Trading Recovery for Interpretability in CS

An heuristic and why it works

Kristiaan Pelckmans, Liu Yang

UU/IT, Sweden

Januari 2013



Context?



- System Identification (with Liang, ...)
- PID and ILC (with Johannes, Bengt)
- Machine Learning (with Yang, ...)
- Medical Applications (with Alexander, Raf, ...)
- Adaptive Filtering (with Toon, ...)
- Gossip and Distributed Computing (with Alexander, ...)

$\rightarrow$ Tools are common factor (optimization, concentration ineq., cauchy-Schwarz).

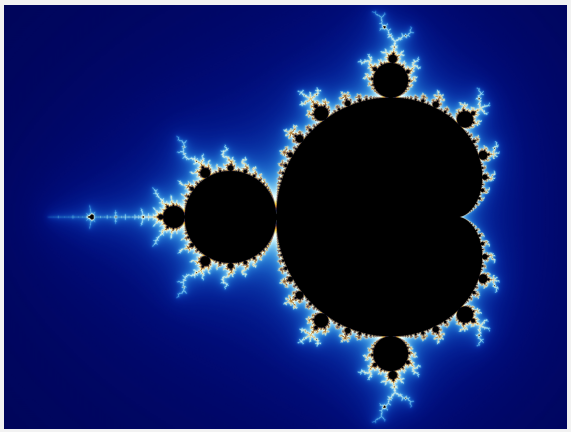
Overview



Today's talk:

1. Sparsity and Compressed Sensing.
2. Some theory.
3. An example.
4. Conclusions and outlooks.

Sparse Estimation (CS)?



- Suppose model

$$y = f(\mathbf{z}; \theta)$$

- Nonlinear (non-convex, non-smooth) in θ?
- But keep (physical) meaning θ.
- trick: choose

