*-Lasso Therapy: a sparse synthesis approach.

Matthieu KOWALSKI

Univ Paris-Sud
L2S (GPI)

Introduction: sparse approximation

" *It is futile to do with more things that which can be done with fewer*"

William of Ockham

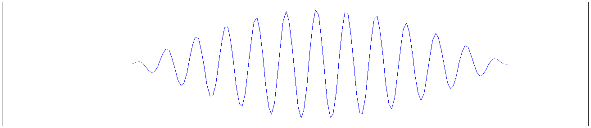
But

Analyse, explain, represent. . . signals.

Exemples

Automatic transciption, source separation, coding. . .

Problem: How to represent a signal and select relevant "information" ?
Sparsity principle: explain a signal with few elements.



Examples of representation of an audio signal

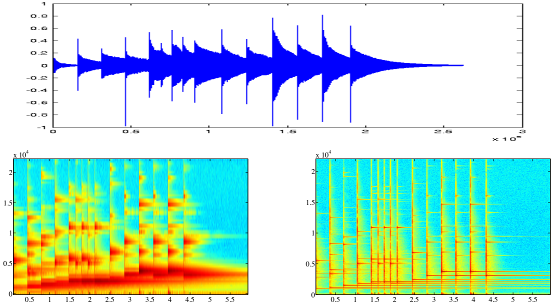


FIGURE : *Time-frequency images. Top: signal, bottom-left: representation adapted to transceents. Bottom-right, representation adapted to tonals.*

The characteristics of interest are rarely directly observable.

Notations and definitions

Some notations

- Let $s \in \mathbb{C}^M$ a signal.
- Let $\Phi \in \mathbb{C}^{M \times N}$, $M \leq N$ the matrix of a dictionnary $\{\varphi_k\}$ (*ie* an over-complete set), constructed as a set of time-frequency atoms.
- Let $y = s + b$ a noisy measure of a signal s.

Definition: synthesis coefficients

Let $\alpha \in \mathbb{C}^N$ such that $s = \Phi\alpha = \sum_k \alpha_k \varphi_k$.
α_k are called *synthesis coefficients*.

if $N > M$, there exists an infinity of such a representation

Definition: analysis coefficients

We call *analysis coefficients*: $\{\langle y, \varphi_k \rangle\} = \Phi^T y$

Sparsity: synthesis approach

Goal: find a "god repsentation" $\hat{s}$ of s such that $\hat{s} = \Phi\hat{\alpha}$

Hypothesis: s admits a sparse representation in the choosen dictionnary.
Ideal solution:

$$\hat{\alpha} = \underset{\alpha}{\operatorname{argmin}} \|\alpha\|_0 \quad \text{sc} \quad s = \Phi\alpha$$

Noisy observation:

$$\hat{\alpha} = \underset{\alpha}{\operatorname{argmin}} \|y - \Phi\alpha\|_2^2 + \lambda\|\alpha\|_0$$

Probleme very hard to solve in a finite time $\Rightarrow$ we relax theℓ_0 constraint into ℓ_1

LASSO [Tibshirani 96] or **Basis Pursuit Denoising** [Chen et al. 98]:

$$\hat{\alpha} = \underset{\alpha}{\operatorname{argmin}} \|y - \Phi\alpha\|_2^2 + \lambda\|\alpha\|_1$$

Frameworks

Mathematical framework

- $\mathbf{y} \in \mathbb{R}^M$
- $\mathbf{x} \in \mathbb{R}^N$
- $A \in \mathbb{R}^{M.N}$

Optimization framework

$$\mathbf{x} = \operatorname{argmin} \mathcal{L}(\mathbf{y}, A, \mathbf{x}) + P(\mathbf{x}; \lambda)$$

1. A convex loss or data term $\mathcal{L}(\mathbf{y}, A, \mathbf{x})$ measuring the fit between the observed mixture $\mathbf{y}$ and the source signal $\mathbf{x}$ given the mixing system A;
2. A regularization term P modeling the assumptions about the sources,
3. An hyperparameter $\lambda \in \mathbb{R}_+$ governing the balance between the data term and the regularization term.

The Loss

Traditional assumption: Gaussian noise

$$\mathcal{L}(\mathbf{y}, A, \mathbf{x}) = \frac{1}{2}\|\mathbf{y} - A\mathbf{x}\|_2^2$$

But other possible choices

- Impulsive noise:

$$\mathcal{L}(\mathbf{y}, A, \mathbf{x}) = \frac{1}{2}\|\mathbf{y} - A\mathbf{x}\|_1$$

- Poisson noise:

$$\mathcal{L}(\mathbf{y}, A, \mathbf{x}) = A\mathbf{x} - \mathbf{y} + \mathbf{y}\ln\left(\frac{\mathbf{y}}{A\mathbf{x}}\right)$$

The Penalty

Goal: Model the prior on the sources.

