

Sparse Representations: from Source Separation to Compressed Sensing

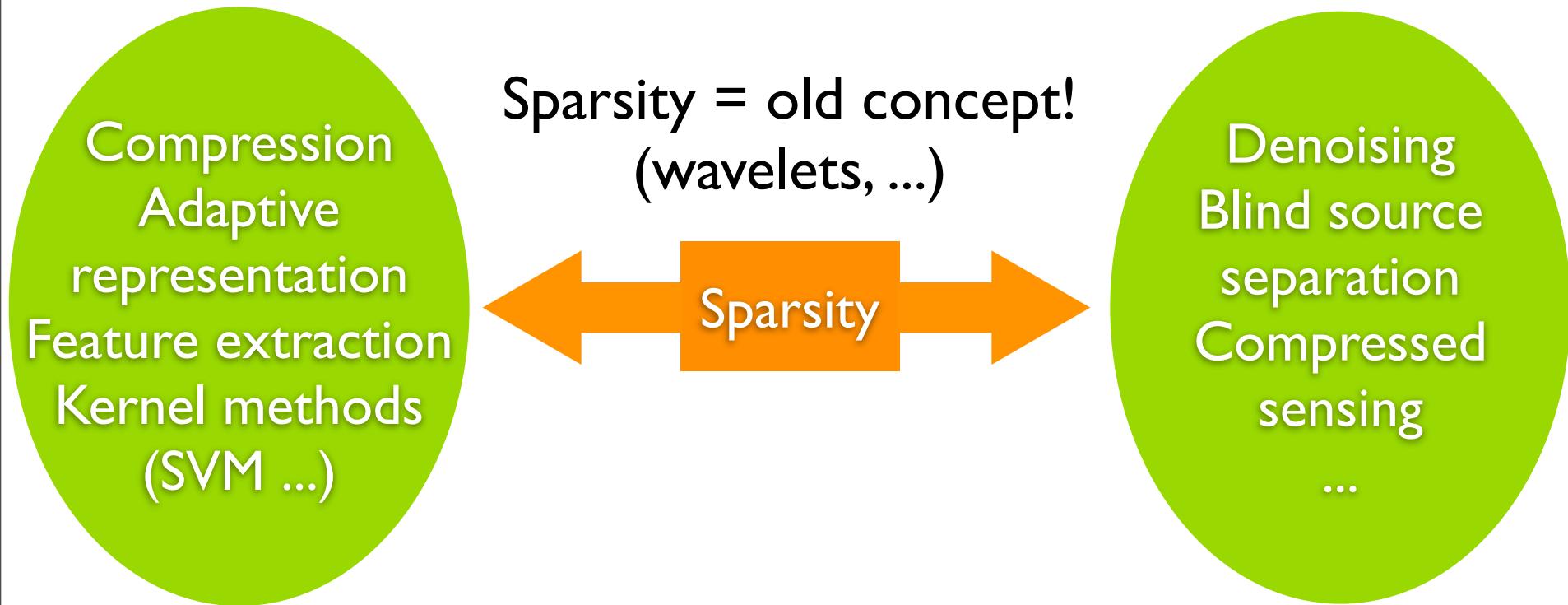
Rémi Gribonval
METISS project-team
IRISA/INRIA



UNE UNITÉ DE RECHERCHE À LA POINTE DES SCIENCES
ET DES TECHNOLOGIES DE L'INFORMATION
ET DE LA COMMUNICATION

HDR Defense - October 24th, 2007

Introduction



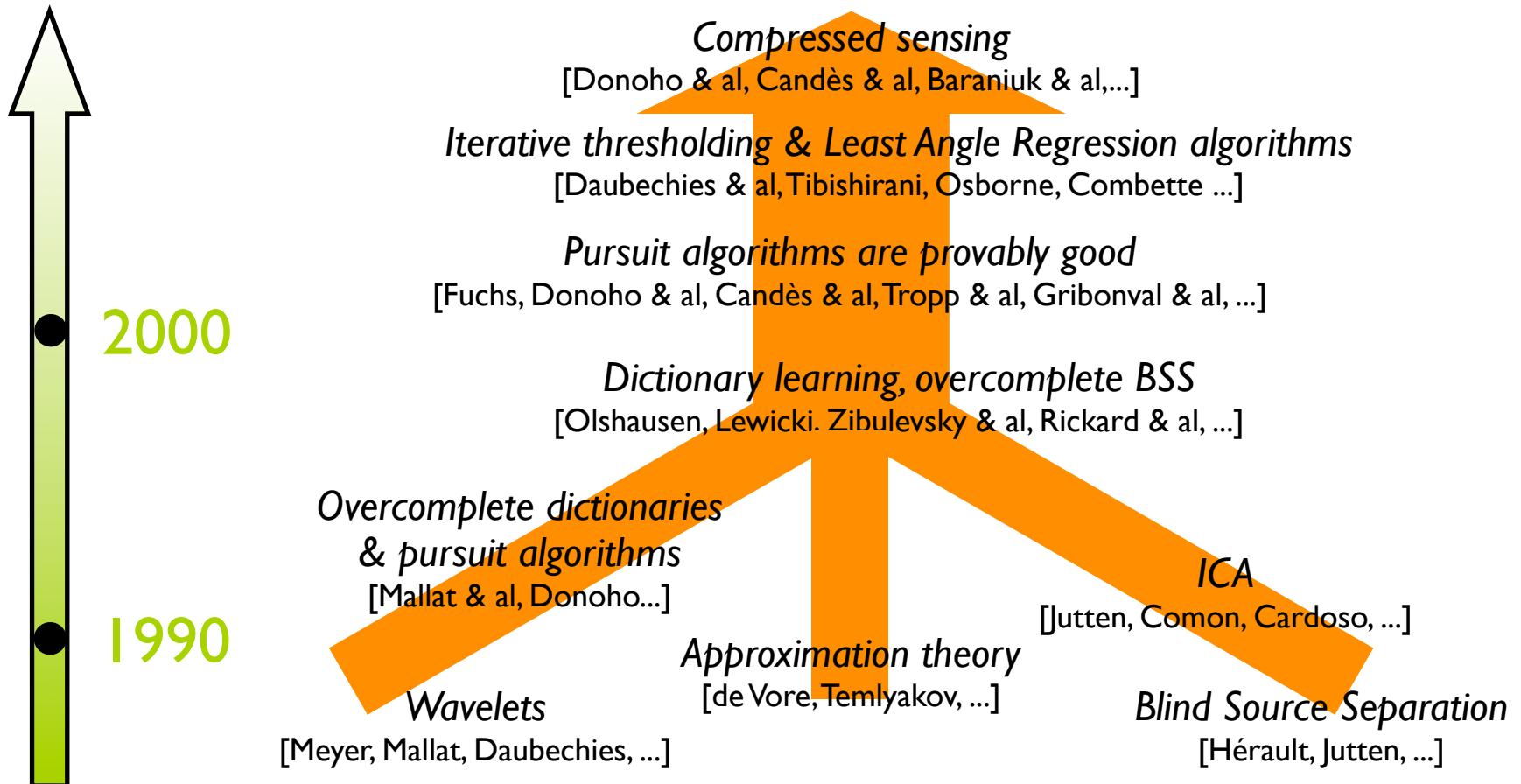
Natural / traditional role :

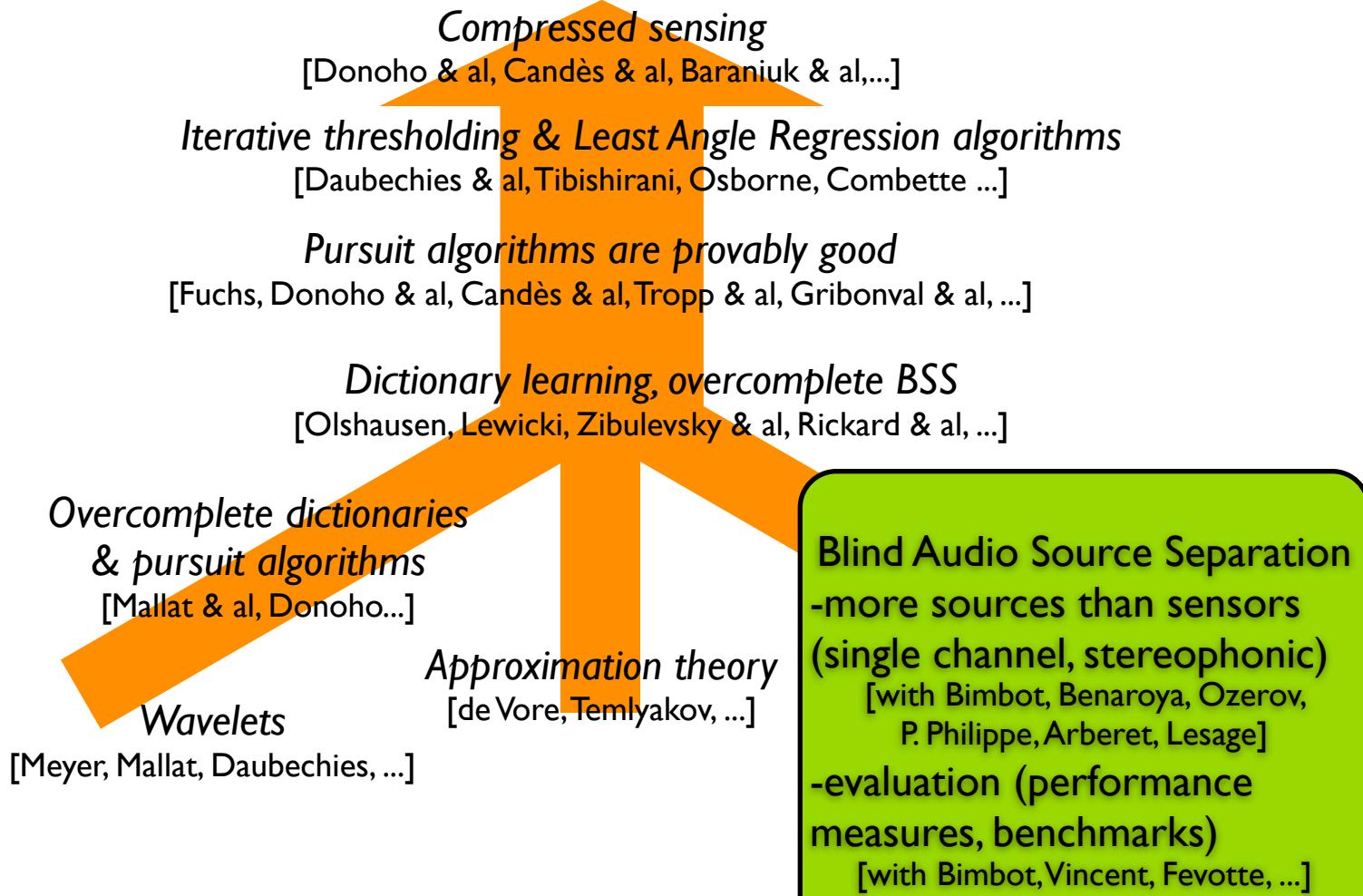
Sparsity = low cost (bits, computations, ...)
direct goal

