Dictionary learning for sparse representations
A statistical analysis using L1 minimization

Rémi Gribonval
METISS team (audio signal processing, indexing, source separation)
INRIA, Rennes, France

Karin Schnass
LTS2, EPFL, Switzerland

Journées STAR 2009
IRISA, 22-23 octobre 2009
Outline

1. Preliminaries:
   ✦ sparsity & (overcomplete) dictionaries of atoms
   ✦ blind source separation & dictionary learning

2. Objectives of (theoretical) dictionary learning

3. L1 minimization for dictionary learning

4. Main results
   ✦ geometric “local” identifiability condition
   ✦ random model and finite sample size analysis

5. Discussion, conclusion & challenges
Sparse data representations

- Short Time Fourier Transform (for time-varying harmonic signals)

- Wavelet transform (for piecewise smooth images)

Black = zero

Grey = zero
Sparse signal models

- An image / a signal = sum of few atoms $a_k$

$$b = \sum_k x_k a_k = A x$$

- Sparsity of $x$: enables compression, separation ...

- Sparsity of $x$? Only if dictionary $A$ is “well chosen”
  - Pre-chosen atoms: wavelets, Gabor, etc.
  - Learned dictionary = from collection of signals / images

$$b_n = A x_n, \ 1 \leq n \leq N$$
Dictionary learning for sparse representations

- Sparse modeling: choose a dictionary

Training image database
Dictionary learning for sparse representations

- Sparse modeling: choose a dictionary

Training patches

\[ b_n = A x_n, \ 1 \leq n \leq N \]

Training image database
Dictionary learning for sparse representations

- Sparse modeling: choose a dictionary

Training image database

Training patches

\[ b_n = A \alpha_n, \quad 1 \leq n \leq N \]

Unknown dictionary

Unknown sparse coefficients
Dictionary learning for sparse representations

- Sparse modeling: choose a dictionary

Training image database → patch extraction

Training patches

\[ b_n = \hat{A} \alpha_n, \quad 1 \leq n \leq N \]

Unknown dictionary

Unknown sparse coefficients

\[ \hat{A} = \text{edge-like atoms} \quad [\text{Olshausen \\ & Field 96}] \]

\[ = \text{shifts of edge-like motifs} \quad [\text{Jost, Vanderheynst, Lesage \\ & Gribonval 2005}] \]
Dictionary learning?

• Problem: estimate a matrix $A$ given observed samples

$$b_n = Ax_n, \ 1 \leq n \leq N$$
Dictionary learning?

- Problem: estimate a matrix $A$ given observed samples

$$b_n = A x_n, \ 1 \leq n \leq N$$
Dictionary learning?

- Problem: estimate a matrix $A$ given observed samples

\[ b_n = A x_n, \quad 1 \leq n \leq N \]

\[ B = AX \]
Dictionary learning?

- Problem: estimate a matrix $A$ given observed samples

$$b_n = Ax_n, \ 1 \leq n \leq N$$

$A$:
- Unknown mixing matrix (blind source separation)
- Unknown dictionary (sparse signal approximation)
- Unknown channel filter (blind channel estimation) ...
Dictionary learning?

- Problem: estimate a matrix $A$ given observed samples

\[ b_n = A x_n, \ 1 \leq n \leq N \]

$B = AX$

$A \{ \text{Unknown mixing matrix (blind source separation)} \}$

$A \{ \text{Unknown dictionary (sparse signal approximation)} \}$

$A \{ \text{Unknown channel filter (blind channel estimation)} \}$ ...

$X \{ \text{Unknown sources / signal representations / ...} \}$
Dictionary learning?

• Problem: estimate a matrix $A$ given observed samples $b_n = Ax_n$, $1 \leq n \leq N$.

$B = AX$

$A$ \{ Unknown mixing matrix (blind source separation) \\
Unknown dictionary (sparse signal approximation) \\
Unknown channel filter (blind channel estimation) \ ...

$X$ Unknown sources / signal representations / ... 

