Regularized Multiplicative Algorithms for Nonnegative Matrix Factorization

Christine De Mol
(joint work with Loïc Lecharlier)

Université Libre de Bruxelles
Dept Math. and ECARES

MAHI 2013 Workshop
"Methodological Aspects of Hyperspectral Imaging"
Nice, October 14, 2013

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

Ill-conditioning $\longrightarrow$ 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 of ($\|x\|_1 = \sum_i |x_i|$) (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, favoring the recovery of sparse groups of possibly correlated components)

Positivity and multiplicative iterative algorithms

- Poisson noise $\longrightarrow$ 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: $x^{(k+1)} = \frac{x^{(k)}}{K^T \mathbf{1}} \circ K^T \frac{y}{Kx^{(k)}} \qquad (k = 0, 1, \dots)$
  (using the Hadamard (entrywise) product $\circ$ and division;
  1 is a vector of ones)
- Positivity automatically preserved if $x^{(0)} > 0$
- Unregularized $\longrightarrow$ semi-convergence $\longrightarrow$ usually early stopping
- Can be easily derived through separable surrogates

Surrogating

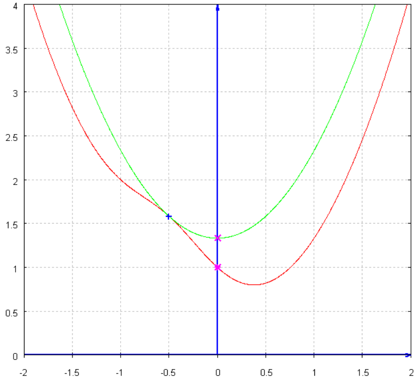


Figure: The function in red and his surrogate in green

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

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

Special case: Blind Deconvolution

- When the imaging operator K in 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 $\longrightarrow$ 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 proofs 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...

Work in progress!

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 (O is a matrix of ones)

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

- 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)

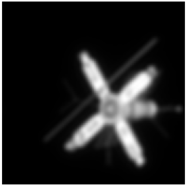
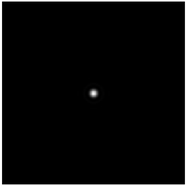
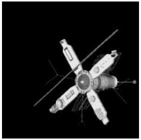
- Assume μ and either ν or λ strictly positive
- Monotonicity is strict iff $(K^{(k+1)}, X^{(k+1)}) \neq (K^{(k)}, X^{(k)})$
- The iterates $(K^{(k)}, X^{(k)})$ converge to a stationary point (K^*, X^*) (satisfying the first-order KKT conditions)
- If (K^*, X^*) is a stationary point then

$$\mu \|K^*\|_F^2 = \lambda \|X^*\|_1 + \nu \|X^*\|_F^2$$

- The ambiguity due to rescaling of (K^*, X^*) is frozen by the penalty as well as the ambiguity due to rotation (provided $\lambda \neq 0$)
- The algorithm can be accelerated using an Armijo rule along the "projection arc"

Application (Gaussian noise)

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

 =  $*$ 

$$Y = K \qquad X$$

Application (Gaussian noise): no noise added

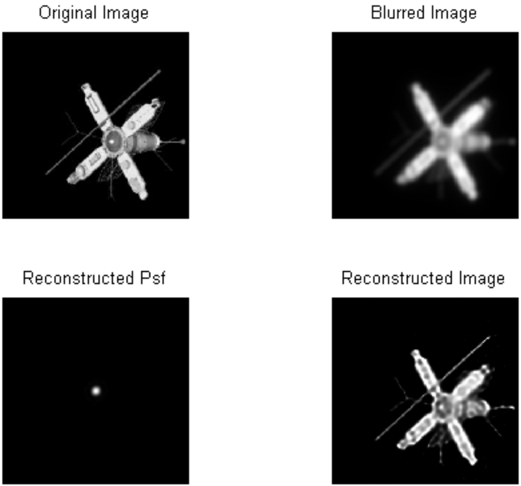
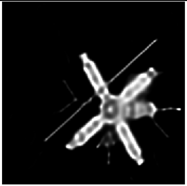
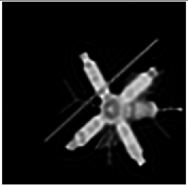
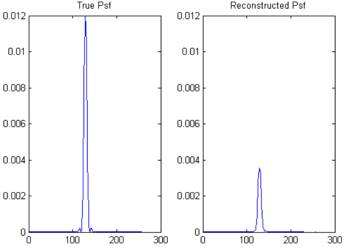
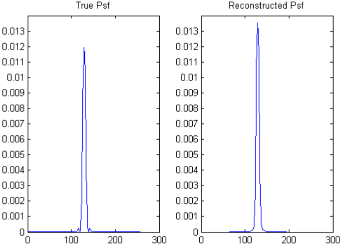


