Nonnegative Matrix Factorization and Applications

Christine De Mol
(joint work with Michel Defrise and Loïc Lecharlier)

Université Libre de Bruxelles
Dept Math. and ECARES

100 Years of the Radon Transform
Linz, March 27-31, 2017

Linear Inverse Problem

- Solve
$$Kx = y$$
in discrete setting
- $x \in \mathbb{R}^p$ = vector of coefficients describing the unknown object
- $y \in \mathbb{R}^n$ = vector of (noisy) data
- K = linear operator ($n \times p$ matrix) modelling the link between the two

Regularization

Noisy data $\rightarrow$ solve approximately by minimizing contrast (discrepancy) function, e.g. $\|Kx - y\|_2^2$

Ill-conditioning $\rightarrow$ regularize by adding constraints/penalties on the unknown vector x e.g.

- on its squared L^2-norm $\|x\|_2^2 = \sum_i |x_i|^2$ (classical quadratic regularization)
- on its L^1-norm $\|x\|_1 = \sum_i |x_i|$ (sparsity-enforcing or "lasso" regularization, favoring the recovery of sparse solutions, i.e. the presence of many zero components in x)
- on a linear combination of both $\|x\|_1$ and $\|x\|_2^2$ norms ("elastic-net" regularization)
- plus here positivity constraints (hold true in many applications)

Positivity and multiplicative iterative algorithms

- Poisson noise $\rightarrow$ minimize (log-likelihood) cost function subject to $x \geq 0$ (assuming $K \geq 0$ and $y \geq 0$)

$$F(x) = KL(y, Kx) \equiv \sum_{i=1}^{n} \left[y_i \ln \left(\frac{y_i}{(Kx)_i} \right) - y_i + (Kx)_i \right]$$

(Kullback-Leibler – generalized – divergence)

- Richardson (1972) - Lucy (1974) (an astronomer's favorite) = EM(ML) in medical imaging
- Algorithm: $x^{(k+1)} = \frac{x^{(k)}}{K^T \mathbf{1}} \circ K^T \frac{y}{Kx^{(k)}} \qquad (k = 0, 1, \ldots)$
(using the Hadamard (entrywise) product $\circ$ and division;
1 is a vector of ones)
- Positivity automatically preserved if $x^{(0)} > 0$
- Unregularized $\rightarrow$ semi-convergence $\rightarrow$ usually early stopping
- Can be easily derived through separable surrogates

Surrogating

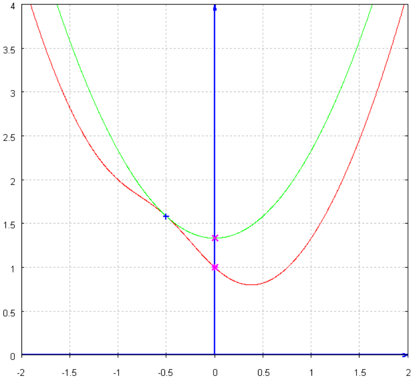


Figure : The function in red and his surrogate in green

Surrogating

- Surrogate cost function $G(x; a)$ for $F(x)$:

$$G(x; a) \geq F(x) \qquad \text{and} \qquad G(a; a) = F(a)$$

  for all x, a

- MM-algorithm (Majorization-Minimization):

$$x^{(k+1)} = \arg\min_{x} G(x; x^{(k)})$$

- Monotonic decrease of the cost function is then ensured:

$$F(x^{(k+1)}) \leq F(x^{(k)})$$

(Lange, Hunter and Yang 2000)

Surrogate for Kullback-Leibler

Cost function ($K \geq 0$ and $y \geq 0$)

$$F(x) = \sum_{i=1}^{n} \left[y_i \ln\left(\frac{y_i}{(Kx)_i}\right) - y_i + (Kx)_i \right]$$

Surrogate cost function (for $x \geq 0$)

$$\begin{aligned} G(x;a) &= \sum_{i=1}^{n} \Bigg[y_i \ln y_i - y_i + (Kx)_i + \\ &- \frac{y_i}{(Ka)_i} \sum_{j=1}^{p} K_{i,j} a_j \ln\left(\frac{x_j}{a_j}(Ka)_i\right) \Bigg] \end{aligned}$$

NB. This surrogate is separable, i.e. it can be written as a sum of terms, where each term depends only on a single unknown component x_j.

