Sparse dictionary learning
in the presence of noise & outliers

Rémi Gribonval
INRIA Rennes - Bretagne Atlantique, France
remi.gribonval@inria.fr
Overview

- Context: sparse signal processing
- Dictionary learning
- Statistical guarantees
- Flavor of the proof
- Conclusion
Sparse signal processing
Sparse Signal / Image Processing

denoising

inpainting

+ Compression, Source Localization, Separation, Compressed Sensing ...
Typical Sparse Models

- Audio: time-frequency representations (MP3)

- Images: wavelet transform (JPEG2000)

Black = zero

White = zero
Mathematical expression

• Signal / image = high dimensional vector
  \( x \in \mathbb{R}^d \)

• **Model** = linear combination of basis vectors
  (ex: *time-frequency atoms, wavelets*)

  \[ x \approx \sum_k z_k d_k = Dz \]
  (Dictionary of atoms
  (Mallat & Zhang 93)

• **Sparsity** = small \( L_0 \) (quasi)-norm

  \[ \| z \|_0 = \sum_k \left| z_k \right|^0 = \text{card}\{k, z_k \neq 0\} \]
Sparse models and inverse problems

Coefficient Domain

Signal Domain

Observation Domain

Sparse coefficient

Synthesis Dictionary $x = Dz$

Measurement System $y = Mx$
Acoustic Imaging

- **Ground truth: laser vibrometry**
  - ✓ direct optical measures
  - ✓ sequential
  - ✓ 2000 measures

- **Nearfield Acoustic Holography**
  - ✓ indirect acoustic measures
  - ✓ 120 microphones at a time
  - ✓ 120 x 16 = 1920 measures
  - ✓ *Tikhonov regularization*
Compressive Nearfield Acoustic Holography

- One shot with 120 micros
- Sparse regularization
Dictionary learning

with K. Schnass, F. Bach, R. Jenatton

small-project.eu
Sparse Atomic Decompositions

\[ x \approx Dz \]

Signal Image

(Overcomplete)
dictionary of atoms

Sparse Representation Coefficients
Data Deluge

• Sparsity: historically for signals & images
  ✓ bottleneck = **large-scale** algorithms

• New “exotic” or composite data
  ✓ bottleneck = **dictionary/operator** design/learning

Signals

Images

Hyperspectral
Satellite imaging

Spherical geometry
Cosmology, HRTF (3D audio)

Graph data
Social networks
Brain connectivity

Vector valued
Diffusion tensor

Signals

Images

Hyperspectral
Satellite imaging

Spherical geometry
Cosmology, HRTF (3D audio)

Graph data
Social networks
Brain connectivity

Vector valued
Diffusion tensor
A quest for the perfect sparse model

Training database

\[ \text{patch extraction} \]

Training patches

\[ x_n \]

\[ 1 \leq n \leq N \]

Unknown dictionary

Unknown sparse coefficients

\[ \hat{\mathbf{D}} = \text{edge-like atoms} \]

[Olshausen & Field 96, Aharon et al 06, Mairal et al 09, ...]

\[ = \text{shifts of edge-like motifs} \]

[Blumensath 05, Jost et al 05, ...]
Dictionary Learning
= Sparse Matrix Factorization

\[ X \approx DZ \]

\[ d \times N \quad d \times K \quad K \times N \]

with s-sparse columns
Many approaches

- Independent component analysis
  - [see e.g. book by Comon & Jutten 2011]

- Convex
  - [Bach et al., 2008; Bradley and Bagnell, 2009]

- Submodular
  - [Krause and Cevher, 2010]

- Bayesian
  - [Zhou et al., 2009]

- **Non-convex matrix-factorization**
  - [Olshausen and Field, 1997; Pearlmutter & Zibulevsky 2001, Aharon et al. 2006; Lee et al., 2007; Mairal et al., 2010 (... and many other authors)]
Sparse coding objective function

