Generalized Concomitant Multi-Task Lasso for sparse multimodal regression

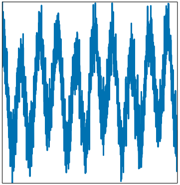
Joseph Salmon
http://josephsalmon.eu
IMAG, Univ Montpellier, CNRS
Montpellier, France

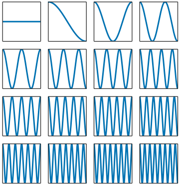
Joint work with:
Mathurin Massias (INRIA, Parietal Team)
Olivier Fercoq (Télécom ParisTech)
Alexandre Gramfort (INRIA, Parietal Team)

Sparsity is all around

Signals can often be represented combining few **atoms** / **features** :

- Fourier decomposition for sounds

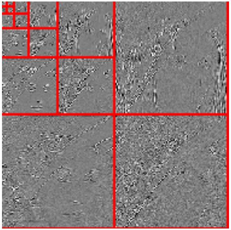




Sparsity is all around

Signals can often be represented combining few **atoms** / **features** :

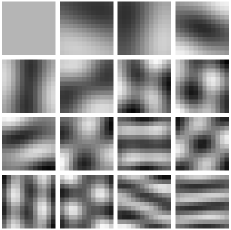
- Fourier decomposition for sounds
- Wavelet for images (1990's)



Sparsity is all around

Signals can often be represented combining 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 combining 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

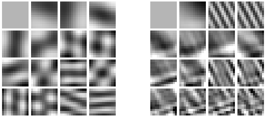
$X = [\mathbf{x}_1, \dots, \mathbf{x}_p] \in \mathbb{R}^{n \times p}$:
dictionary of atoms/features

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

<u>Objective(s)</u>: find $\hat{\beta}$

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

<u>Constraints</u>: large p, sparse β^*



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

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

The ℓ_0 penalty to enforce sparsity

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

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

Combinatorial problem: "NP-hard"[(1)]

$\hookrightarrow$ 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
- $p = 30 \hookrightarrow$ hard: $\approx 10^{10}$ least squares

Rem: mixed integer programming (MIP) fine for small problems[(2)]

(1) B. K. Natarajan. "Sparse approximate solutions to linear systems". In: *SIAM J. Comput.* 24.2 (1995), pp. 227–234.

(2) D. Bertsimas, A. King, and R. Mazumder. "Best subset selection via a modern optimization lens". In: *Ann. Statist.* 44.2 (2016), pp. 813–852.

The ℓ_1 penalty: Lasso and variants

Vocabulary: the "Modern least square"[3]

- Statistics: **Lasso**[4]
- Signal processing variant: **Basis Pursuit**[5]

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

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

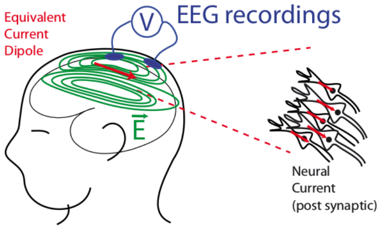
(3) E. J. Candès, M. B. Wakin, and S. P. Boyd. "Enhancing Sparsity by Reweighted l_1 Minimization". In: *J. Fourier Anal. Applicat.* 14.5-6 (2008), pp. 877–905.

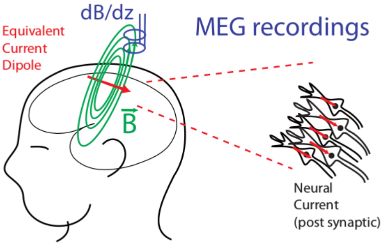
(4) R. Tibshirani. "Regression Shrinkage and Selection via the Lasso". In: *J. R. Stat. Soc. Ser. B Stat. Methodol.* 58.1 (1996), pp. 267–288.

(5) S. S. Chen, D. L. Donoho, and M. A. Saunders. "Atomic decomposition by basis pursuit". In: *SIAM J. Sci. Comput.* 20.1 (1998), pp. 33–61.

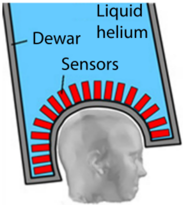
M/EEG inverse problem for brain imaging

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





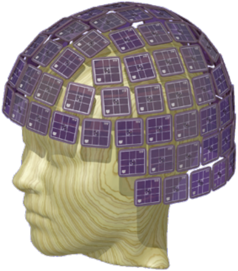
First EEG
recordings
in 1929
by H. Berger