Positivity and multiplicative iterative algorithms

- Gaussian noise $\rightarrow$ minimize (log-likelihood) cost function subject to $x \geq 0$

$$F(x) = \frac{1}{2}\|Kx - y\|_2^2$$

assuming $K \geq 0$ and $y \geq 0$

- ISRA (Image Space Reconstruction Algorithm) (Daube-Witherspoon and Muehllehner 1986; De Pierro 1987)
- Iterative updates

$$x^{(k+1)} = x^{(k)} \circ \frac{K^T y}{K^T K x^{(k)}}$$

- Positivity automatically preserved if $x^{(0)} > 0$
- Unregularized $\rightarrow$ semi-convergence $\rightarrow$ usually early stopping
- Easily derived through separable surrogates

Surrogate for Least Squares

Cost function ($K \geq 0$ and $y \geq 0$)

$$F(x) = \frac{1}{2} \|Kx - y\|_2^2$$

Surrogate cost function (for $x \geq 0$)

$$G(x; a) = \frac{1}{2} \sum_{i=1}^{n} \frac{1}{(Ka)_i} \sum_{j=1}^{p} K_{i,j} a_j \left[y_i - \frac{x_j}{a_j} (Ka)_i \right]^2$$

NB. This surrogate is separable, i.e. it can be written as a sum of terms, where each term depends only on a single unknown component x_j

Blind Inverse Imaging

- In many instances, the operator is unknown ("blind") or only partially known ("myopic" imaging/deconvolution)
- The resulting functional is convex w.r.t. x or K separately but is not jointly convex $\rightarrow$ possibility of local minima
- Usual strategy: alternate minimization on x (with K fixed) and K (with x fixed)
- The problem can be easily generalized to include multiple inputs/unknowns (x becomes a $p \times m$ matrix X) and multiple outputs/measurements (y becomes a $n \times m$ matrix Y) e.g. for Hyperspectral Imaging

$$\longrightarrow \qquad \text{solve} \quad KX = Y$$

Special case: Blind Deconvolution

- When the imaging operator K is translation-invariant, the problem is also referred to as "Blind Deconvolution"
- Alternating minimization approaches using (regularized) least-squares (Ayers and Dainty 1988; You and Kaveh 1996; Chan and Wong 1998, 2000) or Richardson-Lucy (Fish, Brinicombe, Pike and Walker 1996)
- Bayesian approaches are also available
- An interesting non-iterative and nonlinear inversion method has been proposed by Justen and Ramlau (2006) with a uniqueness result. Unfortunately, their solution has been shown to be unrealistic from a physical point of view by Carasso (2009)

Blind Inverse Imaging, Positivity and NMF

- Blind imaging is difficult $\rightarrow$ use as much a priori information and constraints as you can
- In particular, positivity constraints have proved very powerful when available, e.g. in incoherent imaging as for astronomical images
- The special case where all elements of K, X (and Y) are nonnegative ($K \geq 0$, $X \geq 0$) is also referred to as "Nonnegative Matrix Factorization" (NMF)
- There is a lot of recent activity on NMF, as an alternative to SVD/PCA for dimension reduction
- Alternating (ISRA or RL) multiplicative algorithms have been popularized by Lee and Seung (1999, 2000)
  See also Donoho and Stodden (2004)

Our goal

- Develop a general and versatile framework for
- blind deconvolution/inverse imaging with positivity,
- equivalently for Nonnegative Matrix Factorization,
- with convergence results to control not only the decay of the cost function but also the convergence of the iterates,
- with algorithms simple to implement
- and reasonably fast...

Applications

We will consider

- Blind deconvolution with positivity
  (from single or multiple images)
- Hyperspectral Imaging
- Dynamic Positron Emission Tomography (PET)

  NB. There are many other applications of NMF

Regularized least-squares (Gaussian noise)

- Minimize the cost function, for K, X nonnegative (assuming Y nonnegative too),

$$F(K,X) = \frac{1}{2}\|Y - KX\|_F^2 + \frac{\mu}{2}\|K\|_F^2 + \lambda\|X\|_1 + \frac{\nu}{2}\|X\|_F^2$$

  where $\|\cdot\|_F^2$ denotes the Frobenius norm $\|K\|_F^2 = \sum_{i,j} K_{i,j}^2$

- The minimization can be done column by column on X and line by line on K

Regularized least-squares (Gaussian noise)

- Alternating multiplicative algorithm
  ($\mathbf{1}_{p\times m}$ is a $p\times m$ matrix of ones)

$$
\begin{aligned}
K^{(k+1)} &= K^{(k)} \circ \frac{Y(X^{(k)})^T}{K^{(k)}X^{(k)}(X^{(k)})^T + \mu K^{(k)}} \\
X^{(k+1)} &= X^{(k)} \circ \frac{(K^{(k+1)})^T Y}{(K^{(k+1)})^T K^{(k+1)} X^{(k)} + \nu X^{(k)} + \lambda \mathbf{1}_{p\times m}}
\end{aligned}
$$

- to be initialized with arbitrary but strictly positive $K^{(0)}$ and $X^{(0)}$
- Can be derived through surrogates $\rightarrow$ provides a monotonic decrease of the cost function at each iteration
- Special cases:
  • a blind algorithm proposed by Hoyer (2002, 2004) for $\mu = 0, \nu = 0$
  • ISRA for K fixed and $\lambda = \mu = \nu = 0$

Regularized least-squares (Gaussian noise)