Figure: $K^{(0)}$ Unif, $X^{(0)} =$ Blurred Image; $\mu = 0$, $\lambda = 0$, $\nu = 0$, 1000 it

Application (Gaussian noise): no noise added

$K^{(0)}$ Uniform	$K^{(0)}$ Gaussian
	
	

Application (Gaussian noise): 2.5% noise added

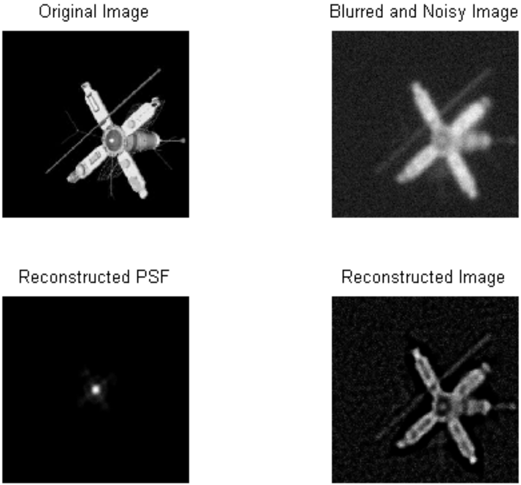


Figure: $K^{(0)}$ Gaussian, $X^{(0)}$ = Noisy Image; $\mu = 2.25 \cdot 10^8$, $\lambda = 0.03$, $\nu = 0.008$; 200 it

Application (Gaussian noise): 2.5% noise added

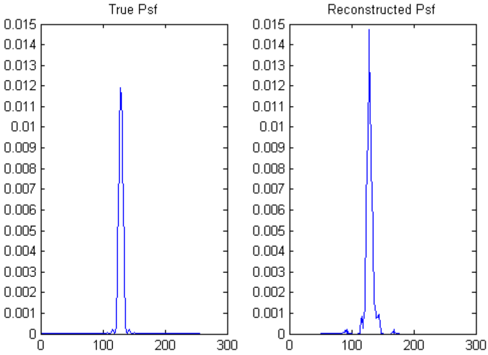
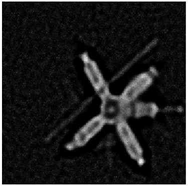
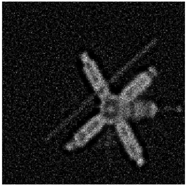
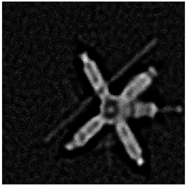


Figure: Point Spread Function

Application (Gaussian noise): 2.5% noise added

$\lambda = 0.03$, $\nu = 0.008$	$\lambda = 0.03$, $\nu = 0$	$\lambda = 0$, $\nu = 0.008$
		

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)

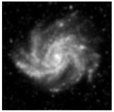
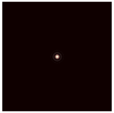
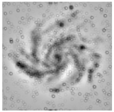
- Can be derived through surrogates $\rightarrow$ provides a monotonic decrease of the cost function at each iteration
- Special case for $\lambda = \mu = \nu = 0$: the blind algorithm proposed by Lee and Seung (1999) which reduces to the EM/Richardson-Lucy algorithm for K fixed
- Properties as above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