$$\theta \in \{\theta_1, \theta_2, \ldots, \theta_n\}$$

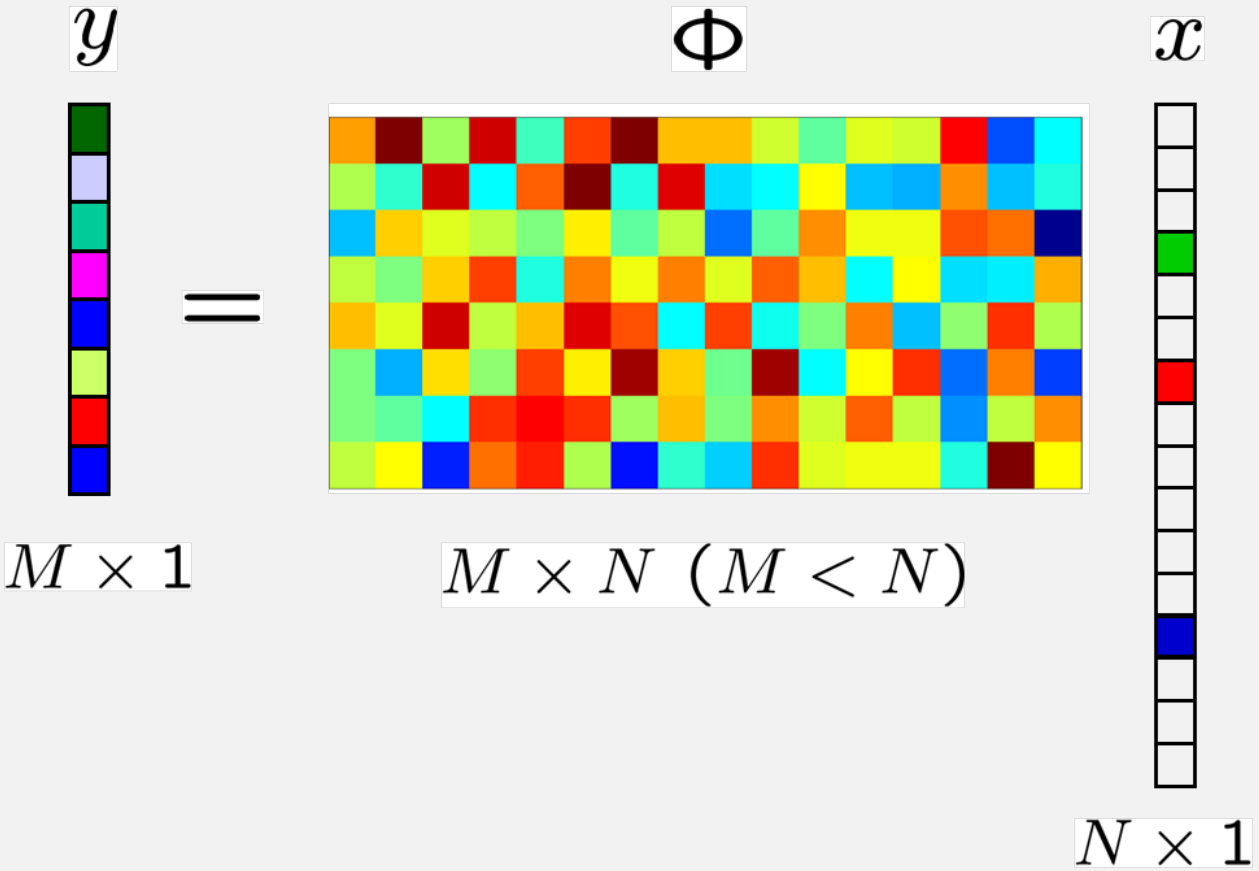
and let

$$y = \sum_{i=1}^{n} a_i f(\mathbf{z}; \theta_i)$$

and find 'sparse' vector $\mathbf{a} \in \mathbb{R}^n$

$$\mathbf{a} = (0, \ldots, 0, 1, 0, \ldots, 0)^T \in \mathbb{R}^n$$

Setup



Setup (Ct'd)

Let $k \ll n$. Given $\mathbf{A} \in \mathbb{R}^{k \times n}$ ($\mathbf{A}$ is *fat*). Assume that there exists $\mathbf{x}_0 \in \mathbb{R}^n$ with $\|\mathbf{x}_0\|_0 \leq s \ll k, n$ and let

$$(\textit{True}): \qquad \mathbf{y} = \mathbf{A}\mathbf{x}_0,$$

then

$$(L_0): \quad \tilde{\mathbf{x}} \in \arg\min_{\mathbf{x}} \|\mathbf{x}\|_0 \text{ s.t. } \mathbf{A}\mathbf{x} = \mathbf{y}$$

and BP

$$(L_1): \quad \hat{\mathbf{x}} \in \arg\min_{\mathbf{x}} \|\mathbf{x}\|_1 \text{ s.t. } \mathbf{A}\mathbf{x} = \mathbf{y}$$

What is the relation between $\mathbf{x}_0, \tilde{\mathbf{x}}, \hat{\mathbf{x}}$?

- Recovery: $\hat{\mathbf{x}} \sim \mathbf{x}_0$.
- Interpretability: $\hat{\mathbf{x}} \sim \tilde{\mathbf{x}}$.
- Stable: (approximately sparse).
- Robust: (towards measurement noise).

Spark

Definition (Spark)

The spark of $\mathbf{A}$ equals the minimal size of a linearly dependent set of columns of $\mathbf{A}$.

