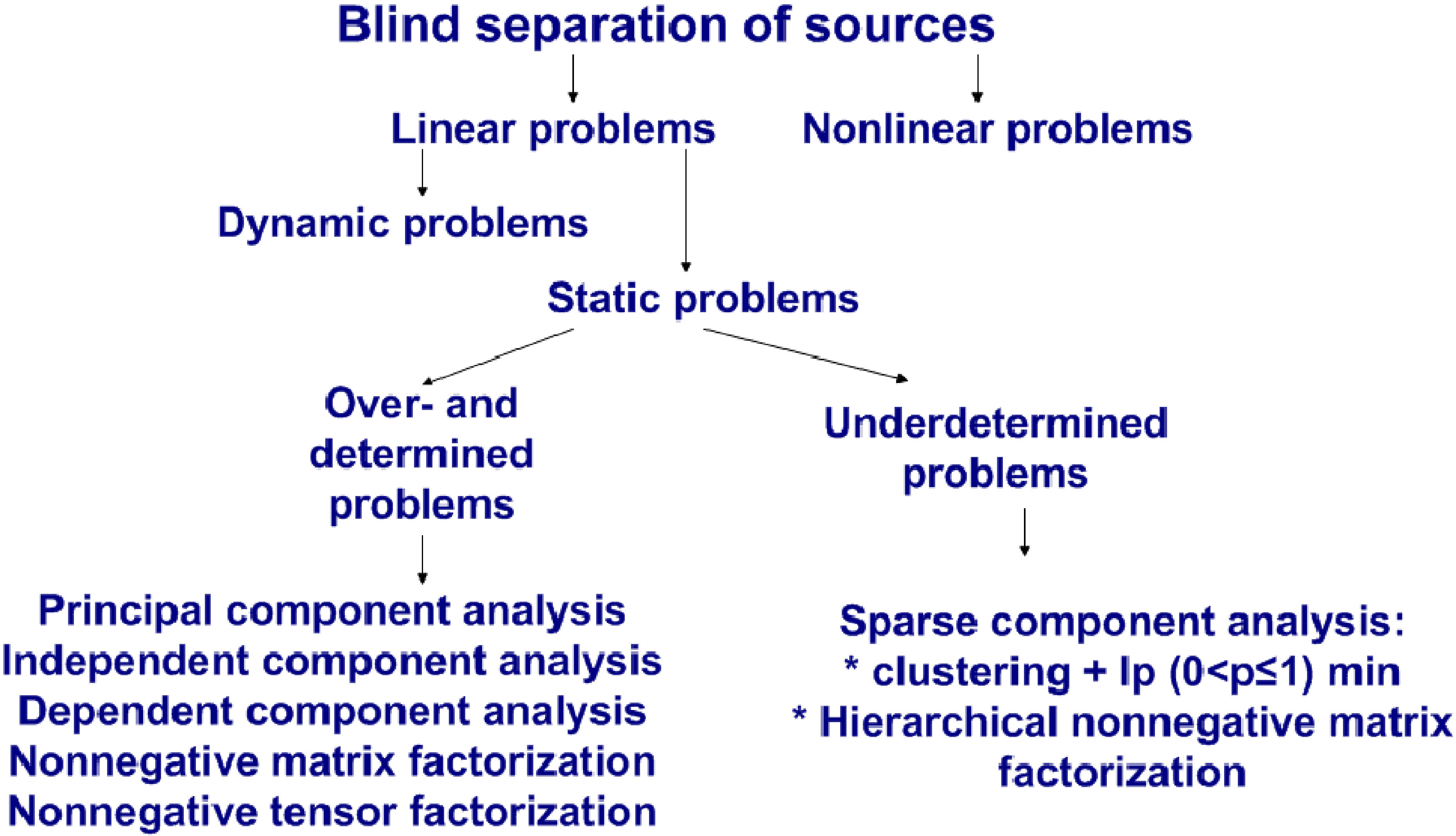
Talk outline

- Instantaneous blind source separation (BSS): problem definition and overview of main methods.

- Nonlinear underdetermined BSS (uBSS): motivation, conversion to linear uBSS.

- uBSS and sparse component analysis (SCA):
    - asymptotic results from compressed sensing theory,
    - SCA by sparseness constrained non-negative matrix factorization (NMF),
    - SCA/NMF in reproducible kernel Hilbert spaces (RKHS).

- Applications: (*i*) unsupervised decomposition of multispectral, CT and PET images; (*ii*) pure components extraction from mass spectra of nonlinear chemical reactions; (*iii*) variable selection in genomics and proteomics.



Blind separation of sources

Linear problems

Nonlinear problems

Dynamic problems

Static problems

Over- and determined problems

Underdetermined problems

Principal component analysis
Independent component analysis
Dependent component analysis
Nonnegative matrix factorization
Nonnegative tensor factorization

Sparse component analysis:
* clustering + lp $(0<p\leq1)$ min
* Hierarchical nonnegative matrix factorization

Blind Source Separation – linear static problem

Recovery of signals from their multichannel linear superposition using minimum of *a priori* information i.e. multichannel measurements only [1-3].

Problem:

$\mathbf{X}=\mathbf{AS}$ $\mathbf{X}\in R^{N\mathrm{x}T}$, $\mathbf{A}\in R^{\mathrm{NxM}}$, $\mathbf{S}\in R^{M\mathrm{x}T}$

N - number of sensors/mixtures;
M - unknown number of sources
T - number of samples/observations

Goal: find **S**, **A** and number of sources *M* based on **X** only.

1. A. Hyvarinen, J. Karhunen, E. Oja, “Independent Component Analysis,” John Wiley, 2001.
2. A. Cichocki, S. Amari, “Adaptive Blind Signal and Image Processing,” John Wiley, 2002.
3. P. Comon, C. Jutten, editors, “Handbook of Blind Source Separation,” Elsevier, 2010.

Blind Source Separation – linear static problem

$\mathbf{X}=\mathbf{AS}$ and $\mathbf{X}=\mathbf{ATT}^{-1}\mathbf{S}$ are equivalent for any square invertible matrix **T**. There are infinitely many pairs ($\mathbf{AT}$, $\mathbf{T}^{-1}\mathbf{S}$) satisfying linear mixture model $\mathbf{X}=\mathbf{AS}$. Solutions unique up to permutation and scaling indeterminacies, $\mathbf{T}=\mathbf{P\Lambda}$, are meaningful. For such solutions constraints must be imposed on **A** and/or **S**.

Independent component analysis (ICA) solves BSS problem provided that: source signals **S** are statistically independent and non-Gaussian; mixing matrix **A** is full column rank i.e. $M\leq N$.

Dependent component analysis (DCA) improves accuracy of ICA when sources are not statistically independent. Linear high-pass filtering type of preprocessing transform is applied row-wise to **X**: $L(\mathbf{X})=\mathbf{A}L(\mathbf{S})$. ICA is applied to $L(\mathbf{X})$ to estimate **A** and $L(\mathbf{S})$. **S** is estimated from $\mathbf{S}\approx\mathbf{A}^{-1}\mathbf{X}$.

Matlab implementation of many ICA algorithms can be found in the ICALAB:
http://www.bsp.brain.riken.go.jp/ICALAB/

Blind Source Separation – linear static problem

Sparse component analysis (SCA) solves BSS problem imposing sparseness constraints on source signals **S**. *M* can be less than, equal to or greater than *N*.

Thus, SCA can be used to solve underdetermined BSS problems where number of source signals is greater than number of mixtures.

Nonnegative matrix factorization (NMF) solves BSS problem imposing nonnegativity, sparseness, smoothness or constraints on source signals. NMF algorithms that enforce sparse decomposition of **X** can be seen as SCA algorithms [4]

Matlab implementation of many NMF algorithms can be found in the NMFLAB:
http://www.bsp.brain.riken.jp/ICALAB/nmflab.html

4. 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.

Underdetermined BSS: (nonlinear) static problem [3,2,5,6]

$\mathbf{x}_t = \mathbf{f}\ \ \mathbf{s}_t \quad t=1,...,T$; $\mathbf{x}_t \in \mathbb{R}^{N\times 1}_{0+}$ stands for nonnegative vector comprised of measurements acquired at *T* independent variables (pixel positions, *m/z* ratios, genes, etc.).

$\mathbf{s}_t \in \mathbb{R}^{M\times 1}_{0+}$ stands for unknown vector of *M* sources. ***M>N* → uBSS problem**

$\mathbf{f}:\mathbb{R}^{M}_{0+} \mapsto \mathbb{R}^{N}_{0+}$ is an unknown multivariate mapping such that:

$\mathbf{f}\ \ \mathbf{s}_t\ = \left[\ f_1\ \ \mathbf{s}_t\ \ \ldots\ f_N\ \ \mathbf{s}_t\ \right]^T$ and $f_n:\mathbb{R}^{M}_{0+} \rightarrow \mathbb{R}_{0+}\ {}_{n=1}^{N}$.

Linear problem: $\mathbf{f}\ \ \mathbf{s}_t\ = \mathbf{A}\mathbf{s}_t$.

5. I. Kopriva, I. Jerić, M. Filipović, L. Brkljačić (2014). Empirical Kernel Map Approach to Nonlinear Underdetermined Blind Separation of Sparse Nonnegative Dependent Sources: Pure Components Extraction from Nonlinear Mixtures Mass Spectra. *J. of Chemometrics* , vol. 28, pp. 704-715.
6. I. Kopriva, I. Jerić, L. Brkljačić, (2013). Nonlinear Mixture-wise Expansion Approach to Underdetermined Blind Separation of Nonnegative Dependent Sources. *J. of Chemometrics*, vol. 27, pp.189-197

Underdetermined Blind Source Separation: motivation

In biomarker identification studies number of mixture spectra of biological samples (urine, blood, tissue extract, saliva, etc.) is rather small, while number of components/analytes (some of them are candidates for biomarkers) can be large.

For example, 326 analytes were quantified in extracts of Arabidopsis thaliana leaf tissue [7], while the independent gas chromatography-mass spectrometry (GC-MS) study of Arabidopsis thaliana leaves detected 497 unique chemical components [8].

Analysis of human adult urinary metabolome by liquid chromatography-mass spectrometry (LC-MS) revealed presence of 1484 components, while 384 of them were characterized by matching their spectra with references stored in libraries [9].

7. Fiehn O, Kopka J, Dörmann P, Altmann T, Trethewey R N, Willmitzer L. Metabolite profiling for plant functional genomics. *Nature Biotechnology* 2000; 18: 1157-1161.
8. Jonsson P, Johansson A I, Gullberg J, Trygg J, Jiye A, Grung B, Marklund S, Sjöström M, Antti H, Moritz T. High-throughput data analysis for detecting and identifying differences between samples in GC/MS-based metabolomic analyses," *Analytical Chem*. 2005; 77: 5635-5642.
9. Roux A, Xu Y, Heilier J-F, Olivier M-F, Ezan E, Tabet J-C, Junot C. Annotation of the human adult urinary metabolome and metabolite identification using ultra high performance liquid chromatography coupled to a linear quadrupole ion trap-orbitrap mass spectrometer. *Anal. Chem*. 2012; 84: 6429−6437.

Nonlinear u-Blind Source Separation: motivation

While linear mixture model is adequate for many scenarios, nonlinear model offers more accurate description of processes and interactions occurring in biological systems.

Living organisms are best examples of complex nonlinear systems that function far from equilibrium. Internal and external stimuli (disease, drug treatment, environmental changes) cause perturbations in the system as a result of highly synchronized molecular interactions, [10].

Furthermore, interactions within genes in components that are parts of gene regulating networks are nonlinear, [11].

10. Walleczek J (ed). Self-organized biological dynamics and non-linear control. Cambridge University Press: Cambridge, UK. 2000.
11. Yuh, C. H., Bolouri, H., Davidson, E. H.: Genomic cis-regulatory logic: experimental and computational analysis of a sea urchin gene. *Science* 279, 1896-1902 (1998).

Linear Underdetermined BSS

- SCA-based solution of the linear uBSS problem is obtained in two stages:

  1) estimate basis or mixing matrix **A** using data clustering.

  2) estimating sources, with estimated **A**, one at a time $\mathbf{s}_t$, t=1,…,T or simultaneously solving underdetermined linear systems of equations $\mathbf{x}_t$ =$\mathbf{A}\mathbf{s}_t$. Provided that $\mathbf{s}_t$ is sparse enough, solution is obtained at the minimum of L_p-norm, $\left\|\mathbf{s}_t\right\|_p$, 0≤ p ≤1.

Here: $\left\|\mathbf{s}_t\right\|_p = \left(\sum_{m=1}^{M}\left|s_{mt}\right|^p\right)^{1/p}.$

- NMF-based solution yields **A** and **S** simulatneously through sparseness and nonnegativity constrained factorization of **X**.

When uBSS problems can(not) be solved?

Let us focus on underdetermined linear system:

$$\mathbf{x}=\mathbf{A}\mathbf{s},\ \mathbf{x}\in \mathrm{R}^{N},\ \mathbf{s}\in \mathrm{R}^{M}\ ,\ M>N$$

Let **s** be *K*-sparse i.e. $K=||\mathbf{s}||_{\mathbf{0}}$.

Provided that **A** is *random*, with entries from Gaussian or Bernoulli distributions, compressed sensing theory has established necessary and sufficient condition on *N, M* and *K* to obtain, with probability one, unique solution at the minimum of L_1-norm of **s,** [12]:

$$N\approx K\log(M/K)$$

12. Candès E, Tao T. Near optimal signal recovery from random projections: universal encoding strategy?. *IEEE Trans. Information Theory* 2006; **52**: 5406-5425.

When uBSS problems can(not) be solved?

However in BSS problems **A** is not random matrix but *deterministic* matrix with a structure. For example, in multispectral imaging it contains spectral profiles of the objects/materials present in the image, [13].In chemometrics **A** contains concentration profiles of pure components present in the mixtures, [14].

One result for deterministic **A** is given in [15]. For cyclic polynomial matrix **A** it applies **$N=O(K^2)$**. That is significantly worse than $N \approx K\log(M/K)$ for random **A**. K correponds with number of sources that are active/present at the specific coordinate t (time, pixel, m/z variable, frequency, etc). Thus, K is application dependent.

13. Kopriva I, Cichocki A. Blind decomposition of low-dimensional multi-spectral image by sparse component analysis. *J. Chemometrics* 2009; **23** (11): 590-597.
14. Kopriva I, Jerić I. Blind separation of analytes in nuclear magnetic resonance spectroscopy and mass spectrometry: sparseness-based robust multicomponent analysis. *Anal. Chem*. 2010; **82**: 1911-1920.
15. DeVore R A. Deterministic constructions of compressed sensing matrices. *Journal of Complexity* 2007; **23**: 918-925.

uBSS – L_p norm minimization: 0< $p \leq 1$

- Signal **s** is *K*-sparse if it has K non-zero components, i.e. $K=\|\mathbf{s}\|_0$. Thereby,

$$\|\mathbf{s}\|_0 = \sum_{m=1}^{M} |s_m|^0 \quad \text{By definition}: 0^0 = 0.$$

- If uBSS problem is not sparse in original domain <u>it ought to be transformed</u> in domain where enough level of sparseness can be achieved: $T(\mathbf{x})=\mathbf{A}T(\mathbf{s})$.

- Time-frequency and time-scale (wavelet) bases are employed for this purpose quite often.

- In addition to sparseness requirement on **s** certain degree of incoherence of the mixing matrix **A** is required as well. Mutual coherence is defined as the largest absolute and normalized inner product between different columns in **A**, what reads as

$$\mu \;\; \mathbf{A} \; = \max_{1\leq i,j\leq M \text{ and } i\neq j} \frac{\left|\mathbf{a}_i^T \mathbf{a}_j\right|}{\left\|\mathbf{a}_i\right\| \left\|\mathbf{a}_j\right\|}$$

uBSS – L_p norm minimization: $0< p \leq 1$

The mutual coherence provides a <u>worst case </u>measure of similarity between the basis vectors. It indicates how much two closely related vectors may confuse any pursuit algorithm (solver of the underdetermined linear system of equations). The worst-case perfect recovery condition for **s** relates sparseness requirement on **s** and coherence of **A**, [16,17]:

$$\|\mathbf{s}\|_0 < \frac{1}{2}\left(1+\frac{1}{\mu\;\mathbf{A}}\right)$$

In [18] another uniqueness theorem has been stated. If **A** has unique representation property, that is if all *N*×*N* sub-matrices are full rank, the unique solution of **x**=**As** exists if: $\|\mathbf{s}\|_0 \leq N/2$.

16. R. Gribonval and M. Nielsen, "Sparse representations in unions of bases," *IEEE Transactions on Information Theory* **49**, 3320-3325 (2003).
17. J. A. Tropp, "Greed is good: Algorithmic results for sparse approximation," *IEEE Transactions on Information Theory* **50**, 2231-2242 (2004).
18. I. F. Gorodnitsky and B. D. Rao, “Sparse signal reconstruction from limited data using FOCUSS, a re-weighted minimum norm algorithm,” *IEEE Trans. Signal Process*., vol.45, no.3, pp. 600–616, Mar. 1997.

uBSS – L_p norm minimization: $0< p \leq 1$

In BSS scenario properties of the mixing matrix **A** can not be predefined, i.e. they are problem dependent. Yet, **A** dictates a level of sparseness of **s** that is necessary to obtain possibly unique solution of the uBSS problem: **x**=**As**. To obtain such solution it is necessary to:

- estimate **A** as accurately as possible.

- find representation (transformation) $T(\mathbf{x})=\mathbf{A}T(\mathbf{s})$ where $T(\mathbf{s})$ is as sparse as possible.

