

*-Lasso Therapy: a sparse synthesis approach.

Matthieu KOWALSKI

Univiversity Paris-Saclay
L2S (GPI)

- 1 Introduction: sparse approximation
- 2 An optimization framework
 - General approach
 - Mixed Norms
 - Hybrid (or multilayers) models
- 3 Iterative Thresholding
 - Thresholding functions
 - Neighborhood thresholding
- 4 Numerical results
 - Application to tonal/transient separation
 - Simulations
 - Audio declipping
- 5 Conclusion

Introduction: sparse approximation

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

William of Ockham

But

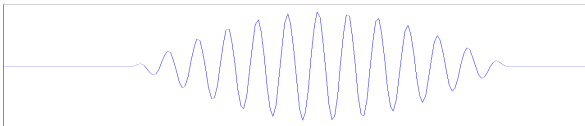
Analyse, explain, represent. . . signals.

Exemples

Automatic transcription, source separation, coding. . .

Problem: How to represent a signal and select relevant "information" ?

Sparsity principle: explain a signal with few elements.



- 1 Introduction: sparse approximation
- 2 An optimization framework
 - General approach
 - Mixed Norms
 - Hybrid (or multilayers) models
- 3 Iterative Thresholding
 - Thresholding functions
 - Neighborhood thresholding
- 4 Numerical results
 - Application to tonal/transient separation
 - Simulations
 - Audio declipping
- 5 Conclusion

Examples of representation of an audio signal

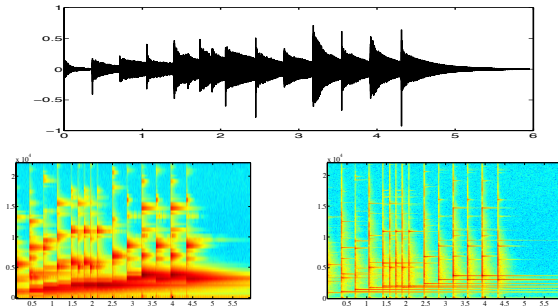


FIGURE : *Time-frequency images. Top: signal, bottom-left: representation adapted to transients. Bottom-right, representation adapted to tonals.*

The characteristics of interest are rarely directly observable.

Information concentration

pdf of samples of a signal and its MDCT coefficients.

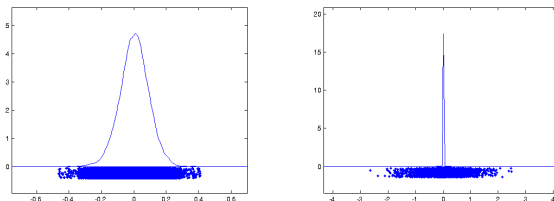


FIGURE : *Left: pdf of samples. Right: pdf MDCT coefficients.*

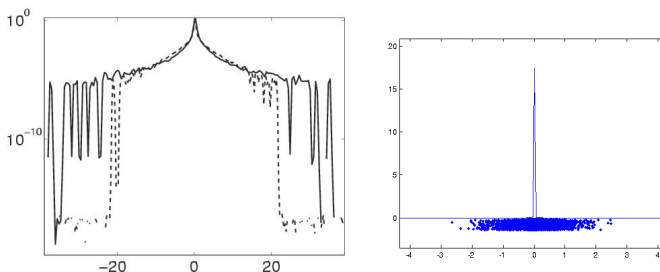
A “good representation” allows to concentrate the information in a few coefficients

Some transforms

- Fourier ;
- Gabor or STFT ;
- MDCT ;
- wavelets ;
- *-lets ;

Observed pdf

We focused on this particularity : the sparsity of the extension for a given component depend on the bases.



pdf of various representations of two sample signals : castagnet (solid line) and organ (dotted line) : wavelet (left) and MDCT (right) coefficients.

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 dictionary $\{\varphi_k\}$ (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 representation” $\hat{s}$ of s such that $\hat{s} = \Phi \hat{\alpha}$

Hypothesis: s admits a sparse representation in the chosen dictionary.

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 the ℓ_0 constraint into ℓ_1

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

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

Examples of representation of an audio signal

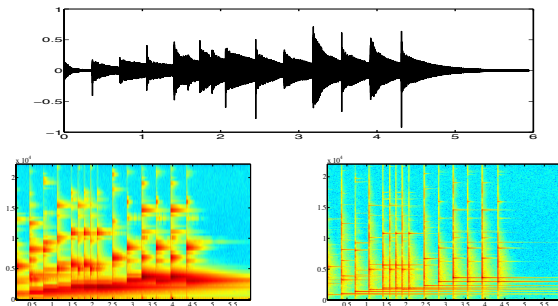


FIGURE : *Time-frequency images. Top: signal, bottom-left: representation adapted to transients. Bottom-right, representation adapted to tonals.*

Links between analysis/synthesis and maximum likelihood

Bayesian point of view

$$\begin{aligned} & \min_{\alpha} \{ \|y - \Phi\alpha\|_2^2 + \lambda \|\alpha\|_1 \} \\ \Leftrightarrow & \max_{\alpha} \left\{ \underbrace{e^{-\|y - \Phi\alpha\|_2^2}}_{\text{"bruit gaussien"}} \underbrace{\prod_k e^{-\lambda |\alpha_k|}}_{\text{a priori Laplacien}} \right\} \end{aligned}$$

Hypothesis: Φ corresponds to an orthonormal basis $\Phi^T = \Phi^{-1}$.

