Sparse Representations: from Source Separation to Compressed Sensing

Rémi Gribonval
METISS project-team
IRISA/INRIA

HDR Defense - October 24th, 2007
**Introduction**

Sparsity = old concept!
(wavelets, ...)

Natural / traditional role:
Sparsity = low cost (bits, computations, ...)
**direct goal**

Novel indirect role:
Sparsity = prior knowledge
Tool for inverse problems
Milestones

2000

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

Iterative thresholding & Least Angle Regression algorithms
[Daubechies & al, Tibishirani, 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, ...]

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, ...]

1990

2000
Compressed sensing
[Donoho & al, Candès & al, Baraniuk & al,...]

Iterative thresholding & Least Angle Regression algorithms
[Daubechies & al, Tibishirani, 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, ...]

Overcomplete dictionaries & pursuit algorithms
[Mallat & al, Donoho...]

Approximation theory
[de Vore, Temlyakov, ...]

Wavelets
[Meyer, Mallat, Daubechies, ...]

Blind Audio Source Separation
-more sources than sensors
(single channel, stereophonic)
[with Bimbot, Benaroya, Ozerov, P. Philippe, Arberet, Lesage]
-evaluation (performance measures, benchmarks)
[with Bimbot, Vincent, Fevotte, ...]
« 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

\[
\begin{pmatrix}
  y_{\text{right}}(t) \\
  y_{\text{left}}(t)
\end{pmatrix}
= A
\begin{pmatrix}
  s_1(t) \\
  s_2(t) \\
  s_3(t)
\end{pmatrix}
\]

- 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

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

Observed data $y(t) \approx A \cdot s(t)$ Unknown
Blind Source Separation: two-step approach

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

- **Estimate mixing matrix** $\hat{\mathbf{A}}$
  - Coarse source model (independence, sparsity, ...)

- **Estimate the sources** $\hat{s}(t) = \hat{\mathbf{A}}^{-1} \cdot \mathbf{y}(t)$

- 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

\[
\begin{pmatrix}
    y_{right}(t) \\
    y_{left}(t)
\end{pmatrix}
= A
\begin{pmatrix}
    s_1(t) \\
    s_2(t) \\
    s_3(t)
\end{pmatrix}
\]

- 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

\[
\begin{bmatrix}
y_{\text{right}}(t) \\
y_{\text{left}}(t)
\end{bmatrix}
= A
\begin{bmatrix}
s_1(t) \\
s_2(t) \\
s_3(t)
\end{bmatrix}
\]

- In practice...
Time-Frequency Masking

- Mixing model in the time-frequency domain

\[
\begin{align*}
Y_{\text{right}}(\tau, f) &\quad (\text{right}) \\
Y_{\text{left}}(\tau, f) &\quad (\text{left})
\end{align*}
\]

\[= A \ S(\tau, f)\]

- And “miraculously” ...

... time-frequency representations of audio signals are (often) **almost disjoint**.
Disjoint Time-Frequency Representations

frequency

time

Source 1
Source 2
Source 3
Disjoint Time-Frequency Representations

frequency

time

Source 1
Source 2
Source 3
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, Tibishirani, 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

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

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

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

\[ y(t) = \sum_{\tau,f} Y(\tau, f) g_{\tau,f}(t) \]

Time-frequency atom

zero = black
Multiscale Time-Frequency Structures

- Audio = superimposition of structures
- Example: glockenspiel

- transients = small scale
- harmonic part = large scale

- Gabor atoms \[ g_{s,\tau,f}(t) = \frac{1}{\sqrt{s}} w\left(\frac{t-\tau}{s}\right) e^{2i\pi ft} \]
Multiscale Time-Frequency Structures

- Audio = superimposition of structures
- Example: glockenspiel

- transients = small scale
- harmonic part = large scale

- Gabor atoms \( \left\{ g_{s,\tau,f}(t) = \frac{1}{\sqrt{s}} w \left( \frac{t - \tau}{s} \right) e^{2i\pi ft} \right\}_{s,\tau,f} \)
Multiscale Time-Frequency Structures

- Audio = superimposition of structures
- Example: glockenspiel

- Transients = small scale
- Harmonic part = large scale

- Gabor atoms

\[
\left\{ g_{s,\tau,f}(t) = \frac{1}{\sqrt{s}} w\left( \frac{t - \tau}{s} \right) e^{2i\pi ft} \right\}_{s,\tau,f}
\]
Multiscale Time-Frequency Structures

• Audio = superimposition of structures

• Example: glockenspiel

✦ transients = small scale
✦ harmonic part = large scale

• Gabor atoms
  \[ g_{s,\tau,f}(t) = \frac{1}{\sqrt{s}} w\left( \frac{t - \tau}{s} \right) e^{2i\pi ft} \]
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

\[ y(t) \approx b \approx A \cdot x \]

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

Known linear system: dictionary, (estimated) mixing matrix

Unknown
representation, sources, ...
Global Algorithms

• Approximation quality

\[ \| A x - b \|_2 \]