Novel indirect role

Sparsity = prior knowledge
Tool for inverse problems

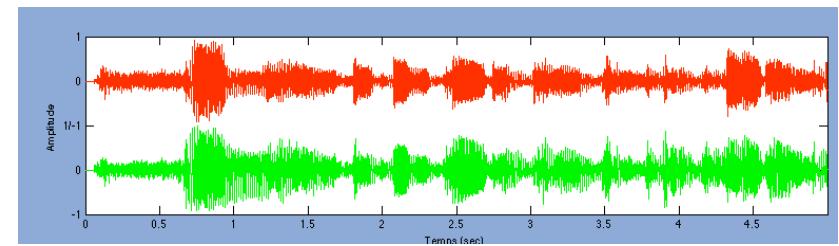
Milestones





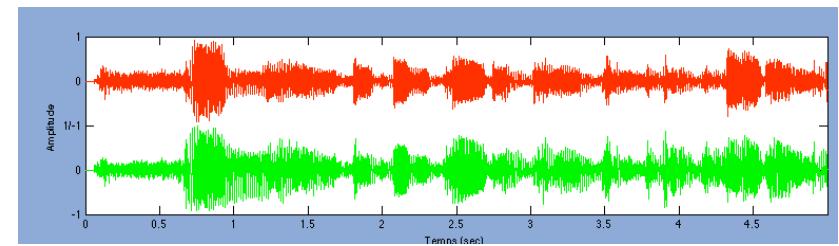
« Blind » Audio Source Separation

- « Softly as in a morning sunrise »



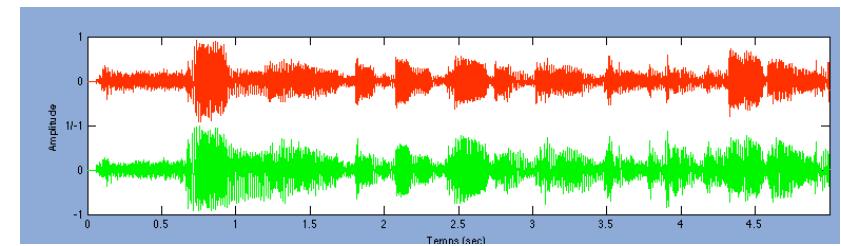
« Blind » Audio Source Separation

- « Softly as in a morning sunrise »



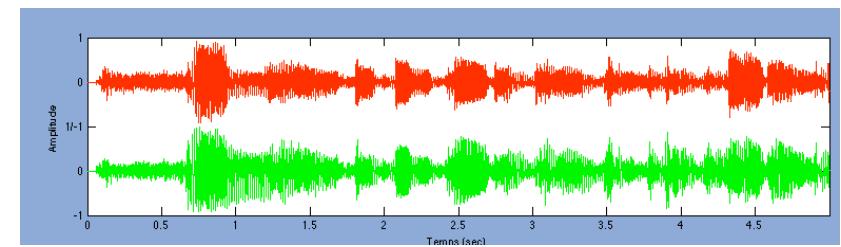
« Blind » Audio Source Separation

- « Softly as in a morning sunrise »



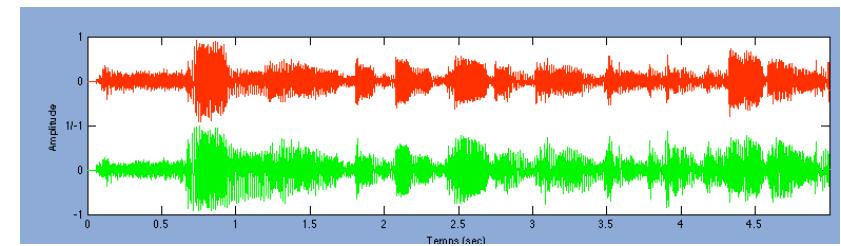
« Blind » Audio Source Separation

- « Softly as in a morning sunrise »



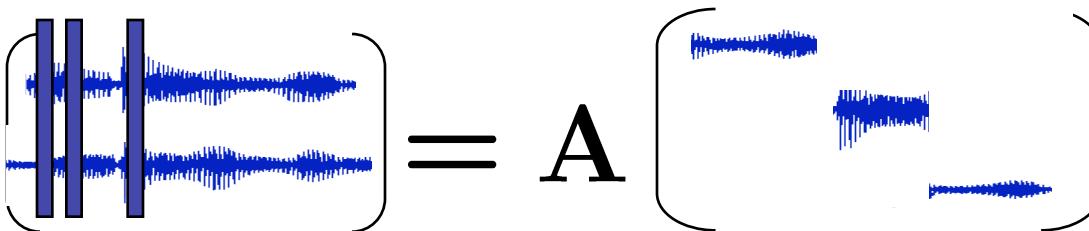
« Blind » Audio Source Separation

- « Softly as in a morning sunrise »

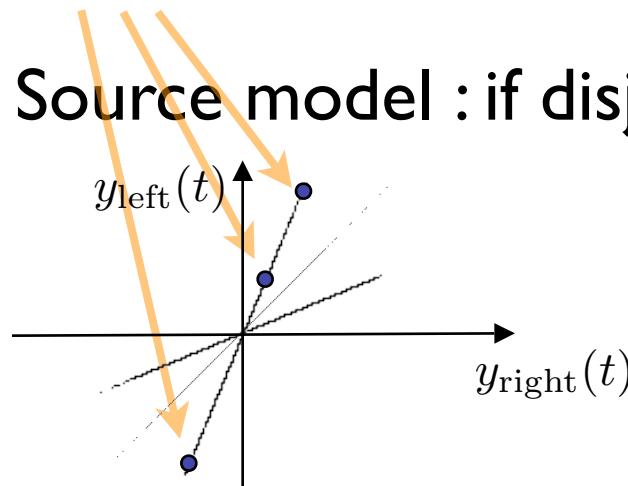


Blind Source Separation

- Mixing model : linear instantaneous mixture

$$\begin{matrix} y_{\text{right}}(t) \\ y_{\text{left}}(t) \end{matrix} = \mathbf{A} \begin{matrix} s_1(t) \\ s_2(t) \\ s_3(t) \end{matrix}$$


- Source model : if disjoint time-supports ...



... then clustering to :
1- identify (columns of) the mixing matrix
2- recover sources

Blind Source Separation: two-step approach

$$\text{Observed data } \mathbf{y}(t) \approx \mathbf{A} \cdot \mathbf{s}(t) \text{ Unknown}$$

- Estimate mixing matrix $\hat{\mathbf{A}}$
 - ◆ Coarse source model (independence, sparsity, ...)
- Estimate the sources $\hat{\mathbf{s}}(t) = \hat{\mathbf{A}}^{-1} \cdot \mathbf{y}(t)$

Blind Source Separation: two-step approach

$$\text{Observed data } \mathbf{y}(t) \approx \mathbf{A} \cdot \mathbf{s}(t) \text{ Unknown}$$

- Estimate mixing matrix $\hat{\mathbf{A}}$
 - ◆ Coarse source model (independence, sparsity, ...)
- Estimate the sources $\hat{\mathbf{s}}(t) = \hat{\mathbf{A}}^{-1} \cdot \mathbf{y}(t)$

More sources than sensors = underdetermined
need finer source model = sparse / disjoint / structured representations

Multichannel recordings

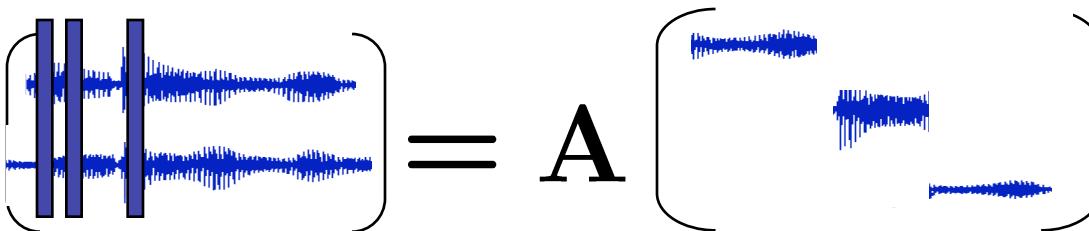
[Ph.D. Lesage, Ph.D. Arberet, with Bimbot]
Matching Pursuits + Clustering

Monophonic recordings

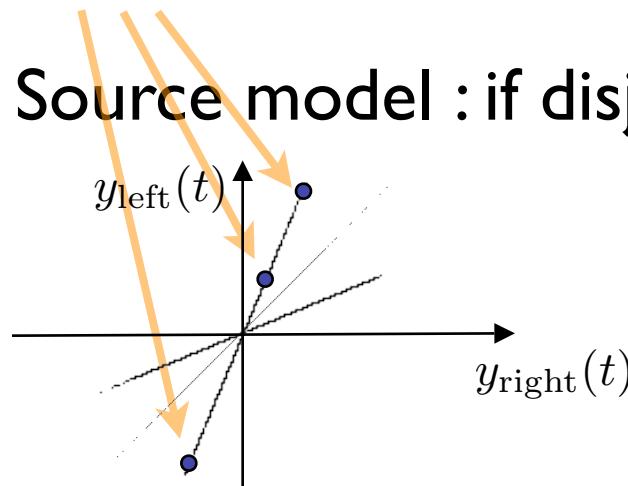
[with Benaroya, Bimbot, Philippe, Ph.D. Ozerov]
Adaptive Wiener Filtering

Blind Source Separation

- Mixing model : linear instantaneous mixture

$$\begin{matrix} y_{\text{right}}(t) \\ y_{\text{left}}(t) \end{matrix} = \mathbf{A} \begin{matrix} s_1(t) \\ s_2(t) \\ s_3(t) \end{matrix}$$


- Source model : if disjoint time-supports ...