Solution

Let $\tilde{y} = \Phi^T y$. The solution is given by soft thresholding *coefficient by coefficient* :

$$\hat{\alpha}_k = \arg(\tilde{y}_k) (|\tilde{y}_k| - \lambda)^+$$

and $\hat{s} = \Phi \hat{\alpha}$

- 1 Introduction: sparse approximation
- 2 An optimization framework
 - General approach
 - Mixed Norms
 - Hybrid (or multilayers) models
- 3 Iterative Thresholding
 - Thresholding functions
 - Neighborhood thresholding
- 4 Numerical results
 - Application to tonal/transient separation
 - Simulations
 - Audio declipping
- 5 Conclusion

Frameworks

Mathematical framework

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

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 Φ
- ② A convex loss or data term $\mathcal{L}(\mathbf{y}, A, \alpha)$ measuring the fit between the observed mixture $\mathbf{y}$ and some synthesis coefficients α , such that $\mathbf{x} = \Phi\alpha$, 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

$$\begin{aligned}\mathbf{x} &= \sum_{k=1}^K \alpha_k \varphi_k \\ &= \Phi \alpha\end{aligned}$$

with

$$\Phi \in \mathbb{C}^{N \times K}, \quad K \geq N$$

Examples

- Gabor
- wavelets
- Union of Gabor (hybrid model or Morphological Component Analysis): $\mathbf{x} = \mathbf{x}_1 + \mathbf{x}_2 = \Phi_1 \alpha_1 + \Phi_2 \alpha_2$
- Frames ([Balazs *et al.*, 2013])

The penalty (returns)

Sparse approximation: key idea

$\mathbf{x} \in \mathbb{R}^N$ admits a sparse decomposition inside a dictionary 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

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

- 1 Introduction: sparse approximation
- 2 An optimization framework
 - General approach
 - **Mixed Norms**
 - Hybrid (or multilayers) models
- 3 Iterative Thresholding
 - Thresholding functions
 - Neighborhood thresholding
- 4 Numerical results
 - Application to tonal/transient separation
 - Simulations
 - Audio declipping
- 5 Conclusion

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

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

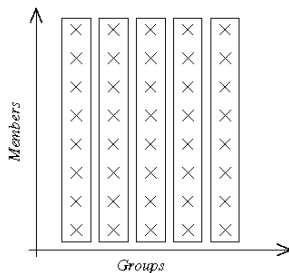


FIGURE : A grouping organisation doubly indexed.

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 without 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: sparsity 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 interested by the following optimization problem

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

Remark

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

Bayesian point of view

$$\begin{aligned} \hat{\alpha} &= \underset{\alpha}{\operatorname{argmax}} e^{-\|\mathbf{y} - \Phi\alpha\|_2^2} e^{-\lambda \|\alpha\|_{p,q}^q} \\ &= \underset{\alpha}{\operatorname{argmax}} e^{-\|\mathbf{y} - \Phi\alpha\|_2^2} \prod_g \exp\{-\lambda \|\alpha_g\|_p^q\} \end{aligned}$$

$\Rightarrow$ Independance is between the groups of coefficients

$\alpha_g = (\alpha_{g,1} \dots \alpha_{g,m} \dots)$ **Decoupling on the groups, not on coefficients**

- 1 Introduction: sparse approximation
- 2 An optimization framework
 - General approach
 - Mixed Norms
 - Hybrid (or multilayers) models
- 3 Iterative Thresholding
 - Thresholding functions
 - Neighborhood thresholding
- 4 Numerical results
 - Application to tonal/transient separation
 - Simulations
 - Audio declipping
- 5 Conclusion

Hybrid/morphological decompositions

We guess that the signal s can be written as

$$\mathbf{s} = \mathbf{s}_1 + \mathbf{s}_2$$

And each layer admit a sparse expansion in an adapted dictionary (Hybrid model [Daudet *et al.* 02] or morphological model [Starck *et al.* 05]).

A new dictionary is build as the union of the two dictionaries

$$\mathcal{D} = \mathcal{D}_1 \cup \mathcal{D}_2$$

Then, the synthesis model becomes:

$$\mathbf{s} = \Phi_1 \alpha_1 + \Phi_2 \alpha_2$$

There still exists an infinity of such an expansion.

Problem: finding α and β .