• Ideal sparsity measure: \( \ell^0 \) “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
  \[ \min_x \frac{1}{2} \| A x - b \|_2^2 + \lambda \| x \|_p \]
  - Sparse representation \( \lambda \to 0 \)
  - Sparse approximation \( \lambda > 0 \)

NP-hard combinatorial | FOCUSS / IRLS | Iterative thresholding | Linear
\[-\quad p = 0\quad \rightarrow \quad p = 1\quad \rightarrow \quad p = 2\]
Global Algorithms

- **Global optimization**
  \[
  \min_x \frac{1}{2} \| A x - b \|_2^2 + \lambda \| x \|_p
  \]

  - Sparse representation \( \lambda \to 0 \)
  - Sparse approximation \( \lambda > 0 \)

Lasso [Tibshirani 1996], Basis Pursuit (Denoising) [Chen, Donoho & Saunders, 1999]
Linear/Quadratic programming (interior point, etc.)

NP-hard combinatorial
FOCUSS / IRLS
Iterative thresholding
Linear

- Local minima
- Convex: global minimum

\( p = 0 \) \( p = 1 \) \( p = 2 \)
Matching Pursuits

- **Matching Pursuit Algorithm** [Mallat & Zhang 1993]
  - initialize residual \( b^{(0)} := b \)
  - find best atom \( k_m := \arg \max_k |\langle b^{(m-1)}, a_k \rangle| \)
  - update residual \( b^{(m)} := b^{(m-1)} - \langle b^{(m-1)}, a_{k_m} \rangle a_{k_m} \)
Matching Pursuits

- Matching Pursuit Algorithm \cite{Mallat:93}
  - initialize residual $b^{(0)} := b$
  - find best atom $k_m := \arg \max_k |\langle b^{(m-1)}, a_k \rangle|$
  - update residual $b^{(m)} := b^{(m-1)} - \langle b^{(m-1)}, a_{k_m} \rangle a_{k_m}$

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 \( b^{(0)} := b \)
  ✦ find best atom \( k_m := \arg \max_k |\langle b^{(m-1)}, a_k \rangle| \)
  ✦ update residual \( b^{(m)} := b^{(m-1)} - \langle b^{(m-1)}, a_{k_m} \rangle a_{k_m} \)

**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 1 hour audio signals instead of 5 seconds
✦ experiments on database of > 2000 songs
**Overcomplete dictionaries & pursuit algorithms**

- [Mallat & al, Donoho & al, Baraniuk & al, ...]

**Wavelets**

- [Meyer, Mallat, Daubechies, ...]

**Compressed sensing**

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

**Iterative thresholding & Least Angle Regression algorithms**

**Approximation theory**

- [de Vore, Temlyakov, ...]

**Blind Source Separation**

- [Hérault, Jutten, ...]

**Matching Pursuit**

- \( \ell^p \) optimization \( 0 \leq p \leq 1 \)

**Provably good algorithms**

- [with Nielsen, Vandergheynst, Rauhut, Schnass]

**ICA**

- [Jutten, Comon, Cardoso, ...]

**Overcomplete dictionaries & pursuit algorithms**

- [Mallat & al, Donoho & al, ...]

**Union of bases, incoherent dictionaries, structured dictionaries**
Sparsity and Ill-Posed Inverse Problems

• Ill-posedness if more unknowns than equations

\[ Ax_0 = Ax_1 \nRightarrow x_0 = x_1 \]

• Uniqueness of sparse solutions:
  ✦ if \( x_0, x_1 \) are “sufficiently sparse”,
  ✦ then \( Ax_0 = Ax_1 \Rightarrow x_0 = x_1 \)
Example: 1-sparse Representations

• Uniqueness of 1-sparse representations

\[ \mathbf{b} = \mathbf{A} \mathbf{x}_0 = \mathbf{A} \mathbf{x}_1 \]

\[ \mathbf{x}_0 = \mathbf{x}_1 \]

• 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 \[\text{[Natarajan, Davies \& al]}\]

\[
x^* := \arg \min_x \|x\|_0 \text{ subject to } Ax = b
\]

- If any 2M columns of \( A \) are linearly independent then

\[
Ax = Ay, \|x\|_0 \leq M, \|y\|_0 \leq M \quad \Rightarrow \quad x = y
\]

- Proof: \( \|x - y\|_0 \leq 2M \) and \( A(x - y) = 0 \)
Coherence of a Dictionary

- **Definition (easily computable)** \( \mu = \mu(A) := \max_{k \neq l} |\langle a_k, a_l \rangle| \)
- **Property** \( \|x\|_0 \leq 2M \Rightarrow \|Ax\|_2^2 \geq (1 - (2M - 1)\mu) \cdot \|x\|_2^2 \)

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

1. **Uniqueness of M-sparse representations whenever**
   \[ \|x\|_0 \leq M < \frac{1 + 1/\mu}{2} \]

2. **Recovery with Basis Pursuit & Matching Pursuit if**
   \[ \|x\|_0 \leq M < \frac{1 + 1/\mu}{2} \]

