The smoothed multivariate square-root Lasso:
an optimization lens on concomitant estimation

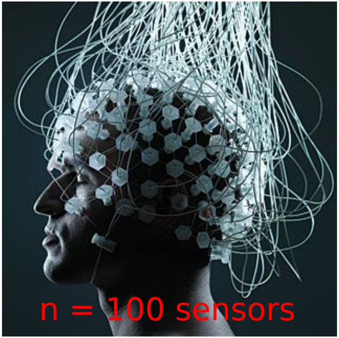
Joseph Salmon
http://josephsalmon.eu
IMAG, Univ. Montpellier, CNRS

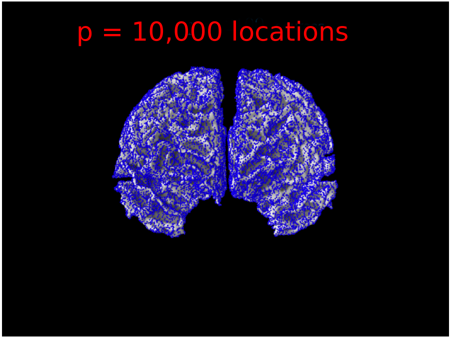
Series of works with:
Quentin Bertrand (INRIA)
Mathurin Massias (University of Genova)
Olivier Fercoq (Institut Polytechnique de Paris)
Alexandre Gramfort (INRIA)

Table of Contents

The M/EEG inverse problem

- observe magnetoelectric field outside the scalp (100 sensors)
- reconstruct cerebral activity inside the brain (10,000 locations)



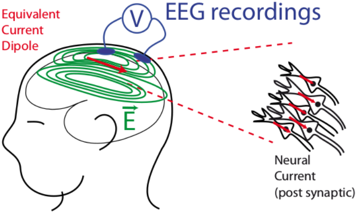


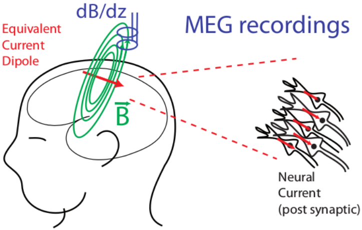
$n \ll p$: ill-posed problem

- **Motivation**: identify brain regions responsible for the signals
- **Applications**: epilepsy treatment, brain aging, anesthesia risks

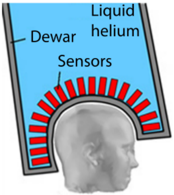
M/EEG inverse problem for brain imaging

- sensors: electric and magnetic fields during a cognitive task





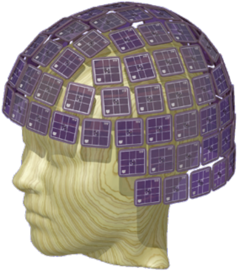
First EEG
recordings
in 1929
by H. Berger