"Analysis" prior

Models the "physical" assumptions on the sources

- Minimum energy : $\frac{1}{2}\|\mathbf{x}\|_2^2$ [Tikhonov, 77]
- Total variation (images) : $\|\nabla\mathbf{x}\|_1$ [ROF, 92]

Sometimes, we need more flexibility: priors are not always in the "samples" domain

Optimization framework with dictionary

1. A Dictionary $\mathbf{\Phi}$
2. A convex loss or data term $\mathcal{L}(\mathbf{y}, A, \boldsymbol{\alpha})$ measuring the fit between the observed mixture $\mathbf{y}$ and some synthesis coefficients $\boldsymbol{\alpha}$, such that $\mathbf{x} = \mathbf{\Phi}\boldsymbol{\alpha}$, given the mixing system A;
3. A regularization term P modeling the assumptions about the sources, in the synthesis coefficient domain
4. An hyperparameter $\lambda \in \mathbb{R}_+$ governing the balance between the data term and the regularization term.

The Dictionary

Synthesis point of view

Assume $\mathbf{x}$ can be written as

$$\mathbf{x} = \sum_{k=1}^{K} \alpha_k \boldsymbol{\varphi}_k$$
$$= \boldsymbol{\Phi}\boldsymbol{\alpha}$$

with

$$\boldsymbol{\Phi} \in \mathbb{C}^{N.K}, \quad k \geq N$$

Examples

- Gabor
- wavelets
- Union of Gabor (hybrid model or Morphological Component Analysis): $\mathbf{x} = \mathbf{x}_1 + \mathbf{x}_2 = \boldsymbol{\Phi}_1\boldsymbol{\alpha}_1 + \boldsymbol{\Phi}_2\boldsymbol{\alpha}_2$
- Frames ([Balazs *et al.*, 2013])

The penalty (returns)

Sparse approximation: key idea
$\mathbf{x} \in \mathbb{R}^N$ admits a sparse decomposition inside a dictionnary of waveforms $\{\varphi_k\}_{k=1}^K$:

$$\mathbf{x} = \sum_{k \in \Lambda} \alpha_k \varphi_k$$

with $\Lambda \subset \{1, \dots, K\}$

Given a (noisy) observation $\mathbf{y} = A\mathbf{x} + \mathbf{n}$, the Lasso/Basis Pursuit Denoising [Tibshirani, 96], [Chen *et al.* 98] estimate reads:

$$\hat{\boldsymbol{\alpha}} = \underset{\boldsymbol{\alpha}}{\operatorname{argmin}} \frac{1}{2} \|\mathbf{y} - A\boldsymbol{\Phi}\boldsymbol{\alpha}\|^2 + \lambda \|\boldsymbol{\alpha}\|_1$$

and

$$\hat{\mathbf{x}} = \boldsymbol{\Phi}\hat{\boldsymbol{\alpha}}$$

Mixed norms: definition

Definition [Benedek *et al.* 61]

Let $\{\alpha_{g,m}\}$ a double indexed sentence. We call mixed norm $\ell_{p,q}$ of α the norm

$$\|\boldsymbol{\alpha}\|_{p,q} = \left(\sum_g \left(\sum_m |\alpha_{g,m}|^p \right)^{q/p} \right)^{1/q}$$

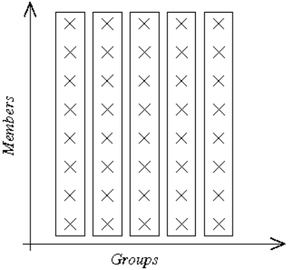


FIGURE : *A grouping organisation doubly indexed.*

Mixed norms: remarks

General remarks

- $\ell_{p,q}$ is a true norm for $p, q \geq 1$.
- Cases $p = +\infty$ ou $q = \infty$ are obtained by replacing the corresponding norm by the supremum.
- We can define corresponding quasi-normes for $p, q < 1$.
- We generalize it on several levels [MK & AG 10].

Some particular case in regression

- $p = q = 2$ **Ridge regression:** no sparsity, no structure
- $p = q = 1$ **LASSO** (or BPDN) regression: sparsity whithout structure
- $p = 1$ and $q = 2$ **Group-LASSO** [Yuan *et al.* 06] (or *joint sparsity* [Fornasier *et al.* 08], or *Multiple measurement vector* [Cotter *et al* 05]) regression: sparisty between groups.
- $p = 2$ and $q = 1$ **Elitist-LASSO** [MK 09, MK & BT 09] regression: sparsity *inside* the groups.

