

Concomitant Lasso with Repetitions (CLaR): beyond averaging multiple realizations of heteroscedastic noise

Joseph Salmon

<http://josephsalmon.eu>

IMAG, Univ Montpellier, CNRS
Montpellier, France

Joint works with:

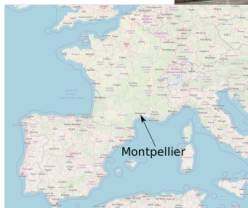
Quentin Bertrand (INRIA, Parietal Team)

Mathurin Massias (INRIA, Parietal Team)

Olivier Fercoq (Institut Polytechnique de Paris)

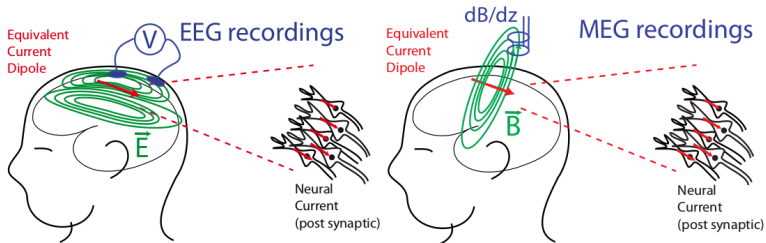
Alexandre Gramfort (INRIA, Parietal Team)

Montpellier: come, visit, work, etc.

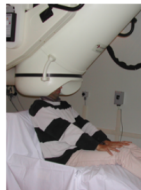
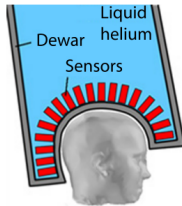


M/EEG inverse problem for brain imaging

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



First EEG recordings in 1929 by H. Berger

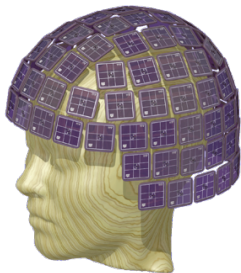


Hôpital La Timone
Marseille, France

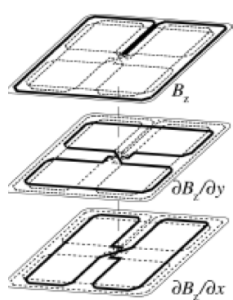
MEG elements: magnetometers and gradiometers



Device

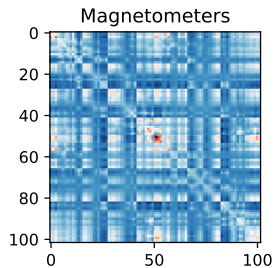
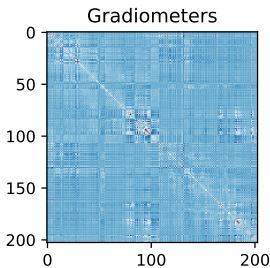
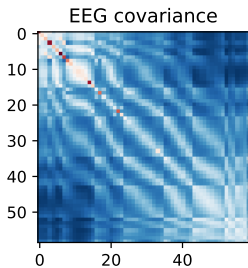


Sensors



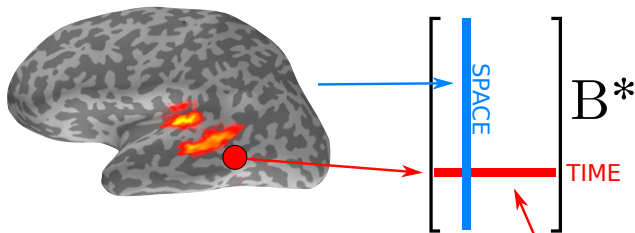
Detail of a sensor

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

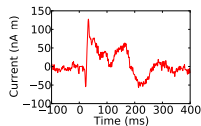


► 3 different sensors \implies 3 different noise structures

Source modeling

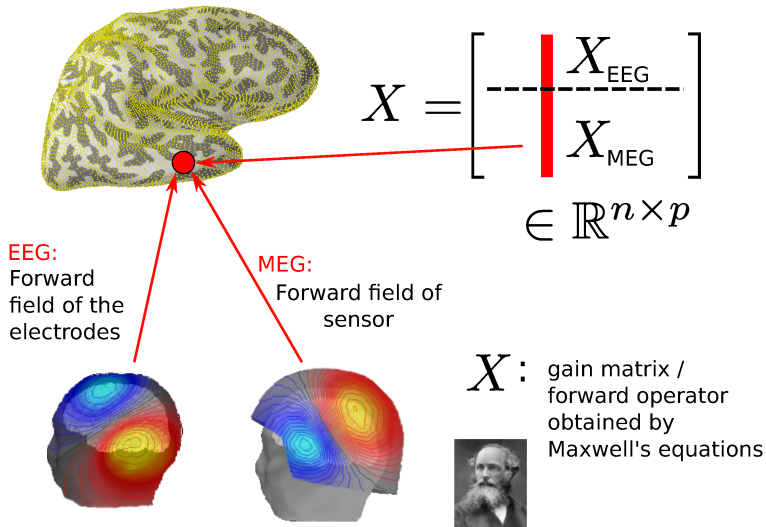


Position a few thousands candidate sources over the brain (e.g., every 5mm)

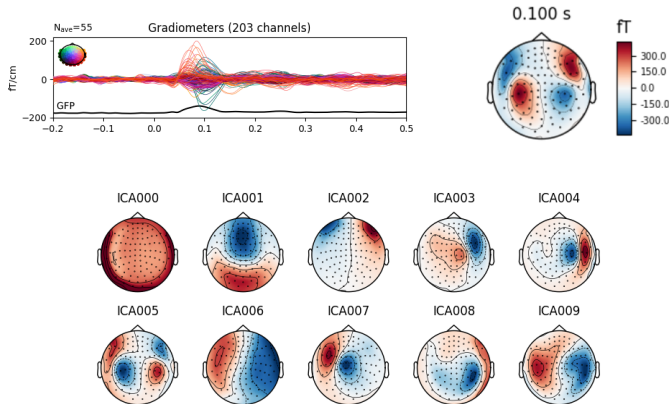


$$B^* \in \mathbb{R}^{p \times q}$$

Design matrix - Forward operator



Sparsity assumption: cortical sources produce dipolar patterns well modelled by focal sources