- construct algorithms for solving underdetermined system of equations $T(\mathbf{x})=\mathbf{A}T(\mathbf{s})$ that are robust with respect to the presence of noise i.e. errors in sparse approximation of $T(\mathbf{s})$: $T(\mathbf{s})$ is approximately *K*-sparse with *K* dominant and number of small coefficients. If possible performance of the algorithm should remain robust if *K* increases.

uBSS – L_p norm minimization: 0< p ≤1

Solving underdetermined system of linear equations **x=As** amounts to solve:

$$\hat{\mathbf{s}}(t) = \underset{\mathbf{s}(t)}{\arg\min} \left\|\mathbf{s}(t)\right\|_0 \quad \text{s.t. } \hat{\mathbf{A}}\mathbf{s}(t) = \mathbf{x}(t) \quad \forall t = 1,\ldots,T$$

or for problems with noise or approximation error:

$$\hat{\mathbf{s}}(t) = \underset{\mathbf{s}(t)}{\arg\min} \frac{1}{2}\left\|\hat{\mathbf{A}}\mathbf{s}(t) - \mathbf{x}(t)\right\|_2^2 + \lambda\left\|\mathbf{s}(t)\right\|_0 \quad \forall t = 1,\ldots,T$$

$$\hat{\mathbf{s}}(t) = \underset{\mathbf{s}(t)}{\arg\min} \left\|\mathbf{s}(t)\right\|_0 \quad \text{s.t. } \left\|\hat{\mathbf{A}}\mathbf{s}(t) - \mathbf{x}(t)\right\|_2^2 \le \varepsilon \quad \forall t = 1,\ldots,T$$

Direct minimization of L_0–norm of **s** is combinatorial problem that is NP-hard. For larger dimension *M* it becomes computationally infeasible.

uBSS – *L*$_1$ norm minimization

Replacement of *L*$_0$-norm by *L*$_1$-norm is done quite often. That is known as *convex relaxation* of the minimum *L*$_0$-norm problem. It leads to linear program:

$$\hat{\mathbf{s}}(t) = \underset{\mathbf{s}(t)}{\arg\min} \sum_{m=1}^{\hat{M}} s_m \;\; t \quad \text{s.t. } \hat{\mathbf{A}}\mathbf{s}(t) = \mathbf{x}(t) \quad \forall t = 1,..., \quad \text{s.t. } \mathbf{s}(t) \geq 0$$

L$_1$-regularized least square problem ref.[19,20]:

$$\hat{\mathbf{s}}(t) = \underset{\mathbf{s}(t)}{\arg\min} \frac{1}{2} \left\| \hat{\mathbf{A}}\mathbf{s}(t) - \mathbf{x}(t) \right\|_2^2 + \lambda \left\| \mathbf{s}(t) \right\|_1 \quad \forall t = 1,...,T$$

and *L*$_2$-regularized linear problem [20,21]:

$$\hat{\mathbf{s}}(t) = \underset{\mathbf{s}(t)}{\arg\min} \; \left\| \mathbf{s}(t) \right\|_1 \quad \text{s.t. } \left\| \hat{\mathbf{A}}\mathbf{s}(t) - \mathbf{x}(t) \right\|_2^2 \leq \varepsilon \quad \forall t = 1,...,T$$

19. S..J. Kim, K. Koh, M. Lustig, S. Boyd, D. Gorinevsky, "An Interior-Point Method for Large-Scale -Regularized Least Squares," *IEEE Journal of Selected Topics in Signal Processing* **1**, 606-617 (2007), **http://www.stanford.edu/~boyd/l1_ls/.**
20. E. van den Berg, M.P. Friedlander, "Probing the Pareto Frontier for Basis Pursuit Solutions," *SIAM J. Sci. Comput*. 31, 890-912 (2008).
21. M.A.T. Figuiredo, R.D. Nowak, S.J. Wright, "Gradient Projection for Sparse Reconstruction: Application to Compressed Sensing and Other Inverse Problems," *IEEE Journal on Selected Topics in Signal Processing* **1**, 586-597 (2007).

Linear uBSS: summary

Linear uBSS problem is characterized with a triplet (*N*, *M*, *K*). Under L_1-norm constraints unique solution is possible if *N*≈*K*log(*M*/*K*) .

In biological experiments *M* corresponds with number of analytes (metabolites) present in mixture spectra and, thus, can be large. *K* represents maximal number of overlapping components. Depending on the resolution of the spectrometer it can be large as well. *N* stands for number of biological samples and is usually (very) small. Thus, requirement *N*≈*K*log(*M*/*K*) can often failed to be fulfilled!!!

In [6] solution was proposed to transform original uBSS problem **x**=**As** into new one $\Psi \;\; \mathbf{x} \;\; = \bar{\mathbf{A}}\bar{\mathbf{s}}, \; \bar{\mathbf{A}} \in \mathbb{R}_{0+}^{D\times P}, \; \bar{\mathbf{s}} \in \mathbb{R}_{0+}^{P\times 1}$, with maximal number of overlapping components equal to *Q*. Thus, uniqueness condition becomes: *D*≈*Q*log(*P*/*Q*). That is fulfilled if: (*D*/*N*) >>(*P*/*M*) as well as (*D*/*N*) >>(*Q*/*K*).

6. I. Kopriva, I. Jerić, L. Brkljačić, (2013). Nonlinear Mixture-wise Expansion Approach to Underdetermined Blind Separation of Nonnegative Dependent Sources. *J. of Chemometrics*, vol. 27, pp.189-197 .

Nonlinear uBSS: Taylor expansion up to aribtrary order *J*, [5]

$$\mathbf{x}_t = \mathbf{f}\ \ \mathbf{s}_t \qquad t=1,\ldots,T$$

$$\mathbf{x}_t \in \mathbb{R}_{0+}^{N\times 1}, \mathbf{s}_t \in \mathbb{R}_{0+}^{M\times 1}\ ,\ \ M > N.$$

$$\mathbf{f} : \mathbb{R}_{0+}^{M} \mapsto \mathbb{R}_{0+}^{N}$$

$$\mathbf{f}\ \ \mathbf{s}_t\ \ = \Big[\, f_1\ \ \mathbf{s}_t\ \ \ldots\ f_N\ \ \mathbf{s}_t\ \ \Big]^T \qquad f_n : \mathbb{R}_{0+}^{M} \rightarrow \mathbb{R}_{0+}\ \ {}_{n=1}^{N}$$

Nonlinear uBSS problem can be expanded into Taylor series around reference point $\mathbf{s}_0$. Without loss of generality let us assume $\mathbf{s}_0=\mathbf{0}_{M\times 1}$. Let us also assume $\mathbf{f}(\mathbf{s}_0)=\mathbf{0}_{N\times 1}$.

5. I. Kopriva, I. Jerić, M. Filipović, L. Brkljačić (2014). Empirical Kernel Map Approach to Nonlinear Underdetermined Blind Separation of Sparse Nonnegative Dependent Sources: Pure Components Extraction from Nonlinear Mixtures Mass Spectra. *J. of Chemometrics* , vol. 28, pp. 704-715.

Nonlinear uBSS: Taylor expansion up to aribtrary order J

In the literature one mostly finds Taylor expansion based on first- (Jacobian) and second (Hessian) order derivatives of vector valued function with vector argument and first derivative of matrix function with matrix argument.

It is argued in [22] that very occasionally one might need third- and higher-order derivatives of vector- or matrix-valued functions with vector or matrix arguments. The main reason for that is notational complexity. It is argued in [23] and [24], in chapters 9 and 10, to use procedure based on differentials when calculating first and second derivatives of discussed functions. For higher order terms no recommendation is given.

22. Magnus J R. On the concept of matrix derivative. *J. Multivariate Analysis* 2010; 101, 2200-2206.
23. Magnus J R, Neudecker H. Matrix Differential Calculus with Applications to Simple, Hadamard, and Kronecker Products. *J. Mathematical Psychology* 1985; 29: 474-492.
24. Magnus J R, Neudecker H. *Matrix Differential Calculus with Applications in Statistics and Economics*. Revised edition. John Wiley: Chichester, UK, 1999.

Nonlinear uBSS: Taylor expansion up to aribtrary order *J*

In [5] an approach is presented to derivation of the Taylor expansion of vector valued function with vector argument up to arbitrary order *J* by using tensorial notation, [25].

That is legitimate given the fact that j^{th} term, j=1,…,*J*, in Taylor expnasion of vector valued function with vector argument is a tensor of order *j*+1.

To this end, higher-order arrays (tensors) will be denoted with underlined uppercase bold letters. For example $\underline{\mathbf{X}} \in \mathbb{R}_{0+}^{I_1 \times I_2 \times I_3}$ refers to a third order nonnegative tensor with dimensions I_1, I_2 and I_3. Uppercase bold letters, **X**, denote matrices, lowercase bold letters, **x**, denote vectors and italic lowercase letters, *x*, denote scalars.

[25] Kiers H A L. Towards a standardized notation and terminology in multiway analysis. J*. Chemometrics* 2000; 14: 105-122.

Nonlinear uBSS: Taylor expansion up to aribtrary order *J*

We can write *j*th order derivative as a tensor of the order *j*+1:

$$\underline{\mathbf{G}}^{j} \in \mathbb{R}_{0+}^{N\times \overbrace{M\times \ldots \times M}^{j\ times}}$$

Element of the derivative tensor indexed by (n, $m_{1,}$ ... , m_k), where n=1, ... , N, m_1=1,..., M, ... , m_k=1,..., M is given as:

$$\left[\underline{\mathbf{G}}^{j}\right]_{nm_1\ldots m_j} = \frac{\partial^{j} f_n \ \ \mathbf{s}}{\partial s_{m_1}\ldots\partial s_{m_j}}$$

We now introduce mode-*r* product of an *R*th order tensor $\underline{\mathbf{T}} \in \mathbb{R}^{I_1\times I_2\times\ldots\times I_R}$ and a matrix $\mathbf{W} \in \mathbb{R}^{J_1\times J_2}$ that is defined when number of columns of matrix is equal to the dimension of the tensor in mode *r*, that is J_2=I_r. That yields a new tensor $\underline{\mathbf{Y}} = \underline{\mathbf{T}}\times_r \mathbf{W}$ such that $\underline{\mathbf{Y}} \in \mathbb{R}^{I_1\times\ldots I_{r-1}\times J_2\times I_{r+1}\times\ldots\times I_R}$, [26].

26. Kolda T G, Bader B W. Tensor Decompositions and Applications. *SIAM Review* 2009; 51: 455-500.

Nonlinear uBSS: Taylor expansion up to aribtrary order *J*

For example, mode-2 product of a 3-way tensor $\underline{\mathbf{T}} \in \mathbb{R}^{I_1\times I_2\times I_3}$ and a matrix $\mathbf{W} \in \mathbb{R}^{D\times I_2}$ is a 3-way tensor $\underline{\mathbf{Y}} = \underline{\mathbf{T}} \times_2 \mathbf{W} \in \mathbb{R}^{I_1\times D\times I_3}$, calculated element-wise as: $y_{i_1,d,i_3} = \sum_{i_2=1}^{I_2} t_{i_1,i_2,i_3} \cdot w_{d,i_2}$.

We can now express contribution of the *j*th order term in Taylor expansion as:

$$\mathbf{x}^j = \frac{1}{j!} \underline{\mathbf{G}}^j \times_2 \mathbf{s}^T \ldots \times_{j+1} \mathbf{s}^T$$

where *T* denotes transpose operation and expression above is known as Tucker tensor model [26, 27]. Thereby, $1/j!\,\underline{\mathbf{G}}^j$ stands for a core tensor and $\mathbf{s}^T$ stand for factors.

Since s^T is 1×*M* row vector mode-2 to mode-*(j*+1) multiplications of derivative tensor $\underline{\mathbf{G}}^j$ with row vectors s^T contracts higher order dimensions yielding as final result an *N*×1 column vector.

27. Tucker L R. Some mathematical notes on three-mode factor analysis. *Psychometrika* 1966; 31: 279-311.

Nonlinear uBSS: Taylor expansion up to aribtrary order *J*

We can also use mode-1 unfolding of $\underline{\mathbf{G}}^{j}\times_{2}\mathbf{s}^{T}\ldots\times_{j+1}\mathbf{s}^{T}$ that yields:

$$\mathbf{x}^{j}=\frac{1}{j!}\mathbf{G}_{(1)}^{j}\left[\underbrace{\mathbf{s}\otimes\ldots\otimes\mathbf{s}}_{j-1\ \ times}\right]$$

where $\mathbf{G}_{(1)}^{j}\in\mathbb{R}_{0+}^{N\times M^{(j)}}$ denotes a matrix obtained by mode-1 unfolding of tensor $\underline{\mathbf{G}}^{j}$, $\otimes$ denotes Kronecker product, and $\underbrace{\mathbf{s}\otimes\ldots\otimes\mathbf{s}}_{j-1\ \ times}$ yields $M^{(j)}\times1$ vector, such that $M^{(j)}=\binom{M+j-1}{j}$.

Hence, we can formally write a *J*th order Taylor expansion of vector valued function with vector argument as:

$$\mathbf{x}=\sum_{j=1}^{J}\frac{1}{j!}\underline{\mathbf{G}}^{j}\times_{2}\mathbf{s}^{T}\ldots\times_{j+1}\mathbf{s}^{T}=\sum_{j=1}^{J}\frac{1}{j!}\mathbf{G}_{(1)}^{j}\left[\underbrace{\mathbf{s}\otimes\ldots\otimes\mathbf{s}}_{j-1\ \ times}\right]$$

Nonlinear uBSS: Taylor expansion up to aribtrary order *J*

Elements of $\underset{j-1\ \ times}{\mathbf{s}\otimes\ldots\otimes\mathbf{s}}$ are monomials of order *j*:

$$\underset{j-1\ \ times}{\mathbf{s}\otimes\ldots\otimes\mathbf{s}} = \left[\ s_{m_1}^{q_1}\ldots s_{m_p}^{q_p}\qquad {}_{m_1,\ldots,m_p=1;\,p=1}^{M;\ \ j}\ \right]\ s.t.\ q_1,\ldots,q_p \in\ 0,1,\ldots,p \quad \text{and} \sum_{i=1}^{p} q_i = j$$

Thus, nonlinear mapping **f**(**s**) induces higher order (nonlinear) terms. For linear mapping, *j*=1, above expression becomes **s**. Formally, Taylor expansion of **f**(**s**) can be written as:

$$\mathbf{x} = \mathbf{G}\bar{\mathbf{s}}$$

where $\mathbf{G}\in\mathbb{R}_{0+}^{N\times\sum_{j=1}^{J}M(j)}$ is a block matrix $\left[\mathbf{G}_{(1)}^{1}\middle|\frac{1}{2}\mathbf{G}_{(1)}^{2}\middle|\ldots\middle|\frac{1}{J\,!}\mathbf{G}_{(1)}^{J}\right]$. $\bar{\mathbf{s}}\in\mathbb{R}_{0+}^{\Sigma_{j=1}^{J}M(j)\times 1}$ is a column vector:

$$\bar{\mathbf{s}} = \left[\mathbf{s}\ \ \mathbf{s}\otimes\mathbf{s}\ \ldots\ \underset{J-1\ \ times}{\mathbf{s}\otimes\ldots\otimes\mathbf{s}}\right]^T$$

Nonlinear uBSS: Taylor expansion up to aribtrary order *J*

Nonlinear uBSS problem **x=f(s)** characterized with a triplet (*N*, *M*, *K*), *M>N,* is represented with linear uBSS problem $\mathbf{x}=\mathbf{G}\bar{\mathbf{s}}$ characterized with a triplet (*N*, $\bar{M}$, *Q*), whereat $\bar{M}=\sum_{j=1}^{J}\binom{M+j-1}{j}$ and *Q>>K*.

If linear uBSS problem **x=As** characterized with (*N*, *M*, *K*), *M>N,* is impossible to solve without additional constraints imposed on **s**, solution of linear uBSS problem $\mathbf{x}=\mathbf{G}\bar{\mathbf{s}}$ characterized with (*N*,$\bar{M}$, *Q*), demands even harder constraints to be imposed on **s**.

However, Taylor expansion of **x=f(s)** up to arbitrary order *J* substitutes study of nonlinear uBSS problem **x=f(s)** with its linear equivalent $\mathbf{x}=\mathbf{G}\bar{\mathbf{s}}$.

Linear uBSS

Linear uBSS problem **x**=**As** is characterized with a triplet (*N*, *M*, *K*). Under L_1-norm constraints on **s** unique solution is possible if *N*≈*K*log(*M*/*K*) .

In biological experiments *M* corresponds with number of analytes (metabolites) present in mixture spectra and, thus, can be large. *K* represents maximal number of overlapping components. Depending on the resolution of the spectrometer it can be large as well. *N* stands for number of biological samples and is small. Thus, requirement *N*≈*K*log(*M*/*K*) can often failed to be fulfilled!!!

In [6] solution was proposed to transform original uBSS problem **x**=**As** into new one $\Psi\ \mathbf{x}\ = \bar{\mathbf{A}}\bar{\mathbf{s}},\ \bar{\mathbf{A}} \in \mathbb{R}_{0+}^{D\times P},\ \bar{\mathbf{s}} \in \mathbb{R}_{0+}^{P\times 1}$, with maximal number of overlapping components equal to *Q*. Thus, uniqueness condition becomes: *D*≈*Q*log(*P*/*Q*). That is fulfilled if: (*D*/*N*) >>(*P*/*M*) as well as (*D*/*N*) >>(*Q*/*K*).

6. I. Kopriva, I. Jerić, L. Brkljačić, (2013). Nonlinear Mixture-wise Expansion Approach to Underdetermined Blind Separation of Nonnegative Dependent Sources. *J. of Chemometrics*, vol. 27, pp.189-197 .

Nonlinear mapping of linear uBSS problem?

In [6] a new concept was proposed by mapping original uBSS problem **X=AS** nonlinearly into new one:

$$\mathbf{x}(t) \rightarrow \phi\ \ \mathbf{x}(t)\ \ {}_{t=1}^{T} \quad \text{s.t.} \quad \mathbf{x}(t) \in \mathbb{R}_{0+}^{N},\ \phi\ \ \mathbf{x}(t)\ \ \in \mathbb{R}_{0+}^{\bar{N}}\ \ and\ \ \bar{N} \gg N$$

since mapping $\phi(\mathbf{x}(t))$ is nonlinear new measurements are linearly independent.

