Compressive Source Separation: Algorithms and Applications

Mohammad Golbabaee
Simon Arberet
Pierre Vandergheynst

Ecole Polytechnique Fédérale de Lausanne (EPFL)

CIMI Workshop 2013, Toulouse
Motivation: MALDI Imaging

Molecular Mass Spectroscopy
Matrix-assisted laser desorption/ionization

Very high-dimensional: 3D X spectra!
High-Dimensional Multichannel Data

\[ X \in \mathbb{R}^{n_1 \times n_2} \]
Often very structured: spatial, spectral

\[ X = \Phi \Theta \]
Very special structure: (linear) mixture model
Very special structure: (linear) mixture model

\[ X = \mathbf{S} \mathbf{H}^T \]
Very special structure: (linear) mixture model

\[ X = S \cdot H^T \]

See for instance [Bioucas-Dias et al., 2012]
Very special structure: (linear) mixture model

\[ X = SHT^T \]

See for instance [Bioucas-Dias et al., 2012]
Very special structure: (linear) mixture model

\[ X = SH^T \]

See for instance [Bioucas-Dias et al., 2012]
Very special structure: (linear) mixture model

See for instance [Bioucas-Dias et al., 2012]
Very special structure: (linear) mixture model

\[ X = \Psi \Theta H^T \]

See for instance [Bioucas-Dias et al., 2012]
Maps to low-dimensional projections
Maps to low-dimensional projections

\[ Y = A(X) \]
Maps to low-dimensional projections
Maps to low-dimensional projections

\[ Y = A(SH^T) \]
Maps to low-dimensional projections
Maps to low-dimensional projections

\[ Y = A(\Psi \Theta H^T) \]
Maps to low-dimensional projections

\[ Y = A(\Psi \Theta H^T) \]

Questions:
- How many projections?
- Design of \( A \)?
Outline

• 2 Problems
  - Sparse regression = dictionary of spectra known
    • Is it interesting in some applications?
    • Can we use this information? Obtain theoretical guarantees?
  - Sparse coding = blind: learn spectra and abundances

• CS: observe projections to low dimension
  - can we directly recover model parameters?
  - can we use knowledge of spectra
CS of Multichannel Signals

Baseline: no structure \( X \in \mathbb{R}^{n_1 \times n_2} \) \( X_{vec} \in \mathbb{R}^{n_1 n_2} \)

\[
AX_{vec} := A(X) \quad y = AX_{vec} + z \quad A \in \mathbb{R}^{m \times n_1 n_2}
\]

\[
\arg\min_{X_{vec}} \|X_{vec}\|_1 \quad \text{s.t.} \quad \|y - AX_{vec}\|_2 \leq \varepsilon
\]

Recovery for \( K(\ll n_1 n_2) \)-sparse signals when \( m = \mathcal{O}(K \log(n_1 n_2/K)) \)

[Donoho, Candès-Romberg, Tao]
CS of Multichannel Signals

Typically impose spatial structure on sparsity model
CS of Multichannel Signals

Typically impose spatial structure on sparsity model

Via the dictionary: \( \Psi = \mathbb{I}_{n_2} \otimes \Psi_{2D} \)

\[
\arg \min_{\Theta_{vec}} \| \Theta_{vec} \|_1 \quad s.t. \quad \| y - A\Psi \Theta_{vec} \|_2 \leq \varepsilon,
\]
CS of Multichannel Signals

Typically impose spatial structure on sparsity model

Via the dictionary: \( \Psi = \mathbb{I}_{n_2} \otimes \Psi_{2D} \)

\[
\arg \min_{\Theta_{vec}} \|\Theta_{vec}\|_1 \quad s.t. \quad \|y - A\Psi \Theta_{vec}\|_2 \leq \varepsilon,
\]

\( k \) sparse per channel and RIP holds \( m \geq cn_2k \log(n_1/k) \)
CS of Multichannel Signals

Typically impose spatial structure on sparsity model

Via the dictionary: $\Psi = \mathbb{I}_{n_2} \otimes \Psi_{2D}$

$$\arg\min_{\Theta_{vec}} \|\Theta_{vec}\|_1 \quad s.t. \quad \|y - A\Psi\Theta_{vec}\|_2 \leq \varepsilon,$$

$k$ sparse per channel and RIP holds $m \geq c n_2 k \log(n_1/k)$
CS of Multichannel Signals

Typically impose spatial structure on sparsity model

Via the dictionary: \( \Psi = \mathbb{I}_{n_2} \otimes \Psi_{2D} \)

\[
\arg\min_{\Theta_{vec}} \|\Theta_{vec}\|_1 \quad s.t. \quad \|y - A\Psi\Theta_{vec}\|_2 \leq \varepsilon,
\]

\( k \) sparse per channel and RIP holds \( m \geq c n_2 k \log(n_1/k) \)

Not exploiting spectral redundancies

Via other (block structured) sparsity penalties ?

