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

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

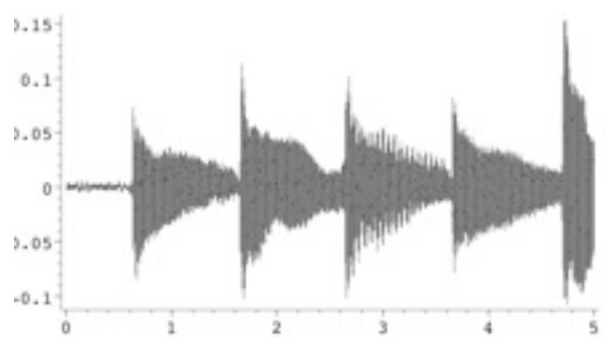
Journées STAR 2009
IRISA, 22-23 octobre 2009

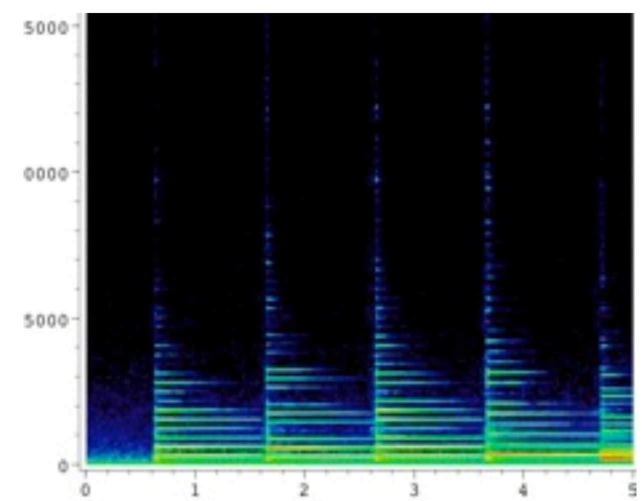
Outline

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

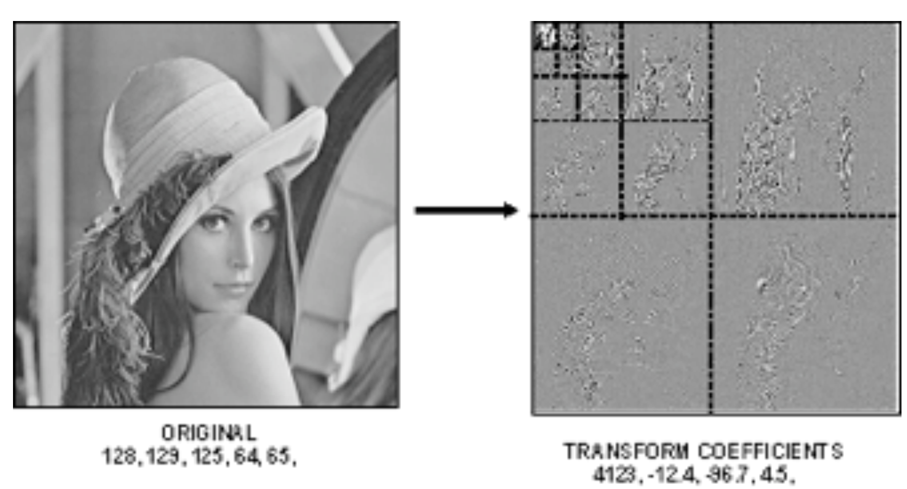
- Short Time Fourier Transform (for time-varying harmonic signals





Black = zero

- Wavelet transform (for piecewise smooth images)



Grey = zero

Sparse signal models

- An image / a signal = sum of few atoms $\mathbf{a}_k$

$$\mathbf{b} = \sum_k x_k \mathbf{a}_k = \mathbf{A}x$$

  - *Dictionary* = collection of atoms = $\mathbf{A}$
  - *Representation* = coefficient vector = x

- Sparsity of x : enables compression, separation ...
- Sparsity of x ? Only if dictionary $\mathbf{A}$ is "well chosen"
  - Pre-chosen atoms: wavelets, Gabor, etc.
  - Learned dictionary = from collection of signals / images

$$\mathbf{b}_n = \mathbf{A}x_n,\ 1 \leq n \leq N$$

Dictionary learning for sparse representations

- Sparse modeling : choose a dictionary

Training image database

Dictionary learning for sparse representations

- Sparse modeling : choose a dictionary

Training image database

patch extraction

Training patches

$$\mathbf{b}_n = \mathbf{A}x_n,\ 1 \leq n \leq N$$

IRISA

Dictionary learning for sparse representations

- Sparse modeling : choose a dictionary

Training image database

patch extraction

Training patches

$$\mathbf{b}_n = \mathbf{A} x_n, \quad 1 \leq n \leq N$$

Unknown dictionary	Unknown sparse coefficients

Dictionary learning for sparse representations

- Sparse modeling : choose a dictionary

Training image database

patch extraction

Training patches

$$\mathbf{b}_n = \mathbf{A} x_n, \quad 1 \leq n \leq N$$

Unknown dictionary	Unknown sparse coefficients

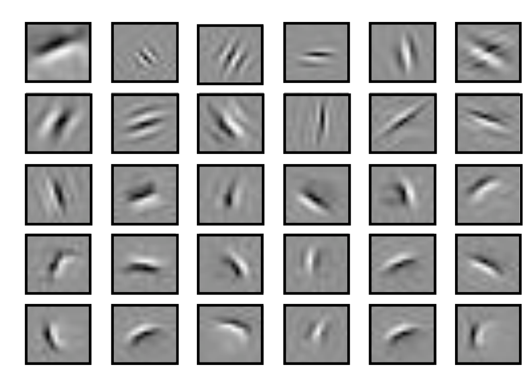
sparse learning

$\hat{\mathbf{A}}$ = edge-like atoms

[Olshausen & Field 96]

= shifts of edge-like motifs

[Jost, Vandergheynst, Lesage & Gribonval 2005]



Dictionary learning ?

- Problem : estimate a matrix $\mathbf{A}$ given observed samples

$$\mathbf{b}_n = \mathbf{A}x_n,\ 1 \leq n \leq N$$

IRISA

Dictionary learning ?

- Problem : estimate a matrix $\mathbf{A}$ given observed samples

$$\mathbf{b}_n = \mathbf{A}x_n,\ 1 \leq n \leq N$$

Dictionary learning ?

- Problem : estimate a matrix $\mathbf{A}$ given observed samples

$$\mathbf{b}_n = \mathbf{A}x_n,\ 1 \leq n \leq N \quad \Rightarrow \quad \mathbf{B} = \mathbf{A}X$$

IRISA

Dictionary learning ?

- Problem : estimate a matrix $\mathbf{A}$ given observed samples

$$\mathbf{b}_n = \mathbf{A}x_n,\ 1 \leq n \leq N \quad \Rightarrow \quad \mathbf{B} = \mathbf{A}X$$

$$\mathbf{A} \begin{cases} \text{Unknown mixing matrix (blind source separation)} \\ \text{Unknown dictionary (sparse signal approximation)} \\ \text{Unknown channel filter (blind channel estimation) ...} \end{cases}$$

Dictionary learning ?

- Problem : estimate a matrix $\mathbf{A}$ given observed samples

$$\mathbf{b}_n = \mathbf{A}x_n, \; 1 \leq n \leq N \quad \Rightarrow \quad \mathbf{B} = \mathbf{A}X$$

$\mathbf{A}$ $\left\{ \begin{array}{l} \text{Unknown mixing matrix (blind source separation)} \\ \text{Unknown dictionary (sparse signal approximation)} \\ \text{Unknown channel filter (blind channel estimation) ...} \end{array} \right.$

X Unknown sources / signal representations / ...

IRISA

jeudi 22 octobre 2009

Dictionary learning ?

- Problem : estimate a matrix $\mathbf{A}$ given observed samples

$$\mathbf{b}_n = \mathbf{A}x_n, \; 1 \le n \le N \quad \Rightarrow \quad \mathbf{B} = \mathbf{A}X$$

$\mathbf{A}$ $\left\{\begin{array}{l}\text{Unknown mixing matrix (blind source separation)} \\ \text{Unknown dictionary (sparse signal approximation)} \\ \text{Unknown channel filter (blind channel estimation) ...}\end{array}\right.$

X Unknown sources / signal representations / ...

- Fundamentally ill-posed **factorization problem** : need (weak) model on unknown coefficients X and / or matrix $\mathbf{A}$

Theoretical dictionary learning

- Problem : estimate a matrix $\mathbf{A}$ given samples

$$\mathbf{b}_n = \mathbf{A}x_n,\ 1 \leq n \leq N \qquad \mathbf{B} = \mathbf{A}X$$

	ICA (Independent Component Analysis)
Model of ...	probability density function $p(X)$
Assumption	**Independence** $p(X) = \prod_{nk} p(x_n(k))$
Identifiability	Darmois theorem
Identification	Contrast functions $\mathbf{A} \sim \hat{\mathbf{W}}^{-1}$ $\hat{\mathbf{W}} := \arg\min_{\mathbf{W}} E_X(f(\mathbf{WA}X))$
Issues	In practice : **finite** training sets ~~expectation~~ $\longrightarrow$ sample average

Theoretical dictionary learning

- Problem : estimate a matrix $\mathbf{A}$ given samples

$$\mathbf{b}_n = \mathbf{A}x_n, \; 1 \leq n \leq N \qquad \mathbf{B} = \mathbf{A}X$$

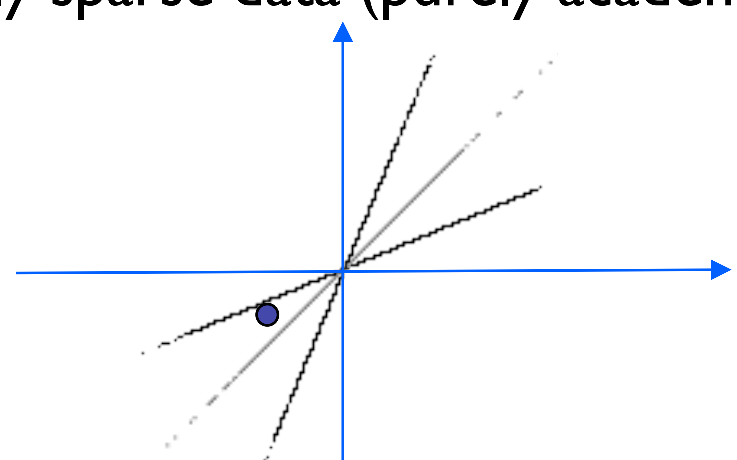
	ICA (Independent Component Analysis)	**SCA** (Sparse Component Analysis)
Model of ...	probability density function $p(X)$	sample matrix X
Assumption	**Independence** $p(X) = \prod_{nk} p(x_n(k))$	Sparsity / geometry ★ many zeroes in X ★ x_n and $\mathbf{b}_n$ concentrate around **union of low dimensional subspaces**
Identifiability	Darmois theorem	[Georgiev, Theis & Cichocki 05] [Aharon, Elad & Bruckstein 06]
Identification	Contrast functions $\mathbf{A} \sim \hat{\mathbf{W}}^{-1}$ $\quad \hat{\mathbf{W}} := \arg\min_{\mathbf{W}} E_X(f(\mathbf{WA}X))$	Combinatorial algorithms
Issues	In practice : **finite** training sets ~~expectation~~ → sample average	Identifiability assumes: ★ highly sparse coefficients ★ (combinatorially ?) many training examples