... then clustering to :
1- identify (columns of) the mixing matrix
2- recover sources

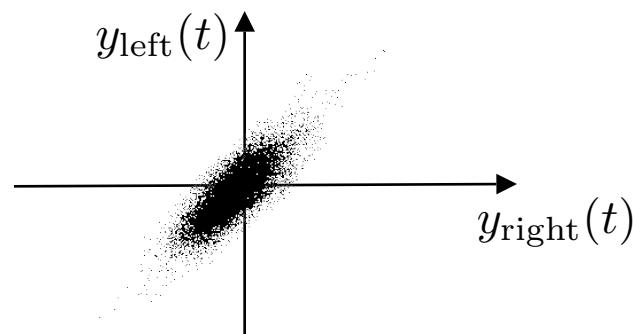
Blind Source Separation

- Mixing model : linear instantaneous mixture

$$\begin{matrix} y_{\text{right}}(t) \\ y_{\text{left}}(t) \end{matrix} = A \quad \begin{matrix} s_1(t) \\ s_2(t) \\ s_3(t) \end{matrix}$$

The diagram illustrates the linear instantaneous mixing model. On the left, two output signals, $y_{\text{right}}(t)$ and $y_{\text{left}}(t)$, are shown as blue waveforms. These signals are multiplied by a matrix A to produce three input signals, $s_1(t)$, $s_2(t)$, and $s_3(t)$, which are also represented as blue waveforms on the right.

- In practice ...

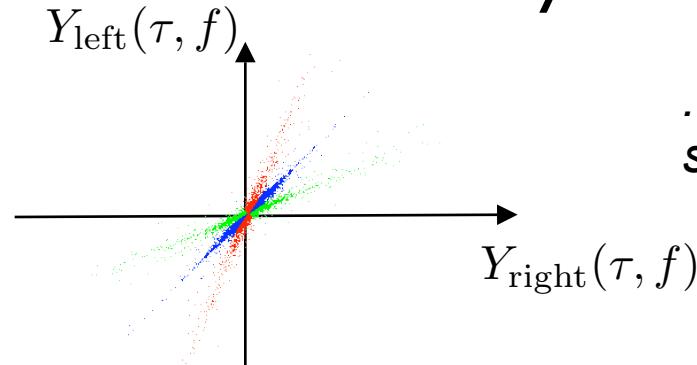


Time-Frequency Masking

- Mixing model in the time-frequency domain

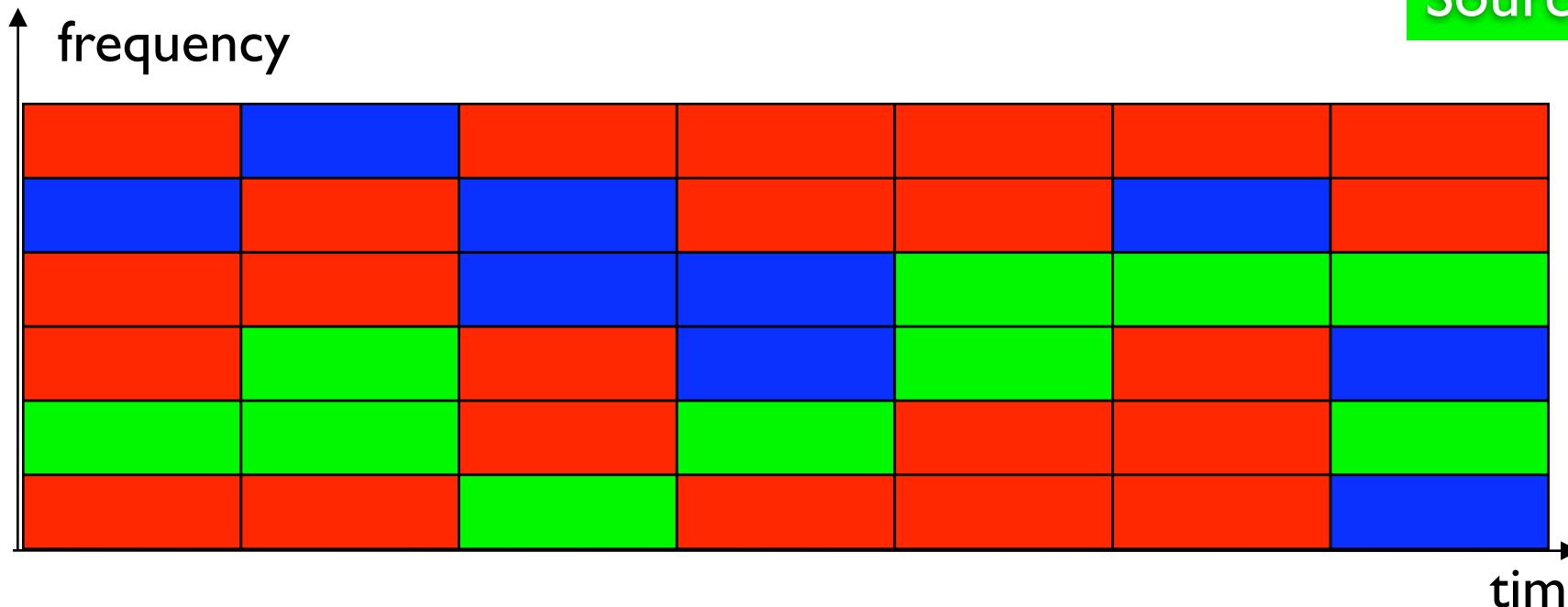
$$\begin{matrix} Y_{\text{right}}(\tau, f) \\ Y_{\text{left}}(\tau, f) \end{matrix} \left(\begin{array}{c} \text{[Heatmap]} \\ \text{[Heatmap]} \end{array} \right) = \mathbf{A} \mathbf{S}(\tau, f)$$

- And “miraculously” ...

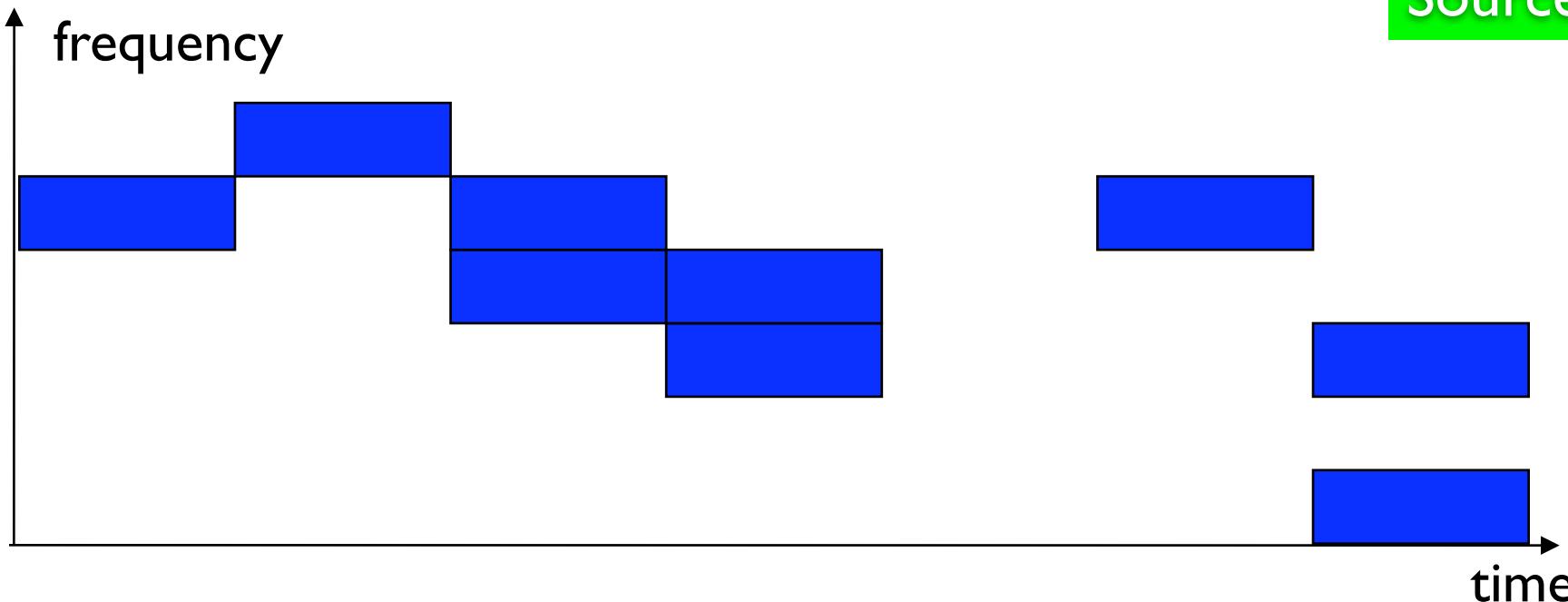


*... time-frequency representations of audio signals are (often) **almost disjoint**.*

