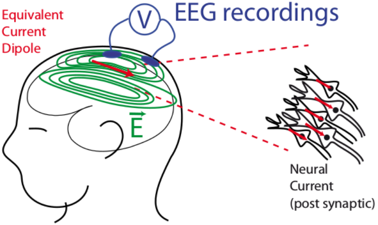
Correlated and repeated measurements with the smoothed Multivariate square-root Lasso

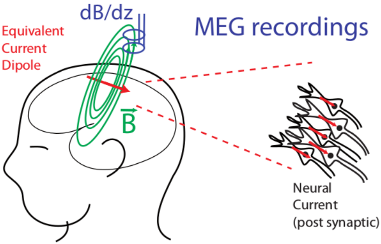
Joseph Salmon
http://josephsalmon.eu
IMAG, Univ Montpellier, CNRS
Montpellier, France

Joint works with:
Quentin Bertrand (INRIA, Parietal Team)
Mathurin Massias (INRIA, Parietal Team)
Olivier Fercoq (Institut Polytechnique de Paris)
Alexandre Gramfort (INRIA, Parietal Team)

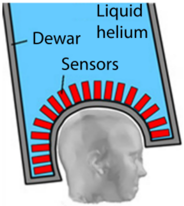
M/EEG inverse problem for brain imaging

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





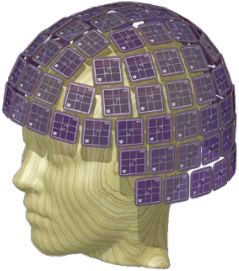
First EEG
recordings
in 1929
by H. Berger



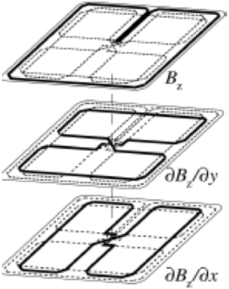
Hôpital La Timone
Marseille, France

MEG elements: magnometers and gradiometers

Device

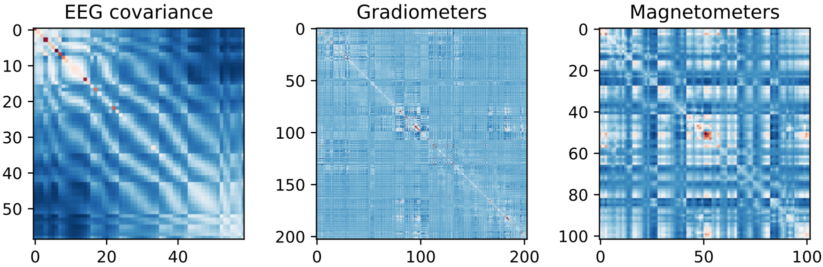


Sensors



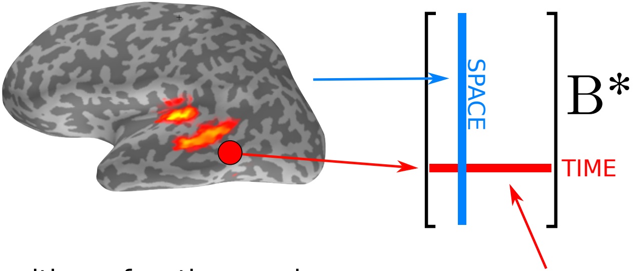
Detail of a sensor

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

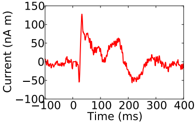


- 3 type of sensors $\Longrightarrow$ 3 different noise structures

Source modeling

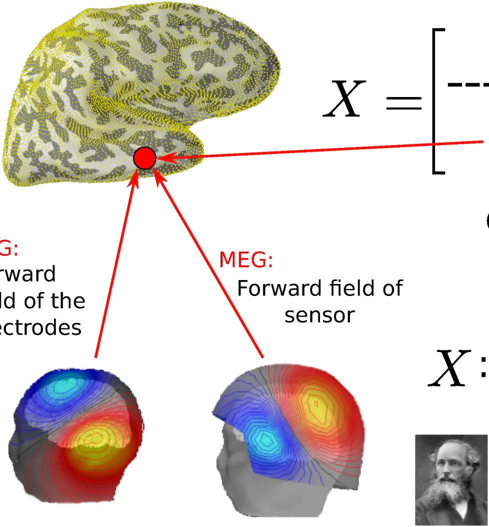


Position a few thousands candidate sources over the brain (*e.g.*, every 5mm)