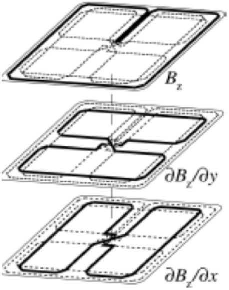
Hôpital La Timone
Marseille, France

MEG elements: magnometers and gradiometers

Device

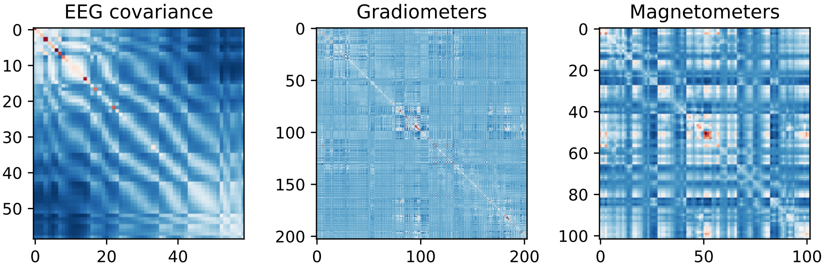


Sensors



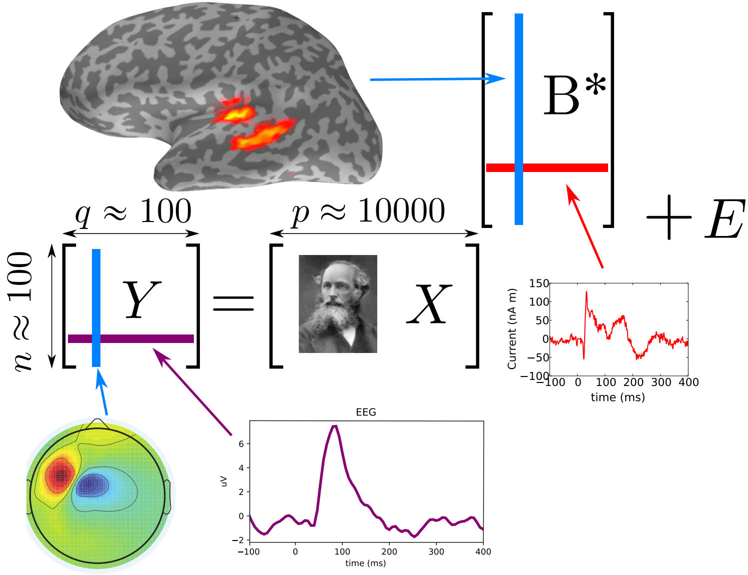
Detail of a sensor

Noise is different for EEG / MEG (magnometers and gradiometers)



- Different sensors $\Longrightarrow$ different noise structures

The M/EEG inverse problem: modeling



A multi-task framework

Multi-task regression:

- n observations (*e.g.*, number of sensors)
- q tasks (*e.g.*, 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

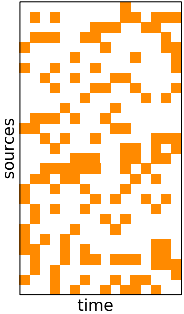
- $\mathrm{B}^* \in \mathbb{R}^{p \times q}$: true source activity matrix
- $E \in \mathbb{R}^{n \times q}$: additive white Gaussian noise (for simplicity)

Notation remark: capital letters refer to matrices

Multi-tasks penalties[6]

Popular convex penalties considered:

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



Sparse support: no structure

Penalty: **Lasso type**

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

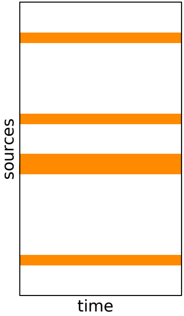
Parameter $\hat{\mathrm{B}} \in \mathbb{R}^{p \times q}$

[6] G. Obozinski, B. Taskar, and M. I. Jordan. "Joint covariate selection and joint subspace selection for multiple classification problems". In: *Statistics and Computing* 20.2 (2010), pp. 231–252.

Multi-tasks penalties[6]

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 type**

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

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

[6] G. Obozinski, B. Taskar, and M. I. Jordan. "Joint covariate selection and joint subspace selection for multiple classification problems". In: *Statistics and Computing* 20.2 (2010), pp. 231–252.

Table of Contents

Step back on the Lasso case ($q = 1$)

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 :

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

Lasso theory[7],[8] : (fairly) well understood

Theorem

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

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

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: $\kappa_{s^*}^2$ controls the conditioning of X when extracting the s^*columns of X associated to the true support

BUT σ_* is unknown in practice !

[7] P. J. Bickel, Y. Ritov, and A. B. Tsybakov. "Simultaneous analysis of Lasso and Dantzig selector". In: *Ann. Statist.* 37.4 (2009), pp. 1705–1732.

[8] A. S. Dalalyan, M. Hebiri, and J. Lederer. "On the Prediction Performance of the Lasso". In: *Bernoulli* 23.1 (2017), pp. 552–581.

Lasso theory[7],[8] : (fairly) well understood

Theorem

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

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

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: $\kappa_{s^*}^2$ controls the conditioning of X when extracting the s^*columns of X associated to the true support

BUT σ_* is unknown in practice !

[7] P. J. Bickel, Y. Ritov, and A. B. Tsybakov. "Simultaneous analysis of Lasso and Dantzig selector". In: *Ann. Statist.* 37.4 (2009), pp. 1705–1732.

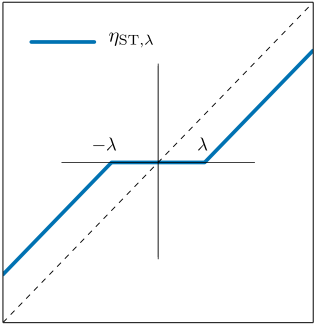
[8] A. S. Dalalyan, M. Hebiri, and J. Lederer. "On the Prediction Performance of the Lasso". In: *Bernoulli* 23.1 (2017), pp. 552–581.

