Introduction to Machine Learning

Lecture 6: Sparse processing

Sparse processing

- Linear regression (with sparsity constraints)
- Sparse algorithms : convex optimization, greedy search, Bayesian analysis
- Applications : compression, parameter estimation, signal reconstruction, classification, Ex. Beamforming

Low-dimensional understanding of high-dimensional data sets

Sparse signals /compressive signals are important

- We don’t need to sample at the Nyquist rate
- Many signals are sparse, but we have solved them under non-sparse assumptions
  - Beamforming
  - Fourier transform
  - Layered structure
- Inverse methods are inherently sparse: We seek the simplest way to describe the data

But all this requires **new developments**

- Mathematical theory
- New algorithms (interior point solvers, convex optimization)
- Signal processing
- New applications/demonstrations

Slide 3 from Lecture-3 (Regression)

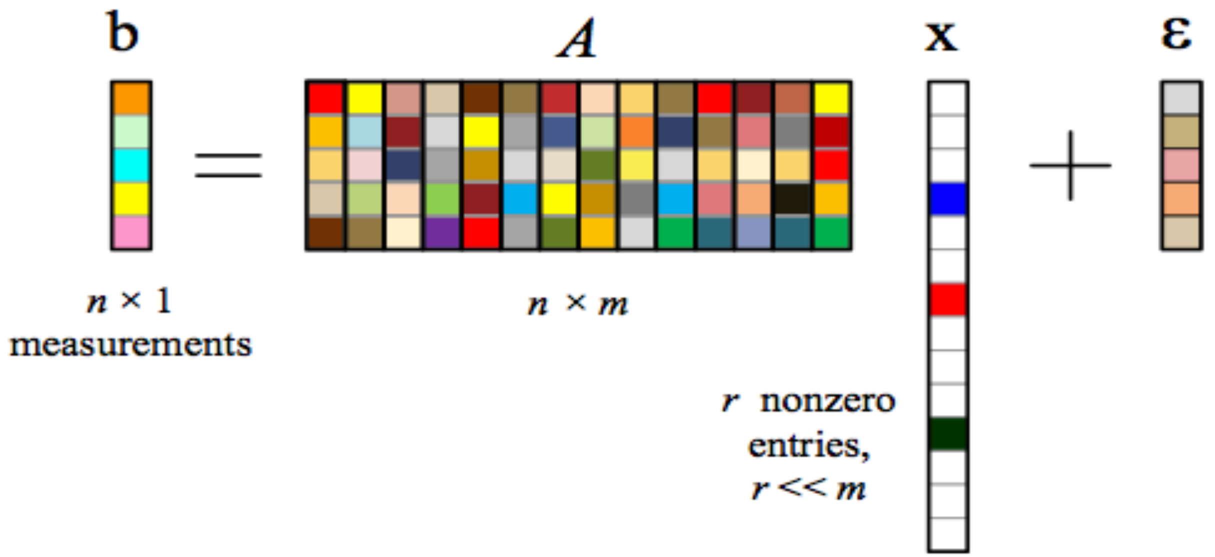
Linear Basis Function Models (2)

- Generally

$$y(\mathbf{x}, \mathbf{w}) = \sum_{j=0}^{M-1} w_j \phi_j(\mathbf{x}) = \mathbf{w}^{\mathrm{T}} \boldsymbol{\phi}(\mathbf{x})$$

- where $\phi_j(x)$ are known as *basis functions*.
- Typically, $\phi_0(x) = 1$, so that w_0 acts as a bias.
- In the simplest case, we use linear basis functions : $\phi_d(x) = x_d$.

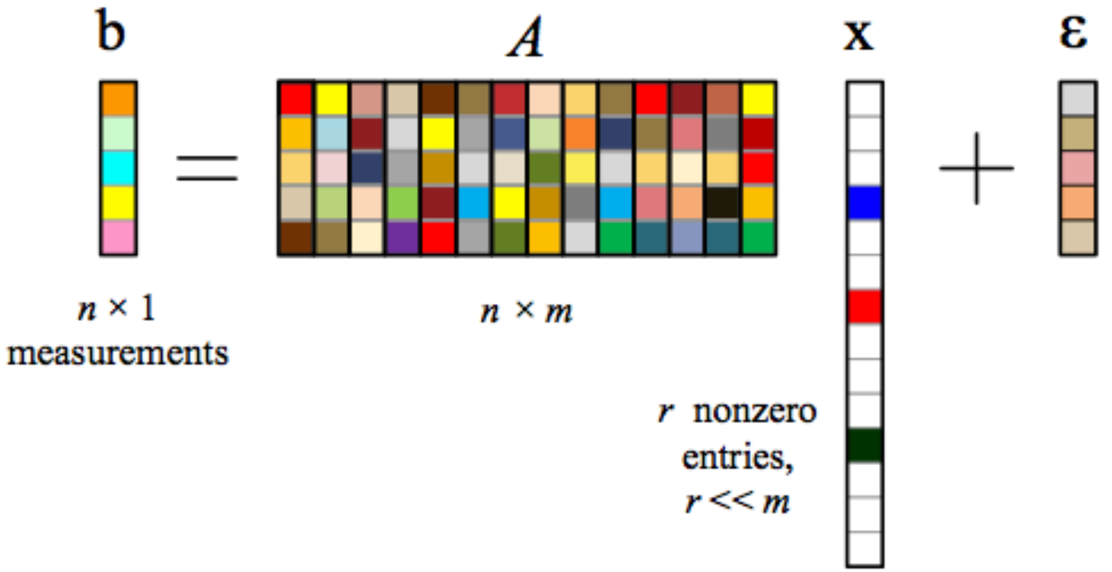
Compressed sensing formulation



- b is $n \times 1$ measurement vector
- A is $n \times m$ measurement/Dictionary matrix, $m >> n$
- x is $m \times 1$ desired vector which is sparse with r nonzero entries
- ε is the measurement noise

- An underdetermined system of equations has many solutions
- Utilizing x is sparse it can often be solved
- This depends on the structure of A (RIP!)

Different applications, but the same math



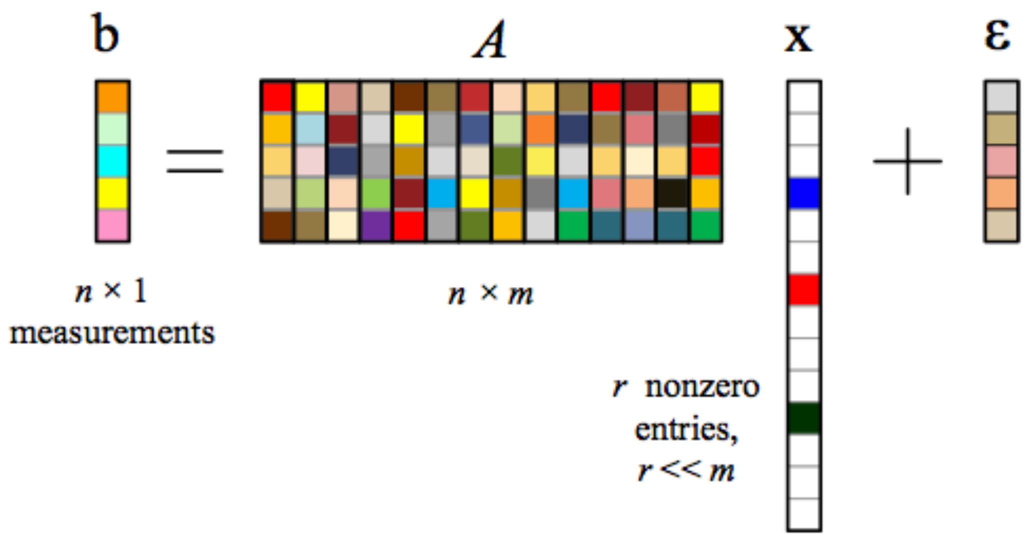
b	A	x
Frequency signal	DFT matrix	Time-signal
Compressed-Image	Random matrix	Pixel-image
signals	Beam weight	Source-location
Reflection sequence	Time delay	Layer-reflector

Compressive Sensing / Sparse Recovery

- **Alternative viewpoint:** We try to find the sparsest solution which explains our noisy measurements

$$\min_{x} \|\mathbf{x}\|_0 \qquad \text{subject to } \|\mathbf{Ax}-\mathbf{b}\|_2 < \varepsilon$$

- Here, the l_0-norm is a shorthand notation for *counting the number of non-zero elements in* x.