Regression and mixed norms

We are interrested by the following optimization problem

$$\hat{\boldsymbol{\alpha}} = \underset{\boldsymbol{\alpha}}{\operatorname{argmin}} \|\mathbf{y} - \boldsymbol{\Phi}\boldsymbol{\alpha}\|_2^2 + \lambda \|\boldsymbol{\alpha}\|_{p,q}^q$$

Remark

This problem is convex for $p, q \geq 1$ and strictly convex for $p, q > 1$.

Decoupling on the groups, not on coefficients

Proximity operators

we suppose that Φ is *orthogonal*. We denote by $\tilde{y} = \Phi^T y$

LASSO solution $\min_\alpha \|y - \Phi\alpha\|_2^2 + \lambda\|\alpha\|_1$

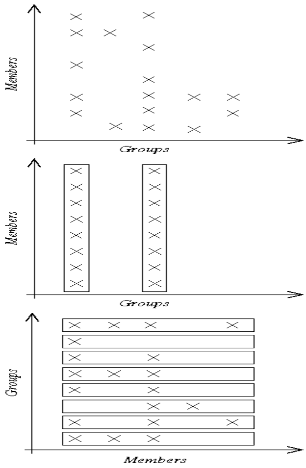
$$\hat{\alpha}_{g,m} = \arg(\tilde{y}_{g,m}) \left(|\tilde{y}_{g,m}| - \lambda\right)^+$$

G-LASSO solution $\min_\alpha \|y - \Phi\alpha\|_2^2 + \lambda\|\alpha\|_{2,1}$

$$\hat{\alpha}_{g,m} = \tilde{y}_{g,m} \left(1 - \frac{\lambda}{\|\tilde{y}_g\|_2}\right)^+$$

E-LASSO solution $\min_\alpha \|y - \Phi\alpha\|_2^2 + \lambda\|\alpha\|_{1,2}^2$

$$\hat{\alpha}_{g,m} = \arg(\tilde{y}_{g,m}) \left(|\tilde{y}_{g,m}| - \frac{\lambda}{1 + \lambda L_g}\|\|\tilde{y}_g\|\|\right)^+$$



(Relaxed) ISTA

- Let $\boldsymbol{\alpha}^{(0)} = \mathbf{0}$, $L \geq \frac{1}{\|\boldsymbol{\Phi}^*\boldsymbol{\Phi}\|}$, $0 \leq \mu < 1$, and $t_{max} \in \mathbb{N}$.
- **For** $t = 0$ to t_{max}

$$\begin{aligned} \boldsymbol{\alpha}^{(t+1/2)} &= \boldsymbol{\gamma}^{(t)} + \boldsymbol{\Phi}^*(\mathbf{y} - \boldsymbol{\Phi}\boldsymbol{\gamma}^{(t)})/L \\ \boldsymbol{\alpha}^{(t+1)} &= \mathbb{S}(\boldsymbol{\alpha}^{(t+1/2)}, \lambda/L) \\ \boldsymbol{\gamma}^{t+1} &= \boldsymbol{\alpha}^{(t+1)} + \mu^{(t+1)}(\boldsymbol{\alpha}^{(t+1)} - \boldsymbol{\alpha}^{(t)}) \end{aligned}$$

End For

with $\mathbb{S}$ a proximity operator (soft thresholding for ℓ_1).

Convergence proved by several authors

- [Combettes & Wajs 05] forward-backward (proximity operators);
- [Daubechies & al 04] Opial's fixed point theorem;
- [Figuereido & Nowak 03] EM algorithm;

Accelerated version by [Nesterov 07], [Beck & Teboulle 09] (FISTA).

Limitations

- Biased coefficients: large coefficients are shrinked [Gao, Bruce 97]
- Lake of flexibility for structures: needs to define an adequate convex penalty (not always simple)

Could we play directly on the thresholding step ?