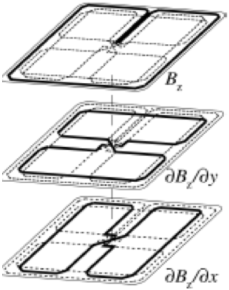
Hôpital La Timone
Marseille, France

MEG elements: magnometers and gradiometers

Device



Sensors



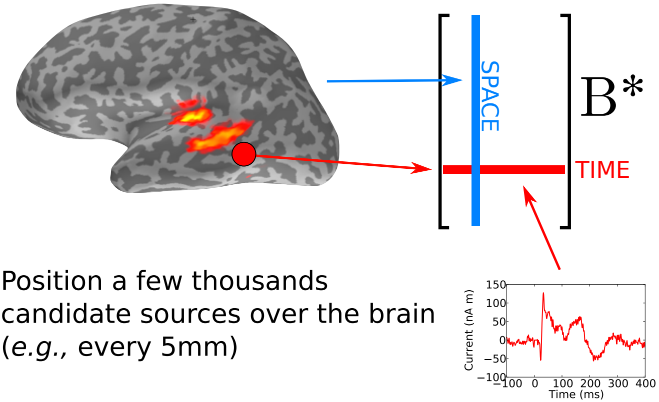
Detail of a sensor

M/EEG = MEG + EEG

Photo Credit: Stephen Whitmarsh

Table of Contents

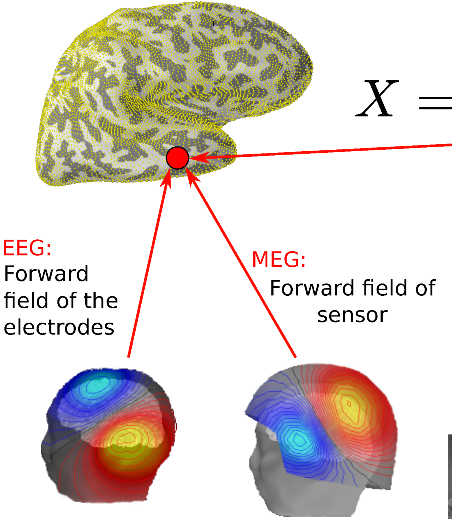
Source modeling



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

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

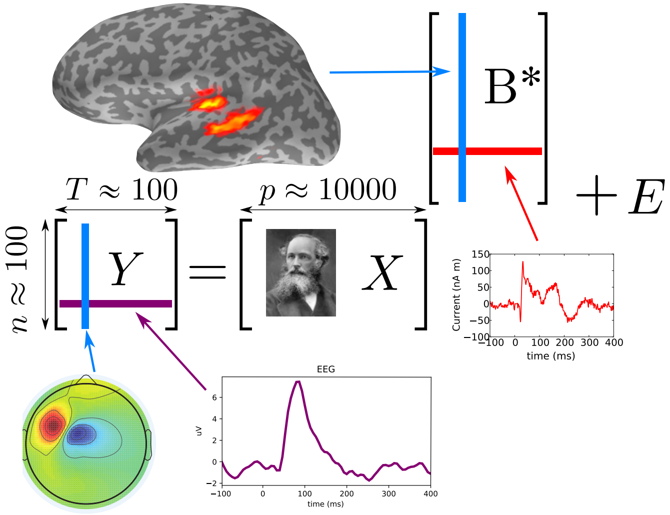
Design matrix - Forward operator



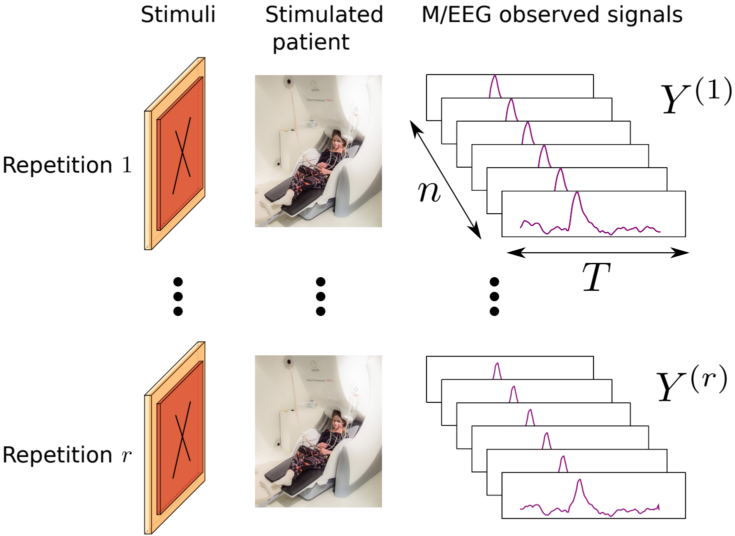
$$X = \begin{bmatrix} X_{\text{EEG}} \\ \hdashline X_{\text{MEG}} \end{bmatrix} \in \mathbb{R}^{n \times p}$$

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

Mathematical model: linear regression

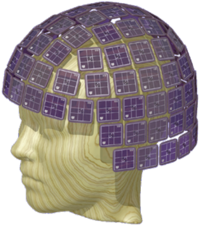


Experiments repeated r times

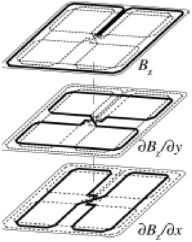


M/EEG specifity #1: combined measurements

Device



Sensors

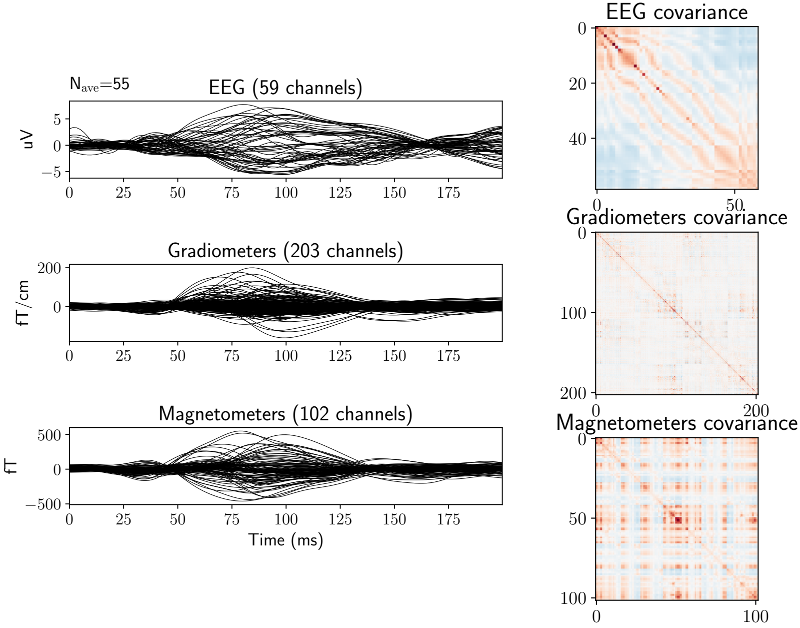


Sensor detail

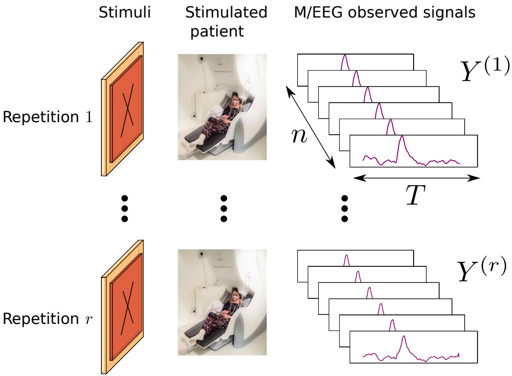
Structure of Y and X:

$$\begin{pmatrix} X_{\text{EEG}} \\ \hdashline X_{\text{grad}} \\ \hdashline X_{\text{mag}} \end{pmatrix} \qquad \begin{pmatrix} Y_{\text{EEG}} \\ \hdashline Y_{\text{grad}} \\ \hdashline Y_{\text{mag}} \end{pmatrix}$$

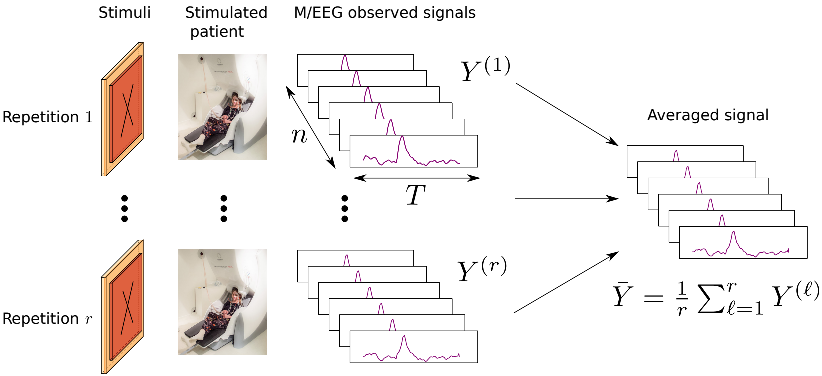
Sensor types & noise structure



M/EEG specificity #2: averaging repetitions of experiment

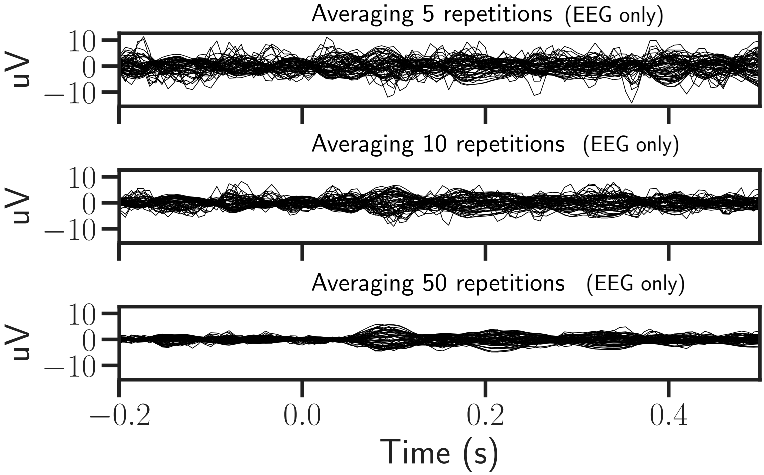


M/EEG specificity #2: averaging repetitions of experiment



$$\bar{Y} = \frac{1}{r}\sum_{\ell=1}^{r} Y^{(\ell)}$$

M/EEG specificity #2: averaged signals



Limit on the repetitions: subject/patient fatigue

A multi-task framework

Multi-task regression notation:

- n observations (number of sensors)
- T tasks (temporal information)
- p features (spatial description)
- r number of repetitions for the experiment
- $Y^{(1)}, \ldots, Y^{(r)} \in \mathbb{R}^{n \times T}$ 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[1] (**unknown**)
- $\mathrm{E}^{(1)}, \ldots, \mathrm{E}^{(r)} \in \mathbb{R}^{n \times T}$: white noise (standard Gaussian)

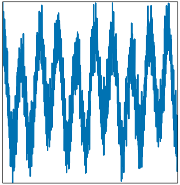
[1] $S \succeq \underline{\sigma}$ means $S - \underline{\sigma}\, \mathrm{Id}_n$ is Semi-Definite Positive

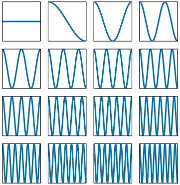
Table of Contents

Sparsity everywhere

Signals can often be represented combining few atoms/features:

- Fourier decomposition for sounds





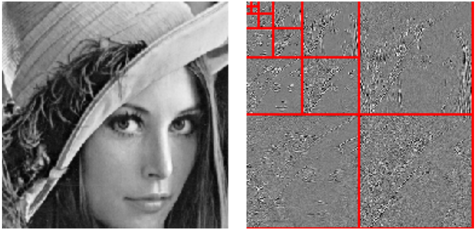
[(2)] I. Daubechies. *Ten lectures on wavelets*. SIAM, 1992.

[(3)] B. A. Olshausen and D. J. Field. "Sparse coding with an overcomplete basis set: A strategy employed by V1?" In: *Vision research* (1997).