IRISA

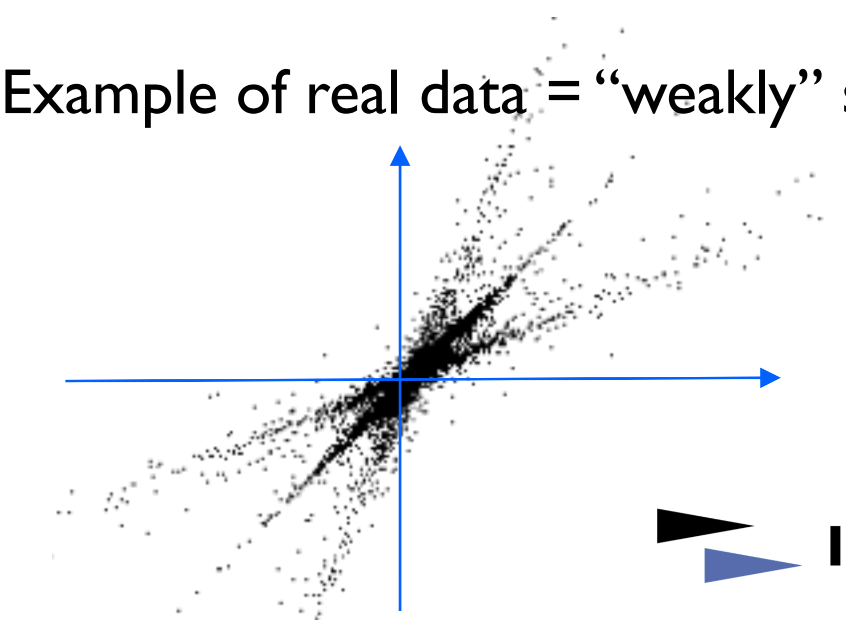
Objectives

- Long term goal :
  - ✦ identifiability conditions on X to recover **A** from $\mathbf{B} = \mathbf{A}X$
  - ✦ 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



IRISA

L1 minimization for dictionary learning

Holy grail: provably good + efficient sparse learning

- **Sparse representations**
  - Known matrix $\mathbf{A}$
  - Data model $\mathbf{b} = \mathbf{A}x_0$
  - Identifiability theorems:

$$\|x_0\|_0 \leq k_1(\mathbf{A})$$

$$x_0 = \arg\min_{\mathbf{A}x = \mathbf{b}} \|x\|_1$$

  - Much literature since 2001
    (Donoho & Huo, Elad & Bruckstein, Gribonval & Nielsen, Candès & Romberg & Tao, Tropp, Donoho & Tanner, ... and many others)

- **Dictionary learning**
  - Unknown matrix $\mathbf{A}_0$
  - Data model $\mathbf{B} = \mathbf{A}_0 X_0$
  - Identifiability theorem ?

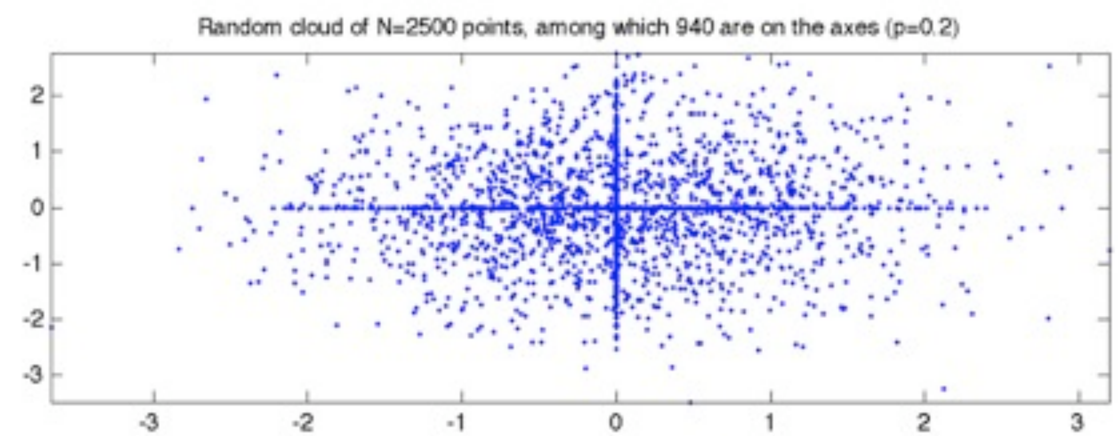
$$\mathbf{A}_0, X_0 \in \boxed{?}$$

$$(\mathbf{A}_0, X_0) \in \arg\min_{\mathbf{A}X = \mathbf{B}} \|X\|_1$$

  - Most literature on Independent Component Analysis (ICA), density model rather than finite sample size geometric model

Numerical example

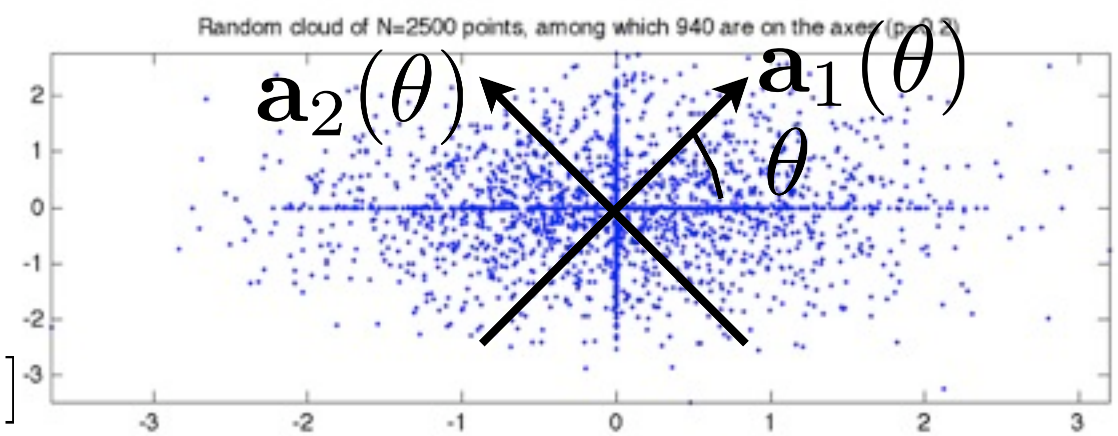
- Cloud of 2500 training samples in $\mathbb{R}^2$
  - ✦ ~1000 sparse [= on axes]
  - ✦ ~1500 non-sparse



IRISA

Numerical example

- Cloud of 2500 training samples in $\mathbb{R}^2$
  - ✦ ~1000 sparse [= on axes]
  - ✦ ~1500 non-sparse
- Orthonormal basis
  - ✦ Angle $\theta \longleftrightarrow \mathbf{A}_\theta = [\mathbf{a}_1(\theta), \mathbf{a}_2(\theta)]$

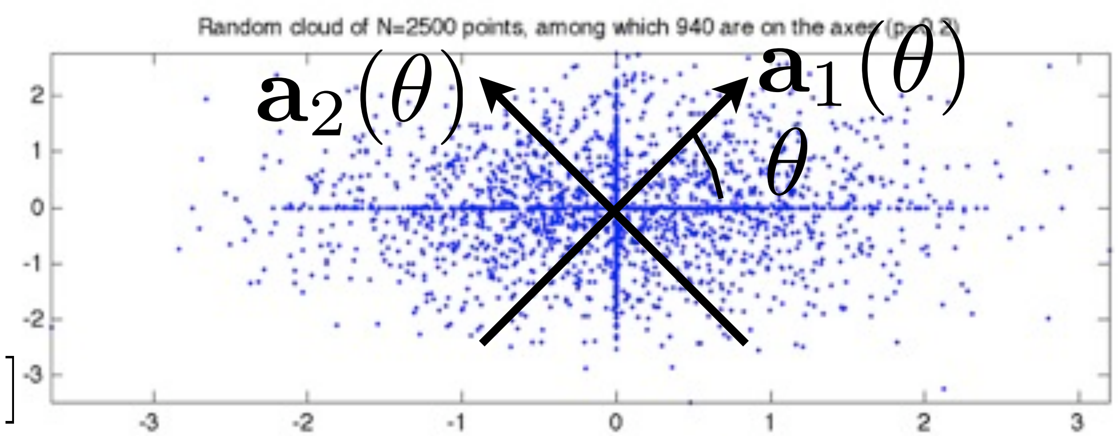


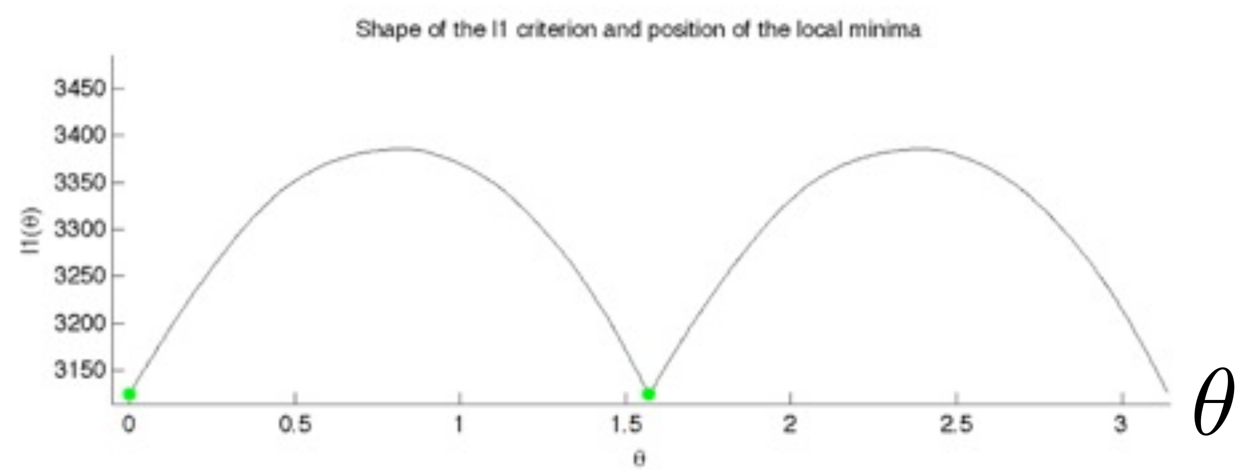
IRISA

Numerical example

- Cloud of 2500 training samples in $\mathbb{R}^2$
  - ✦ ~1000 sparse [= on axes]
  - ✦ ~1500 non-sparse
- Orthonormal basis
  - ✦ Angle $\theta \longleftrightarrow \mathbf{A}_\theta = [\mathbf{a}_1(\theta), \mathbf{a}_2(\theta)]$
- L1 criterion