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

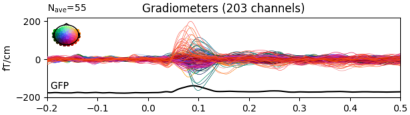
Design matrix - Forward operator

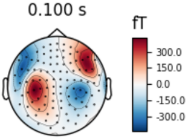


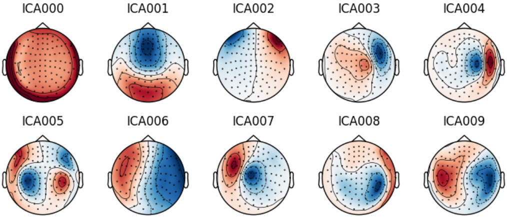
$$X = \left[\begin{array}{c} X_{\text{EEG}} \\ \hdashline X_{\text{MEG}} \end{array}\right] \in \mathbb{R}^{n \times p}$$

X : gain matrix / forward operator obtained by Maxwell's equations

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





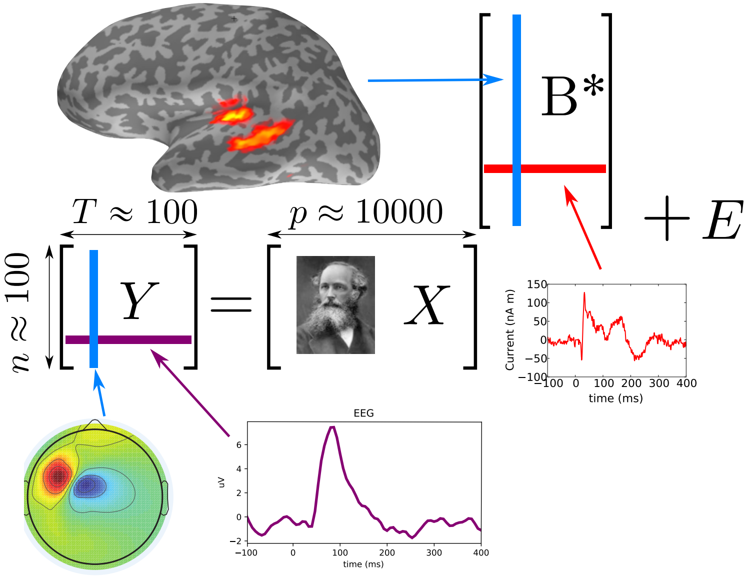


ICA : Blind source separation recovers dipolar patterns[1]

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



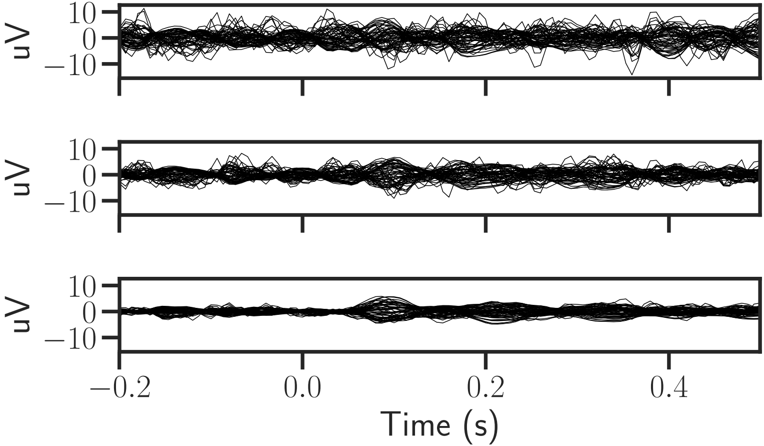
$T \approx 100$

$p \approx 10000$

$n \approx 100$

$$Y = X \, B^* + E$$

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

$$\boxed{Y^{(l)} = X\mathrm{B}^* + S_* \mathrm{E}^{(l)}}, \quad \text{where}$$

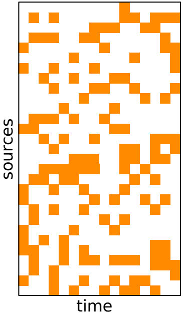
- $\mathrm{B}^* \in \mathbb{R}^{p \times T}$: true source activity matrix (**unknown**)
- $S_* \in \mathbb{S}^n_{++}$ co-standard deviation matrix[2] (**unknown**)
- $\mathrm{E}^{(1)}, \dots, \mathrm{E}^{(r)} \in \mathbb{R}^{n \times T}$: white Gaussian noise

[2] $S \succeq \underline{\sigma}$ means $S - \underline{\sigma}\, \mathrm{Id}_n$ is Semi-Definite Positive

Multi-tasks penalties[3]

Popular convex penalties considered:

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



Sparse support: no structure

Penalty: **Lasso type**

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

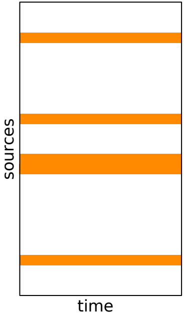
Parameter $\hat{\mathrm{B}} \in \mathbb{R}^{p \times T}$