The nonlinear mapping has the following algebraic structure:

$$\phi\ \ \mathbf{x}(t)\ \ = \Big[\ \ c_{q_1 \ldots q_N} x_1^{q_1}(t) \ldots x_N^{q_N}(t)\ \ {}_{q_1,\ldots,q_N=0}^{\bar{N}}\ \Big]^T \ \text{such that} \ \sum_{n=1}^{N} q_n \leq \bar{N}, \quad \forall t = 1, \ldots, T.$$

6. Kopriva I, Jerić I, Brkljačić, L. Nonlinear mixtures-wise expansion approach to underdetermined blind separation of nonnegative depedent sources. *J. Chemometrics* 2013; **27**: 189-197.

Nonlinear mapping of linear uBSS problem?

The mapped problem becomes:

$$\phi\left(\mathbf{x}(t)\right)=c_0\mathbf{e}_1+\mathbf{B}\begin{bmatrix}0\\ \mathbf{s}(t)\end{bmatrix}+\mathbf{B}_{HOT}\begin{bmatrix}0\\ \mathbf{0}_{M\times1}\\ \mathbf{s}(t)_{HOT}\end{bmatrix}\quad\forall t=1,\dots,T$$

where **s**(*t*)$_{HOT}$ is $\bar{N}-M-1$ column vector comprised of: $\left\{s_1^{q_1}(t)\times..\times s_M^{q_M}(t)\right\}_{q_1,\dots,q_M=2}^{\bar{N}}$

such that: $\sum_{m=1}^{M}q_m\leq\bar{N}$.

Sparse probabilistic model of sources

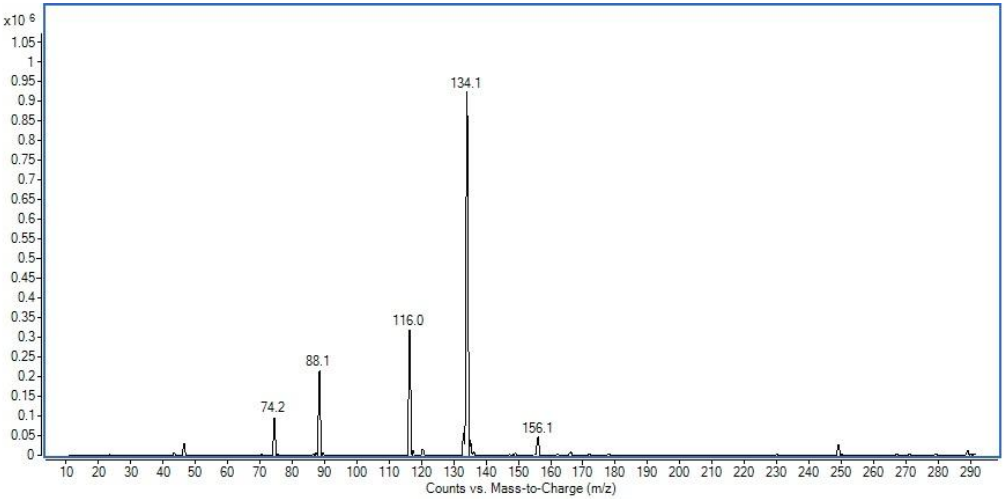
Let us assume sparse probabilistic model of the sources, i.e. each source signal is distributed according to p.d.f. based on mixed state random variable model [28, 29, 5]:

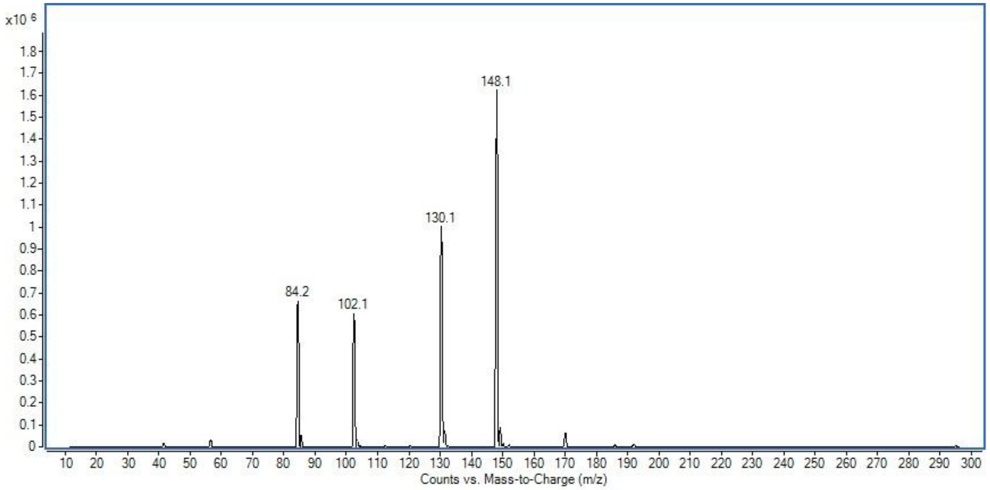
$$p(s_{mt}) = \rho\delta \quad s_{mt} \quad + 1-\rho \quad \delta^{*} \quad s_{mt} \quad f \quad s_{mt} \quad \forall m=1,\ldots,M \ \forall t=1,\ldots,T$$

where $\delta(s_{mt})$ is an indicator function and $\delta^{*}(s_{mt})$ =1- $\delta(s_{mt})$ is its complementary function. $\rho = \ P \ \mathbf{s}_{mt} = 0 \ {}_{t=1}^{T}$ Thus, $P \ s_{mt} > 0 \ = 1-\rho \ {}_{t=1}^{T}$.

28. Bouthemy P, Piriou C H G, Yao J. Mixed-state auto-models and motion texture modeling. *J. Math Imaging Vision* 2006; 25: 387-402.
29. Caifa C, Cichocki A. Estimation of Sparse Nonnegative Sources from Noisy Overcomplete Mixtures Using MAP. *Neural Comput.* 2009; 21: 3487-3518.
5. I. Kopriva, I. Jerić, M. Filipović, L. Brkljačić (2014). Empirical Kernel Map Approach to Nonlinear Underdetermined Blind Separation of Sparse Nonnegative Dependent Sources: Pure Components Extraction from Nonlinear Mixtures Mass Spectra. *J. of Chemometrics* , vol. 28, pp. 704-715.

Sparse probabilistic model of sources





Examples of mass spectra of sources (pure components, analytes,…). They are sparse in support and amplitude. We can use exponential distribution for:

$$f\left(s_{mt}\right) = \left(1/\mu_m\right)\exp\left(-s_{mt}/\mu_m\right)$$

In [5] using mass spectra of 25 pure components it has been estimated:

$$\hat{\rho}_m \in \left[0.27, 0.74\right]$$

$$\hat{\mu}_m \in \left[0.0012, 0.0014\right]$$

Nonlinear mapping of linear uBSS problem?

Thus, with high probability at least one source will not be present at location *t*. Thus, many cross-products will vanish. Also, by assuming $0 \leq s_{mt} \leq 1$ it follows that $s_m^{q_m}(t) \rightarrow 0$ when q_m grows.

Thus, by *hard* or *soft thresholding of* $\phi(\mathbf{x}(t))$ higher-order terms can be suppressed. Under sparse probabilistic prior validated on experimental mass spectra mostly second order terms will survive. That yields:

$$\phi \;\; \mathbf{X} \;\; _{\tau} \approx \left[\underbrace{c_0\mathbf{e}_1 \ldots c_0\mathbf{e}_1}_{\times T\, times} \right] + \bar{\mathbf{B}} \begin{bmatrix} 0 \\ \mathbf{S} \\ \mathbf{s}_{m_1}\mathbf{s}_{m_2} \quad {}_{m_1,m_2=1}^{M} \end{bmatrix}$$

where : $\phi \;\; \mathbf{X} \;\; _{\tau} \in \mathbb{R}_{0+}^{\bar{N}\times T}$ $\bar{\mathbf{B}} \in \mathbb{R}_{0+}^{\bar{N}\times P+1}$ and *P≈2M + M(M-1)/2.*

Nonlinear mapping of linear uBSS problem?

Thus, linear uBSS problem characterized by (*N*,*M*,*K*) is converted into new one characterized by $(\bar{N}, P, Q)$, where *Q* denotes maximal number of overlapping sources in mapped domain. If sources do not overlapp heavily and higher-order terms are suppressed we have:

$$(\bar{N}/N) \gg (P/M) \quad \text{and} \quad (\bar{N}/N) \gg (Q/K)$$

where *Q*≈2*K* + *K(K*-1)/2. *P*≈ 2*M* + *M*(*M*-1)/2 above condition becomes:

$$(\bar{N}/N) \gg (M/2-3/2) \quad \text{and} \quad (\bar{N}/N) \gg (K/2-3/2)$$

The same procedure can be applied to equivalent linear representation $\mathbf{x} = \mathbf{G}\bar{\mathbf{s}}$ of the nonlinear BSS problem **x**=**f**(**s**).

Nonlinear mapping of linear uBSS problem?

The problem with using explicit feature maps $\phi(\mathbf{x}(t))$ is that $\bar{N}$ can be very large or even infinite. Thus, factorization problem:

$$\phi\ \mathbf{X}\ _{\tau} \approx \left[\underbrace{c_0\mathbf{e}_1 \ldots c_0\mathbf{e}_1}_{\times T\, times} \right] + \bar{\mathbf{B}} \left[\begin{matrix} 0 \\ \mathbf{S} \\ \mathbf{s}_{m_1}\mathbf{s}_{m_2} \quad {}^{M}_{m_1,m_2=1} \end{matrix} \right]$$

becomes computationally intractable.

Reproducible kernel Hilbert spaces

Definition 1. A real function $\kappa : \mathbb{R}^N \times \mathbb{R}^N \rightarrow \mathbb{R}$ is positive semi-definite if it is symmetric and satisfies for any finite set of points $\mathbf{x}_t \in \mathbb{R}^N \; {}_{t=1}^{T}$ and real numbers $\alpha_t \; {}_{t=1}^{T}$: $\sum_{i,j=1}^{T} \alpha_i \alpha_j \kappa \;\; \mathbf{x}_i , \mathbf{x}_j \;\; \geq 0$.

Theorem 1. The Morre-Aronszjan theorem [30]. Given any nonnegative definite function κ (**x**,**y**) there exists a uniquely determined RKHS H_κ consisting of real valued functions on set $\mathbf{X} \subset \mathbb{R}^N$ such that: (*i*) $\forall \mathbf{x} \in \mathbf{X},\; \kappa \;\; \circ , \mathbf{x} \;\; \in H_\kappa$;
(*ii*) $\forall \mathbf{x} \in \mathbf{X},\; \forall f \in H_\kappa ,\; f \;\; \mathbf{x} \;\; = \left\langle f , \kappa \;\; \circ , \mathbf{x} \;\; \right\rangle_{H_\kappa}$. Here, $\langle \circ , \circ \rangle$ denotes inner product associated with H_κ .

30. Aronszajn, N., "The theory of reproducing kernels," *Trans. of the Amer. Math. Soc*. 68, 337-404 (1950).

Reproducible kernel Hilbert spaces

Definition 2. Replacing *f*(**x**) in (*ii*) in Theorem 1 by κ(°,**x**) it follows $\kappa\ \mathbf{x}_t, \mathbf{x}\ = \langle \kappa\ \circ, \mathbf{x}_t\ , \kappa\ \circ, \mathbf{x}\ \rangle_{H_\kappa}$. By selecting the nonlinear map as ϕ(**x**)= κ(°,**x**) it follows $\kappa\ \mathbf{x}_t, \mathbf{x}\ = \langle \phi\ \mathbf{x}_t\ , \phi\ \mathbf{x}\ \rangle_{H_\kappa}$. That is known as ***kernel trick***. The nonlinear mapping ϕ(**x**) is known as as explicit feature map (EFM) associated with kernel κ(°,**x**).

Definition 3. Empirical kernel map (EKM), [31]. For a given set of patterns $\mathbf{v}_d \in \mathbb{R}^N\ {}_{d=1}^{D} \subset \mathbf{X}\ ,\ D \in \mathbb{N}$, we call $\psi : \mathbb{R}^N \rightarrow \mathbb{R}^D$:

$\mathbf{x}_t \mapsto \kappa\ \circ, \mathbf{x}_t\ \Big|_{\mathbf{v}_d\ {}_{d=1}^{D}} = \left[\kappa\ \mathbf{v}_1, \mathbf{x}_t\ , \ldots, \kappa\ \mathbf{v}_D, \mathbf{x}_t\ \right]^T\ {}_{t=1}^{T}$ the EKM with respect to $\mathbf{v}_d\ {}_{d=1}^{D}$.

31. Schölkopf, B., and Smola, A., *Learning with kernels*, MIT Press, 2002, pp. 42-45.

Nonlinear mapping of linear uBSS problem?

The problem with using explicit feature maps $\phi(\mathbf{x}(t))$ is that $\bar{N}$ can be very large or even infinite. Thus, factorization problem:

$$\phi\ \ \mathbf{X}\ \ _{\tau} \approx \left[\underbrace{c_0\mathbf{e}_1 \ldots c_0\mathbf{e}_1}_{\times T\,times}\right] + \bar{\mathbf{B}} \begin{bmatrix} 0 \\ \mathbf{S} \\ \mathbf{s}_{m_1}\mathbf{s}_{m_2} \ \ {}_{m_1,m_2=1}^{M} \end{bmatrix}$$

becomes computationally intractable. That is fixed by projecting $\phi(\mathbf{x}(t))$ onto $\phi(\mathbf{V})$ where $\mathbf{V} = \ \mathbf{v}_d \in \mathbb{R}^{N\times 1}\ {}_{d=1}^{D}$ stands for basis such that:

Then:

$$span\ \ \mathbf{v}_d\ \ {}_{d=1}^{D} \approx span\ \ \mathbf{x}_t\ \ {}_{t=1}^{T}$$

$$span\ \ \phi\ \ \mathbf{v}_d\ \ \ {}_{d=1}^{D} \approx span\ \ \phi\ \ \mathbf{x}_t\ \ \ {}_{t=1}^{T}$$

Nonlinear mapping of linear uBSS problem?

Projection yields:

$$\phi\ \mathbf{V}^{T}\phi\ \mathbf{x}_t\ = \psi\ \mathbf{x}_t\ {}_{\mathbf{V}} = \Big[\big\langle \phi\ \mathbf{v}_1\ ,\phi\ \mathbf{x}_t\ \big\rangle \ldots \big\langle \phi\ \mathbf{v}_D\ ,\phi\ \mathbf{x}_t\ \big\rangle\Big]^T$$

When ϕ(**x**)=*k*(°,**x**) it follows: <ϕ(**v**), ϕ(**x**)>=*k*(**v,x**). It is shown in [6] that when sources comply with sparse probabilistic model it applies:

$$\psi\ \mathbf{X}\ _{\tau} \approx \begin{bmatrix} \underbrace{c_0\mathbf{e}_1 \ldots c_0\mathbf{e}_1}_{\times T\ times} \end{bmatrix} + \bar{\mathbf{B}} \begin{bmatrix} 0 \\ \mathbf{S} \\ \mathbf{s}_{m_1}\mathbf{s}_{m_2}\ {}_{m_1,m_2=1}^{M} \end{bmatrix}$$

$\psi\ \mathbf{X}\ _{\tau} \in \mathbb{R}_{0+}^{D\times T}$ $\quad \bar{\mathbf{B}} \in \mathbb{R}_{0+}^{D\times P+1}$ and $P \approx 2M + M(M-1)/2.$

Subscript τ inidcates that some type of thresholding was applied on $\psi(\mathbf{X})$ to suppress *HOT*.

Nonlinear mapping of linear uBSS problem?

Nonlinear uBSS problem (*N*, *M*, *K*) is substituted by the linear BSS problem (*D*, 2*M* + *M*(*M*-1)/2, *Q*), *Q*≈2K + *K*(*K*-1)/2. Equvalent linear BSS problem is solvable when:

$$(D/N) \gg (P/M) \quad \text{and} \quad (D/N) \gg (Q/K)$$

Since *P*≈ 2*M* + *M*(*M*-1)/2 and *Q*≈2K + *K*(*K*-1)/2 above condition becomes:

$$(D/N) \gg (M/2-3/2) \quad \text{and} \quad (D/N) \gg (K/2-3/2)$$

That is possible to fulfill by finding basis $\mathbf{V} = \ \mathbf{v}_d \in \mathbb{R}^{N\times 1} \ {}_{d=1}^{D}$ with sufficiently large dimension *D*.

Nonlinear mapping of linear uBSS problem?

Basis $\mathbf{V} = \ \mathbf{v}_d \in \mathbb{R}^{N\times 1}\ {}_{d=1}^{D}$ needs to fulfill:

$$span\ \ \mathbf{v}_d\ {}_{d=1}^{D} \approx span\ \ \mathbf{x}_t\ {}_{t=1}^{T}$$

Thus, **V** can be found by clustering $\mathbf{x}_t\ {}_{t=1}^{T}$ into *D*≤*T* clusters. That, for example, can be accomplished by *kmeans* algorithm.

For *D*=*T* each data sample is a basis vector and clustering is not necessary. But, computational costs in matrix factorization stage (that follows) is very large.

When in addition to sparseness constraint nonnegativity constraints apply as well (that is the case in applications in imaging and/or mass spectrometry) sparseness constrained NMF algorithms can be applied to $\psi\ \ \mathbf{X}\ {}_{\tau}$ to estimate source components.

Nonnegative matrix factorization

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

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

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.