$$\|\mathbf{A}_\theta^{-1}\mathbf{A}_0 X\|_1$$

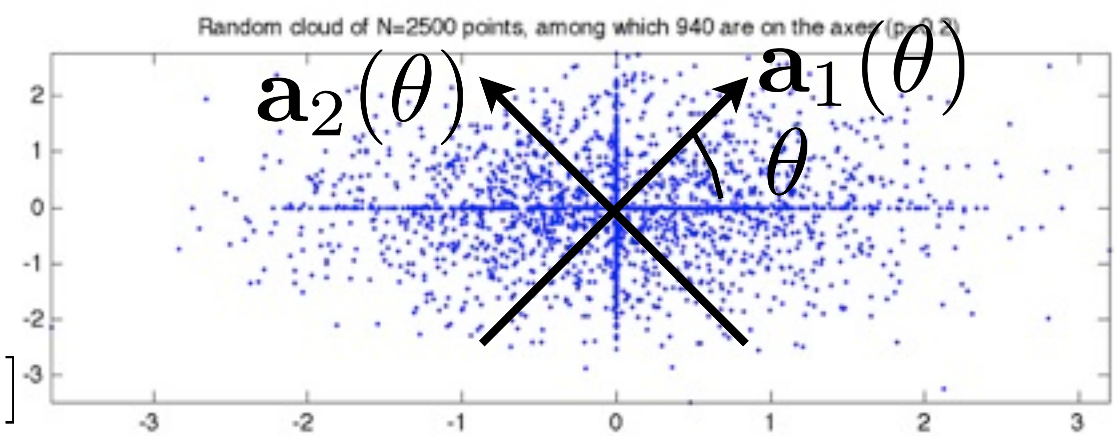




IRISA

Numerical example

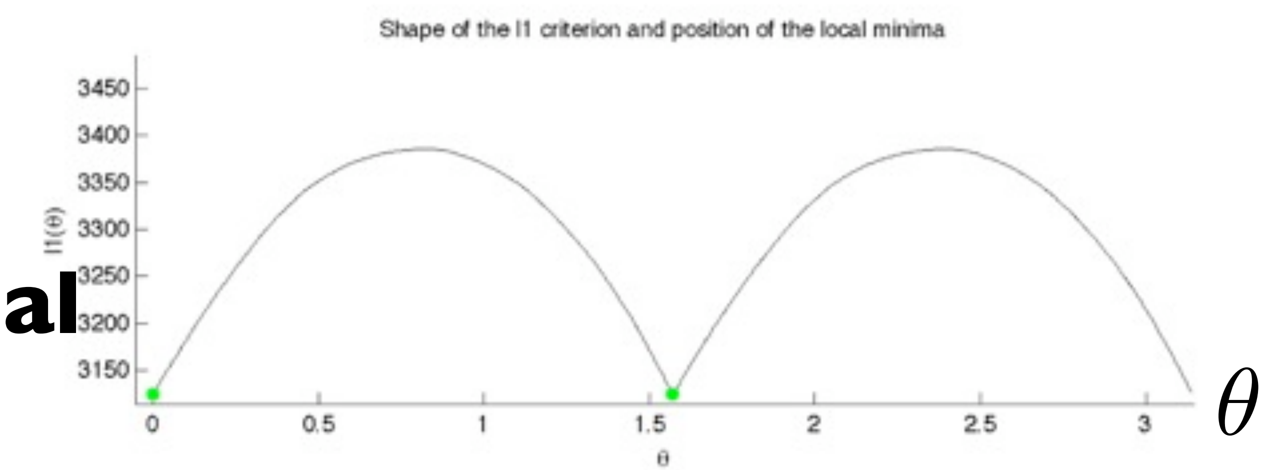
- Cloud of 2500 training samples in $\mathbb{R}^2$
  - ~1000 sparse [= on axes]
  - ~1500 non-sparse



- Orthonormal basis
  - Angle $\theta \longleftrightarrow \mathbf{A}_\theta = [\mathbf{a}_1(\theta), \mathbf{a}_2(\theta)]$
- L1 criterion

$$\|\mathbf{A}_\theta^{-1}\mathbf{A}_0 X\|_1$$

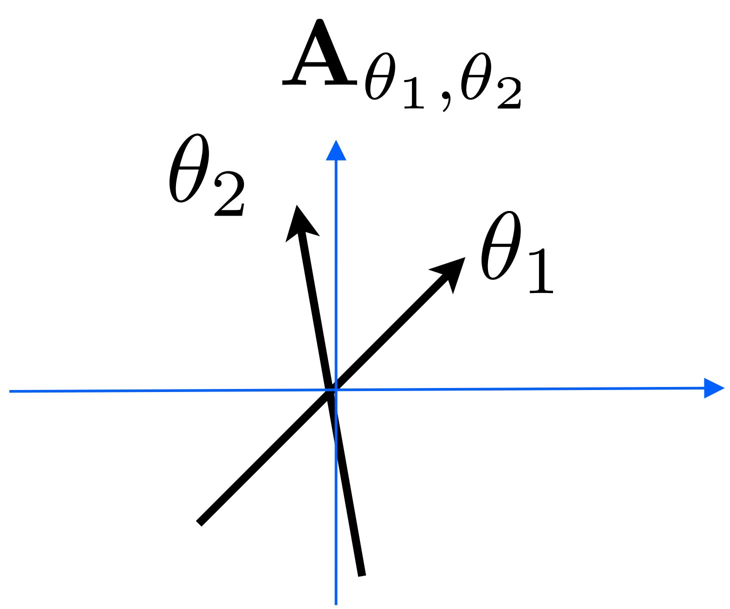
  - global optimum=**original**
  - *no other local minimum*