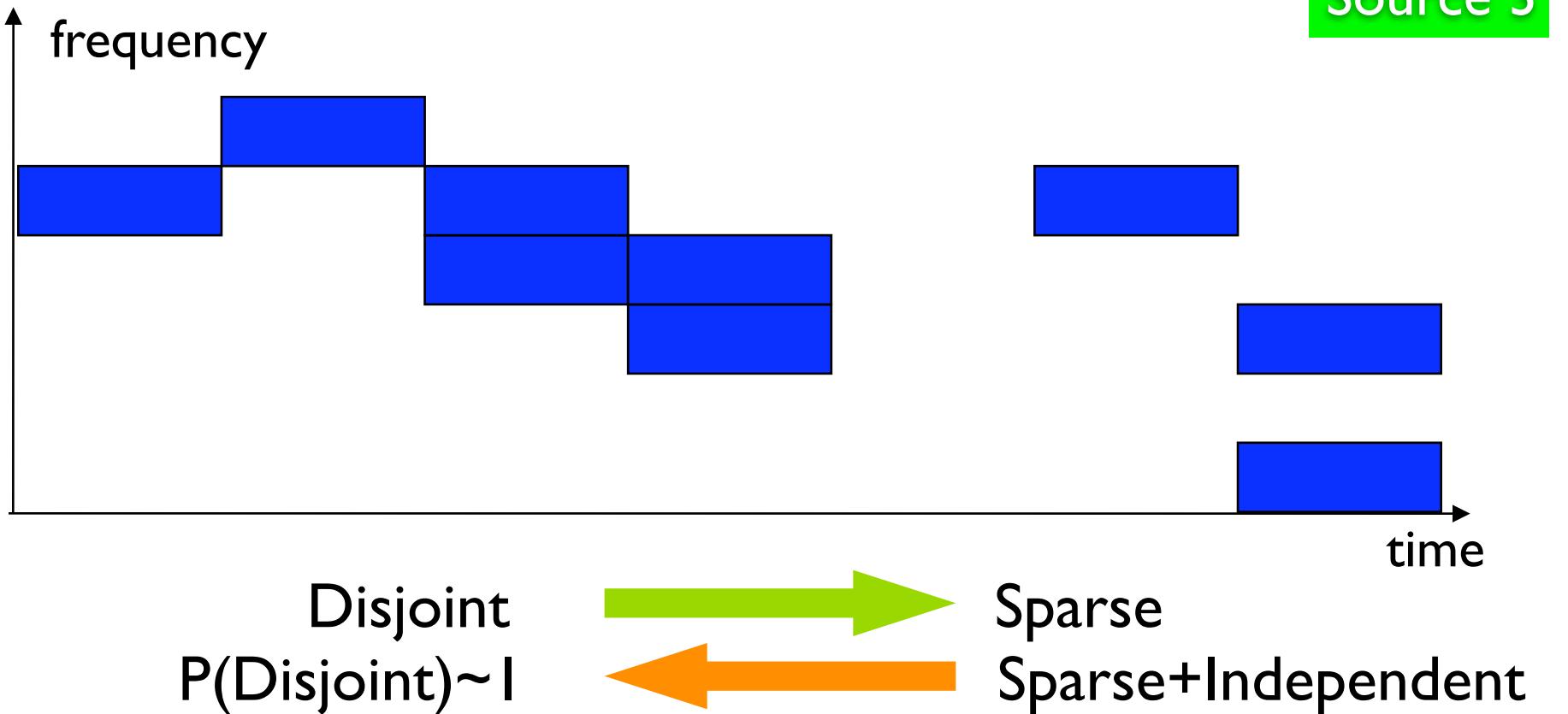
Disjoint Time-Frequency Representations



Disjoint Time-Frequency Representations



Disjoint Time-Frequency Representations



Sparse representations
-time-frequency dictionaries
(chirps, harmonic structures,
multichannel, ...)
[with Bacry, Mallat, Lesage, Bimbot]
-fast Matching Pursuit algorithms
[with Bacry, Mallat, Krstulovic, Roy]

Compressed sensing

[Donoho & al, Candès & al, Baraniuk & al,...]

Iterative thresholding & Least Angle Regression algorithms

[Daubechies & al, Tibshirani, Osborne, Combette ...]

Pursuit algorithms are provably good

[Fuchs, Donoho & al, Candès & al, Tropp & al, Gribonval & al, ...]

Dictionary learning, overcomplete BSS

[Olshausen, Lewicki, Zibulevsky & al, Rickard & al, ...]

ICA

[Jutten, Comon, Cardoso, ...]

Approximation theory

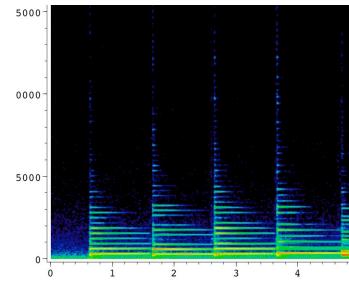
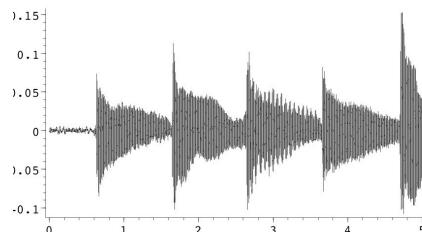
[de Vore, Temlyakov, ...]

Blind Source Separation

[Hérault, Jutten, ...]

Sparse Time-Frequency Representations

- Short Time-Fourier Transform of audio = sparse



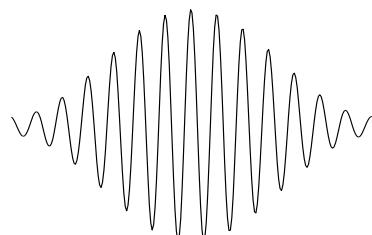
zero = black

- Analysis $Y(\tau, f) = \langle y, g_{\tau,f} \rangle$

Time-frequency atom

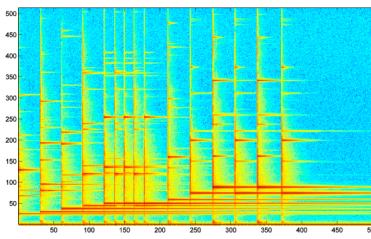
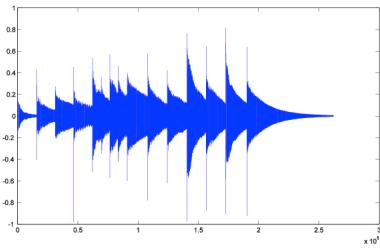
- Reconstruction $y(t) = \sum_{\tau,f} Y(\tau, f) g_{\tau,f}(t)$

$$g_{\tau,f}(t) := w(t - \tau) e^{2i\pi f t}$$



Multiscale Time-Frequency Structures

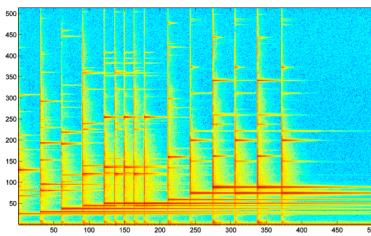
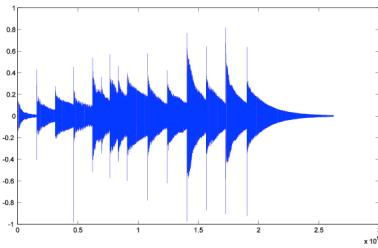
- Audio = superimposition of structures
- Example : glockenspiel



- ◆ transients = small scale
- ◆ harmonic part = large scale
- Gabor atoms $\left\{ g_{s,\tau,f}(t) = \frac{1}{\sqrt{s}} w\left(\frac{t-\tau}{s}\right) e^{2i\pi f t} \right\}_{s,\tau,f}$

Multiscale Time-Frequency Structures

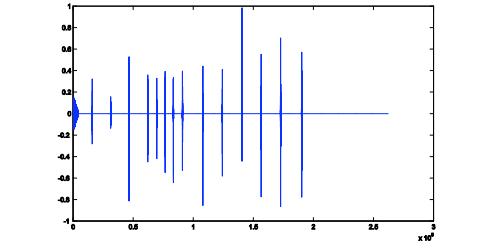
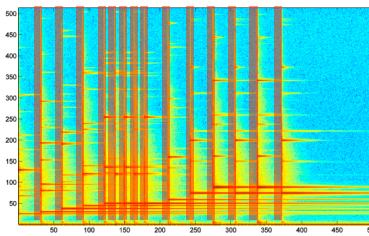
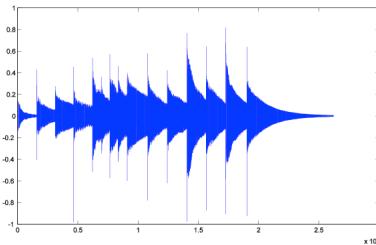
- Audio = superimposition of structures
- Example : glockenspiel



- ◆ transients = small scale
- ◆ harmonic part = large scale
- Gabor atoms $\left\{ g_{s,\tau,f}(t) = \frac{1}{\sqrt{s}} w\left(\frac{t-\tau}{s}\right) e^{2i\pi f t} \right\}_{s,\tau,f}$