Nonnegative matrix factorization

Modern approaches to NMF problems have been initiated by Lee-Seung' Nature paper, [32], where it is proposed to estimate **A** and **S** through alternative minimization procedure of the two possibly different cost functions:

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

For k=1,2,…, until convergence do

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

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

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

32. 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:

$$\mathbf{A}^{*},\mathbf{S}^{*} \;=\underset{\mathbf{A},\mathbf{S}}{\arg\min}\, D\;\; \mathbf{X}\|\mathbf{AS} \;=\frac{1}{2}\left\|\mathbf{X}-\mathbf{AS}\right\|_F^2 \quad s.t.\, \mathbf{A}\geq\mathbf{0},\mathbf{S}\geq\mathbf{0}$$

- 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:

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

- 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 additional constraints are necessary to obtain factorization unique up to permutation and scaling. Generalization that involves constraints is given in [33]:

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

where $J_{\mathbf{S}}(\mathbf{S})=\sum_{m,t} s_{mt}$ and $J_{\mathbf{A}}(\mathbf{A})=\sum_{n,m} a_{nm}$ 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:

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

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

33. 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***(A**,**SIX**) yields:

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

Above formulation is equivalent to maximizing likelihood P**(XIA**,**S)** and maximizing prior probabilities P(**A**) and P(**S**). Assuming normal distribution of approximation error **E=X-AS** this yields:

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

Maximum a posteriori probability BSS/NMF

Assuming non-informative prior on **A**: *P*(**A**)=const and Laplacian (sparse) prior on **S**: $P(\mathbf{S}) = \exp - \ \left|\mathbf{s}_1\right| + \ldots + \left|\mathbf{s}_M\right|$ yields:

$$\mathbf{A}^*, \mathbf{S}^* \ = \underset{\mathbf{A},\mathbf{S}}{\arg\min} \frac{1}{2}\left\|\mathbf{X} - \mathbf{AS}\right\|_F^2 + \alpha_{\mathbf{S}} \left\|\mathbf{S}\right\|_1 \ s.t.\, \mathbf{A} \geq \mathbf{0}, \mathbf{S} \geq \mathbf{0}.$$

It is possible to select for *P*(**S**) prior other than Laplacian. That leads to general sparseness constrained factorization:

$$\mathbf{A}^*, \mathbf{S}^* \ = \underset{\mathbf{A},\mathbf{S}}{\arg\min} \frac{1}{2}\left\|\mathbf{X} - \mathbf{AS}\right\|_F^2 + \alpha_{\mathbf{S}} \left\|\mathbf{S}\right\|_p \ s.t.\ 0 < p \leq 1, \mathbf{A} \geq \mathbf{0}, \mathbf{S} \geq \mathbf{0}.$$

Nonnegative matrix factorization

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

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

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.:

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

Non-negative matrix under-approximation (NMU)

NMF algorithms outlined befor require a priori knowledge of sparseness related regularization (trade off) constant.

A sequential approach to NMF has been recently proposed in [34] by estimating rank-1 one factors $\mathbf{a}_m\mathbf{s}_\mathbf{m}$ one at a time. Each time $\mathbf{a}_m\mathbf{s}_\mathbf{m}$ is estimated it is removed from $\mathbf{X} \rightarrow \mathbf{X}\text{-}\mathbf{a}_m\mathbf{s}_\mathbf{m}$. To prevent subtraction from being negative the under-approximation constraint is imposed on $\mathbf{a}_m\mathbf{s}_\mathbf{m}$: $\mathbf{a}_m\mathbf{s}_\mathbf{m} \leq \mathbf{X}$.

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

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

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

Non-negative matrix under-approximation (NMU)

Theorem 1 in [34] proves that number of nonzero entries in **A** and **S** is less than in **X**. Thus, the underapproximation constraint ensures sparse (parts based) factorization of **X**. This, however, does not imply that **A** and **S** obtained by enforcing underapproximation constrain yields the sparseset decomposition of **X**.

However, 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:
https://sites.google.com/site/nicolasgillis/code

Non-negative matrix factorization with L_0-constraint (NMF_L0)

The NMF_L0 algorithm, [35], imposes explicit L_0-constraint on entries of **S**, i.e. number of nonzero entries is tried to be minimized explicitly by integrating nonnegativity constraint in the OMP algorithm. That is achieved through modifications of the nonnegative least square (NNLS) algorithm, [36], called sparse NNLS and recursive sparse NNLS. The mixing matrix is updated by some of standards dictionary update methods.

The „weak" side of the NMF_L0 algorithm is that, in addition to number of sources *M*, the maximal number of overlapped sources *K* has to be known *a priori*. Quite often that is hard to achieve in practice.

MATLAB code for NMF_L0 algorithm is available at:
http://www3.spsc.tugraz.at/people/robert-peharz.

35. R. Peharz, F. Pernkopf, "Sparse nonnegative matrix factorization with ℓ^0 constraints," *Neurocomputing*, vol. 80, pp. 38-46, 2012.
36. C. Lawson, R. Hanson, *Solving Least Squares Problems*, Prentice-Hall, 1974.

Nonlinear underdetermined blind source separation: numerical experiments and separation of pure components mass spectra from mixtures of nonlinear chemical reactions [5]

5. I. Kopriva, I. Jerić, M. Filipović, L. Brkljačić (2014). Empirical Kernel Map Approach to Nonlinear Underdetermined Blind Separation of Sparse Nonnegative Dependent Sources: Pure Components Extraction from Nonlinear Mixtures Mass Spectra. *J. of Chemometrics* , vol. 28, pp. 704-715.

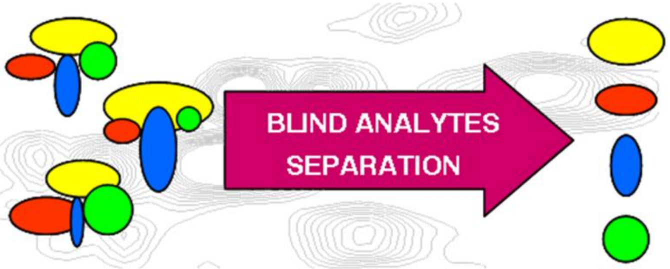
Linear mixing model

X=AS

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

In chemometrics (NMR spectroscopy or mass spectrometry) rows of **X** represent spectra of mixture samples, columns of **A** represent concentration profiles of analytes (a.k.a. pure components) present in mixture spectra **X**, while rows of **S** represent spectra of analytes present in mixture spectra **X**.

The (u)BSS problem relates to extraction of anlytes (and their concentratios) using mixture spectra **X** only:



Pure components can represent compounds indicative for disease. Thus, they can be useful for biomarker analysis. They can be isolated from spectra (NMR, mass) of biological samples (urine, blood, tissues).

Implementation details

Studies on numerical and experimental data reported below were executed on personal computer running under Windows 64-bit operating system with 64GB of RAM using Intel Core i7-3930K processor and operating with a clock speed of 3.2 GHz. MATLAB 2012b environment has been used for programming.

Electrospray ionization-mass spectrometry (ESI-MS) measurements operating in a positive ion mode were performed on a HPLC-MS triple quadrupole instrument equipped with an autosampler (Agilent Technologies, Palo Alto, CA, USA). The desolvation gas temperature was 300^0C with flow rate of 8.0 L/min. The fragmentor voltage was 135 V and capillary voltage was 4.0 kV. Mass spectra were recorded in m/z segment of 10-2000. All data acquisition and processing was performed using Agilent MassHunter software. Acquired mass spectra are composed of intensities at T=9901 m/z coordinates.

Numerical experiment

Nonlinear uBSS problem characterized by *N*=3,*M*=8, *K=3* and*T*=1000 is simulated:

$$f_1(\mathbf{s}) = s_1^3 + s_2^2 + \tan^{-1}(s_3) + s_4^2 + s_5^3 + s_6^3 + \tanh(s_7) + \sin(s_8)$$

$$f_2(\mathbf{s}) = \tanh(s_1) + s_2^3 + s_3^3 + \tan^{-1}(s_4) + \tanh(s_5) + \sin(s_6) + s_7^2 + s_8^2$$

$$f_3(\mathbf{s}) = \sin(s_1) + \tan^{-1}(s_2) + s_3^2 + s_4^3 + \tanh(s_5) + \sin(s_6) + s_7^3 + \tan^{-1}(s_8)$$

Each source signal is according to p.d.f. based on mixed state random variable model with exponential prior [5]:

$$p(s_{mt}) = \rho\delta\;\; s_{mt}\;\; + \; 1-\rho\;\; \delta^*\;\; s_{mt}\;\; f\;\; s_{mt}\quad \forall m = 1,\ldots,M\;\forall t = 1,\ldots,T$$

$$f\;\; s_{mt}\;\; = \;\; 1/\mu_m\;\; \exp\;\; -s_{mt}/\mu_m$$

where ρ_m=0.8 and μ_m=1.5×10^{-3} ∀*m*=1,...,*M*. .

Numerical experiment

Comparative performance analysis of NMU, NMF_L0, EKM-NMU, EKM-NMF_L0, PTs-EKM-NMU and PTs-EKM-NMF_L0 algorithms. Probability of zero state was ρ_m=0.8.

Four metrics used in comparative performance analysis were: number of associated components with normalized correlation coefficient greater than or equal to 0.6, mean value of correlation coefficient over all associated components, minimal value of correlation coefficient and number of pure components assigned incorrectly (that occurs due to poor separation).

All four metrics were calculated with respect to predefined labeling of the pure components stored in the library. Incorrect assignment implies that, based on maximal correlation criterion, two or more pure components are assigned to the same separated component.

Mean values and variance are reported and estimated over 10 Monte Carlo runs. The best result in each metric is in bold. The first three metrics are calculated only for correctly assigned components. That is why NMU and NMF_L0 appear to have comparable performance.

Numerical experiment

	NMU	**NMF_L0**	**EKM-NMU**	**EKM-NMF_L0**	**PTs_EKM-NMU**	**PTs-EKM-NMF_L0**
correlation G.E. 0.6	2.8±0.92	2.3±1.34	3.7±0.48	3.2±0.63	**3.8±0.42**	3.7±0.48
mean correllation	**0.70±0.03**	0.61±0.11	0.69±0.02	0.64±0.03	**0.70±0.03**	0.69±0.04
minimal correlation	**0.53±0.04**	0.42±0.08	0.51±0.03	0.45±0.04	0.52±.04	0.49±0.06
inccorect assignments	3.4±0.70	3.1±0.57	2.4±0.97	2.2±0.63	2.0±0.88	**1.5±1.43**

Nonlinear chemical reaction

9 nonlinear mixtures mass spectra were recorded in nonlinear chemical reaction related to peptide bond synthesis.

25 pure components were present in the mixtures. They were separated chromatographically which enabled formation of pure components library and validation of the algorithms' performances.

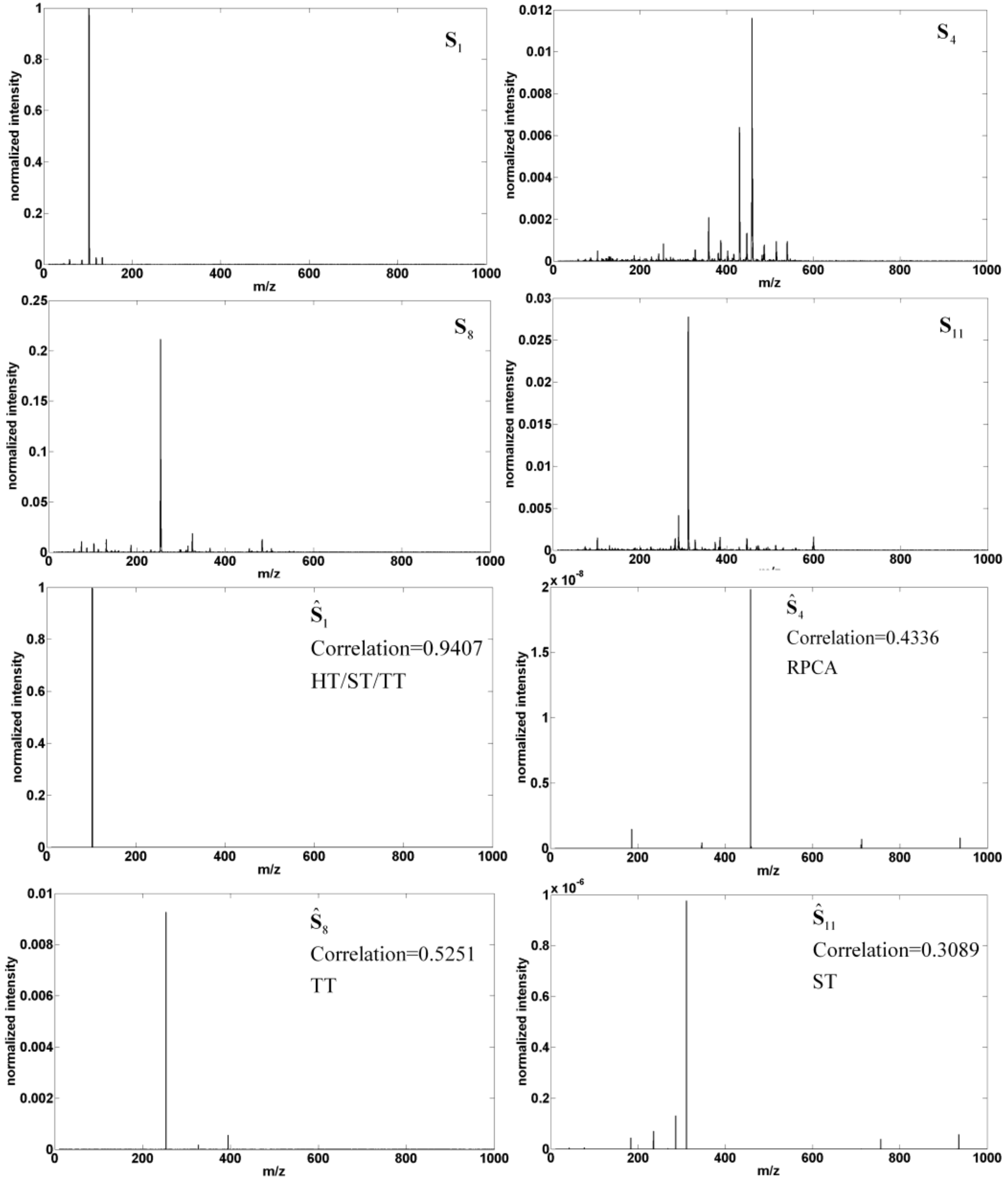
Pure components correlation matrix. 30 pairs of pure components have correlation greater than or equal to 0.1

	s_2	s_6	s_7	s_9	s_{10}	s_{12}
s_1	**0.9839**	0.1416	0.1218	0.1796	0.1072	0.3343
	s_6	s_7	s_9	s_{10}	s_{12}	
s_2	0.1418	0.1268	0.1797	0.1075	0.3305	
	s_{16}	s_{17}	s_{18}			
s_3	0.3575	0.3103	0.1716			
	s_6	s_{19}	s_{21}			
s_4	0.3077	0.3947	0.4005			
	s_7					
s_5	**0.7824**					
	s_9					
s_7	0.3297					
	s_{13}					
s_8	0.1293					
	s_{12}	s_{22}				
s_{11}	0.2666	0.1622				
	s_{17}					
s_{14}	0.1024					
	s_{22}					
s_{15}	0.1349					
	s_{17}					
s_{16}	**0.9783**					
	s_{18}					
s_{17}	0.1186					
	s_{21}					
s_{19}	**0.9962**					
	s_{24}	s_{25}				
s_{23}	0.4409	0.4339				
	s_{25}					
s_{24}	0.3008					

Nonlinear chemical reaction

	NMU	**NMF_L0**	**EKM-NMU**	**PTs_EKM-NMU** *D*=*T*=9901	**PTs-EKM-NMU** *D*=4000
correlation G.E. 0.6	8	14	16	**18**	**18**
mean correlation	0.342	0.518	0.673	**0.702**	**0.708**
minimal correlation	0.038	0.039	0.267	**0.419**	0.283
inccorect assignments	15	7	**0**	**0**	1
CPU time	1.3s	40 s	78.78h	4×78h*	4×13.7h*

* Sparseness constrained NMF had to be executed 4 times because 4 methods for supression of *HOT* have been applied to $\psi(\mathbf{X})$: hard, soft and trimmed threshodling as well as robust PCA



Mass spectra of several true and estimated pure components.

A Nonlinear Mixture Model Based Unsupervised Variable Selection in Genomics and Proteomics [37]

37. I. Kopriva, "A Nonlinear Mixture Model Based Unsupervised Variable Selection in Genomics and Proteomics," Bioinformatics 2015 - *6th Int. Conf. on Bioinformatics Models, Methods and Algorithms*, pp. 85-92, Lisbon, Portugal, January 12-15, 2015. DOI: 10.5220/0005161700850092.

Motivation

Disease diagnosis in proteomics and genomics is characterized by *small number* of samples (experiments) and *large number* of features (variables).

That results in classical “small *N* large *p* problem”, in which case classifiers and regression models are overly tuned to the training data (ovefitting).

Linear mixture models, often used in bioinformatics data analysis, represent samples as additive mixture of components.

State-of-the-art matrix factorization methods are used to extract those components using mixture samples only.

SCA in bioinformatics

Sparseness constrained NMF for BSS problems with sufficiently sparse sources is applied to microarray data analysis [38, 39].

Sparseness constrained NMF is used to decompose set of *N* gene expression profiles (mixtures in BSS) into *M* metagenes (sources in BSS). That yields metagenes comprised of small number of co-expressed genes. This indicates that they can be involved in disease (cancer) [40]. Thus, sparseness constraint is biologically justified. Extracted metagenes were confimed meaningful through subsequent biological analysis.

38. Stadtlthanner K, Theis FJ, Lang EW, Tomé AM, Puntonet CG, Górriz JM: Hybridizing Sparse Component Analysis with Genetic Algorithms for Microarray Analysis. *Neurocomputing* 2008, **71**: 2356-2376.
39. Gao Y, Church G: Improving molecular cancer class discovery through sparse non-negative matrix factorization. *Bioinformatics* 2005, **21**: 3970-3975.
40. Lee SI, Batzoglou S: Application of independent component analysis to microarrays. *Genome Biology* 2003, **4**: R76.

SCA in bioinformatics

How to *automatically* select/recognize metagene (component) comprised of disease relevant genes?

In state-of-the-art SCA/ICA algorithms [38-40] component associated with basis vector that is most colinear with the vector of labels (diagnosis) is selected as disease relevant component.

However, use of label information in component selection process prohibits usage of selected component for learning/training prediction models (label information cannot be used twice: for component selection and cross-validation).