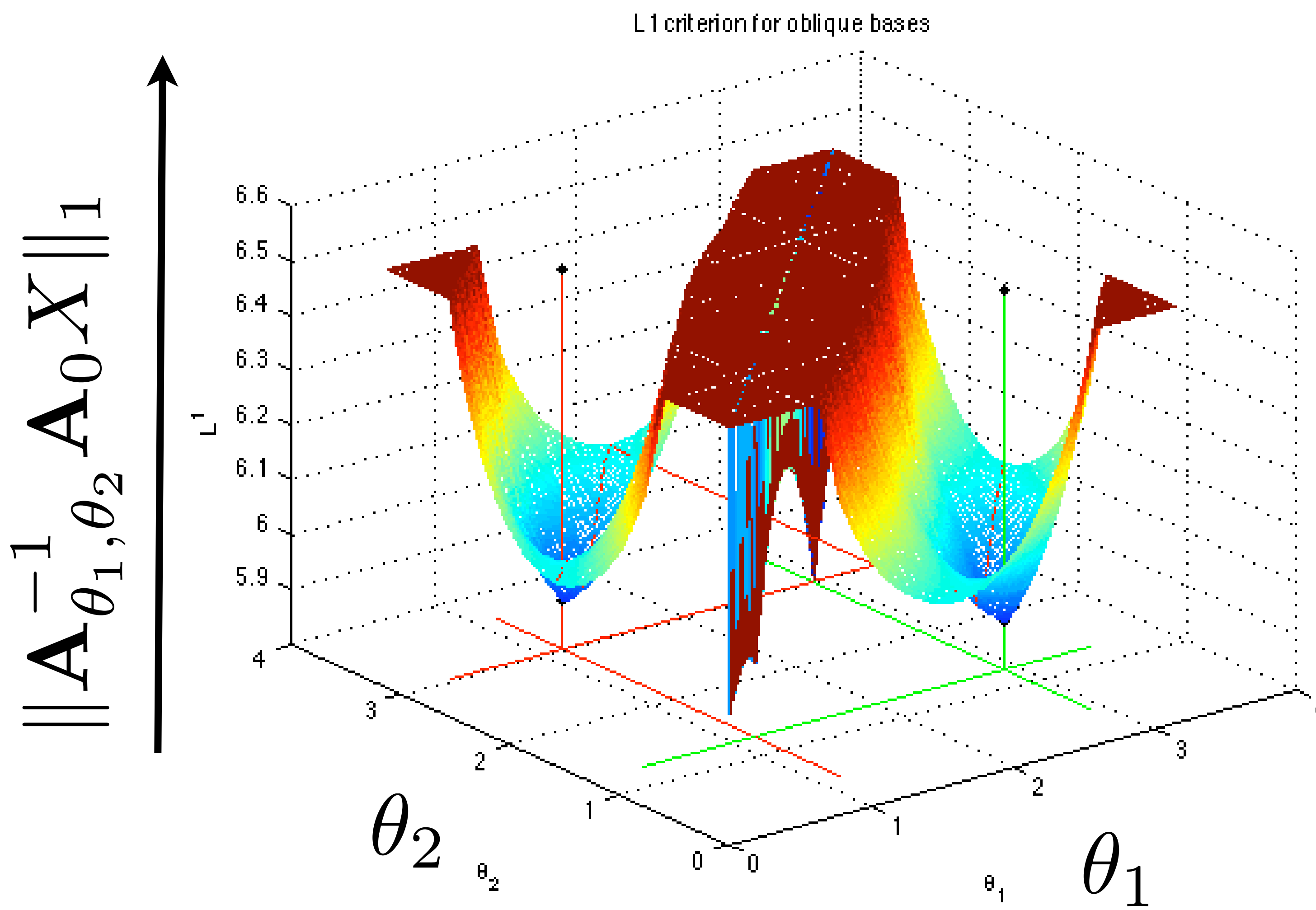


IRISA

Numerical example

Non orthogonal bases

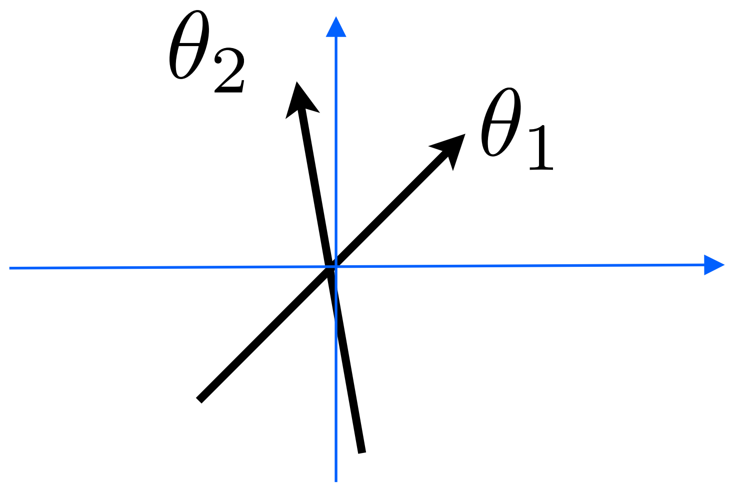


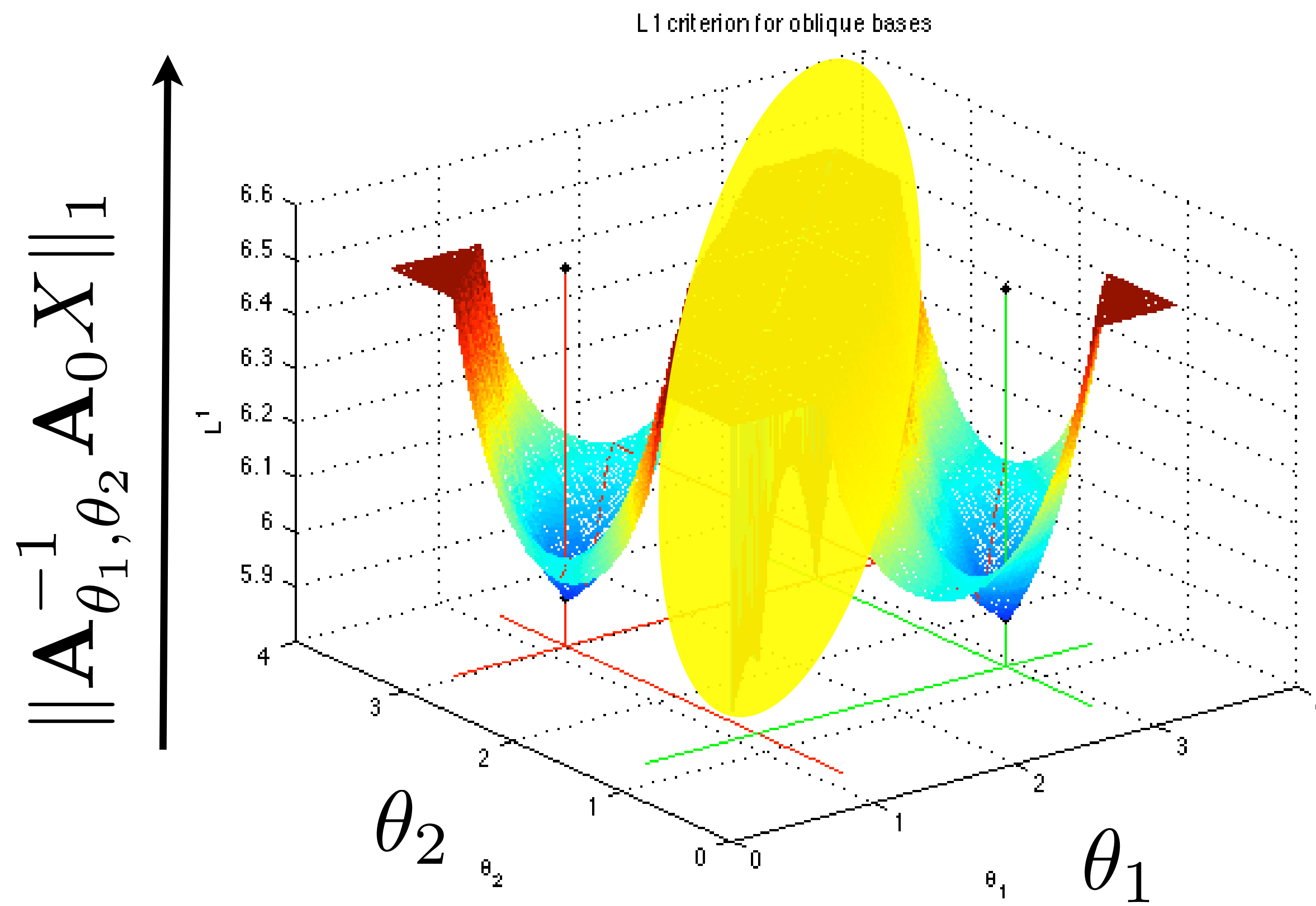


Numerical example

Non orthogonal bases

$\mathbf{A}_{\theta_1,\theta_2}$

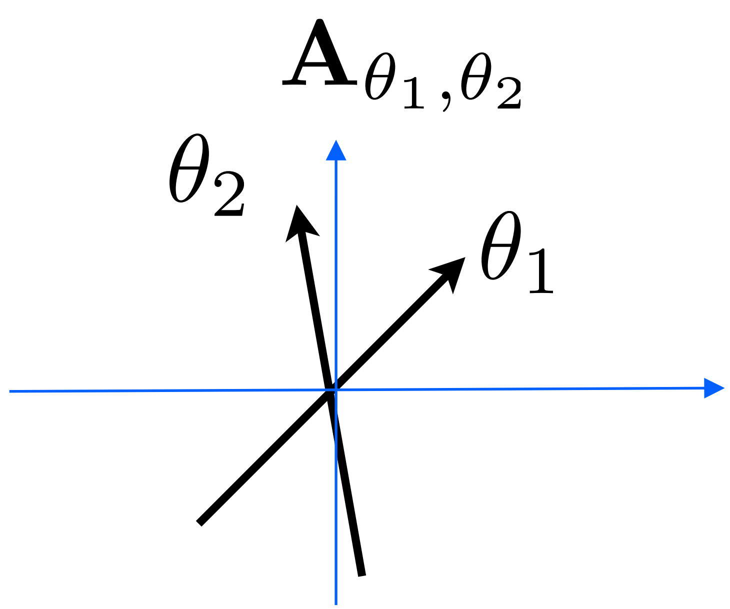


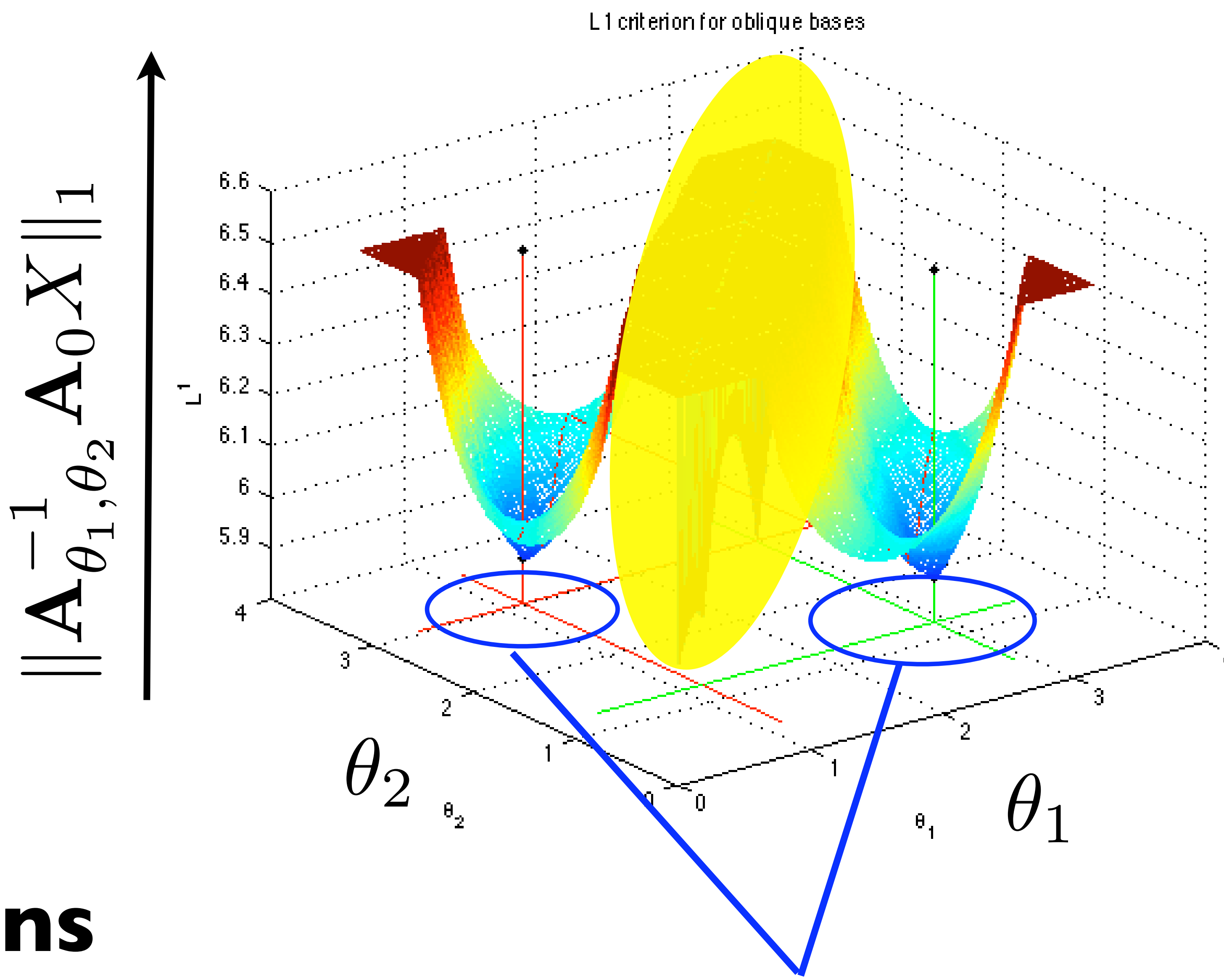


IRISA

Numerical example

Non orthogonal bases





Empirical observations

a) Global minima match the original basis

b) There is no other local minimum.

IRISA

Theoretical results

1. "Local identifiability" for (non overcomplete) L1 dictionary learning
   - algebraic / geometric *characterization of local minima*

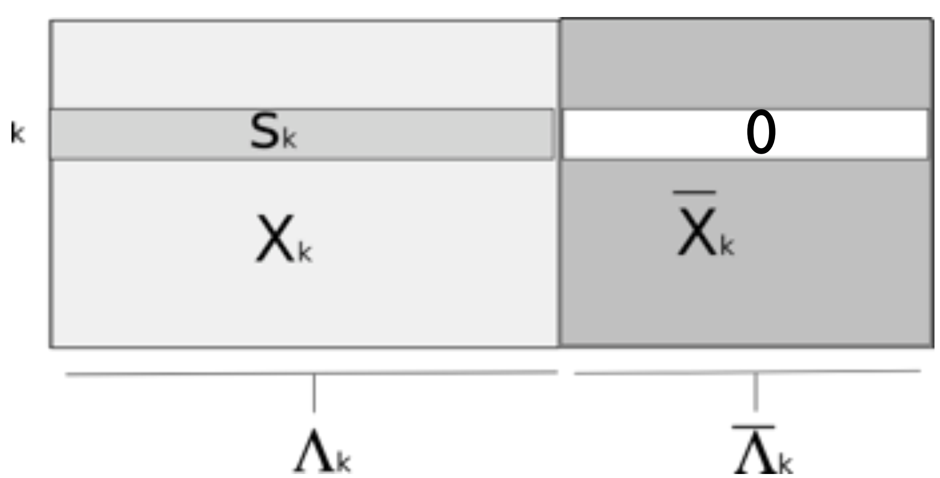
2. Probability of identifiability
   - model on X: random, weakly-sparse
   - analysis of identifiability *for (small) finite sample size*

IRISA

Local identifiability result

- **Two assumptions**:
  - ✦ X : for each row k, up to column permutation, has decomposition

$X =$



and there exists $d_k, \ \|d_k\|_\infty < 1,$

$$X_k s_k^T = \bar{X}_k d_k$$

  - ✦ $\mathbf{A}_0$ = **basis of sufficiently incoherent unit atoms**

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

- **Conclusion** :
  - ✦ $\mathbf{A}_0$ = **local minimum** of L1 among (*not necessarily orthonormal*) bases

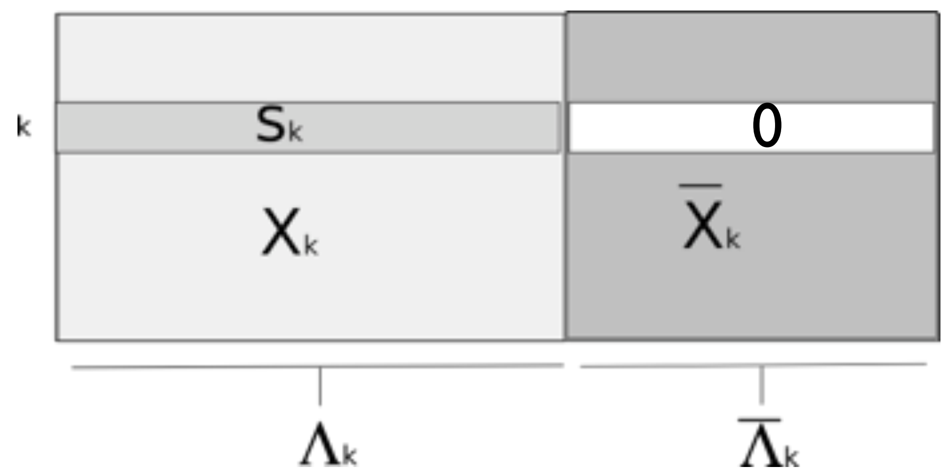
$$\begin{cases} (\mathbf{A}', X') \approx (\mathbf{A}_0, X) \\ \mathbf{A}'X' = \mathbf{A}_0 X \end{cases}$$

$$\|X'\|_1 \geq \|X\|_1$$

Trivial example

- **Two assumptions**:
  - ✦ X : for each row k, up to column permutation, has decomposition