Sparsity everywhere

Signals can often be represented combining few atoms/features:

- Fourier decomposition for sounds
- Wavelets for images (1990's)[2]



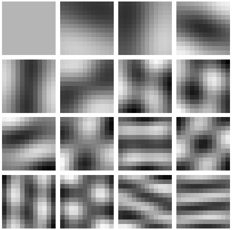
[2] I. Daubechies. *Ten lectures on wavelets*. SIAM, 1992.

[3] B. A. Olshausen and D. J. Field. “Sparse coding with an overcomplete basis set: A strategy employed by V1?” In: *Vision research* (1997).

Sparsity everywhere

Signals can often be represented combining few atoms/features:

- Fourier decomposition for sounds
- Wavelets for images (1990's)[2]
- Dictionary learning for images (2000's)[3]



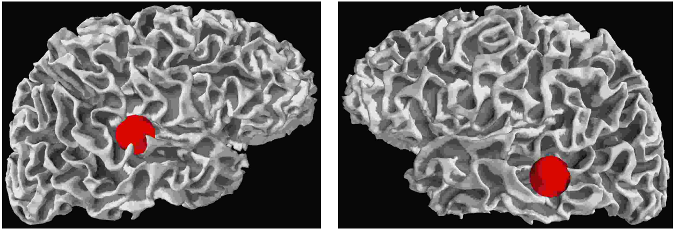
[2] I. Daubechies. *Ten lectures on wavelets*. SIAM, 1992.

[3] B. A. Olshausen and D. J. Field. "Sparse coding with an overcomplete basis set: A strategy employed by V1?" In: *Vision research* (1997).

Sparsity everywhere

Signals can often be represented combining few atoms/features:

- Fourier decomposition for sounds
- Wavelets for images (1990's)[2]
- Dictionary learning for images (2000's)[3]
- Neuroimaging: measurements assumed to be explained by a few active brain sources



---

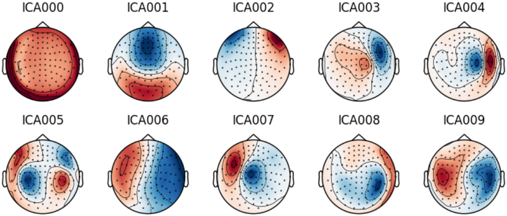
(2) I. Daubechies. *Ten lectures on wavelets*. SIAM, 1992.

(3) B. A. Olshausen and D. J. Field. "Sparse coding with an overcomplete basis set: A strategy employed by V1?" In: *Vision research* (1997).

Justification for dipolarity assumption

Sparsity holds: dipolar patterns equivalent to focal sources

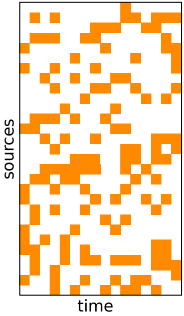
- short duration
- simple cognitive task
- repetitions of experiment average out other sources
- ICA recovers dipolar patterns,[4] well modeled by focal sources:



[4] A. Delorme et al. "Independent EEG sources are dipolar". In: *PloS one* 7.2 (2012), e30135.

(Structured) Sparsity inducing penalties[5]

$$\hat{\mathrm{B}} \in \operatorname*{arg\,min}_{\mathrm{B} \in \mathbb{R}^{p \times T}} \left(\frac{1}{2nT} \|Y - X\mathrm{B}\|_F^2 + \lambda \|\mathrm{B}\|_1 \right)$$



Sparse support: no structure ✗

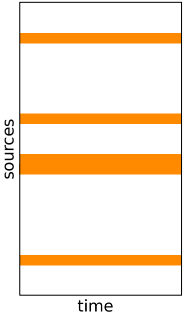
Lasso penalty

$$\|\mathrm{B}\|_1 \triangleq \sum_{j=1}^{p} \sum_{t=1}^{T} |\mathrm{B}_{jt}|$$

[5] 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.

(Structured) Sparsity inducing penalties[5]

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



Sparse support: group structure ✓

Group-Lasso penalty

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

with $\mathrm{B}_{j:}$, j-th row of B

[5] 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.

Data-fitting term and experiment repetitions

- Classical estimator: use averaged[6] signal $\bar{Y}$

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

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

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

[6] & whitened, say using baseline data

Data-fitting term and experiment repetitions

- Classical estimator: use averaged[6] signal $\bar{Y}$

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

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

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

- Fail: $\hat{\mathrm{B}}^{\text{repet}} = \hat{\mathrm{B}}$ (because of datafit $\|\cdot\|_F^2$)

[6] & whitened, say using baseline data

Data-fitting term and experiment repetitions

- Classical estimator: use averaged[(6)] signal $\bar{Y}$

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

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

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

- Fail: $\hat{\mathrm{B}}^{\text{repet}} = \hat{\mathrm{B}}$ (because of datafit $\|\cdot\|_F^2$)

$\hookrightarrow$ investigate other datafits

(6) & whitened, say using baseline data

Table of Contents

Lasso[7],[8]: the "modern least-squares"[9]

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

- $y \in \mathbb{R}^n$: observations
- $X \in \mathbb{R}^{n \times p}$: design matrix
- **sparsity**: for λ large enough, $\|\hat{\beta}\|_0 \ll p$

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

[8] S. S. Chen and D. L. Donoho. "Atomic decomposition by basis pursuit". In: *SPIE*. 1995.