More structure ? [MMV, Davies-Eldar]
The Linear Mixture Case

\[ X = S H^T \]

\[ S \in \mathbb{R}^{n_1 \times \rho} \quad \text{Spatial abundance maps} \]

\[ H \in \mathbb{R}^{n_2 \times \rho} \quad \text{Spectra or endmembers} \]

Each channel is a mixture

\[ X_j = \sum_{i=1}^{\rho} [H]_{j,i} S_i \]

Typically the number of endmembers is very small compared to the spatial and spectral dimensions

\[ \rho \ll n_1, \rho \ll n_2 \]
Spectra as Side Information?

Assume we have a dictionary of spectra/endmembers

\[ \Phi = H \otimes \text{Id}_{n_1} \quad y = A\Phi S_{vec} + z \]
Spectra as Side Information?

Assume we have a dictionary of spectra/endmembers

\[ \Phi = H \otimes \text{Id}_{n_1} \quad \text{and} \quad y = A \Phi S_{\text{vec}} + z \]

New sensing operator
Spectra as Side Information?

Assume we have a dictionary of spectra/endmembers

\[ \Phi = H \otimes \text{Id}_{n_1} \quad y = A \Phi S_{vec} + z \]

New sensing operator

\[ S_{vec} = \Psi \Theta_{vec} \]
Spectra as Side Information?

Assume we have a dictionary of spectra/endmembers

\[
\Phi = \mathbf{H} \otimes \text{Id}_{n_1}
\]

\[
y = A\Phi \mathbf{S}_{vec} + \mathbf{z}
\]

New sensing operator

\[
\mathbf{S}_{vec} = \Psi \Theta_{vec}
\]

directly recover abundances
Spectra as Side Information?

Assume we have a dictionary of spectra/endmembers

$$\Phi = \mathbf{H} \otimes \text{Id}_{n_1}$$

$$y = A\Phi\mathbf{S}_{vec} + z$$

New sensing operator

$$\mathbf{S}_{vec} = \Psi \Theta_{vec}$$

directly recover abundances

$$\arg\min_{\Theta_{vec}} \|\Theta_{vec}\|_1 \quad \text{s.t.} \quad \|y - A\Phi\Psi\Theta_{vec}\|_2 \leq \varepsilon$$

How do we choose $A$? Influence of $\mathbf{H}$?

How can we use the knowledge of $\mathbf{H}$?
Our problem is of the form:

$$\arg\min_{\theta} \|\theta\|_1 \quad s.t. \quad \|y - AD\theta\|_2 \leq \varepsilon$$

Compressive Sensing with a coherent dictionary $D$

But ...
Fundamental Limits - 1

Our problem is of the form:

$$\arg \min_{\theta} \| \theta \|_1 \quad s.t. \quad \| y - AD\theta \|_2 \leq \varepsilon$$

Compressive Sensing with a coherent dictionary $D$

But ...

RIP on $AD$ will impose very strong restrictions on the underlying $H$: (recall $D = (H \otimes \text{Id}_{n_1})\Psi$)
Our problem is of the form:
\[
\arg\min_{\theta} \|\theta\|_1 \quad s.t. \quad \|y - A D \theta\|_2 \leq \varepsilon
\]

Compressive Sensing with a coherent dictionary \( D \)

But ...

RIP on \( AD \) will impose very strong restrictions on the underlying \( H \): (recall \( D = (H \otimes \text{Id}_{n_1}) \Psi \) )

\[
\delta_k(AD) < \sqrt{2} - 1 \implies \xi(H) < \sqrt{\sqrt{2} + 1}
\]

\[
\xi(H) = \frac{\sigma_{\text{max}}(H)}{\sigma_{\text{min}}(H)}
\]
Fundamental Limits - 2

Instead analyze with D-RIP for the sensing matrix
Instead analyze with D-RIP for the sensing matrix
\[
(1 - \delta_k^*) \|Dx\|_2^2 \leq \|ADx\|_2^2 \leq (1 + \delta_k^*) \|Dx\|_2^2
\]
[Candès, Eldar, Needel, Randall]
Instead analyze with D-RIP for the sensing matrix

\[(1 - \delta_k^*) \|Dx\|_2^2 \leq \|ADx\|_2^2 \leq (1 + \delta_k^*) \|Dx\|_2^2\]

