Generalized Concomitant Multi-Task Lasso for sparse multimodal regression

Joseph Salmon
http://josephsalmon.eu
LTCI, Télécom Paristech, Université Paris-Saclay

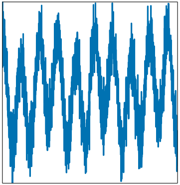
Joint work with:
Mathurin Massias (INRIA, Parietal Team)
Olivier Fercoq (Télécom ParisTech)
Alexandre Gramfort (INRIA, Parietal Team)

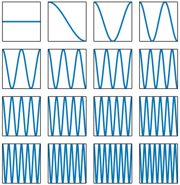
Table of Contents

Sparsity is all around

Signals can often be represented through a combination of a few **atoms** / **features** :

- Fourier decomposition for sounds

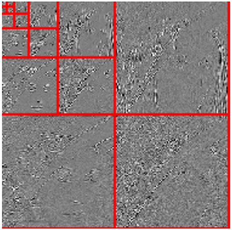




Sparsity is all around

Signals can often be represented through a combination of a few **atoms** / **features** :

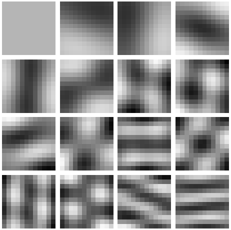
- Fourier decomposition for sounds
- Wavelet for images (1990's)



Sparsity is all around

Signals can often be represented through a combination of a few **atoms** / **features** :

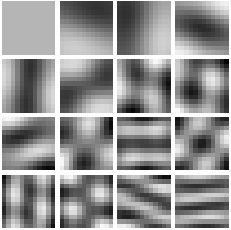
- Fourier decomposition for sounds
- Wavelet for images (1990's)
- Dictionary learning for images (late 2000's)



Sparsity is all around

Signals can often be represented through a combination of a few **atoms** / **features** :

- Fourier decomposition for sounds
- Wavelet for images (1990's)
- Dictionary learning for images (late 2000's)
- More inverse problems

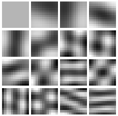
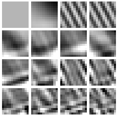


Simplest model: standard sparse regression

$y \in \mathbb{R}^n$: a signal

$X = [\mathbf{x}_1, \dots, \mathbf{x}_p] \in \mathbb{R}^{n \times p}$: **dictionary** of atoms/features

Assumption : signal well approximated by a **sparse** combination $\beta^* \in \mathbb{R}^p$: $y \approx X\beta^*$

$$\underbrace{\begin{bmatrix} y \end{bmatrix}}_{y \in \mathbb{R}^n} \approx \underbrace{\begin{bmatrix} \mathbf{x}_1 & \Big| & \dots & \Big| & \mathbf{x}_p \end{bmatrix}}_{X \in \mathbb{R}^{n \times p}} \cdot \underbrace{\begin{bmatrix} \beta_1^* \\ \vdots \\ \beta_p^* \end{bmatrix}}_{\beta \in \mathbb{R}^p}$$

Objective(s): find $\hat{\beta}$

- Estimation: $\hat{\beta} \approx \beta^*$
- Prediction: $X\hat{\beta} \approx X\beta^*$
- Support recovery: $\mathrm{supp}(\hat{\beta}) \approx \mathrm{supp}(\beta^*)$

$$y \approx \sum_{j=1}^{p} \beta_j^* \mathbf{x}_j$$

Constraints: large p, sparse β^*

The ℓ_0 penalty

Objective: use Least-Squares with an ℓ_0 penalty to enforce sparsity

$$\boxed{\underset{\beta \in \mathbb{R}^p}{\arg\min} \left(\underbrace{\frac{1}{2}\|y - X\beta\|_2^2}_{\text{data fitting}} + \underbrace{\lambda\|\beta\|_0}_{\text{regularization}} \right)}$$

where $\|\beta\|_0 = \mathrm{card}(\{j \in [\![1, p]\!], \beta_j \neq 0\}) = \mathrm{card}(\mathrm{supp}(\beta))$

Combinatorial problem; "NP-hard" Natarajan (1995)

$\hookrightarrow$ Exact resolution requires Least-Squares (LS) solutions for all sub-models, *i.e.*, compute LS for all possible supports (up to 2^p)

- $p = 10 \hookrightarrow$ possible: $\approx 10^3$ least squares
- $p = 30 \hookrightarrow$ impossible: $\approx 10^{10}$ least squares

The ℓ_1 penalty: Lasso and variants

Vocabulary: the "Modern least square" Candès *et al.* (2008)

- Statistics: **Lasso** Tibshirani (1996)
- Signal processing variant: **Basis Pursuit** Chen *et al.* (1998)

$$\boxed{\hat{\beta}^{(\lambda)} \in \operatorname*{arg\,min}_{\beta \in \mathbb{R}^p} \left(\underbrace{\frac{1}{2}\|y - X\beta\|^2}_{\textbf{data fitting term}} + \underbrace{\lambda\|\beta\|_1}_{\textbf{sparsity-inducing penalty}} \right)}$$

- Solutions are **sparse** (sparsity level controlled by λ)

A multi-task framework

Multi-task regression:

- n observations
- q tasks (hereafter: temporal information)
- p features
- $Y \in \mathbb{R}^{n \times q}$ observation matrix
- $X \in \mathbb{R}^{n \times p}$ forward matrix

$$\boxed{Y = X\mathrm{B}^* + E}$$

where

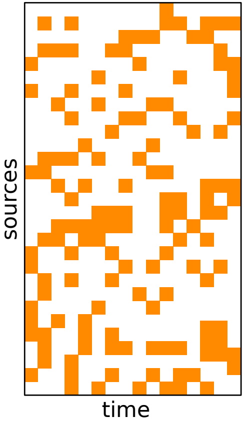
- $\mathrm{B}^* \in \mathbb{R}^{p \times q}$: true source activity matrix
- $E \in \mathbb{R}^{n \times q}$: additive white Gaussian noise; no additional assumption yet

Notation point: capital letters refer to matrices

Multi-tasks penalties Obozinski *et al.* (2010)

Popular convex penalties considered:

$$\hat{\mathrm{B}} \in \operatorname*{arg\,min}_{\mathrm{B} \in \mathbb{R}^{p \times q}} \left(\frac{1}{2nq} \left\| Y - X\mathrm{B} \right\|^2 + \lambda \Omega(\mathrm{B}) \right)$$



Parameter $\hat{\mathrm{B}} \in \mathbb{R}^{p \times q}$

Sparse support: no structure

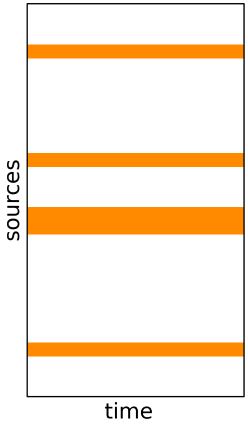
Penalty: Lasso

$$\|\mathrm{B}\|_1 = \sum_{j=1}^{p} \sum_{k=1}^{q} |\mathrm{B}_{j,k}|$$

Multi-tasks penalties Obozinski *et al.* (2010)