[9] 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.

Lasso and optimal $\lambda^{(10),(11)}$

Theorem

For $y = X\beta^* + \sigma_* \varepsilon$, $\varepsilon \sim \mathcal{N}(0, \mathrm{Id}_n)$ and X satisfying the "Restricted Eigenvalue" property, if $\lambda = 2\sigma_* \sqrt{\frac{2\log(p/\delta)}{n}}$, then

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

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

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

BUT σ_* **is unknown in practice !**

[10] 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.

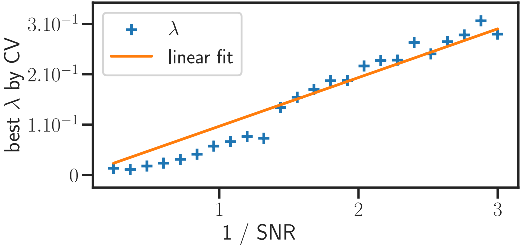
[11] A. S. Dalalyan, M. Hebiri, and J. Lederer. "On the Prediction Performance of the Lasso". In: *Bernoulli* 23.1 (2017), pp. 552–581.

Other datafit: the $\sqrt{\text{Lasso}}$[12]

$$\hat{\beta}_{\text{Lasso}} \in \underset{\beta \in \mathbb{R}^p}{\arg\min} \left(\frac{1}{2n} \|y - X\beta\|^2 + \lambda \|\beta\|_1 \right)$$

optimal $\lambda \propto \sigma_*$

Confirmed in practice:



Lasso

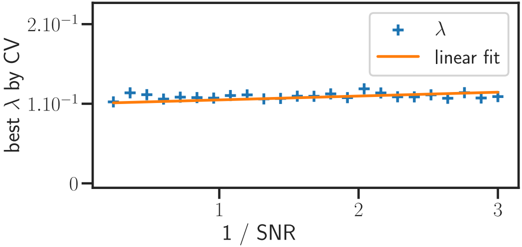
[12] A. Belloni, V. Chernozhukov, and L. Wang. "Square-root Lasso: pivotal recovery of sparse signals via conic programming". In: *Biometrika* 98.4 (2011), pp. 791–806.

Other datafit: the $\sqrt{\text{Lasso}}$[12]

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

optimal λ adaptive to σ_*

Confirmed in practice:



Square-root Lasso

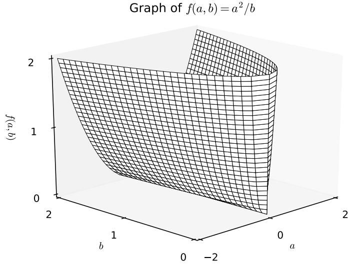
[12] A. Belloni, V. Chernozhukov, and L. Wang. "Square-root Lasso: pivotal recovery of sparse signals via conic programming". In: *Biometrika* 98.4 (2011), pp. 791–806.

Unhappy optimizer

$\sqrt{\text{Lasso}}$: non-smooth+non-smooth $\hookrightarrow$ use *Concomitant Lasso*[13]:

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

same solutions when $\|y - X\hat{\beta}_{\sqrt{\text{Lasso}}}\| \neq 0$, but **jointly convex**,
non smooth + separable: solvable by alternate min.[14] in β and σ



---

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

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

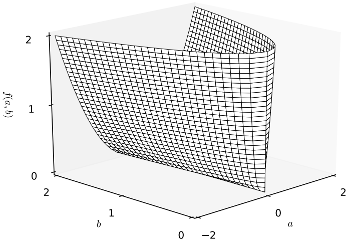
Unhappy optimizer

$\sqrt{\text{Lasso}}$: non-smooth+non-smooth $\hookrightarrow$ use *Concomitant Lasso*[13]:

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

same solutions when $\|y - X\hat{\beta}_{\sqrt{\text{Lasso}}}\| \neq 0$, but **jointly convex**, smooth + separable: solvable by alternate min.[14] in β and σ

Graph of $f(a, b) = a^2/b$

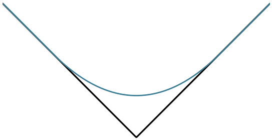


[13] A. B. Owen. "A robust hybrid of lasso and ridge regression". In: *Contemporary Mathematics* 443 (2007), pp. 59–72.