- Assume μ and either ν or λ are strictly positive and Y has at least one strictly positive element in each row and each column
- Monotonicity is strict iff $(K^{(k+1)}, X^{(k+1)}) \neq (K^{(k)}, X^{(k)})$
- The sequence $F(K^{(k)}, X^{(k)})$ converges
- Asymptotic regularity holds: $\forall i, j$ $\lim_{k \to +\infty} \left(K_{i,j}^{(k+1)} - K_{i,j}^{(k)} \right) = 0$; $\lim_{k \to +\infty} \left(X_{i,j}^{(k+1)} - X_{i,j}^{(k)} \right) = 0$
- $\Rightarrow$ the set of accumulation points of the sequence of iterates $(K^{(k)}, X^{(k)})$ is compact and connected
- If this set is finite, the iterates $(K^{(k)}, X^{(k)})$ converge to a stationary point (K^*, X^*) (satisfying the first-order KKT conditions)

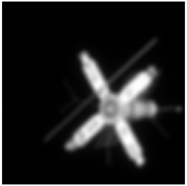
Some recent related (methodological) work

(with convergence results, possibly positivity constraints)

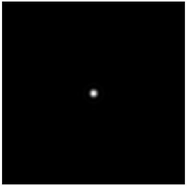
- Algorithm based on the SGP algorithm by Bonettini, Zanella, Zanni (2009) and inexact block coordinate descent (Bonettini 2011): Prato, La Camera, Bonettini, Bertero (2013)
  For a space-variant PSF, see also Ben Hadj, Blanc-Féraud and Aubert (2012)
- Proximal Alternating Minimization and Projection Methods for Nonconvex Problems
  (Attouch, Bolte, Redont, Soubeyran 2010; Bolte, Combettes and Pesquet 2010; Bolte, Sabach and Teboulle 2013)
- Underapproximations for Sparse Nonnegative Matrix Factorization
  (Gillis and Glineur 2010)

Application of NMF to Blind Deconvolution

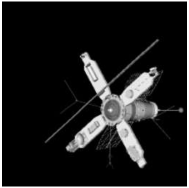
- X : 256×256 positive image
- K : Convolution with the Airy function (circular low-pass filter)



=



$*$



$$Y = K \quad X$$

Application (Gaussian noise): no noise added

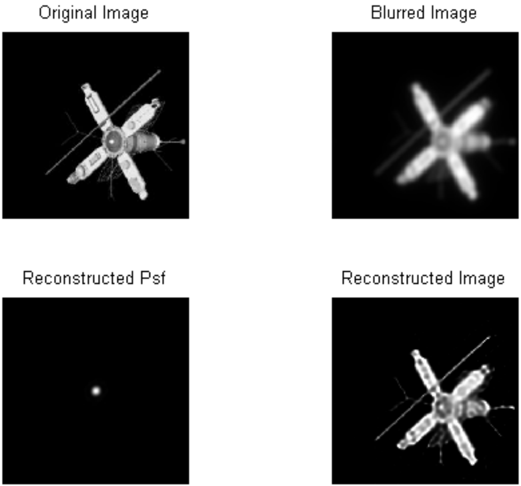


Figure : $K^{(0)}$ Unif, $X^{(0)} =$ Blurred Image; $\mu = 0$, $\lambda = 0$, $\nu = 0$, 1000 it

Application (Gaussian noise): 2.5% noise added

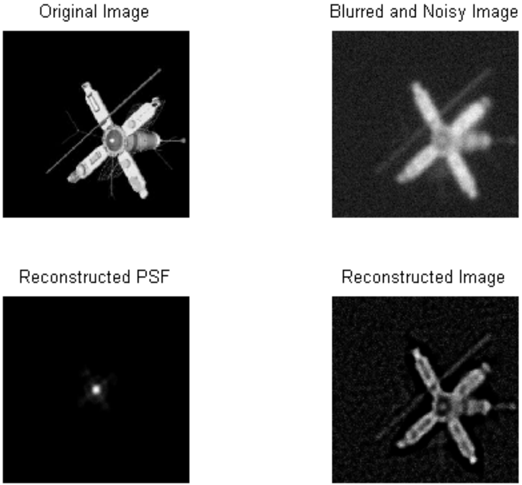


Figure : $K^{(0)}$ Gaussian, $X^{(0)} =$ Noisy Image; $\mu = 2.25 \cdot 10^8$, $\lambda = 0.03$, $\nu = 0.008$; 200 it

Regularized Kullback-Leibler (Poisson noise)

• Minimize the cost function, for K, X nonnegative (assuming Y nonnegative too),

$$F(K,X) = KL(Y,KX) + \frac{\mu}{2}\,\|K\|_F^2 + \lambda\,\|X\|_1 + \frac{\nu}{2}\,\|X\|_F^2$$

with

$$KL(Y,KX) = \sum_{i=1}^{n}\sum_{j=1}^{m}\left[(Y)_{i,j}\ln\left(\frac{(Y)_{i,j}}{(KX)_{i,j}}\right) - (Y)_{i,j} + (KX)_{i,j}\right]$$