[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)

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



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

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 row of B

[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

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

- **How to take advantage of the number of repetitions ?**

Multi-tasks data-fitting term

- Classical multi-tasks estimator: use averaged signal

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

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

$$\hat{\mathrm{B}}^{\text{repet}} \in \underset{\mathrm{B} \in \mathbb{R}^{p \times T}}{\arg\min} \left(\frac{1}{2nTr} \sum_{l=1}^{r} \left\| Y^{(l)} - X\mathrm{B} \right\|_F^2 + \lambda \Omega(\mathrm{B}) \right)$$

Multi-tasks data-fitting term

- Classical multi-tasks estimator: use averaged signal

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

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

$$\hat{\mathrm{B}}^{\mathsf{repet}} \in \underset{\mathrm{B} \in \mathbb{R}^{p \times T}}{\arg\min} \left(\frac{1}{2nTr} \sum_{l=1}^{r} \left\| Y^{(l)} - X\mathrm{B} \right\|_F^2 + \lambda \Omega(\mathrm{B}) \right)$$

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

Multi-tasks data-fitting term

- Classical multi-tasks estimator: use averaged signal

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

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

$$\hat{\mathrm{B}}^{\text{repet}} \in \operatorname*{arg\,min}_{\mathrm{B} \in \mathbb{R}^{p \times T}} \left(\frac{1}{2nTr} \sum_{l=1}^{r} \left\| Y^{(l)} - X\mathrm{B} \right\|_F^2 + \lambda \Omega(\mathrm{B}) \right)$$

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

Multi-tasks data-fitting term

- Classical multi-tasks estimator: use averaged signal

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

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

$$\hat{\mathrm{B}}^{\text{repet}} \in \underset{\mathrm{B} \in \mathbb{R}^{p \times T}}{\arg\min} \left(\frac{1}{2nTr} \sum_{l=1}^{r} \left\| Y^{(l)} - X\mathrm{B} \right\|_F^2 + \lambda \Omega(\mathrm{B}) \right)$$

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

Multi-tasks data-fitting term

- Classical multi-tasks estimator: use averaged signal

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

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

$$\hat{\mathrm{B}}^{\text{repet}} \in \operatorname*{arg\,min}_{\mathrm{B} \in \mathbb{R}^{p \times T}} \left(\frac{1}{2nTr} \sum_{l=1}^{r} \left\| Y^{(l)} - X\mathrm{B} \right\|_F^2 + \lambda \Omega(\mathrm{B}) \right)$$

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

Multi-tasks data-fitting term

- Classical multi-tasks estimator: use averaged signal

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

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

$$\hat{\mathrm{B}}^{\text{repet}} \in \mathop{\arg\min}_{\mathrm{B}\in\mathbb{R}^{p\times T}} \left(\frac{1}{2nTr} \sum_{l=1}^{r} \left\| Y^{(l)} - X\mathrm{B} \right\|_F^2 + \lambda\Omega(\mathrm{B}) \right)$$

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

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

Lasso reminder : $$\boxed{\hat{\beta}^{(\lambda)} \in \operatorname*{arg\,min}_{\beta \in \mathbb{R}^p} \frac{1}{2n} \left\|y - X\beta\right\|^2 + \lambda \left\|\beta\right\|_1}$$

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

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

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 !

[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

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

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 !

[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

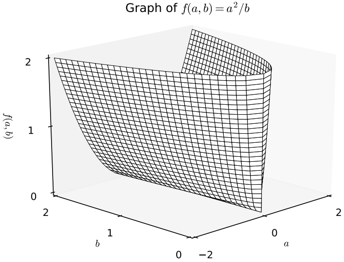
[6] A. B. Owen. "A robust hybrid of lasso and ridge regression". In: *Contemporary Mathematics* 443 (2007), pp. 59–72.

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

Concomitant Lasso

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

- $\frac{\sigma}{2}$: 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 the 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: 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

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

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

[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]

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

- 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

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

- 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 : $\mathcal{F}(g) = g$	$\omega = \frac{\lVert\cdot\rVert^2}{2}$: $\quad \omega^* = \omega$
	$f_h = \frac{1}{h} g\left(\frac{\cdot}{h}\right) \star f$	$f_\mu = \mu\omega\left(\frac{\cdot}{\mu}\right)\Box f$