[with Nielsen, Vandergheynst & Figueras]

\( \ell^p \) optimisation for any \( 0 \leq p \leq 1 \)
Restricted Isometry Constants

- **Definition**: isometry constant $\delta_M$ is the smallest number such that

\[ \|x\|_0 \leq M \Rightarrow 1 - \delta_M \leq \frac{\|Ax\|_2^2}{\|x\|_2^2} \leq 1 + \delta_M \]

**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

\[ A^T A - \text{Id} \]

max over \( K(K-1) \) entries

\[ \mu = \mu(A) := \max_{k \neq l} |\langle a_k, a_l \rangle| \]

(Cumulative) coherence
Low cost
Coarse / pessimistic

\[ \delta_M := \sup_{\#I \leq M, c \in \mathbb{R}^M} \left| \frac{\|A_I c\|_2^2}{\|c\|_2^2} - 1 \right| \]

Isometry constants
Hard to compute
~Sharp

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

• Dirac-Fourier dictionary

\[ K = 2T \]

\[ T \]

\[ \delta_k(t) = \frac{1}{\sqrt{T}} e^{2i\pi kt/T} \]

• Coherence

\[ \mu = 1/\sqrt{T} \]

• “Generic” (random) dictionary

[Canèdes & al, Vershynin, ...]

\[ a_{tk} \sim P(a), \text{ i.i.d.} \]

• Isometry constants

if \[ T \geq CM \log K / M \]

then \[ P(\delta_{2M} + \delta_{3M} < 1) \approx 1 \]

Recovery by Basis Pursuit

\[ M_{\text{Basis Pursuit}}(A) \approx 0.914\sqrt{T} \]

\[ C' \gg 1 \]

\[ M_{\text{Basis Pursuit}}(A) \gtrsim C'T / \log^a(T) \]
Compressed Sensing

- MRI from incomplete measures

[Candès, Romberg & Tao]

Data → Lossy measurement = tomography

Measured data (FFT minus lost data) → Sparse L1 decomposition (Candès et al. 2004) → Reconstruction

FFT⁻¹
Compressed Sensing

- MRI from incomplete measures
  [Candès, Romberg & Tao]

\[ y = Ax \]
\[ z = KA \min \| x \|_1, \text{ subject to } z = KAx \]
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, ...]

Dictionary learning, overcomplete BSS
[Olshausen, Lewicki, Zibulevsky & al, Rickard & al, ...]

Pursuit algorithms are provably good
[Fuchs, Donoho & al, Candès & al, Tropp & al, Gribonval & al, ...]

Iterative thresholding & Least Angle Regression algorithms
[Daubechies & al, Tibishirani, Osborne, Combette ...]

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

Large scale, multimodal data

Wave field sampling

Kernel methods (SVM), databases

Theoretical dictionary learning
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

  Time-frequency scatter plot
  Clustering
  Mixing matrix
  [Ph.D. S. Arberet]

  Training image database
  Learning
  Dictionary of edge atoms
  [Ph.Ds S. Lesage & B. Mailhé]

IRISA
Perspectives & Challenges

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

(Sparse) large-dimensional data

Random projection

Reconstruction

Low-dimensional observation

Kernel trick

k-nearest neighbors

Fast database queries

Large-dimensional features
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
- ...
Overcomplete dictionaries & pursuit algorithms
[Mallat & al, Donoho & al, ...]

Wavelets
[Meyer, Mallat, Daubechies, ...]

Dictionary learning, overcomplete BSS
[Olshausen, Lewicki, Zibulevsky & al, Rickard & al, ...]

Iterative thresholding & Least Angle Regression algorithms
[Daubechies & al, Tibishirani, Osborne, Combette ...]

Pursuit algorithms are provably good
[Fuchs, Donoho & al, Candès & al, Tropp & al, Gribonval & al, ...]

Approximation theory
[de Vore, Temlyakov, ...]

Approximation theory with redundant dictionaries
[with Nielsen]

Jackson inequality
Bernstein inequality

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

ICA
[Hyvarinen, Jutten, ...]

Blind Source Separation
[Hérault, Jutten, ...]
Theoretical Nonlinear Approximation

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

\[ b = Ax, \|x\|_p < C \]

\[ \|b - A\hat{x}_M\|_2 \leq C' M^{-\alpha} \]
Theoretical Nonlinear Approximation

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

**Representation properties**

\[ b = Ax, \|x\|_p < C \]

**Approximation properties**

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

- 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

\[ b = Ax, \|x\|_p < C \]

Approximation properties

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

[Gribonval & Nielsen]

Jackson inequality

\[ \alpha = \frac{1}{p} - \frac{1}{2} \]

• Overcomplete “Hilbertian” dictionary
Theoretical Nonlinear Approximation

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

Representation properties

\[ b = Ax, \|x\|_p < C \]

Approximation properties

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

[Gribonval & Nielsen]

Bernstein inequality

\[ \alpha = 2(1/p - 1/2) \]

• Decomposable incoherent dictionary