Soft-Thresholding: Lasso for orthogonal design

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

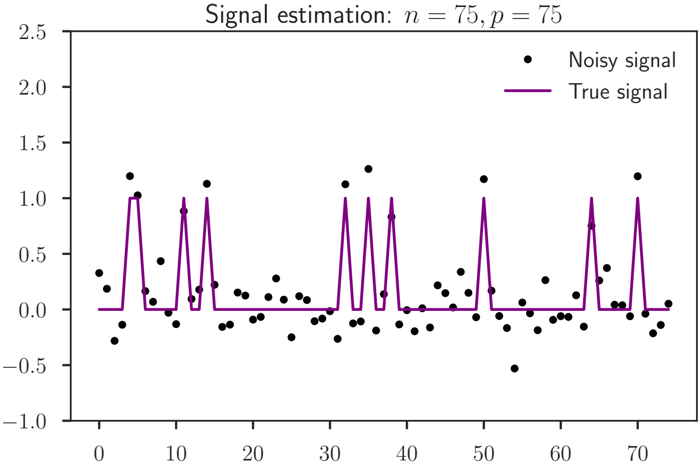
$$\eta_{\mathrm{ST},\lambda}(y) := \underset{\beta \in \mathbb{R}}{\arg\min} \left(\frac{(y-\beta)^2}{2} + \lambda|\beta| \right)$$
$$= \mathrm{sign}(y)(|y| - \lambda)_+$$

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



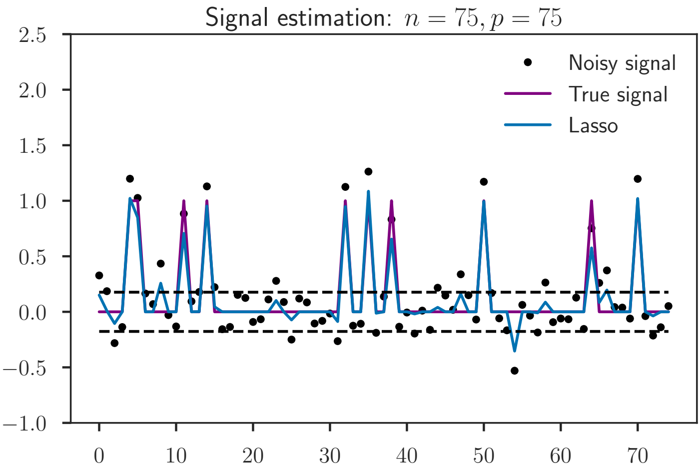
Extension for $X = \mathrm{Id}_p$: component-wise soft thresholding

“Universal”[9] λ choice ($X = \mathrm{Id}_n$)



[9] D. L. Donoho and I. M. Johnstone. “Adapting to unknown smoothness via wavelet shrinkage”. In: *J. Amer. Statist. Assoc.* 90.432 (1995), pp. 1200–1224.

"Universal"[9] λ choice ($X = \mathrm{Id}_n$)



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

[9] D. L. Donoho and I. M. Johnstone. "Adapting to unknown smoothness via wavelet shrinkage". In: *J. Amer. Statist. Assoc.* 90.432 (1995), pp. 1200–1224.

Joint estimation of β and σ

How to calibrate (theoretically) λ when σ_* is unknown?

Intuitive idea: initialize λ

- run Lasso with λ; get β
- estimate σ with residuals: $\sigma = \|y - X\beta\| / \sqrt{n}$
- re-run Lasso with $\lambda \propto \sigma$
- iterate (until convergence?)

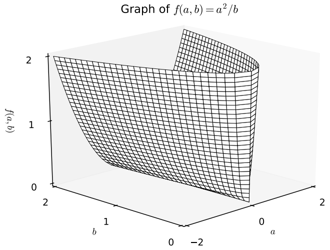
N.B.: exactly the Scaled-Lasso[10] implementation

[10] T. Sun and C.-H. Zhang. "Scaled sparse linear regression". In: *Biometrika* 99.4 (2012), pp. 879–898.

Concomitant Lasso[12]

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

- $\frac{\sigma}{2}$: penalty over the noise level, roots in robust estimation[11]
- jointly convex program: $(a, b) \mapsto a^2/b$ is convex



[11] P. J. Huber. *Robust Statistics*. John Wiley & Sons Inc., 1981.

[12] A. B. Owen. "A robust hybrid of lasso and ridge regression". In: *Contemporary Mathematics* 443 (2007), pp. 59–72.

Concomitant performance

Theorem [13]

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

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

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

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

Rem: theoretically important, though λ still has to be calibrated...

[13] T. Sun and C.-H. Zhang. "Scaled sparse linear regression". In: *Biometrika* 99.4 (2012), pp. 879–898.

Link with $\sqrt{\text{Lasso}}$[14]

- Independently, $\sqrt{\text{Lasso}}$ analyzed to get "σ free" choice of λ

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

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

$$\hat{\sigma}^{(\lambda)}_{\sqrt{\text{Lasso}}} = \frac{\left\| y - X\hat{\beta}^{(\lambda)}_{\sqrt{\text{Lasso}}} \right\|}{\sqrt{n}} \neq 0$$

Rem: non-smooth data fitting term with non-smooth regularization

[14] A. Belloni, V. Chernozhukov, and L. Wang. "Square-root Lasso: pivotal recovery of sparse signals via conic programming". In: *Biometrika* 98.4 (2011), pp. 791–806.