Popular convex penalties considered: Multi-Task Lasso (MTL)

$$\hat{\mathrm{B}} \in \mathop{\arg\min}_{\mathrm{B} \in \mathbb{R}^{p \times q}} \left(\frac{1}{2nq} \|Y - X\mathrm{B}\|^2 + \lambda \Omega(\mathrm{B}) \right)$$



Parameter $\hat{\mathrm{B}} \in \mathbb{R}^{p \times q}$

Sparse support: group structure

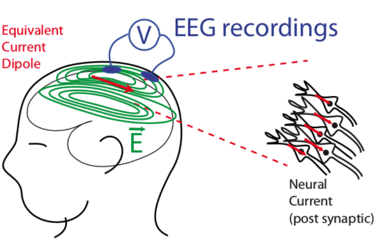
Penalty: Group-Lasso

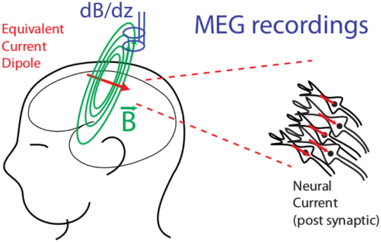
$$\|\mathrm{B}\|_{2,1} = \sum_{j=1}^{p} \|\mathrm{B}_{j,:}\|_2$$

where $\mathrm{B}_{j,:}$ the j-th line of B

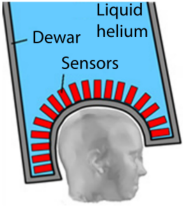
M/EEG inverse problem for brain imaging

- sensors: magneto- and electro-encephalogram measurements during a cognitive experiment
- sources: brain locations





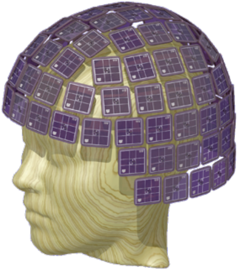
First EEG
recordings
in 1929
by H. Berger



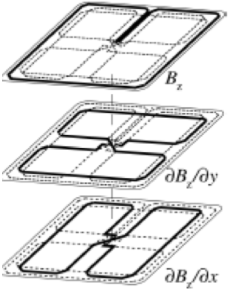
Hôpital La Timone
Marseille, France

MEG elements

Device



Sensors



Detail of a sensor

The M/EEG inverse problem: modeling

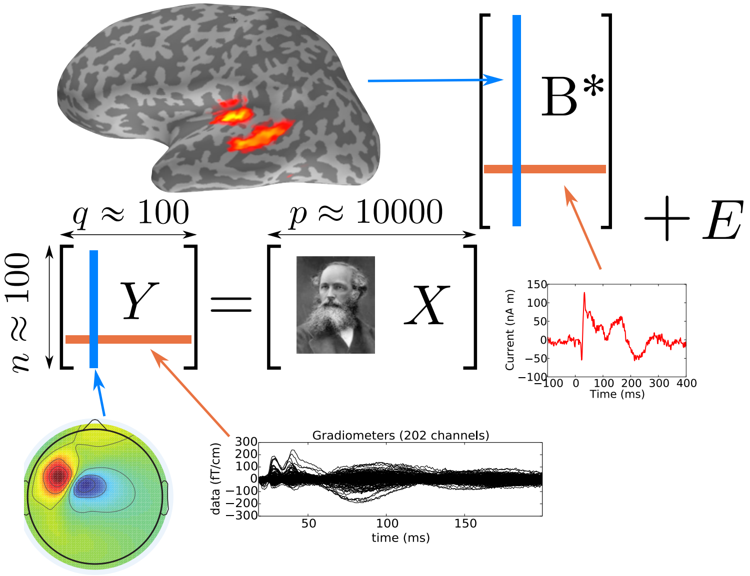


Table of Contents

Gaussian model and Lasso (single task, $q = 1$)

Sparse Gaussian model: $y = X\beta^* + \sigma_* \varepsilon$

- $y \in \mathbb{R}^n$: observation
- $X \in \mathbb{R}^{n \times p}$: design matrix
- $\beta^* \in \mathbb{R}^p$: signal to recover; unknown
- $\|\beta^*\|_0 = s^*$: sparsity level (small w.r.t. p); s^* unknown
- $\varepsilon \sim \mathcal{N}(0, \sigma_*^2 \operatorname{Id}_n)$; σ_* unknown

Lasso reminder : $\boxed{\hat{\beta}^{(\lambda)} \in \underset{\beta \in \mathbb{R}^p}{\arg\min} \frac{\|y - X\beta\|^2}{2n} + \lambda \|\beta\|_1}$

Lasso theory : (fairly) well understood

Theorem Bickel *et al.* (2009), Dalalyan *et al.* (2017)

For Gaussian noise model with X satisfying the "Restricted Eigenvalue" property and $\lambda = 2\sigma_* \sqrt{\frac{2\log(p/\delta)}{n}}$, then

$$\frac{1}{n}\left\| X(\beta^* - \hat{\beta}^{(\lambda)}) \right\|^2 \leq \frac{18}{\kappa_{s^*}^2} \frac{\sigma_*^2 s^*}{n} \log\left(\frac{p}{\delta}\right)$$

with probability $1-\delta$, where $\hat{\beta}^{(\lambda)}$ is a Lasso solution

Rem: optimal rate in the minimax sense (up to constant/log term)

Rem: under the "Restricted Eigenvalue" property, $\kappa_{s^*}^2$ controls strong convexity of the (quadratic part of the) objective function obtained when extracting s^* columns of X

Yet σ_* is unknown in practice !

Lasso theory : (fairly) well understood

Theorem Bickel *et al.* (2009), Dalalyan *et al.* (2017)

For Gaussian noise model with X satisfying the "Restricted Eigenvalue" property and $\lambda = 2\sigma_* \sqrt{\frac{2\log(p/\delta)}{n}}$, then

$$\frac{1}{n} \left\| X(\beta^* - \hat{\beta}^{(\lambda)}) \right\|^2 \leq \frac{18}{\kappa_{s^*}^2} \frac{\sigma_*^2 s^*}{n} \log\left(\frac{p}{\delta}\right)$$

with probability $1-\delta$, where $\hat{\beta}^{(\lambda)}$ is a Lasso solution

Rem: optimal rate in the minimax sense (up to constant/log term)

Rem: under the "Restricted Eigenvalue" property, $\kappa_{s^*}^2$ controls strong convexity of the (quadratic part of the) objective function obtained when extracting s^* columns of X

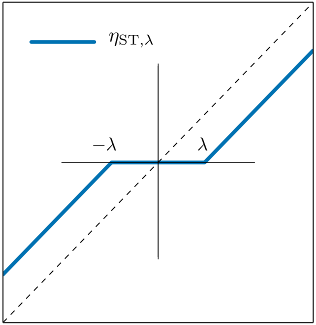
Yet σ_* is unknown in practice !