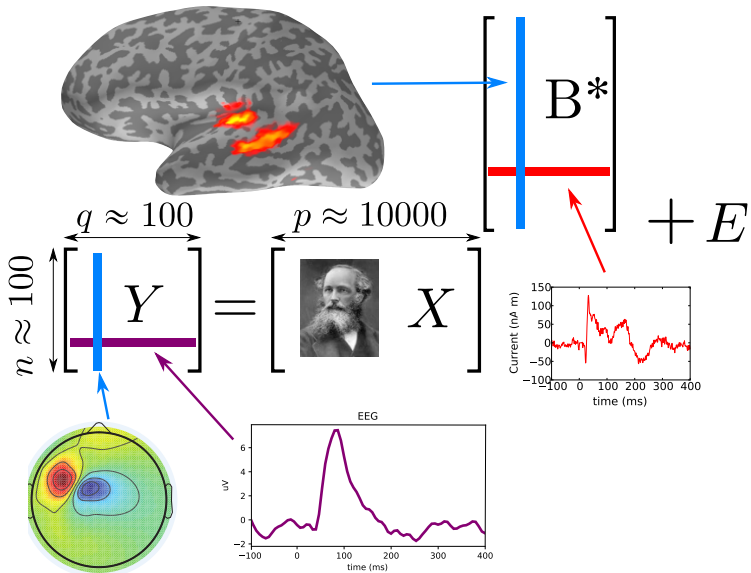
ICA : Blind source separation recovers dipolar patterns⁽¹⁾

http://martinos.org/mne/stable/auto_tutorials/plot_visualize_evoked.html

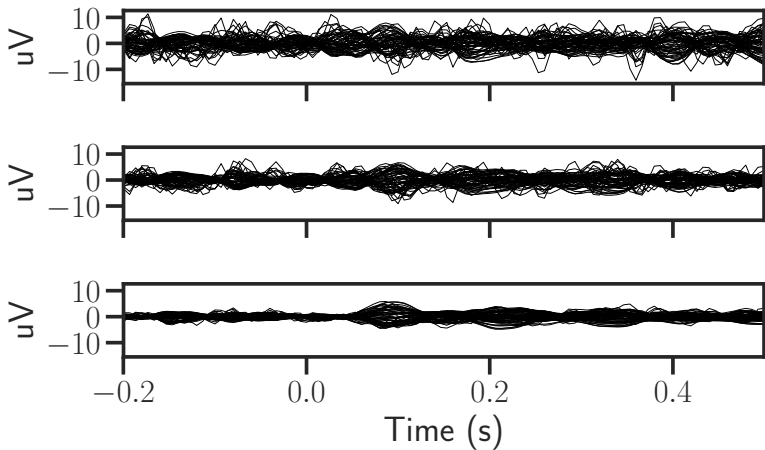
http://martinos.org/mne/stable/auto_tutorials/plot_artifacts_correction_ica.html

⁽¹⁾ A. Delorme et al. "Independent EEG sources are dipolar". In: *PLoS one* 7.2 (2012), e30135.

The M/EEG inverse problem: modeling



Multiple repetitions (r) of experiments:
 $r = 5$ (top), $r = 10$ (middle), $r = 50$ (bottom)
repetitions



A multi-task framework

Multi-task regression notation:

- ▶ n observations (e.g., number of sensors)
- ▶ q tasks (e.g., temporal information)
- ▶ p features (e.g., spatial description)
- ▶ r number of repetitions for the experiment
- ▶ $Y^{(1)}, \dots, Y^{(r)} \in \mathbb{R}^{n \times q}$ observation matrices; $\bar{Y} = \frac{1}{r} \sum_l Y^{(l)}$
- ▶ $X \in \mathbb{R}^{n \times p}$ forward matrix

$$\boxed{Y^{(l)} = XB^* + SE^{(l)}}, \quad \text{where}$$

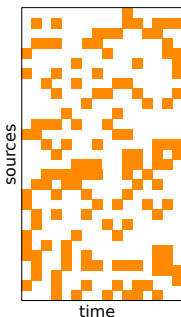
- ▶ $B^* \in \mathbb{R}^{p \times q}$: true source activity matrix (**unknown**)
- ▶ $S \in \mathbb{S}_{++}^n$ co-standard deviation matrix⁽²⁾ (**unknown**)
- ▶ $E^{(1)}, \dots, E^{(r)} \in \mathbb{R}^{n \times q}$: white Gaussian noise

⁽²⁾ $S \succeq \underline{\sigma}$ means $S - \underline{\sigma}$ is Semi-Definite Positive

Multi-tasks penalties⁽³⁾

Popular convex penalties considered:

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



Sparse support: no structure

Penalty: **Lasso type**

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

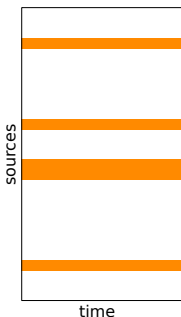
Parameter $\hat{B} \in \mathbb{R}^{p \times q}$

⁽³⁾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⁽³⁾

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

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



Sparse support: group structure

Penalty: **Group-Lasso type**

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

where $B_{j,:}$: the j -th row of B

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

⁽³⁾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 data-fitting term

- Classical multi-tasks estimator: use averaged signal

$$\hat{B} \in \arg \min_{B \in \mathbb{R}^{p \times q}} \left(\frac{1}{2nq} \left\| \bar{Y} - XB \right\|_F^2 + \lambda \Omega(B) \right)$$

- How to take advantage of the number of repetitions ?

Multi-tasks data-fitting term

- Classical multi-tasks estimator: use averaged signal

$$\hat{B} \in \arg \min_{B \in \mathbb{R}^{p \times q}} \left(\frac{1}{2nq} \left\| \bar{Y} - XB \right\|_F^2 + \lambda \Omega(B) \right)$$

- **How to take advantage of the number of repetitions ?**
- Intuitive estimator:

$$\hat{B}^{\text{repet}} \in \arg \min_{B \in \mathbb{R}^{p \times q}} \left(\frac{1}{2nqr} \sum_{l=1}^r \left\| Y^{(l)} - XB \right\|_F^2 + \lambda \Omega(B) \right)$$

Multi-tasks data-fitting term

- Classical multi-tasks estimator: use averaged signal

$$\hat{B} \in \arg \min_{B \in \mathbb{R}^{p \times q}} \left(\frac{1}{2nq} \left\| \bar{Y} - XB \right\|_F^2 + \lambda \Omega(B) \right)$$

- **How to take advantage of the number of repetitions ?**
- Intuitive estimator:

$$\hat{B}^{\text{repet}} \in \arg \min_{B \in \mathbb{R}^{p \times q}} \left(\frac{1}{2nqr} \sum_{l=1}^r \left\| Y^{(l)} - XB \right\|_F^2 + \lambda \Omega(B) \right)$$

- It's a fail! $\hat{B}^{\text{repet}} = \hat{B}$ (because of data-fitting loss $\|\cdot\|_F^2$)

Multi-tasks data-fitting term

- Classical multi-tasks estimator: use averaged signal

$$\hat{B} \in \arg \min_{B \in \mathbb{R}^{p \times q}} \left(\frac{1}{2nq} \left\| \bar{Y} - XB \right\|_F^2 + \lambda \Omega(B) \right)$$

- **How to take advantage of the number of repetitions ?**
- Intuitive estimator:

$$\hat{B}^{\text{repet}} \in \arg \min_{B \in \mathbb{R}^{p \times q}} \left(\frac{1}{2nqr} \sum_{l=1}^r \left\| Y^{(l)} - XB \right\|_F^2 + \lambda \Omega(B) \right)$$

- It's a fail! $\hat{B}^{\text{repet}} = \hat{B}$ (because of data-fitting loss $\|\cdot\|_F^2$)
- A priori $\hat{B}^{\text{repet}} \neq \hat{B}$ if the data-fitting is not $\|\cdot\|_F^2$, for instance $\|\cdot\|_F$ or $\|\cdot\|_{s,1}$ (Schatten-1 norm = trace / nuclear norm)