Regularized Kullback-Leibler (Poisson noise)

- Alternating multiplicative algorithm

$$K^{(k+1)} = \frac{2A^{(k)}}{B^{(k)} + \sqrt{B^{(k)} \circ B^{(k)} + 4\mu A^{(k)}}}$$

where

$$A^{(k)} = K^{(k)} \circ \frac{Y}{K^{(k)}X^{(k)}} (X^{(k)})^T$$

$$B^{(k)} = \mathbf{1}_{n\times m} (X^{(k)})^T$$

($\mathbf{1}_{n\times m}$ is a $n \times m$ matrix of ones)

Regularized Kullback-Leibler (Poisson noise)

$$X^{(k+1)} = \frac{2C^{(k+1)}}{D^{(k+1)} + \sqrt{D^{(k+1)} \circ D^{(k+1)} + 4\nu C^{(k+1)}}}$$

where

$$C^{(k+1)} = X^{(k)} \circ (K^{(k+1)})^T \frac{Y}{K^{(k+1)} X^{(k)}}$$

$$D^{(k+1)} = \lambda \mathbf{1}_{p\times m} + (K^{(k+1)})^T \mathbf{1}_{n\times m}$$

to be initialized with arbitrary but strictly positive $K^{(0)}$ and $X^{(0)}$

Regularized Kullback-Leibler (Poisson noise)

- Can be derived through surrogates $\rightarrow$ provides a monotonic decrease of the cost function at each iteration
- Special case for $\lambda = \mu = \nu = 0$: the blind algorithm proposed by Lee and Seung (1999) which reduces to the EM/Richardson-Lucy algorithm for K fixed
- Properties as above for the least-squares case

Normalization constraint

- At each iteration, one can enforce a normalization constraint on the PSF (line of K), imposing that its values sum to one
- To do this a Lagrange multiplier is introduced and its value is determined by means of a few Newton-Raphson iterations
- The convergence results can be adapted to cope with this case

Application : 1% (equiv. rmse) Poisson Noise; PSF normalized

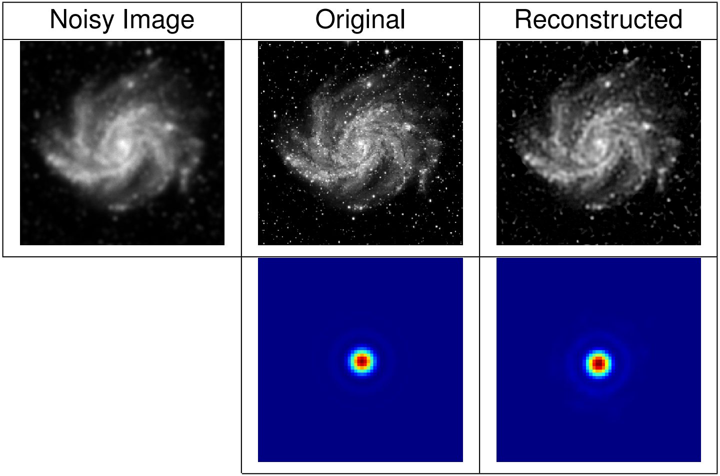


Figure : $K^{(0)}$ = Unif, $X^{(0)}$ = Noisy Image, $\mu = 10^9$, $\lambda = 10^{-7}$, $\nu = 6 \cdot 10^{-8}$, 2000 it in 12m37s

Extension to TV regularization

- Total Variation: use discrete differentiable approximation

$$\|X\|_{TV} = \sum_{i,j} \sqrt{\varepsilon^2 + (X_{i+1,j} - X_{i,j})^2 + (X_{i,j+1} - X_{i,j})^2}$$

  for 2D images

- Use penalty $\lambda\|X\|_{TV}$ instead of $\lambda\|X\|_1$
- Use separable surrogate proposed by Defrise, Vanhove and Liu (2011) to derive explicit update rules both for Gaussian and Poisson noise

Application KL-TV: 1% (equiv. rmse) Poisson Noise; PSF normalized

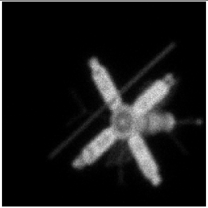
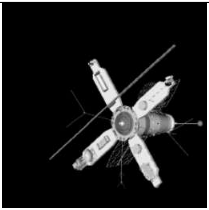
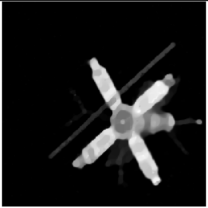
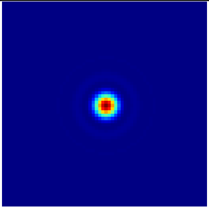
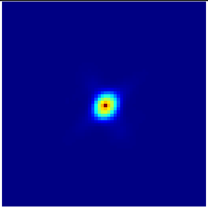
Noisy Image	Original	Reconstructed
		