$X =$



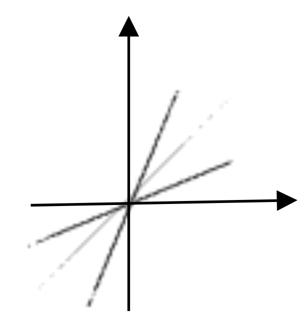
and there exists $d_k, \quad \|d_k\|_\infty < 1,$

$$X_k s_k^T = \bar{X}_k d_k$$

  - ✦ $\mathbf{A}_0$ = **basis of sufficiently incoherent unit atoms**

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

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



  - ✦ *Simply choose*

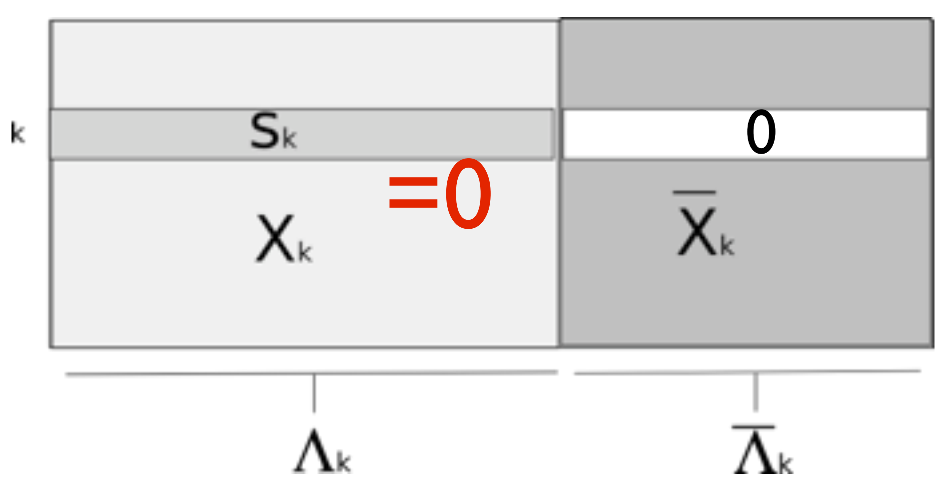
- *How robust is the condition to weakly-sparse outliers ?*
- *How many samples N does it then typically require ?*

IRISA

Trivial example

- **Two assumptions**:
  - ✦ X : for each row k, up to column permutation, has decomposition

$X =$



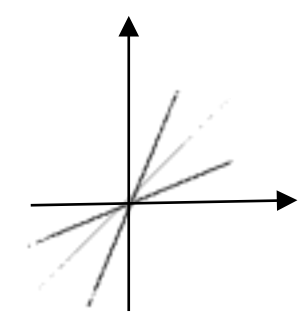
and there exists $d_k, \quad \|d_k\|_\infty < 1,$

$$0 = X_k s_k^T = \bar{X}_k d_k$$

  - ✦ $\mathbf{A}_0$ = **basis of sufficiently incoherent unit atoms**

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

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



  - ✦ *Simply choose*

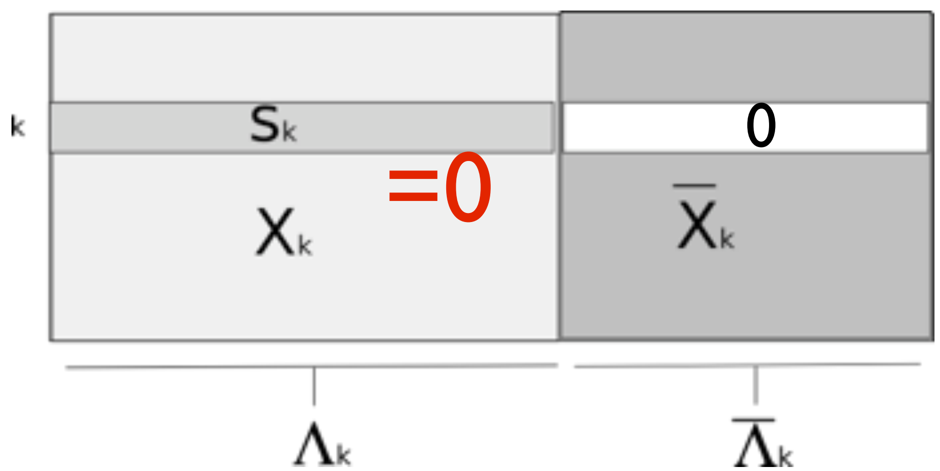
- *How robust is the condition to weakly-sparse outliers ?*
- *How many samples N does it then typically require ?*

IRISA

Trivial example

- **Two assumptions**:
  - X : for each row k, up to column permutation, has decomposition

$X =$



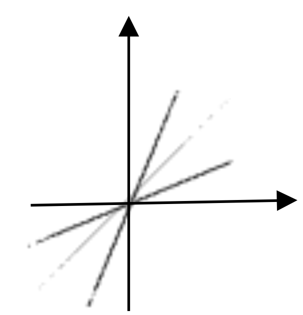
and there exists $d_k, \ \|d_k\|_\infty < 1,$

$$0 = X_k s_k^T = \bar{X}_k d_k$$

  - $\mathbf{A}_0$ = **basis of sufficiently incoherent unit atoms**

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

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



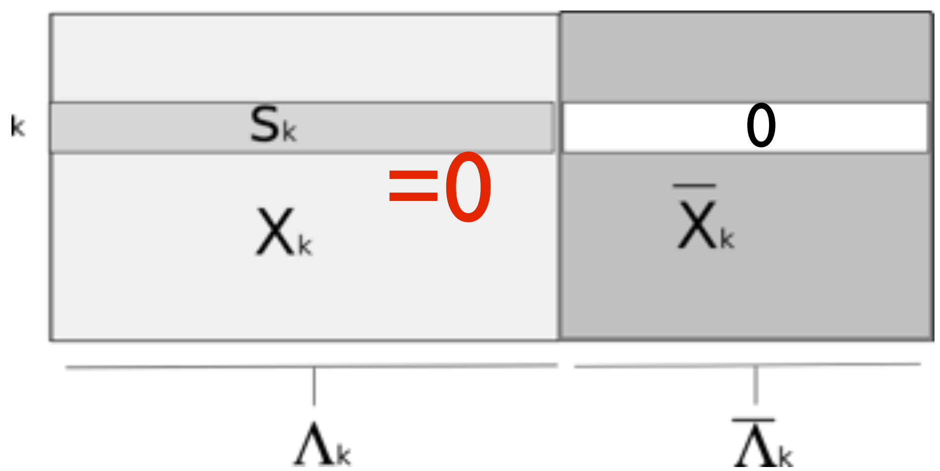
  - *Simply choose* $d_k = 0$

IRISA

Trivial example

- **Two assumptions**:
  - ✦ X : for each row k, up to column permutation, has decomposition

$X =$



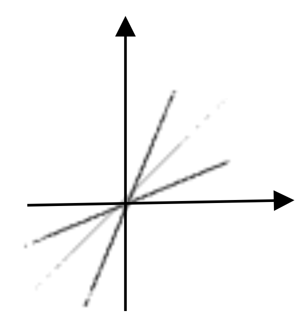
and there exists $d_k, \ \|d_k\|_\infty < 1,$

$0 = X_k s_k^T = \bar{X}_k d_k$

  - ✦ $\mathbf{A}_0$ = **basis of sufficiently incoherent unit atoms**

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

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



  - ✦ *Simply choose* $d_k = 0$

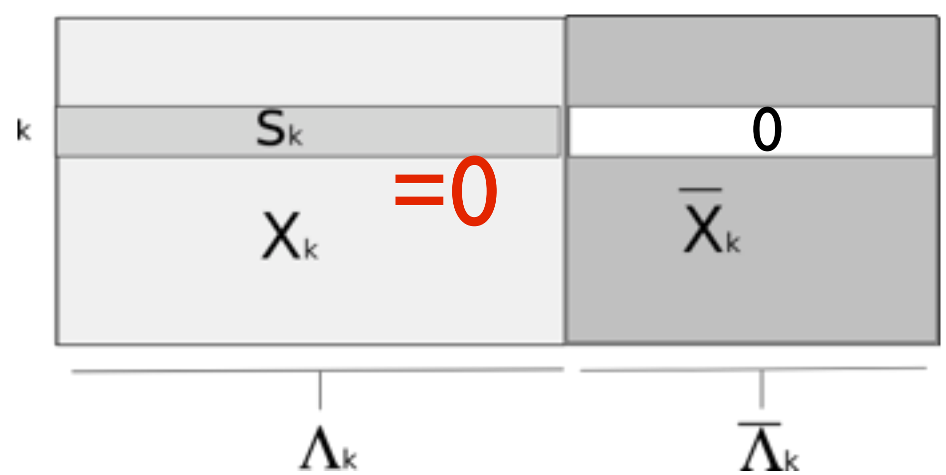
- *How robust is the condition to weakly-sparse outliers ?*

IRISA

Trivial example

- **Two assumptions**:
  - ✦ X : for each row k, up to column permutation, has decomposition

$X =$



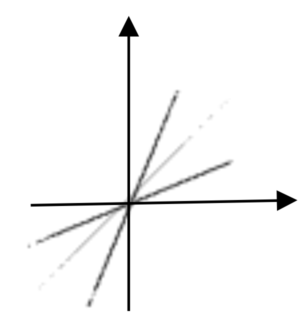
and there exists $d_k, \ \|d_k\|_\infty < 1,$