Multi-tasks data-fitting term

- Classical multi-tasks estimator: use averaged signal

$$\hat{B} \in \arg \min_{B \in \mathbb{R}^{p \times q}} \left(\frac{1}{2nq} \left\| \bar{Y} - XB \right\|_F^2 + \lambda \Omega(B) \right)$$

- **How to take advantage of the number of repetitions ?**
- Intuitive estimator:

$$\hat{B}^{\text{repet}} \in \arg \min_{B \in \mathbb{R}^{p \times q}} \left(\frac{1}{2nqr} \sum_{l=1}^r \left\| Y^{(l)} - XB \right\|_F^2 + \lambda \Omega(B) \right)$$

- It's a fail! $\hat{B}^{\text{repet}} = \hat{B}$ (because of data-fitting loss $\|\cdot\|_F^2$)
- A priori $\hat{B}^{\text{repet}} \neq \hat{B}$ if the data-fitting is not $\|\cdot\|_F^2$, for instance $\|\cdot\|_F$ or $\|\cdot\|_{s,1}$ (Schatten-1 norm = trace / nuclear norm)
- However $\|\cdot\|_F$ and $\|\cdot\|_{s,1}$ are non-smooth

Multi-tasks data-fitting term

- Classical multi-tasks estimator: use averaged signal

$$\hat{B} \in \arg \min_{B \in \mathbb{R}^{p \times q}} \left(\frac{1}{2nq} \left\| \bar{Y} - XB \right\|_F^2 + \lambda \Omega(B) \right)$$

- **How to take advantage of the number of repetitions ?**
- Intuitive estimator:

$$\hat{B}^{\text{repet}} \in \arg \min_{B \in \mathbb{R}^{p \times q}} \left(\frac{1}{2nqr} \sum_{l=1}^r \left\| Y^{(l)} - XB \right\|_F^2 + \lambda \Omega(B) \right)$$

- It's a fail! $\hat{B}^{\text{repet}} = \hat{B}$ (because of data-fitting loss $\|\cdot\|_F^2$)
- A priori $\hat{B}^{\text{repet}} \neq \hat{B}$ if the data-fitting is not $\|\cdot\|_F^2$, for instance $\|\cdot\|_F$ or $\|\cdot\|_{s,1}$ (Schatten-1 norm = trace / nuclear norm)
- However $\|\cdot\|_F$ and $\|\cdot\|_{s,1}$ are non-smooth
- Need for smoothing!

Multi-tasks data-fitting term

- Classical multi-tasks estimator: use averaged signal

$$\hat{B} \in \arg \min_{B \in \mathbb{R}^{p \times q}} \left(\frac{1}{2nq} \left\| \bar{Y} - XB \right\|_F^2 + \lambda \Omega(B) \right)$$

- **How to take advantage of the number of repetitions ?**
- Intuitive estimator:

$$\hat{B}^{\text{repet}} \in \arg \min_{B \in \mathbb{R}^{p \times q}} \left(\frac{1}{2nqr} \sum_{l=1}^r \left\| Y^{(l)} - XB \right\|_F^2 + \lambda \Omega(B) \right)$$

- It's a fail! $\hat{B}^{\text{repet}} = \hat{B}$ (because of data-fitting loss $\|\cdot\|_F^2$)
- A priori $\hat{B}^{\text{repet}} \neq \hat{B}$ if the data-fitting is not $\|\cdot\|_F^2$, for instance $\|\cdot\|_F$ or $\|\cdot\|_{s,1}$ (Schatten-1 norm = trace / nuclear norm)
- However $\|\cdot\|_F$ and $\|\cdot\|_{s,1}$ are non-smooth
- Need for smoothing!

Table of Contents

Calibrating λ and noise level estimation

Multi-task case and noise structure

Experiments

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**
- ▶ $\varepsilon \sim \mathcal{N}(0, \text{Id}_n)$ and σ_* **unknown**

Lasso reminder :

$$\hat{\beta}^{(\lambda)} \in \arg \min_{\beta \in \mathbb{R}^p} \frac{1}{2n} \|y - X\beta\|^2 + \lambda \|\beta\|_1$$

Lasso theory^{(4),(5)}

Theorem

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

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

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 !

⁽⁴⁾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.

⁽⁵⁾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^{(4),(5)}

Theorem

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

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

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 !

⁽⁴⁾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.

⁽⁵⁾A. S. Dalalyan, M. Hebiri, and J. Lederer. “On the Prediction Performance of the Lasso”. In: *Bernoulli* 23.1 (2017), pp. 552–581.

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 , e.g., with residual $\sigma \leftarrow \frac{\|y - X\beta\|}{\sqrt{n}}$
- ▶ re-scale $\lambda \propto \sigma$, and run Lasso with it
- ▶ iterate (until convergence)

Rem: exactly the Concomitant Lasso⁽⁶⁾ / Scaled-Lasso⁽⁷⁾ implementation

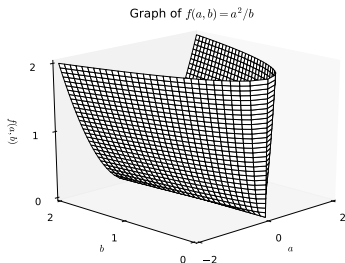
⁽⁶⁾A. B. Owen. "A robust hybrid of lasso and ridge regression". In: *Contemporary Mathematics* 443 (2007), pp. 59–72.

⁽⁷⁾T. Sun and C.-H. Zhang. "Scaled sparse linear regression". In: *Biometrika* 99.4 (2012), pp. 879–898.

Concomitant Lasso

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

- ▶ $\frac{\sigma}{2}$: penalty on noise level, roots in robust estimation^{(8), (9)}
- ▶ jointly convex program: $(a, b) \mapsto a^2/b$ is convex



⁽⁸⁾P. J. Huber and R. Dutter. "Numerical solution of robust regression problems". In: *Compstat 1974 (Proc. Sympos. Computational Statist., Univ. Vienna, Vienna, 1974)*. Physica Verlag, Vienna, 1974, pp. 165–172.

⁽⁹⁾P. J. Huber. *Robust Statistics*. John Wiley & Sons Inc., 1981.

Concomitant performance

Theorem^{(10), (11)}

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

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

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

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

Rem: λ has no dimension, but calibration still needed in practice...

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

⁽¹¹⁾ C. Giraud. *Introduction to high-dimensional statistics*. Vol. 138. CRC Press, 2014.

Link with $\sqrt{\text{Lasso}}^{(12)}$

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

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

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

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

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

⁽¹²⁾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⁽¹⁴⁾

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

- ▶ useful for optimization with small λ
- ▶ 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 $\sqrt{\text{Lasso}}$ ⁽¹³⁾

⁽¹³⁾Y. Nesterov. "Smooth minimization of non-smooth functions". In: *Math. Program.* 103.1 (2005), pp. 127–152.