Normalization constraint

- At each iteration, one can enforce a normalization constraint on the PSF, 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 proof can be adapted to cope with this case

Application (Poisson noise)

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

 =  $*$  $+$ 

$$Y = K \quad X + E$$

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

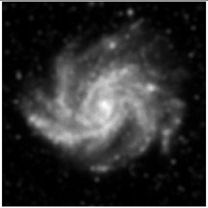
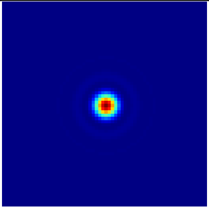
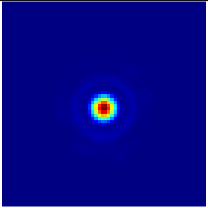
Noisy Image	Original	Reconstructed
		
		

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

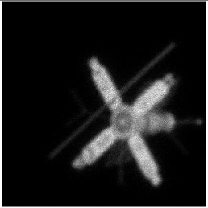
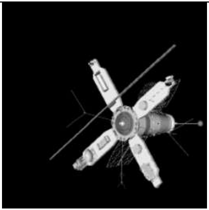
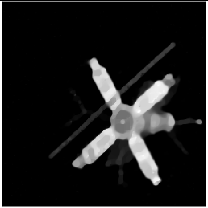
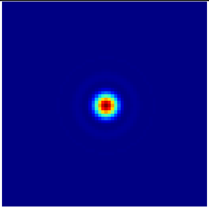
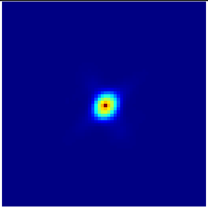
Noisy Image	Original	Reconstructed
		
		

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

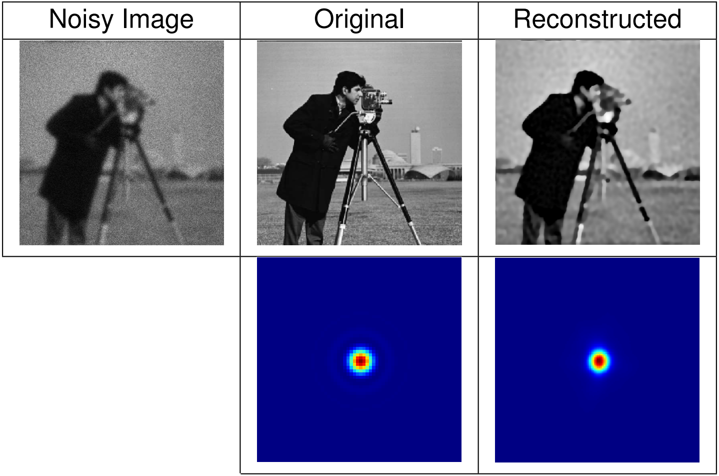
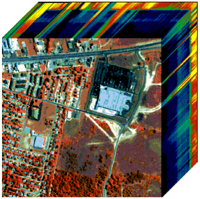


Figure: $K^{(0)}$ = Unif, $X^{(0)} =$ Noisy Image, $\mu = 10^7$, $\lambda = 0.03$, $\varepsilon = \sqrt{10}$, 2000 it in 54m30s

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

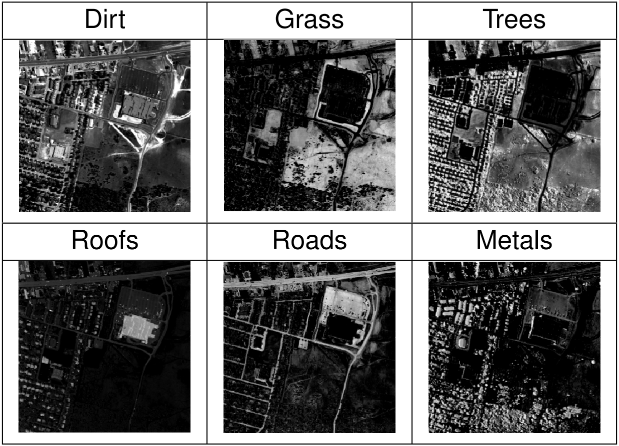


Figure: Abundances with $p = 6$, $\mu = 10^{-10}$, $\lambda = 0$, $\nu = 1.1$, random $K^{(0)}$ and $X^{(0)}$, 1000 it in 1h19min12s