		

Figure : $K^{(0)}$ = Unif, $X^{(0)} =$ Noisy Image, $\mu = 1.5 \cdot 10^6$, $\lambda = 0.0485$, $\varepsilon = 6 \cdot 10^{-7}$, 200 it in 1m46s

Application KL-TV: 2.5% (equiv. rmse) Poisson Noise; normalized PSF

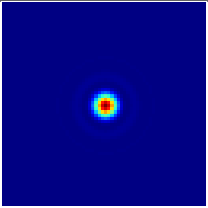
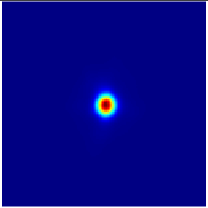
Noisy Image	Original	Reconstructed
		
		

Figure : $K^{(0)}$ = Unif, $X^{(0)}$ = Noisy Image, $\mu = 10^7$, $\lambda = 0.03$, $\varepsilon = \sqrt{10}$, 2000 it in 54m30s

Application to AO astronomical images

- Test images used by Prato, La Camera, Bonettini and Bertero (2013) for post-adaptive-optics astronomical imaging (PSF with Strehl Ratio 0.40), with a restoration method based on the SGP algorithm by Bonettini, Zanella, Zanni (2009) and inexact block coordinate descent (Bonettini 2011)
- Necessity to incorporate a background term in the cost function ($KX \to KX + B$) and in the KL restoration algorithm
- Quality of restorations comparable to those of Prato, La Camera, Bonettini and Bertero (2013)

Planetary Nebula NGC 7027

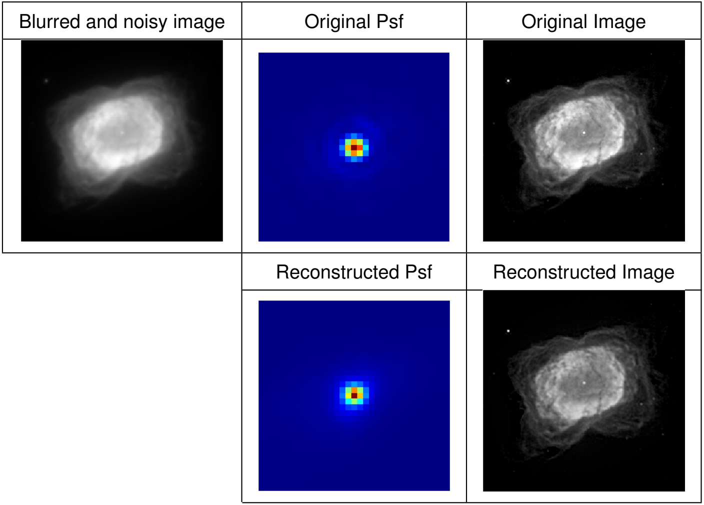


Figure :

Reconstruction with $\mu = 10^6$, $\lambda = 6 \cdot 10^{-8}$, $\nu = 10^{-9}$, uniform $K^{(0)}$, $X^{(0)} = Y$, 5500 it in 17min40s, $RMSE_K = 22.61\%$, $RMSE_X = 6.66\%$

Galaxy NGC 6946

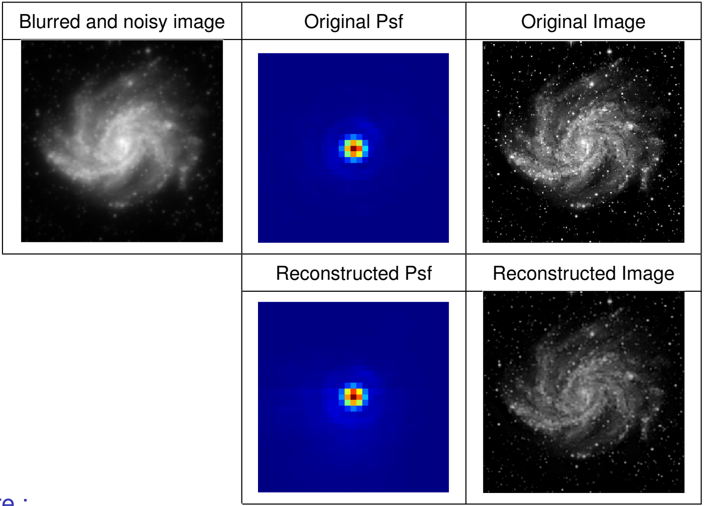


Figure : Reconstruction with $\mu = 2.75 \cdot 10^8$, $\lambda = 5 \cdot 10^{-8}$, $\nu = 1.668 \cdot 10^{-8}$, uniform $K^{(0)}$, $X^{(0)} = Y$, 5000 it in 15min45s, $RMSE_K = 14.24\%$, $RMSE_X = 22.83\%$