[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

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

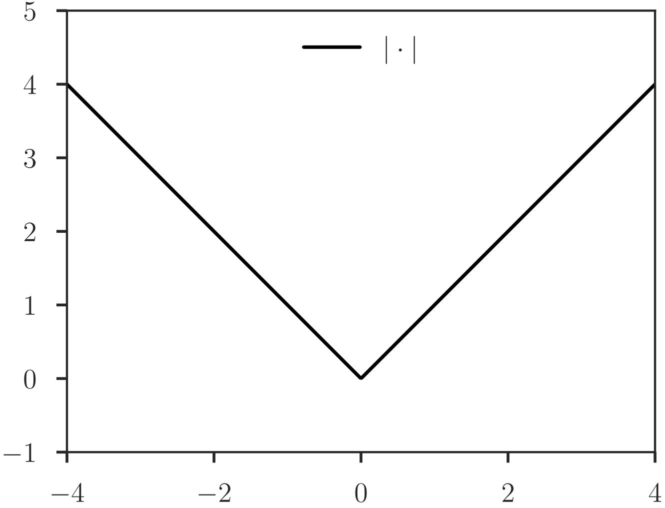
- ω is a predefined smooth function (s.t. $\nabla\omega$ is Lipschitz)

	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 : $\mathcal{F}(g) = g$	$\omega = \frac{\lVert\cdot\rVert^2}{2}$: $\quad \omega^* = \omega$
	$f_h = \frac{1}{h} g\left(\frac{\cdot}{h}\right) \star f$	$f_\mu = \mu\omega\left(\frac{\cdot}{\mu}\right) \Box f$

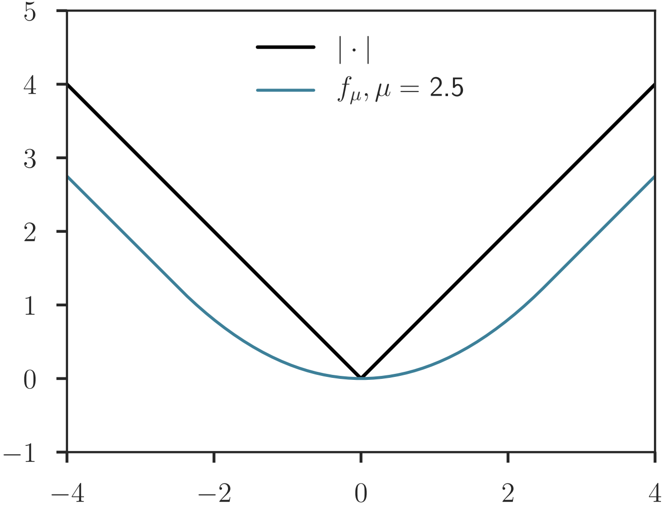
[15] Y. Nesterov. "Smooth minimization of non-smooth functions". In: *Math. Program.* 103.1 (2005), pp. 127–152.

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

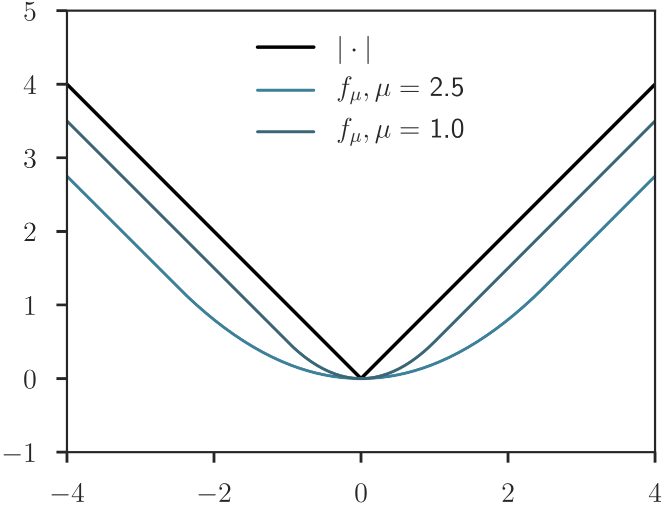
Huber function: $\omega(t) = \frac{t^2}{2}$



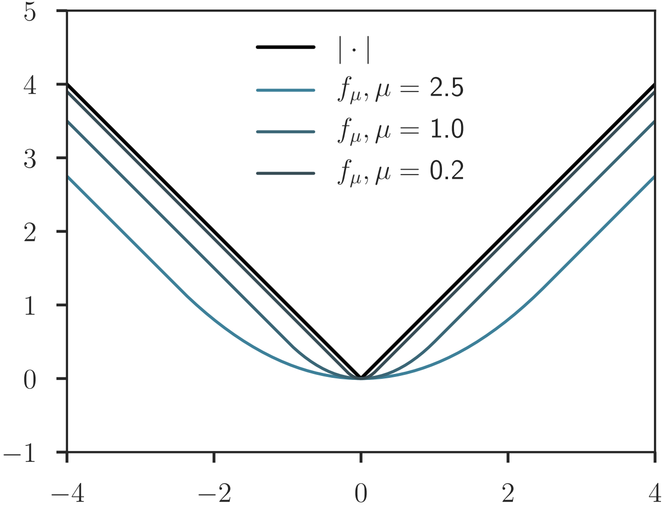
Huber function: $\omega(t) = \frac{t^2}{2}$