$$0 = X_k s_k^T = \bar{X}_k d_k$$

  - ✦ $\mathbf{A}_0$ = **basis of sufficiently incoherent unit atoms**

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

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



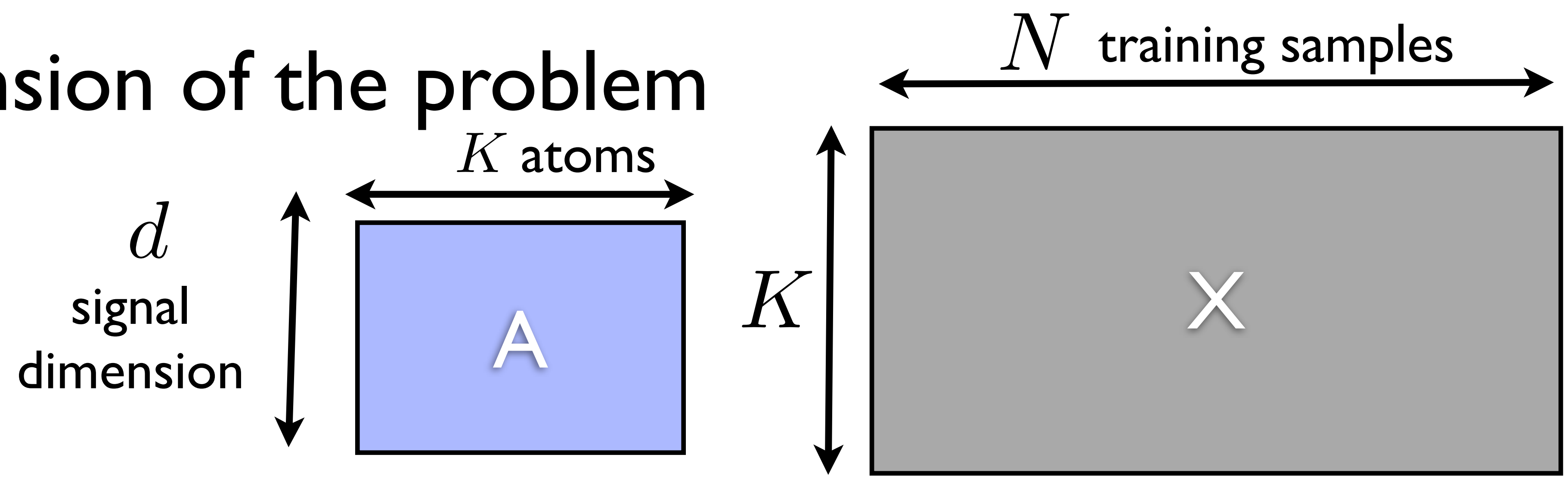
  - ✦ *Simply choose* $d_k = 0$

- *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

$\mathbf{B} = \mathbf{A}X =$



- General dictionary $K \geq d$, basis $K = d$

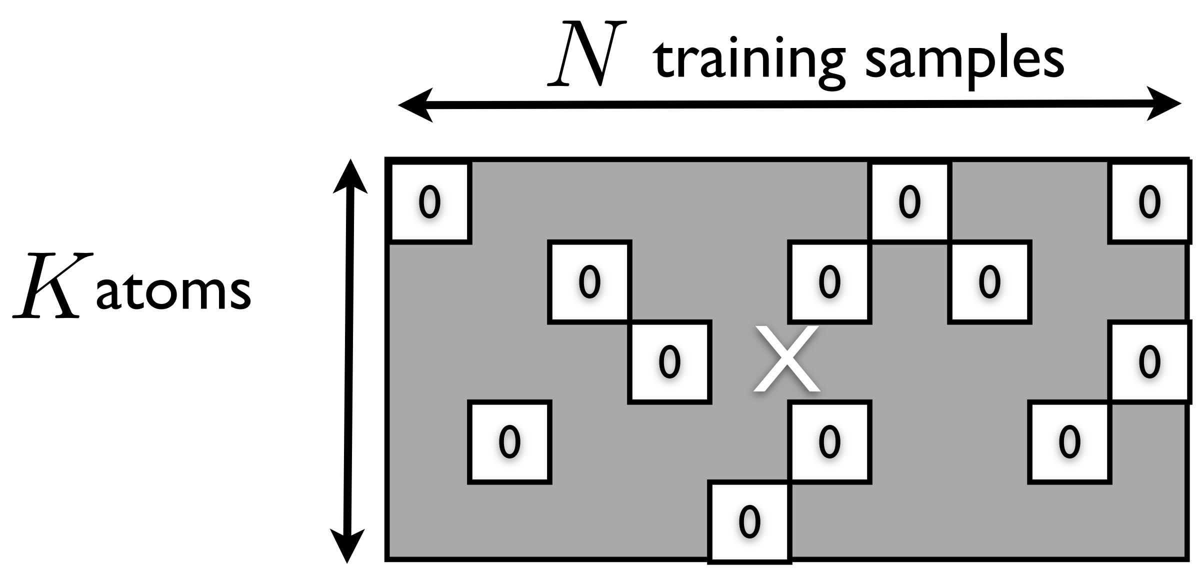
- Required number of training samples:
  - ✤ With $N = K$, maximum sparsity achieved for $\hat{\mathbf{A}} = \mathbf{B} \neq \mathbf{A}$ $\quad \hat{X} = \mathbf{Id}$

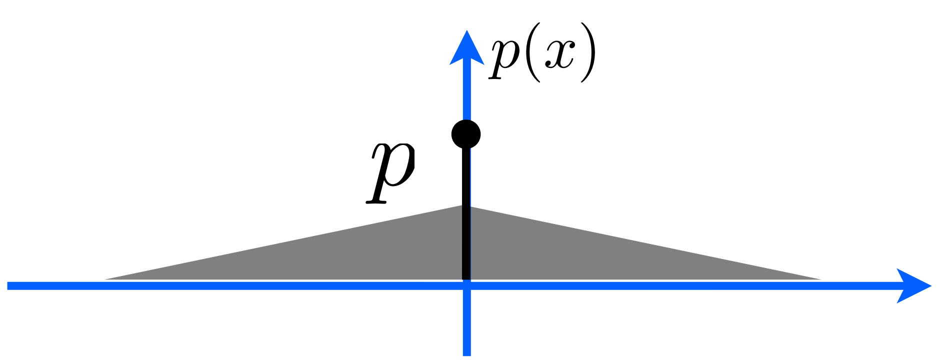
    1 atom $\hat{\mathbf{a}}_k$ = 1 training sample $\mathbf{b}_n$

  - ✤ Identifiability from $N \leq CK \log K$ samples for all "nice" $\mathbf{A}$?
  - ✤ 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 :

Probability of failure ...

$$P(☹) \leq C \exp(aK \log K - bN)$$

Conclusion

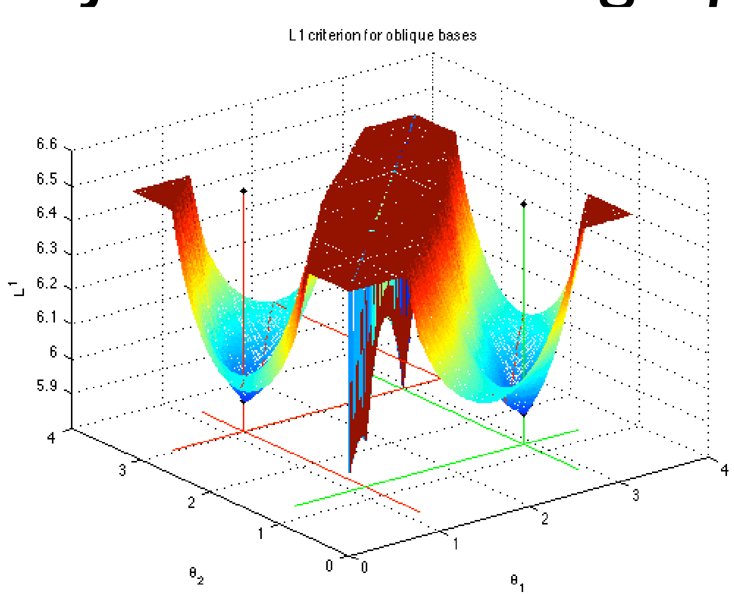
Local identifiability guaranteed with high probability from only **"few"** training samples:

$$N \geq C(p) \cdot K \log K$$

(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) $N \geq C(p) \cdot K \log K$
- Consequence :
  - ideal convergence of descent algorithms *conditionally on good initialization*
  - *conjecture : with high probability, no spurious local minima*



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 arrangments / manifolds [cf Yi Ma]