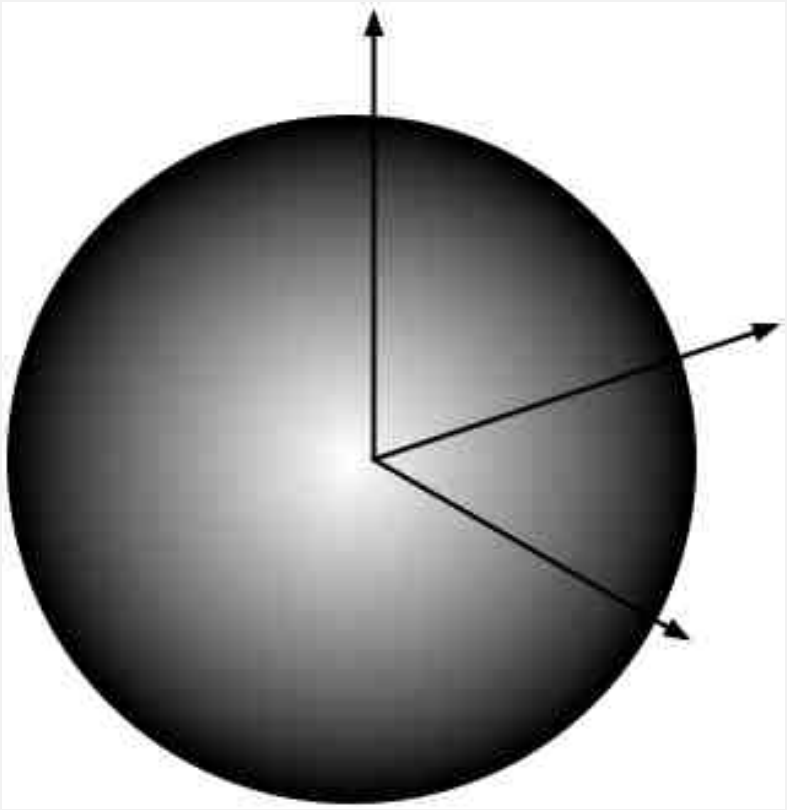
Theorem

Any $\mathbf{x}_0$ *with* $\|\mathbf{x}_0\|_0 \leq s$ *is a unique s-sparse solution* $\mathbf{z}$ *of* $\mathbf{Az} = \mathbf{Ax}_0$ *iff spark*$(\mathbf{A}) > 2s$.

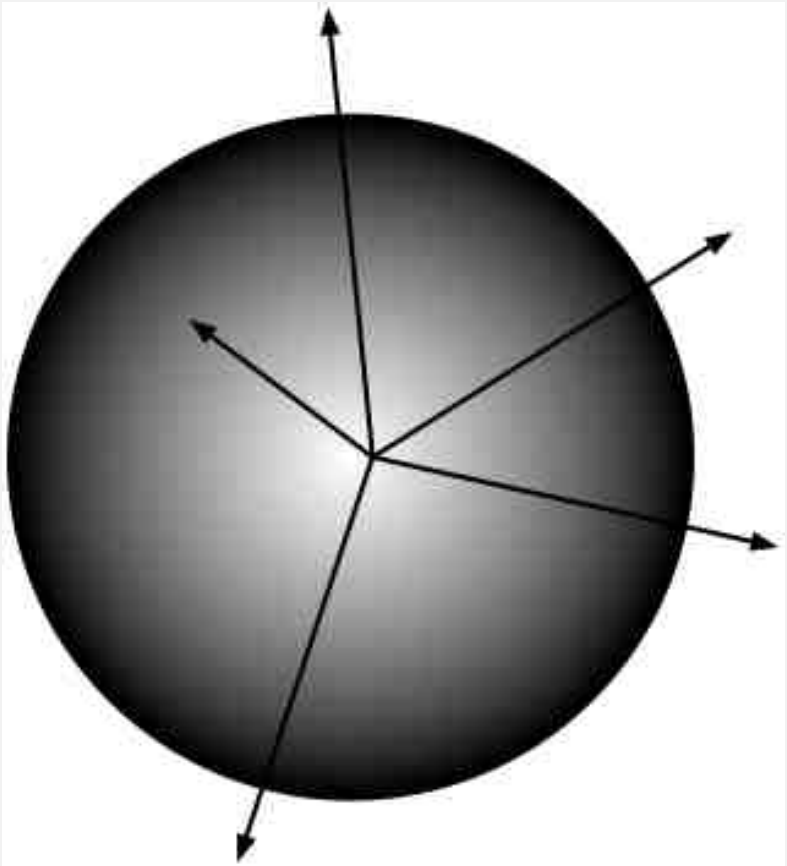
I.e. $\mathbf{x}_0 \leftrightarrow \tilde{\mathbf{x}}$.

Idea

Consequence of Johnsson-Lindenstrauss (1988):



3D $\rightarrow$ 3 linear independent vectors



3D $\rightarrow$ 5 approximate independent vectors

Coherence

Definition (Coherence)

The coherence of a matrix $\mathbf{A} \in \mathbb{R}^{k \times n}$ is defined as

$$\mu = \max_{i \neq j} |\mathbf{A}_i^T \mathbf{A}_j|$$

if $\mathbf{A}_i$ denotes the ith column of $\mathbf{A}$ and $\|\mathbf{A}_i\|_2 = 1$

Sufficient for $\tilde{\mathbf{x}} \sim \hat{\mathbf{x}}$.