• Fundamentally ill-posed factorization problem: need (weak) model on unknown coefficients $X$ and / or matrix $A$. 

jeudi 22 octobre 2009
## Theoretical dictionary learning

- Problem: estimate a matrix $A$ given samples

$$b_n = A x_n, \quad 1 \leq n \leq N \quad \quad B = A X$$

<table>
<thead>
<tr>
<th>Model of ...</th>
<th><strong>ICA</strong> (Independent Component Analysis)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Assumption</td>
<td>independence</td>
</tr>
<tr>
<td></td>
<td>$p(X) = \prod_{nk} p(x_n(k))$</td>
</tr>
<tr>
<td>Identifiability</td>
<td>Darmois theorem</td>
</tr>
<tr>
<td>Identification</td>
<td>Contrast functions</td>
</tr>
<tr>
<td></td>
<td>$A \sim \hat{W}^{-1} \quad \hat{W} := \arg \min_W E_X(f(WAX))$</td>
</tr>
<tr>
<td>Issues</td>
<td>In practice: <strong>finite</strong> training sets</td>
</tr>
<tr>
<td></td>
<td>expectation $\rightarrow$ sample average</td>
</tr>
</tbody>
</table>

$\hat{W}$
Theoretical dictionary learning

- Problem: estimate a matrix $A$ given samples
  $$b_n = Ax_n, \ 1 \leq n \leq N$$
  $$B = AX$$

<table>
<thead>
<tr>
<th></th>
<th>ICA (Independent Component Analysis)</th>
<th>SCA (Sparse Component Analysis)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model of ...</td>
<td>probability density function $p(X)$</td>
<td>sample matrix $X$</td>
</tr>
<tr>
<td>Assumption</td>
<td>Independence</td>
<td>Sparsity / geometry</td>
</tr>
<tr>
<td></td>
<td>$p(X) = \prod_{nk} p(x_n(k))$</td>
<td>$\star$ many zeroes in $X$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\star$ $x_n$ and $b_n$ concentrate around union of low dimensional subspaces</td>
</tr>
<tr>
<td>Identifiability</td>
<td>Darmois theorem</td>
<td>[Georgiev, Theis &amp; Cichocki 05]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>[Aharon, Elad &amp; Bruckstein 06]</td>
</tr>
<tr>
<td>Identification</td>
<td>Contrast functions</td>
<td>Combinatorial algorithms</td>
</tr>
<tr>
<td></td>
<td>$A \sim \hat{W}^{-1}$ $\hat{W} := \arg\min_{W} \mathbb{E}_X(f(WAX))$</td>
<td></td>
</tr>
<tr>
<td>Issues</td>
<td>In practice: finite training sets</td>
<td>Identifiability assumes:</td>
</tr>
<tr>
<td></td>
<td>expectation $\rightarrow$ sample average</td>
<td>$\star$ highly sparse coefficients</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\star$ (combinatorially ?) many training examples</td>
</tr>
</tbody>
</table>
Objectives

• Long term goal:
  ✦ identifiability conditions on $X$ to recover $A$ from $B = AX$
  ✦ provably good + efficient identification algorithms

• Focus: exploit sparsity of $X$

• Desirable features
  ✷ geometric understanding of identifiability conditions
  ✷ robustness to “weakly-sparse” data
  ✷ identifiability with limited number of training samples
  ✷ non-combinatorial algorithms

Exactly sparse data (purely academic)  Example of real data = “weakly” sparse
L1 minimization for dictionary learning
Holy grail: provably good + efficient sparse learning

- **Sparse representations**
  - Known matrix \( A \)
  - Data model \( b = Ax_0 \)
  - Identifiability theorems:
    \[
    \|x_0\|_0 \leq k_1(A)
    \]
  - Much literature since 2001
    (Donoho & Huo, Elad & Bruckstein, Gribonval & Nielsen, Candès & Romberg & Tao, Tropp, Donoho & Tanner, ... and many others)

- **Dictionary learning**
  - Unknown matrix \( A_0 \)
  - Data model \( B = A_0X_0 \)
  - Identifiability theorem?
    \[
    A_0, X_0 \in ?
    \]
  - Most literature on Independent Component Analysis (ICA), density model rather than finite sample size geometric model
Numerical example

- Cloud of 2500 training samples in $\mathbb{R}^2$
  - ~1000 sparse [= on axes]
  - ~1500 non-sparse
Numerical example

- Cloud of 2500 training samples in $\mathbb{R}^2$
  - $\sim 1000$ sparse [= on axes]
  - $\sim 1500$ non-sparse
- Orthonormal basis
  - Angle $\theta \leftrightarrow A_\theta = [a_1(\theta), a_2(\theta)]$
Numerical example

- Cloud of 2500 training samples in $\mathbb{R}^2$
  - $\sim 1000$ sparse [on axes]
  - $\sim 1500$ non-sparse