[Candès, Eldar, Needel, Randall]

Control on D: \(\mathcal{L}_k(D) \|x\|_2 \leq \|Dx\|_2 \leq \mathcal{U}_k(D) \|x\|_2\)
Instead analyze with D-RIP for the sensing matrix

\[ (1 - \delta_k^*) \| Dx \|_2^2 \leq \| ADx \|_2^2 \leq (1 + \delta_k^*) \| Dx \|_2^2 \]

[Candès, Eldar, Needel, Randall]

Control on D: \( \mathcal{L}_k(D) \| x \|_2 \leq \| Dx \|_2 \leq \mathcal{U}_k(D) \| x \|_2 \)

If A satisfies the D-RIP for \( H \otimes \text{Id}_{n_1} \) with constant:

\( \delta_{\gamma',k} < 1/3 \) where \( \gamma' = 1 + 2\xi^2(H) \)

Then \( \| \Theta_{vec} - \hat{\Theta}_{vec} \|_2 \leq c'_0 k^{-1/2} \| \Theta_{vec} - (\Theta_{vec})_k \|_1 + c'_1 \varepsilon \)
Instead analyze with D-RIP for the sensing matrix

\[ (1 - \delta_k^*) \| \mathbf{D} \mathbf{x} \|_2^2 \leq \| \mathbf{A} \mathbf{D} \mathbf{x} \|_2^2 \leq (1 + \delta_k^*) \| \mathbf{D} \mathbf{x} \|_2^2 \]

[Candès, Eldar, Needel, Randall]

Control on \( \mathbf{D} \): \( \mathcal{L}_k(\mathbf{D}) \| \mathbf{x} \|_2 \leq \| \mathbf{D} \mathbf{x} \|_2 \leq \mathcal{U}_k(\mathbf{D}) \| \mathbf{x} \|_2 \)

If \( \mathbf{A} \) satisfies the D-RIP for \( \mathbf{H} \otimes \text{Id}_{n_1} \) with constant:

\[ \delta_{\gamma'k}^* < 1/3 \quad \text{where} \quad \gamma' = 1 + 2\xi^2(\mathbf{H}) \]

Then \( \| \Theta_{vec} - \hat{\Theta}_{vec} \|_2 \leq c'_0 k^{-1/2} \| \Theta_{vec} - (\Theta_{vec})_k \|_1 + c'_1 \varepsilon \)

Note that: \( \gamma' k < n_1 n_2 \Rightarrow \xi(\mathbf{H}) \leq \sqrt{\frac{n_1 n_2}{k} - 1} \)
“Decorrelation” sampling: \( A = \mathbf{H}^\dagger \otimes \tilde{A} \)

\[
y = A \Phi \mathbf{S}_{vec} + z \\
= (\mathbf{H}^\dagger \otimes \tilde{A}) (\mathbf{H} \otimes \text{Id}_{n_1}) \mathbf{S}_{vec} + z, \\
= (\text{Id}_\rho \otimes \tilde{A}) \mathbf{S}_{vec} + z.
\]

The analysis is then standard since \( \mathbf{H} \) has disappeared.

Constants will not depend on \( \mathbf{H} \), but effect on noise.
With “decorrelation” sampling, the effect of $H$ disappears.

<table>
<thead>
<tr>
<th>CS Acquisition Scheme</th>
<th>Dense</th>
<th>Dense</th>
<th>Uniform</th>
<th>Decorrelating</th>
</tr>
</thead>
<tbody>
<tr>
<td>CS Recovery Approach</td>
<td>BPDN</td>
<td>SS-$\ell_1$</td>
<td>SS-$\ell_1$</td>
<td>SS-$\ell_1$</td>
</tr>
<tr>
<td>CS measurements $m \gtrsim$</td>
<td>$O(n_2 k \log(n_1/k))$</td>
<td>$O(k \log(\rho n_1/k))$</td>
<td>$O(n_2 k \log(n_1/k))$</td>
<td>$O(k \log(\rho n_1/k))$</td>
</tr>
<tr>
<td>Constant depends on $H$</td>
<td>-</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
</tr>
</tbody>
</table>

- first case: $n_2 n_1$ matrix with sparsity $k$ on spatial dimension
- second case: mixture model: reduces sparsity by taking into account all channels and dense sampling uses it. But attention to constant
- third case: each channel separately and each channel is $k$ sparse, so roughly $kn_2$ sparse
- fourth case: mixture model reduces sparsity, decorrelation step removes effect of $H$.  