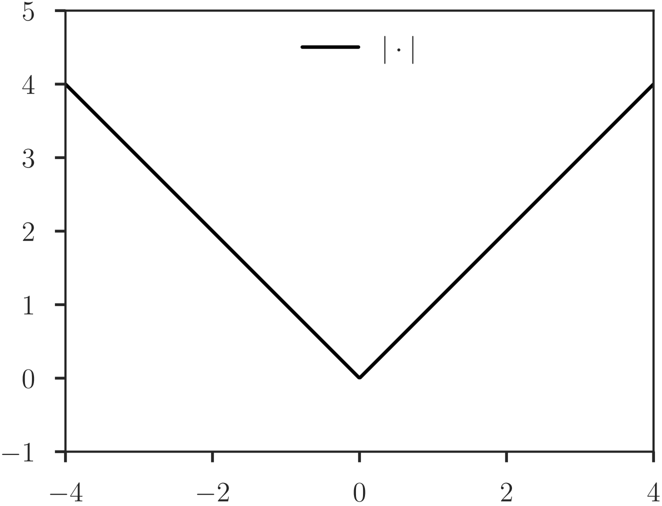
Huber function: $\omega(t) = \frac{t^2}{2}$



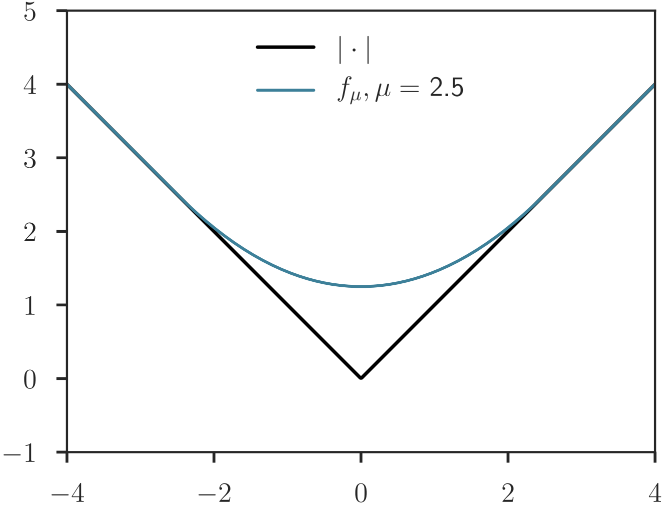
Huber function: $\omega(t) = \frac{t^2}{2}$



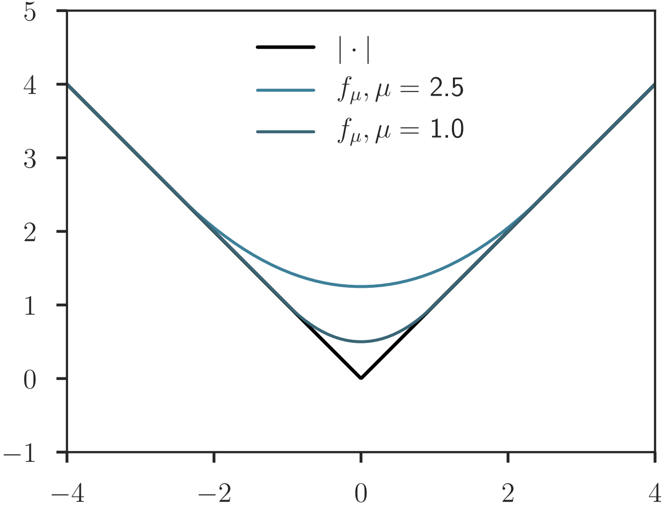
Huber function (bis): $\omega(t) = \frac{t^2}{2} + \frac{1}{2}$



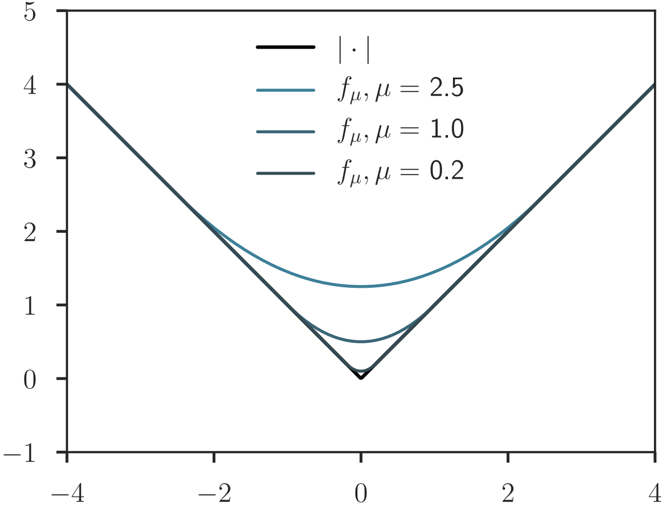
Huber function (bis): $\omega(t) = \frac{t^2}{2} + \frac{1}{2}$