Multiscale Time-Frequency Structures

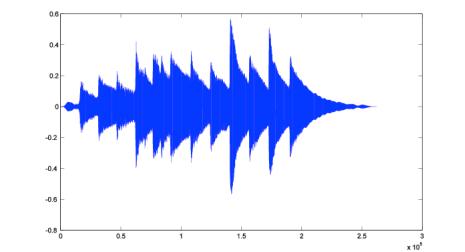
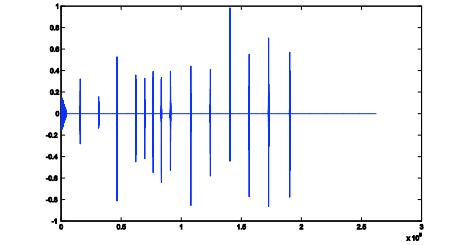
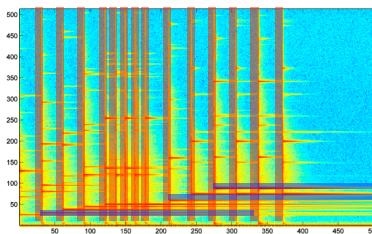
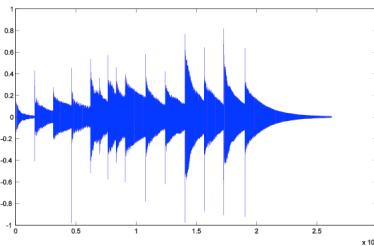
- Audio = superimposition of structures
- Example : glockenspiel



- ◆ transients = small scale
- ◆ harmonic part = large scale
- Gabor atoms $\left\{ g_{s,\tau,f}(t) = \frac{1}{\sqrt{s}} w\left(\frac{t-\tau}{s}\right) e^{2i\pi f t} \right\}_{s,\tau,f}$

Multiscale Time-Frequency Structures

- Audio = superimposition of structures
- Example : glockenspiel



- ◆ transients = small scale
- ◆ harmonic part = large scale
- Gabor atoms $\left\{ g_{s,\tau,f}(t) = \frac{1}{\sqrt{s}} w\left(\frac{t-\tau}{s}\right) e^{2i\pi f t} \right\}_{s,\tau,f}$

Sparse Redundant Representations

- Sparse signal model

$$y(t) \approx \sum_{s,\tau,f} c_{s,\tau,f} \cdot g_{s,\tau,f}(t)$$

Sparse representation =
unknown, but few
significant coefficients

Gabor dictionary =
redundant (more atoms
than signal dimension)

- Infinitely many representations
 - ◆ can choose a preferred one, e.g. the “sparsest”
 - ◆ how to compute it ?

Inverse Linear Problems

$$y(t) \approx \sum_k c_k g_k(t)$$

$$\mathbf{y}(t) \approx \mathbf{A} \cdot \mathbf{s}(t)$$



Observed data:
signal, image, mixture of sources,...

$$\mathbf{b} \approx \mathbf{A}x$$

Known linear system:
dictionary, (estimated) mixing matrix

Unknown
representation, sources, ...

Global Algorithms

- Approximation quality

$$\|\mathbf{A}x - \mathbf{b}\|_2$$

- Ideal sparsity measure : ℓ^0 “norm”

$$\|x\|_0 := \#\{k, x_k \neq 0\} = \sum_k |x_k|^0$$

- Relaxed sparsity measure

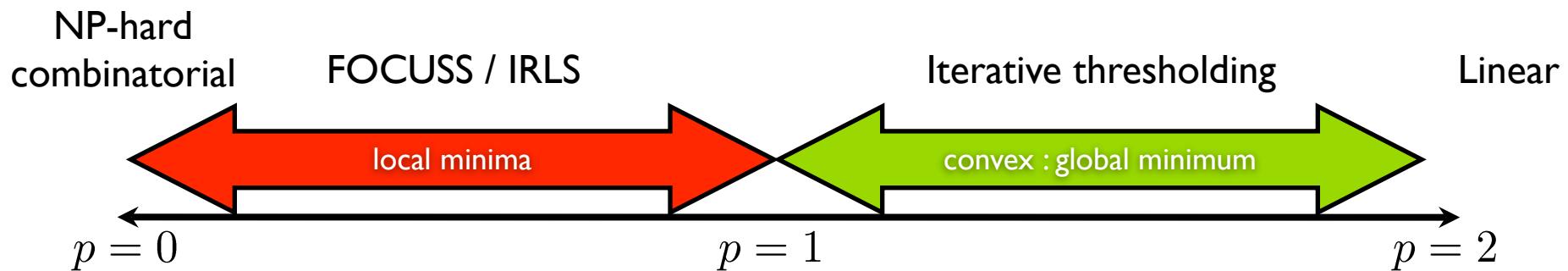
$$\|x\|_p := \sum_k |x_k|^p$$

Global Algorithms

- Global optimization

$$\min_x \frac{1}{2} \|\mathbf{A}x - \mathbf{b}\|_2^2 + \lambda \|x\|_p$$

- ◆ Sparse representation $\lambda \rightarrow 0$
- ◆ Sparse approximation $\lambda > 0$



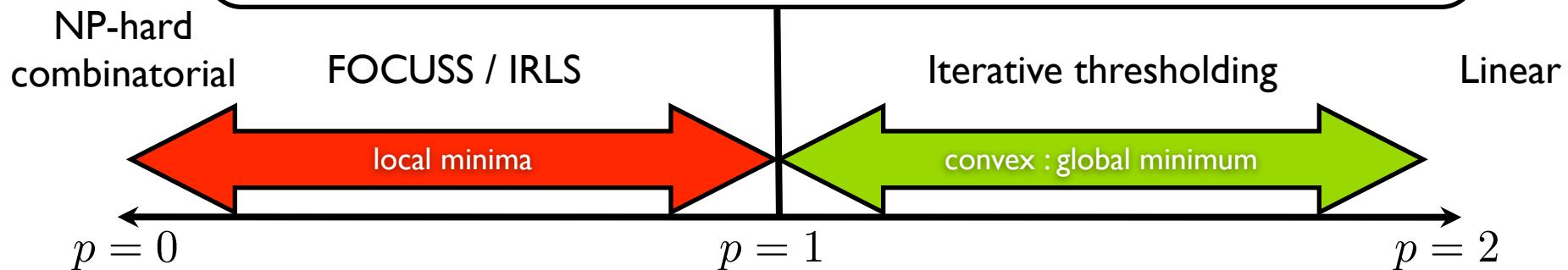
Global Algorithms

- Global optimization

$$\min_x \frac{1}{2} \|\mathbf{A}x - \mathbf{b}\|_2^2 + \lambda \|x\|_p$$

- ◆ Sparse representation $\lambda \rightarrow 0$
- ◆ Sparse approximation $\lambda > 0$

Lasso [Tibshirani 1996], Basis Pursuit (Denoising) [Chen, Donoho & Saunders, 1999]
Linear/Quadratic programming (interior point, etc.)
Homotopy method [Osborne 2000] / Least Angle Regression [Efron & al 2002]



Matching Pursuit

- Matching Pursuit Algorithm [Mallat & Zhang 1993]
 - ◆ initialize residual $\mathbf{b}^{(0)} := \mathbf{b}$ $\mathbf{A} = [\mathbf{a}_1, \dots, \mathbf{a}_K]$
 - ◆ find best atom $k_m := \arg \max_k |\langle \mathbf{b}^{(m-1)}, \mathbf{a}_k \rangle|$
 - ◆ update residual $\mathbf{b}^{(m)} := \mathbf{b}^{(m-1)} - \langle \mathbf{b}^{(m-1)}, \mathbf{a}_{k_m} \rangle \mathbf{a}_{k_m}$

Matching Pursuits

- Matching Pursuit Algorithm [Mallat & Zhang 1993]
 - ◆ initialize residual $\mathbf{b}^{(0)} := \mathbf{b}$ $\mathbf{A} = [\mathbf{a}_1, \dots, \mathbf{a}_K]$
 - ◆ find best atom $k_m := \arg \max_k |\langle \mathbf{b}^{(m-1)}, \mathbf{a}_k \rangle|$
 - ◆ update residual $\mathbf{b}^{(m)} := \mathbf{b}^{(m-1)} - \langle \mathbf{b}^{(m-1)}, \mathbf{a}_{k_m} \rangle \mathbf{a}_{k_m}$

Multichannel Matching Pursuit

-theoretical analysis [with Rauhut, Schnass & Vandergheynst]
-application to source separation [Ph.D. Lesage, with Bimbot]

Demixing Pursuit

-theoretical analysis [with Nielsen]
-application to source separation [Ph.D. Lesage, with Bimbot]

Matching Pursuits

- Matching Pursuit Algorithm [Mallat & Zhang 1993]
 - ◆ initialize residual $\mathbf{b}^{(0)} := \mathbf{b}$ $\mathbf{A} = [\mathbf{a}_1, \dots, \mathbf{a}_K]$
 - ◆ find best atom $k_m := \arg \max_k |\langle \mathbf{b}^{(m-1)}, \mathbf{a}_k \rangle|$
 - ◆ update residual $\mathbf{b}^{(m)} := \mathbf{b}^{(m-1)} - \langle \mathbf{b}^{(m-1)}, \mathbf{a}_{k_m} \rangle \mathbf{a}_{k_m}$

Matching Pursuit ToolKit [with Krstulovic, Lesage, Roy]
= efficient Matching Pursuit for large scale data

Multichannel Matching Pursuit

-theoretical analysis [with Rauhut, Schnass & Vandergheynst]
-application to source separation [Ph.D. Lesage, with Bimbot]

Demixing Pursuit

-theoretical analysis [with Nielsen]
-application to source separation [Ph.D. Lesage, with Bimbot]

- ◆ handle 1hour audio signals instead of 5 seconds
- ◆ experiments on database of > 2000 songs

Compressed sensing

[Donoho & al, Candès & al, Baraniuk & al,...]

Iterative thresholding & Least Angle Regression algorithms

Provably good algorithms

[with Nielsen, Vandergheynst, Rauhut, Schnass]

Matching Pursuit

ℓ^p optimization $0 \leq p \leq 1$

Union of bases, incoherent dictionaries, structured dictionaries

Overcomplete dictionaries

& pursuit algorithms

[Mallat & al, Donoho...]