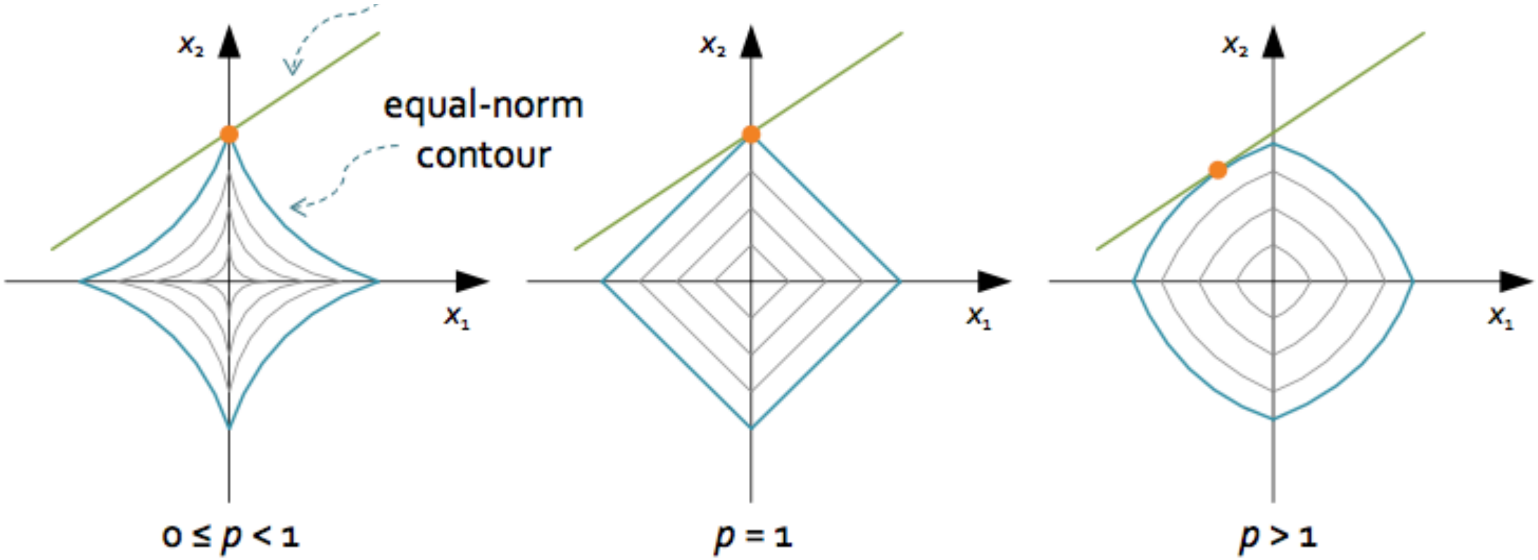
l_p-Norms

$$\| x \|_p = \left(\sum_{m=1}^{M} | x_m |^p \right)^{1/p} \quad \text{for} \quad p > 0$$

- Classic choices for p are 1, 2, and ∞.
- We will abuse notation and allow also $p = 0$.

Norms

$$\|x\|_p = \left(\sum_{m=1}^{M} |x_m|^p \right)^{1/p}$$



Solutions

- Regularized Inverse
- Orthogonal matching pursuit (OMP)
- Basis pursuit denoising
- Sparse Bayesian Learning

Inverse Techniques

- For the systems of equations Ax = b, the solution set is characterized by $\{x_s : x_s = A^+ y + v;\ v \in N(A)\}$, where N(A) denotes the null space of A and $A^+ = A^T(AA^T)^{-1}$.

- Minimum Norm solution: The minimum ℓ_2 norm solution

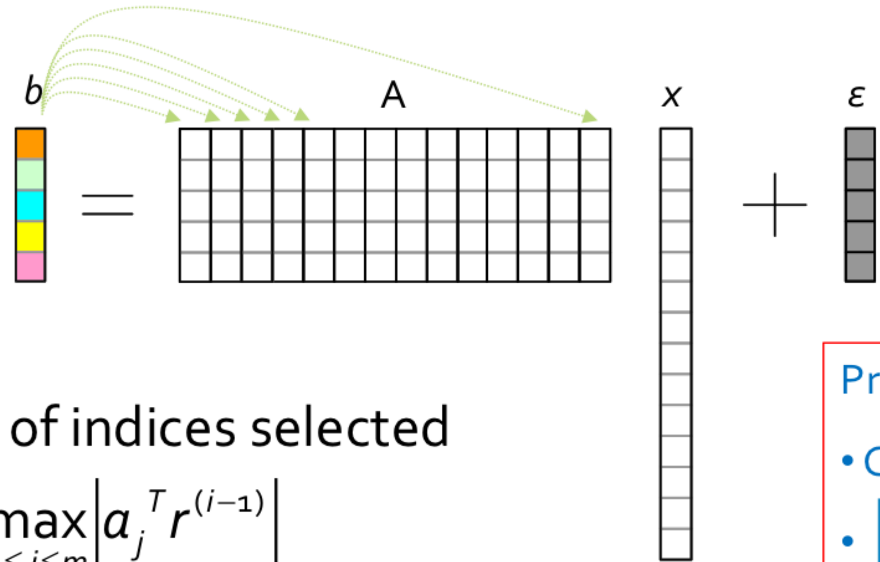
  $x_{mn} = A^+b$ is a popular solution

- Noisy Case: regularized ℓ_2 norm solution often employed and is given by

$$x_{reg} = A^T(AA^T + \lambda I)^{-1} b$$

Greedy Search Method: Matching Pursuit

- Select a column that is most aligned with the current residual



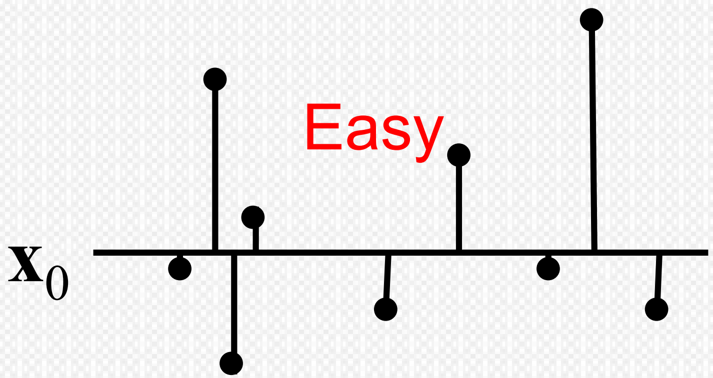
  - $r^{(0)} = b$
  - $S^{(i)}$: set of indices selected
  - $l = \underset{1 \le j \le m}{\operatorname{argmax}} \left| a_j^T r^{(i-1)} \right|$
- Remove its contribution from the residual
  - Update $S^{(i)}$: If $l \notin S^{(i-1)}, S^{(i)} = S^{(i-1)} \bigcup \{l\}$. Or, keep $S^{(i)}$ the same
  - Update $r^{(i)}$: $r^{(i)} = P_{a_l}^{\perp} r^{(i-1)} = r^{(i-1)} - a_l a_l^T r^{(i-1)}$

Practical stop criteria:

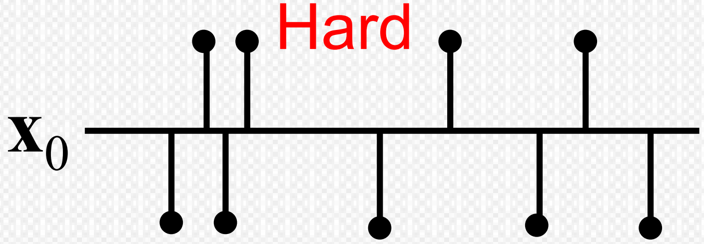
- Certain # iterations
- $\left\| r^{(i)} \right\|_2$ smaller than threshold

Amplitude Distribution

- If the magnitudes of the non-zero elements in $\mathbf{x}_0$ are highly scaled, then the canonical sparse recovery problem should be easier.



scaled coefficients (easy)



uniform coefficients (hard)

For strongly scaled coefficients, Matching Pursuit (or Orthogonal MP) works better. It picks one coefficient at a time.

Basis Pursuit / LASSO

- The l_0-norm minimization is not convex and requires combinatorial search.
- We convexify by substituting the l_1-norm in place of the l_0-norm.

$$\min_x \|\mathbf{x}\|_1 \qquad \text{subject to} \quad \|\mathbf{Ax}-\mathbf{b}\|_2 < \varepsilon$$

- This can also be formulated as

$$\min_x \|\mathbf{x}\|_1 + \lambda \|\mathbf{Ax}-\mathbf{b}\|_2$$

$$\min_x \|\mathbf{Ax}-\mathbf{b}\|_2 + \mu \|\mathbf{x}\|_1$$

$$\min_x \|\mathbf{Ax}-\mathbf{b}\|_2 \qquad \text{subject to} \quad \|\mathbf{x}\|_1 < \delta$$

Basis Pursuit / LASSO

- Why is it legal to substitute the l_1-norm for the l_0-norm?
- What are the conditions such that the two problems have the same solution?

$$\min_x \|x\|_1 \quad \text{subject to} \quad \|Ax - b\|_2 < \varepsilon$$

$$\min_x \|x\|_0 \quad \text{subject to} \quad \|Ax - b\|_2 < \varepsilon$$

Restricted Isometry Property (RIP)

$$(1 - \delta_s)\|\boldsymbol{u}\|_2 \leq \|\boldsymbol{A}_{\boldsymbol{S}}\, u\|_2 \leq (1 + \delta_s)\|\boldsymbol{u}\|_2$$

The unconstrained -LASSO- formulation

Constrained formulation of the ℓ_1-norm minimization problem:

$$\widehat{\mathbf{x}}_{\ell_1}(\epsilon) = \arg\min_{\mathbf{x}\in\mathbb{C}^N} \|\mathbf{x}\|_1 \text{ subject to } \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2 \leq \epsilon$$

Unconstrained formulation in the form of least squares optimization with an ℓ_1-norm regularizer:

$$\widehat{\mathbf{x}}_{\text{LASSO}}(\mu) = \arg\min_{\mathbf{x}\in\mathbb{C}^N} \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2 + \mu\|\mathbf{x}\|_1$$

For every ϵ exists a μ so that the two formulations are equivalent

Regularization parameter : μ

Regularization parameter selection

The objective function of the LASSO problem:

$$L(\mathbf{x}, \mu) = \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2 + \mu\|\mathbf{x}\|_1$$

is minimized if

$$\mathbf{0} \in \partial_{\mathbf{x}} L(\mathbf{x}, \mu)$$

where the subgradient is

$$\partial_{\mathbf{x}} L(\mathbf{x}, \mu) = 2\mathbf{A}^H (\mathbf{A}\mathbf{x} - \mathbf{y}) + \mu\partial_{\mathbf{x}}\|\mathbf{x}\|_1$$

thus, the global minimum is attained if

$$\mu^{-1}\mathbf{r} \quad \in \quad \partial_{\mathbf{x}}\|\mathbf{x}\|_1, \quad \mathbf{r} = 2\mathbf{A}^H (\mathbf{y} - \mathbf{A}\widehat{\mathbf{x}})$$