Definition (RIP of order s)

A matrix $\mathbf{A} \in \mathbb{R}^{k \times n}$ is Restricted Iso-Perimetric (RIP) of order s if there exists a constant $\delta_s \geq 0$ such that

$$(1 - \delta_s)\|\mathbf{x}_0\|_2 \leq \|\mathbf{A}\mathbf{x}_0\|_2 \leq (1 + \delta_s)\|\mathbf{x}_0\|_2$$

for any $\mathbf{x}_s \in \mathbb{R}^n$ with $\|\mathbf{x}_0\|_s \leq s$.

Sufficient for $\hat{\mathbf{x}} \sim \mathbf{x}_0$.

Null-Space Property

Definition (Null-space Property)

A matrix $\mathbf{A} \in \mathbb{R}^{k \times n}$ is said to obey the Null Space Property (NSP) iff

$$\forall \mathbf{w} \neq 0, \mathbf{w} \in \ker(\mathbf{A}) : \|\mathbf{w}_S\|_1 < \|\mathbf{w}_{\bar{S}}\|_1$$

where $|S| \leq s$.

Necessary and Sufficient for $\hat{\mathbf{x}} \sim \mathbf{x}_0$ (Recovery).

Birds-eye View on Theory

Constructions of **A**

- Random ensembles.
- Structured matrices.
- Expanders.
- ...

$\leftrightarrow$

- Coh.
- RIP
- NSP

$\leftrightarrow$

Algorithms for Recovery:

- Thresholding (IHT, IST).
- Basis Pursuit.
- Matching Pursuit.
- OMP.
- CoSamp.
- ...

Trick

Basis Pursuit:

$$(A) : \mathbf{x} \in \arg\min \|\mathbf{x}\|_1 \;\; \text{s.t.} \;\; \mathbf{A}\mathbf{x} = \mathbf{y}$$

Let $\Delta \in \mathbb{R}^{k \times n}$ be a *small* perturbation. Hence $(\Delta\mathbf{x}) \approx 0$ for any $\mathbf{x}$ (concentration) if Δ i.i.d.

$$(B) : \mathbf{x}' \in \arg\min \|\mathbf{x}\|_1 \;\; \text{s.t.} \;\; (\mathbf{A} + \Delta)\mathbf{x} \approx \mathbf{y}$$

with 'error' $\Delta\mathbf{x}$. The Spark, NSP, RIP, ... of $(\mathbf{A} + \Delta)$ is however entirely different than from original $\mathbf{A}$.

An Example

$$\mathbf{A}\mathbf{x}_0 = \begin{bmatrix} 1 & 1 & \dots & 1 & -1 & -1 & \dots & -1 \\ -1 & -1 & \dots & -1 & 1 & 1 & \dots & 1 \\ -1 & -1 & \dots & -1 & -1 & -1 & \dots & -1 \\ \vdots & & & \vdots & & & & \\ 1 & 1 & \dots & 1 & -1 & -1 & \dots & -1 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_k \end{bmatrix}$$