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

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

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

Leads to the Smoothed Concomitant Lasso formulation:

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

Solving the Smooth Concomitant Lasso

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

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

Alternate iteratively:

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

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

Solving the Smooth Concomitant Lasso

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

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

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

- Fix β: closed form solution to update σ

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

Solving the Smooth Concomitant Lasso

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

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

Alternate iteratively:

- Fix σ: (approximatively) solve a Lasso problem to update β
$$\hat{\beta} \leftarrow \underset{\beta \in \mathbb{R}^p}{\arg\min} \frac{\|y - X\beta\|^2}{2n} + \lambda\sigma \|\beta\|_1 \quad \text{(Lasso step)}$$

- Fix β: closed form solution to update σ
$$\hat{\sigma} \leftarrow \max\left(\frac{\|y - X\beta\|}{\sqrt{n}}, \underline{\sigma}\right) \quad \text{(Noise estimation step)}$$

Table of Contents

Calibrating λ and noise level estimation

Multi-task case and noise structure

Experiments

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

General case: $Y \in \mathbb{R}^{n\times T}$, $\mathrm{B} \in \mathbb{R}^{p\times T}$, and the noise $\mathrm{E} \in \mathbb{R}^{n\times T}$ 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:

$$\boxed{(\hat{\mathrm{B}}^{(\lambda)}, \hat{\sigma}^{(\lambda)}) \in \operatorname*{arg\,min}_{\mathrm{B}\in\mathbb{R}^{p\times T}, \sigma \geq \underline{\sigma}} \left(\frac{\|Y - X\mathrm{B}\|_F^2}{2n\sigma} + \frac{\sigma}{2} + \lambda \|\mathrm{B}\|_{2,1}\right)}$$

and similar efficient algorithms.

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

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

"**Huberization of the Frobenius norm**"

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

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

Leads to the Smoothed Concomitant Lasso formulation:

$$\boxed{(\hat{\mathrm{B}}^{(\lambda)}, \hat{\sigma}^{(\lambda)}) \in \mathop{\arg\min}_{\mathrm{B}\in\mathbb{R}^{p\times T}, \sigma \geq \underline{\sigma}} \left(\frac{\|Y - X\mathrm{B}\|_F^2}{2n\sigma} + \frac{\sigma}{2} + \lambda \left\|\mathrm{B}\right\|_{2,1}\right)}$$

and similar efficient algorithms.

What about estimating jointly $\hat{\mathrm{B}}, \hat{S}$?

First: generalizing with the Frobenius norm is not enough (only a scalar counterpart for the noise level)

Concomitant motivation: alternate iteratively

- Fix S: (approximatively) solve a Multi-task Lasso problem to update B
- Fix B: to update S

What about estimating jointly $\hat{\mathrm{B}}, \hat{S}$?

First: generalizing with the Frobenius norm is not enough (only a scalar counterpart for the noise level)

Concomitant motivation: alternate iteratively

- Fix S: (approximatively) solve a Multi-task Lasso problem to update B
- Fix B: to update S

What about other norms ?

Trace norm (Schatten-1 norm, or nuclear norm): $Z \in \mathbb{R}^{n \times T}$

$$\|Z\|_{\mathrm{tr}} = \sum_{i=1}^{n \wedge T} \gamma_i$$

where the γ_i's are the singular values of Z (obtained by SVD)

$$\begin{aligned} \|\cdot\|_{\mathrm{tr}} \,\square\, \omega_{\underline{\sigma}}(Z) &= \begin{cases} \frac{1}{2\underline{\sigma}} \sum_i \gamma_i^2 - (\gamma_i \wedge \underline{\sigma} - \gamma_i)^2, & \text{if } \|Z\|_{s,1} \le \underline{\sigma} \\ \|Z\|_F \,, & \text{if } \|Z\|_{s,1} > \underline{\sigma} \end{cases} \\ &= \min_{S \succeq \underline{\sigma}} \left(\frac{1}{2} \|Z\|_{S^{-1}}^2 + \frac{1}{2} \operatorname{Tr}(S) \right) \end{aligned}$$

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

Smoothing of the nuclear/trace norm

Smoothed Generalized Concomitant Lasso (SGCL)[17]:

$$\boxed{(\hat{\mathrm{B}}^{\mathrm{SGCL}}, \hat{S}^{\mathrm{SGCL}}) \in \mathop{\arg\min}_{\substack{\mathrm{B} \in \mathbb{R}^{p \times T} \\ S \in \mathbb{S}^n_{++}, S \succeq \underline{\sigma}}} \frac{\left\| \bar{Y} - X\mathrm{B} \right\|^2_{S^{-1}}}{2nT} + \frac{\mathrm{Tr}(S)}{2n} + \lambda \left\| \mathrm{B} \right\|_{2,1}}$$

[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.