Hyperspectral Imaging

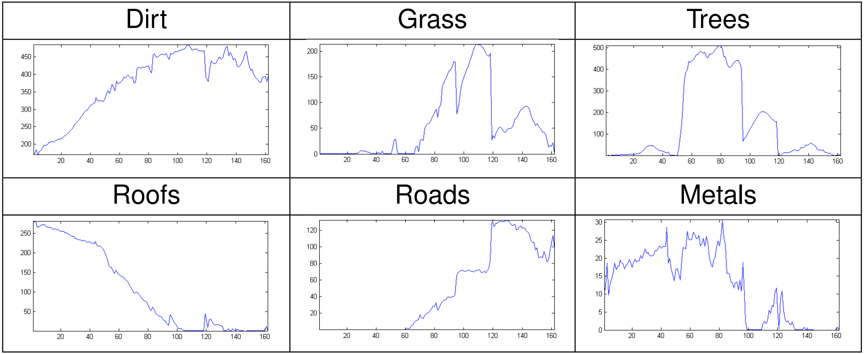


Figure: Spectra

Hyperspectral Imaging

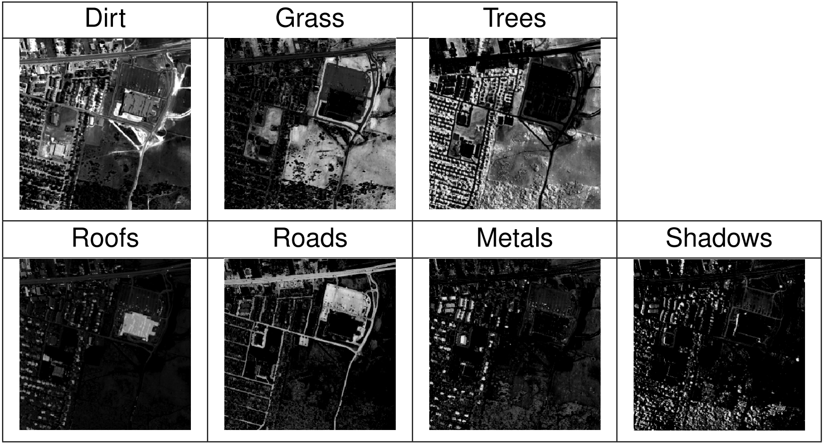


Figure: Abundances with $p = 7$, $\mu = 10^{-10}$, $\lambda = 0$, $\nu = 1.1$, uniform $K^{(0)}$, random $X^{(0)}$, 500 it in 39min10s