Regularization parameter selection

The global minimum is attained if

$$\mu^{-1}\mathbf{r} \quad \in \quad \partial_{\mathbf{x}}\|\mathbf{x}\|_1, \quad \mathbf{r} = 2\mathbf{A}^H(\mathbf{y} - \mathbf{A}\widehat{\mathbf{x}})$$

The subgradient for the ℓ_1-norm is the set of vectors

$$\partial_{\mathbf{x}}\|\mathbf{x}\|_1 = \left\{\mathbf{s}: \ \|\mathbf{s}\|_\infty \leq 1, \ \mathbf{s}^H\mathbf{x} = \|\mathbf{x}\|_1\right\}$$

which implies

$$\begin{aligned} s_i &= \tfrac{x_i}{|x_i|}, & x_i &\neq 0 \\ |s_i| &\leq 1, & x_i &= 0, \end{aligned}$$

thus,

$$\begin{aligned} |r_i| &= \mu, & \widehat{x}_i &\neq 0 \\ |r_i| &\leq \mu, & \widehat{x}_i &= 0 \end{aligned}$$

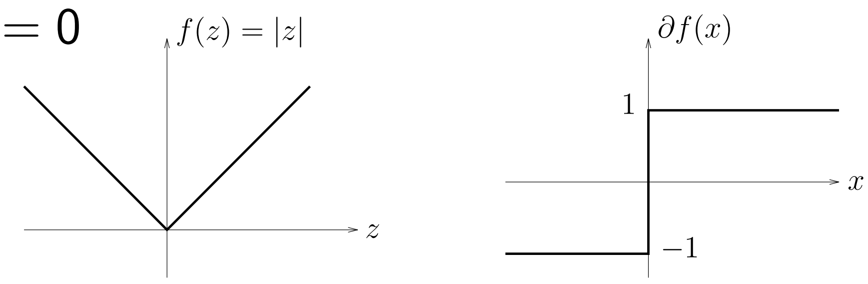
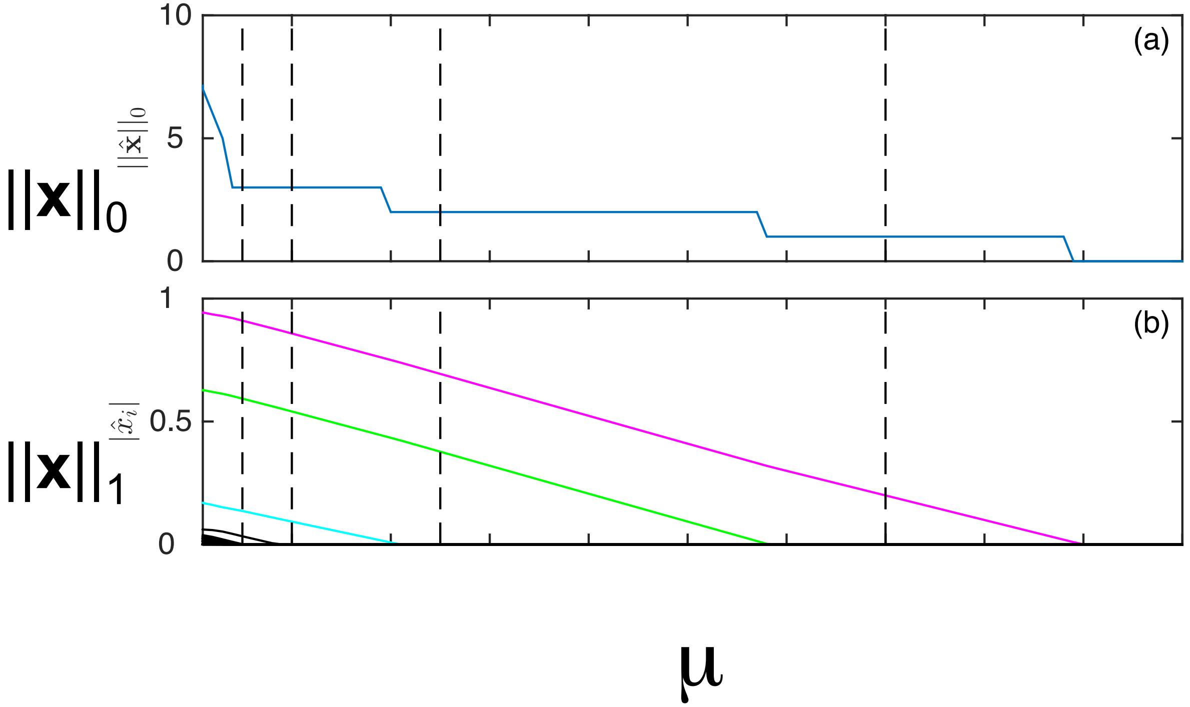


Figure 3: The absolute value function (left), and its subdifferential $\partial f(x)$ as a function of x (right).

Lasso Path



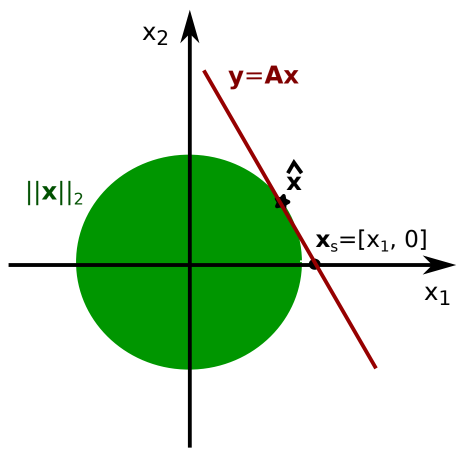
Solving an underdetermined problem

$$\mathbf{y} = \mathbf{A}_{M\times N}\mathbf{x}, \qquad M < N$$

$\mathbf{x}$: K-sparse, $K \ll N$

l_2-norm minimization (min energy)

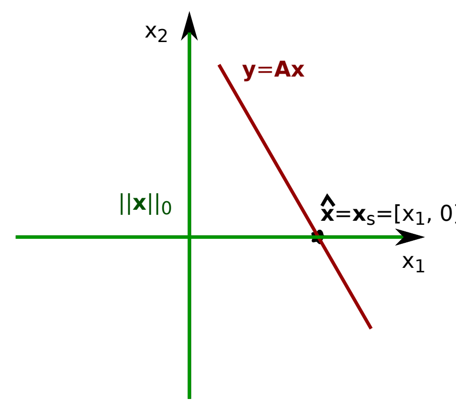
$$\min_{\mathbf{x}\in\mathbb{C}^N} \|\mathbf{x}\|_2 \text{ subject to } \mathbf{y} = \mathbf{A}\mathbf{x}$$



$$\hat{\mathbf{x}} = \mathbf{A}^H\left(\mathbf{A}\mathbf{A}^H\right)^{-1}\mathbf{y}$$

l_0-norm minimization (min sparsity)

$$\min_{\mathbf{x}\in\mathbb{C}^N} \|\mathbf{x}\|_0 \text{ subject to } \mathbf{y} = \mathbf{A}\mathbf{x}$$



$\hat{\mathbf{x}}$: combinatorial intractable problem

The l_2-solution has minimum energy while the l_0-solution is sparse

Compressive sensing

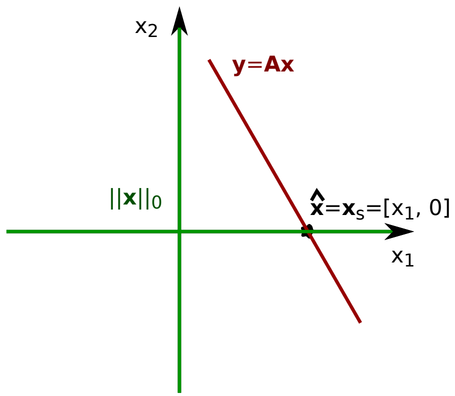
$$\mathbf{y} = \mathbf{A}_{M\times N}\mathbf{x},\ \ M < N,$$

$\mathbf{x}$: K-sparse, $K \ll N,\ K < M$

$\mathbf{A} = [\mathbf{a}_1, \cdots, \mathbf{a}_N] :\ |\mathbf{a}_i^H \mathbf{a}_j|_{i\neq j} < 1$

l_0-norm minimization (min sparsity)

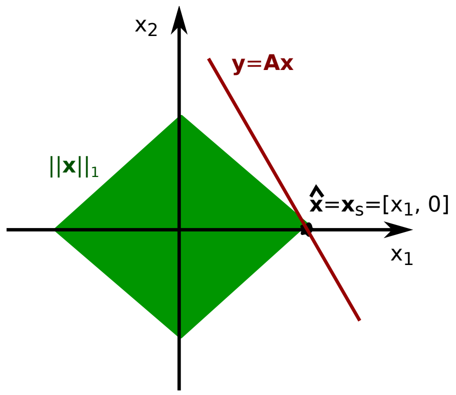
$$\min_{\mathbf{x}\in\mathbb{C}^N} \|\mathbf{x}\|_0 \text{ subject to } \mathbf{y} = \mathbf{A}\mathbf{x}$$



$\hat{\mathbf{x}}$: combinatorial intractable problem

l_1-norm convex relaxation

$$\min_{\mathbf{x}\in\mathbb{C}^N} \|\mathbf{x}\|_1 \text{ subject to } \mathbf{y} = \mathbf{A}\mathbf{x}$$



$$\hat{\mathbf{x}} = \underset{\mathbf{x}\in\mathbb{C}^N}{\operatorname{argmin}} \|\mathbf{x}\|_1 \text{ subject to } \mathbf{y} = \mathbf{A}\mathbf{x}$$