Soft-Thresholding: Lasso for orthogonal design

Closed form solution for 1D-problem $(p = 1)$: **Soft-Thresholding**

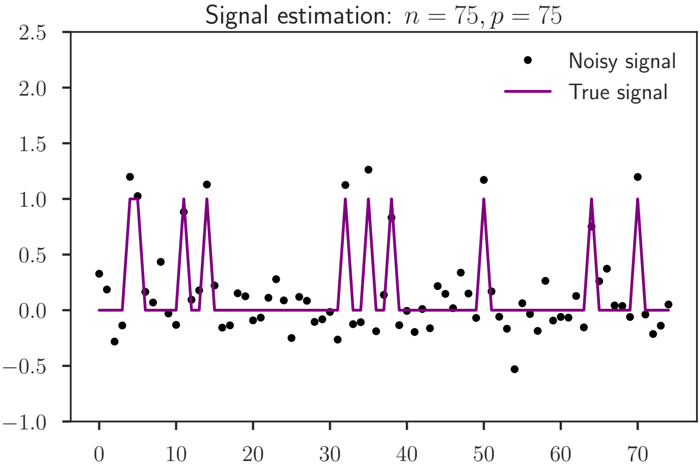
$$\eta_{\mathrm{ST},\lambda}(y) := \underset{\beta \in \mathbb{R}}{\arg\min} \left(\frac{(y-\beta)^2}{2} + \lambda|\beta| \right)$$
$$= \mathrm{sign}(y)(|y| - \lambda)_+$$

with $(\cdot)_+ := \max(0, \cdot)$

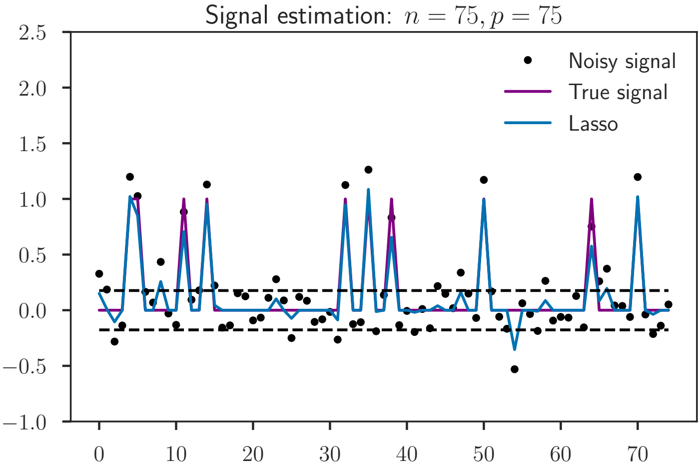


Extension for $X = \mathrm{Id}_p$: component-wise soft thresholding

"Universal" λ (orthogonal design $X = \mathrm{Id}_n$)



"Universal" λ (orthogonal design $X = \mathrm{Id}_n$)



Dash lines : $\pm\lambda = 2\sigma_* \sqrt{\frac{2\log(p/\delta)}{n}}$ ($\sigma_* = 0.2$ known, $\delta = 0.05$)

Joint estimation of β and σ

How to perform λ calibration when σ_* is unknown?

Intuitive idea:

- run Lasso with some λ, get $\hat{\beta}$
- estimate σ with residuals: $\sigma = \|y - X\hat{\beta}\| / \sqrt{n}$
- relaunch Lasso with $\lambda \propto \sigma$
- iterate, ...

Note: this is the original implementation proposed for the Scaled-Lasso Sun and Zhang (2012)

Concomitant Lasso Owen (2007)

$$\boxed{(\beta^{(\lambda)}, \sigma^{(\lambda)}) \in \underset{\beta \in \mathbb{R}^p, \sigma > 0}{\arg\min} \left(\frac{\|y - X\beta\|^2}{2n\sigma} + \frac{\sigma}{2} + \lambda \|\beta\|_1 \right)}$$

- $\frac{\sigma}{2}$ acts as a penalty over the noise level

Concomitant Lasso Owen (2007)

$$\boxed{(\beta^{(\lambda)}, \sigma^{(\lambda)}) \in \operatorname*{arg\,min}_{\beta \in \mathbb{R}^p, \sigma > 0} \left(\frac{\|y - X\beta\|^2}{2n\sigma} + \frac{\sigma}{2} + \lambda \|\beta\|_1 \right)}$$

- $\frac{\sigma}{2}$ acts as a penalty over the noise level
- Roots in Huber (1981)'s work on robust estimation

Concomitant Lasso Owen (2007)

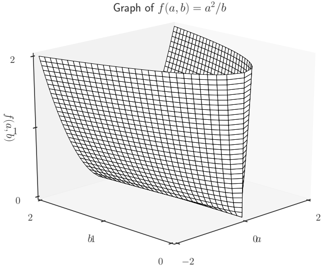
$$\boxed{(\beta^{(\lambda)}, \sigma^{(\lambda)}) \in \mathop{\arg\min}_{\beta \in \mathbb{R}^p, \sigma > 0} \left(\frac{\|y - X\beta\|^2}{2n\sigma} + \frac{\sigma}{2} + \lambda \|\beta\|_1 \right)}$$

- $\frac{\sigma}{2}$ acts as a penalty over the noise level
- Roots in Huber (1981)'s work on robust estimation
- jointly convex program: $(a, b) \mapsto a^2/b$ is convex

Concomitant Lasso Owen (2007)

$$\boxed{(\beta^{(\lambda)}, \sigma^{(\lambda)}) \in \underset{\beta \in \mathbb{R}^p, \sigma > 0}{\arg\min} \left(\frac{\|y - X\beta\|^2}{2n\sigma} + \frac{\sigma}{2} + \lambda \|\beta\|_1 \right)}$$

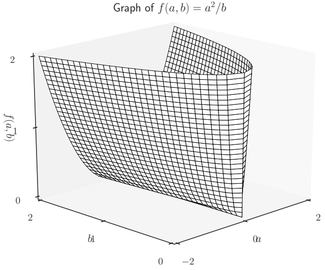
- $\frac{\sigma}{2}$ acts as a penalty over the noise level
- Roots in Huber (1981)'s work on robust estimation
- jointly convex program: $(a, b) \mapsto a^2/b$ is convex



Concomitant Lasso Owen (2007)

$$\boxed{(\beta^{(\lambda)}, \sigma^{(\lambda)}) \in \underset{\beta \in \mathbb{R}^p, \sigma > 0}{\arg\min} \left(\frac{\|y - X\beta\|^2}{2n\sigma} + \frac{\sigma}{2} + \lambda \|\beta\|_1 \right)}$$

- $\frac{\sigma}{2}$ acts as a penalty over the noise level
- Roots in Huber (1981)'s work on robust estimation
- jointly convex program: $(a, b) \mapsto a^2/b$ is convex



Concomitant performance

Theorem Sun and Zhang (2012)

For Gaussian noise model with X satisfying the "Restricted Eigenvalue" property and $\lambda = 2\sqrt{\frac{2\log(p/\delta)}{n}}$, then

$$\frac{1}{n}\left\|X(\beta^* - \hat{\beta}^{(\lambda)})\right\|^2 \leq \frac{18}{\kappa_{s^*}^2}\frac{\sigma_*^2 s_*}{n}\log\left(\frac{p}{\delta}\right)$$

with "high" probability, where $\hat{\beta}^{(\lambda)}$ is a Concomitant Lasso solution