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

"Concomitant": smoothing the $\sqrt{\text{Lasso}}$[17]

"**Huberization**":
replace $\frac{\|\cdot\|}{\sqrt{n}}$ by a smooth approximation



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

$$= \min_{\sigma \geq \underline{\sigma}} \left(\frac{\|z\|^2}{2n\sigma} + \frac{\sigma}{2} \right) = \frac{1}{\sqrt{n}} \|\cdot\| \,\square\, \left(\tfrac{1}{2n\underline{\sigma}} \|\cdot\|^2 + \tfrac{\underline{\sigma}}{2} \right)(z)$$

Leads to the Smoothed[15],[16] Concomitant Lasso formulation:

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

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

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

[17] 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[18],[19]

Smoothing: for $\sigma > 0$, a "smoothed" version of f is f_σ

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

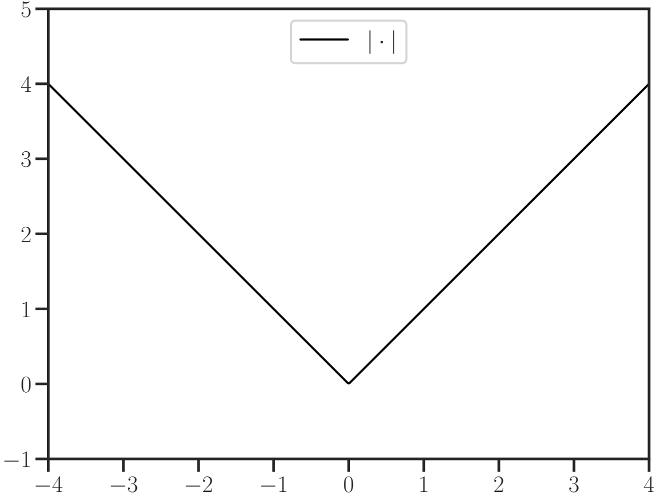
- ω is a predefined smooth function (s.t. $\nabla\omega$ is Lipschitz)

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

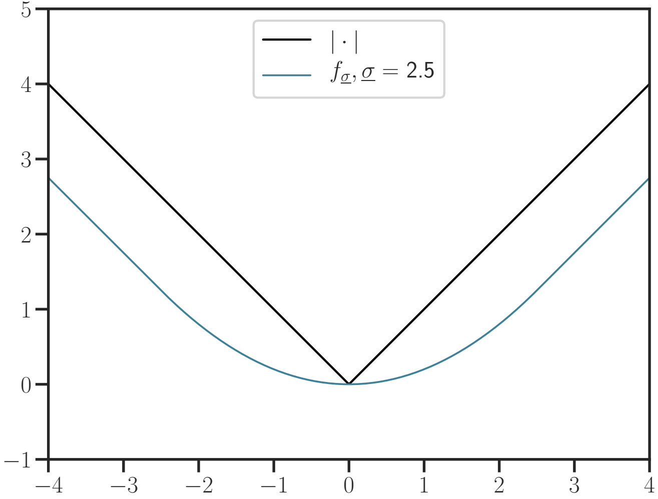
[18] Y. Nesterov. "Smooth minimization of non-smooth functions". In: *Math. Program.* 103.1 (2005), pp. 127–152.

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

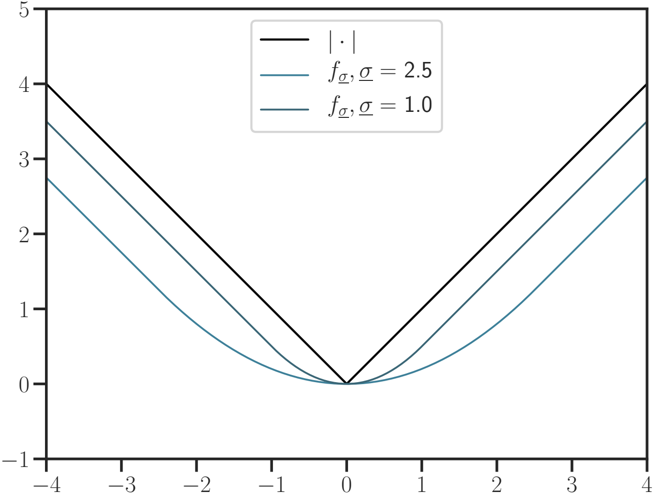
Huber function: $\omega(t) = \frac{t^2}{2}$



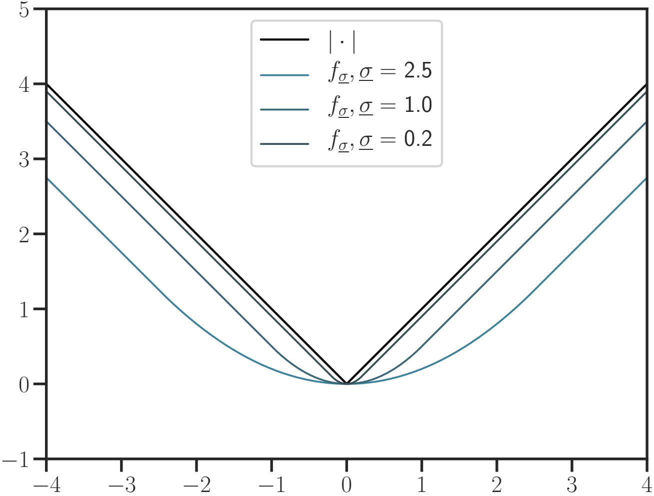
Huber function: $\omega(t) = \frac{t^2}{2}$