Novel type of additive linear mixture model comprised of test and reference sample has been proposed in [41] to enable automatic selection of component with disease specific features on a *sample-by-sample* basis.

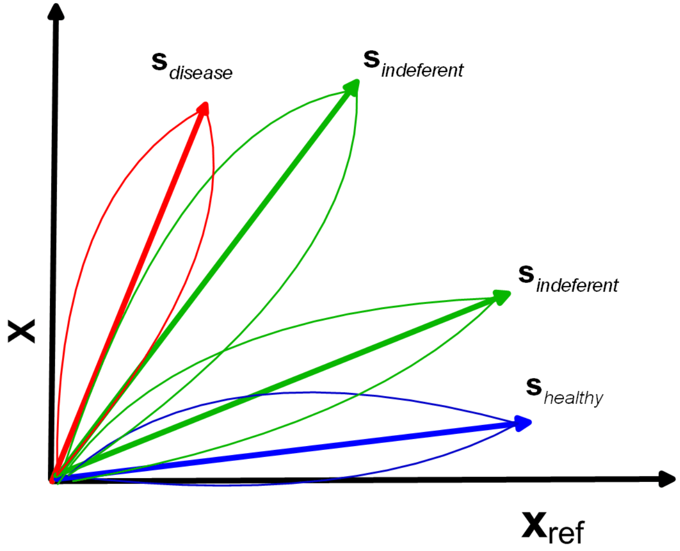
41. I. Kopriva, M. Filipović, "A mixture model with a reference-based automatic selection of components for disease classification from protein and/or gene expression levels," *BMC Bioinformatics*, vol. **12**, pp. 496 (17 pages), 2011.

Linear mixture model with a reference sample

Novel linear mixture model is comprised of actual test sample under consideration and a reference sample representing disease and/or control group. Number of additive components *M* is unknown and is estimated by cross-validation.

$$\begin{bmatrix} \mathbf{x}_{\text{control}} \\ \mathbf{x} \end{bmatrix} = \mathbf{A}_{\text{control}}\mathbf{S}_{\text{control}} :$$

$$\mathbf{X} \in R^{2\times T}, \mathbf{A}_{\text{control}} \in R^{2\times M}, \mathbf{S}_{\text{control}} \in R^{M\times T} \quad M \geq 2$$



Component with specific features is selected automatically in mixing angles domain by exploiting geometry of linear mixture model.

Nonlinear mixture model with a reference sample

However, it is known that interactions within gene regulatory networks can be nonlinear [11].

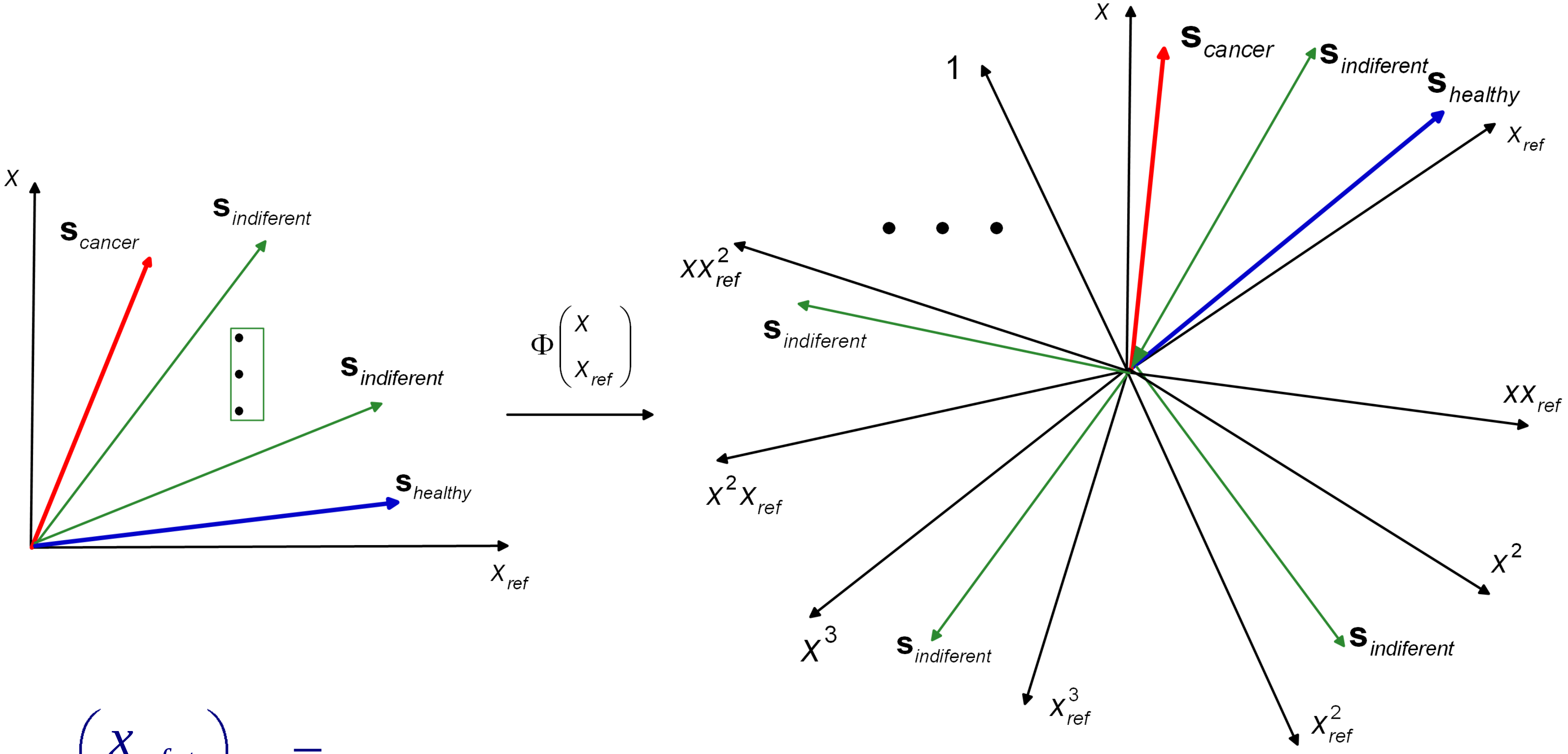
In metabolomics living organisms are examples of complex nonlinear systems that function far from equilibrium. Internal and external stimuli (disease, drug treatment, environmental changes) cause perturbations in the system as a result of highly synchronized molecular interactions [10].

Thus, the question is can linear mixture model with a reference sample [41] be generalized to the nonlinear one?

11. Yuh, C. H., Bolouri, H., and Davidson, E. H, “Genomic cis-regulatory logic: experimental and computational analysis of a sea urchin gene,” *Science*, vol. **279**, pp.1896-1902, 1998.

10. Walleczek J(ed). Self-organized biological dynamics and non-linear control. Cambridge University Press: Cambridge, UK. 2000

Nonlinear mixture model with a reference sample


x
$\mathbf{s}_{cancer}$
$\mathbf{s}_{indiferent}$
$\mathbf{s}_{indiferent}$
$\mathbf{s}_{healthy}$
x_{ref}
$\Phi\begin{pmatrix} x \\ x_{ref}\end{pmatrix}$
x
1
$\mathbf{s}_{cancer}$
$\mathbf{s}_{indiferent}$
$\mathbf{s}_{healthy}$
x_{ref}
xx_{ref}^2
$\mathbf{s}_{indiferent}$
xx_{ref}
x^2x_{ref}
x^2
x^3
$\mathbf{s}_{indiferent}$
$\mathbf{s}_{indiferent}$
x_{ref}^3
x_{ref}^2

$$\phi\begin{pmatrix} x_{ref,t} \\ x_{nt}\end{pmatrix} \approx \bar{\mathbf{A}}_n \bar{\mathbf{s}}_{t;n} \quad t=1,\dots,T$$

Nonlinear mapping of linear uBSS problem?

How to choose nonlinear mapping ? One, smart (?), way is to select ϕ by factorizing positive semi-definite symmetric kernel function *k*(**x**,**y**) on the basis of reproducibility condition:

$$k(\mathbf{x},\mathbf{y}) = \left\langle \phi(\mathbf{x}), \phi(\mathbf{y}) \right\rangle$$

For Gaussian kernel $k(\mathbf{x},\mathbf{y}) = \exp\left(-\left\|\mathbf{x}-\mathbf{y}\right\|^2 \big/ \sigma^2\right)$ we obtain:

$$\phi(\mathbf{x}) = \mathrm{e}^{-\frac{\left\|\mathbf{x}\right\|^2}{\sigma^2}} \sum_{r=0}^{d} \sum_{|\boldsymbol{\alpha}|=r} \frac{1}{\sigma^r} \sqrt{\frac{2^r}{\boldsymbol{\alpha}!}} \mathbf{x}^{\boldsymbol{\alpha}}$$

$$\boldsymbol{\alpha} \in \mathbb{N}_0^N \qquad |\boldsymbol{\alpha}| = \sum_{n=1}^{N} \alpha_n \qquad \boldsymbol{\alpha}! = \prod_{n=1}^{N} \alpha_n! \qquad \mathbf{x}^{\boldsymbol{\alpha}} = \prod_{n=1}^{N} x_n^{\alpha_n}$$

Nonlinear mapping of linear uBSS problem?

Example. For $\mathbf{x} \in \mathbb{R}^3$ approximate EFM of order *d*=3 associated with Gaussian kernel:

$$\hat{\Phi}_{\kappa}\ \mathbf{x}\ = e^{-\frac{\|\mathbf{x}\|_2^2}{\sigma^2}} \begin{bmatrix} 1 \frac{\sqrt{2}}{\sigma}x_1 \frac{\sqrt{2}}{\sigma}x_2 \frac{\sqrt{2}}{\sigma}x_3 \frac{\sqrt{2}}{\sigma^2}x_1^2 \frac{\sqrt{2}}{\sigma^2}x_2^2 \frac{\sqrt{2}}{\sigma^2}x_3^2 \frac{2}{\sigma^2}x_1x_2 \frac{2}{\sigma^2}x_1x_3 \frac{2}{\sigma^2}x_2x_3 \frac{1}{\sigma^3}\ldots \\ \sqrt{\frac{4}{3}}x_1^3 \frac{1}{\sigma^3}\sqrt{\frac{4}{3}}x_2^3 \frac{1}{\sigma^3}\sqrt{\frac{4}{3}}x_3^3 \frac{2}{\sigma^3}x_1x_2^2 \frac{2}{\sigma^3}x_1x_3^2 \frac{2}{\sigma^3}x_1^2x_3 \frac{2}{\sigma^3}x_1^2x_2 \frac{2}{\sigma^3}x_2^2x_3 \frac{2}{\sigma^3}x_2x_3^2 \frac{\sqrt{8}}{\sigma^3}x_1x_2x_3 \end{bmatrix}^{\mathrm{T}}$$

Thus, 3D vector is mapped into 20D vector. Second order mapping yields 10D vector.

Nonlinear mixture model with a reference sample

Regarding σ we have found that when data are scaled to [-1, 1] interval, σ can be approximately set to 1.

Dimension *D* of mapping induced space depends on order of the mapping d through: *D*=(d+2)(d+1)/2.

Mapped sample can also be written in Cartesian coordinate system as:

$$\phi\begin{pmatrix} x_{ref,k} \\ x_{nk} \end{pmatrix} = \mathbf{e}_1 + c_1 x_{ref,k}\mathbf{e}_2 + c_2 x_{nk}\mathbf{e}_3 + c_3 x_{ref,k} x_{nk}\mathbf{e}_4 + \ldots$$

where $\mathbf{e}_j \in R^D$, *j*=1,…,*D* are unit vector in standard Euclidean basis.

Nonlinear mixture model with a reference sample

We can calculate cosines of the angles that mixing vectors $\bar{\mathbf{a}}_{m;n} \in R_{0+}^{D}$ close with a reference sample according to:

$$\cos\angle\ \bar{\mathbf{a}}_{m;n},\mathbf{x}_{ref}\ \ =\left\langle \bar{\mathbf{a}}_{m;n},\mathbf{e}_{2}\right\rangle \Big/ \left\|\bar{\mathbf{a}}_{m;n}\right\|$$

When reference sample represents negative (healthy) class, component comprised of disease (cancer) relevant features is associated with a mixing vector that closes largest angle with a reference sample. Hence after executing sparseness constrained factorization of

$$\phi\begin{pmatrix}\mathbf{x}_{ref}\\ \mathbf{x}_{n}\end{pmatrix}\approx \bar{\mathbf{A}}_{n}\bar{\mathbf{S}}_{n}$$

component with disease relevant features is selected automatically:

$$\bar{\mathbf{s}}_{cancer;n}=\arg\min_{m}\cos\angle\ \bar{\mathbf{a}}_{m;n},\mathbf{x}_{ref}$$

Nonlinear mixture model with a reference sample

After each sample is decomposed components comprised of cancer relevant variables are stored row-wise in a matrix $\bar{\mathbf{S}}_{cancer} \in R^{N\times T}$.

Variables (columns of $\bar{\mathbf{S}}_{cancer}$) are ranked by their variance accross sample dimension yielding: $\bar{\mathbf{S}}_{cancer}^{ranked} \in R^{N\times K}$.

Let us denote by *I* a corresponding index set. Variables ranked in original space of samples are obtained by indexing each sample by *I*, that is: $\mathbf{x}_n^{ranked} = \mathbf{x}_n$, *n*=1,…,*N*.

Samples with ranked variance form rows of the matrix: $\mathbf{X}^{ranked} \in R^{N\times T}$. That matrix when paired with a vector of labels **y** is used to learn SVM-based diagnostic models.

Sparseness constrained decomposition

Sample dependent sparseness constrained factorization of linear mixtutre model in mapped space:

$$\phi\begin{pmatrix}\mathbf{x}_{ref}\\ \mathbf{x}_n\end{pmatrix}\approx\bar{\mathbf{A}}_n\bar{\mathbf{S}}_n$$

is peformed in two steps.

Step 1. Mixing matrix $\bar{\mathbf{A}}_n$ is estimated first. That is achieved by separable NMF algorithm [42] which also estimates number of components M_n. MATAB code is available at: https://sites.google.com/ site/nicolasgillis/publications. There is no parameter reuquired to be tuned of defined *a priori*.

42. Gillis, N., and Vavanis, S. A., “Fast and Robust Recursive Algorithms for Separable Nonnegative Matrix Factorization,” *IEEE Trans. on Pattern Analysis and Machine Intelligence* **36**, 698--714 (2014).

Sparseness constrained decomposition

Step 2. Based on estimated mixing matrix $\hat{\bar{\mathbf{A}}}_n$ matrix of componnets (sources) $\bar{\mathbf{S}}_n$ is estimaed by solving sparseness constrained optimization problem:

$$\hat{\bar{\mathbf{S}}}_n = \min_{\bar{\mathbf{S}}_n} \left\{ \frac{1}{2} \left\| \hat{\bar{\mathbf{A}}}_{\mathrm{n}} \bar{\mathbf{S}}_n - \phi\begin{pmatrix} \mathbf{x}_{ref} \\ \mathbf{x}_n \end{pmatrix} \right\|_F^2 + \lambda \left\| \bar{\mathbf{S}}_n \right\|_1 \right\}$$

where λ stands for regularization constant and has to be cross-validated. We have used the iterative shrinkage thresholding (IST) type of method [43] to solve this optimization problem. MATLAB code is available at: http://ie.technion.ac.il/Home/Users/ becka.html.

[43] Beck, A., and Teboulle, M., "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM J. on Imag. Sci*., vol. 2, pp.183-202, 2009.

Comparative performance analysis

Proposed unsupervised variable selection method has been compared with 3 supervised vairable selection methods: maximum mutual information minimal redundancy (MIMR) [44] , HITTON_PC and HITTON_MB methods [45, 46] , and its linear counterpart [41].

To comply with reproducible research principles Gene Expression Model Selector (GEMS) software system [47], has been used for cross-validation and learning of SVM-based diagnostic models with polynomial and Gaussian kernels the parameters of which were optimized in cross-validation loop as well. The system is available online at: http://www.gems-system.org/. HITON_PC and HITON_MB algorithms are implemented in GEMS software system while implementation of the MIMR algorithm is available at MATLAB File Exchange.

44. Brown, G., "A New Perspective for Information Theoretic Feature Selection," *J. Mach. Learn. Res*., vol. 5, pp. 49-56, 2009.
45. Aliferis, C. F., *et al*., "Local Causal and Markov Blanket Induction for Causal Discovery and Feature Selection for Classification - Part I: Algorithms and Empirical Evaluation," *J. Mach. Learn. Res*., vol. 11, pp.171-234, 2010.
46. Aliferis, C. F., *et al*., "Local Causal and Markov Blanket Induction for Causal Discovery and Feature Selection for Classification - Part II: Analysis and Extensions," *J. Mach. Learn. Res*., vol. 11, pp. 235-284, 2010.
47. Statnikov, A., *et al*., "GEMS: A system for automated cancer diagnosis and biomarker discovery from microarray gene expression data," *Int. J. Med. Informatics*, vol. 74, pp.491-503, 2005.

Comparative performance analysis

Variable selection methods were compared on three well known datasets in genomics: colon cancer [48], diffuse large b-cell lymphoma and follicular lymphomas (DLBCL/FL) [49] and prostate cancer [50], and two well known datasets in proteomics: ovarian cancer [51] and prostate cancer [52].

Dataset	Number of samples (cancer/normal)	Number of variables	Reference
1. Prostate cancer	52 /50	10509	[50]
2. Colon cancer	40/22	2000	[48]
3.DLBCL/FL	58/19	5469	[49]
4. Ovarian cancer	100/100	15152	[51]
5. Prostate cancer	69/63	15154	[51]

48. Alon, U., *et al*., "Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays," *Proc. Natl. Acad. Sci. USA*, vol. 96, pp.6745-6750, 1999.

49. Shipp, M. A., *et al*., "Diffuse large B-cell lymphoma outcome prediction by gene expression profiling and supervised machine learning. *Nature Med*., vol. 8, pp.68-74, 2002.

50. Singh, D., *et al*., "Gene expression correlates of clinical prostate cancer behavior," *Cancer Cell*, vol. 1, pp.203-209, 2002.