Rem: provide same rate as Lasso, without knowing σ_*

Rem: "high" refers to the (complex) dependency on δ

Link with the $\sqrt{\text{Lasso}}$ Belloni *et al.* (2011)

- Independently, Belloni *et al.* (2011) analyzed $\sqrt{\text{Lasso}}$ to get "σ free" choice of λ

$$\boxed{\hat{\beta}^{(\lambda)}_{\sqrt{\text{Lasso}}} \in \mathop{\arg\min}_{\beta \in \mathbb{R}^p} \left(\frac{1}{\sqrt{n}} \|y - X\beta\| + \lambda \|\beta\|_1 \right)}$$

- Connections with Concomitant Lasso: $\left(\hat{\beta}^{(\lambda)}_{\sqrt{\text{Lasso}}}, \hat{\sigma}^{(\lambda)}_{\sqrt{\text{Lasso}}}\right)$ is solution of the Concomitant Lasso for

$$\hat{\sigma}^{(\lambda)}_{\sqrt{\text{Lasso}}} = \frac{\left\| y - X\hat{\beta}^{(\lambda)}_{\sqrt{\text{Lasso}}} \right\|}{\sqrt{n}}$$

<u>Rem</u>: non-smooth data fitting term with non-smooth regularization

The Smoothed Concomitant Lasso
Ndiaye *et al.* (2016)

To remove issues for small λ (and σ), we have introduced:

$$\boxed{(\hat{\beta}^{(\lambda)}, \hat{\sigma}^{(\lambda)}) \in \underset{\beta \in \mathbb{R}^p, \sigma \geq \underline{\sigma}}{\arg\min} \ \frac{\|y - X\beta\|^2}{2n\sigma} + \frac{\sigma}{2} + \lambda \|\beta\|_1}$$

- With prior information on the minimal noise level, one can set $\underline{\sigma}$ as this bound (and both estimators are the same)
- Setting $\underline{\sigma} = \epsilon$, smoothing theory asserts that $\frac{\epsilon}{2}$-solutions for the smoothed problem provide ϵ-solutions for the $\sqrt{\text{Lasso}}$ problem Nesterov (2005)

Smoothing aparté

Nesterov (2005), Beck and Teboulle (2012)

Motivation: smooth a non-smooth function f to ease optimization

Smoothing step: for $\mu > 0$, a "smoothed" version of f is f_μ

$$f_\mu = \mu\omega\left(\frac{\cdot}{\mu}\right)\Box f$$

- **inf-convolution**: $f\Box g(x) = \inf_u \{f(u) + g(x-u)\}$
- ω is a predefined smooth function (such that $\nabla\omega$ is Lipschitz)

Analogy with "kernel smoothing":

- usual convolution "$\star$" $\to$ inf-convolution "$\Box$"
- Fourier transform exchange "$\star$" and "$\times$" $\to$ Legendre transform exchange "$\Box$" and "$+$"
- Gaussian kernel $\to \|\cdot\|^2/2$
- in both cases μ controls the scaling (bandwidth)

Smoothing aparté
Nesterov (2005), Beck and Teboulle (2012)

Motivation: smooth a non-smooth function f to ease optimization

Smoothing step: for $\mu > 0$, a "smoothed" version of f is f_μ

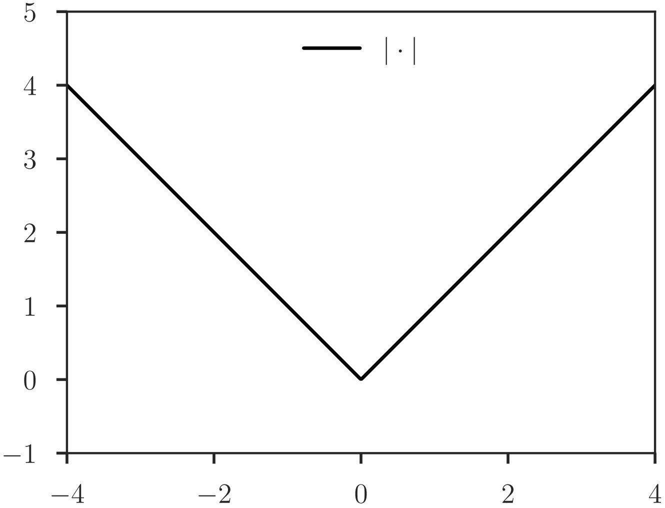
$$f_\mu = \mu\omega\left(\frac{\cdot}{\mu}\right)\Box f$$

- **inf-convolution**: $f\Box g(x) = \inf_u \{f(u) + g(x-u)\}$
- ω is a predefined smooth function (such that $\nabla\omega$ is Lipschitz)

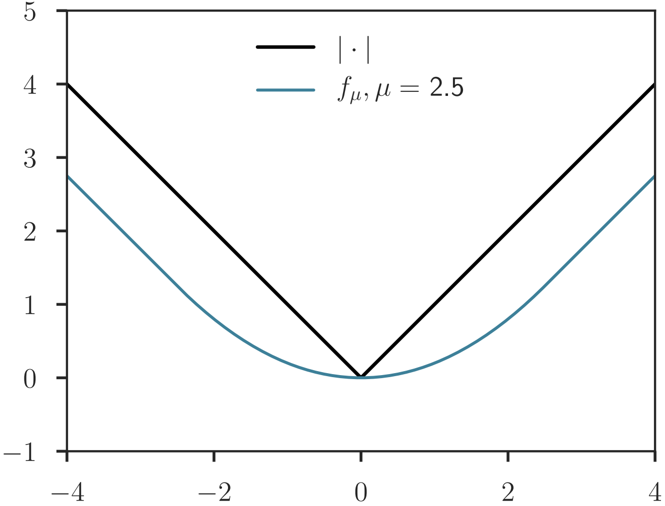
Analogy with "kernel smoothing":

- usual convolution "$\star$" $\rightarrow$ inf-convolution "$\Box$"
- Fourier transform exchange "$\star$" and "$\times$" $\rightarrow$ Legendre transform exchange "$\Box$" and "$+$"
- Gaussian kernel $\rightarrow \|\cdot\|^2/2$
- in both cases μ controls the scaling (bandwidth)

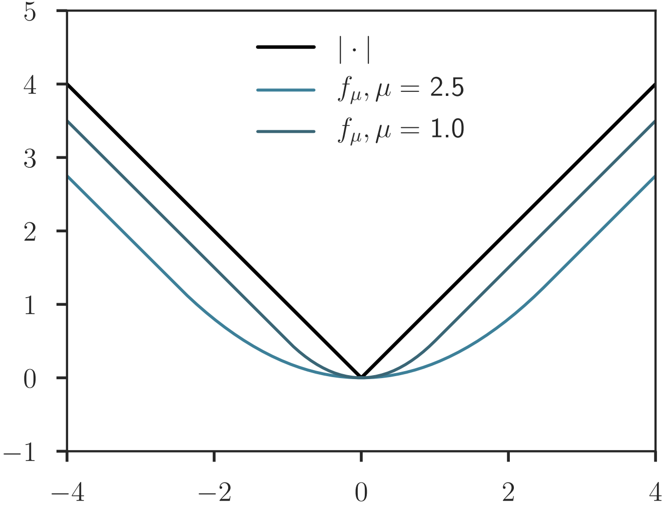
Huber function: $\omega(t) = \frac{t^2}{2}$