- Orthonormal basis
  - Angle $\theta \leftrightarrow A_\theta = [a_1(\theta), a_2(\theta)]$

- L1 criterion
  $$\| A_\theta^{-1} A_0 X \|_1$$
Numerical example

- Cloud of 2500 training samples in $\mathbb{R}^2$
  - $\sim$1000 sparse [on axes]
  - $\sim$1500 non-sparse
- Orthonormal basis
  - Angle $\theta \leftrightarrow A_\theta = [a_1(\theta), a_2(\theta)]$
- $\ell_1$ criterion
  - $\left\| A_\theta^{-1} A_0 X \right\|_1$
  - global optimum = original
  - no other local minimum
Numerical example

Non orthogonal bases

\[ \mathbf{A}_{\theta_1, \theta_2} \]

\[ \| \mathbf{A}_{\theta_1, \theta_2}^{-1} \mathbf{A}_{0} \mathbf{X} \|_1 \]

L1 criterion for oblique bases
Numerical example

Non orthogonal bases

\[ A_{\theta_1, \theta_2} \]

\[ \| A_{\theta_1, \theta_2}^{-1} A_{0, X} \|_1 \]

L1 criterion for oblique bases

jeudi 22 octobre 2009
Numerical example

Non orthogonal bases

\[ A_{\theta_1, \theta_2} \]

\[ \| A_{\theta_1, \theta_2}^{-1} A_0 X \|_1 \]

Empirical observations

a) Global minima match the original basis
b) There is no other local minimum.

jeudi 22 octobre 2009
Theoretical results

1. “Local identifiability” for (non overcomplete) L1 dictionary learning
   ✦ algebraic / geometric characterization of local minima

2. Probability of identifiability
   ✦ model on X: random, weakly-sparse
   ✦ analysis of identifiability for (small) finite sample size
Local identifiability result

- **Two assumptions:**
  - $X$: for each row $k$, up to column permutation, has decomposition
  
  \[
  X = \begin{bmatrix}
  S_k & 0 \\
  X_k & \bar{X}_k
  \end{bmatrix}
  \]
  
  and there exists $d_k$, $\|d_k\|_\infty < 1$, 
  \[
  X_k s_k^T = \bar{X}_k d_k
  \]

- **Conclusion:**
  - $A_0 = \text{local minimum}$ of $L_1$ among (not necessarily orthonormal) bases
  
  \[
  (A', X') \approx (A_0, X) \\
  A'X' = A_0X
  \]

  \[
  \|X'\|_1 \geq \|X\|_1
  \]
Trivial example

- **Two assumptions:**
  - $\bar{X}$: for each row $k$, up to column permutation, has decomposition
  - $X = \begin{bmatrix} S_k & 0 \\ X_k & \bar{X}_k \end{bmatrix}$

- If $X$ has at most one nonzero entry per column (at unknown positions)

- Simply choose

- How robust is the condition to weakly-sparse outliers?
- How many samples $N$ does it then typically require?

and there exists $d_k$, $\|d_k\|_\infty < 1$, $X_k s_k^T = \bar{X}_k d_k$

- $A_0 = \text{basis of sufficiently incoherent unit atoms}$

$\forall k \|a_k\|_2 = 1$, $\max_{k \neq l} |\langle a_k, a_l \rangle| \ll 1$
Trivial example

- **Two assumptions:**
  - $X$: for each row $k$, up to column permutation, has decomposition
  - $X = S_k \circ \Lambda_k \oplus 0$,

  and there exists $d_k, \|d_k\|_\infty < 1$,
  
  $0 = X_k S_k^T = \bar{X}_k d_k$

  - $A_0 = \text{basis of sufficiently incoherent unit atoms}$

  $\forall k \|a_k\|_2 = 1 \quad \max_{k \neq l} |\langle a_k, a_l \rangle| \ll 1$

- If $X$ has at most one nonzero entry per column (at unknown positions)

  - Simply choose

- **How robust is the condition to weakly-sparse outliers?**
- **How many samples $N$ does it then typically require?**
Trivial example