Given $\mathbf{A}, \mathbf{y}$, a minimal L_0 solutions are also

$$\tilde{\mathbf{x}} = (0, 1 \dots, 0, 1, \dots, 0, 0)$$

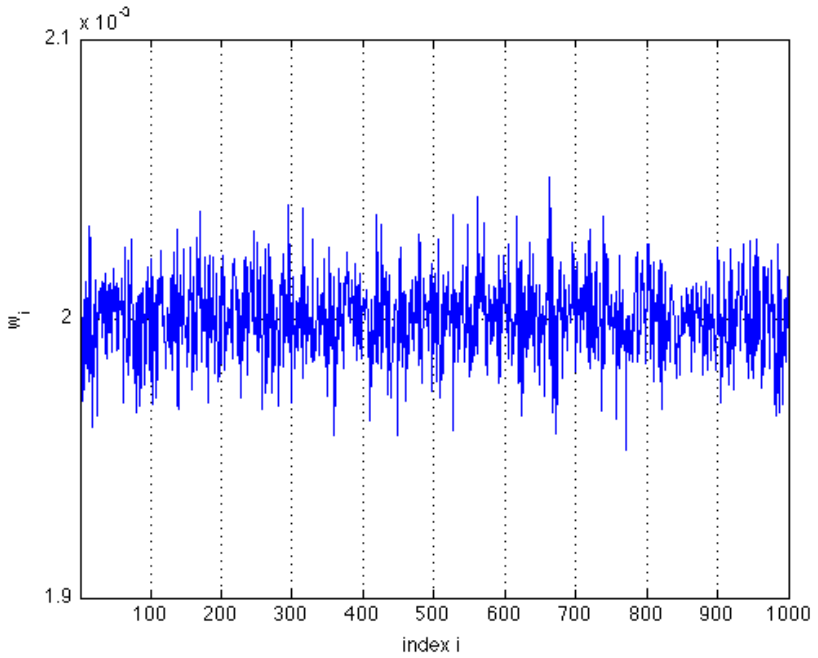
or

$$\tilde{\mathbf{x}}' = (1, 0 \dots, 0, 1, \dots, 0, 0)$$

A minimal L_1 solutions

$$\hat{\mathbf{x}} = \left(\frac{1}{n}, \frac{1}{n}, \dots, \frac{1}{n}, \frac{1}{n}, \dots, \frac{1}{n}, \frac{1}{n} \right)$$

An Example (Ct'd)



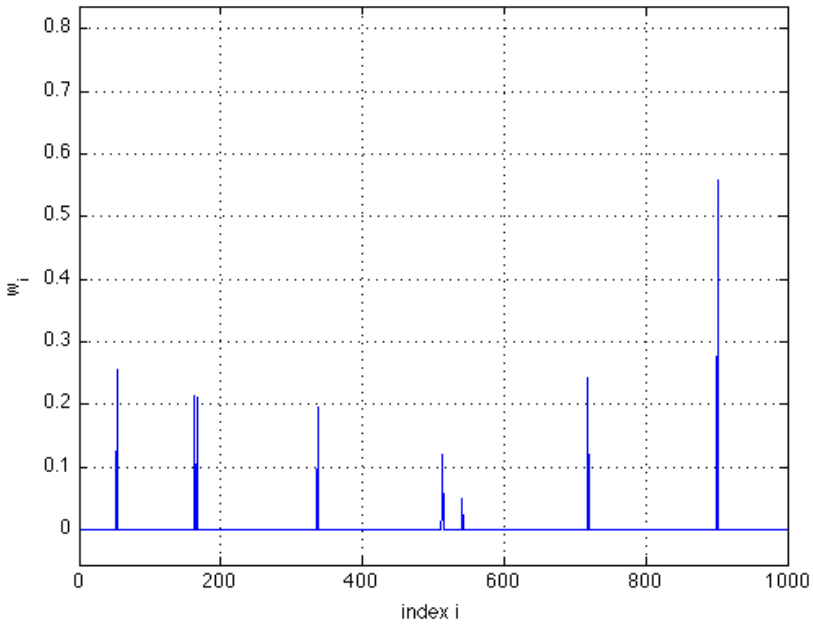


Figure: *The plots display the results of the procedure (A) (left), and of procedure (B). While the* true *solution* $\mathbf{w}_0 = (1, 0, \ldots, 0, 1)$ *is not recovered (there is no hope for this to happen since the data was not sufficiently rich), the solutions given by (B) are much more interpretable. Note that the fluctuations in the first panel are due to numerical issues with the used solver.*

An Example (Ct'd)

(n,k)	A: $\|\|\hat{\mathbf{w}}\|\|_0$	A: L_2 (test)	B: $\|\|\hat{\mathbf{w}}\|\|_0$	B: L_2 (test)
(10,10)	10	0.003873	2.07	0.003504
(10,50)	10	0.0033191	4.38	0.0032394
(10,200)	10	0.0031942	5.47	0.0031811
(100,10)	99	0.0043759	2.13	0.0035118
(100,50)	100	0.0032819	2.17	0.0032062
(100,200)	100	0.0031843	2.25	0.0031678
(1000,10)	1000	0.0038885	2.06	0.0038684
(1000,50)	1000	0.003296	2.07	0.0032293
(1000,200)	1000	0.0031845	2.13	0.0031696
(5000,10)	4850	0.0051792	2.03	0.0034888
(5000,50)	5000	0.0032852	2.06	0.0032314
(5000,200)	5000	0.0031982	2.12	0.0031827
(10000,10)	10000	0.0038838	2.59	0.003486
(10000,50)	10000	0.0032828	2.02	0.0032187
(10000,200)	10000	0.0031815	2.11	0.0031689

Conclusions



- $\leftarrow$ JL and CS.
- $\rightarrow$ Algorithms and Random Constructions of **A**.
- $+$ Unique Representation.
- $+$ Simple trick.
- $+$ Stability.
- ? Compressible and Stable.
- ? stability of spark, RIP and NSP?