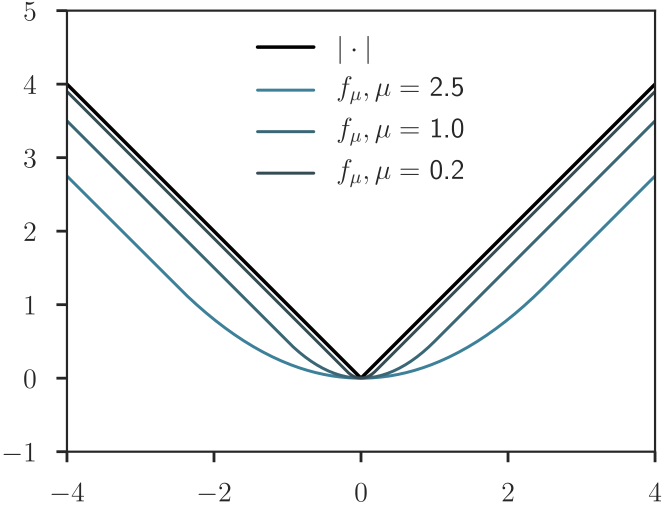
Huber function: $\omega(t) = \frac{t^2}{2}$



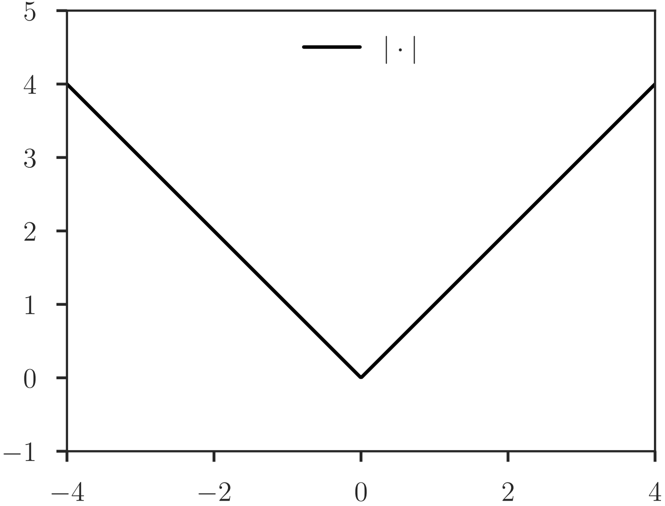
Huber function: $\omega(t) = \frac{t^2}{2}$



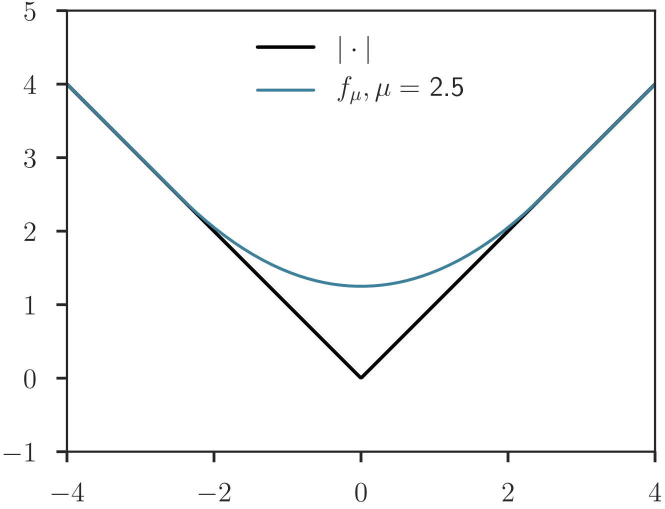
Huber function: $\omega(t) = \frac{t^2}{2}$



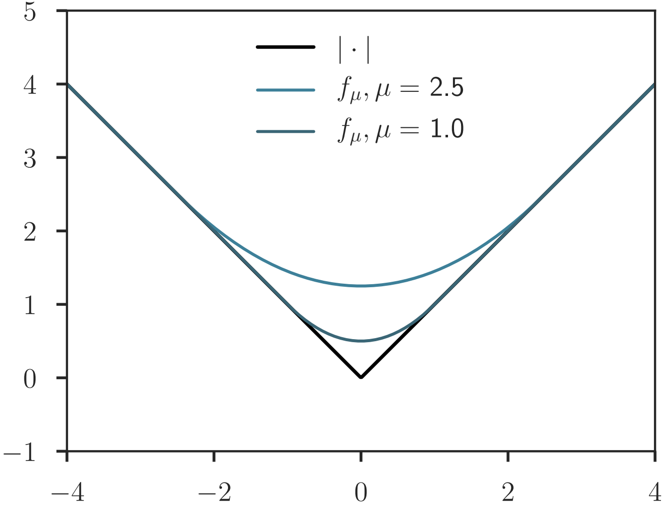
Huber function (bis): $\omega(t) = \frac{t^2}{2} + \frac{1}{2}$



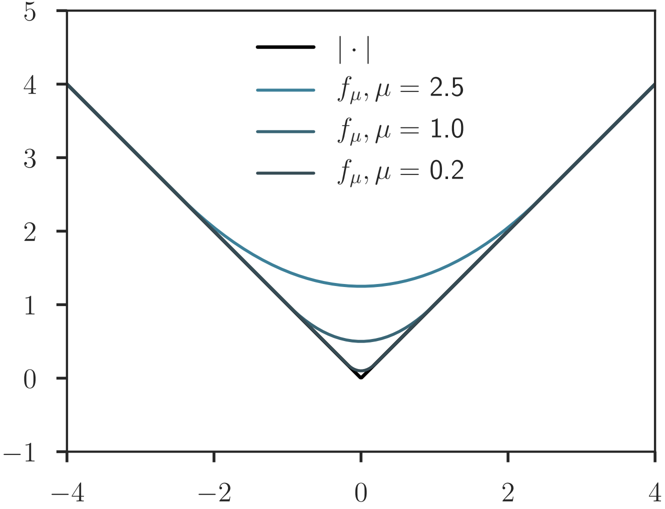
Huber function (bis): $\omega(t) = \frac{t^2}{2} + \frac{1}{2}$