The Smoothed Concomitant Lasso[16]

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

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

- With prior information on the minimal noise level, one can set $\underline{\sigma}$ as this bound (recovers Concomitant Lasso)
- Setting $\underline{\sigma} = \epsilon$, smoothing theory asserts that $\frac{\epsilon}{2}$-solutions for the smoothed problem provide ϵ-solutions for the $\sqrt{\text{Lasso}}$[15]

[15] Y. Nesterov. "Smooth minimization of non-smooth functions". In: *Math. Program.* 103.1 (2005), pp. 127–152.

[16] E. Ndiaye et al. "Efficient Smoothed Concomitant Lasso Estimation for High Dimensional Regression". In: *NCMIP*. 2017.

Smoothing aparté[17],[18]

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

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

$$f_\mu = \mu\omega\left(\frac{\cdot}{\mu}\right)\Box f, \quad \text{where} \quad f\Box g(x) = \inf_u \{f(u) + g(x-u)\}$$

- ω is a predefined smooth function (s.t. $\nabla\omega$ is Lipschitz)

	Fourier: $\mathcal{F}(f)$	Fenchel/Legendre: f^*
	convolution: $\star$	**inf-convolution**: $\Box$
Kernel smoothing analogy:	$\mathcal{F}(f \star g) = \mathcal{F}(f) \cdot \mathcal{F}(g)$	$(f\Box g)^* = f^* + g^*$
	Gaussian : $\mathcal{F}(g) = g$	$\omega = \frac{\|\cdot\|^2}{2}$: $\quad \omega^* = \omega$
	$f_h = \frac{1}{h} g\left(\frac{\cdot}{h}\right) \star f$	$f_\mu = \mu\omega\left(\frac{\cdot}{\mu}\right)\Box f$

[17] Y. Nesterov. "Smooth minimization of non-smooth functions". In: *Math. Program.* 103.1 (2005), pp. 127–152.

[18] A. Beck and M. Teboulle. "Smoothing and first order methods: A unified framework". In: *SIAM J. Optim.* 22.2 (2012), pp. 557–580.

Smoothing aparté[17],[18]

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

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

$$f_\mu = \mu\omega\left(\frac{\cdot}{\mu}\right)\Box f, \quad \text{where} \quad f\Box g(x) = \inf_u \{f(u) + g(x-u)\}$$

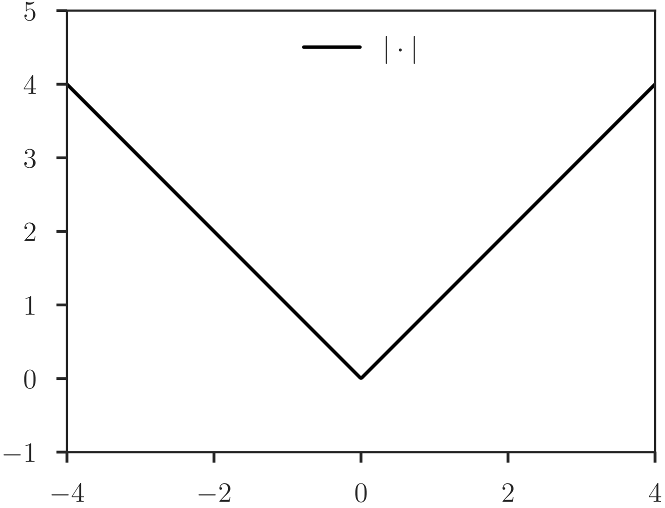
- ω is a predefined smooth function (s.t. $\nabla\omega$ is Lipschitz)

	Fourier: $\mathcal{F}(f)$	Fenchel/Legendre: f^*
	convolution: $\star$	**inf-convolution**: $\Box$
Kernel smoothing analogy:	$\mathcal{F}(f \star g) = \mathcal{F}(f) \cdot \mathcal{F}(g)$	$(f\Box g)^* = f^* + g^*$
	Gaussian : $\mathcal{F}(g) = g$	$\omega = \frac{\lVert\cdot\rVert^2}{2}$: $\quad \omega^* = \omega$
	$f_h = \frac{1}{h} g\left(\frac{\cdot}{h}\right) \star f$	$f_\mu = \mu\omega\left(\frac{\cdot}{\mu}\right)\Box f$

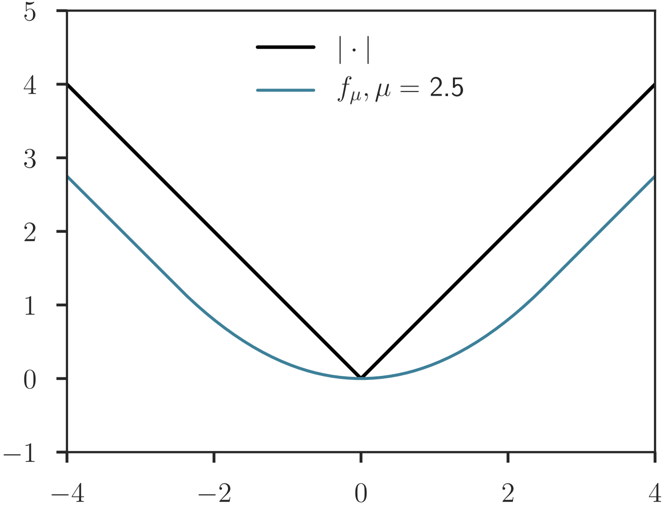
[17] Y. Nesterov. "Smooth minimization of non-smooth functions". In: *Math. Program.* 103.1 (2005), pp. 127–152.