⁽¹⁴⁾E. Ndiaye et al. "Efficient Smoothed Concomitant Lasso Estimation for High Dimensional Regression". In: *Journal of Physics: Conference Series* 904.1 (2017), p. 012006.

Smoothing aparté^{(15),(16)}

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

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

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

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

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

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

⁽¹⁶⁾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^{(15),(16)}

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

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

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

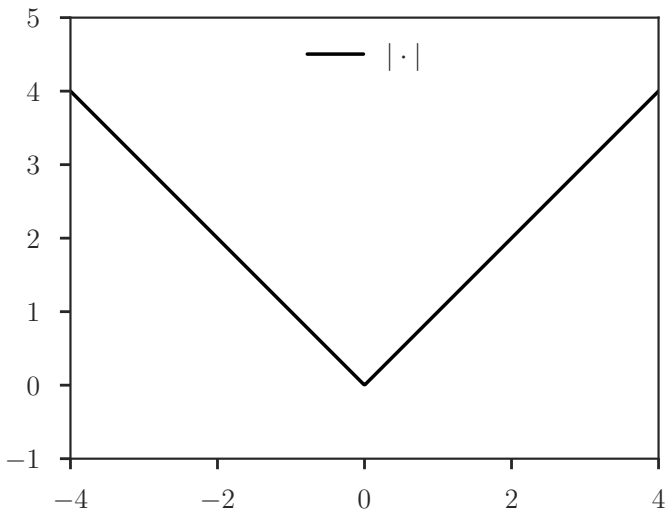
► ω is a predefined smooth function (s.t. $\nabla\omega$ is Lipschitz)

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

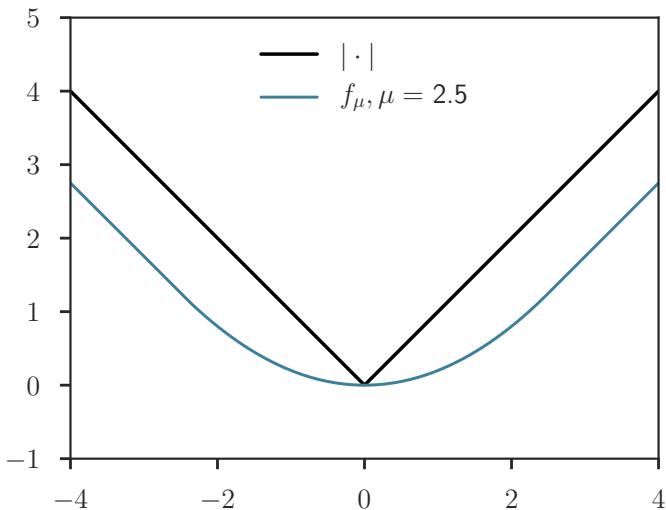
⁽¹⁵⁾Y. Nesterov. “Smooth minimization of non-smooth functions”. In: *Math. Program.* 103.1 (2005), pp. 127–152.

⁽¹⁶⁾A. Beck and M. Teboulle. “Smoothing and first order methods: A unified framework”. In: *SIAM J. Optim.* 22.2 (2012), pp. 557–580.

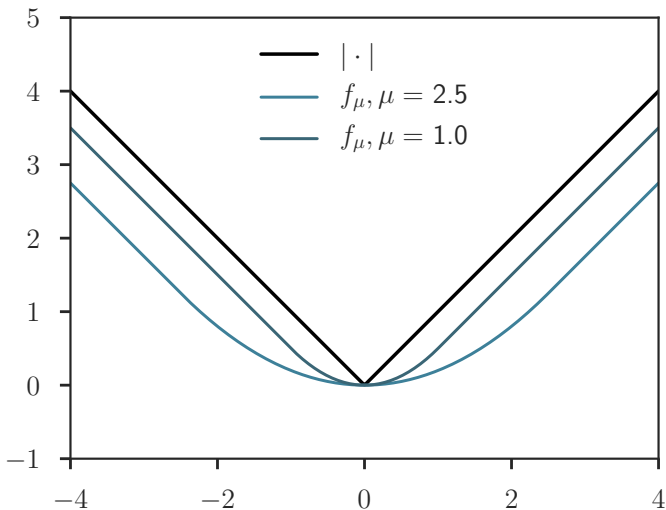
Huber function: $\omega(t) = \frac{t^2}{2}$



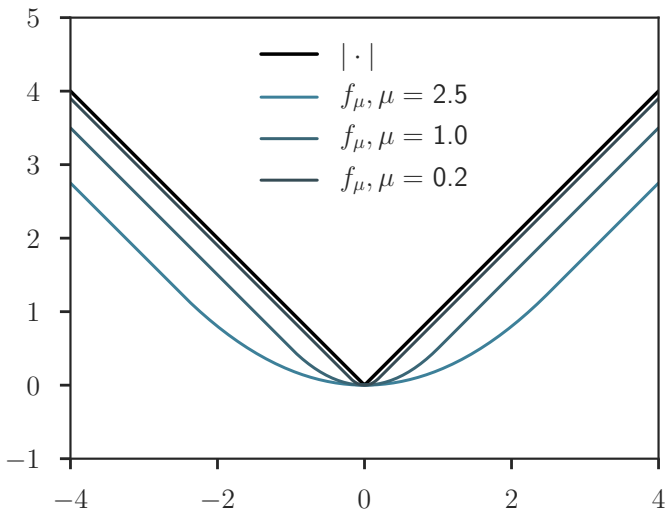
Huber function: $\omega(t) = \frac{t^2}{2}$



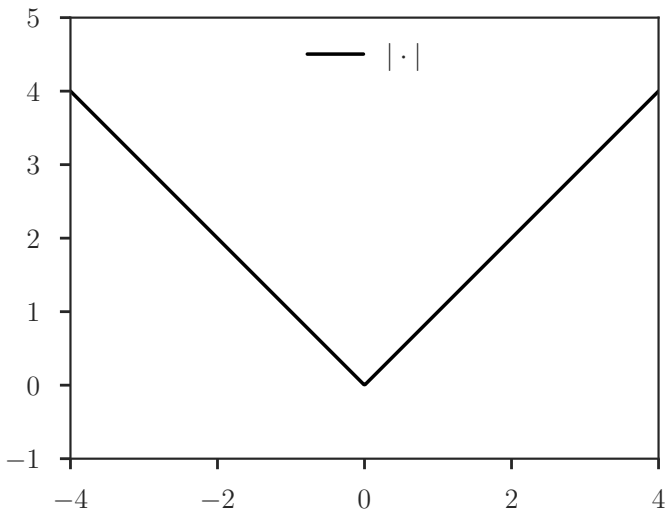
Huber function: $\omega(t) = \frac{t^2}{2}$