Crab Nebula NGC 1952

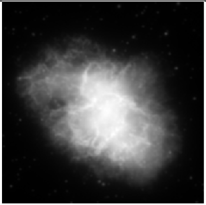
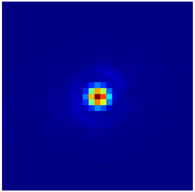
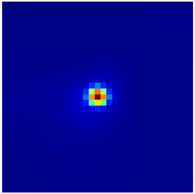
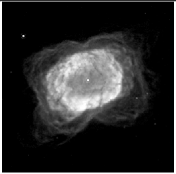
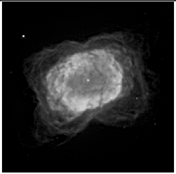
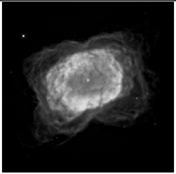
Blurred and noisy image	Original Psf	Original Image
		
	Reconstructed Psf	Reconstructed Image
		

Figure : Reconstruction with $\mu = 7.4 \cdot 10^7$, $\lambda = 2.5 \cdot 10^{-8}$, $\nu = 1.5 \cdot 10^{-8}$, uniform $K^{(0)}$, $X^{(0)} = Y$, 4000 it in 11min48s, $RMSE_K = 15.72\%$, $RMSE_X = 15.96\%$

Reconstruction from multiple images (same PSF)

	Original	Single Image	Multiple (3) Images
Galaxy			
Planetary Nebula			
Crab Nebula			

Reconstruction from multiple images (same PSF)

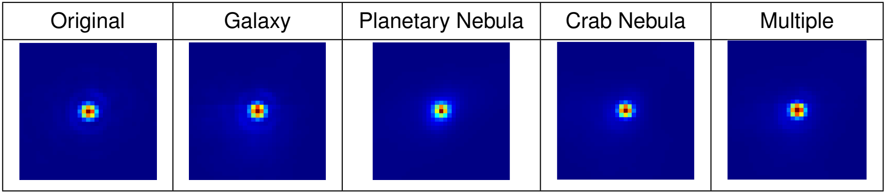


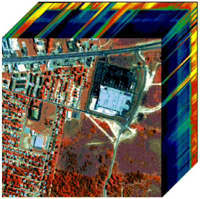
Figure : Reconstructed PSF

Object	μ	λ	ν	**Nb It.**	**Time**	$RMSE_K$	$RMSE_X$	$RMSE_M$
Plan.	10^6	10^{-8}	10^{-9}	5500	17m40s	0.23	0.06	0.05
Gal.	$2.75 \cdot 10^8$	$5 \cdot 10^{-8}$	$1.67 \cdot 10^{-8}$	5000	15m45s	0.14	0.23	0.19
Crab	$7.4 \cdot 10^7$	$2.5 \cdot 10^{-8}$	$1.5 \cdot 10^{-8}$	4000	11m48s	0.15	0.16	0.14
All 3	$2.4 \cdot 10^8$	$6.75 \cdot 10^{-8}$	$4.36 \cdot 10^{-9}$	4000	21m27s	0.12		

Table : Values used for the parameters and RMSE

Application of NMF to Hyperspectral Imaging

Example: Urban HYDICE HyperCube: $307 \times 307 \times 162$
containing the images of an urban zone recorded for 162 different wavelength/frequencies



- Factorize the $Y : 307^2 \times 162$ data matrix as $Y = KX$ where K is a $307^2 \times p$ (relative) abundances matrix of some basis elements to be determined and X is a $p \times 162$ matrix containing the spectra of those basis elements
- Penalized Kullback-Leibler divergence used as cost function
- The sum of the relative abundances is normalized to one

Hyperspectral Imaging

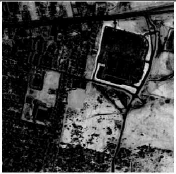
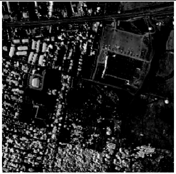
Dirt	Grass	Trees
		
Roofs	Roads	Metals
		

Figure : Abundances with $p = 6$, $\mu = 10^{-10}$, $\lambda = 0$, $\nu = 1.1$, random $K^{(0)}$ and $X^{(0)}$, 1000 it in 1h19min12s