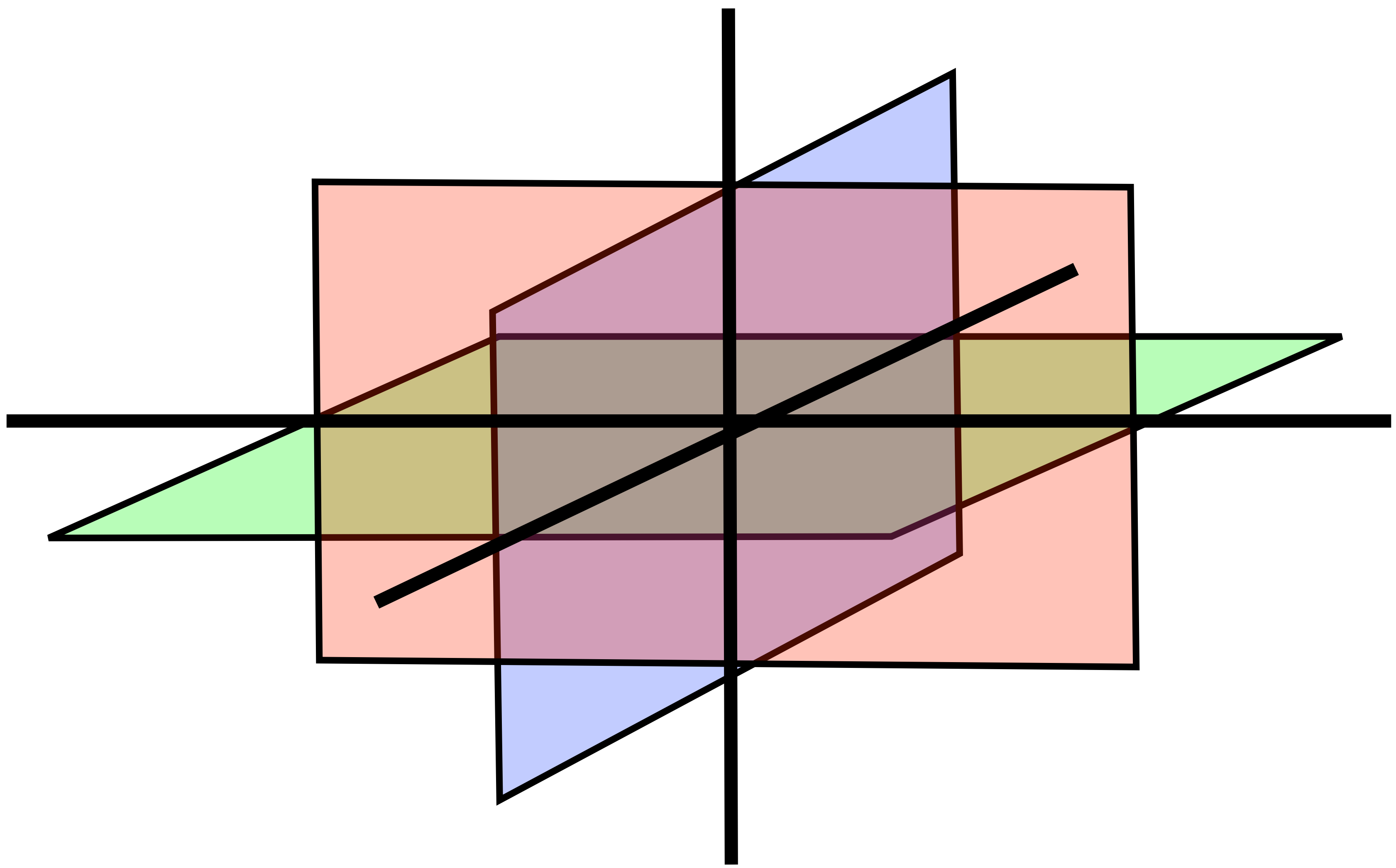
THE END

remi.gribonval@inria.fr

IRISA

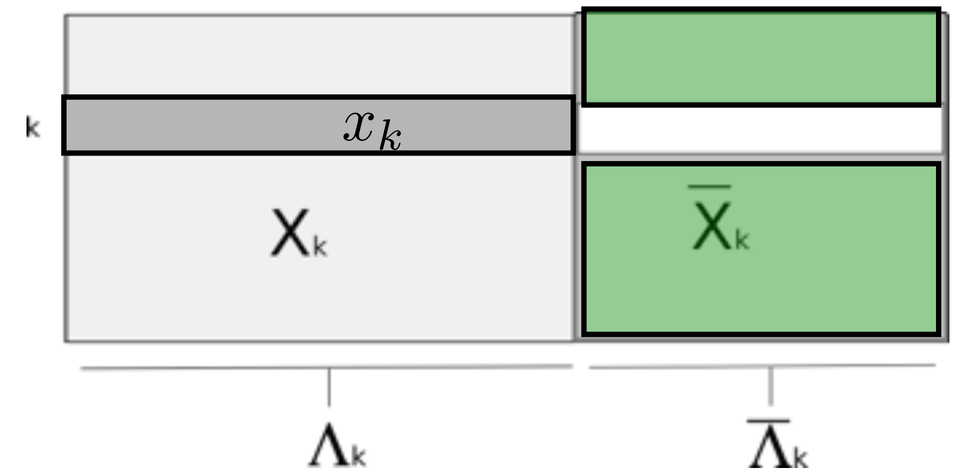
Geometric interpretation

- Many sparse training examples lie on low-dimensional subspaces

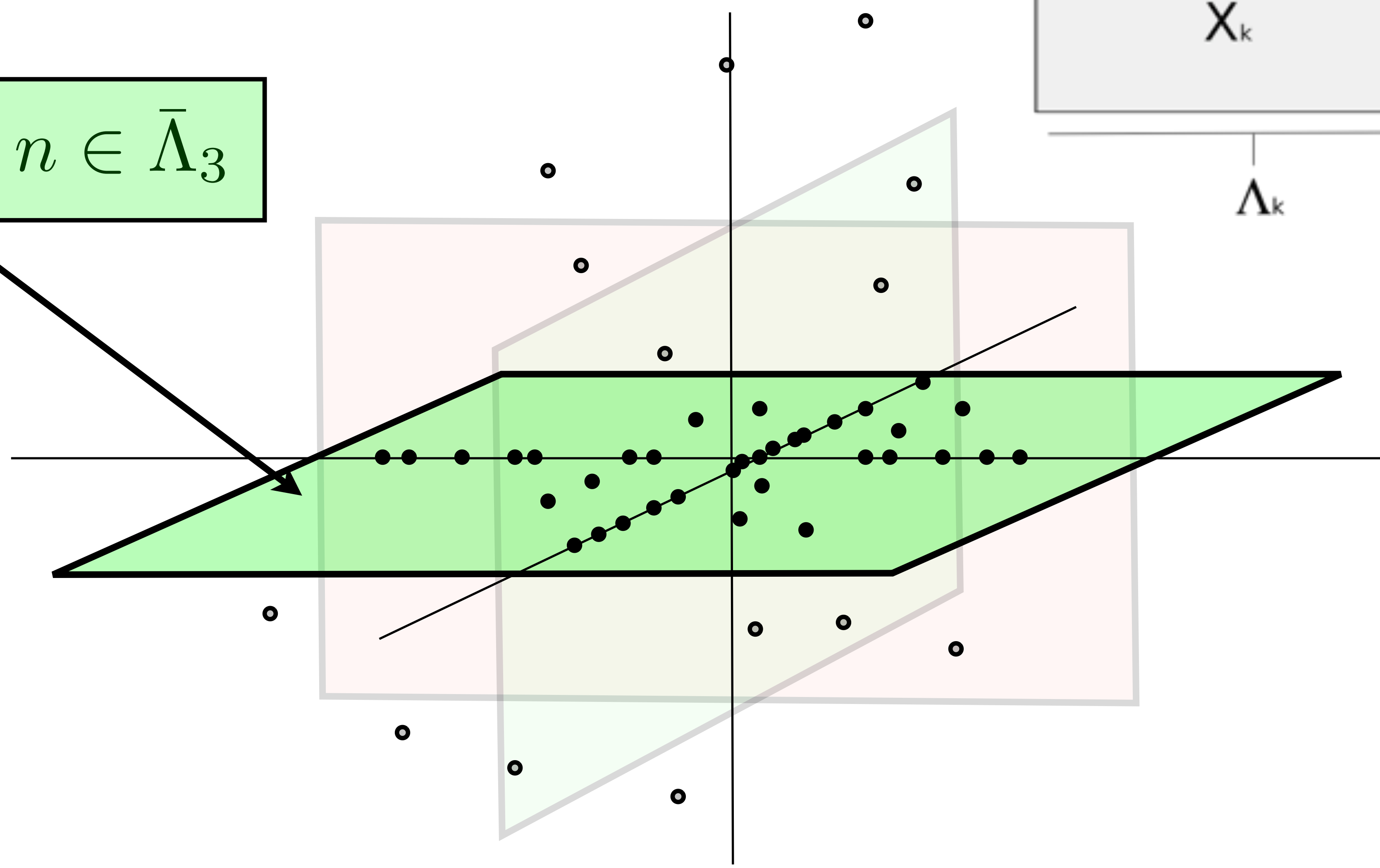


IRISA

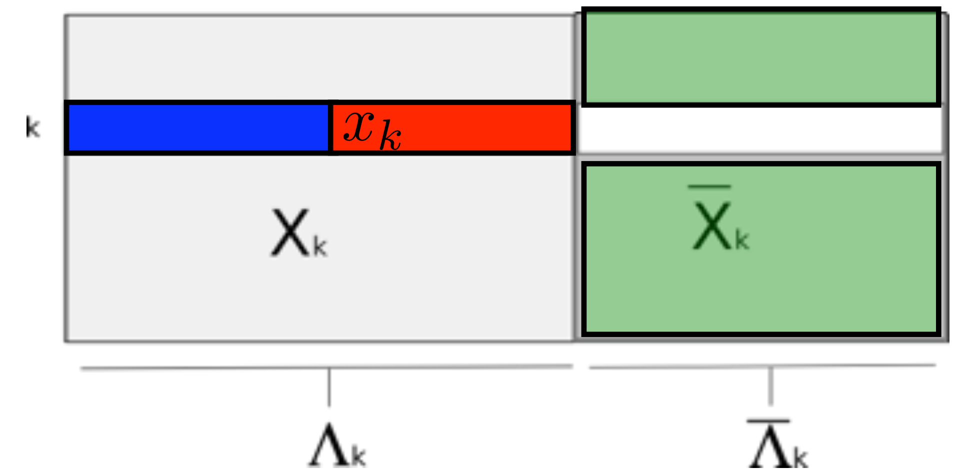
Geometric interpretation

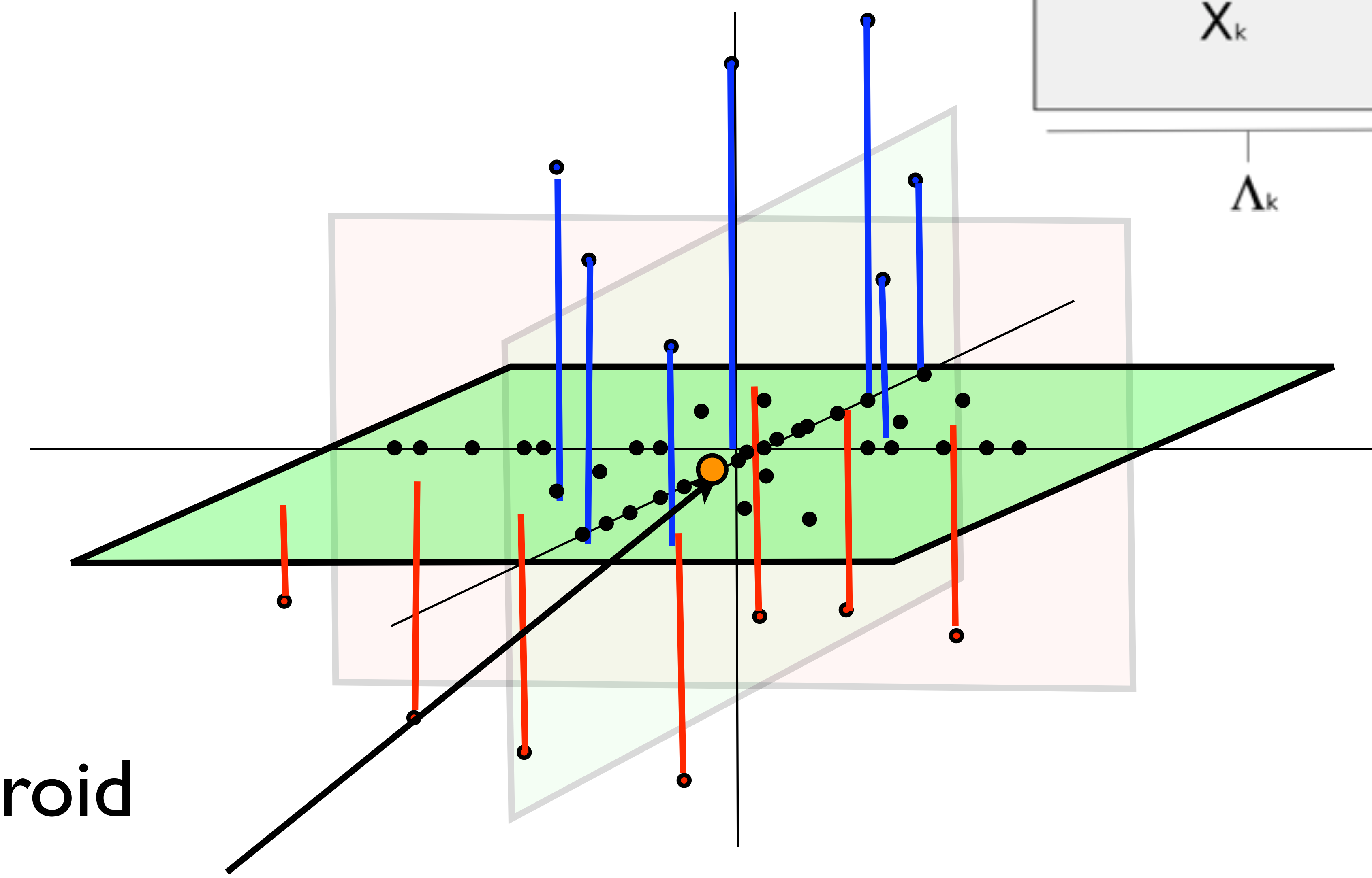


$$x_n(3) = 0 \Leftrightarrow n \in \bar{\Lambda}_3$$



Geometric interpretation

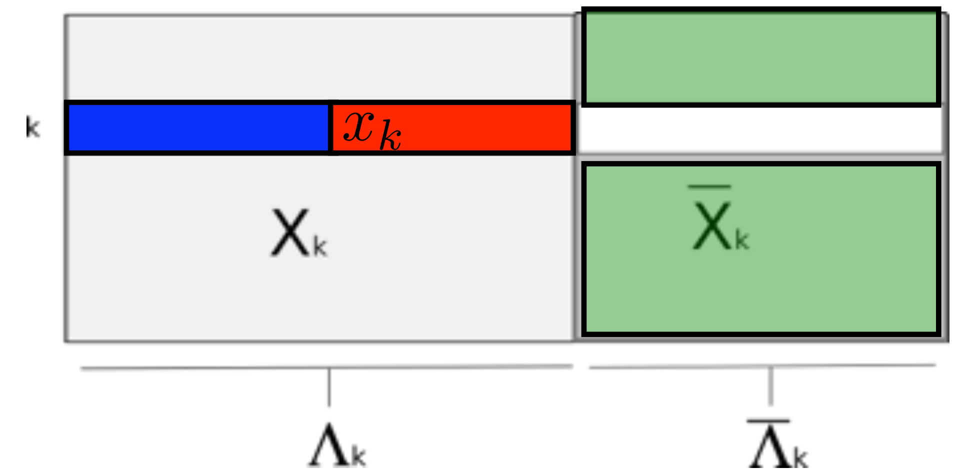




~ centroid

$$\boxed{X_k \cdot \mathrm{sign}(x_k)^T}$$

Geometric interpretation



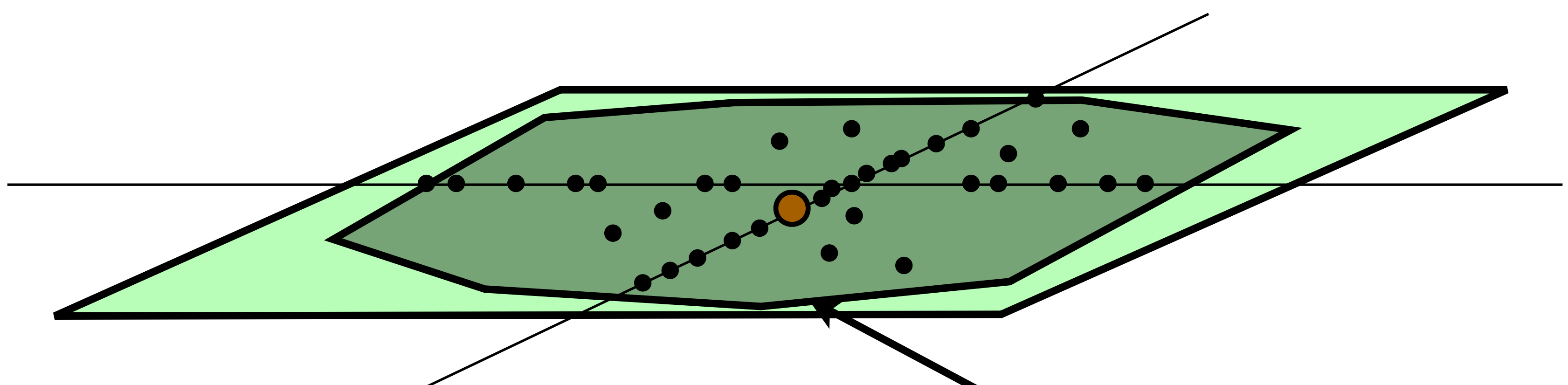


Image of unit hypercube by $\bar{X}_3$

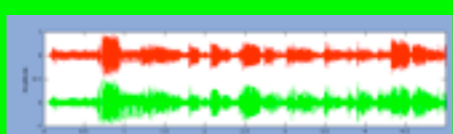
$$\boxed{X_k \cdot \mathrm{sign}(x_k)^T} = \bar{X}_k d_k, \|d_k\|_\infty < 1$$

IRISA

Sparse data models ?