Variational formulation

We seek to estimate s by optimizing the following functional

$$\hat{\alpha}_1, \hat{\alpha}_2 = \operatorname{argmin}_{\alpha_1, \alpha_2} \left\| \mathbf{y} - \sum_k \alpha_{1_k} \varphi_{1_k} - \sum_\ell \alpha_{2_\ell} \varphi_{2_\ell} \right\|_2^2 + \lambda_1 \|\alpha\|_{p_1, q_1}^{q_1} + \lambda_2 \|\alpha\|_{p_2, q_2}^{q_2}$$

$$\hat{\alpha}_1, \hat{\alpha}_2 = \operatorname{argmin}_{\alpha_1, \alpha_2} \left\| \mathbf{y} - \Phi_1 \alpha_1 - \Phi_2 \alpha_2 \right\|_2^2 + \lambda_1 \|\alpha_1\|_{p_1, q_1}^{q_1} + \lambda_2 \|\alpha_2\|_{p_2, q_2}^{q_2}$$

We obtain the decomposition into two layers:

$$\hat{\mathbf{s}}_1 = \Phi_1 \hat{\alpha}_1 \quad \hat{\mathbf{s}}_2 = \Phi_2 \hat{\alpha}_2$$

The penalty (summary)

Structured penalties

- Structured sparsity via mixed norm [K,Torrésani 2008], [K, 2009]:
 - Group-Lasso [Yuan, Lin 2006]

$$P(\alpha; \lambda) = \lambda \|\alpha\|_{2;1} = \lambda \sum_g \sqrt{\sum_m |\alpha_{g,m}|^2}$$
 - Elitist-Lasso [K,Torrésani 2008]

$$P(\alpha; \lambda) = \lambda \|\alpha\|_{1;2}^2 = \lambda \sum_g (\sum_m |\alpha_{g,m}|)^2$$
- Hi-Lasso [Jenatton *et al.* 2011], [Sprechmann *et al.* 2011]

$$P(\alpha; \lambda) = \lambda ((1 - \nu) \|\alpha\|_{2;1} + \nu \|\alpha\|_1)$$
- sub-modular functions etc. [Bach 2012]

$$\hat{\alpha}_1, \hat{\alpha}_2 = \underset{\alpha_1, \alpha_2}{\operatorname{argmin}} \frac{1}{2} \|\mathbf{y} - A(\Phi_1 \alpha_1 + \Phi_2 \alpha_2)\|^2 + P(\alpha_1; \lambda_1) + P(\alpha_2; \lambda_2)$$

and

$$\hat{\mathbf{x}} = \Phi_1 \hat{\alpha}_1 + \Phi_2 \hat{\alpha}_2$$

- 1 Introduction: sparse approximation
- 2 An optimization framework
 - General approach
 - Mixed Norms
 - Hybrid (or multilayers) models
- 3 Iterative Thresholding**
 - Thresholding functions
 - Neighborhood thresholding
- 4 Numerical results
 - Application to tonal/transient separation
 - Simulations
 - Audio declipping
- 5 Conclusion

Proximity operators

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

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

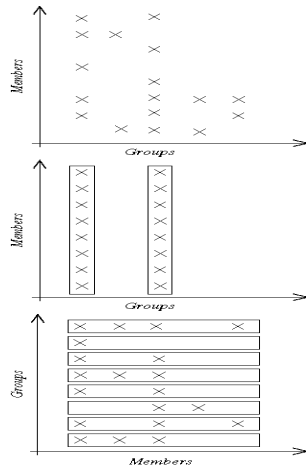
$$\hat{\alpha}_{g,m} = \arg(\tilde{y}_{g,m}) (|\tilde{y}_{g,m}| - \lambda)^+$$

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

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

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 shrunked [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 ?

- 1 Introduction: sparse approximation
- 2 An optimization framework
 - General approach
 - Mixed Norms
 - Hybrid (or multilayers) models
- 3 **Iterative Thresholding**
 - Thresholding functions
 - Neighborhood thresholding
- 4 Numerical results
 - Application to tonal/transient separation
 - Simulations
 - Audio declipping
- 5 Conclusion

Thresholding rules

Definition [Antoniadis 07]

- ① $\mathbb{S}(\cdot; \lambda)$ is an odd function. ($\mathbb{S}_+(\cdot; \lambda)$ is used to denote the $\mathbb{S}(\cdot; \lambda)$ restricted to $\mathbb{R}_+$.)
- ② $\mathbb{S}(\cdot; \lambda)$ 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 $\lim_{t \rightarrow +\infty} \mathbb{S}(t; \lambda) = +\infty$

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]