[18] A. Beck and M. Teboulle. "Smoothing and first order methods: A unified framework". In: *SIAM J. Optim.* 22.2 (2012), pp. 557–580.

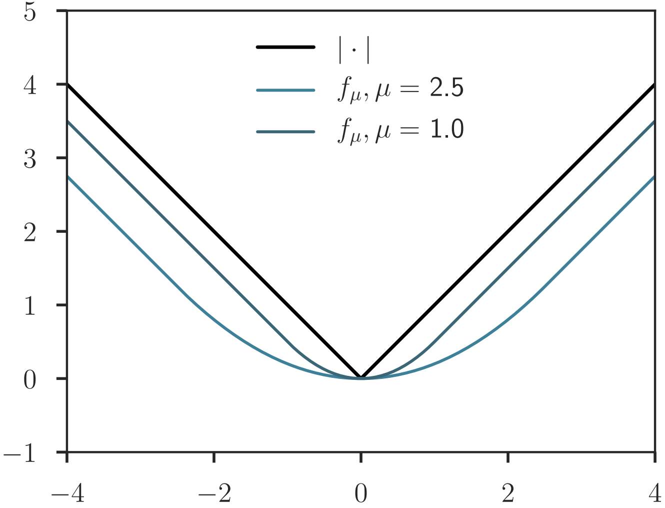
Huber function: $\omega(t) = \frac{t^2}{2}$



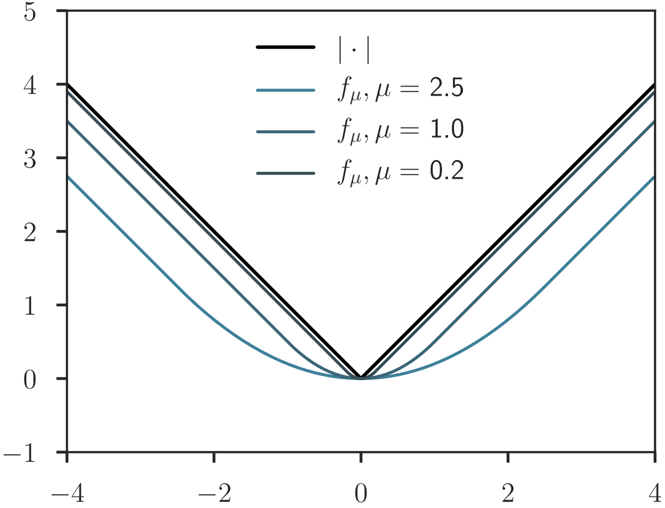
Huber function: $\omega(t) = \frac{t^2}{2}$



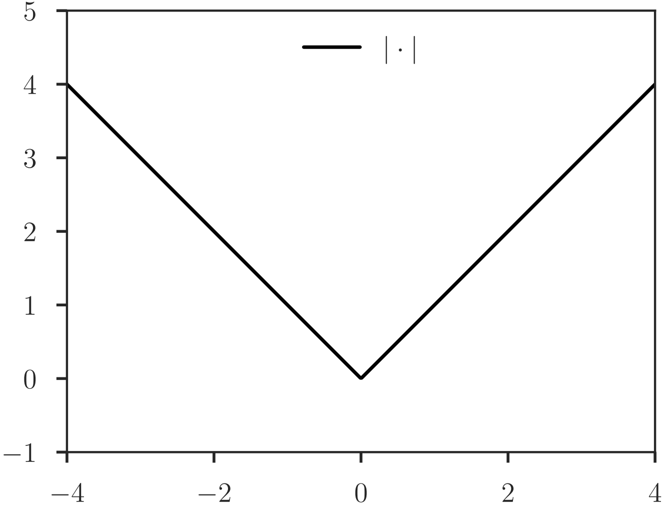
Huber function: $\omega(t) = \frac{t^2}{2}$



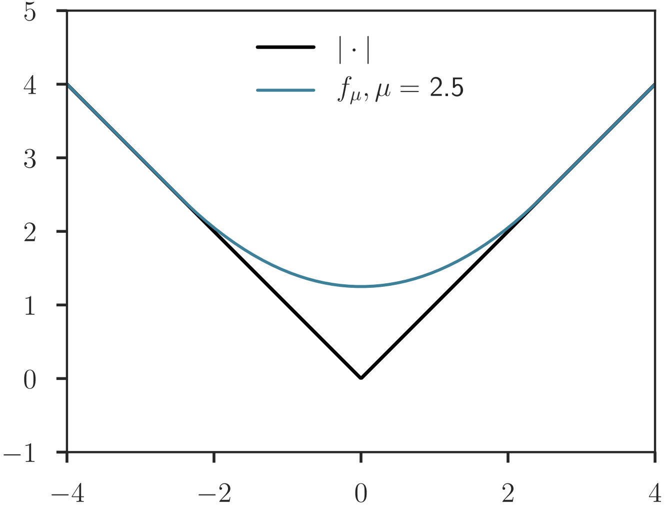
Huber function: $\omega(t) = \frac{t^2}{2}$