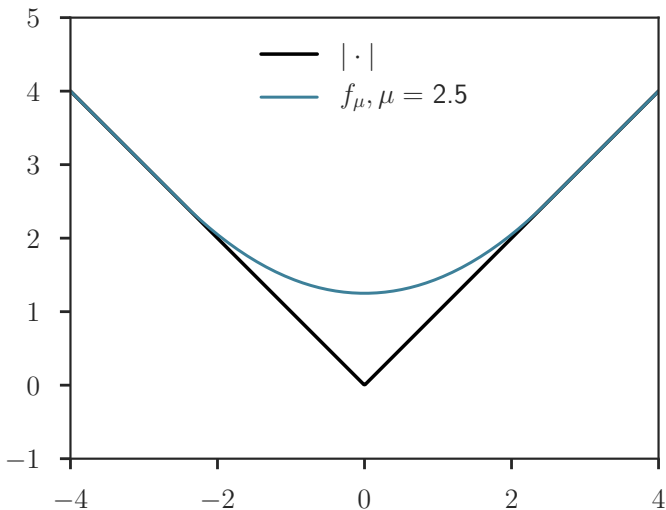
Huber function: $\omega(t) = \frac{t^2}{2}$



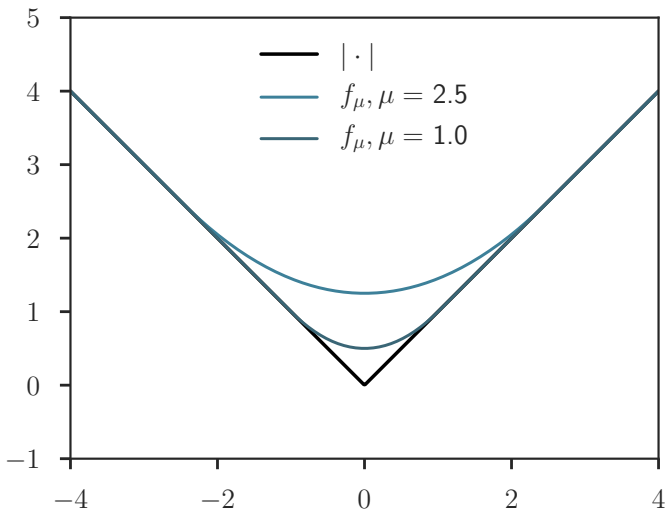
Huber function (bis): $\omega(t) = \frac{t^2}{2} + \frac{1}{2}$



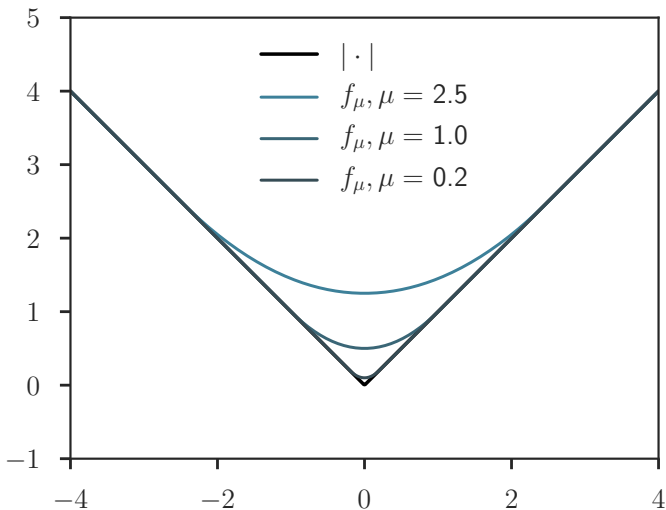
Huber function (bis): $\omega(t) = \frac{t^2}{2} + \frac{1}{2}$



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

$$\begin{aligned}\|\cdot\| \square_{\underline{\sigma}} \omega \left(\frac{\cdot}{\underline{\sigma}} \right) (z) &= \begin{cases} \frac{\|z\|^2}{2\underline{\sigma}} + \frac{\underline{\sigma}}{2}, & \text{if } \|z\| \leq \underline{\sigma} \\ \|z\|, & \text{if } \|z\| > \underline{\sigma} \end{cases} \\ &= \min_{\sigma \geq \underline{\sigma}} \left(\frac{\|z\|^2}{2\sigma} + \frac{\sigma}{2} \right)\end{aligned}$$

Leads to the Smoothed Concomitant Lasso formulation:

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

Solving the Smooth Concomitant Lasso

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

Jointly convex formulation : can be optimized by alternate minimization w.r.t. β and σ (gradient Lipschitz)

Alternate iteratively:

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

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

Solving the Smooth Concomitant Lasso

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

Jointly convex formulation : can be optimized by alternate minimization w.r.t. β and σ (gradient Lipschitz)

Alternate iteratively:

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

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

- 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

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

Jointly convex formulation : can be optimized by alternate minimization w.r.t. β and σ (gradient Lipschitz)

Alternate iteratively:

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

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

- 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

Calibrating λ and noise level estimation

Multi-task case and noise structure

Experiments

Back to multi-task : $Y^{(l)} = XB^* + SE^{(l)}$

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

“Huberization of the Frobenius norm”

$$\|\cdot\|_F \square_{\underline{\sigma}} \omega \left(\frac{\cdot}{\underline{\sigma}} \right) (Z) = \begin{cases} \frac{\|Z\|_F^2}{2\underline{\sigma}} + \frac{\underline{\sigma}}{2}, & \text{if } \|Z\| \leq \underline{\sigma} \\ \|Z\|_F, & \text{if } \|Z\| > \underline{\sigma} \end{cases}$$
$$= \min_{\sigma \geq \underline{\sigma}} \left(\frac{\|Z\|_F^2}{2\sigma} + \frac{\sigma}{2} \right)$$

Leads to the Smoothed Concomitant Lasso formulation:

$$(\hat{B}^{(\lambda)}, \hat{\sigma}^{(\lambda)}) \in \arg \min_{B \in \mathbb{R}^p, \sigma \geq \underline{\sigma}} \left(\frac{\|Y - XB\|_F^2}{2n\sigma} + \frac{\sigma}{2} + \lambda \|B\|_{2,1} \right)$$

and similar efficient algorithms.

Back to multi-task : $Y^{(l)} = XB^* + SE^{(l)}$

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

“Huberization of the Frobenius norm”

$$\|\cdot\|_F \square_{\underline{\sigma}\omega} \left(\frac{\cdot}{\underline{\sigma}} \right) (Z) = \begin{cases} \frac{\|Z\|_F^2}{2\underline{\sigma}} + \frac{\underline{\sigma}}{2}, & \text{if } \|Z\| \leq \underline{\sigma} \\ \|Z\|_F, & \text{if } \|Z\| > \underline{\sigma} \end{cases}$$
$$= \min_{\sigma \geq \underline{\sigma}} \left(\frac{\|Z\|_F^2}{2\sigma} + \frac{\sigma}{2} \right)$$

Leads to the Smoothed Concomitant Lasso formulation:

$$(\hat{B}^{(\lambda)}, \hat{\sigma}^{(\lambda)}) \in \arg \min_{B \in \mathbb{R}^p, \sigma \geq \underline{\sigma}} \left(\frac{\|Y - XB\|_F^2}{2n\sigma} + \frac{\sigma}{2} + \lambda \|B\|_{2,1} \right)$$

and similar efficient algorithms.

What about other norms ?