Huber function (bis): $\omega(t) = \frac{t^2}{2} + \frac{1}{2}$



Huber function (bis): $\omega(t) = \frac{t^2}{2} + \frac{1}{2}$



Huberization of the $\sqrt{\text{Lasso}}$

"**Huberization**": $f(\beta) = \frac{\|y - X\beta\|}{\sqrt{n}}$, $\mu = \underline{\sigma}$, $\omega(\beta) = \frac{\|\beta\|^2}{2} + \frac{1}{2}$

$$
\begin{aligned}
f_{\underline{\sigma}}(\beta) &= \begin{cases} \frac{\|y - X\beta\|^2}{2n\underline{\sigma}} + \frac{\underline{\sigma}}{2} & \text{if } \frac{\|y - X\beta\|}{\sqrt{n}} \leq \underline{\sigma} \\ \frac{\|y - X\beta\|}{\sqrt{n}} & \text{if } \frac{\|y - X\beta\|}{\sqrt{n}} > \underline{\sigma} \end{cases} \\
&= \min_{\sigma \geq \underline{\sigma}} \left(\frac{\|y - X\beta\|^2}{2n\sigma} + \frac{\sigma}{2} \right)
\end{aligned}
$$

Leads to the Smoothed Concomitant Lasso formulation

$$
\boxed{(\hat{\beta}^{(\lambda)}, \hat{\sigma}^{(\lambda)}) \in \underset{\beta \in \mathbb{R}^p, \sigma \geq \underline{\sigma}}{\arg\min} \left(\frac{\|y - X\beta\|^2}{2n\sigma} + \frac{\sigma}{2} + \lambda \|\beta\|_1 \right)}
$$

Solving the Smooth Concomitant Lasso

$$\boxed{(\hat{\beta}^{(\lambda)}, \hat{\sigma}^{(\lambda)}) \in \mathop{\arg\min}_{\beta \in \mathbb{R}^p, \sigma \geq \underline{\sigma}} \frac{\|y - X\beta\|^2}{2n\sigma} + \frac{\sigma}{2} + \lambda \|\beta\|_1}$$

Jointly convex formulation : can be optimized by alternating β and σ optimization (the other parameter being fixed)

Alternate:

- Fix σ: solve a Lasso problem to update β

$$\beta \in \mathop{\arg\min}_{\beta \in \mathbb{R}^p} \frac{\|y - X\beta\|^2}{2n} + \lambda\sigma \|\beta\|_1 \quad \text{(Lasso step)}$$

- Fix β: closed form solution to get σ

$$\sigma = \max\left(\frac{\|y - X\beta\|}{\sqrt{n}}, \underline{\sigma}\right) \quad \text{(Noise estimation step)}$$

Solving the Smooth Concomitant Lasso

$$\boxed{(\hat{\beta}^{(\lambda)}, \hat{\sigma}^{(\lambda)}) \in \mathop{\arg\min}_{\beta \in \mathbb{R}^p, \sigma \geq \underline{\sigma}} \frac{\|y - X\beta\|^2}{2n\sigma} + \frac{\sigma}{2} + \lambda \|\beta\|_1}$$

Jointly convex formulation : can be optimized by alternating β and σ optimization (the other parameter being fixed)

<u>Alternate</u>:

- Fix σ: solve a Lasso problem to update β

$$\beta \in \mathop{\arg\min}_{\beta \in \mathbb{R}^p} \frac{\|y - X\beta\|^2}{2n} + \lambda\sigma \|\beta\|_1 \quad \text{(Lasso step)}$$

- Fix β: closed form solution to get σ

$$\sigma = \max\left(\frac{\|y - X\beta\|}{\sqrt{n}}, \underline{\sigma}\right) \quad \text{(Noise estimation step)}$$

Table of Contents

Back to multi-task framework

General case: $Y \in \mathbb{R}^{n \times q}$, $\mathrm{B} \in \mathbb{R}^{p \times q}$, and the noise $E \in \mathbb{R}^{n \times q}$ might have some structure evolving along the n samples

Back to multi-task framework

General case: $Y \in \mathbb{R}^{n \times q}$, $\mathrm{B} \in \mathbb{R}^{p \times q}$, and the noise $E \in \mathbb{R}^{n \times q}$ might have some structure evolving along the n samples

Smoothed Generalized Concomitant Lasso (SGCL):

$$(\hat{\mathrm{B}}, \hat{\Sigma}) \in \underset{\substack{\mathrm{B} \in \mathbb{R}^{p \times q} \\ \Sigma \in \mathbb{S}^n_{++}, \Sigma \succeq \underline{\Sigma}}}{\arg\min} \frac{\|Y - X\mathrm{B}\|^2_{\Sigma^{-1}}}{2nq} + \frac{\mathrm{Tr}(\Sigma)}{2n} + \lambda \|\mathrm{B}\|_{2,1}$$

with $\|Z\|^2_A := \mathrm{Tr}(Z^\top A Z)$, and $\underline{\Sigma} := \underline{\sigma}\, \mathrm{Id}_n$ (for simplicity)

Back to multi-task framework

General case: $Y \in \mathbb{R}^{n \times q}$, $\mathrm{B} \in \mathbb{R}^{p \times q}$, and the noise $E \in \mathbb{R}^{n \times q}$ might have some structure evolving along the n samples

Smoothed Generalized Concomitant Lasso (SGCL):

$$(\hat{\mathrm{B}}, \hat{\Sigma}) \in \underset{\substack{\mathrm{B} \in \mathbb{R}^{p \times q} \\ \Sigma \in \mathbb{S}^n_{++}, \Sigma \succeq \underline{\Sigma}}}{\arg\min} \; \frac{\|Y - X\mathrm{B}\|^2_{\Sigma^{-1}}}{2nq} + \frac{\mathrm{Tr}(\Sigma)}{2n} + \lambda \|\mathrm{B}\|_{2,1}$$

with $\|Z\|^2_A := \mathrm{Tr}(Z^\top A Z)$, and $\underline{\Sigma} := \underline{\sigma}\, \mathrm{Id}_n$ (for simplicity)

- the formulation remains jointly convex

Back to multi-task framework

General case: $Y \in \mathbb{R}^{n \times q}$, $\mathrm{B} \in \mathbb{R}^{p \times q}$, and the noise $E \in \mathbb{R}^{n \times q}$ might have some structure evolving along the n samples

Smoothed Generalized Concomitant Lasso (SGCL):

$$(\hat{\mathrm{B}}, \hat{\Sigma}) \in \underset{\substack{\mathrm{B} \in \mathbb{R}^{p \times q} \\ \Sigma \in \mathbb{S}^n_{++}, \Sigma \succeq \underline{\Sigma}}}{\arg\min} \frac{\|Y - X\mathrm{B}\|^2_{\Sigma^{-1}}}{2nq} + \frac{\mathrm{Tr}(\Sigma)}{2n} + \lambda \|\mathrm{B}\|_{2,1}$$

with $\|Z\|^2_A := \mathrm{Tr}(Z^\top A Z)$, and $\underline{\Sigma} := \underline{\sigma}\, \mathrm{Id}_n$ (for simplicity)

- the formulation remains jointly convex
- the noise penalty is now on the sum of the eigenvalues of Σ

Back to multi-task framework

General case: $Y \in \mathbb{R}^{n \times q}$, $\mathrm{B} \in \mathbb{R}^{p \times q}$, and the noise $E \in \mathbb{R}^{n \times q}$ might have some structure evolving along the n samples