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

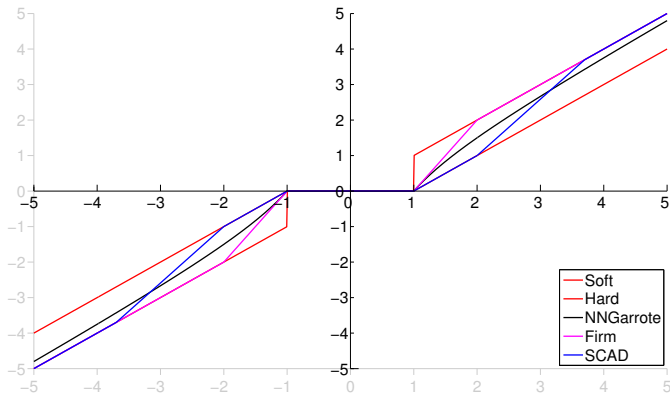
- Firm [Gao, Bruce 97]

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

- SCAD [Antoniadis, Fan 01]

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

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 $P(\cdot; \lambda)$, with $c \leq 1$ to any thresholding rules. Moreover, $\frac{1}{1-c}$ is an upper-bound of $S'(\cdot; \lambda)$.

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 $\mathbb{S}'(\cdot; \lambda)$

Examples

- NNGarrote ($c = 1/2$)

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

- 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
 - Hybrid (or multilayers) models
- 3 Iterative Thresholding
 - Thresholding functions
 - Neighborhood thresholding
- 4 Numerical results
 - Application to tonal/transient separation
 - Simulations
 - Audio declipping
- 5 Conclusion

Windowed Group-LASSO

Back to the model $\mathbf{y} = \Phi\alpha + \mathbf{b}$, with Φ 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]

$$\begin{aligned}\hat{\alpha}_k &= \tilde{y}_k \left(1 - \frac{\lambda}{\sqrt{\sum_{m \in g(k)} |\tilde{y}_m|^2}} \right)^+ \\ &= \tilde{y}_k \left(1 - \frac{\lambda}{\|\tilde{y}_{g(k)}\|_2} \right)^+\end{aligned}$$

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

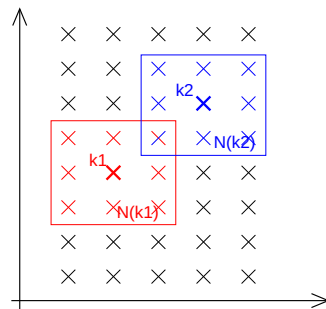


FIGURE : WG-LASSO. Two overlapping groups: neighborhood of k_1 and k_2 .

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

Neighborhood with latents variables

Can we define the WG-Lasso by using proximity operator ?

thanks to the following strategy

- map the original coefficients into a bigger space;
- define independent groups over the neighborhood of the coefficients;
- apply the (group-lasso) proximity operator;
- go back to the original space.

Moreover, can we use the WG-Lasso inside ISTA ?

Expanded operators

Definition : Expanding operator

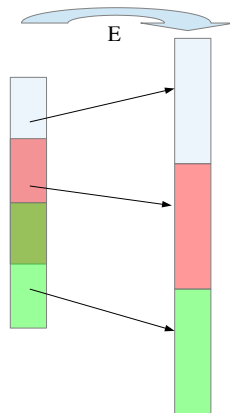
Let $\alpha \in \mathbb{C}^N$. Let $\mathbf{E} : \mathbb{C}^N \rightarrow \mathbb{C}^{N \times N}$ be an expanded operator such that

$$\alpha = (\alpha_1, \dots, \alpha_N) \mapsto (w_1^1 \alpha_1, w_2^1 \alpha_2, \dots, w_N^1 \alpha_N, \dots, w_1^N \alpha_1, \dots, w_N^N \alpha_N)$$

with $w_i^j \geq 0$, $\sum_j |w_i^j|^2 = 1$ and $w_i^i > 0$

proposition

$\mathbf{E}$ is isometrical, and then $\mathbf{E}^T \mathbf{E} = \mathbf{I}$.



A left inverse

Definition : a natural left inverse

$$\begin{aligned}
 \mathbf{D} &: \mathbb{C}^{N \times N} \rightarrow \mathbb{C}^N \\
 \mathbf{z} = (z_1^1, \dots, z_N^1, \dots, z_1^N, \dots, z_N^N) &\mapsto \mathbf{x} \\
 \text{such that } \forall k, x_k &= \frac{1}{w_k^k} z_k^k
 \end{aligned} \tag{1}$$

$\mathbf{DE} = \mathbf{I}$ and then $\mathbf{DE}$ is a bi-orthogonal (oblique) projection.