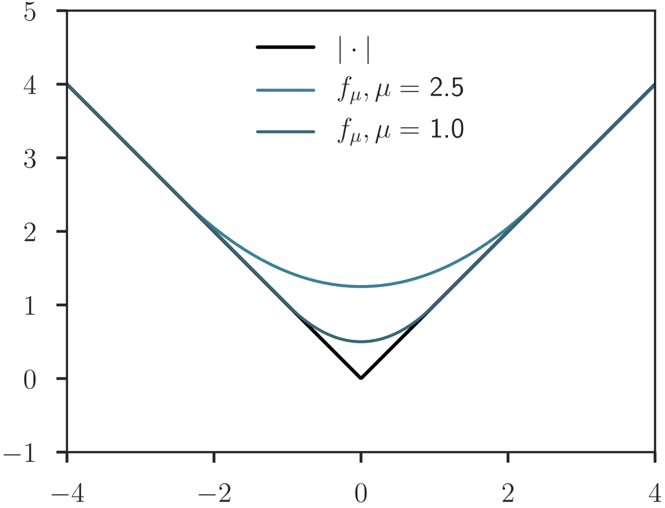
Huber function (bis): $\omega(t) = \frac{t^2}{2} + \frac{1}{2}$



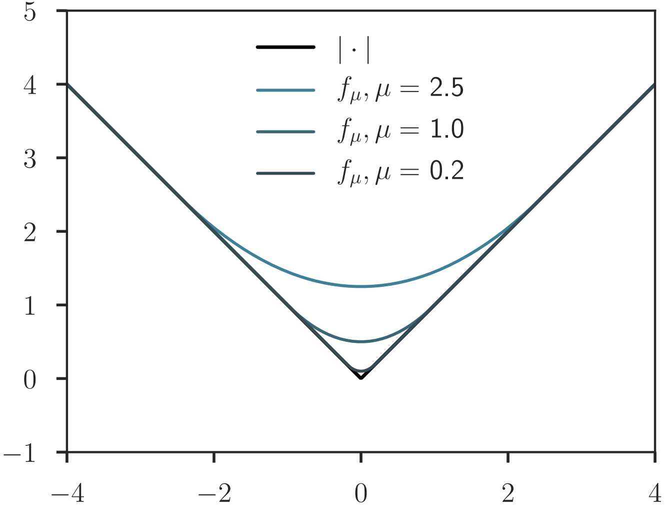
Huber function (bis): $\omega(t) = \frac{t^2}{2} + \frac{1}{2}$



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

$$f_{\underline{\sigma}}(z) = \begin{cases} \frac{\|z\|^2}{2n\underline{\sigma}} + \frac{\underline{\sigma}}{2}, & \text{if } \frac{\|z\|}{\sqrt{n}} \leq \underline{\sigma} \\ \frac{\|z\|}{\sqrt{n}}, & \text{if } \frac{\|z\|}{\sqrt{n}} > \underline{\sigma} \end{cases}$$

$$= \min_{\sigma \geq \underline{\sigma}} \left(\frac{\|z\|^2}{2n\sigma} + \frac{\sigma}{2} \right)$$

Leads to the Smoothed Concomitant Lasso formulation:

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

Solving the Smooth Concomitant Lasso

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

Jointly convex formulation : can be optimized by alternate minimization w.r.t. β and σ (the other parameter being fixed)

Alternate iteratively:

- Fix σ: (approximatively) solve a Lasso problem in β

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

Solving the Smooth Concomitant Lasso

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

Jointly convex formulation : can be optimized by alternate minimization w.r.t. β and σ (the other parameter being fixed)

Alternate iteratively:

- Fix σ: (approximatively) solve a Lasso problem in β

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

- Fix β: closed form solution to update σ

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

Solving the Smooth Concomitant Lasso

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

Jointly convex formulation : can be optimized by alternate minimization w.r.t. β and σ (the other parameter being fixed)

Alternate iteratively:

- Fix σ: (approximatively) solve a Lasso problem in β

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

- Fix β: closed form solution to update σ

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

Table of Contents

Back to multi-task : $Y = X\mathrm{B}^* + E$

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 (sensors)

Back to multi-task : $Y = X\mathrm{B}^* + E$

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 (sensors)

Smoothed Generalized Concomitant Lasso (SGCL):

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

with $\|R\|^2_{\Sigma^{-1}} := \mathrm{Tr}(R^\top \Sigma^{-1} R)$, and $\underline{\Sigma} := \underline{\sigma}\, \mathrm{Id}_n$ (for simplicity)

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

Beware: Σ not a covariance, more a generalized standard deviation

Solving the SGCL

Jointly convex formulation: alternate minimization still converging

B **Update - Σ fixed**:

smooth + ℓ_1-type optimization problem, *e.g.*, use Block Coordinate Descent (BCD) to update B row by row

Possible refinements:

- (Gap safe) screening rules[19],[20]
- Stong rules[21]
- Active sets methods[22] etc.

---

[19] L. El Ghaoui, V. Viallon, and T. Rabbani. “Safe feature elimination in sparse supervised learning”. In: *J. Pacific Optim.* 8.4 (2012), pp. 667–698.

[20] E. Ndiaye et al. “Gap Safe screening rules for sparsity enforcing penalties”. In: *J. Mach. Learn. Res.* 18.128 (2017), pp. 1–33.

[21] R. Tibshirani et al. “Strong rules for discarding predictors in lasso-type problems”. In: *J. R. Stat. Soc. Ser. B Stat. Methodol.* 74.2 (2012), pp. 245–266.

[22] T. B. Johnson and C. Guestrin. “BLITZ: A Principled Meta-Algorithm for Scaling Sparse Optimization”. In: *ICML*. 2015, pp. 1171–1179.