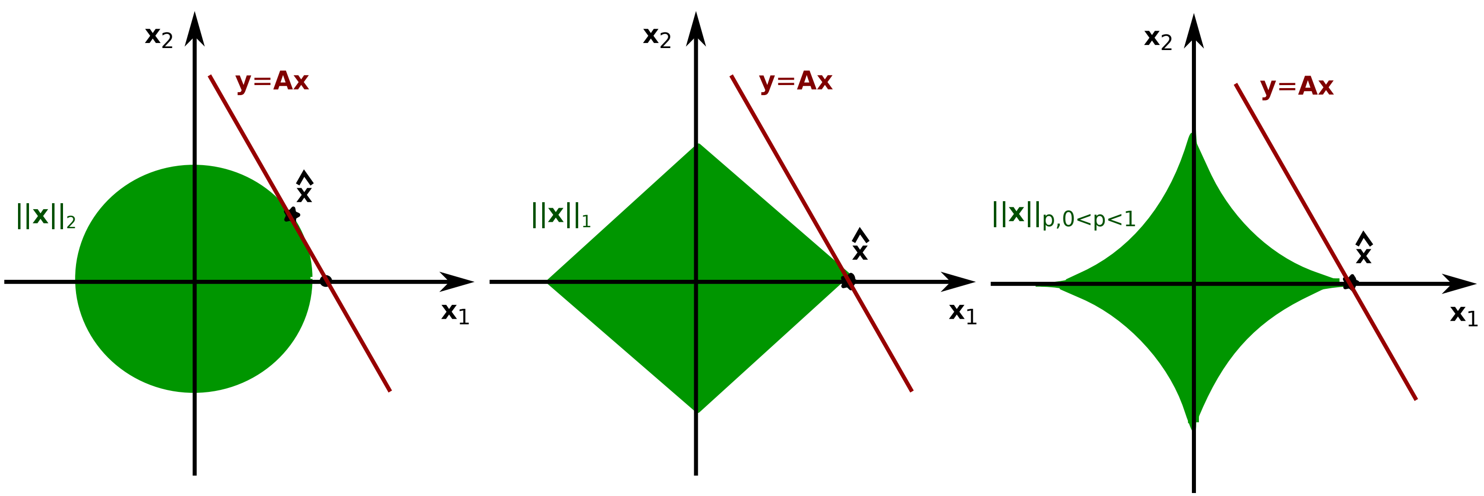
The l_1-problem is both convex and promotes sparse solutions

Enhancing sparsity

$$\underset{\mathbf{x}\in\mathbb{C}^n}{\arg\min} J(\mathbf{x}) \text{ subject to } \|\mathbf{A}\mathbf{x}-\mathbf{y}\|_2 \leq \epsilon$$

$$J(\mathbf{x}) = \begin{cases} \|\mathbf{x}\|_p^p = \sum_{i=1}^{N} |x_i|^p, \ 0 < p < 1 \\ \sum_{i=1}^{N} \ln\left(|x_i|\right) \end{cases}, \text{ concave}$$

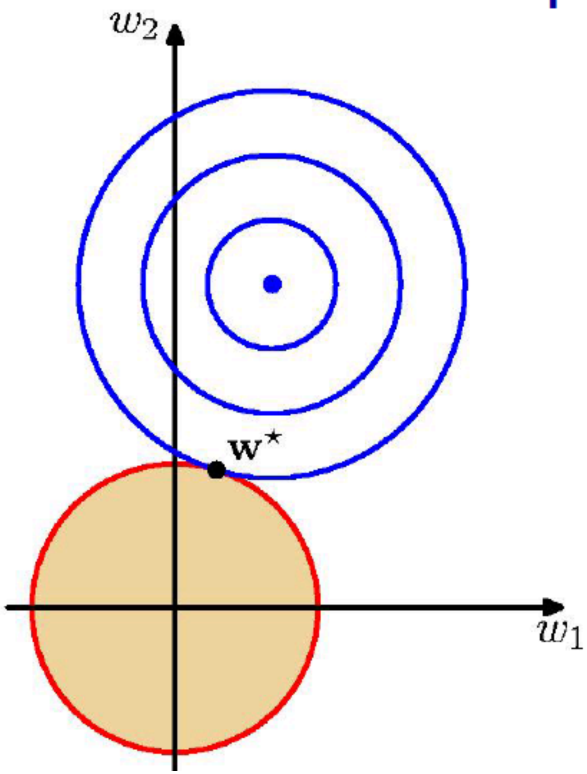
$p = 2$ $\qquad$ $p = 1$ $\qquad$ $0 < p < 1$

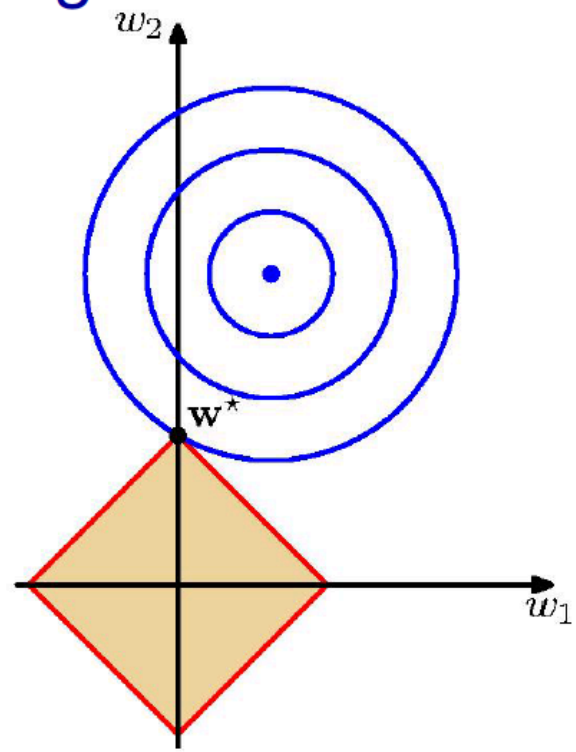


Minimization of a concave function with an iterative majorization-minimization algorithm

Geometrical view

Geometrical view of the lasso compared with a penalty on the squared weights



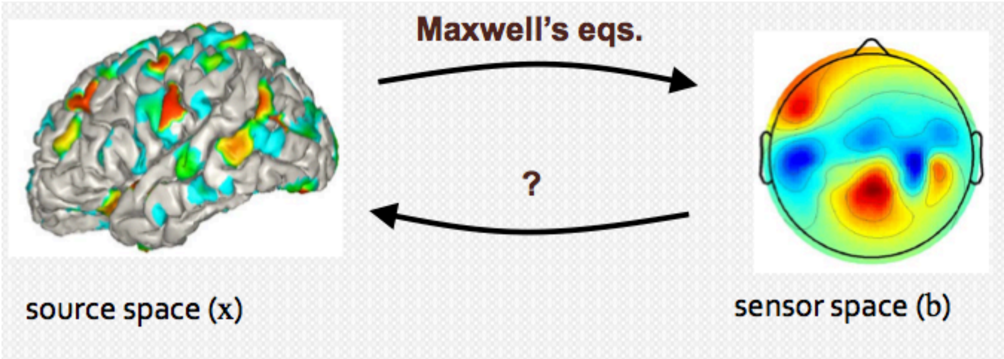


Applications

- MEG/EEG/MRI source location (earthquake location)
- Channel equalization
- Compressive sampling (beyond Nyquist sampling!)
- Compressive camera!

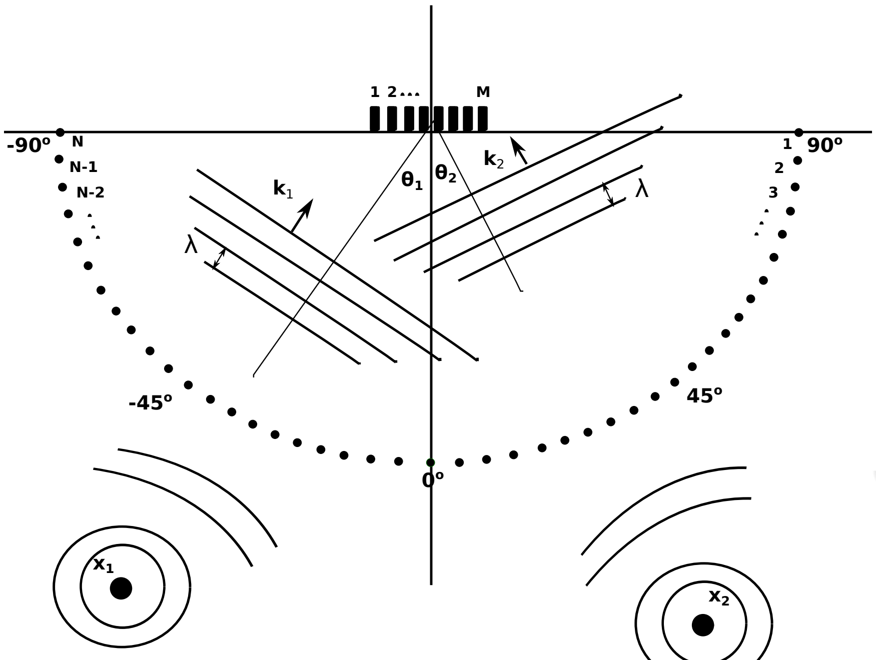
Lots of low hanging fruits

- Beamforming
- Fathometer
- Geoacoustic inversion
- Sequential estimation
- Bayesian
- Grid free methods



Beamforming / DOA estimation

DOA estimation with sensor arrays



$p_1(\mathbf{r},t) = x_1\, e^{j(\omega t - \mathbf{k}_1 \mathbf{r})}$ $p_2(\mathbf{r},t) = x_2\, e^{j(\omega t - \mathbf{k}_2 \mathbf{r})}$

$$x \in \mathbb{C},\ \theta \in [-90^\circ, 90^\circ]$$

$$\mathbf{k} = -\frac{2\pi}{\lambda}\sin\theta,\ \lambda\text{:wavelength}$$

$$y_m = \sum_n x_n e^{j\frac{2\pi}{\lambda} r_m \sin\theta_n}$$

$m \in [1, \cdots, M]$: sensor
$n \in [1, \cdots, N]$: look direction

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

$$\mathbf{y} = [y_1, \cdots, y_M]^T, \quad \mathbf{x} = [x_1, \cdots, x_N]^T$$

$$\mathbf{A} = [\mathbf{a}_1, \cdots, \mathbf{a}_N]$$

$$\mathbf{a}_n = \frac{1}{\sqrt{M}}[e^{j\frac{2\pi}{\lambda} r_1 \sin\theta_n}, \cdots, e^{j\frac{2\pi}{\lambda} r_M \sin\theta_n}]^T$$