Smoothing of the nuclear/trace norm

Smoothed Generalized Concomitant Lasso (SGCL)[17]:

$$\boxed{(\hat{\mathrm{B}}^{\mathrm{SGCL}}, \hat{S}^{\mathrm{SGCL}}) \in \mathop{\arg\min}_{\substack{\mathrm{B} \in \mathbb{R}^{p \times T} \\ S \in \mathbb{S}^n_{++}, S \succeq \underline{\sigma}}} \frac{\left\| \bar{Y} - X\mathrm{B} \right\|^2_{S^{-1}}}{2nT} + \frac{\mathrm{Tr}(S)}{2n} + \lambda \left\| \mathrm{B} \right\|_{2,1}}$$

Concomitant Lasso with Repetitions (CLaR)[18]:

$$\boxed{(\hat{\mathrm{B}}^{\mathrm{CLaR}}, \hat{S}^{\mathrm{CLaR}}) \in \mathop{\arg\min}_{\substack{\mathrm{B} \in \mathbb{R}^{p \times T} \\ S \in \mathbb{S}^n_{++}, S \succeq \underline{\sigma}}} \frac{\sum_{l=1}^{r} \left\| Y^{(l)} - X\mathrm{B} \right\|^2_{S^{-1}}}{2nTr} + \frac{\mathrm{Tr}(S)}{2n} + \lambda \left\| \mathrm{B} \right\|_{2,1}}$$

[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: $Y^{(l)} \in \mathbb{R}^{n \times T}$, $\mathrm{B} \in \mathbb{R}^{p \times T}$, and the noise $\mathrm{E}^{(l)} \in \mathbb{R}^{n \times T}$

[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-Four, 2015, École d'Été de Probabilités de Saint-Flour. Springer, 2016, pp. xiii+274.

Efficient solvers for SGCL and CLaR

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

SGCL:

$$(\hat{\mathrm{B}}^{\mathrm{SGCL}}, \hat{S}^{\mathrm{SGCL}}) \in \mathop{\arg\min}_{\substack{\mathrm{B}\in\mathbb{R}^{p\times T} \\ S\in\mathbb{S}^n_{++}, S\succeq \underline{\sigma}}} \frac{\left\| \bar{Y} - X\mathrm{B} \right\|^2_{S^{-1}}}{2nT} + \frac{\mathrm{Tr}(S)}{2n} + \lambda \left\| \mathrm{B} \right\|_{2,1}$$

CLaR:

$$(\hat{\mathrm{B}}^{\mathrm{CLaR}}, \hat{S}^{\mathrm{CLaR}}) \in \mathop{\arg\min}_{\substack{\mathrm{B}\in\mathbb{R}^{p\times T} \\ S\in\mathbb{S}^n_{++}, S\succeq \underline{\sigma}}} \frac{\sum_{l=1}^{r} \left\| Y^{(l)} - X\mathrm{B} \right\|^2_{S^{-1}}}{2nTr} + \frac{\mathrm{Tr}(S)}{2n} + \lambda \left\| \mathrm{B} \right\|_{2,1}$$

with $\|Z\|^2_{S^{-1}} := \mathrm{Tr}(Z^\top S^{-1} Z)$ (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-Four, 2015, École d'Été de Probabilités de Saint-Flour. Springer, 2016, pp. xiii+274.

Efficient solvers for SGCL and CLaR

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

SGCL:

$$(\hat{\mathrm{B}}^{\mathrm{SGCL}}, \hat{S}^{\mathrm{SGCL}}) \in \mathop{\arg\min}_{\substack{\mathrm{B} \in \mathbb{R}^{p \times T} \\ S \in \mathbb{S}^n_{++}, S \succeq \underline{\sigma}}} \frac{\left\| \bar{Y} - X\mathrm{B} \right\|^2_{S^{-1}}}{2nT} + \frac{\mathrm{Tr}(S)}{2n} + \lambda \left\| \mathrm{B} \right\|_{2,1}$$

CLaR:

$$(\hat{\mathrm{B}}^{\mathrm{CLaR}}, \hat{S}^{\mathrm{CLaR}}) \in \mathop{\arg\min}_{\substack{\mathrm{B} \in \mathbb{R}^{p \times T} \\ S \in \mathbb{S}^n_{++}, S \succeq \underline{\sigma}}} \frac{\sum_{l=1}^{r} \left\| Y^{(l)} - X\mathrm{B} \right\|^2_{S^{-1}}}{2nTr} + \frac{\mathrm{Tr}(S)}{2n} + \lambda \left\| \mathrm{B} \right\|_{2,1}$$

with $\|Z\|^2_{S^{-1}} := \mathrm{Tr}(Z^\top S^{-1} Z)$ (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-Four, 2015, École d'Été de Probabilités de Saint-Flour. Springer, 2016, pp. xiii+274.