Wavelets

[Meyer, Mallat, Daubechies, ...]

Approximation theory

[de Vore, Temlyakov, ...]

ICA

[Jutten, Comon, Cardoso, ...]

Blind Source Separation

[Hérault, Jutten, ...]

Sparsity and Ill-Posed Inverse Problems

- Ill-posedness if more unknowns than equations

$$\mathbf{A}x_0 = \mathbf{A}x_1 \not\Rightarrow x_0 = x_1$$

- Uniqueness of sparse solutions:
 - ◆ if x_0, x_1 are “sufficiently sparse”,
 - ◆ then $\mathbf{A}x_0 = \mathbf{A}x_1 \Rightarrow x_0 = x_1$

Example : l-sparse Representations

- Uniqueness of l-sparse representations

$$\mathbf{b} = \mathbf{A}x_0 \quad \mathbf{b} = \mathbf{A}x_1 \quad \rightarrow \quad x_0 = x_1$$

The diagram shows two sparse vectors x_0 and x_1 . Both vectors have a single green nonzero component. In x_0 , this component is at the first position. In x_1 , it is at the second position. A red 'X' is placed over the second position of x_1 to indicate it is incorrect or irrelevant. An orange arrow points from the equations to the conclusion $x_0 = x_1$.

- Recovery = correlation with atoms \mathbf{a}_n
 - ◆ index of nonzero component $\hat{k} := \arg \max_k |\langle \mathbf{b}, \mathbf{a}_k \rangle|$
 - ◆ principle of DUET source separation [Jourjine & al 2000]
 - ◆ similar to first step of Matching Pursuit
- Extension to M-sparse representations ?

Ideal Sparse Representation

- Optimization problem = NP-hard [Natarajan, Davies &al]

$$x^* := \arg \min_x \|x\|_0 \text{ subject to } \mathbf{A}x = \mathbf{b}$$

- If any $2M$ columns of \mathbf{A} are linearly independent then

$$\mathbf{A}x = \mathbf{A}y, \|x\|_0 \leq M, \|y\|_0 \leq M \quad \rightarrow \quad x = y$$

- Proof : $\|x - y\|_0 \leq 2M$ and $\mathbf{A}(x - y) = 0$

Coherence of a Dictionary

- Definition (easily computable) $\mu = \mu(\mathbf{A}) := \max_{k \neq l} |\langle \mathbf{a}_k, \mathbf{a}_l \rangle|$
- Property $\|x\|_0 \leq 2M \Rightarrow \|\mathbf{A}x\|_2^2 \geq (1 - (2M - 1)\mu) \cdot \|x\|_2^2$

Theorem [Fuchs, G. & Nielsen, Donoho & Elad, Tropp, G. & Vandergheynst]

I. Uniqueness of M-sparse representations whenever

$$\|x\|_0 \leq M < (1 + 1/\mu)/2$$

2. Recovery with Basis Pursuit & Matching Pursuit if

$$\|x\|_0 \leq M < (1 + 1/\mu)/2$$

ℓ^p optimisation for any $0 \leq p \leq 1$ [with Nielsen, Vandergheynst & Figueras]

Restricted Isometry Constants

- Definition : isometry constant = smallest δ_M such that $\|x\|_0 \leq M \Rightarrow 1 - \delta_M \leq \frac{\|\mathbf{A}x\|_2^2}{\|x\|_2^2} \leq 1 + \delta_M$

Theorem [Candès, Romberg & Tao]

1. Uniqueness of M-sparse representations whenever

$$\|x\|_0 \leq M \quad \delta_{2M} < 1$$

2. Recovery by Basis Pursuit if

$$\|x\|_0 \leq M \quad \delta_{2M} + \delta_{3M} < 1$$

Coherence vs Isometry

Constants

max over $K(K-1)$ entries

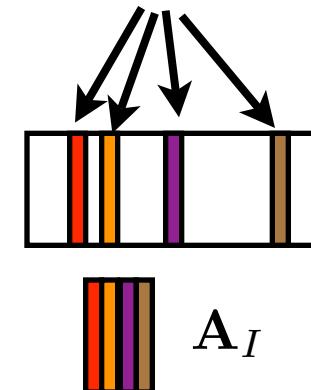
$$\mu = \mu(\mathbf{A}) := \max_{k \neq l} |\langle \mathbf{a}_k, \mathbf{a}_l \rangle|$$

(Cumulative) coherence

Low cost
Coarse / pessimistic

K columns

$$k \in I, \#I \leq M$$



max over $\frac{K!}{M!(K-M)!}$ subsets I

$$\delta_M := \sup_{\#I \leq M, c \in \mathbb{R}^M} \left| \frac{\|\mathbf{A}_I c\|_2^2}{\|c\|_2^2} - 1 \right|$$

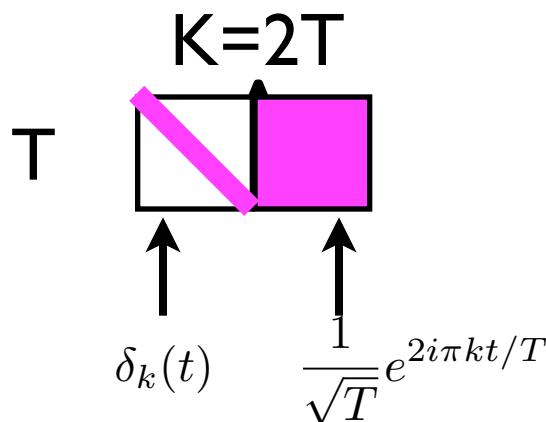
Isometry constants

Hard to compute
~Sharp

[with Rauhut, Schnass & Vanderghenst]
average case analysis for multichannel algorithms

Examples

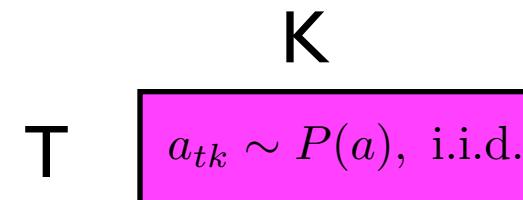
- Dirac-Fourier dictionary



- Coherence

$$\mu = 1/\sqrt{T}$$

- “Generic” (random) dictionary
[Candès & al, Vershynin, ...]



- Isometry constants
if $T \geq CM \log K/M$
then $P(\delta_{2M} + \delta_{3M} < 1) \approx 1$

Recovery by Basis Pursuit

$C' \gg 1$

$$M_{\text{BasisPursuit}}(\mathbf{A}) \approx 0.914\sqrt{T}$$

$$M_{\text{BasisPursuit}}(\mathbf{A}) \gtrsim C'T / \log^a(T)$$

Compressed Sensing

- MRI from incomplete measures

[Candès, Romberg & Tao]

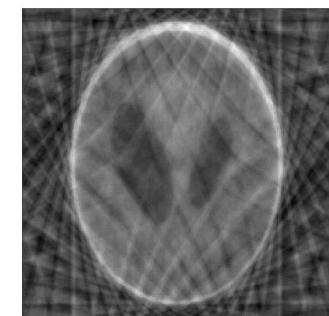
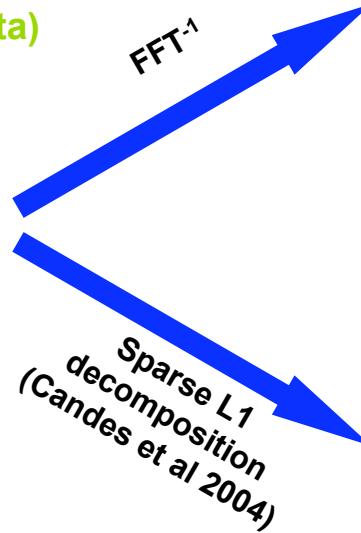
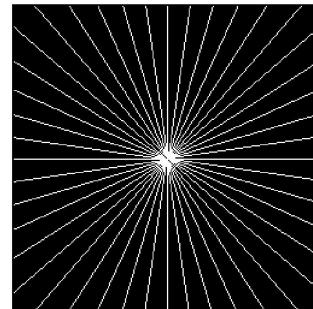
Data



Lossy
measurement
= tomography



Measured data
(FFT minus lost data)



Reconstruction



Compressed Sensing

- MRI from incomplete measures

[Candès, Romberg & Tao]

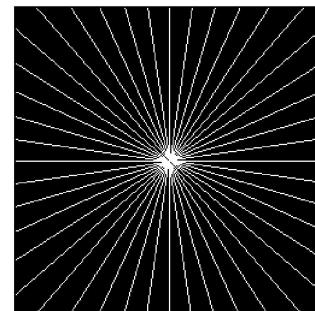
Data



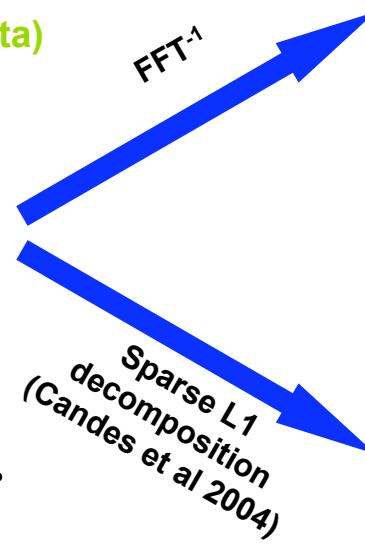
Lossy
measurement
= tomography



Measured data
(FFT minus lost data)