The DOA estimation is formulated as a linear problem

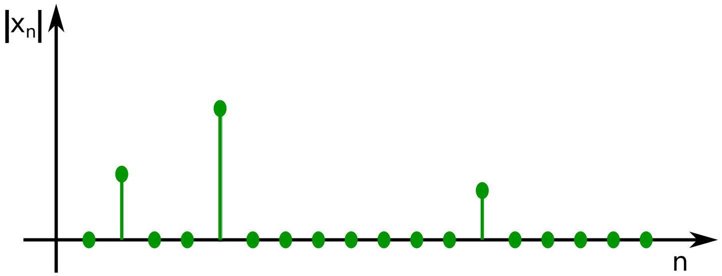
Sparse representation of the DOA estimation problem

Underdetermined problem

$$\mathbf{y} = \mathbf{A}\mathbf{x}, \quad M < N$$

Prior information

$$\mathbf{x}\text{: K-sparse, } K \ll N$$

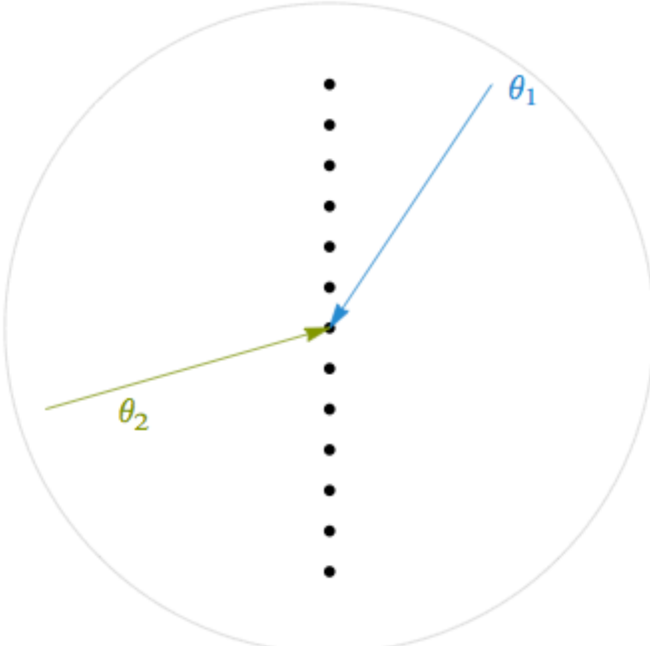


$$\|\mathbf{x}\|_0 = \sum_{n=1}^{N} 1_{x_n \neq 0} = K$$

Not really a norm: $\|a\mathbf{x}\|_0 = \|\mathbf{x}\|_0 \neq |a|\|\mathbf{x}\|_0$

There are only few sources with unknown locations and amplitudes

Direction of arrival estimation



Plane waves from a source/interferer impinging on an array/antenna

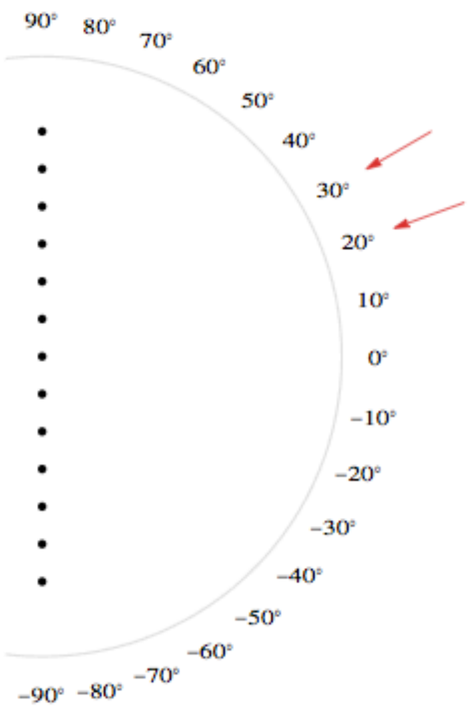
True DOA is sparse in the angle domain

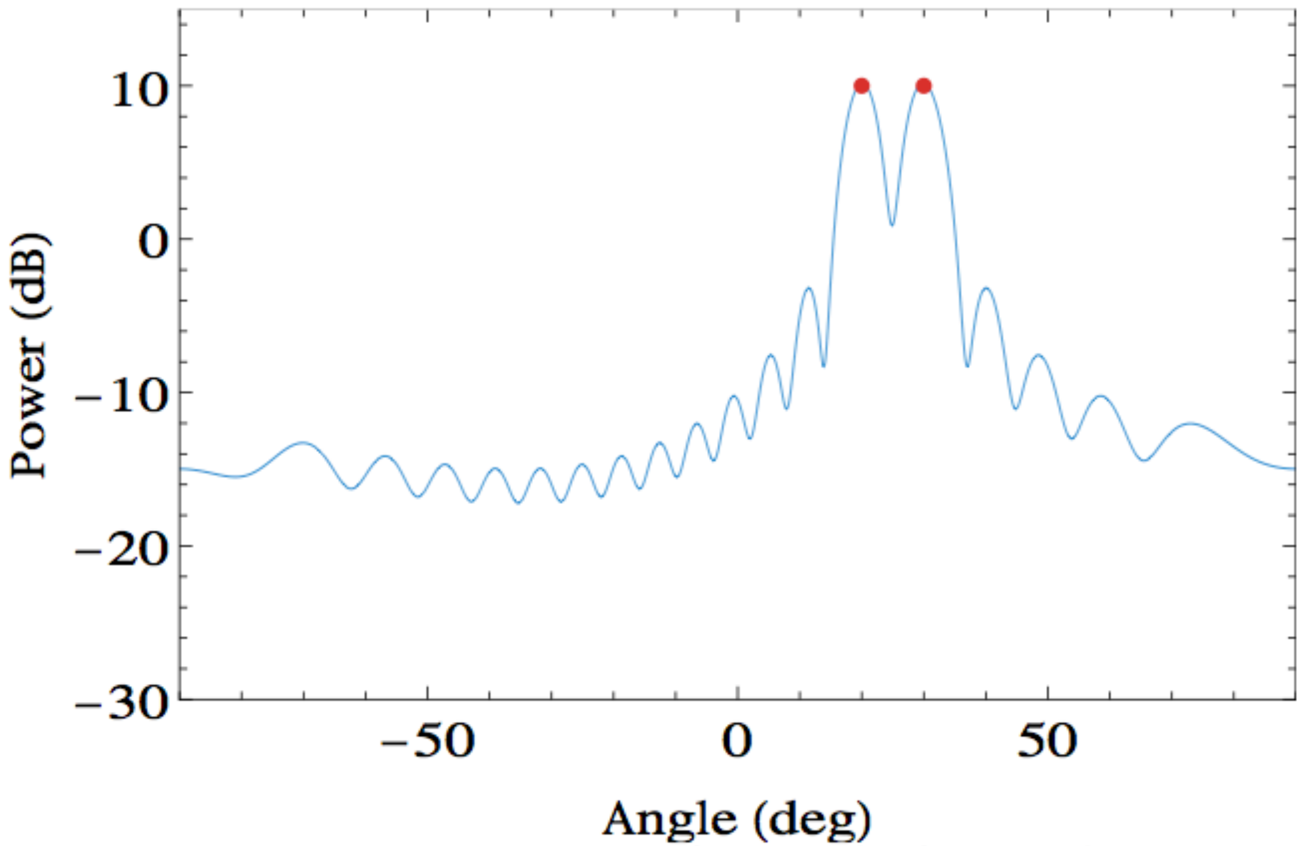
$\boldsymbol{\Theta} = \{0, \cdots, 0, \theta_1, 0, \cdots, 0, \theta_2, 0, \cdots, 0\}$

Conventional beamforming

Plane wave weight vector $\mathbf{w}_i = [1, e^{-\imath \sin(\theta_i)}, \cdots, e^{-\imath (N-1) \sin(\theta_i)}]^T$