Structured shrinkage and proximity operators

proposition

Let $\mathbb{S}$ be the shrinkage operator of the WG-Lasso and $\Omega = \|\cdot\|_{21}$ the regularizer of the G-lasso. Let $\mathbf{E}$ be the expanded operator as previously defined and $\mathbf{D}$ its left inverse. Then

$$\mathbb{S}(\cdot, \lambda) = \mathbf{D} \circ \text{prox}_{\lambda\Omega} \circ \mathbf{E}$$

$$\hat{\alpha}_k = \tilde{y}_k \left(1 - \frac{\lambda}{\sqrt{\sum_{m \in g(k)} |\tilde{y}_m|^2}} \right)^+ = \tilde{y}_k \left(1 - \frac{\lambda}{\|\tilde{y}_{g(k)}\|_2} \right)^+$$

Remark

$\mathbb{S}$ cannot be a proximity operator (it is even not a nonexpansive operator).

Neighborhood as a convex prior

social sparsity convex regularizers

Let $\alpha \in \mathbb{C}^N$ and let $\mathbf{E}$ be the expanded operator.

cvx windowed group lasso:

$$\begin{aligned}\Omega_{wgl}(\alpha) &= \sum_{k=1}^N \sqrt{\sum_{\ell \in \mathcal{N}(k)} w_{\ell}^{(k)} |\alpha_{\ell}|^2} \\ &= \|\mathbf{E}\alpha\|_{21}\end{aligned}$$

cvx windowed elitist lasso:

$$\begin{aligned}\Omega_{wel}(\alpha) &= \sum_{k=1}^N \left(\sum_{\ell \in \mathcal{N}(k)} w_{\ell}^{(k)} |\alpha_{\ell}| \right)^2 \\ &= \|\mathbf{E}\alpha\|_{12}^2\end{aligned}$$

A convex functional for social sparsity

A natural convex functional is (aka group-Lasso with overlaps [Bayram 11])

$$F(\alpha) = \frac{1}{2} \|\mathbf{y} - \Phi \alpha\|^2 + \lambda \|\mathbf{E} \alpha\|_{21}$$

one can look for

$$\begin{aligned} \hat{\alpha} &= \operatorname{argmin}_{\alpha \in \mathbb{C}^N} F(\alpha) \\ &= \mathbf{E}^T \operatorname{argmin}_{\mathbf{z}} \frac{1}{2} \|\mathbf{y} - \Phi \mathbf{E}^T \mathbf{z}\|^2 + \lambda \|\mathbf{z}\|_{21} \\ &\quad \text{s.t. } \mathbf{E} \mathbf{E}^T \mathbf{z} = \mathbf{z} \end{aligned}$$

- Similar functional introduced by [Peyré & Fadili 11].
- several approach can be used to minimize F (ISTA + Douglas Rachford, augmented lagrangian. . .)

But: this penalty acts as a *discarding* procedure, not a *selection*.

G-Lasso with overlaps VS latent-G-Lasso

Instead of

$$F(\alpha) = \frac{1}{2} \|\mathbf{y} - \Phi \alpha\|^2 + \lambda \|\mathbf{E} \alpha\|_{21}$$

[Jacob & al. 09] propose to minimize

$$F(\tilde{\alpha}) = \frac{1}{2} \|\mathbf{y} - \Phi \mathbf{E}^T \tilde{\alpha}\|^2 + \lambda \|\tilde{\alpha}\|_{21}$$

to obtain a *selection* of active groups.

Curse of dimension in both cases !

Link between the convex functional and our shrinkages

ISTA with WG-Lasso becomes:

$$\mathbf{z}^{(k)} = \mathbf{E}\mathbf{D} \operatorname{prox}_{\frac{\lambda}{\gamma}\|\cdot\|_*} \left(\left(\tilde{\mathbf{z}}^{(k-1)} \right) \right)$$

$$\boldsymbol{\alpha}^k = \mathbf{D}\mathbf{z}^k$$

$$\text{where } \tilde{\mathbf{z}}^{(k-1)} = \mathbf{z}^{(k-1)} + \frac{\mathbf{E}}{\gamma} \boldsymbol{\Phi}^* (\mathbf{y} - \boldsymbol{\Phi}\mathbf{E}^T \mathbf{z}^{(k-1)})$$

It is a proximal descent followed by an *oblique* projection on $\operatorname{Im}(\mathbf{E})$.

conjecture

ISTA with WG-Lasso converges to a fixed point.

Orthogonal social sparsity

An Orthogonal version

$$\mathbf{z}^{(k)} = \mathbf{E}\mathbf{E}^T \operatorname{prox}_{\frac{\lambda}{\gamma}\|\cdot\|_*} \left(\left(\tilde{\mathbf{z}}^{(k-1)} \right) \right)$$

where $\tilde{\mathbf{z}}^{(k-1)} = \mathbf{z}^{(k-1)} + \frac{\mathbf{E}}{\gamma} \boldsymbol{\Phi}^* (\mathbf{y} - \boldsymbol{\Phi}\mathbf{E}^T \mathbf{z}^{(k-1)})$

orth-WG-Lasso

$$\alpha_k = \tilde{y}_k \sum_j \frac{1}{w_j^j} \left(1 - \frac{\lambda}{\sqrt{\sum_{j' \in \mathcal{N}(j)} w_{j'}^{(j)} |\tilde{y}_{k'}|^2}} \right)^+.$$