- **Given one training sample**: Basis Pursuit / LASSO

\[ f_{x_n}(D) = \min_{z_n} \frac{1}{2} \| x_n - Dz_n \|_2^2 + \lambda \| z_n \|_1 \]

- **Given N training samples**

\[ F_X(D) = \frac{1}{N} \sum_{n=1}^{N} f_{x_n}(D) \]

\[ \propto \min_Z \frac{1}{2} \| X - DZ \|_F^2 + \lambda \| Z \|_1 \]
Learning = constrained minimization

\[ \hat{D} = \arg \min_{D \in \mathcal{D}} F_x(D) \]

✓ Online learning with SPAMS library (Mairal & al)
✓ Constraint = dictionary with unit columns

\[ \mathcal{D} = \{ D = [d_1, \ldots, d_D], \ \forall k \ \|d_k\|_2 = 1 \} \]
Empirical findings
Numerical example (2D)

\[ X = D_0 Z_0 \]

Empirical observations

a) Global minima match angles of the original basis
b) There is no other local minimum.
Sparsity vs coherence (2D)

\[ \mu = |\cos(\theta_1 - \theta_0)| \]

Empirical probability of success

- ground truth=local min
- ground truth=global min
- no spurious local min

**Rule of thumb**: perfect recovery if:

a) Incoherence \( \mu < 1 - p \)
b) Enough training samples (N large enough)
Empirical findings

- **Stable & robust dictionary identification**
  - ✓ Global minima often match ground truth
  - ✓ Often, there is no spurious local minimum

- **Role of parameters?**
  - ✓ *sparsity* of $Z$?
  - ✓ *incoherence* of $D$?
  - ✓ *noise* level?
  - ✓ presence / nature of *outliers*?
  - ✓ *sample complexity* (number of training samples)?
Theoretical guarantees
Theoretical guarantees

- **Excess risk analysis** (~Machine Learning)
  - [Maurer and Pontil, 2010; Vainsencher et al., 2010; Mehta and Gray, 2012]
  
  \[ F_X(\hat{D}) - \min_D \mathbb{E}_X F_X(D) \]

- **Identifiability analysis** (~Signal Processing)
  - [Independent Component Analysis, e.g. book Comon & Jutten 2011]

\[ \| \hat{D} - D_0 \|_F \]

- **Array processing perspective**
  - Dictionary ~ directions of arrival
  - Identification ~ source localization

- **Neural coding perspective**:
  - Dictionaries ~ receptive fields
### Theoretical guarantees: overview

<table>
<thead>
<tr>
<th></th>
<th>[G. &amp; Schnass 2010]</th>
<th>[Geng &amp; al 2011]</th>
</tr>
</thead>
<tbody>
<tr>
<td>signal model</td>
<td><img src="signal_model.png" alt="Signal Model Image" /></td>
<td><img src="signal_model.png" alt="Signal Model Image" /></td>
</tr>
<tr>
<td>overcomplete (d&lt;K)</td>
<td>no</td>
<td>yes</td>
</tr>
<tr>
<td>outliers</td>
<td>yes</td>
<td>no</td>
</tr>
<tr>
<td>noise</td>
<td>no</td>
<td></td>
</tr>
<tr>
<td>cost function</td>
<td>( \min_{D, Z} |Z|_1 \ s.t. DZ = X )</td>
<td></td>
</tr>
</tbody>
</table>

**Theoretical guarantees**

\[
\min_{D, Z} \|Z\|_1 \quad \text{s.t.} \quad DZ = X
\]
Sparse Signal Model

- Random support

\[ J \subset [1, K], \quad \|J\| = s \]

- Sub-Gaussian iid coefficients, bounded below

\[ P(|z_i| < \bar{z}) = 0 \]

- Sub-Gaussian additive noise

\[ x = \sum_{i \in J} z_i d_i + \varepsilon = D_J z_J + \varepsilon \]
Local stability & robustness