Solving the SGCL

Jointly convex formulation: alternate minimization still converging

Σ **Update** - B **fixed**:

with $R = Y - X\mathrm{B}$ (**residuals**), 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 (**Spectral clipping**):

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

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

Main drawbacks

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

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

Key point: observations divided into 3 blocks (known partition)

Block Homoscedastic model

K groups of observations (K sensors modalities)

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

$$\Sigma^* = \mathrm{diag}(\sigma_1^* \,\mathrm{Id}_{n_1}, \ldots, \sigma_K^* \,\mathrm{Id}_{n_K}) \text{ where } n = n_1 + \cdots + n_K$$

For each block, the entries $E_{i,j}^k \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$ (homoscedastic):

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

MEG/EEG case: $K = 3$ corresponding to 3 physical signals
1) EEG, 2) MEG magnetometers, 3) MEG gradiometers

Smoothed Block Homoscedastic Concomitant (SBHCL)

Additional constraints: Σ piecewise constant **diagonal**, *i.e.*,

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

Block Homoscedastic Concomitant:

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

Benefit: number of parameters reduced $\frac{n(n+1)}{2} \to K$

Solving the SBHCL

- B **update**: (approximately) solve a Multi-Task Lasso problem *e.g.*, by Block Coordinate Descent (BCD) over rows, etc.

- Σ **update**: simply update the σ_k's, potentially at each row B_j update (cheap : residuals are stored!)

Table of Contents

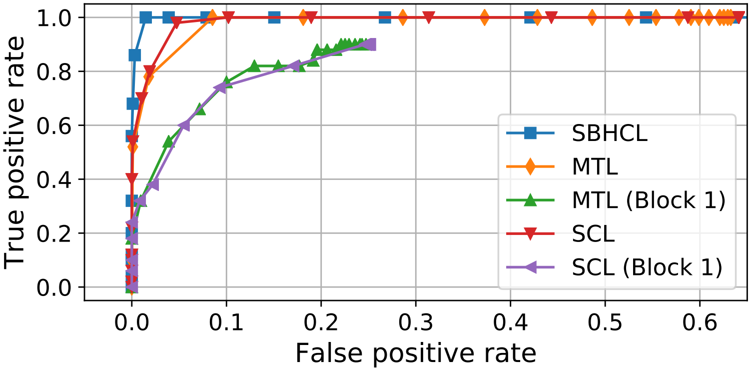
Simulated scenario

Simulated block homoscedastic design:

- $n = 300$, with equals block sizes $n_1 = n_2 = n_3 = 100$
- $p = 1000$
- $q = 100$
- X Toeplitz-correlated: $\mathrm{Cov}(X_i, X_j) = \rho^{|i-j|}$, $\rho \in]0, 1[$
- 3 blocks with standard deviation in ratio 1, 2, 5

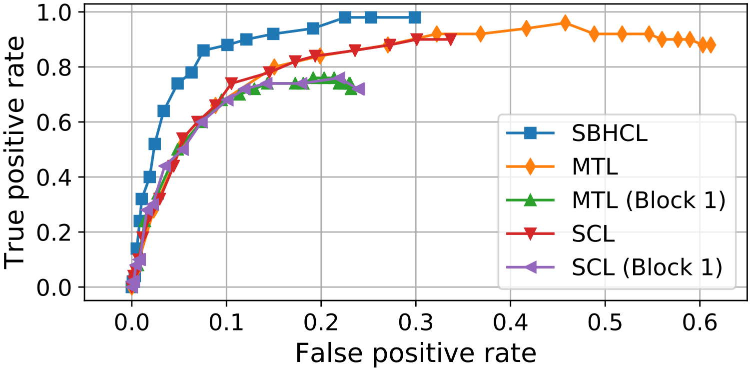
Rem: Block 1 has smallest standard deviation

Support recovery: ROC curve w.r.t. λ, $\rho = 0.1$



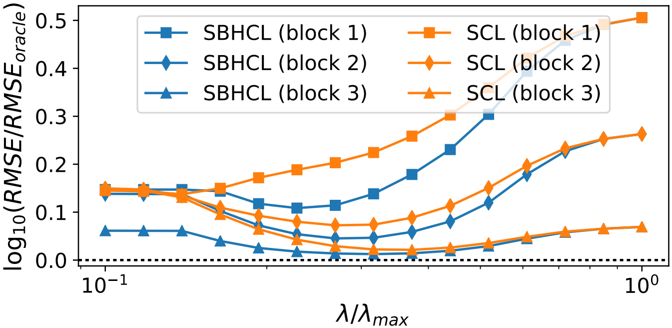
SBHCL:	Smoothed Block Homoscedastic Concomitant
MTL:	Multi-Task Lasso
SCL:	Smooth Concomitant Lasso (single σ for all blocks)
MTL (Block 1):	MTL on least noisy block
SCL (Block 1):	SCL on least noisy block

Support recovery: ROC curve w.r.t. λ, $\rho = 0.9$



SBHCL:	Smoothed Block Homoscedastic Concomitant
MTL:	Multi-Task Lasso
SCL:	Smooth Concomitant Lasso (single σ for all blocks)
MTL (Block 1):	MTL on least noisy block
SCL (Block 1):	SCL on least noisy block

Prediction error: RMSE curve w.r.t. λ, $\rho = 0.7$



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

Conclusion: align best λ's for all modalities

Conclusion and perspectives

- New insights for handling (structured) noise in multi-task
- Cost equivalent to Multi-Task Lasso for“simple” noise structure (*e.g.*, block homoscedastic)

Conclusion and perspectives

- New insights for handling (structured) noise in multi-task
- Cost equivalent to Multi-Task Lasso for"simple" noise structure (*e.g.*, block homoscedastic)
- Handling multiple noise levels: helps both for prediction and support identification

Conclusion and perspectives

- New insights for handling (structured) noise in multi-task
- Cost equivalent to Multi-Task Lasso for"simple" noise structure (*e.g.*, block homoscedastic)
- Handling multiple noise levels: helps both for prediction and support identification
- Future work: non-convex penalties, repetitive neuro task, etc.