Hyperspectral Imaging

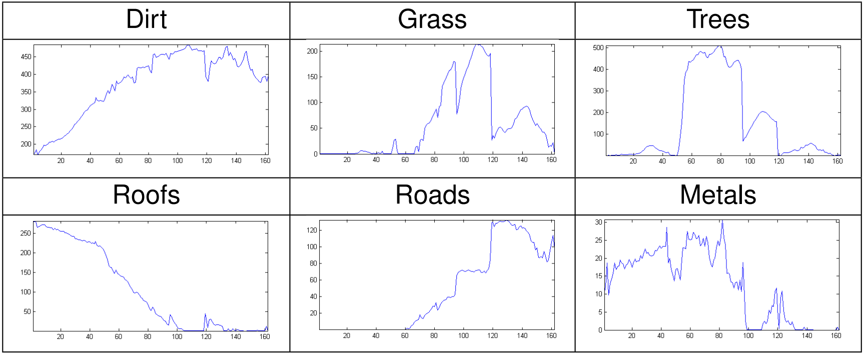
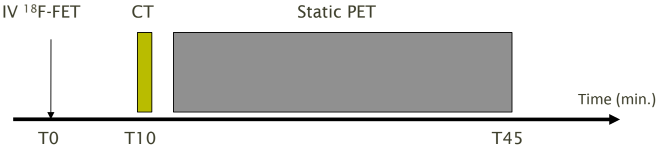


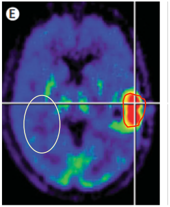
Figure : Spectra

Application of NMF to Dynamic PET

(work in progress with Michel Defrise)

^{18}F-Fluoro-ethyl-tyrosine (^{18}F-FET) =
artificial amino-acid for PET, crossing the Blood Brain Barrier

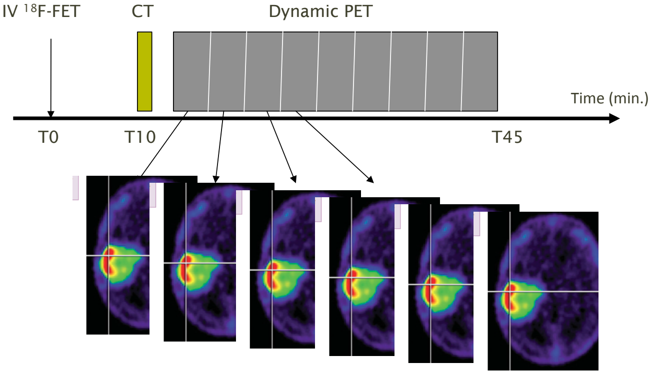




Pixels with FET-uptake ≥ 1.6 of normal brain are considered "tumor associated"

Sum of all tumor associated pixels:
Metabolic Active Volume (MAV in ml)

Dynamic PET



(slides : courtesy by Dr Hendrik Everaert)
NB. Large literature on NNMF and in particular for dynamic PET, see e.g.
Tichy, Smidl, 2014 International Conference on BioMedical Engineering and Informatics,
Lee et al (Seoul) for myocardial H20 PET, …

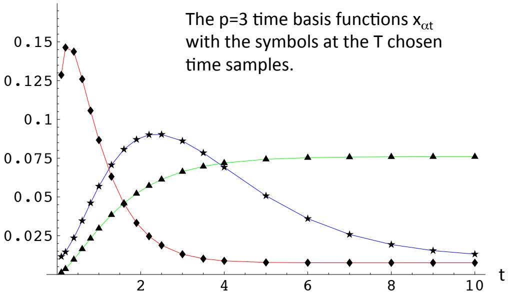
Dynamic PET

PET scan data inversion is done time by time and then the 4D dataset (nb voxels $\times$ nb of time frames) is factored through NMF, with cost function:

- Least-squares discrepancy (Gaussian noise)
- Smoothness of the temporal activity curves (TAC)/factors: quadratic penalty on a high-pass filtered version/differences
- Normalization of the temporal activity curves (TAC)/factors: sum to one over time
- Sparsity-enforcing penalty (L^1-norm) on the coefficients (few factors active in each voxel)
- Spatial correlations: quadratic penalty on differences of coefficients corresponding to neighbouring voxels

Dynamic PET

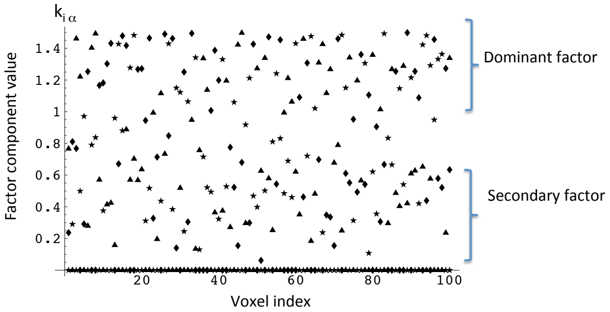
A toy problem with 100 “voxels”, 20 time frames, p=3 factors
No spatial connection between voxels



The p=3 time basis functions $x_{\alpha t}$ with the symbols at the T chosen time samples.