Thresholding rules

Definition [Antoniadis 07]

1. $\mathbb{S}(.;\lambda)$ is an odd function. ($\mathbb{S}_+(.;\lambda)$ is used to denote the $\mathbb{S}(.;\lambda)$ restricted to $\mathbb{R}_+$.)
2. $\mathbb{S}(.;\lambda)$ is a shrinkage rule: $0 \leq \mathbb{S}_+(t;\lambda) \leq t,\ \forall t \in \mathbb{R}_+$.
3. $\mathbb{S}_+$ is nondecreasing on $\mathbb{R}_+$, and $\lim_{t\to+\infty} \mathbb{S}(t;\lambda) = +\infty$

Examples

- Soft [Donoho, Johnstone 94]

$$\mathbb{S}(x;\lambda) = x\left(1 - \frac{\lambda}{|x|}\right)^+$$

- Hard Soft [Donoho, Johnstone 94]

$$\mathbb{S}(x;\lambda) = x\mathbf{1}_{|x|>\lambda}$$

- NonNegativeGarrote (NNGarrote) [Gao 98]

$$\mathbb{S}(x;\lambda) = x\left(1 - \frac{\lambda}{|x|^2}\right)^+$$

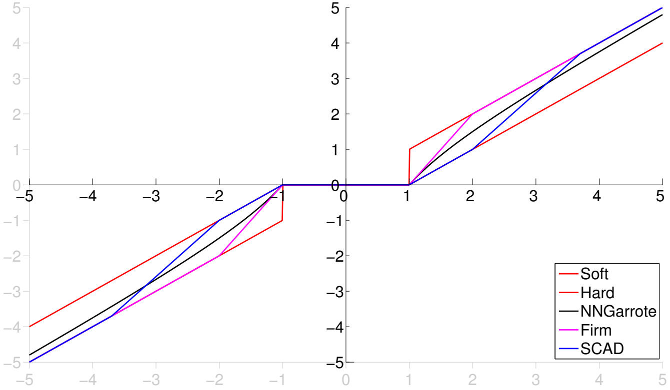
- Firm [Gao, Bruce 97]

$$\mathbb{S}(x;\lambda_1;\lambda 2) = \begin{cases} 0 & \text{if } |x| < \lambda_1 \\ \frac{x\lambda_2\left(1-\frac{\lambda_1}{|x|}\right)}{\lambda_2-\lambda 1} & \text{if } \lambda_1 \leq |x| < \lambda_2 \\ x & |x| > \lambda_2 \end{cases}$$

- SCAD [Antoniadis, Fan 01]

$$\mathbb{S}(x;\lambda;a) = \begin{cases} x(1 - \frac{\lambda}{|x|})^+ & \text{if } |x| < 2\lambda \\ \frac{x\left(a-1-\frac{a\lambda}{|x|}\right)}{a-2} & \text{if } 2\lambda \leq |x| < a\lambda \\ x & \text{if } |x| > a\lambda \end{cases}$$

Examples



Properties of Thresholding rules

Definition: semi-convex fonction

A function f is said to be semi-convex, iff there exists c such that

$$x \mapsto f(x) + \frac{c}{2}\|x\|^2$$

is convex

Proposition

We can associate a semi-convex penalty $P(.;\lambda)$, with $c \leq 1$ to any thresholding rules. Moreover, $\frac{1}{1-c}$ is an upper-bound of $\mathbb{S}'(.;\lambda)$.

Convergence results

Theorem

- ISTA converges with any thresholding rules
- Relaxed ista converges for $0 \leq \mu < 1 - c$, where c is an upper-bound of $\mathbb{S}'(.;\lambda)$

Examples

- NNGarrote ($c = 1/2$)

$$P(x;\lambda) = \lambda^2 + \operatorname{asinh}\left(\frac{|x|}{2\lambda}\right) + \lambda^2 \frac{|x|}{\sqrt{x^2 + 4\lambda^2} + |x|}$$

- SCAD ($c = a - 1$)

$$P(x;\lambda) = \begin{cases} \lambda x & \text{if } x \leq \lambda \\ \frac{(a\lambda x - x^2/2)}{a-1} & \text{if } \lambda < x \leq a\lambda \\ a\lambda & \text{if } x > a\lambda \end{cases}$$

1. Introduction: sparse approximation
2. An optimization framework
   - General approach
   - Mixed Norms
3. Iterative Thresholding
   - Thresholding functions
   - Neighborhood thresholding
4. Numerical illustration
   - Application to tonal/transicent separation

Windowed Group-LASSO

Back to the model $\mathbf{y} = \mathbf{\Phi}\boldsymbol{\alpha} + \mathbf{b}$, with $\mathbf{\Phi}$ orthonormal. Back to a simple indexing, and for each index k, we define a neighborhood $g(k)$.