$$\mathcal{B}(\theta) = |\mathbf{w}^H(\theta)\mathbf{b}|^2$$



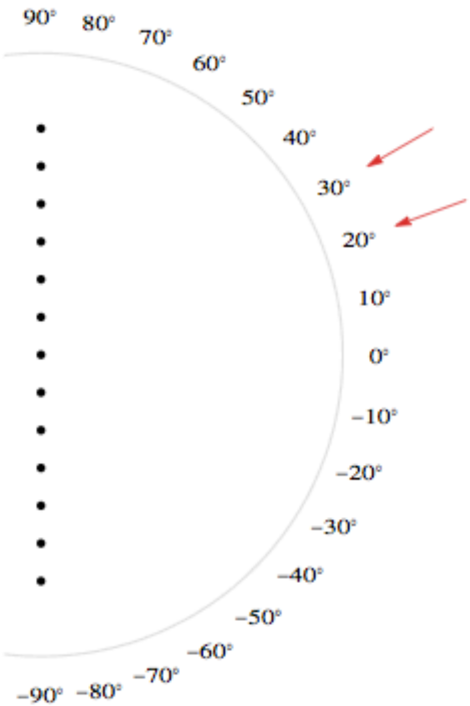


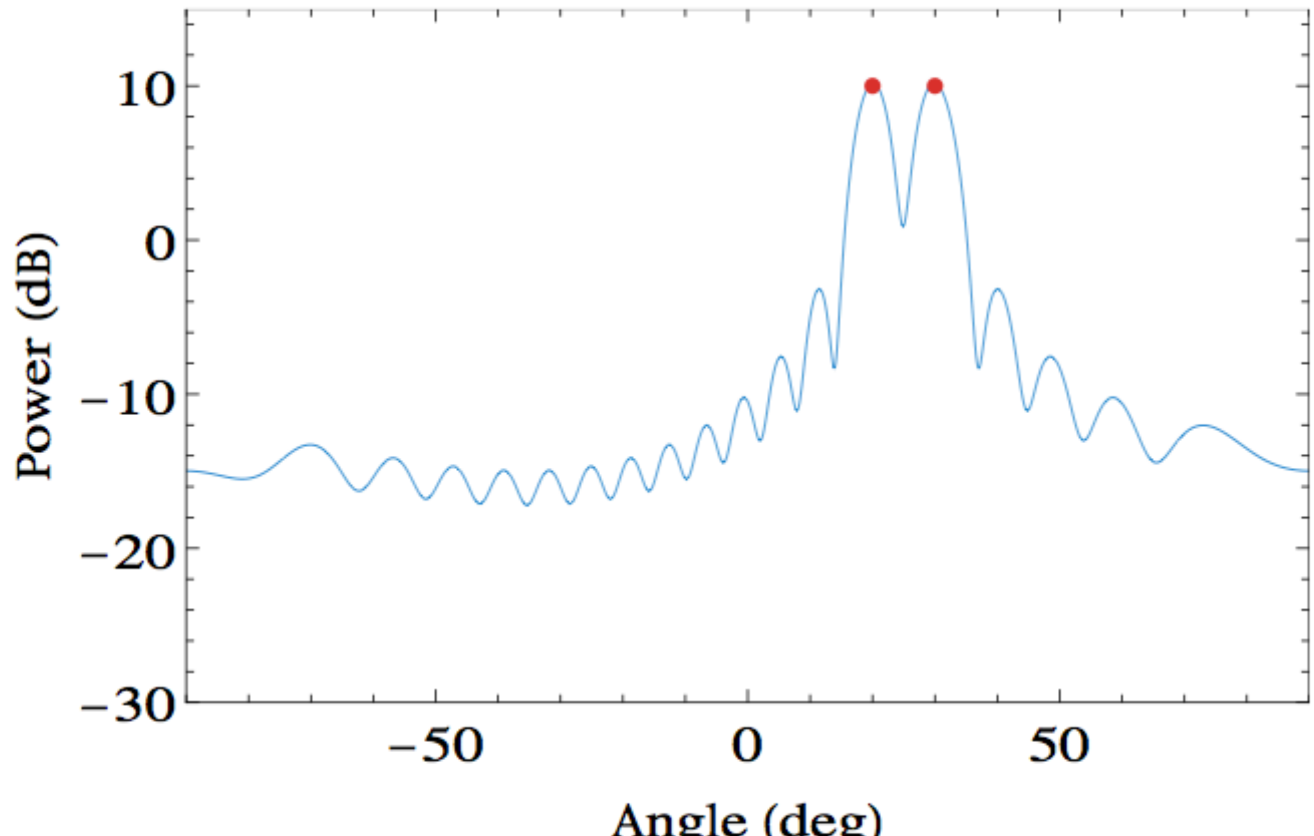
ULA, half-wavelength spacing, $N = 20$ sensors, $\theta_1 = 20°$, $\theta_2 = 30°$,

Conventional beamforming

Equivalent to solving the ℓ_2 problem with $\mathbf{A} = [\mathbf{w}_1, \cdots, \mathbf{w}_M]$, $M > N$.

$$\min \|\mathbf{x}\|_2 \text{ subject to } \mathbf{A}\mathbf{x} = \mathbf{b}$$



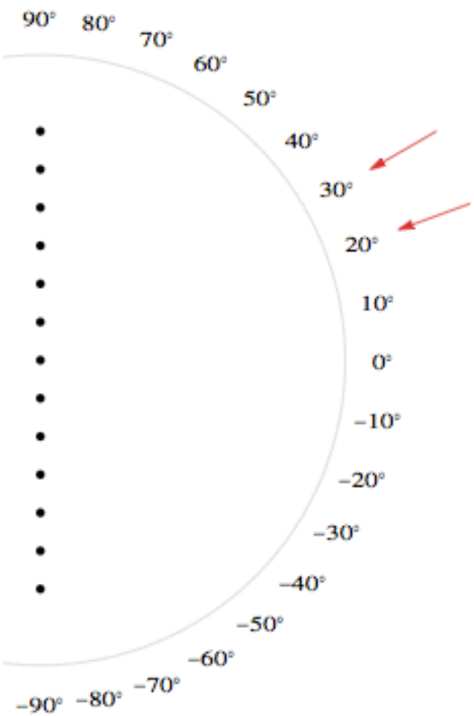


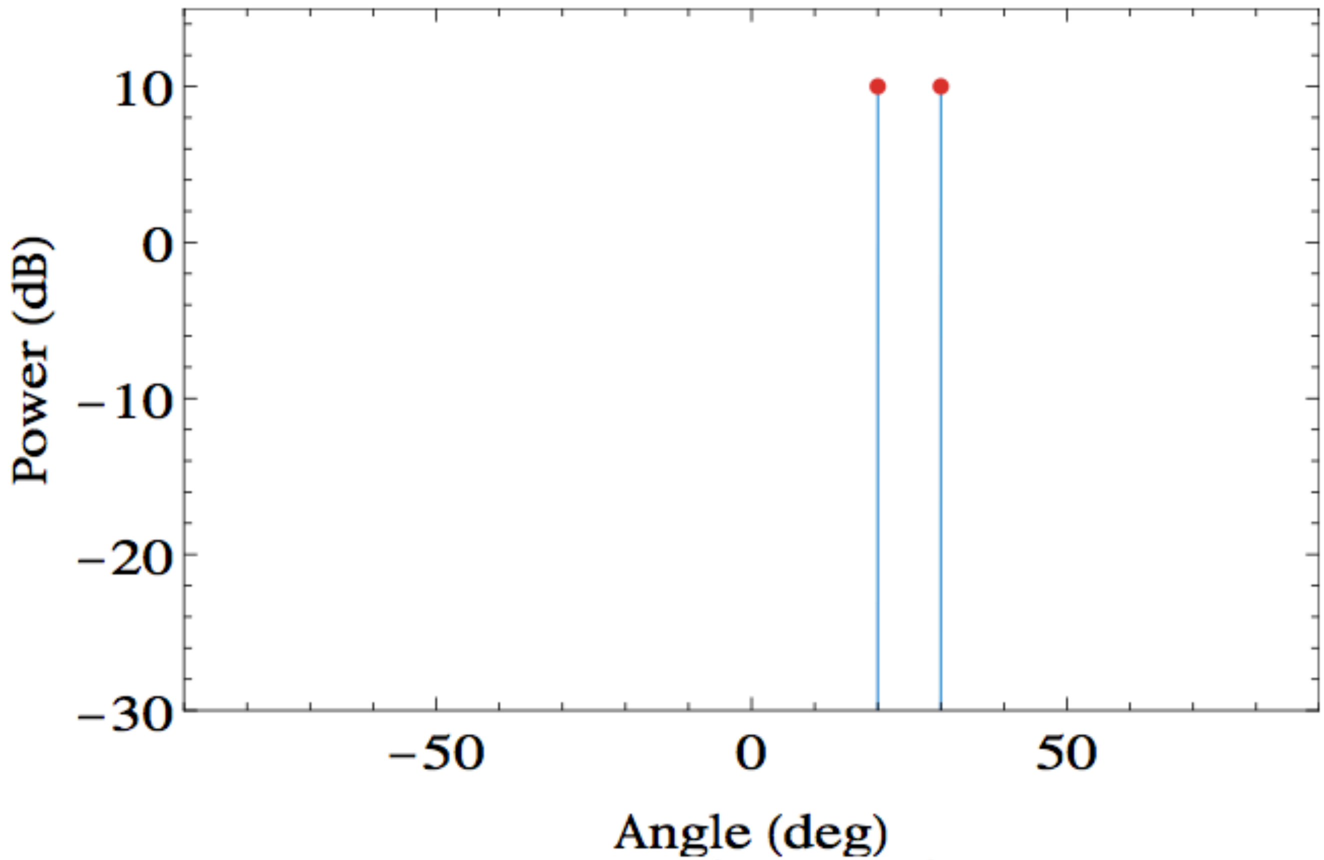
A is an overcomplete dictionary of candidate DOA vectors. Columns span -90° to 90° in steps of 1° ($M = 181$).