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

A family of shrinkage operators

$\alpha = \mathbb{S}(\mathbf{y})$ 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

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

- Empirical Persistent Wiener [Siedenburg 13]

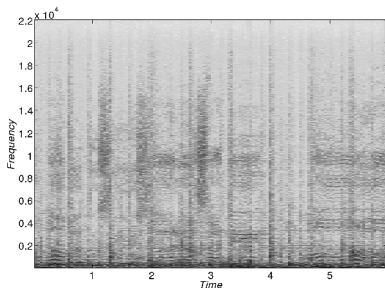
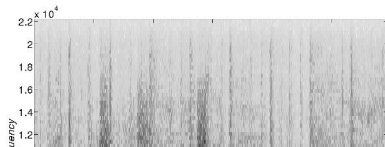
$$\alpha_k = \tilde{y}_k \left(1 - \frac{\lambda}{\|\tilde{\mathbf{y}}_{g(k)}\|_2^2} \right)^+$$

- 1 Introduction: sparse approximation
- 2 An optimization framework
 - General approach
 - Mixed Norms
 - Hybrid (or multilayers) models
- 3 Iterative Thresholding
 - Thresholding functions
 - Neighborhood thresholding
- 4 Numerical results
 - Application to tonal/transient separation
 - Simulations
 - Audio declipping
- 5 Conclusion

Tonal/transient separation - 1

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

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



Tonal/transient separation - 2

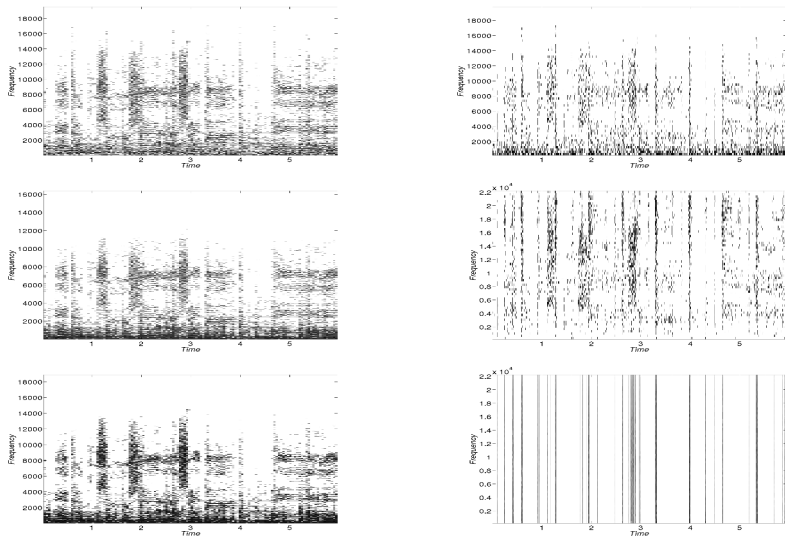


FIGURE : Left: tonal layers. Right: transient layers. From top to bottom: LASSO/LASSO, LASSO/ELASSO, LASSO/GLASSO.

- 1 Introduction: sparse approximation
- 2 An optimization framework
 - General approach
 - Mixed Norms
 - Hybrid (or multilayers) models
- 3 Iterative Thresholding
 - Thresholding functions
 - Neighborhood thresholding
- 4 Numerical results
 - Application to tonal/transient separation
 - Simulations
 - Audio declipping
- 5 Conclusion

Simulation of structured sparse signals

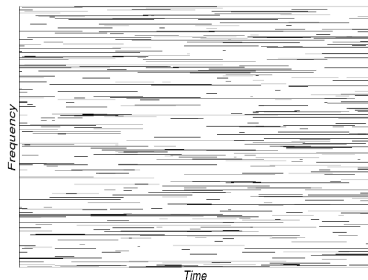


FIGURE : A structured sparse significance map Δ generated by Markov chain.

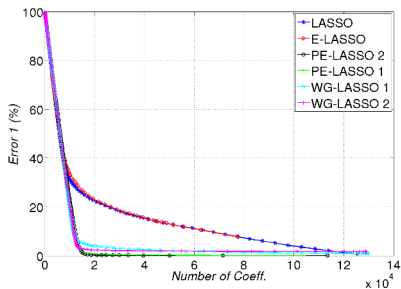
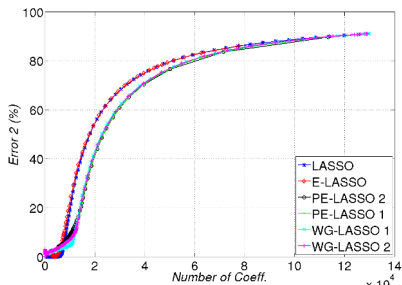
Simulated signal:

$$y = \sum_{(t,f) \in \Delta} x_{(t,f)} \varphi_{(t,f)} + b ,$$

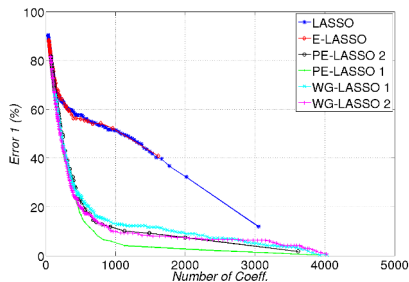
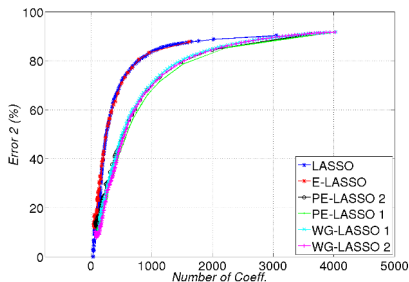
with:

- Δ the **significance map** generated by Markov chain.
- For $(t, f) \in \Delta$, $x_{(t,f)} \sim \mathcal{N}(0, \sigma^2)$.
- b a white gaussian noise.

Orthogonal case

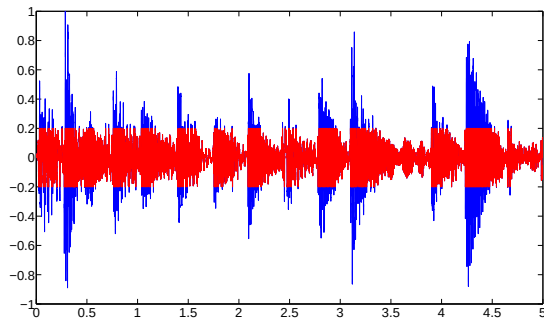


Frame case



- 1 Introduction: sparse approximation
- 2 An optimization framework
 - General approach
 - Mixed Norms
 - Hybrid (or multilayers) models
- 3 Iterative Thresholding
 - Thresholding functions
 - Neighborhood thresholding
- 4 Numerical results
 - Application to tonal/transient separation
 - Simulations
 - Audio declipping
- 5 Conclusion

Audio Declicking



Audio inpainting: forward problem [A. Adler, V. Emiya *et Al*]

$$\mathbf{y}^r = \mathbf{M}^r \mathbf{x}$$

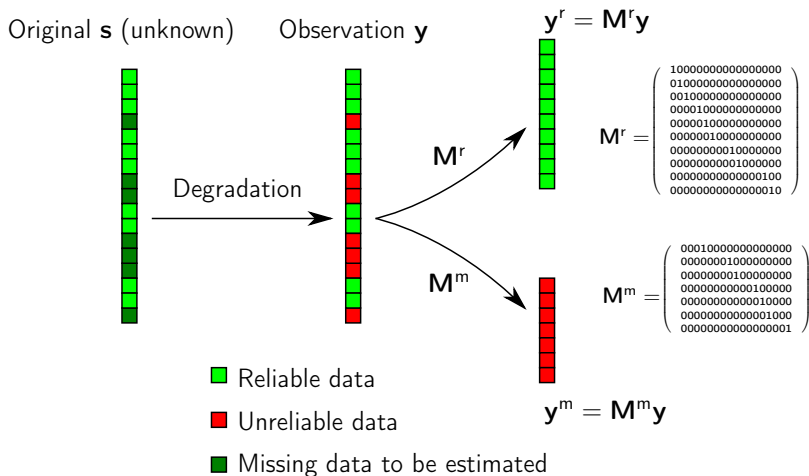
where

- $\mathbf{x} \in \mathbb{R}^N$ is the unknown “clean” signal;
- $\mathbf{y}^r \in \mathbb{R}^M$ are the “reliable” sample of the observed signal
- $\mathbf{M}^r \in \mathbb{R}^{M \times N}$ is the matrix of the reliable support of $\mathbf{x}$

we can also define the missing samples as

$$\mathbf{y}^m = \mathbf{M}^m \mathbf{x}$$

Reliable vs Unreliable coeff.



Audio declipping: (constrained and convex) inverse problem

For audio declipping, we can add the following constraint

$$\begin{aligned}\hat{\alpha} &= \underset{\alpha}{\operatorname{argmin}} \frac{1}{2} \|\mathbf{y}^r - \mathbf{M}^r \Phi \alpha\| + \lambda \|\alpha\|_1 \\ \text{s.t. } \quad & \mathbf{M}^{m^+} \Phi \alpha > \theta^{clip} \\ & \mathbf{M}^{m^-} \Phi \alpha < \theta^{clip}\end{aligned}$$

where $\mathbf{M}^{m^+}$ (resp. $\mathbf{M}^{m^-}$) select the positive (resp. negative) samples.