Windowed G-Lasso [MK & BT 09], [K *et al.* 13]

$$\hat{\alpha}_k = \tilde{y}_k \left(1 - \frac{\lambda}{\sqrt{\sum_{m \in g(k)} |\tilde{y}_m|^2}} \right)^+$$

$$= \tilde{y}_k \left(1 - \frac{\lambda}{\|\tilde{y}_{g(k)}\|_2} \right)^+$$

with $\tilde{y} = \Phi^* y$

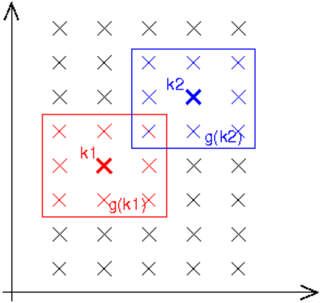


FIGURE : *WG-LASSO. Two overlapping groups: neighborhood of k_1 and k_2.*

Similar thresholding rules introduced by [Cai & Silvermanss 01] for wavelet thresholding.

A family of shrinkage operators

$\boldsymbol{\alpha} = \mathbb{S}(\mathbf{y})$ is given coordinatewise:

- Lasso:

$$\alpha_k = y_k \left(1 - \frac{\lambda}{|y_k|}\right)^+$$

- NNGarrote / Empirical Wiener

$$\alpha_k = y_k \left(1 - \frac{\lambda}{|y_k|^2}\right)^+$$

- Windowed Group Lasso

$$\alpha_k = \tilde{y}_k \left(1 - \frac{\lambda}{\|\tilde{y}_{g(k)}\|_2}\right)^+$$

- Empirical Persistent Wiener [Siedenburg 13]

$$\alpha_k = \tilde{y}_k \left(1 - \frac{\lambda}{\|\tilde{y}_{g(k)}\|_2^2}\right)^+$$

1. Introduction: sparse approximation

2. An optimization framework
   - General approach
   - Mixed Norms

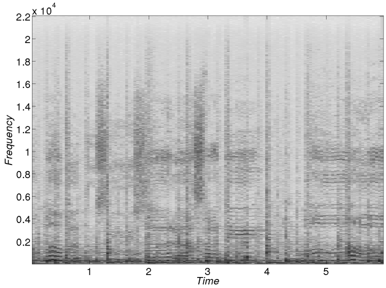
3. Iterative Thresholding
   - Thresholding functions
   - Neighborhood thresholding

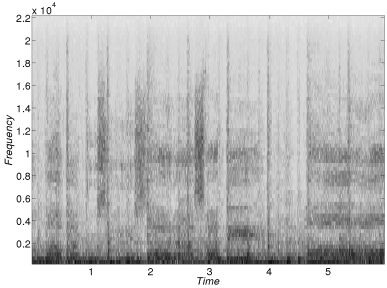
4. Numerical illustration
   - Application to tonal/transicent separation

Tonal/transcient separation - 1

Excerpt of *Mamavatu* from Susheela Raman. Length of windows analysis for MDCT:

- For tonal layer: 4096 samples (93 ms) (Left)
- For transcient layer: 128 samples (3 ms) (Right)





Tonal/transcient separation - 2

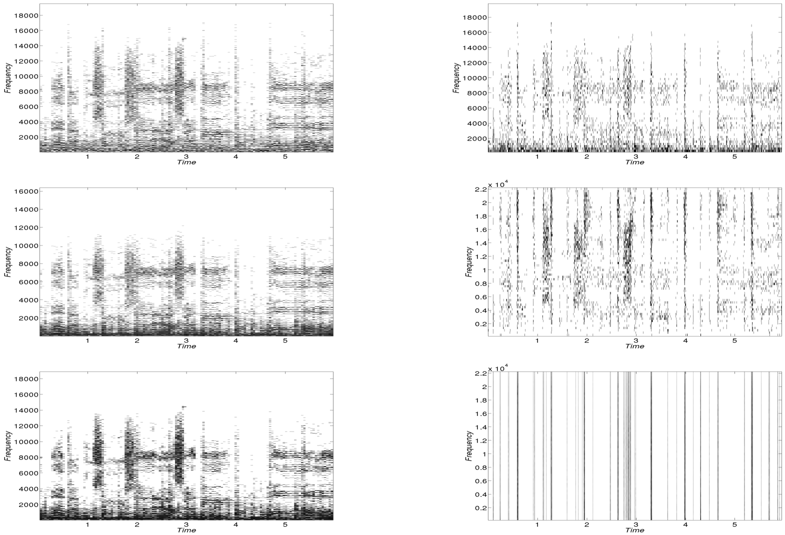


FIGURE : *Left: tonal layers. Right: transcient layers. From top to bottom: LASSO/LASSO, LASSO/ELASSO, LASSO/GLASSO.*