• **Two assumptions:**
  ✦ \( \bar{X} \) : for each row \( k \), up to column permutation, has decomposition
  ✦ \( \bar{X} = \begin{bmatrix} S_k & 0 \\ \hline \Lambda_k & \bar{\Lambda}_k \end{bmatrix} \)

  \[
  \begin{align*}
  X = &
  \begin{bmatrix}
  X_k \\
  \hline \\
  \Lambda_k & \bar{\Lambda}_k
  \end{bmatrix}
  \end{align*}
  \]

  and there exists \( d_k, \|d_k\|_{\infty} < 1 \),
  \[
  0 = X_k \bar{s}_k^T \bar{s}_k = \bar{X}_k d_k
  \]

  ✦ \( \Lambda_0 = \text{basis of sufficiently incoherent unit atoms} \)
  \[
  \forall k \|a_k\|_2 = 1 \quad \max_{k \neq l} |\langle a_k, a_l \rangle| \ll 1
  \]

• If \( X \) has at most one nonzero entry per column (at unknown positions)
  ✦ **Simply choose** \( d_k = 0 \)
**Trivial example**

- **Two assumptions:**
  - $\bar{X}$: for each row $k$, up to column permutation, has decomposition
    \[
    \begin{array}{c|c}
    X_k & S_k \\
    \hline
    X_k & \bar{X}_k
    \end{array}
    \]
  - $\Lambda_k = 0$
  - If $X$ has at most one nonzero entry per column (at unknown positions)
    - Simply choose $d_k = 0$

- **How robust is the condition to weakly-sparse outliers?**

- $A_0 = \text{basis of sufficiently incoherent unit atoms}$
  \[
  \forall k \| a_k \|_2 = 1 \quad \max_{k \neq l} |\langle a_k, a_l \rangle| \ll 1
  \]
Trivial example

- **Two assumptions:**
  - \( \bar{X} \): for each row \( k \), up to column permutation, has decomposition
  \[
  \bar{X} = S_k \begin{bmatrix} X_k & 0 \\ \Lambda_k & \bar{\Lambda}_k \end{bmatrix} = 0
  \]
  - and there exists \( d_k, \|d_k\|_\infty < 1 \),
  \[
  0 = X_k s_k^T = \bar{X}_k d_k
  \]
  - \( A_0 = \) basis of sufficiently incoherent unit atoms
  \[
  \forall k \|a_k\|_2 = 1 \quad \max_{k \neq l} |\langle a_k, a_l \rangle| \ll 1
  \]

- If \( X \) has at most one nonzero entry per column (at unknown positions)
  - Simply choose \( d_k = 0 \)

- How robust is the condition to weakly-sparse outliers?
- How many samples \( N \) does it then typically require?
How many training samples?

- **Dimension of the problem**
  \[ B = AX = \]
  \[ \begin{array}{c}
  d \\
  \text{signal dimension}
  \end{array} \]
  \[ K \text{ atoms} \]

- **General dictionary** \( K \geq d \), basis \( K = d \)

- **Required number of training samples:**
  
  \* With \( N = K \), maximum sparsity achieved for \( \hat{A} = B \neq A \)
  \( \hat{X} = \text{Id} \)
  
  1 atom \( \hat{a}_k = 1 \) training sample \( b_n \)

  \* Identifiability from \( N \leq CK \log K \) samples for all “nice” \( A \)?

  \* Identifiability with weakly-sparse \( X \)?
Second result: probability of identifiability

- Random model \( X = (x_{kn}) \)
  - i.i.d. (sub)Gaussian entries in \( \mathbb{R}^K \)
  - a fraction \( p \) set to zero at random

Using concentration of measure:

\[
    P(\text{failure}) \leq C \exp(aK \log K - bN)
\]

Conclusion

Local identifiability guaranteed with high probability from only "few" training samples:

\[
    N \geq C(p) \cdot K \log K
\]

(almost linear in dimension \( K \), even for small \( p \))
Summary

• L1-minimization for dictionary learning:
  ✦ Sufficient condition for local identifiability of bases
  ✦ Condition typically valid
    ✤ even if only weakly-sparse training samples
    ✤ even with relatively few training samples (non combinatorial training set)

• Consequence:
  ✦ ideal convergence of descent algorithms conditionally on good initialization
  ✦ conjecture: with high probability, no spurious local minima

\[ N \geq C(p) \cdot K \log K \]
Perspectives & challenges

• Main open questions:
  ✤ Probability of spurious local minima
  ✤ Optimization algorithm (L1 criterion is nonconvex ...), in progress
  ✤ Stability/robustness to noise / compressible $X$?