Huber-like formula for the Frobenius norm

$$\|\cdot\|_F \square_{\underline{\sigma}} \omega \left(\frac{\cdot}{\underline{\sigma}} \right) (Z) = \begin{cases} \frac{\|Z\|_F^2}{2\underline{\sigma}} + \frac{\underline{\sigma}}{2}, & \text{if } \|Z\|_F \leq \underline{\sigma} \\ \|Z\|_F, & \text{if } \|Z\|_F > \underline{\sigma} \end{cases}$$
$$= \min_{\sigma \geq \underline{\sigma}} \left(\frac{\|Z\|_F^2}{2\sigma} + \frac{\sigma}{2} \right)$$

What about other norms ?

What about other norms ?

Huber-like formula for the Frobenius norm

$$\|\cdot\|_F \square_{\underline{\sigma}} \omega \left(\frac{\cdot}{\underline{\sigma}} \right) (Z) = \begin{cases} \frac{\|Z\|_F^2}{2\underline{\sigma}} + \frac{\underline{\sigma}}{2}, & \text{if } \|Z\|_F \leq \underline{\sigma} \\ \|Z\|_F, & \text{if } \|Z\|_F > \underline{\sigma} \end{cases}$$

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

What about other norms ? **TO DO: introduce nuclear norm usage better... transition missing!!!** Huber-like formula for the nuclear/trace norm

$$\|\cdot\|_{s,1} \square_{\underline{\sigma}} \omega_{\underline{\sigma}}(Z) = \begin{cases} \frac{1}{2\underline{\sigma}} \sum_i \gamma_i^2 - (\gamma_i \wedge \underline{\sigma} - \gamma_i)^2, & \text{if } \|Z\|_{s,1} \leq \underline{\sigma} \\ \|Z\|_{s,1}, & \text{if } \|Z\|_{s,1} > \underline{\sigma} \end{cases}$$

$$= \min_{S \succeq \underline{\sigma}} \frac{1}{2} \|Z\|_{S^{-1}}^2 + \frac{1}{2} \text{Tr}(S)$$

γ_i : singular values of Z

$\|Z\|_{S^{-1}}^2 := \text{Tr}(Z^\top S^{-1} Z)$ **Mahalanobis distance**

What about other norms ?

Huber-like formula for the Frobenius norm

$$\|\cdot\|_F \square_{\underline{\sigma}} \omega \left(\frac{\cdot}{\underline{\sigma}} \right) (Z) = \begin{cases} \frac{\|Z\|_F^2}{2\underline{\sigma}} + \frac{\underline{\sigma}}{2}, & \text{if } \|Z\|_F \leq \underline{\sigma} \\ \|Z\|_F, & \text{if } \|Z\|_F > \underline{\sigma} \end{cases}$$

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

What about other norms ? **TO DO: introduce nuclear norm usage better... transition missing!!!** Huber-like formula for the nuclear/trace norm

$$\|\cdot\|_{s,1} \square_{\underline{\sigma}} \omega_{\underline{\sigma}}(Z) = \begin{cases} \frac{1}{2\underline{\sigma}} \sum_i \gamma_i^2 - (\gamma_i \wedge \underline{\sigma} - \gamma_i)^2, & \text{if } \|Z\|_{s,1} \leq \underline{\sigma} \\ \|Z\|_F, & \text{if } \|Z\|_{s,1} > \underline{\sigma} \end{cases}$$

$$= \min_{S \succeq \underline{\sigma}} \frac{1}{2} \|Z\|_{S^{-1}}^2 + \frac{1}{2} \text{Tr}(S)$$

γ_i : singular values of Z

$\|Z\|_{S^{-1}}^2 := \text{Tr}(Z^\top S^{-1} Z)$ **Mahalanobis distance**

Smoothing of the nuclear/trace norm

Smoothed Generalized Concomitant Lasso (SGCL)⁽¹⁷⁾:

$$(\hat{B}^{\text{SGCL}}, \hat{S}^{\text{SGCL}}) \in \arg \min_{\substack{B \in \mathbb{R}^{p \times q} \\ S \in \mathbb{S}_{++}^n, S \succeq \underline{\sigma}}} \frac{\|\bar{Y} - XB\|_{S^{-1}}^2}{2nq} + \frac{\text{Tr}(S)}{2n} + \lambda \|B\|_{2,1}$$

Concomitant Lasso with Repetitions (CLaR)⁽¹⁸⁾:

$$(\hat{B}^{\text{CLaR}}, \hat{S}^{\text{CLaR}}) \in \arg \min_{\substack{B \in \mathbb{R}^{p \times q} \\ S \in \mathbb{S}_{++}^n, S \succeq \underline{\sigma}}} \frac{\sum_{l=1}^r \|Y^{(l)} - XB\|_{S^{-1}}^2}{2nq} + \frac{\text{Tr}(S)}{2n} + \lambda \|B\|_{2,1}$$

⁽¹⁷⁾ M. Massias et al. "Generalized Concomitant Multi-Task Lasso for Sparse Multimodal Regression". In: *AISTATS*. vol. 84. 2018, pp. 998–1007.

⁽¹⁸⁾ Q. Bertrand et al. "Handling correlated and repeated measurements with the smoothed multivariate square-root Lasso". In: *NeurIPS*. 2019.

Efficient solvers for SGCL and CLaR

General case: $Y^{(l)} \in \mathbb{R}^{n \times q}$, $B \in \mathbb{R}^{p \times q}$, and the noise $E^{(l)} \in \mathbb{R}^{n \times q}$

⁽¹⁹⁾S. van de Geer. *Estimation and testing under sparsity*. Vol. 2159. Lecture Notes in Mathematics. Lecture notes from the 45th Probability Summer school held in Saint-Flour, 2015, École d'Été de Probabilités de Saint-Flour. Springer, 2016, pp. xiii+274.

Efficient solvers for SGCL and CLaR

General case: $Y^{(l)} \in \mathbb{R}^{n \times q}$, $B \in \mathbb{R}^{p \times q}$, and the noise $E^{(l)} \in \mathbb{R}^{n \times q}$

SGCL:

$$(\hat{B}^{\text{SGCL}}, \hat{S}^{\text{SGCL}}) \in \arg \min_{\substack{B \in \mathbb{R}^{p \times q} \\ S \in \mathbb{S}_{++}^n, S \succeq \underline{\sigma}}} \frac{\|\bar{Y} - XB\|_{S^{-1}}^2}{2nq} + \frac{\text{Tr}(S)}{2n} + \lambda \|B\|_{2,1}$$

CLaR:

$$(\hat{B}^{\text{CLaR}}, \hat{S}^{\text{CLaR}}) \in \arg \min_{\substack{B \in \mathbb{R}^{p \times q} \\ S \in \mathbb{S}_{++}^n, S \succeq \underline{\sigma}}} \frac{\sum_{l=1}^r \|Y^{(l)} - XB\|_{S^{-1}}^2}{2nqr} + \frac{\text{Tr}(S)}{2n} + \lambda \|B\|_{2,1}$$

with $\|Z\|_{S^{-1}}^2 := \text{Tr}(Z^\top S^{-1} Z)$ (Mahalanobis distance)

⁽¹⁹⁾ S. van de Geer. *Estimation and testing under sparsity*. Vol. 2159. Lecture Notes in Mathematics. Lecture notes from the 45th Probability Summer school held in Saint-Flour, 2015, École d'Été de Probabilités de Saint-Flour. Springer, 2016, pp. xiii+274.