ℓ_1 minimization

In contrast ℓ_1 minimization provides a sparse solution with exact recovery:

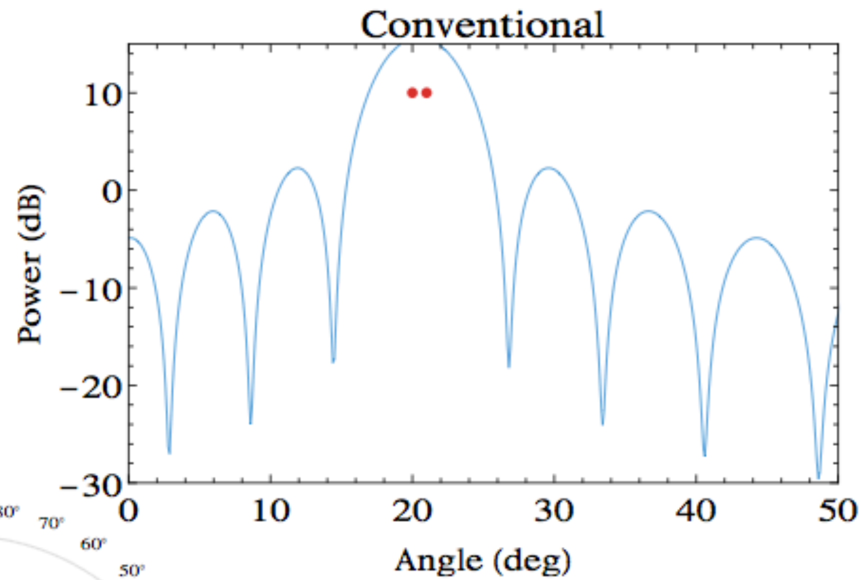
$$\min \|\mathbf{x}\|_1 \text{ subject to } \mathbf{Ax} = \mathbf{b}$$

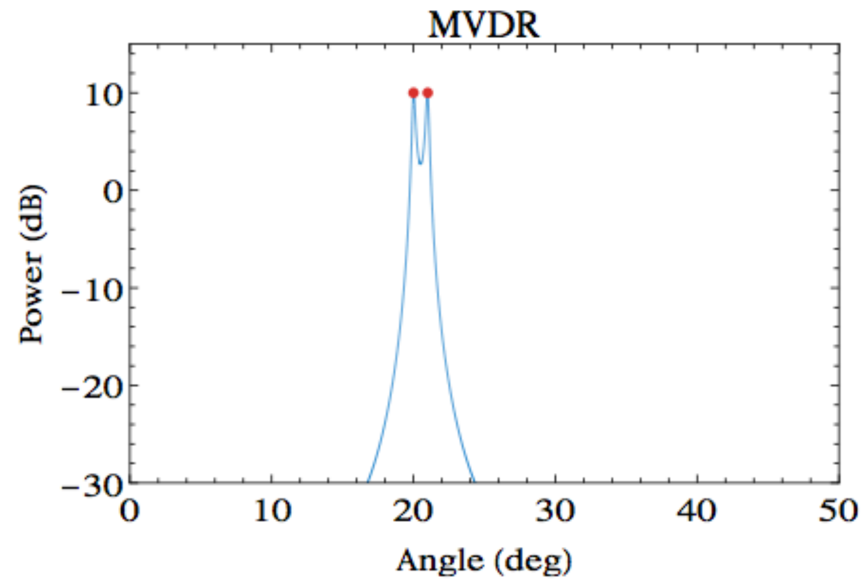


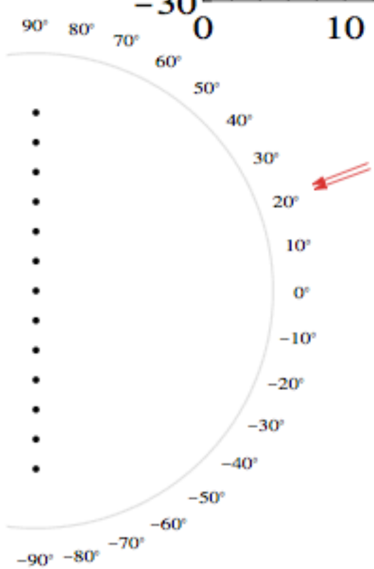


Columns of **A** span −90° to 90° in steps of 1° ($M = 181$).

Resolving closely spaced signals





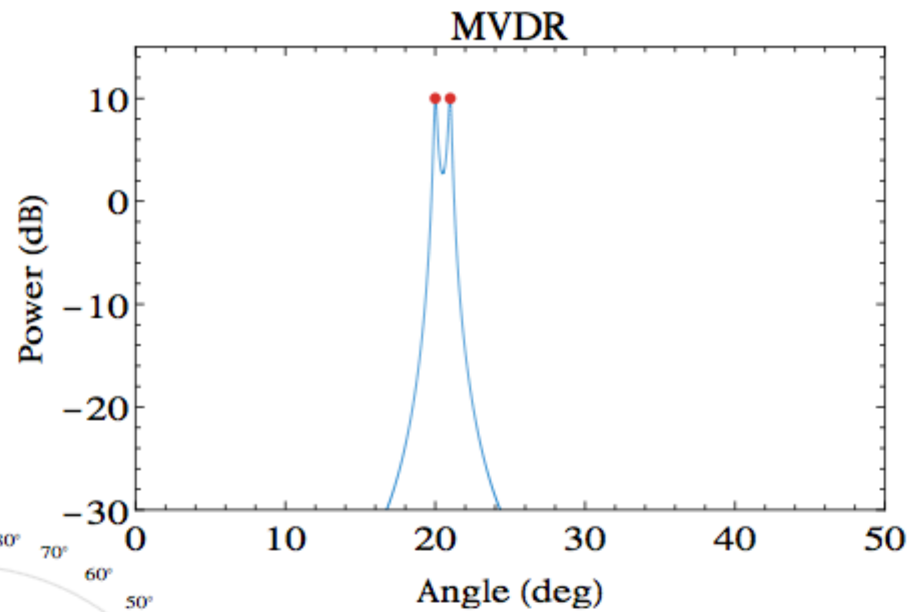


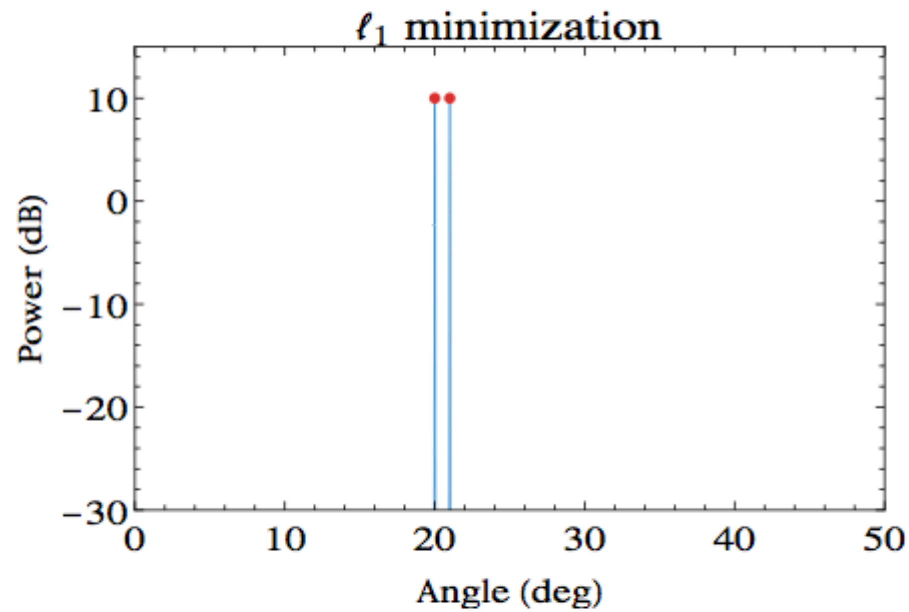
$\theta_1 = 20°$, $\theta_2 = 21°$ (note the change in x-axis)

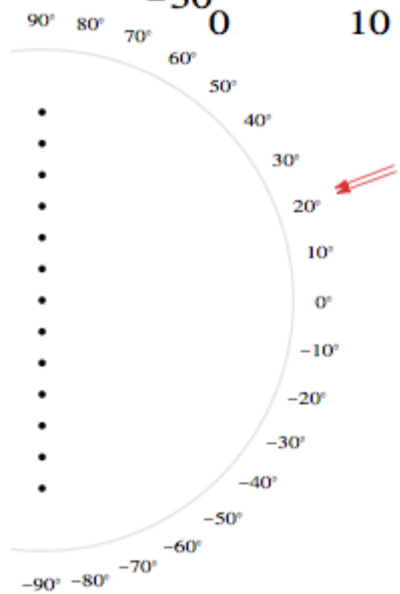
Power overestimated by 6 dB

Resolution of $\sim 5.7°$

Resolving closely spaced signals



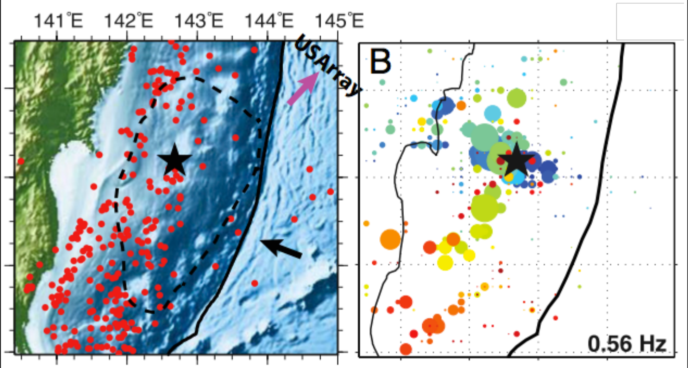




$\theta_1 = 20°$, $\theta_2 = 21°$ (note the change in x-axis)

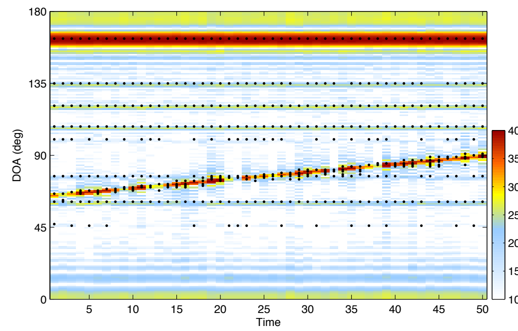
CS approach to geophysical data analysis