Hyperspectral Imaging

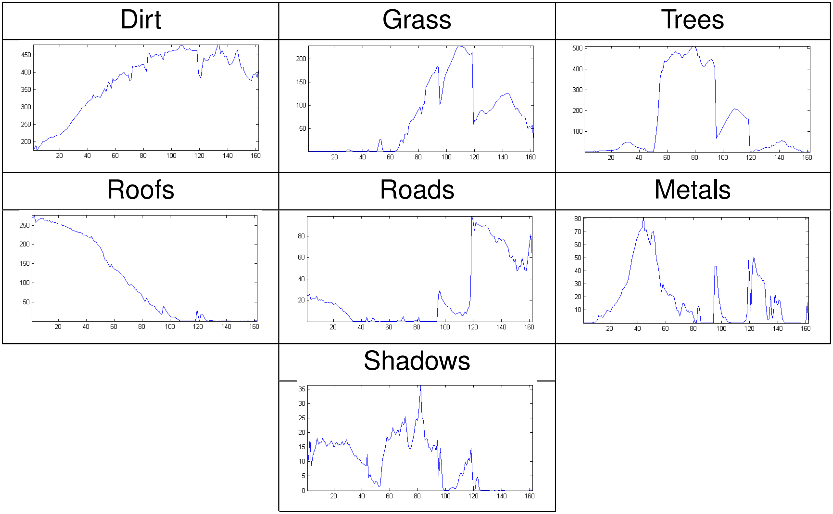


Figure: Spectra

Hyperspectral Imaging

Example: San Diego Airport HYDICE Hypercube $400 \times 400 \times 158$

- Y : $400^2 \times 158$ data matrix
- K : $400^2 \times p$ abundance matrix
- X : $p \times 158$ matrix containing the spectra of the basis elements

Hyperspectral Imaging with TV penalty

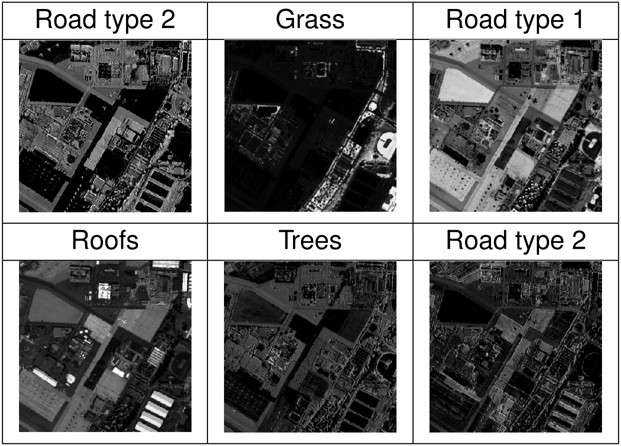


Figure:
Abundances with $p = 6$, $\lambda_{TV} = 0.001$, $\varepsilon = \sqrt{10^{-7}}$, $\lambda = 0$, $\nu = 0.05$, uniform $K^{(0)}$, random $X^{(0)}$, 1000 it in 3h36min59s