Efficient solvers for SGCL and CLaR

General case: $Y^{(l)} \in \mathbb{R}^{n \times q}$, $B \in \mathbb{R}^{p \times q}$, and the noise $E^{(l)} \in \mathbb{R}^{n \times q}$

SGCL:

$$(\hat{B}^{\text{SGCL}}, \hat{S}^{\text{SGCL}}) \in \arg \min_{\substack{B \in \mathbb{R}^{p \times q} \\ S \in \mathbb{S}_{++}^n, S \succeq \underline{\sigma}}} \frac{\|\bar{Y} - XB\|_{S^{-1}}^2}{2nq} + \frac{\text{Tr}(S)}{2n} + \lambda \|B\|_{2,1}$$

CLaR:

$$(\hat{B}^{\text{CLaR}}, \hat{S}^{\text{CLaR}}) \in \arg \min_{\substack{B \in \mathbb{R}^{p \times q} \\ S \in \mathbb{S}_{++}^n, S \succeq \underline{\sigma}}} \frac{\sum_{l=1}^r \|Y^{(l)} - XB\|_{S^{-1}}^2}{2nqr} + \frac{\text{Tr}(S)}{2n} + \lambda \|B\|_{2,1}$$

with $\|Z\|_{S^{-1}}^2 := \text{Tr}(Z^\top S^{-1} Z)$ (Mahalanobis distance)

- ▶ jointly convex formulation (=nuclear norm smoothing⁽¹⁹⁾)
- ▶ noise penalty on the sum of the eigenvalues of S ($S^* = \Sigma^{*\frac{1}{2}}$)

⁽¹⁹⁾ S. van de Geer. *Estimation and testing under sparsity*. Vol. 2159. Lecture Notes in Mathematics. Lecture notes from the 45th Probability Summer school held in Saint-Flour, 2015, École d'Été de Probabilités de Saint-Flour. Springer, 2016, pp. xiii+274.

SGCL and CLaR computations: B update

Jointly convex formulation: alternate minimization still converging

B Update - S fixed:

“smooth + non-smooth” optimization; use Block Coordinate Descent (Iterative Block Soft-Thresholding) to update B row-wise

Possible refinements:

- ▶ (Gap) safe screening rules⁽²⁰⁾, ⁽²¹⁾
- ▶ Strong rules⁽²²⁾
- ▶ Active sets methods⁽²³⁾ etc.

⁽²⁰⁾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.

⁽²¹⁾E. Ndiaye et al. “Efficient Smoothed Concomitant Lasso Estimation for High Dimensional Regression”. In: *Journal of Physics: Conference Series* 904.1 (2017), p. 012006.

⁽²²⁾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.

⁽²³⁾T. B. Johnson and C. Guestrin. “Blitz: A Principled Meta-Algorithm for Scaling Sparse Optimization”. In: *ICML*. 2015, pp. 1171–1179.

SGCL and CLaR computations: B update

Jointly convex formulation: alternate minimization still converging

B Update - S fixed:

“smooth + non-smooth” optimization; use Block Coordinate Descent (Iterative Block Soft-Thresholding) to update B row-wise

Possible refinements:

- ▶ (Gap) safe screening rules⁽²⁰⁾, ⁽²¹⁾
- ▶ Strong rules⁽²²⁾
- ▶ Active sets methods⁽²³⁾ etc.

⁽²⁰⁾ 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.

⁽²¹⁾ E. Ndiaye et al. “Efficient Smoothed Concomitant Lasso Estimation for High Dimensional Regression”. In: *Journal of Physics: Conference Series* 904.1 (2017), p. 012006.

⁽²²⁾ 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.

⁽²³⁾ T. B. Johnson and C. Guestrin. “Blitz: A Principled Meta-Algorithm for Scaling Sparse Optimization”. In: *ICML*. 2015, pp. 1171–1179.

SGCL and CLaR computations: S update

S Update - B fixed:

For SGCL and CLaR the problem can be reformulated as

$$\hat{S} = \arg \min_{S \in \mathbb{S}_{++}^n, S \succeq \underline{\sigma} \text{Id}_n} \left(\frac{1}{2n} \underbrace{\text{Tr}[Z^\top S^{-1} Z]}_{\|Z\|_{S^{-1}}^2} + \frac{1}{2n} \text{Tr}(S) \right)$$

Rem: as in the classical concomitant Lasso, at each step CLaR and SGCL estimate alternatively B and S

SGCL and CLaR computations: S update

S Update - B fixed:

For SGCL and CLaR the problem can be reformulated as

$$\hat{S} = \arg \min_{S \in \mathbb{S}_{++}^n, S \succeq \underline{\sigma} \text{Id}_n} \left(\frac{1}{2n} \underbrace{\text{Tr}[Z^\top S^{-1} Z]}_{\|Z\|_{S^{-1}}^2} + \frac{1}{2n} \text{Tr}(S) \right)$$

Closed-form solution (**Spectral clipping**):

if $U^\top \text{diag}(s_1, \dots, s_n)U$ is the spectral decomposition of ZZ^\top :

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

Rem: as in the classical concomitant Lasso, at each step CLaR and SGCL estimate alternatively B and S

Main drawbacks

- ▶ Statistically⁽²⁴⁾: $\mathcal{O}(n^2)$ parameters to estimate for S
 - SGCL case: only nq observations (need q large w.r.t. n)
 - CLaR case: only nq ^r observations
- ▶ Computationally: S update cost is $\mathcal{O}(n^3)$ too slow in general (SVD computation)
Rem: fine for MEG/EEG problems ($n \approx 300$)

More structure can easily be incorporated to estimate S ,
e.g., block diagonal, etc.

⁽²⁴⁾not to mention that the original model is not identifiable

Main drawbacks

- ▶ Statistically⁽²⁴⁾: $\mathcal{O}(n^2)$ parameters to estimate for S
 - SGCL case: only nq observations (need q large w.r.t. n)
 - CLaR case: only nq ^r observations
- ▶ Computationally: S update cost is $\mathcal{O}(n^3)$ too slow in general (SVD computation)
Rem: fine for MEG/EEG problems ($n \approx 300$)

More structure can easily be incorporated to estimate S ,
e.g., block diagonal, etc.