Conclusion and perspectives

- New insights for handling (structured) noise in multi-task
- Cost equivalent to Multi-Task Lasso for“simple” noise structure (*e.g.*, block homoscedastic)
- Handling multiple noise levels: helps both for prediction and support identification
- Future work: non-convex penalties, repetitive neuro task, etc.

Merci!

"*All models are wrong but some come with good open source implementation and good documentation so use those.*"

A. Gramfort

- Paper online: arXiv, personnal webpages, AISTATS[19]
- Python code online: https://github.com/mathurinm/SHCL

Powered with **MooseTeX**

[19] M. Massias et al. "Generalized Concomitant Multi-Task Lasso for Sparse Multimodal Regression". In: *AISTATS*. vol. 84. 2018, pp. 998–1007.

References I

- Beck, A. and M. Teboulle. "Smoothing and first order methods: A unified framework". In: *SIAM J. Optim.* 22.2 (2012), pp. 557–580.
- Belloni, A., V. Chernozhukov, and L. Wang. "Square-root Lasso: pivotal recovery of sparse signals via conic programming". In: *Biometrika* 98.4 (2011), pp. 791–806.
- Bertsimas, D., A. King, and R. Mazumder. "Best subset selection via a modern optimization lens". In: *Ann. Statist.* 44.2 (2016), pp. 813–852.
- Bickel, P. J., Y. Ritov, and A. B. Tsybakov. "Simultaneous analysis of Lasso and Dantzig selector". In: *Ann. Statist.* 37.4 (2009), pp. 1705–1732.
- Candès, E. J., M. B. Wakin, and S. P. Boyd. "Enhancing Sparsity by Reweighted l_1 Minimization". In: *J. Fourier Anal. Applicat.* 14.5-6 (2008), pp. 877–905.

References II

- Chen, S. S., D. L. Donoho, and M. A. Saunders. "Atomic decomposition by basis pursuit". In: *SIAM J. Sci. Comput.* 20.1 (1998), pp. 33–61.
- Dalalyan, A. S., M. Hebiri, and J. Lederer. "On the Prediction Performance of the Lasso". In: *Bernoulli* 23.1 (2017), pp. 552–581.
- Donoho, D. L. and I. M. Johnstone. "Adapting to unknown smoothness via wavelet shrinkage". In: *J. Amer. Statist. Assoc.* 90.432 (1995), pp. 1200–1224.
- El Ghaoui, L., V. Viallon, and T. Rabbani. "Safe feature elimination in sparse supervised learning". In: *J. Pacific Optim.* 8.4 (2012), pp. 667–698.
- Huber, P. J. *Robust Statistics*. John Wiley & Sons Inc., 1981.
- Johnson, T. B. and C. Guestrin. "BLITZ: A Principled Meta-Algorithm for Scaling Sparse Optimization". In: *ICML*. 2015, pp. 1171–1179.

References III

- Massias, M. et al. "Generalized Concomitant Multi-Task Lasso for Sparse Multimodal Regression". In: *AISTATS*. Vol. 84. 2018, pp. 998–1007.
- Natarajan, B. K. "Sparse approximate solutions to linear systems". In: *SIAM J. Comput.* 24.2 (1995), pp. 227–234.
- Ndiaye, E. et al. "Efficient Smoothed Concomitant Lasso Estimation for High Dimensional Regression". In: *NCMIP*. 2017.
- Ndiaye, E. et al. "Gap Safe screening rules for sparsity enforcing penalties". In: *J. Mach. Learn. Res.* 18.128 (2017), pp. 1–33.
- Nesterov, Y. "Smooth minimization of non-smooth functions". In: *Math. Program.* 103.1 (2005), pp. 127–152.
- Obozinski, G., B. Taskar, and M. I. Jordan. "Joint covariate selection and joint subspace selection for multiple classification problems". In: *Statistics and Computing* 20.2 (2010), pp. 231–252.

References IV

- Owen, A. B. "A robust hybrid of lasso and ridge regression". In: *Contemporary Mathematics* 443 (2007), pp. 59–72.
- Sun, T. and C.-H. Zhang. "Scaled sparse linear regression". In: *Biometrika* 99.4 (2012), pp. 879–898.
- Tibshirani, R. "Regression Shrinkage and Selection via the Lasso". In: *J. R. Stat. Soc. Ser. B Stat. Methodol.* 58.1 (1996), pp. 267–288.
- Tibshirani, R. et al. "Strong rules for discarding predictors in lasso-type problems". In: *J. R. Stat. Soc. Ser. B Stat. Methodol.* 74.2 (2012), pp. 245–266.