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

Sparse models = Fourier, *lets, ...

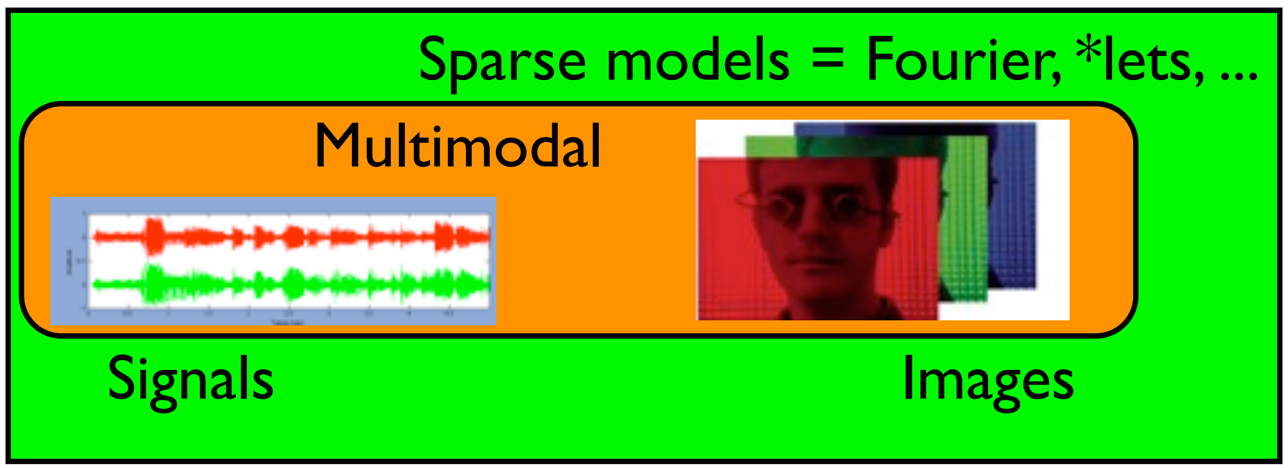


Signals

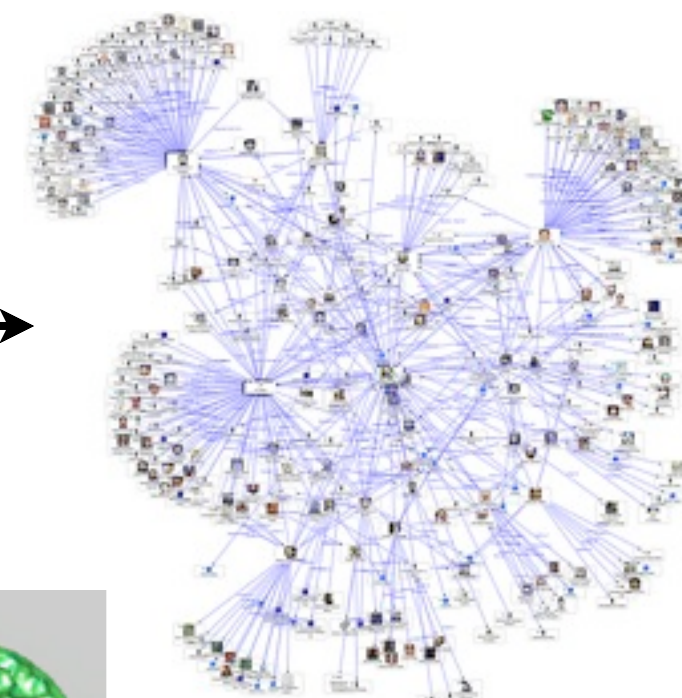
Images

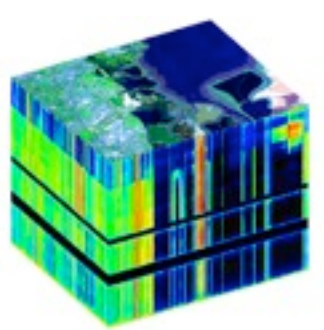
Sparse data models ?

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



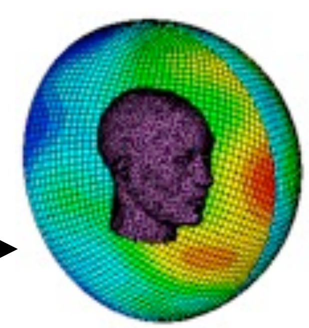
Data on graphs
Social networks
Brain connectivity

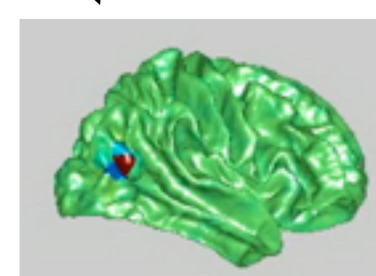




Hyperspectral
Satellite imaging

Spherical geometry
Cosmology, HRTF (3D audio)





Vector valued
Diffusion tensor

- "Exotic" or complex data
  - ✦ challenge = **generic models / model building tools**

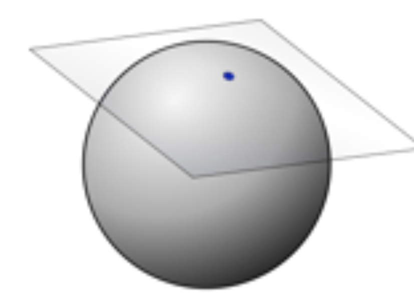
Local identifiability analysis (1)

- Normalization convention needed (as $\mathbf{A}X = \mathbf{B} \Rightarrow (2\mathbf{A})(X/2) = \mathbf{B}$)

$$\mathbf{A} \in \mathcal{A} := \{\|\mathbf{a}_k\|_2 = 1\}$$

- Compatible basis perturbations : tangent plane

$$\mathbf{A} + \delta_{\mathbf{A}} \qquad \delta_{\mathbf{A}} \in T_{\mathbf{A}}\mathcal{A}$$



- Coupling between basis and coefficients

$$\mathbf{A}X = \mathbf{B} \Rightarrow \delta_{\mathbf{A}} \cdot X + \mathbf{A} \cdot \delta_X = 0 \Rightarrow \delta_X = -\mathbf{A}^{-1}\delta_{\mathbf{A}} \cdot X$$

IRISA

Local identifiability analysis (2)

- First order approx. of L1 criterion (Λ = support of X)

$$\|X+\delta_X\|_1 - \|X\|_1 \approx \langle \delta_X, \mathrm{sign}(X)\rangle + \|(\delta_X)_{\bar{\Lambda}}\|_1$$

$$\forall \delta_X,\ |\langle \delta_X, \mathrm{sign}(X)\rangle| < \|(\delta_X)_{\bar{\Lambda}}\|_1 \Rightarrow$$ local minimum

- Admissible perturbations (from previous slide)

$$\delta_X = -\mathbf{A}^{-1}\delta_{\mathbf{A}} \cdot X \qquad \delta_{\mathbf{A}} \in T_{\mathbf{A}}\mathcal{A}$$

- (...) local minimum iff for zero-diagonal matrices Z

$$\forall Z \neq 0, |\langle Z, X\mathrm{sign}(X)^T - \mathrm{diag}(\|x_k\|_1)\mathbf{A}^T\mathbf{A}\rangle| < \|(ZX)_{\bar{\Lambda}}\|_1$$

- Decoupling between rows: for orthonormal $\mathbf{A}$ ...

$$|\langle z, X\mathrm{sign}(x_k)^T\rangle| < \|(X^T z)_{\bar{\Lambda}_k}\|_1$$

IRISA

jeudi 22 octobre 2009

Dictionary learning is *not* about ...

- Channel estimation with known pulse sequence

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

  - ✦ x = known channel input
  - ✦ $\mathbf{b}$ = observed channel output
  - ✦ $\mathbf{A}$ = unknown channel response

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

$$\mathbf{A} = \sum_k \alpha_k \mathbf{A}_k \qquad \mathbf{b} = \sum_k \alpha_k (\mathbf{A}_k x)$$

Dictionary learning is *not* about ...

- Cha[...]ion with know[...]uence
  - ✦ x = known cha[...]
  - ✦ $\mathbf{b}$ = observed [...]
  - ✦ $\mathbf{A}$ = unkn[...]es[...]

- [Ex[...] Identification of Matrices having a Spars[...]

$$[...] = \sum_k \alpha_k \mathbf{A}_k \qquad \mathbf{b} = \sum_k \alpha_k (\mathbf{A}_k [...]$$

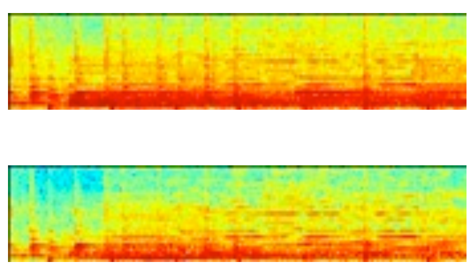
IRISA

Blind Source Separation

- Mixing model in the time-frequency domain

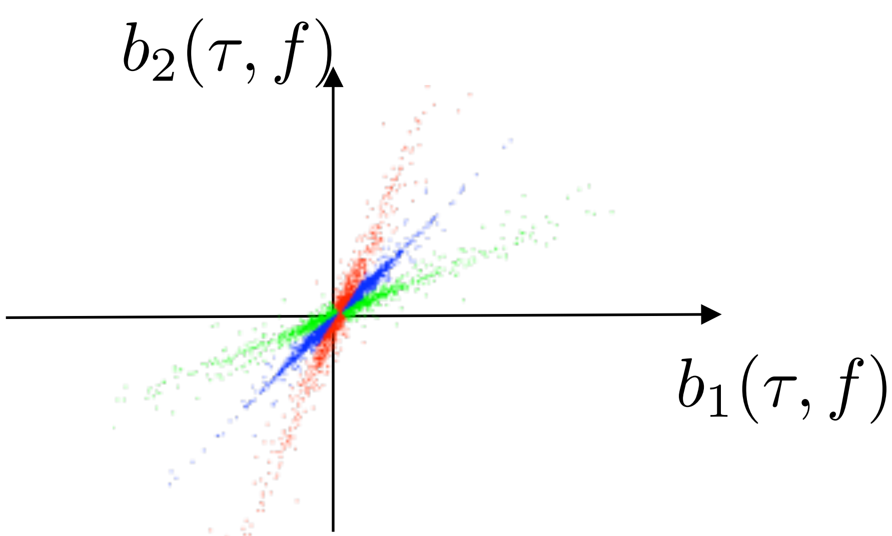
$b_1(\tau, f)$

$b_2(\tau, f)$



$$= \mathbf{A}\,\mathbf{X}(\tau, f)$$

- And "miraculously" ...



... 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

IRISA