SGCL and CLaR computations: B update

Jointly convex: alternate minimization converges

B **Update (**S **fixed)**: "smooth + non-smooth" optimization

(Block) Coordinate Descent (Soft-Threshold.) : 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: alternate minimization converges

B **Update (**S **fixed)**: "smooth + non-smooth" optimization

(Block) Coordinate Descent (Soft-Threshold.) : 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

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

SGCL and CLaR computations: S update

S **Update (**B **fixed)**:

For SGCL and CLaR the problem can be reformulated as

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

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

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

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

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

SGCL and CLaR computations: S update

S **Update (**B **fixed)**:

For SGCL and CLaR the problem can be reformulated as

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

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

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

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

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 nT observations (need T large w.r.t. n)
  - CLaR case: only nTr observations

- Computationally: S update cost is $\mathcal{O}(n^3)$ too slow in general (SVD computation)
  Rem: fine for MEG/EEG problems ($n \approx 300$)

Rem: 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 nT observations (need T large w.r.t. n)
  - CLaR case: only nTr observations

- Computationally: S update cost is $\mathcal{O}(n^3)$ too slow in general (SVD computation)
  Rem: fine for MEG/EEG problems ($n \approx 300$)

Rem: 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

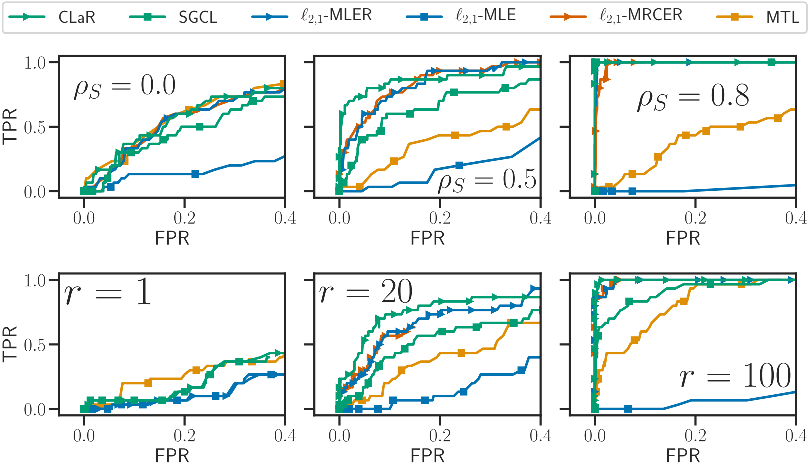
Calibrating λ and noise level estimation

Multi-task case and noise structure

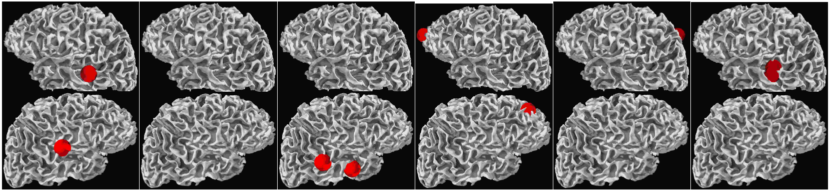
Experiments

Simulated scenarios - ROC curves w.r.t. λ

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



Real data



(a) CLaR (ours) (b) SGCL (c) MLER (d) MLE (e) MRCER (f) MTL

Figure: *Real data, left auditory stimulations (*$n = 102$*,* $p = 7498$*,* $T = 76$*,* $r = 63$*)* 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 $\|\hat{\mathrm{B}}\|_{2,0} = 2$
- 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)
- On-going 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)
- On-going 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

[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

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

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

$$(\hat{\mathrm{B}}, \hat{\Sigma}) \in \operatorname*{arg\,min}_{\substack{\mathrm{B} \in \mathbb{R}^{p \times T} \\ \Sigma \succeq \underline{\sigma}^2}} \sum_{1}^{r} \left\| Y^{(l)} - X\mathrm{B} \right\|_{\Sigma^{-1}}^2 - \log\det(\Sigma^{-1}) + \lambda \left\| \mathrm{B} \right\|_{2,1} \quad .$$

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