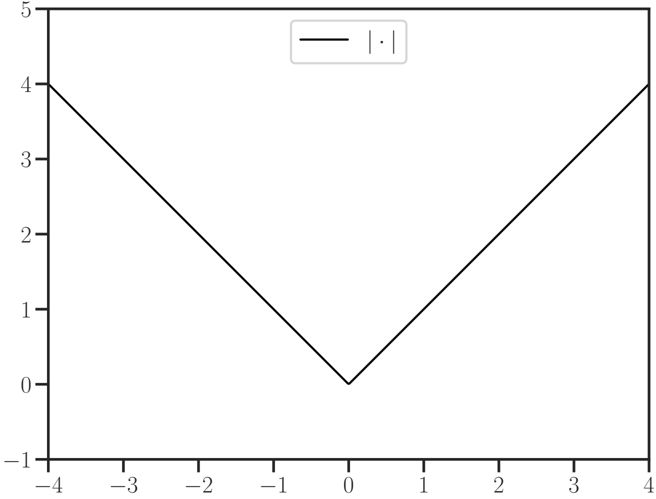
Huber function: $\omega(t) = \frac{t^2}{2}$



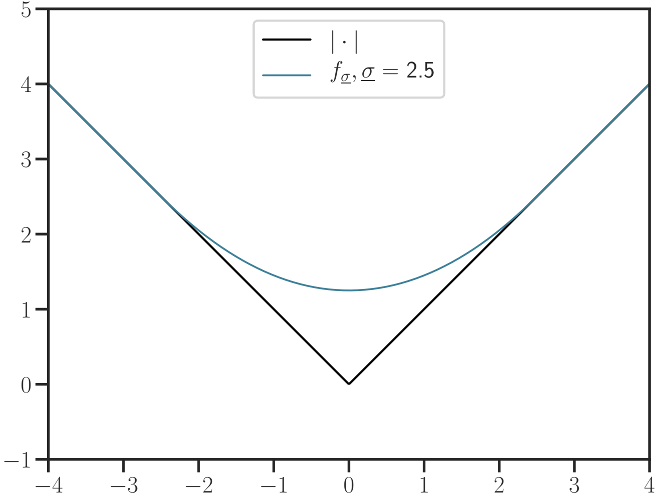
Huber function: $\omega(t) = \frac{t^2}{2}$



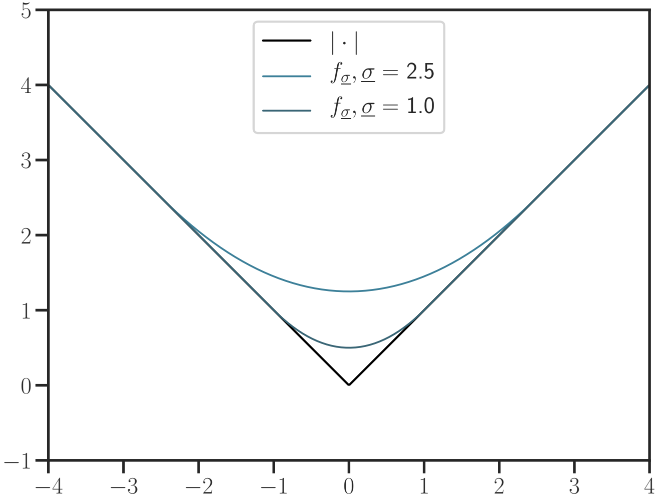
Huber function (bis): $\omega(t) = \frac{t^2}{2} + \frac{1}{2}$



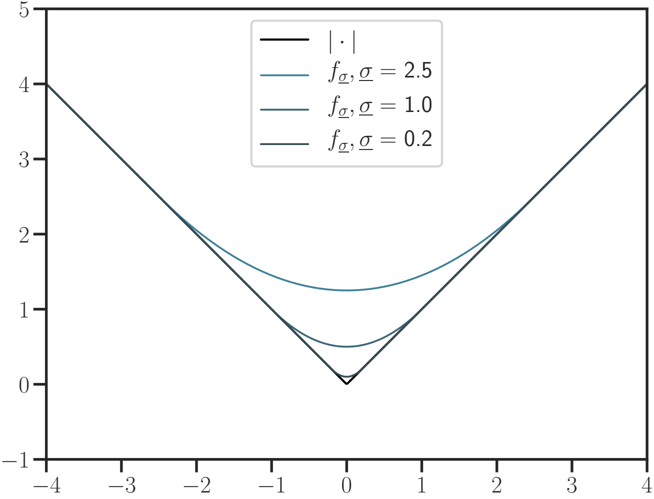
Huber function (bis): $\omega(t) = \frac{t^2}{2} + \frac{1}{2}$



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



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



Smoothing other norms

- Smoothing Frobenius norm yields a trivial gen. of conco Lasso

- More interesting: S. van de Geer introduced the pivotal *multivariate* $\sqrt{\text{Lasso}}$,[20] using trace/nuclear norm for data-fitting

$$\operatorname*{arg\,min}_{\mathrm{B} \in \mathbb{R}^{p \times T}} \frac{1}{n\sqrt{T}} \|Y - X\mathrm{B}\|_* + \lambda \|\mathrm{B}\|_{2,1}$$

  hard to solve, statistical analysis makes stringent assumptions

- Smoothing the datafit makes optim. and stats easier!