51. Petricoin, E.F., *et al*., "Use of proteomic patterns in serum to identify ovarian cancer," *The Lancet*, vol. 359, pp. 572-577, 2002.

52. Petricoin, E.F., *et al.* ,"Serum proteomic patterns for detection of prostate cancer," *J. Natl. Canc. Institute*, vol. 94, pp.1576-1578, 2002.

Comparative performance analysis

For each dataset we report the best result achieved by one of these supervised methods. The results were obtained by 10-fold cross-validation. For each of five datasets proposed method achieves result that is worse than but comparable with the result of state-of-the-art supervised algorithm and much better than its linear unsupervised counterpart. Since reported results are achieved with small number of variables the probability of overfitting is reduced. Thus, it is reasonable to expect that performance on unseen data of the same cancer type by proposed unsupervised method will be better than the one achieved with supervised algorithms.

Dataset	Proposed method	Supervised method	[41]
1. Prostate cancer [50]	91.27% / **38 genes** (*d*=2, λ=0.4).	MIMR: 98.09% / **10 genes**.	94.27% / **477 genes**.
2. Colon cancer [48]	91.91% / **24 genes** (*d*=5, λ=0.1).	HITON_MB: 93.33% **4 genes**.	90.48% / **30 genes**, λ=0.05.
3. DLBCL/FL [49]	96.25% / **14 genes** (*d*=2, λ=0.2).	HITON_PC: 100% / **6 genes**.	98.57% / **169 genes**, λ=0.01.
4. Ovarian cancer [51]	93% / **7 *m/z* lines** (*d*=4, $\lambda \in$[0.4, 0.7]).	HITON_PC: 99.5% / **7 *m/z* lines**.	82% / **25 *m/z* lines**, λ=0.2.
5. Prostate cancer [52]	94.06% / **14 *m/z* lines** (*d*=4, λ=0.2). .	MIMR: 100% / **10 *m/z* lines**	94.01% / **85 *m/z* lines**,λ=0.2.

Nonlinear decomposition of RGB image of unstained specimen in histopatology

53. I. Kopriva**,** M. Hadžija, M. Popović-Hadžija, M. Korolija, A. Cichocki (2011). Rational Variety Mapping for Contrast-Enhanced Nonlinear Unsupervised Segmentation of Multispectral Images of Unstained Specimen, *The American Journal of Pathology*, vol. **179**, No. 2, pp. 547-553.

Decomposition of RGB image in histopatology

Decomposition of objects with spectrally similar profiles is hard problem. That occurs due to poor spectral resolution or due to physiological reasons.

“Standard” way of enhancing visual contrast is by means of staining i.e.using contrast agens to treat a specimen.

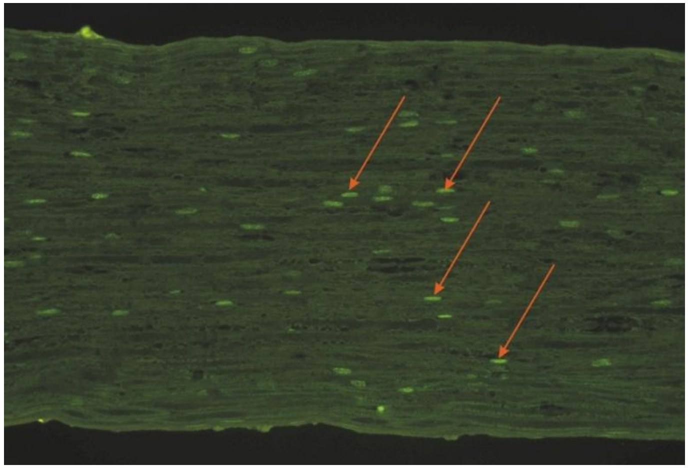
This, possibly, can also be achieved by digital image analysis through nonlinear sparse component analysis (NSCA).

RGB image is first mapped nonlinearly by means of EFM: **X**→ ϕ(**X**). Afterwards, sparseness constrained NMF is executed in induced space: **S**≈sNMF(ϕ(**X**)).

EFM of order *d:*

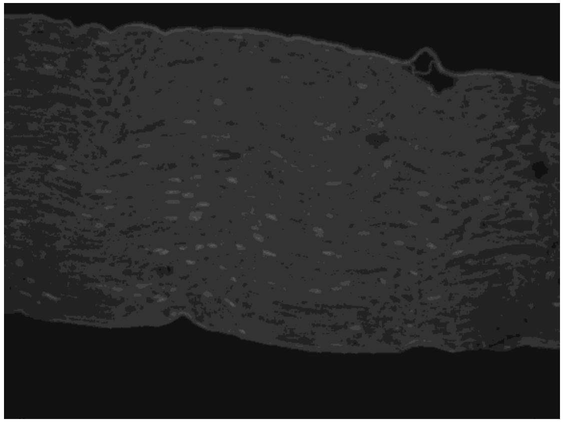
$$\phi \ \mathbf{x}(t) \ = \Big[\ x_1^{q_1}(t)...x_N^{q_N}(t) \ {}_{q_1,...,q_N=0}^{d} \Big]^T \text{ such that } \sum_{n=1}^{N} q_n \le d, \quad \forall t = 1,...,T.$$

RGB image of a nerve (*nervus ischiadicus*)

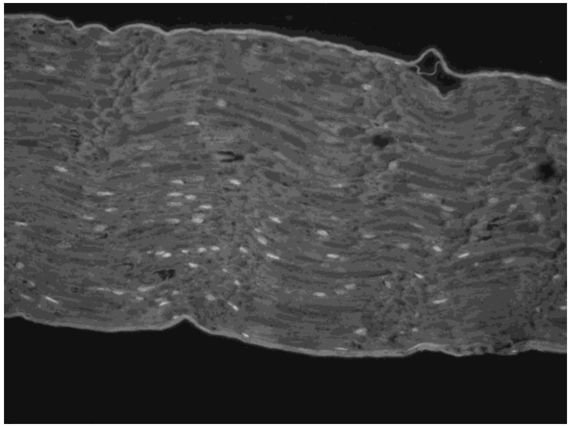


Nerves in RGB image of unstained specimen

Spectral channels of RGB image of unstained specimen



red



green

blue

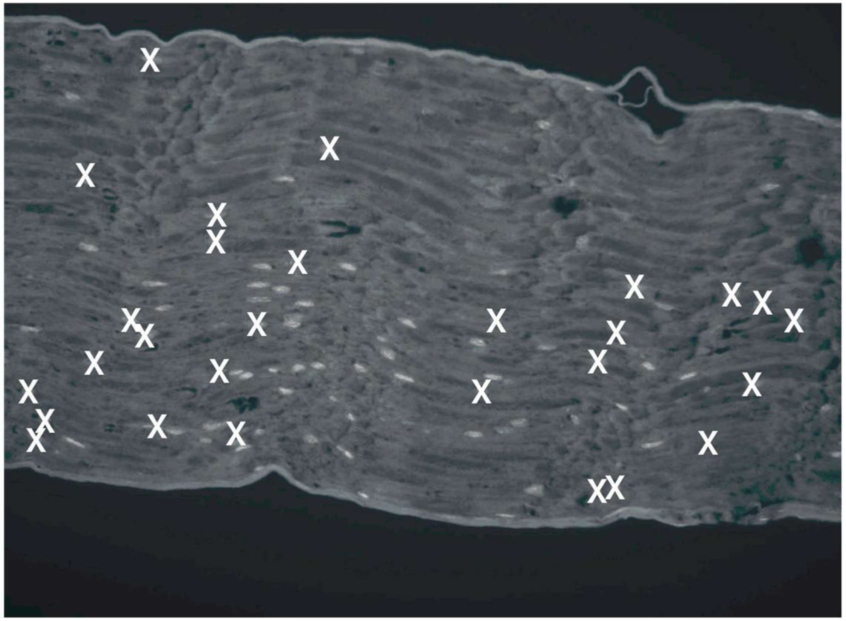
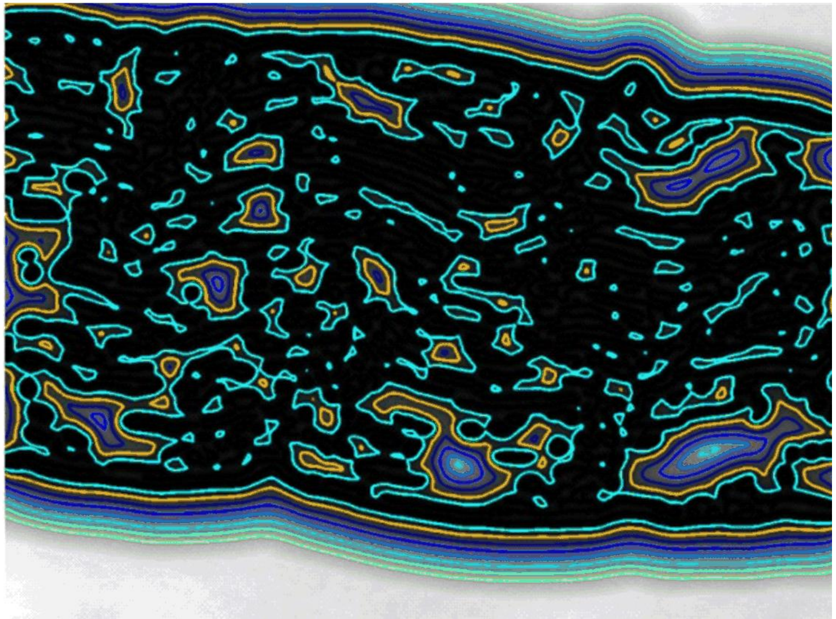
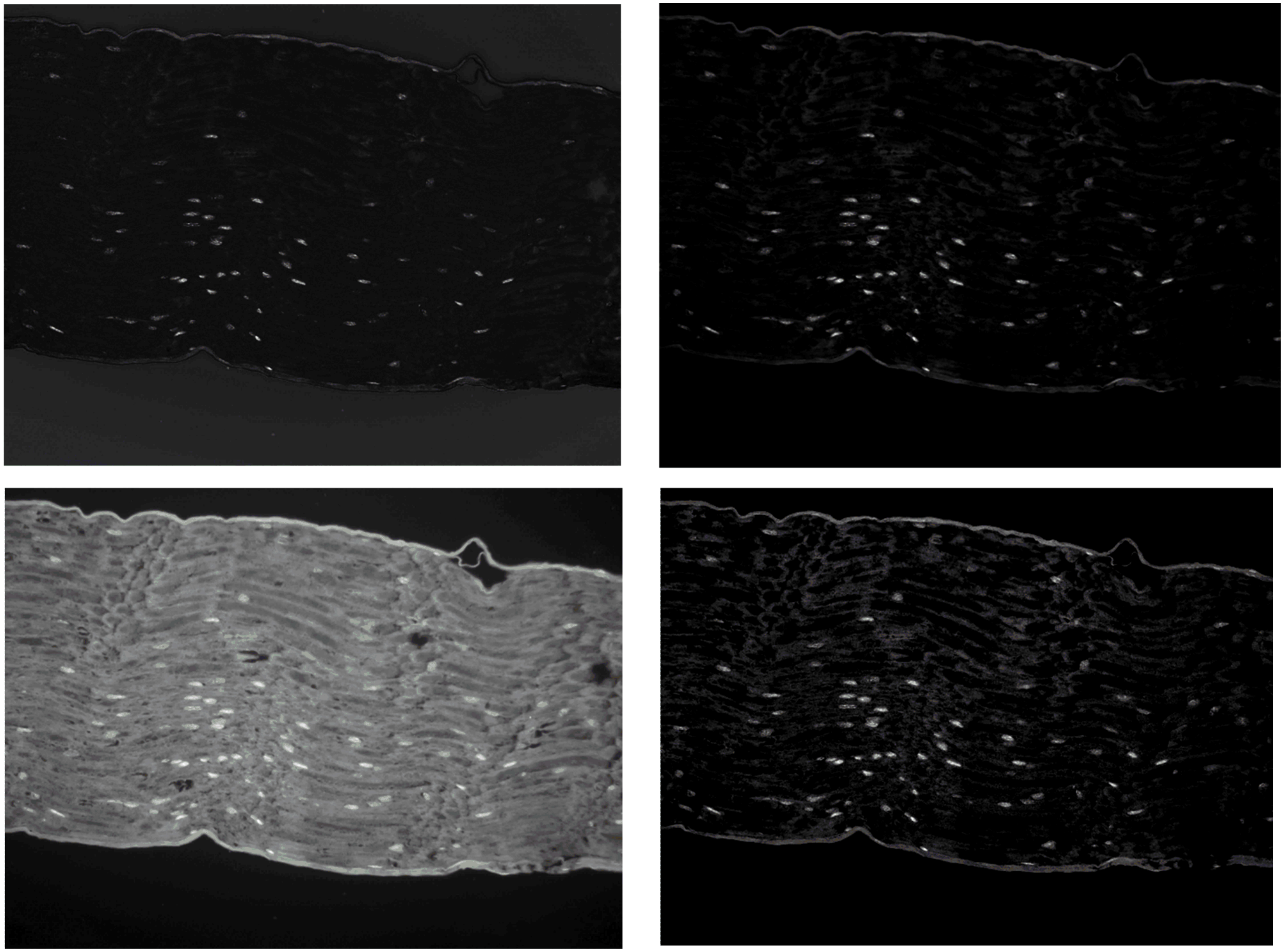


Image of unstained specimen at 510 nm wavelength (green color). White crosses denote false positive spots.



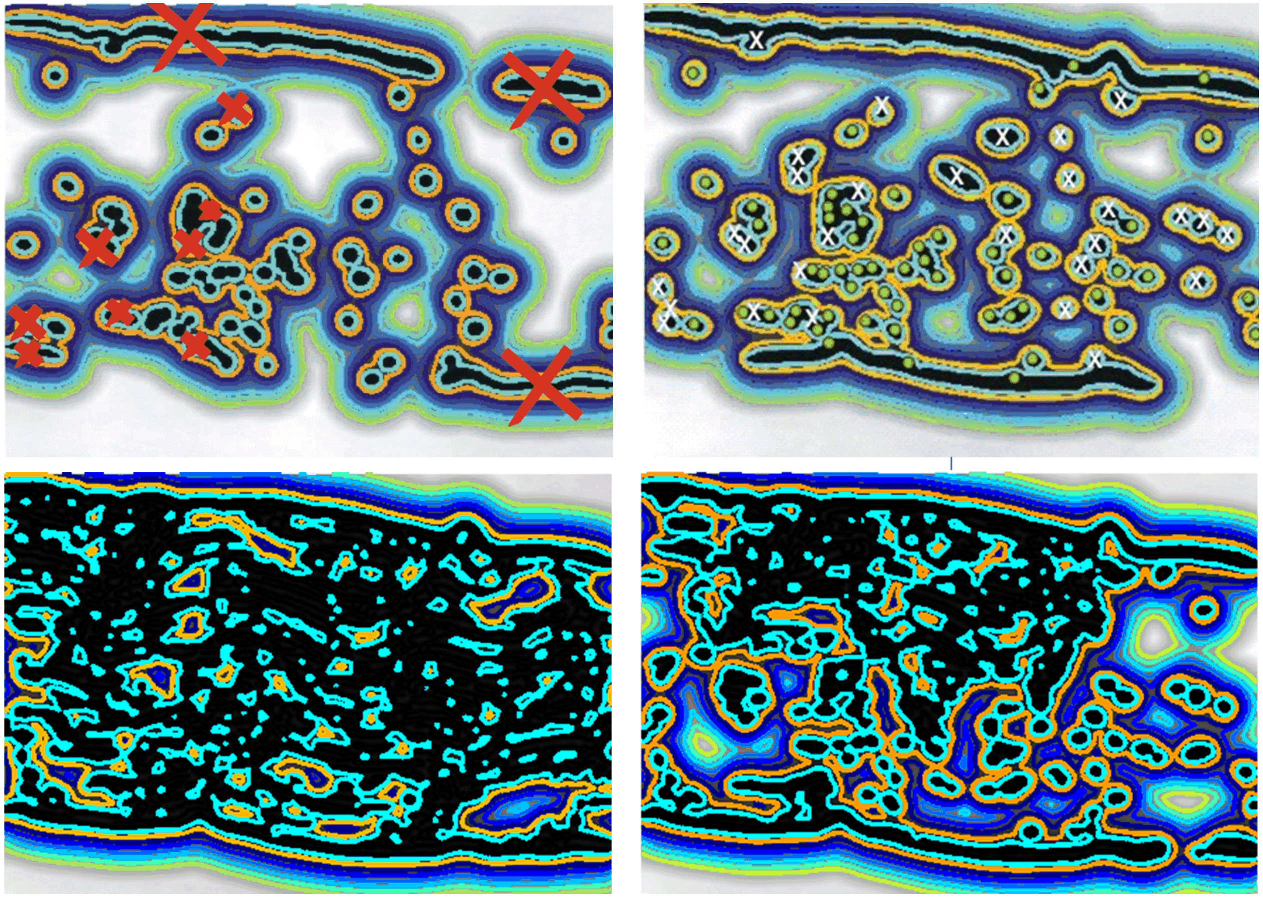
Active contours



Top left: EFM2 and DCA decomposition; Bottom left: DCA decomposition only.

Top right: EFM3 and NMU decompsoition; Bottom right: NMU decomposition only.

Active contours for decomposed nerve component



Top left: EFM2 and DCA decomposition; Bottom left: DCA decomposition only.

Top right: EFM3 and NMU decompsoition; Bottom right: NMU decomposition only.

Nonlinear decomposition of RGB image of skin cancer

54. I. Kopriva**,** A. Peršin (2009) Unsupervised decomposition of low-intensity low-dimensional multi-spectral fluorescent images for tumour demarcation, *Medical Image Analysis* , vol.**13**, pp.507-518.

55. I. Kopriva, Method for real time tumour visualisation and demarcation by means of photodynamic diagnosis, *US Patent* 8,224,427, 17. 7. 2012.