Smoothed Generalized Concomitant Lasso (SGCL):

$$(\hat{\mathrm{B}}, \hat{\Sigma}) \in \underset{\substack{\mathrm{B} \in \mathbb{R}^{p \times q} \\ \Sigma \in \mathbb{S}^n_{++}, \Sigma \succeq \underline{\Sigma}}}{\arg\min} \ \frac{\|Y - X\mathrm{B}\|^2_{\Sigma^{-1}}}{2nq} + \frac{\mathrm{Tr}(\Sigma)}{2n} + \lambda \|\mathrm{B}\|_{2,1}$$

with $\|Z\|^2_A := \mathrm{Tr}(Z^\top A Z)$, and $\underline{\Sigma} := \underline{\sigma}\, \mathrm{Id}_n$ (for simplicity)

- the formulation remains jointly convex
- the noise penalty is now on the sum of the eigenvalues of Σ
- adding the restriction $\Sigma = \sigma \, \mathrm{Id}_n$ recovers the Smoothed Concomitant Lasso

Solving the SGCL

Jointly convex formulation: alternate minimization still possible

Σ fixed: smooth + ℓ_1-type, Block Coordinate Descent (BCD) to update B row by row, *e.g.*, using safe screening rules Fercoq *et al.* (2015), Ndiaye *et al.* (2015)

Solving the SGCL

Jointly convex formulation: alternate minimization still possible

B fixed: with the current **residuals** $R = Y - X\mathrm{B}$, the problem can be reformulated

$$\hat{\Sigma} = \mathop{\arg\min}_{\Sigma \in \mathbb{S}^n_{++}, \Sigma \succeq \underline{\Sigma}} \left(\frac{1}{2nq} \operatorname{Tr}[R^\top \Sigma^{-1} R] + \frac{1}{2n} \operatorname{Tr}(\Sigma) \right)$$

Closed-form solution: if $U^\top \operatorname{diag}(s_1, \ldots, s_n) U$ is the spectral decomposition of $\frac{1}{q} R R^\top$:

$$\hat{\Sigma} = U^\top \operatorname{diag}(\max(\underline{\sigma}, \sqrt{s_1}), \ldots, \max(\underline{\sigma}, \sqrt{s_n})) U$$

Main drawbacks

- Statistically: $\mathcal{O}(n^2)$ parameters to infer for Σ, with only nq observations (works fine for q large w.r.t. n)
- Computationally: Σ update cost is $\mathcal{O}(n^3)$ (SVD computation); too slow in general ... Note: ok for MEG/EEG problems as $n \approx 300$

Table of Contents

Block Homoscedastic model

In the MEG/EEG case : 3 different types of signals are recorded

- electrodes measure the electric potentials
- magnetometers measure the magnetic field
- gradiometers measure the gradient of the magnetic field

$\neq$ physical natures $\Longrightarrow$ different noise levels

Observations are divided into 3 blocks $\&$ <u>the partition is known</u>

Block Homoscedastic model

K groups of observations (due to K sensors modalities)

$$X = \begin{pmatrix} X^1 \\ \vdots \\ X^K \end{pmatrix}, Y = \begin{pmatrix} Y^1 \\ \vdots \\ Y^K \end{pmatrix}, E = \begin{pmatrix} E^1 \\ \vdots \\ E^K \end{pmatrix}$$

$$\Sigma^* = \operatorname{diag}(\sigma_1^* \operatorname{Id}_{n_1}, \dots, \sigma_K^* \operatorname{Id}_{n_K})$$

For each block, homoscedastic model with white noise:

$$Y^k = X^k \mathrm{B}^* + \sigma_k^* E^k$$

and entries of E^k are i.i.d. $\mathcal{N}(0, 1)$

<u>Rem</u>: for MEG/EEG, $K = 3$ corresponding to physical signals:

1. EEG
2. MEG magnetometers
3. MEG gradiometers

Smoothed Block Homoscedastic Concomitant (SBHCL)

Reformulation with additional diagonal constraint on Σ, constant over consecutive blocks:

Block Homoscedastic Concomitant:

$$\underset{\substack{\mathrm{B} \in \mathbb{R}^{p \times q}, \\ \sigma_1, \ldots, \sigma_K \in \mathbb{R}_{++}^K \\ \sigma_k \geq \underline{\sigma}_k, \forall k \in [K]}}{\arg\min} \sum_{k=1}^{K} \left(\frac{\|Y^k - X^k \mathrm{B}\|^2}{2nq\sigma_k} + \frac{n_k \sigma_k}{2n} \right) + \lambda \|\mathrm{B}\|_{2,1}$$

Reduce number of parameters to estimate from $\frac{n(n-1)}{2}$ to K (hopeless otherwise without additional structure)

Solving the SBHCL

- Block Coordinate Descent (BCD) steps remain the same, as for the classical Multi-Task Lasso
- computing $\Sigma^{-1}(Y - X\mathrm{B})$ for the BCD is easier (inverting a diagonal matrix)

Solving the SBHCL

- Block Coordinate Descent (BCD) steps remain the same, as for the classical Multi-Task Lasso
- computing $\Sigma^{-1}(Y - X\mathrm{B})$ for the BCD is easier (inverting a diagonal matrix)
- σ_k's updates are simple and can even be performed at each B_j update (as for the concomitant)

Solving the SBHCL

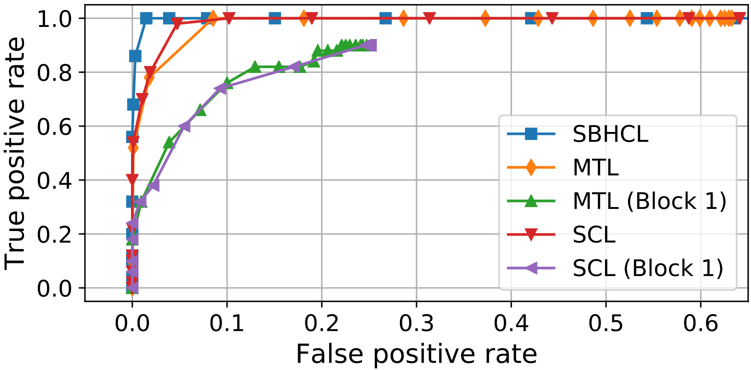
- Block Coordinate Descent (BCD) steps remain the same, as for the classical Multi-Task Lasso
- computing $\Sigma^{-1}(Y - X\mathrm{B})$ for the BCD is easier (inverting a diagonal matrix)
- σ_k's updates are simple and can even be performed at each B_j update (as for the concomitant)

In practice

Simulated block homoscedastic design (similar to real data):

- $(n, p, q) = 300, 1000, 100$
- X Toeplitz-correlated: $Cov(X_i, X_j) = \rho^{|i-j|}$, $\rho \in]0, 1[$
- 3 blocks with standard deviation in ratio 1, 2, 5