• Extensions:
  ✦ other learning paradigms: efficiency? equivalence?
  ✦ greedy approaches ("deflation", ongoing work)
  ✦ alternate optimization (MOD, K-SVD, ...)
  ✦ blind sparse deconvolution
  ✦ learning general subspace arrangements / manifolds [cf Yi Ma]
THE END

remi.gribonval@inria.fr
Geometric interpretation

- Many sparse training examples lie on low-dimensional subspaces
Geometric interpretation

\[ x_n(3) = 0 \iff n \in \overline{\Lambda}_3 \]
Geometric interpretation

\[ X_k \cdot \text{sign}(x_k)^T \]

\sim \text{centroid}
Geometric interpretation

\[ X_k \cdot \text{sign}(x_k)^T = \bar{X}_k d_k, \|d_k\|_\infty < 1 \]
Sparse data models?

- Sparse signals, sparse images ...
  - challenge = large-scale algorithms

Sparse models = Fourier, *lets, ...

Signals

Images
Sparse data models?

- Sparse signals, sparse images ...
  - challenge = large-scale algorithms

Multimodal

- Sparse models = Fourier, *lets, ...

Signals Images

Hyperspectral

- Satellite imaging

Spherical geometry

- Cosmology, HRTF (3D audio)

Data on graphs

- Social networks
- Brain connectivity

Vector valued

- Diffusion tensor

“Exotic” or complex data

- challenge = generic models / model building tools
Local identifiability analysis (1)

• Normalization convention needed (as $AX = B \iff (2A)(X/2) = B$)

\[
A \in \mathcal{A} := \{ \|a_k\|_2 = 1 \}
\]

• Compatible basis perturbations : tangent plane

\[
A + \delta_A, \quad \delta_A \in T_A \mathcal{A}
\]

• Coupling between basis and coefficients

\[
AX = B \iff \delta_A \cdot X + A \cdot \delta_X = 0 \iff \delta_X = -A^{-1} \delta_A \cdot X
\]
Local identifiability analysis (2)

- First order approx. of L1 criterion ($\Lambda = \text{support of } X$)
  \[ \| X + \delta_X \|_1 - \| X \|_1 \approx \langle \delta_X, \text{sign}(X) \rangle + \| (\delta_X)_{\bar{\Lambda}} \|_1 \]
  \[ \forall \delta_X, \ |\langle \delta_X, \text{sign}(X) \rangle| < \| (\delta_X)_{\bar{\Lambda}} \|_1 \implies \text{local minimum} \]

- Admissible perturbations (from previous slide)
  \[ \delta_X = -A^{-1}\delta_A \cdot X \quad \delta_A \in T_{A^T}A \]

- (...) local minimum iff for zero-diagonal matrices $Z$
  \[ \forall Z \neq 0, \ |\langle Z, X\text{sign}(X)^T - \text{diag}([\| x_k \|_1])A^T A \rangle| < \| (ZX)_{\bar{\Lambda}} \|_1 \]

- Decoupling between rows: for orthonormal $A$ ...
  \[ |\langle z, X\text{sign}(x_k)^T \rangle| < \| (X^T z)_{\bar{\Lambda}_k} \|_1 \]
Dictionary learning is not about ...

• Channel estimation with known pulse sequence
  \[ b = Ax \]
  \[ x = \text{known channel input} \]
  \[ b = \text{observed channel output} \]
  \[ A = \text{unknown channel response} \]

• [Ex: Pfander, Rauhut & Tanner, “Identification of Matrices having a Sparse Representation”, 2008]

\[
A = \sum_k \alpha_k A_k \\
b = \sum_k \alpha_k (A_k x)
\]
Dictionary learning is not about ...

- Channel estimation with known pulse sequence
  - $x = \text{known channel input}$
  - $b = \text{observed channel output}$
  - $A = \text{unknown channel response}$

- [Ex: Pfander, Rauhut & Tanner, "Identification of Matrices having a Sparse Representation", 2008]

\[
 b = \sum_k \alpha_k A_k
\]
Blind Source Separation

- Mixing model in the time-frequency domain

\[
\begin{align*}
    b_1(\tau, f) & = A X(\tau, f) \\
    b_2(\tau, f) & = \end{align*}
\]

- And “miraculously” ... \( \cdots \) time-frequency representations of audio signals are (often) almost disjoint.

Identifiability of mixing matrix \( A \)

- geometric properties of the scatter plot
- (concentration along lines)
- thanks to sparsity of \( X \)