⁽²⁴⁾not to mention that the original model is not identifiable

Table of Contents

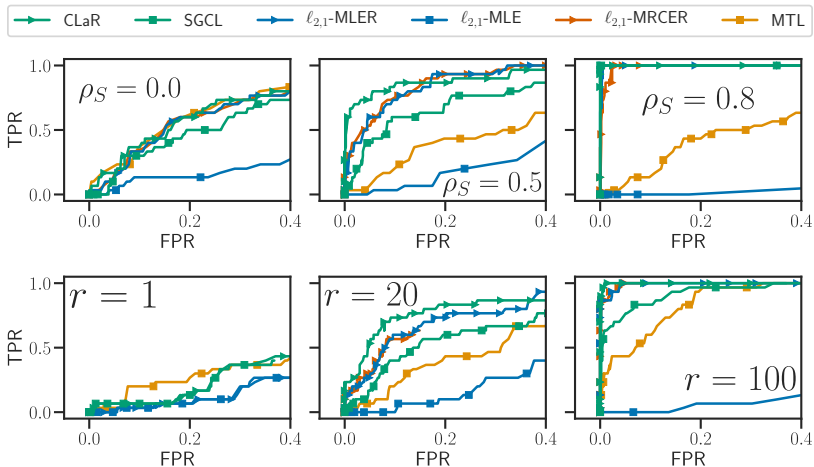
Calibrating λ and noise level estimation

Multi-task case and noise structure

Experiments

Simulated scenarios

- ▶ $n = 150, p = 500, q = 100$
- ▶ X Toeplitz-correlated: $\text{Cov}(X_i, X_j) = \rho^{|i-j|}$, $\rho_X \in]0, 1[$
- ▶ S^* Toeplitz matrix: $S^*_{i,j} = \rho_{S^*}^{|i-j|}$, $\rho_{S^*} \in]0, 1[$



Real data

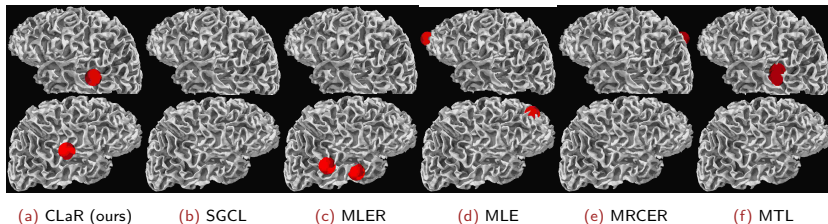


Figure: *Real data, left auditory stimulations* ($n = 102$, $p = 7498$, $q = 76$, $r = 63$) Sources found in the left hemisphere (top) and the right hemisphere (bottom) after left auditory stimulations.

- ▶ expected: 2 sources (one in each auditory cortex)
- ▶ λ chosen such that $\|\hat{\mathbf{B}}\|_{2,0} = 2$
- ▶ deep sources for SGCL and $\ell_{2,1}$ -MRCER (not visible)

Conclusion and perspectives

- ▶ New insights for handling (structured) noise in multi-task
- ▶ Handling refined noise structure benefits:
improve support identification (and prediction)

Conclusion and perspectives

- ▶ New insights for handling (structured) noise in multi-task
- ▶ Handling refined noise structure benefits:
improve support identification (and prediction)
- ▶ Numerical cost equivalent to classical Multi-Task Lasso for
“simple” noise structure (e.g., block homoscedastic)

Conclusion and perspectives

- ▶ New insights for handling (structured) noise in multi-task
- ▶ Handling refined noise structure benefits:
improve support identification (and prediction)
- ▶ Numerical cost equivalent to classical Multi-Task Lasso for
“simple” noise structure (e.g., block homoscedastic)
- ▶ Future work: non-convex penalties, statistical analysis, etc.

Conclusion and perspectives

- ▶ New insights for handling (structured) noise in multi-task
- ▶ Handling refined noise structure benefits:
improve support identification (and prediction)
- ▶ Numerical cost equivalent to classical Multi-Task Lasso for
“simple” noise structure (e.g., block homoscedastic)
- ▶ Future work: non-convex penalties, statistical analysis, etc.

Merci!

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

A. Gramfort

- ▶ Paper: arXiv / personal webpage^{(25), (26)}
- ▶ Python code online for CLaR <https://github.com/QB3/CLaR>
- ▶ Python code online for SGCL <https://github.com/mathurinm/SHCL>



Powered with **MooseTeX**

⁽²⁵⁾ M. Massias et al. "Generalized Concomitant Multi-Task Lasso for Sparse Multimodal Regression". In: *AISTATS*. vol. 84. 2018, pp. 998–1007.

⁽²⁶⁾ Q. Bertrand et al. "Handling correlated and repeated measurements with the smoothed multivariate square-root Lasso". In: *NeurIPS*. 2019.

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.
- ▶ Bertrand, Q. et al. “Handling correlated and repeated measurements with the smoothed multivariate square-root Lasso”. In: *NeurIPS*. 2019.
- ▶ 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.
- ▶ Dalalyan, A. S., M. Hebiri, and J. Lederer. “On the Prediction Performance of the Lasso”. In: *Bernoulli* 23.1 (2017), pp. 552–581.

References II

- ▶ Delorme, A. et al. “Independent EEG sources are dipolar”. In: *PloS one* 7.2 (2012), e30135.
- ▶ 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.
- ▶ Giraud, C. *Introduction to high-dimensional statistics*. Vol. 138. CRC Press, 2014.
- ▶ Huber, P. J. *Robust Statistics*. John Wiley & Sons Inc., 1981.
- ▶ Huber, P. J. and R. Dutter. “Numerical solution of robust regression problems”. In: *Compstat 1974 (Proc. Sympos. Computational Statist., Univ. Vienna, Vienna, 1974)*. Physica Verlag, Vienna, 1974, pp. 165–172.
- ▶ 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.
- ▶ Ndiaye, E. et al. “Efficient Smoothed Concomitant Lasso Estimation for High Dimensional Regression”. In: *Journal of Physics: Conference Series* 904.1 (2017), p. 012006.
- ▶ 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.
- ▶ 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.

References IV

- ▶ 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.
- ▶ van de Geer, S. *Estimation and testing under sparsity*. Vol. 2159. Lecture Notes in Mathematics. Lecture notes from the 45th Probability Summer school held in Saint-Four, 2015, École d'Été de Probabilités de Saint-Flour. Springer, 2016, pp. xiii+274.

Competitors

- (smoothed) $\ell_{2,1}$ -MLE

$$(\hat{\mathbf{B}}, \hat{\Sigma}) \in \arg \min_{\substack{\mathbf{B} \in \mathbb{R}^{p \times q} \\ \Sigma \succeq \underline{\sigma}^2 / r^2}} \left\| \bar{\mathbf{Y}} - X\mathbf{B} \right\|_{\Sigma^{-1}}^2 - \log \det(\Sigma^{-1}) + \lambda \|\mathbf{B}\|_{2,1} \quad ,$$

- and its repetitions version ($\ell_{2,1}$ -MLER):

$$(\hat{\mathbf{B}}, \hat{\Sigma}) \in \arg \min_{\substack{\mathbf{B} \in \mathbb{R}^{p \times q} \\ \Sigma \succeq \underline{\sigma}^2}} \sum_1^r \left\| \mathbf{Y}^{(l)} - X\mathbf{B} \right\|_{\Sigma^{-1}}^2 - \log \det(\Sigma^{-1}) + \lambda \|\mathbf{B}\|_{2,1} \quad .$$

- $\ell_{2,1}$ -MLE and $\ell_{2,1}$ -MLER are bi-convex but not jointly convex