August 21, 2012

DRAFT

no mixing structure

effect of $H$
Applications

Full problem incorporates more constrains:
Applications

Full problem incorporates more constrains:

\[ \sum_{j=1}^{\rho} [S]_{i,j} = 1 \quad \forall i \in \{1, \ldots, n_1\} \quad [S]_{i,j} \geq 0 \]
Applications

Full problem incorporates more constrains:

\[ \sum_{j=1}^{\rho} [S]_{i,j} = 1 \quad \forall i \in \{1, \ldots, n_1\} \quad [S]_{i,j} \geq 0 \]

\[ \arg \min_{\Theta} \| \Theta_{vec} \|_1 \]

subject to \( \| y - A \Phi \Psi \Theta_{vec} \|_2 \leq \varepsilon \)

\( \Psi_{2D} \Theta I_\rho = I_{n_1} \)

\( \Psi \Theta_{vec} \geq 0. \)
Applications

Full problem incorporates more constrains:

\[ \sum_{j=1}^{\rho} [S]_{i,j} = 1 \quad \forall i \in \{1, \ldots, n_1\} \quad [S]_{i,j} \geq 0 \]

\[ \arg \min_{\Theta} \| \Theta_{vec} \|_1 \]

subject to \[ \| y - A\Phi \Psi \Theta_{vec} \|_2 \leq \varepsilon \]
\[ \Psi_{2D} \Theta \mathbb{I}_\rho = \mathbb{I}_{n_1} \]
\[ \Psi \Theta_{vec} \geq 0. \]

\[ \arg \min_{S} f_1(S) + f_2(S) + f_3(S) \]

\[ f_1(S) = \mathcal{P}(S), \quad f_2(S) = i_{\mathbb{B}_2}(S), \quad f_3(S) = i_{\mathbb{B}_{\triangle^+}}(S) \]
Hyper Spectral Imaging

\[ S: \text{Sources (element abundances)} \quad S \in \mathbb{R}^{n_1 \times \rho} \]
**Hyper Spectral Imaging**

\[ \mathbf{S}: \text{Sources (element abundancies)} \quad \mathbf{S} \in \mathbb{R}^{n_1 \times \rho} \]

\[ \mathbf{A}: \text{Spectra (depending on modality)} \quad \mathbf{A} \in \mathbb{R}^{n_2 \times \rho} \]
Hyper Spectral Imaging

\( S: \) Sources (element abundancies) \( S \in \mathbb{R}^{n_1 \times \rho} \)

\( A: \) Spectra (depending on modality) \( A \in \mathbb{R}^{n_2 \times \rho} \)

Each pixel is a weighted combination of source spectra: \( y = SA^T \)
Some experiments

- We have implemented various problems
  - with/without linear mixture model
  - simple sparse wavelet model, TV
- We compared different algorithms
  - PPXA, a variant with IHT, ...
- We used several sensing matrices
  - dense, uniform, decorr, varied the core matrix (random conv, ...)
- We compared on various datasets
  - synthetic, real, CASSI ...
Results

(a) Reconstruction SNR vs. subsampling ratio (noiseless sampling)

(b) Reconstruction SNR vs. sampling SNR (subsampling ratio: 1/16)

<table>
<thead>
<tr>
<th>Noise SNR</th>
<th>Sampling rate</th>
<th>+∞ dB</th>
<th>30 dB</th>
<th>10 dB</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1/4</td>
<td>1/8</td>
<td>1/16</td>
<td>1/32</td>
</tr>
<tr>
<td>SS-IHT (dense sampling)</td>
<td>0.69</td>
<td>0.61</td>
<td>0.57</td>
<td>0.48</td>
</tr>
<tr>
<td>SS-IHT (uniform sampling)</td>
<td>0.43</td>
<td>0.38</td>
<td>0.31</td>
<td>0.25</td>
</tr>
<tr>
<td>SS-TV (dense sampling)</td>
<td>1.0</td>
<td>1.0</td>
<td>0.95</td>
<td>0.81</td>
</tr>
<tr>
<td>SS-TV (uniform sampling)</td>
<td>0.97</td>
<td>0.73</td>
<td>0.43</td>
<td>0.31</td>
</tr>
</tbody>
</table>

We evaluate the different methods, for different sampling rates (Fig. 1(a)), and different noise levels. We acknowledge Xavier Gigandet and Meritxell Bach Cuadra for providing this ground truth map.
Results

You can download all the (GPU accelerated) code, datasets and scripts for experiments on our web page. Code uses UnlocBox: http://unlocbox.sourceforge.net

Have fun :)