Problem: cannot be solved “efficiently” with (F)ISTA

Audio declipping: (convex unconstrained) inverse problem

Let

$$[\theta^{clip} - \mathbf{x}]_+^2 = \sum_{k: \theta_k^{clip} > 0} (\theta_k^{clip} - x_k)_+^2 + \sum_{k: \theta_k^{clip} < 0} (-\theta_k^{clip} + x_k)_+^2$$

We consider the following unconstrained convex problem:

$$\alpha = \underset{\alpha}{\operatorname{argmin}} \frac{1}{2} \|\mathbf{y}^r - \mathbf{M}^r \Phi \alpha\|_2^2 + \frac{1}{2} [\theta^{clip} - \mathbf{M}^m \Phi \alpha]_+^2 + P(\alpha; \lambda)$$

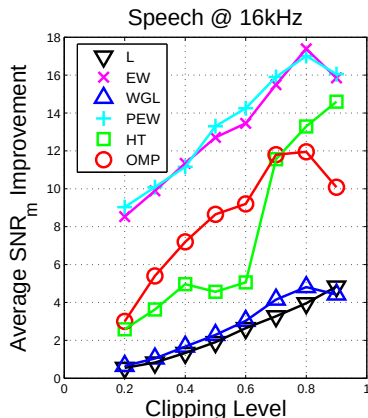
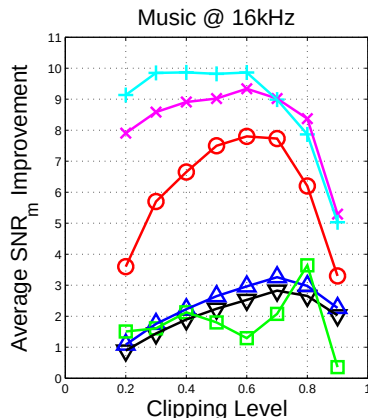
which is under the form

$$f_1(\alpha) + f_2(\alpha)$$

with f_1 Lipschitz-differentiable and f_2 semi-convex.

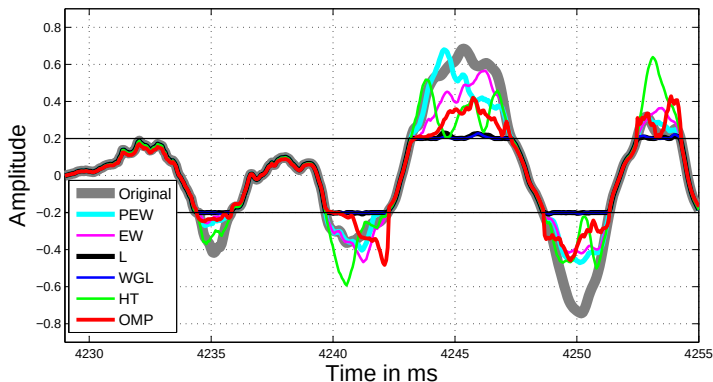
We can apply (relaxed)-ISTA directly !

Numerical results



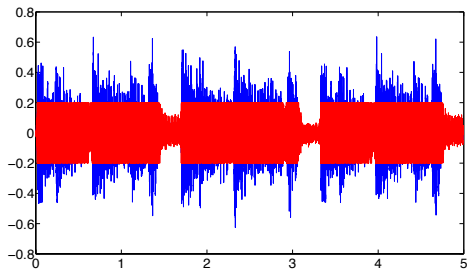
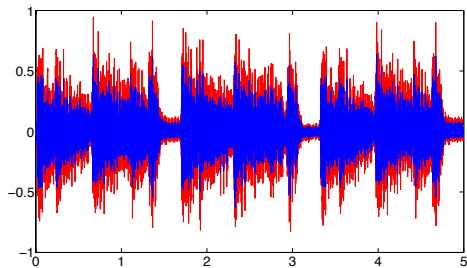
Average SNR_{miss} for 10 speech (left) and music (right) signals over different clipping levels and operators. Neighborhoods extend 3 and 7 coefficients in time for speech and music signals, respectively.

Numerical results: zoom on reconstructions



Declicked music signal using different operators for clip level $\theta^{clip} = 0.2$ using the Lasso, WGL, EW, PEW, HT, and OMP operators. Neighborhood size for WGL and PEW was 7.

Original Vs clipped Vs declipped Signal



- 1 Introduction: sparse approximation
- 2 An optimization framework
 - General approach
 - Mixed Norms
 - Hybrid (or multilayers) models
- 3 Iterative Thresholding
 - Thresholding functions
 - Neighborhood thresholding
- 4 Numerical results
 - Application to tonal/transient separation
 - Simulations
 - Audio declipping
- 5 Conclusion

Conclusion

Take home messages

- Use dictionary to get sparsity
- Play on thresholding rules in ISTA
- Define some neighborhoods for “flexible” structures

Next...

- Some practical issues (warm start: how many iterations, λ)
- Some theoretical issues (more on convergence)