Hyperspectral Imaging with TV penalty

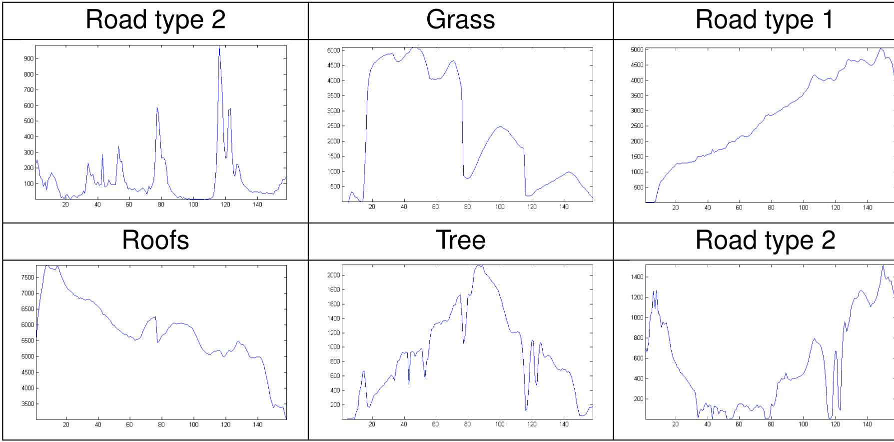


Figure: Spectra

Hyperspectral Imaging with TV penalty

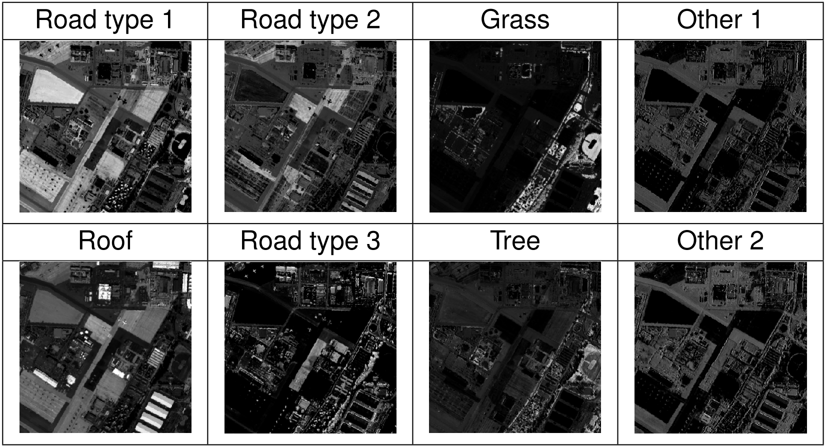


Figure:
Abundances with $p = 8$, $\lambda_{TV} = 0.001$, $\varepsilon = \sqrt{10^{-7}}$, $\lambda = 0$, $\nu = 0.05$
uniform $K^{(0)}$, random $X^{(0)}$, 500 it in 2h17min39s

Hyperspectral Imaging with TV penalty

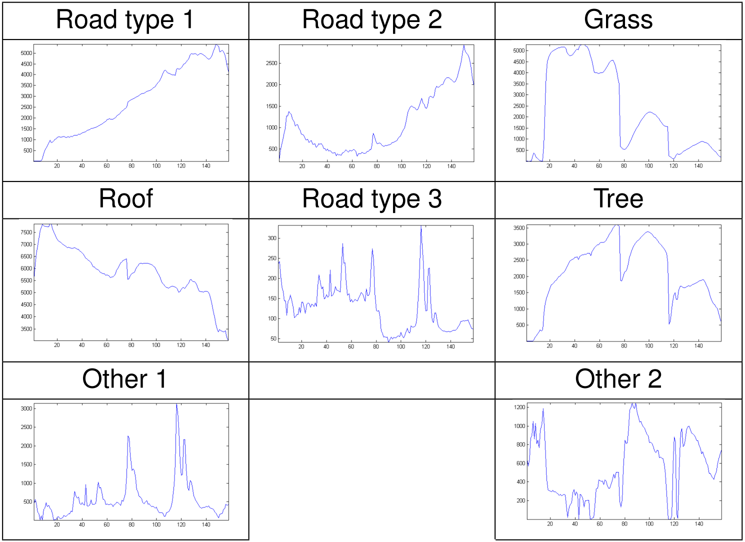


Figure: Spectra

Generalization to the β-divergence

- Some of our convergence results can be extended to the case of the β-divergence considered by Févotte and Idier (2011)

$$D_\beta(Y, HX) = \sum_{i=1}^{n} \sum_{j=1}^{m} d_\beta \left(Y_{i,j}, (HX)_{i,j} \right)$$

with

$$d_\beta(y,x) = \begin{cases} y \ln\left(\frac{y}{x}\right) - y + x & \text{if} \quad \beta = 1 \\ \frac{y}{x} - \ln\left(\frac{y}{x}\right) - 1 & \text{if} \quad \beta = 0 \\ \frac{1}{\beta(\beta-1)}\left(y^\beta + (\beta-1)x^\beta - \beta y x^{\beta-1}\right) & \text{if} \quad \beta \neq 0, \beta \neq 1 \end{cases}.$$

- Special cases (NB. The β-divergence is convex *iff* $1 \leq \beta \leq 2$)

$\beta = 0$: Itakura-Saito divergence ; $\beta = 1$: Kullback-Leibler divergence

$\beta = 2$: least-squares

Recent related (methodological) work

(with convergence proofs)

- Algorithms based on the SGP algorithm by Bonettini, Zanella, Zanni 2009
  (Prato, La Camera, Bonettini, Bertero 2013;
  Ben Hadj, Blanc-Féraud and Aubert 2012)
- Inexact block coordinate descent
  (Bonettini 2011)
- Underapproximations for Sparse Nonnegative Matrix Factorization
  (Gillis and Glineur 2010)
- Proximal Alternating Minimization and Projection Methods for Nonconvex Problems
  (Attouch, Bolte, Redont, Soubeyran 2010; Bolte, Combettes and Pesquet 2010)
- Others?