<table>
<thead>
<tr>
<th>Method</th>
<th>10</th>
<th>20</th>
<th>30</th>
<th>40</th>
<th>50</th>
<th>60</th>
<th>70</th>
<th>80</th>
<th>90</th>
<th>100</th>
</tr>
</thead>
<tbody>
<tr>
<td>SS-IHT (dense)</td>
<td>0.99</td>
<td>0.98</td>
<td>0.96</td>
<td>0.94</td>
<td>0.92</td>
<td>0.91</td>
<td>0.94</td>
<td>0.96</td>
<td>0.98</td>
<td>0.99</td>
</tr>
<tr>
<td>SS-l1 (dense)</td>
<td>1.0</td>
<td>1.0</td>
<td>0.95</td>
<td>0.91</td>
<td>0.86</td>
<td>0.82</td>
<td>0.82</td>
<td>0.82</td>
<td>0.82</td>
<td>0.82</td>
</tr>
<tr>
<td>SS-TV (dense)</td>
<td>1.0</td>
<td>1.0</td>
<td>0.96</td>
<td>0.92</td>
<td>0.88</td>
<td>0.84</td>
<td>0.84</td>
<td>0.84</td>
<td>0.84</td>
<td>0.84</td>
</tr>
<tr>
<td>SS-IHT (uniform)</td>
<td>0.97</td>
<td>0.97</td>
<td>0.96</td>
<td>0.94</td>
<td>0.92</td>
<td>0.91</td>
<td>0.94</td>
<td>0.97</td>
<td>0.99</td>
<td>0.99</td>
</tr>
<tr>
<td>SS-l1 (uniform)</td>
<td>1.0</td>
<td>1.0</td>
<td>0.98</td>
<td>0.96</td>
<td>0.94</td>
<td>0.92</td>
<td>0.94</td>
<td>0.97</td>
<td>0.99</td>
<td>0.99</td>
</tr>
<tr>
<td>SS-TV (uniform)</td>
<td>1.0</td>
<td>1.0</td>
<td>0.98</td>
<td>0.96</td>
<td>0.94</td>
<td>0.92</td>
<td>0.94</td>
<td>0.97</td>
<td>0.99</td>
<td>0.99</td>
</tr>
<tr>
<td>SS-IHT-decorr</td>
<td>0.99</td>
<td>0.98</td>
<td>0.96</td>
<td>0.94</td>
<td>0.92</td>
<td>0.91</td>
<td>0.94</td>
<td>0.97</td>
<td>0.99</td>
<td>0.99</td>
</tr>
<tr>
<td>SS-l1-decorr</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>0.99</td>
</tr>
<tr>
<td>SS-TV-decorr</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>0.99</td>
</tr>
</tbody>
</table>
Hyper-Spectral Imaging

(a) Ground truth

From 3% of the original data:

(e) SS-TV-decorr, source reconstruction SNR: 8.64 dB
Blind Case

Very high dimensional data (thousands of channels)

Typical problem: **Source Separation**

Given the dictionary of spectra $\mathbf{A}$ and the data $\mathbf{y}$

$\mathbf{V}$ Recover the source abundances, factorizing $\mathbf{y} = \mathbf{S}\mathbf{A}^T$
**Blind Case**

Very high dimensional data (thousands of channels)

Typical problem: **Source Separation**

Given the dictionary of spectra $A$ and the data $y$

$y = S A^T$

Recover the source abundances, factorizing $y = S A^T$

Can this be achieved even when?

$y = A(SA^T)$ Indirect/degraded observations

The dictionary of spectra $A$ is unknown
Blind Source Separation

HSI Compressive Blind Source Separation (CS-BSS)

\[
\arg\min_{S, A} ||y - A(SA^T)||_{\ell_2}^2
\]

subject to

\[
\sum_j \|S_j\|_{TV} \leq \tau
\]

\[
\sum_{i=1}^{\rho} [S]_{i,j} = 1
\]

\[
[S]_{i,j} \geq 0
\]

\[
\|\Psi_{1D}^T A\|_{\ell_1} \leq \gamma
\]

\[
[A]_{i,j} \geq 0.
\]

Source image constraints
Spectral signature constraints

Goal: Compute source maps directly from compressed measurements
Separation or Segmentation in the compressed domain
Blind Separation
Blind Separation

S-est1

S-est2

A-est

S-est3

S-est4

Reconstrunction Error

S-est5

S-est6
Outlook

- Significant challenges ahead in signal processing
  - Big Data
  - Ubiquitous but Cheap Sensing (i.e. dirty signals)

- Sometimes, no need to reconstruct
  - clustering
  - classification

- Methods that would allow principled and guaranteed task-based processing very appealing