Dynamic PET

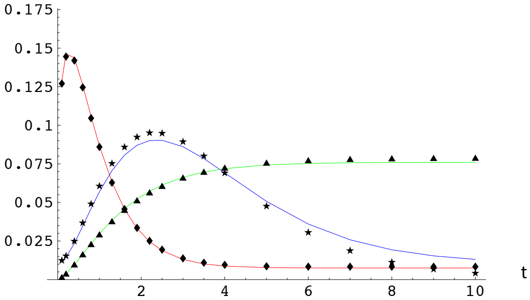
The “true” coefficients $k_{i\,\alpha}$ are selected as follows: each “voxel” has non-zero Coefficients for two factors randomly selected from the three factors. Given two random numbers r1, r2 uniformly distributed on (0,1) the dominant coefficient is (0.5+max(r1,r2)) and the secondary factor modeling a weaker contamination is min(r1,r2). No spatial correlation is modeled.



Dynamic PET

Toy problem: noise free simulation. $\mu = \nu = 0,\ N_{iter} = 10000.\ x_{\alpha t}^{(0)} = 1,\ \ k_{it}^{(0)} = 0.1 + rand(0,1)$

The estimated factors (symbols) and the exact ones (full line).



The dominant factor $\alpha_i^* = \underset{\alpha}{\text{argmax}}\ k_{i\alpha}$ is identified correctly for all voxels.

The mean and standard deviation of the V=100 errors on the coefficients are

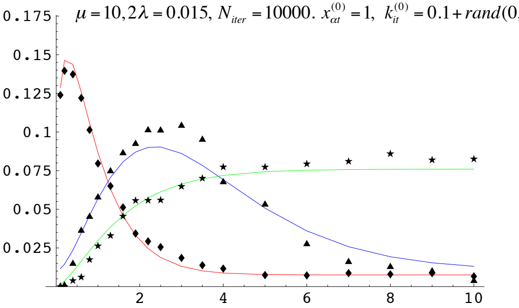
Red factor: 0.01 (0.0069)
Blue factor: -0.05 (0.074)
Green factor: 0.041 (0.074)

Relative RMSE of the reconstructed TAC (all time bins and all voxels): 2.2 10^{-6}
Maximum TAC error/Maximum TAC: 0.00002.

Dynamic PET

Toy problem: noisy simulation.

The estimated basis factors (symbols) and the exact ones (full line).



Despite noise and the important factor mixing, the dominant factor is identified correctly for 99 of the 100 voxels (for this noise realization).

The mean and standard deviation of the V=100 errors on the coefficients are

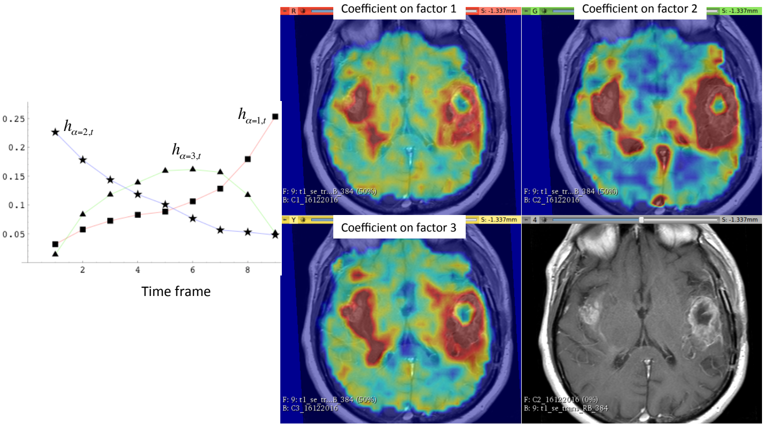
Red factor: -0.007 (0.10)

Blue factor: -0.14 (0.21)

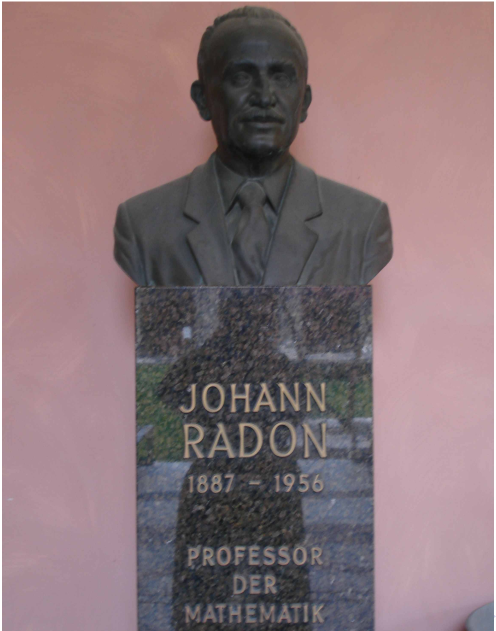
Green factor: -0.15 (0.17)

Dynamic PET: clinical data

Voxels = 24037, T = 9 time frames, we select A = 3 factors.
NNMF, 10000 iterations, (μ=ν=0). Initialization: all factors are constant, coefficients are random.



Happy Anniversary to the Radon Transform



A selfie from the pre-selfie era...

... and an invitation to blind deconvolution!