Nonlinear decomposition of RGB image of skin cancer

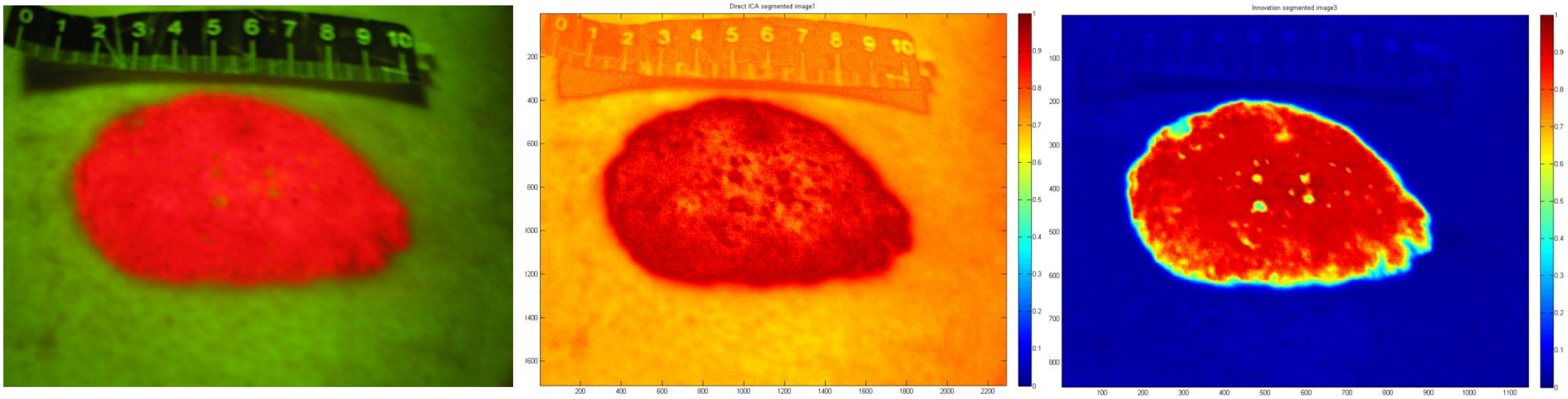
Decomposition of objects with spectrally similar profiles is hard problem. That occurs due to poor spectral resolution or due to physiological reasons.

In fluorescent imaging that occurs when intensity of fluorescence is weak.

As an example that may happen when intensity of illuminating (laser) light is low (we do not to cause damage).

RGB image is first mapped nonlinearly by means of 2nd order explicit feature map (EFM2). Afterwards, dependent component analysis (DCA) is executed in induced space.

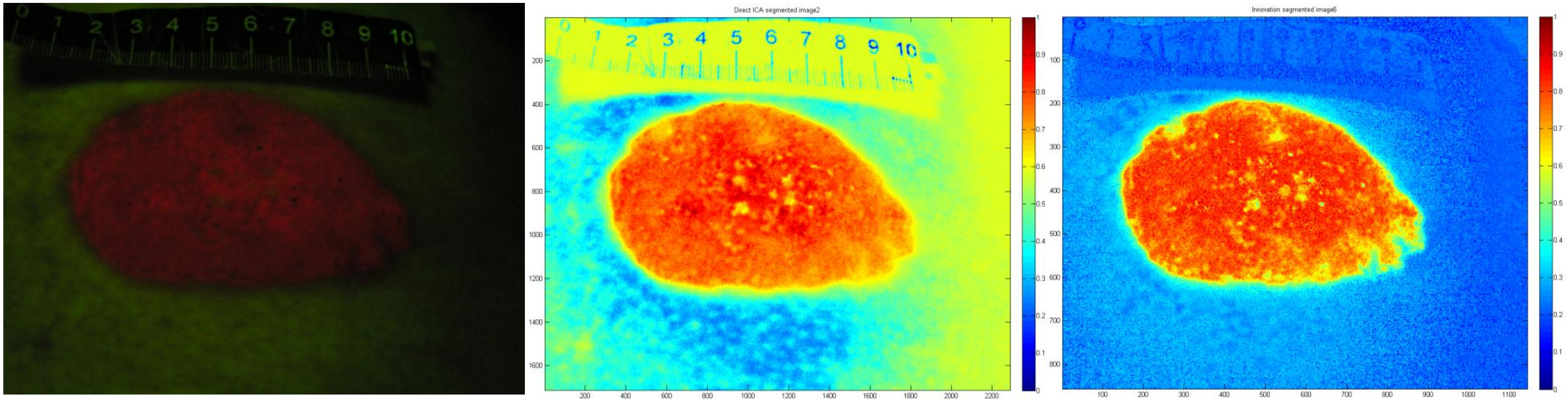
2nd order EFM and dependent component analysis



Left: Experimental high-intensity fluorescent RGB image of the skin tumour (basal cell carcinoma).

Mid: linear ICA algorithm; **Right:** Nonlinear DCA algorithm.

2nd order EFM and dependent component analysis



Left: Experimental low-intensity fluorescent RGB image of the skin tumour (basal cell carcinoma).

Mid: linear ICA algorithm; **Right:** Nonlinear DCA algorithm.

Nonlinear projection on orthonormal basis and ICA?

If centered basis in RKHS is orthogonalized:

$$\psi_c \;\mathbf{V} \;=\; \mathbf{I}_D - 1/D\,\mathbf{1}_D\mathbf{1}_D^T \;\psi\;\mathbf{V}\;\; \mathbf{I}_D - 1/D\,\mathbf{1}_D\mathbf{1}_D^T \;\; = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^T$$

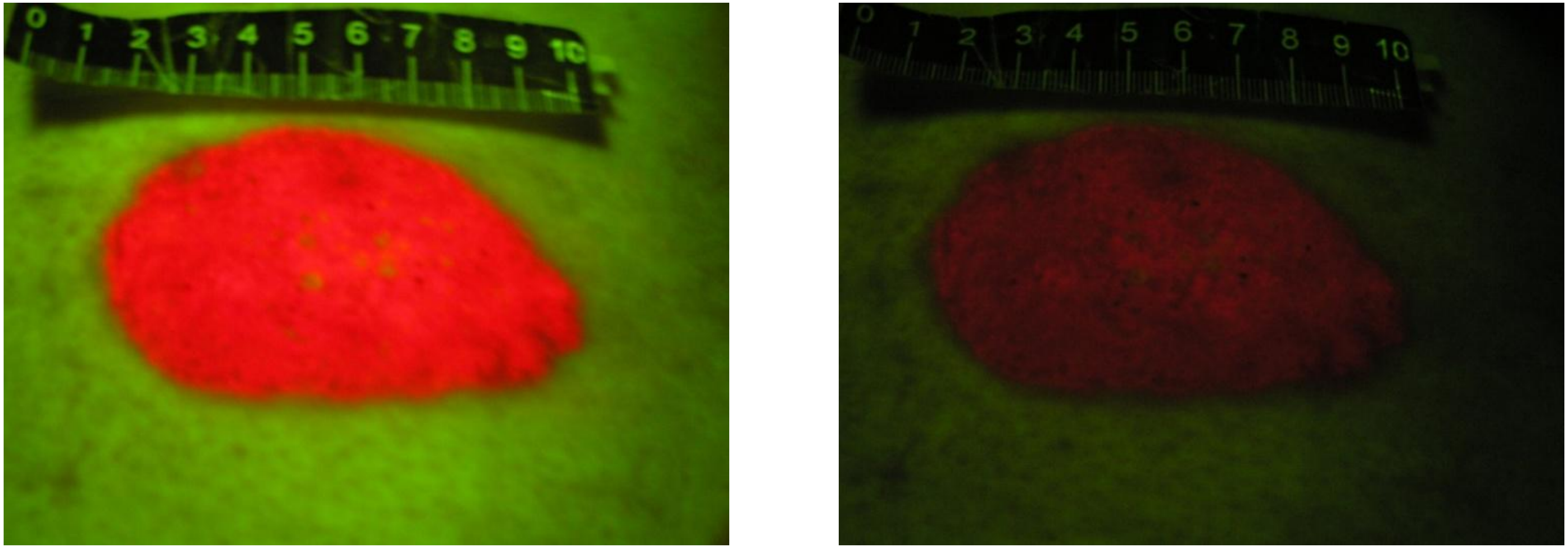
we can project centered data in RKHS onto $\psi_c \;\mathbf{V}$:

$$\mathbf{Y} = \boldsymbol{\Lambda}^{-1/2}\mathbf{U}^T \;\mathbf{I}_D - \;1/D\;\mathbf{1}_D\mathbf{1}_D^T \;\; \psi\;\mathbf{X}\; - \;1/D\;\psi\;\mathbf{V}\;\mathbf{1}_D\mathbf{1}_T^T$$

Thus, **Y** contains decorrelated components. By applying some linear ICA algorithm on **Y** we obtain nonlinear ICA of **X**.

Nonlinear projection on orthonormal basis and ICA?

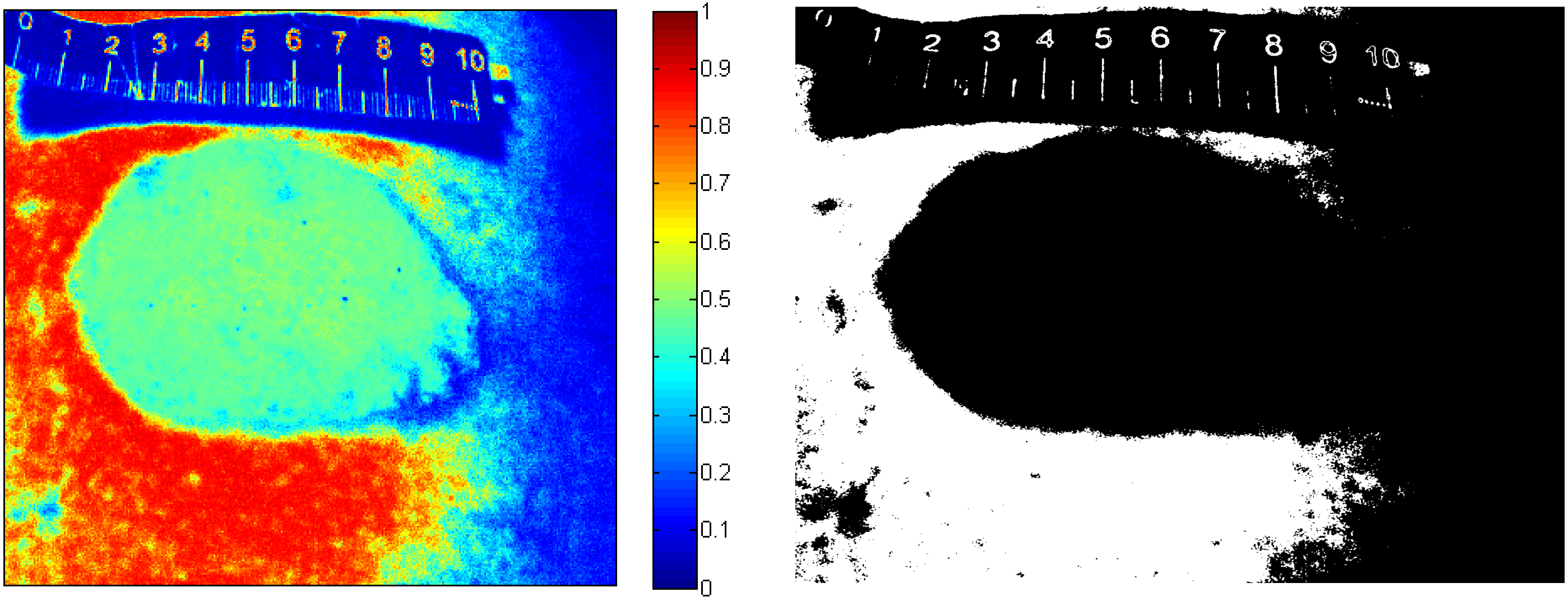
We have applied “temporal predictability” ICA algorithm, [56], on RKHS-deccorelated version of the low-intensity fluorescent RGB image of basal cell carcinoma.



Left: high-intensity fluorescent RGB image of BCC. **Right**: low-intensity fluorescent image of BCC.

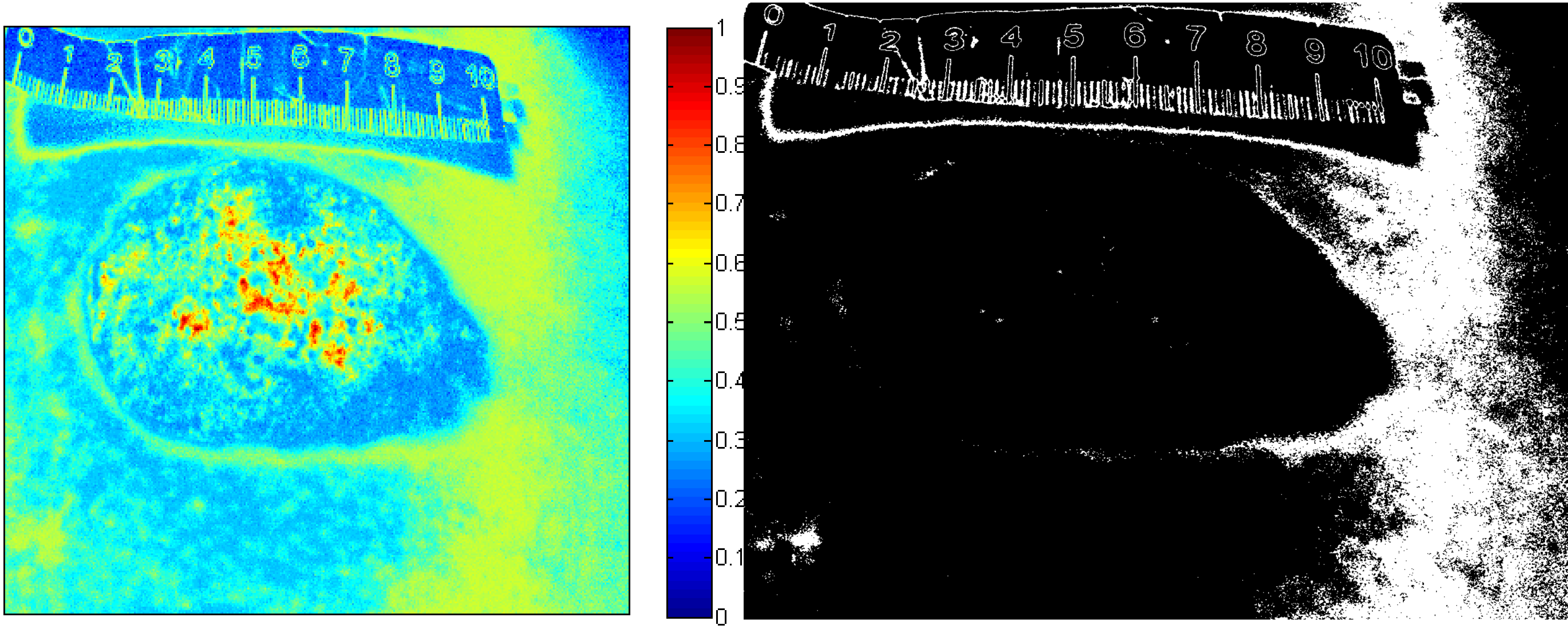
[56] Stone, J.V., 2001. Blind source separation using temporal predictability. *Neural Comput*. 13, 1559-1574.

Nonlinear projection on orthonormal basis and ICA?



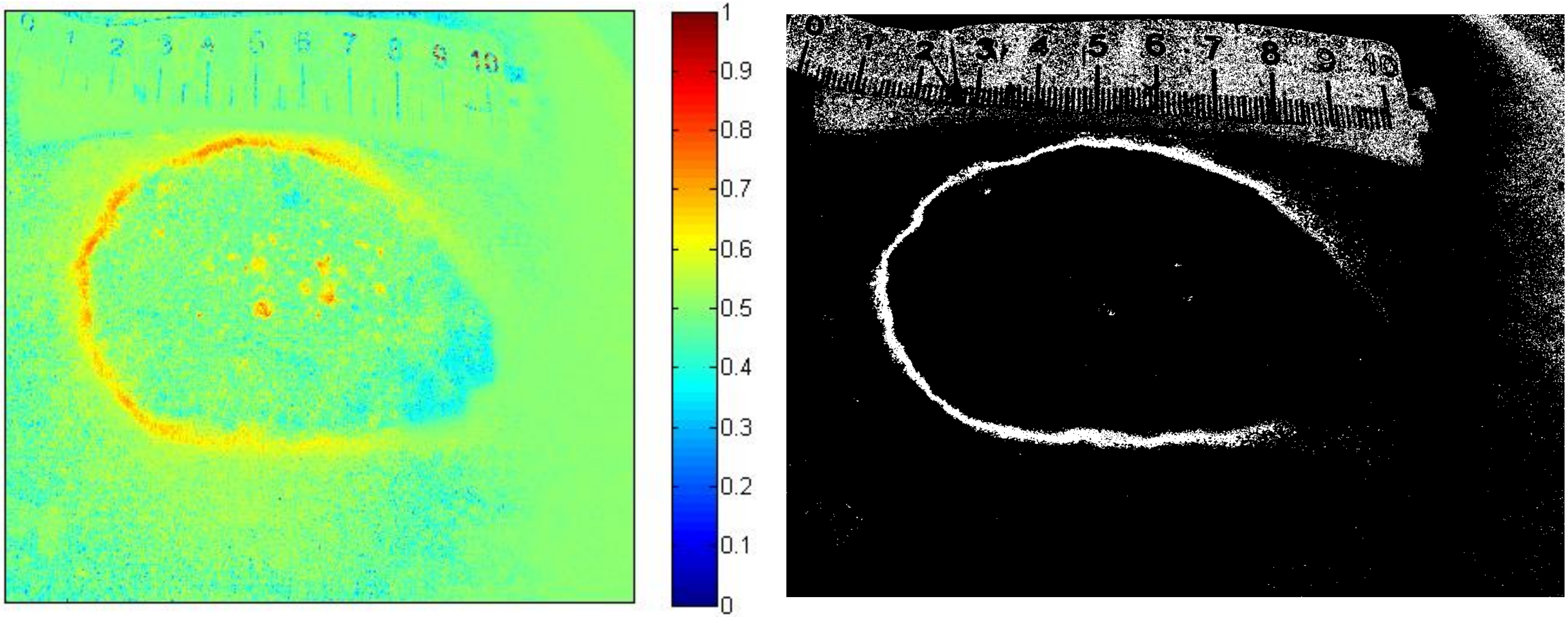
Background and ruler numbers. Left: Nonlinear ICA extracted component; Right: binarized version.