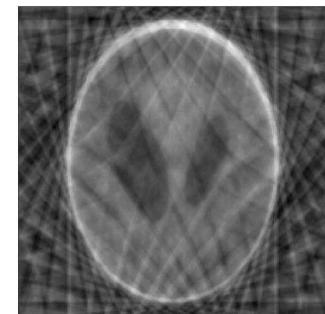
FFT⁻¹



$$y = \mathbf{Ax}$$

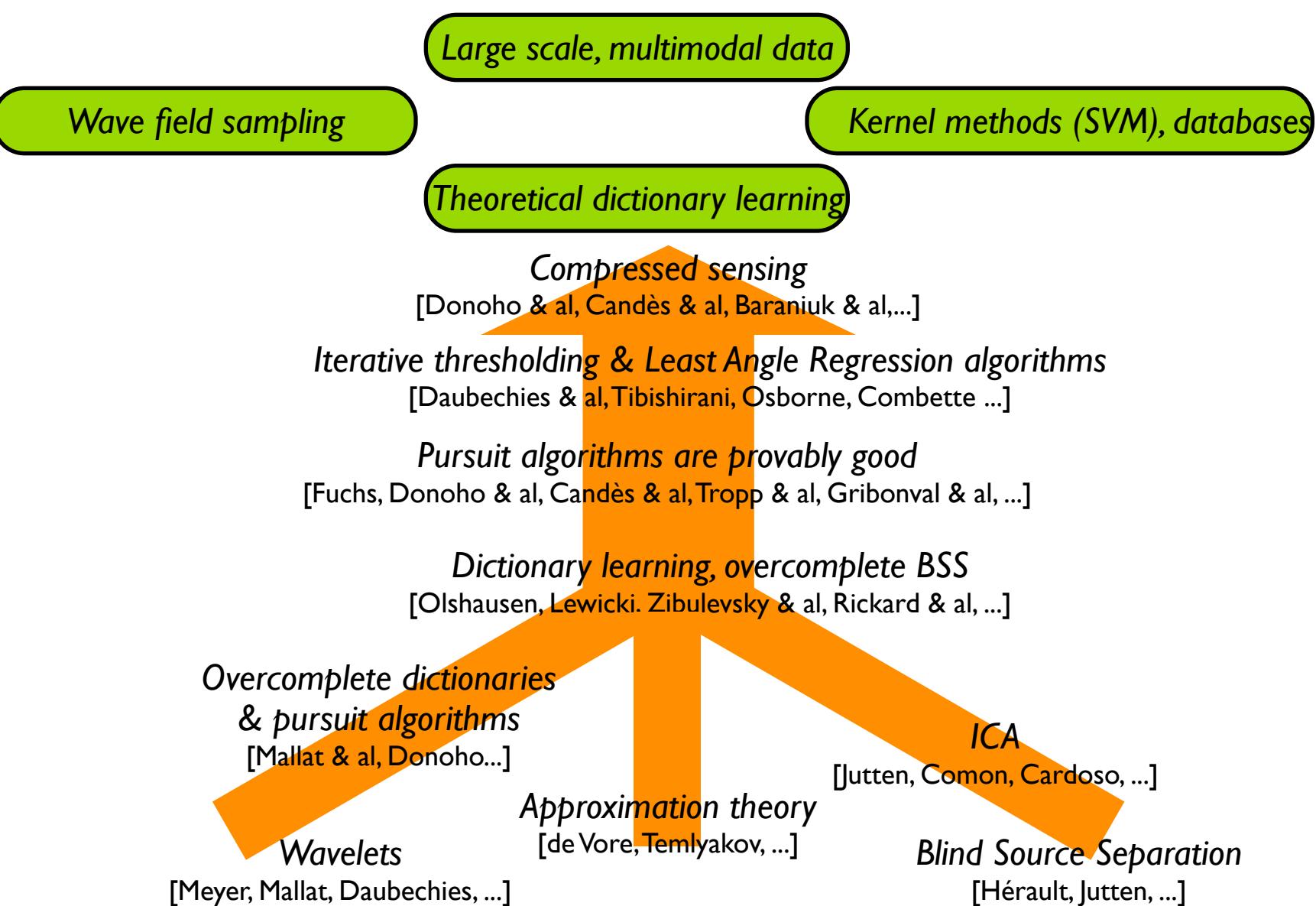
$$z = \mathbf{K}\mathbf{Ax}$$

$$\min \|x\|_1, \text{ subject to } z = \mathbf{K}\mathbf{Ax}$$



Reconstruction

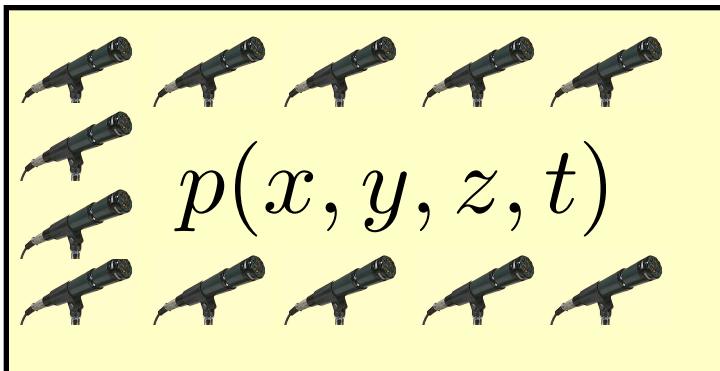




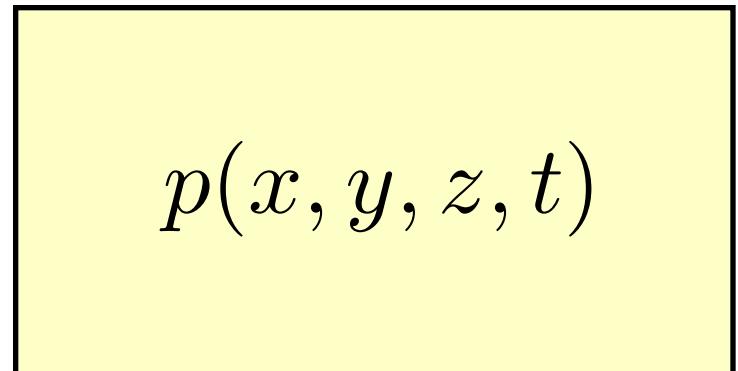
Perspectives & Challenges

- Co-design algorithm / compressed sensing device
 - ◆ which analog measurement?
 - ◆ calibration ?
 - ◆ efficiency & robustness (noise, loss of measures)

Today: pointwise microphone arrays



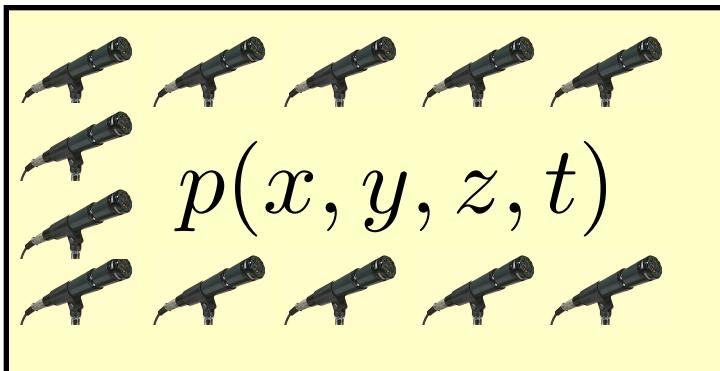
Tomorrow : acoustic field tomography?



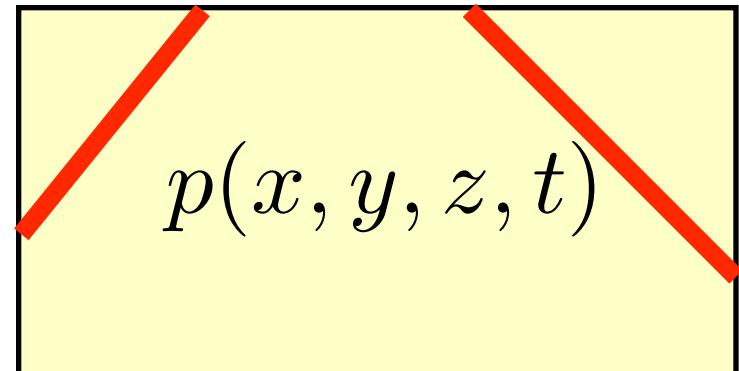
Perspectives & Challenges

- Co-design algorithm / compressed sensing device
 - ◆ which analog measurement?
 - ◆ calibration ?
 - ◆ efficiency & robustness (noise, loss of measures)

Today: pointwise microphone arrays

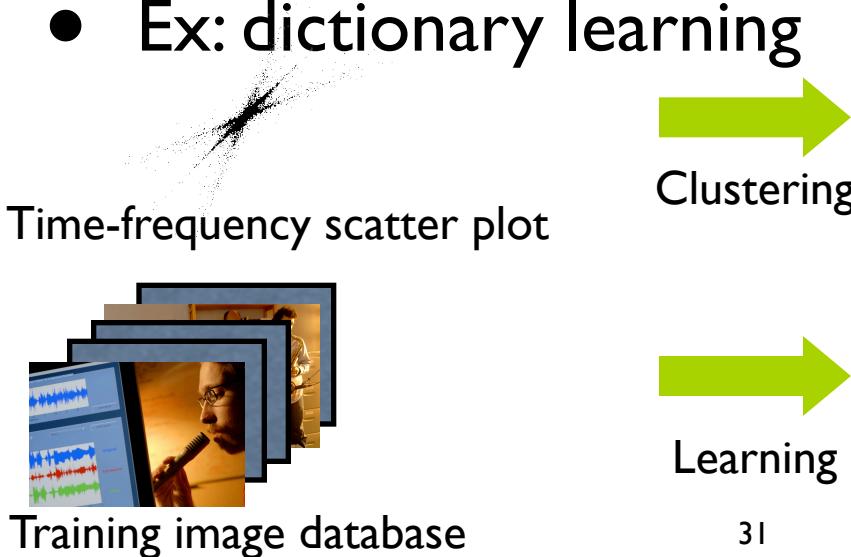


Tomorrow : acoustic field tomography?