CS of Earthquakes



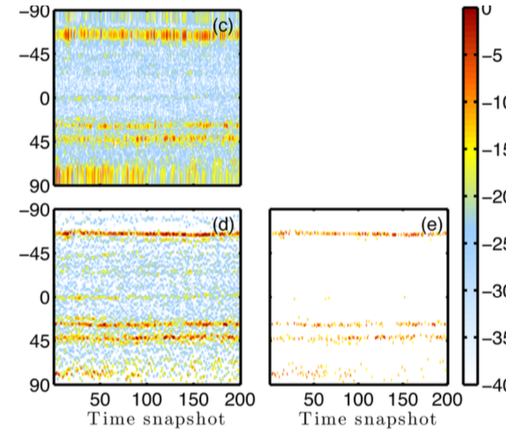
Yao, GRL 2011, PNAS 2013

Sequential CS



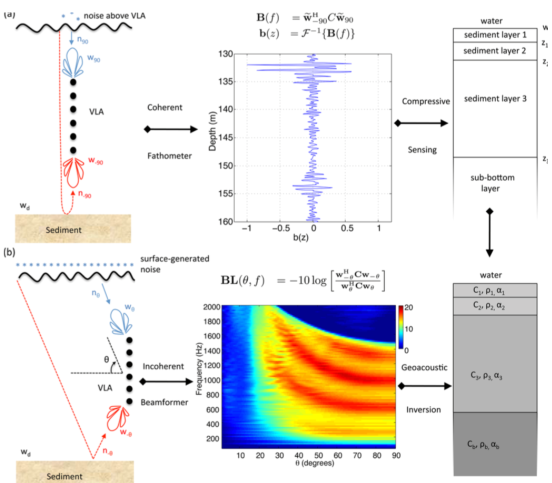
Mecklenbrauker, TSP 2013

CS beamforming



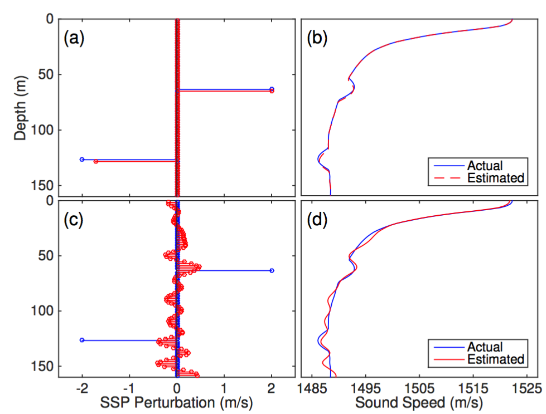
Xenaki, JASA 2014, 2015
Gerstoft JASA 2015

CS fathometer



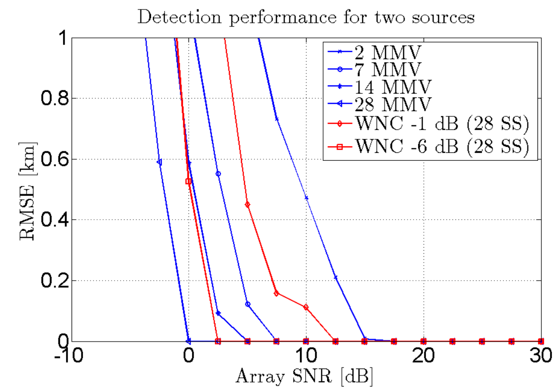
Yardim, JASA 2014

CS Sound speed estimation



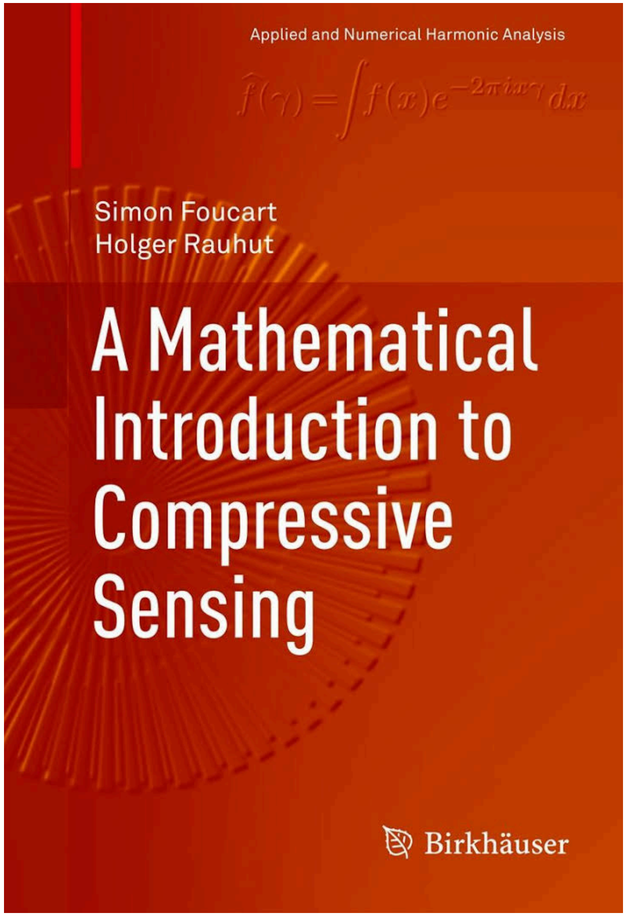
Bianco, JASA 2016

CS matched field

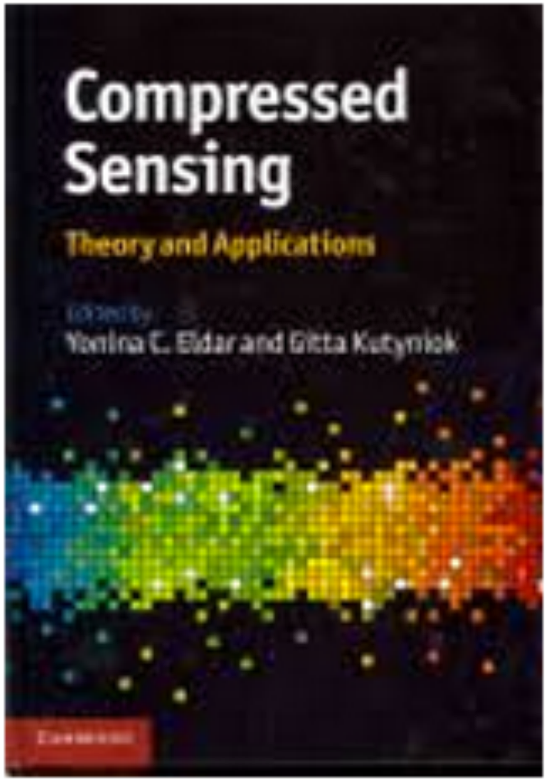


Gemba, JASA 2016

Resources


Applied and Numerical Harmonic Analysis
$\hat{f}(\gamma)=\int f(x)e^{-2\pi i x\gamma}dx$
Simon Foucart
Holger Rauhut
A Mathematical Introduction to Compressive Sensing
Birkhäuser


Michael Elad
Sparse and Redundant Representations
From Theory to Applications in Signal and Image Processing
Springer


Compressed Sensing
Theory and Applications
Yonina C. Eldar and Gitta Kutyniok

Bayesian interpretation of LASSO

MAP estimate via the unconstrained -LASSO- formulation

$$\widehat{\mathbf{x}}_{\text{LASSO}}(\mu) = \arg\min_{\mathbf{x}\in\mathbb{C}^N} \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2 + \mu\|\mathbf{x}\|_1$$

Bayes rule:

$$p(\mathbf{x}|\mathbf{y}) = \frac{p(\mathbf{y}|\mathbf{x})p(\mathbf{x})}{p(\mathbf{y})}$$

MAP estimate:

$$\begin{aligned}
\widehat{\mathbf{x}}_{\text{MAP}} &= \arg\max_{\mathbf{x}} \ \ln p(\mathbf{x}|\mathbf{y}) \\
&= \arg\max_{\mathbf{x}} \ [\ln p(\mathbf{y}|\mathbf{x}) + \ln p(\mathbf{x})] \\
&= \arg\min_{\mathbf{x}} \ [-\ln p(\mathbf{y}|\mathbf{x}) - \ln p(\mathbf{x})]
\end{aligned}$$

MAP estimate via the unconstrained -LASSO- formulation

Bayes rule:

$$p(\mathbf{x}|\mathbf{y}) = \frac{p(\mathbf{y}|\mathbf{x})p(\mathbf{x})}{p(\mathbf{y})}$$

MAP estimate:

$$\widehat{\mathbf{x}}_{\text{MAP}} = \arg\min_{\mathbf{x}} \; [-\ln p(\mathbf{y}|\mathbf{x}) - \ln p(\mathbf{x})]$$

Gaussian likelihood:

$$p(\mathbf{y}|\mathbf{x}) \propto \mathrm{e}^{-\frac{\|\mathbf{y}-\mathbf{A}\mathbf{x}\|_2^2}{\sigma^2}}$$

Laplace-like prior:

$$p(\mathbf{x}) \propto \prod_{i=1}^{N} \mathrm{e}^{-\frac{\sqrt{(\Re x_i)^2+(\Im x_i)^2}}{\nu}} = \mathrm{e}^{-\frac{\|\mathbf{x}\|_1}{\nu}}$$

MAP estimate (LASSO):

$$\widehat{\mathbf{x}}_{\text{MAP}} = \arg\min_{\mathbf{x}} \left[\|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2 + \mu\|\mathbf{x}\|_1\right] = \widehat{\mathbf{x}}_{\text{LASSO}}(\mu), \; \mu = \frac{\sigma^2}{\nu}$$

MAP– LASSO path

Likelihood (noise complex Gaussian) $$p(y \mid x) \propto \exp\left(-\frac{\|Ax-y\|_2^2}{\sigma^2}\right)$$

Prior (Laplacian) $$p(x) \propto \exp\left(-\frac{\|x\|_1}{\nu}\right)$$

Bayes rule $$p(\mathrm{x|y}) \propto \mathrm{p(y|x)p(x)} \propto \exp\left(-\frac{\|Ax-y\|_2^2}{\sigma^2} - \frac{\|x\|_1}{\nu}\right)$$

Maximum A Posteriori (MAP)

$$\hat{\mathbf{x}}_{\mathrm{MAP}} = \arg\min_{\mathbf{x}} \left[\|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2 + \mu\|\mathbf{x}\|_1\right] = \hat{\mathbf{x}}_{\mathrm{LASSO}}(\mu),$$

LASSO=Least Absolute Shrinkage and Selection Operator

$$\mu = \frac{\sigma^2}{\nu}$$

μ large: $\mathbf{x} = \mathbf{0}$

μ small: $\mathbf{x}$ minumum norm