Nonlinear projection on orthonormal basis and ICA?



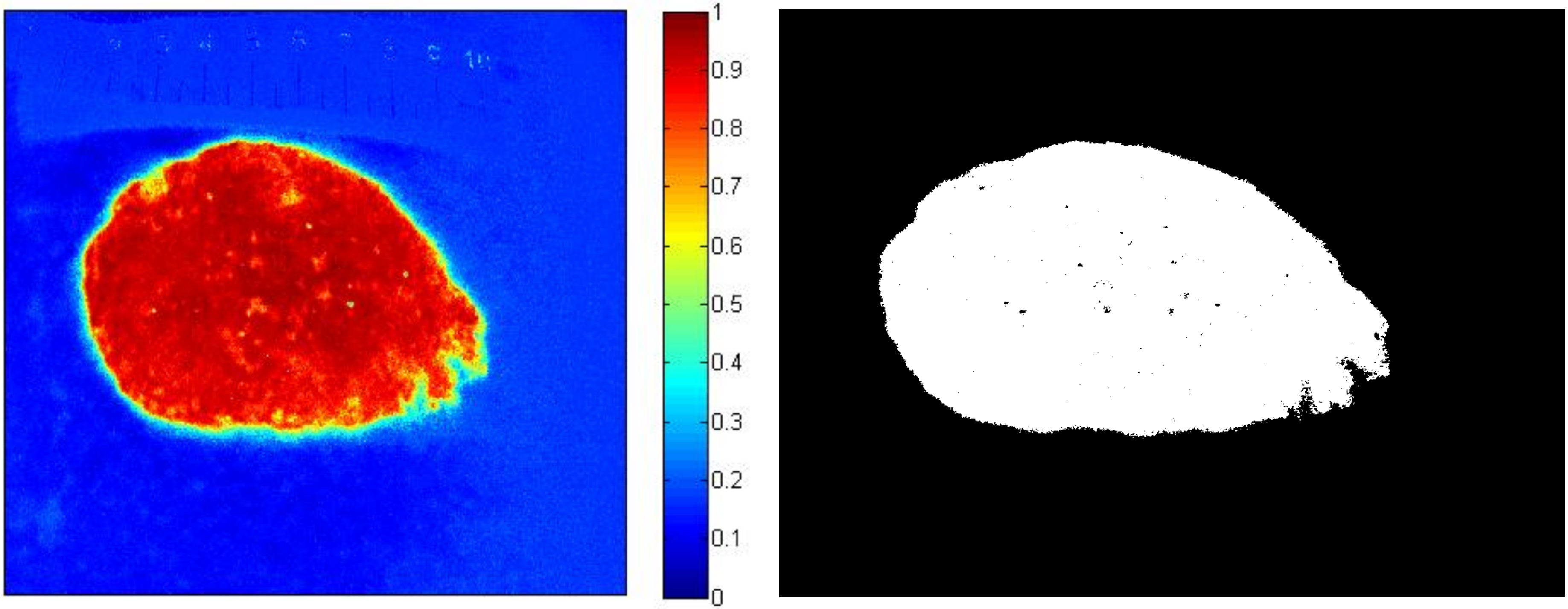
Background and ruler numbers. Left: Nonlinear ICA extracted component; Right: binarized version.

Nonlinear projection on orthonormal basis and ICA?



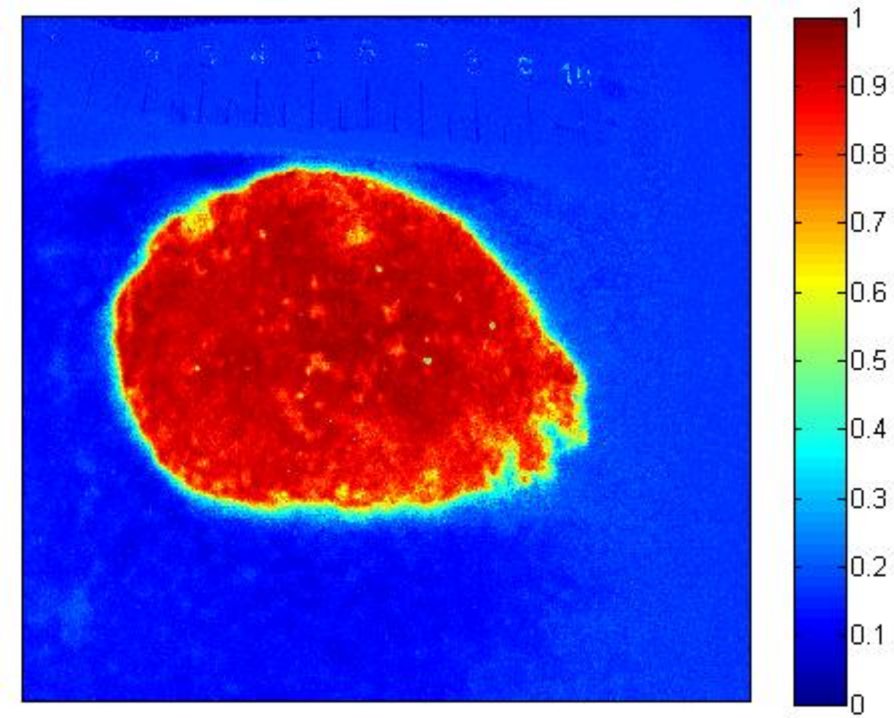
Tumor demarcation line and ruler body. Left: Nonlinear ICA extracted component; Right: binarized version.

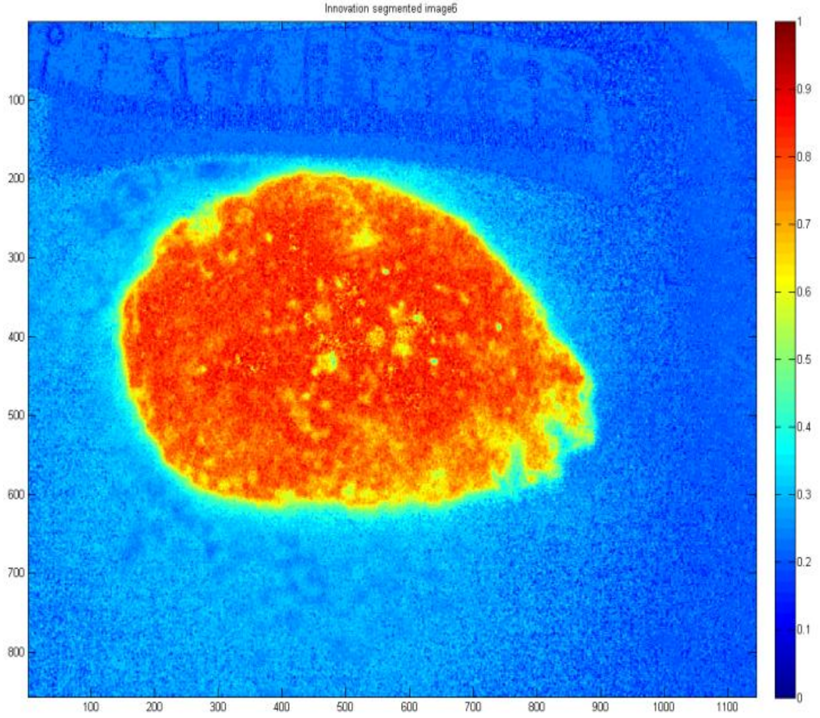
Nonlinear projection on orthonormal basis and ICA?



Basal cell carcinoma. Left: Nonlinear ICA extracted component. Right: binarized version.

Nonlinear projection on orthonormal basis and ICA?



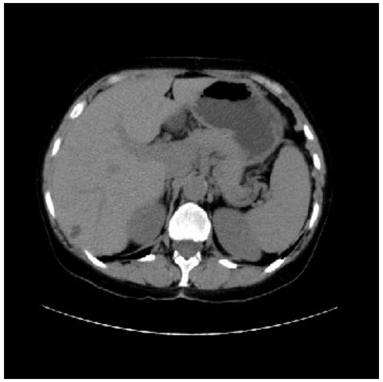


I. Kopriva**,** A. Peršin (2009) *Medical Image Analysis* , vol.**13**, pp.507-518.

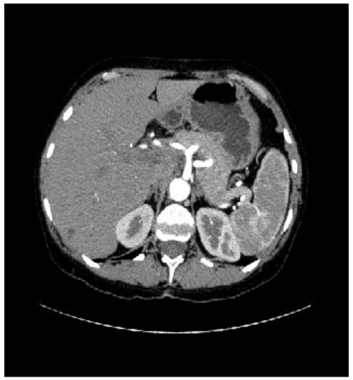
Nonlinear projection on orthonormal basis and ICA?

We have applied “AMUSE” ICA algorithm, [57], on RKHS-deccorelated version of the multi-phase CT of abdomen. (D=100, Gaussian kernel, $\sigma^2=10^3$.

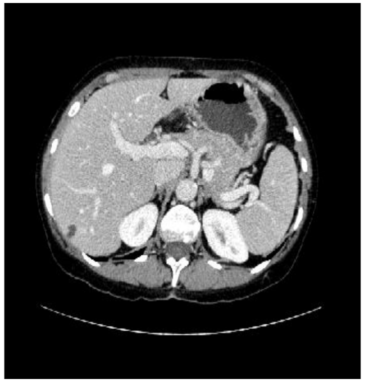
Slide 114 displayed in window [-100 200] Hounsfield unit.



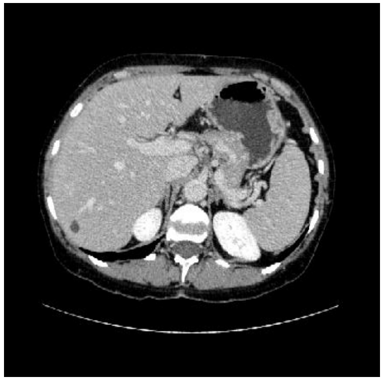
Non-contrast



Arterial



Venous 1

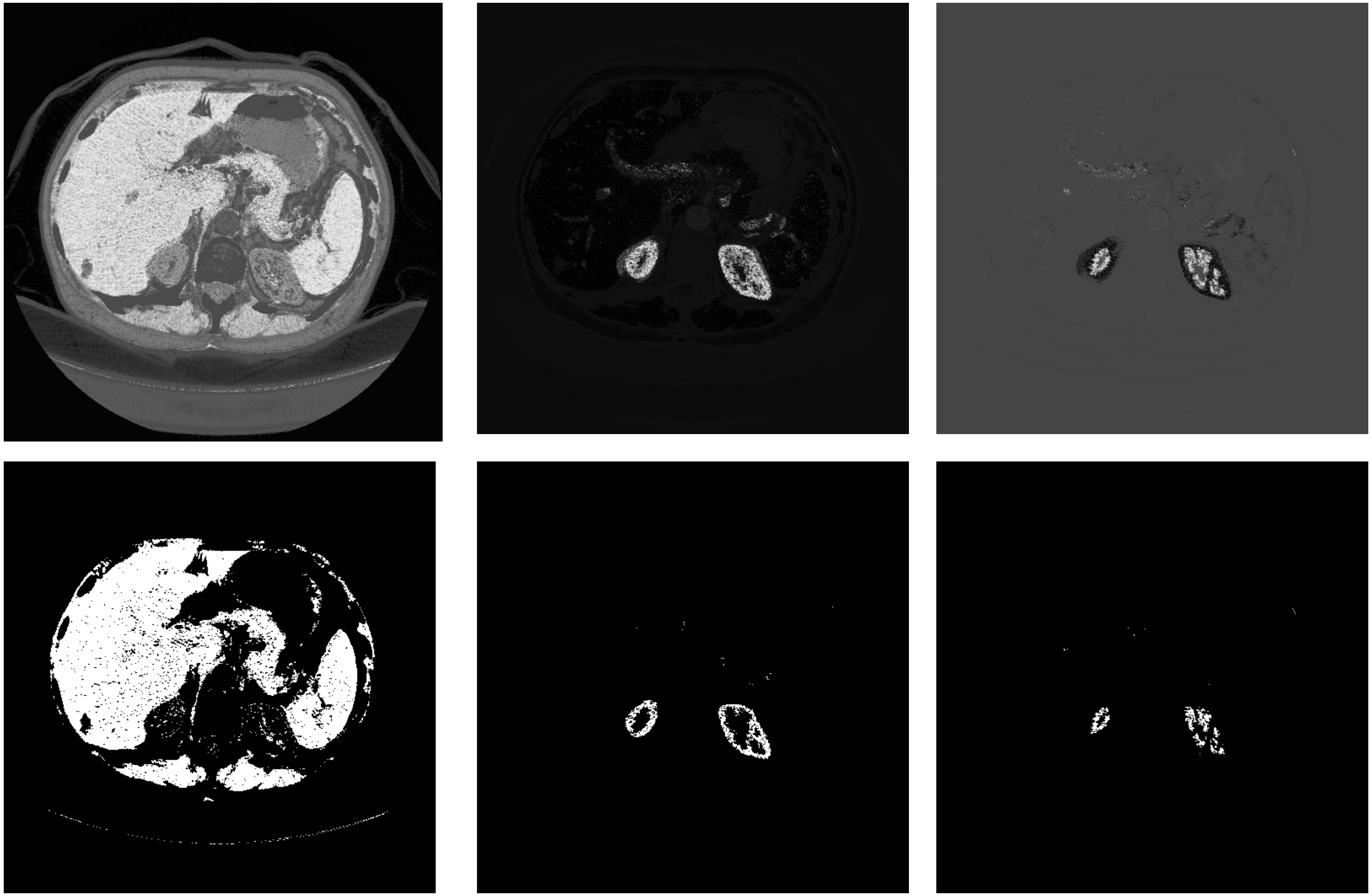


Venous 2

[57] L. Tong, R.W. Liu, V.C. Soon, and Y. F. Huang, “Indeterminacy and identifiability of blind identification,” *IEEE Trans. on Circuits and Systems,* 38:499-509, 1991.

Top row:components decomposed by nonlinear ICA algorithm.

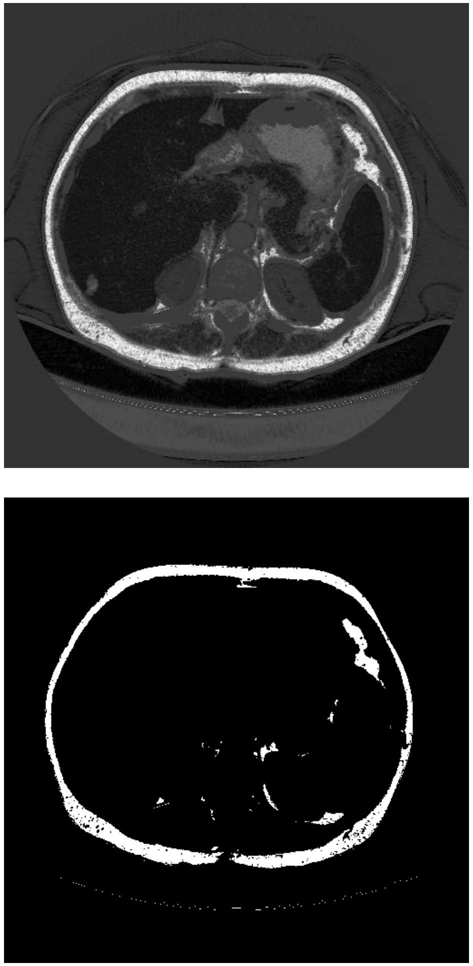
Bottom row: assignment according to maximal value criterion (occlusions).



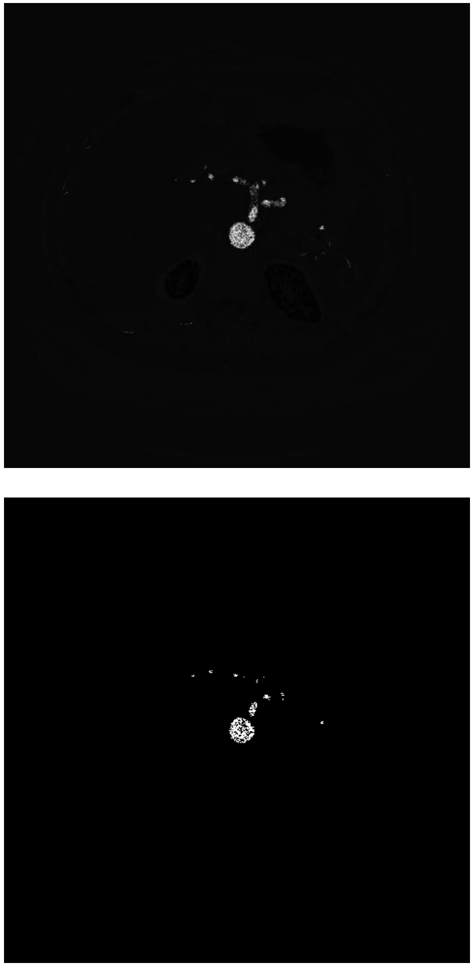
Liver

Kidneys – renal cortex

Kidneys – renal medula



skin



aorta

Nonlinear projection on orthonormal basis and ICA?

Nonlinear ICA can, **possibly**, be applied on unsupervised decomposition (segmentation) of multichannel medical images with **good spatial resolution and low-sensitivity** (contrast).

One imaging modaility of immediate relevance is multi-phase computed tomography (CT) imaging, where soft tissues (liver, kidneys, …) have low contrast.

In particular, nonlinear ICA can be used **to reduce number of phase-contrast** images!!!

Nonlinear ICA can also be used on **multimodal images** (e.g. CT and PET)**?**

THANK YOU !!!!!!!!