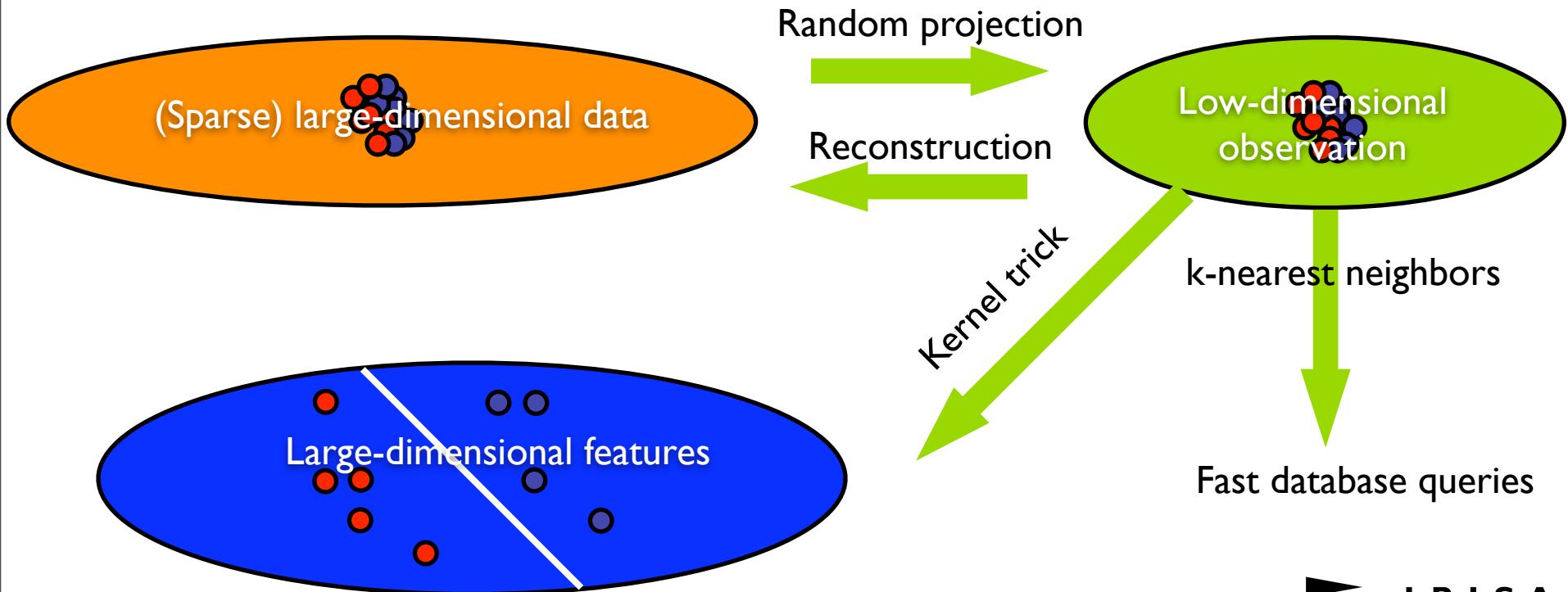
Perspectives & Challenges

- Large scale data (3D, multimodal, time-varying, ...)
 - ◆ efficiency : seismic data = 5 Terabytes / dataset!
 - ◆ models
 - ✿ dictionary learning
 - ✿ multimodal data
 - ✿ structured sparse representation + noise model, Bayesian models
- Ex: dictionary learning



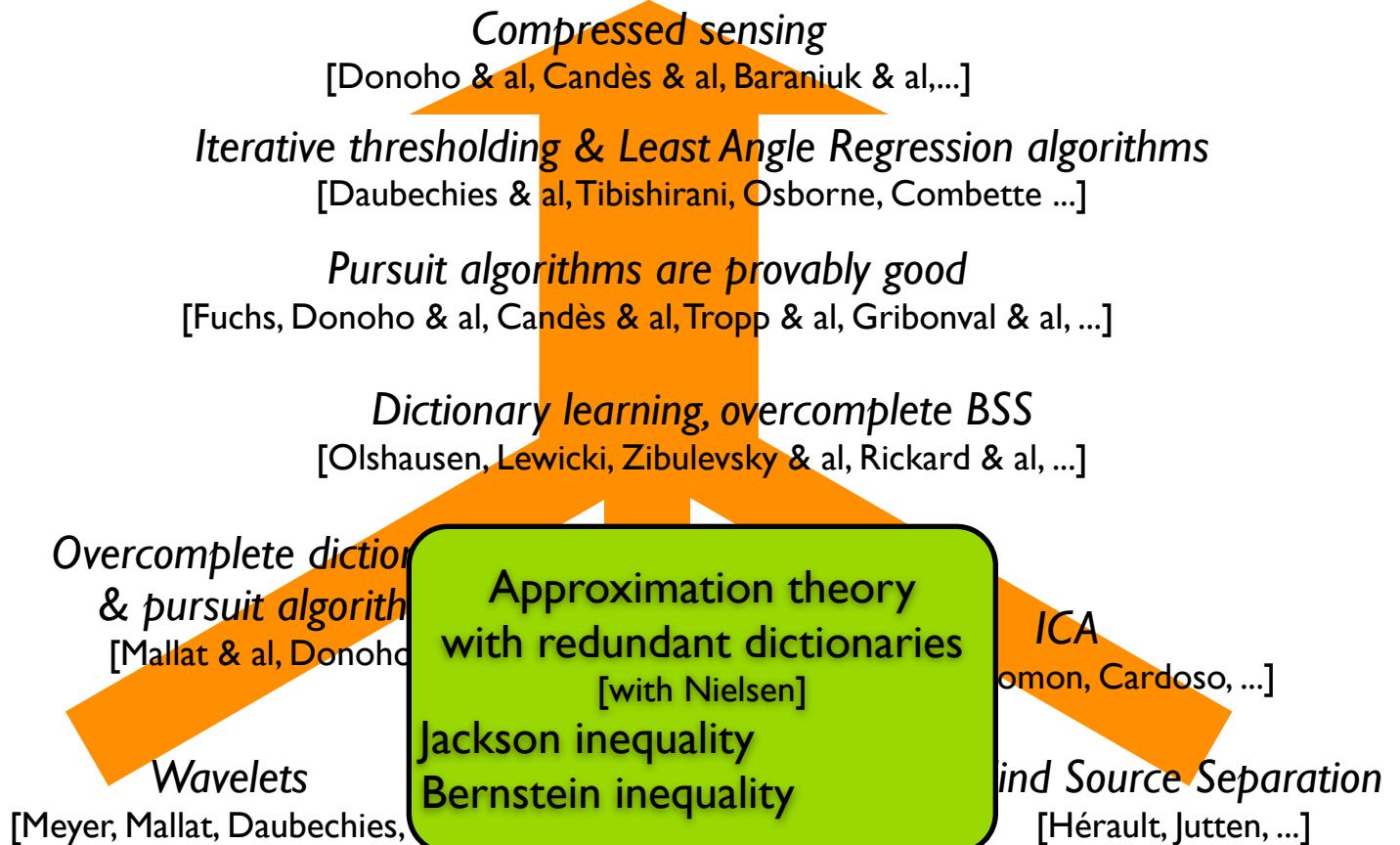
Perspectives & Challenges

- Signal processing in the compressed domain
- Links with kernel methods, databases, ...



Special thanks to

- Simon Arberet
- Laurent Benaroya
- Frédéric Bimbot
- Laurent Daudet
- Cédric Févotte
- Rosa Figueras i Ventura
- Marie-Noëlle Georgeault
- Gilles Gonon
- Guillaume Gravier
- Sacha Krstulovic
- Stéphanie Lemaile
- Sylvain Lesage
- Boris Mailhé
- Morten Nielsen
- Alexey Ozerov
- Karin Schnass
- Bruno Torrésani
- Pierre Vandergheynst
- Emmanuel Vincent
- ...



Theoretical Nonlinear Approximation

- Which signals are well approximated in a given dictionary, for a given algorithm ?
- Two descriptions

Representation
properties

$$\mathbf{b} = \mathbf{A}x, \quad \|x\|_p < C$$

Approximation
properties

$$\|\mathbf{b} - \mathbf{A}\hat{x}_M\|_2 \leq C'M^{-\alpha}$$

Theoretical Nonlinear Approximation

- Which signals are well approximated in a given dictionary, for a given algorithm ?
- Two descriptions

Representation properties

$$\mathbf{b} = \mathbf{A}x, \|x\|_p < C$$

[Stechkin, de Vore, Temlyakov]



Approximation properties

$$\|\mathbf{b} - \mathbf{A}\hat{x}_M\|_2 \leq C'M^{-\alpha}$$

- Orthonormal basis

$$\alpha = 1/p - 1/2$$

Theoretical nonlinear approximation

- Which signals are well approximated in a given dictionary, for a given algorithm ?
- Two descriptions

Representation properties

$$\mathbf{b} = \mathbf{A}x, \|x\|_p < C$$

[Gribonval & Nielsen]



Jackson inequality

Approximation properties

$$\|\mathbf{b} - \mathbf{A}\hat{x}_M\|_2 \leq C'M^{-\alpha}$$

- Overcomplete “Hilbertian” dictionary

$$\alpha = 1/p - 1/2$$

Theoretical Nonlinear Approximation

- Which signals are well approximated in a given dictionary, for a given algorithm ?
- Two descriptions

Representation properties

$$\mathbf{b} = \mathbf{A}x, \|x\|_p < C$$

[Gribonval & Nielsen]



Bernstein inequality

Approximation properties

$$\|\mathbf{b} - \mathbf{A}\hat{x}_M\|_2 \leq C'M^{-\alpha}$$

- Decomposable incoherent dictionary

$$\alpha = 2(1/p - 1/2)$$