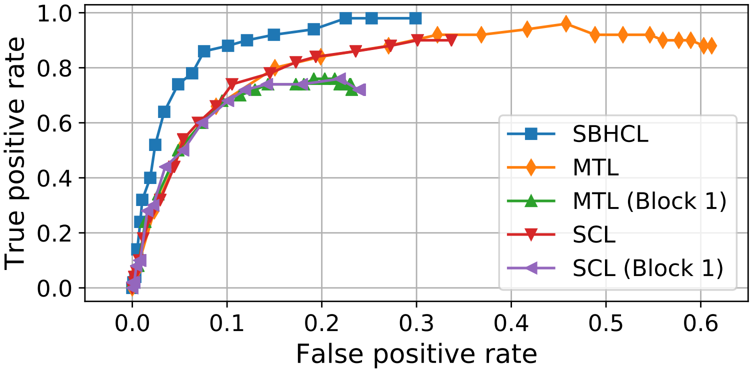
Support recovery: ROC curve w.r.t. λ



SBHCL, MTL (Multi-Task Lasso) and SCL (single noise level) on all blocks, and the MTL and SCL on the least noisy block

$\rho = 0.1$ (low correlation, easy case)

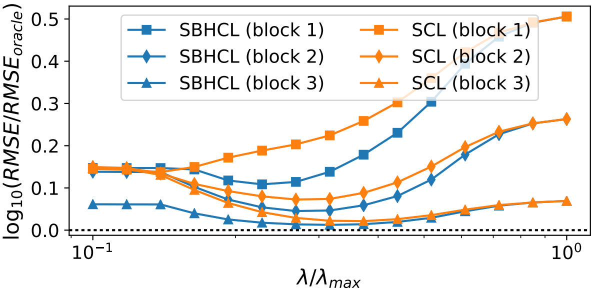
Support recovery: ROC curve w.r.t. λ



SBHCL, MTL (Multi-Task Lasso) and SCL (single noise level) on all blocks, and the MTL and SCL on the least noisy block

$\rho = 0.9$ (high correlation, hard case)

Prediction performance



RMSE (Root Mean Square Error) normalized by oracle RMSE, per block, for the multi-task SBHCL and SCL on testing set, for various values of λ.

Take home message

- more general noise models: possible to estimate full covariance if there are enough tasks
- if more noise structure is known (*e.g.*, block homoscedastic model): not more costly than the Multi-Task Lasso (MTL)

Take home message

- more general noise models: possible to estimate full covariance if there are enough tasks
- if more noise structure is known (*e.g.*, block homoscedastic model): not more costly than the Multi-Task Lasso (MTL)
- taking into account multiple noise levels helps: both for prediction and support identification

Take home message

- more general noise models: possible to estimate full covariance if there are enough tasks
- if more noise structure is known (*e.g.*, block homoscedastic model): not more costly than the Multi-Task Lasso (MTL)
- taking into account multiple noise levels helps: both for prediction and support identification
- using additional (though noisier) data helps!

Take home message

- more general noise models: possible to estimate full covariance if there are enough tasks
- if more noise structure is known (*e.g.*, block homoscedastic model): not more costly than the Multi-Task Lasso (MTL)
- taking into account multiple noise levels helps: both for prediction and support identification
- using additional (though noisier) data helps!
- future work: using non-convex penalties

Take home message

- more general noise models: possible to estimate full covariance if there are enough tasks
- if more noise structure is known (*e.g.*, block homoscedastic model): not more costly than the Multi-Task Lasso (MTL)
- taking into account multiple noise levels helps: both for prediction and support identification
- using additional (though noisier) data helps!
- future work: using non-convex penalties

Python code is available at https://github.com/mathurinm/SHCL

Massias *et al.* (2018): to appear in AISTATS 2018

This work was funded by ERC Starting Grant SLAB ERC-YStG-676943

Take home message

- more general noise models: possible to estimate full covariance if there are enough tasks
- if more noise structure is known (*e.g.*, block homoscedastic model): not more costly than the Multi-Task Lasso (MTL)
- taking into account multiple noise levels helps: both for prediction and support identification
- using additional (though noisier) data helps!
- future work: using non-convex penalties

Python code is available at https://github.com/mathurinm/SHCL

Massias *et al.* (2018): to appear in AISTATS 2018

This work was funded by ERC Starting Grant SLAB ERC-YStG-676943

References I

- B. K. Natarajan.
  Sparse approximate solutions to linear systems.
  SIAM J. Comput., 24(2):227–234, 1995.
- E. J. Candès, M. B. Wakin, and S. P. Boyd.
  Enhancing sparsity by reweighted l_1 minimization.
  J. Fourier Anal. Applicat., 14(5-6):877–905, 2008.
- R. Tibshirani.
  Regression shrinkage and selection via the lasso.
  J. R. Stat. Soc. Ser. B Stat. Methodol., 58(1):267–288, 1996.
- S. S. Chen, D. L. Donoho, and M. A. Saunders.
  Atomic decomposition by basis pursuit.
  SIAM J. Sci. Comput., 20(1):33–61, 1998.
- G. Obozinski, B. Taskar, and M. I. Jordan.
  Joint covariate selection and joint subspace selection for multiple classification problems.
  Statistics and Computing, 20(2):231–252, 2010.

References II

- P. J. Bickel, Y. Ritov, and A. B. Tsybakov.
  Simultaneous analysis of Lasso and Dantzig selector.
  Ann. Statist., 37(4):1705–1732, 2009.
- A. S. Dalalyan, M. Hebiri, and J. Lederer.
  On the prediction performance of the Lasso.
  Bernoulli, 23(1):552–581, 2017.
- T. Sun and C.-H. Zhang.
  Scaled sparse linear regression.
  Biometrika, 99(4):879–898, 2012.
- P. J. Huber.
  Robust Statistics.
  John Wiley & Sons Inc., 1981.
- A. B. Owen.
  A robust hybrid of lasso and ridge regression.
  Contemporary Mathematics, 443:59–72, 2007.

References III

- A. Belloni, V. Chernozhukov, and L. Wang.
  Square-root Lasso: pivotal recovery of sparse signals via conic programming.
  Biometrika, 98(4):791–806, 2011.
- Y. Nesterov.
  Smooth minimization of non-smooth functions.
  Math. Program., 103(1):127–152, 2005.
- E. Ndiaye, O. Fercoq, A. Gramfort, V. Leclère, and J. Salmon.
  Efficient smoothed concomitant Lasso estimation for high dimensional regression.
  In *NCMIP*, 2017.
- A. Beck and M. Teboulle.
  Smoothing and first order methods: A unified framework.
  SIAM J. Optim., 22(2):557–580, 2012.
- O. Fercoq, A. Gramfort, and J. Salmon.
  Mind the duality gap: safer rules for the lasso.
  In *ICML*, pages 333–342, 2015.

References IV

- E. Ndiaye, O. Fercoq, A. Gramfort, and J. Salmon.
  Gap safe screening rules for sparse multi-task and multi-class models.
  In *NIPS*, pages 811–819, 2015.
- M. Massias, O. Fercoq, A. Gramfort, and J. Salmon.
  Heteroscedastic concomitant lasso for sparse multimodal electromagnetic brain imaging.
  Technical report, 2017.
  URL https://arxiv.org/pdf/1705.09778.pdf.