• **Theorem 1: local stability** [Jenatton, Bach & G. 2012]
  ✓ Assumptions:
  ✦ overcomplete *incoherent* dictionary $\mathbf{D}_0$
  ✦ $s$-sparse sub-Gaussian coefficient model (no outlier)
  
  ✓ Conclusion:
  ✦ with high probability there exists a local minimum of $F_X(\mathbf{D})$ such that
  \[
  \|\mathbf{D} - \mathbf{D}_0\|_F \leq C \sqrt{sdK^3} \cdot \frac{\log N}{N}
  \]

• **Theorem 2: robustness to noise**
  ✓ technical assumption: bounded coefficient model

• **Theorem 3: robustness to outliers**

$s = \text{sparsity}$
$d = \text{signal dimension}$
$K = \text{number of atoms}$
$N = \text{number of samples}$
Learning Guarantees vs Empirical Findings

- Robustness to noise
- Sample complexity

**Predicted slope**

- Hadamard dictionary in dimension $d$
- Hadamard–Dirac dictionary in dimension $d$

Graphs showing the relative error versus noise level and number of training signals for different dictionary dimensions and initialization methods.
Flavor of the proof
Characterizing local minima (1)

- **Noiseless setting**
  - Minimum *exactly* at ground truth

\[ F_X(D) - F_X(D_0) \]

- **Noisy setting**
  - Minimum *close to* ground truth

\[ F_X(D) - F_X(D_0) \]

- Zero at ground truth
- Lower bound at radius \( r \)
Controlling the cost function

• **Problem**: \( F_X(D) \) sum of complicated functions!

\[
f_{x_n}(D) = \min_{z_n} \frac{1}{2} \| x_n - Dz_n \|_2^2 + \lambda \| z_n \|_1
\]

• **Solution**: simplified expression if sparse recovery

✦ adaptation from [Fuchs, 2005; Zhao and Yu, 2006; Wainwright, 2009]

\[
f_x(D) = \phi_x(D|\text{sign}(z_0)) \quad x = D_0z_0 + \varepsilon
\]

✓ Approximate cost function \( \Phi_X(D) \approx F_X(D) \)
Controlling the *approximate* cost function

**Problem:**
- Need *uniform* lower bound on the sphere \( \|D - D_0\|_F = r \)
  of the *random* function

\[
\Phi_X(D) - \Phi_X(D_0)
\]

**Solution:**
- Lower bound expectation for a given \( D \)
- Control Lipschitz constant (with high probability)
- Conclude with epsilon-net argument
Putting the pieces together

- **With high probability:**
  - lower-bound on approximate cost function
  - lower-bound on cost function

- **Outliers:** «no model» but total energy bounded

\[
\frac{1}{N} \sum_{n \in \text{outlier}} \|x_n\|_2^2 \leq c
\]

\[
\Phi_X(D) - \Phi_X(D_0)
\]

\[
F_X(D) - F_X(D_0)
\]
From local to global guarantees?

\[ \hat{D} = \arg \min_{D \in D} F_X(D) \]

- Ground truth = local min
- Ground truth = global min
- No spurious local min
To conclude ...
Summary

- **Sparse Dictionary Learning**
  - widely used in image processing and machine learning
  - from **heuristics** ...
    - online algorithms, empirically successful
  - ... to **statistics**
    - local stability and robustness guarantees
  - [http://hal.inria.fr/hal-00737152](http://hal.inria.fr/hal-00737152) [Jenatton, G. & Bach, Local stability and robustness of sparse dictionary learning in the presence of noise, Oct 2012]
What’s next?

● Immediate challenges
  ✓ global guarantees? empirically yes
  ✓ sharp sample complexity
  ✓ guarantees from cost functions to algorithms
    ✦ recent papers [link1], [link2], [link3]

● Sparse learning beyond dictionaries
  ✓ synthesis / analysis flavor (e.g. TV-like)
  ✓ structured models (shift-invariance, etc.)
  ✓ structured sparsity (e.g. trees, graphs)

● More examples = less work to learn?
THANKS

projection, learning and sparsity for efficient data processing