[20] S. van de Geer. *Estimation and testing under sparsity*. École d'Été de Probabilités de Saint-Flour. 2016.

Smoothing the nuclear norm[21]

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

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

where the γ_i's are the singular values of Z

$$\begin{aligned} \|\cdot\|_* \,\square \left(\tfrac{1}{2\underline{\sigma}} \|\cdot\|^2 + \tfrac{n}{2}\right)(Z) &= \sum_i \mathrm{huber}_{\underline{\sigma}}(\gamma_i) \\ &= \min_{S \succeq \underline{\sigma}} \left(\tfrac{1}{2} \|Z\|^2_{S^{-1}} + \tfrac{1}{2} \operatorname{Tr}(S) \right) \end{aligned}$$

where $\|Z\|^2_{S^{-1}} \triangleq \operatorname{Tr}(Z^\top S^{-1} Z)$

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

Smoothing of the multivariate $\sqrt{\text{Lasso}}$

Smoothed Generalized Concomitant Lasso (SGCL)[22]:

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

Concomitant Lasso with Repetitions (CLaR)[23]:

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

[22] M. Massias et al. "Generalized concomitant multi-task Lasso for sparse multimodal regression". In: *AISTATS*. vol. 84. 2018, pp. 998–1007.

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

Simulations : row support identification

- $n = 150$, $p = 500$, $T = 100$
- X Toeplitz-correlated
- S^* Toeplitz matrix: ${S^*}_{i,j} = \rho_{S^*}^{|i-j|}$, $\rho_{S^*} \in]0, 1[$

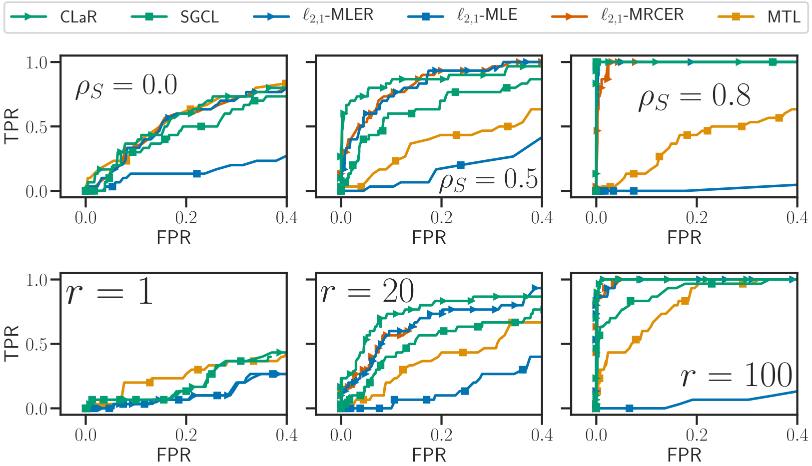


Table of Contents

SGCL and CLaR: alternate updates

Alternate minimization converges

B **update (**S **fixed)**: standard Multi-task Lasso optimization, off-the-shelf techniques and lots of refinements

S **update (**B **fixed)**:

$$\underset{S \succeq \underline{\sigma}}{\arg\min} \left(\frac{1}{2n} \mathrm{Tr}[Z^\top S^{-1} Z] + \frac{1}{2n} \mathrm{Tr}(S) \right)$$

closed-form solution : clipped sqrt of eigen value decomposition of $\frac{1}{T}(\bar{Y} - X\mathrm{B})(\bar{Y} - X\mathrm{B})^\top$ or $\frac{1}{rT}\sum_{l=1}^{r}(Y^{(l)} - X\mathrm{B})(Y^{(l)} - X\mathrm{B})^\top$

Rem: see online Python code https://github.com/QB3/CLaR

Algorithm: Concomitant Lasso w. Repetitions (CLaR)

```
input : X ∈ R^{n×p}, Y^(1), ..., Y^(r) ∈ R^{n×T}, σ > 0, λ > 0
init  : B = 0_{p,q}, R = Ȳ
for iter = 1, ..., do
    S ← SpectralClipping( (1/(Tr)) Σ_l^r (Y^(l) − XB)(Y^(l) − XB)^⊤, σ )
      // closed-form sol. of min. in S: EVD + clipping sqrt of eigenvalues at level σ
    for j = 1, ..., p do
        L_j = X_{:j}^⊤ S^{-1} X_{:j}                              // Lipschitz constants
    for j = 1, ..., p do
        R ← R + X_{:j} B_{j:}                                     // partial residual update
        B_{j:} ← BST( X_{:j}^⊤ S^{-1} R / L_j, λnT / L_j )        // coef. update
        R ← R − X_{:j} B_{j:}                                     // residual update
return B, S
```

Complexity?

Fine, if we store $S^{-1}X$, and $S^{-1}R$ instead of R.

Need eigenvalue decomposition though $\mathcal{O}(n^3)$ (here $n \approx 100$)

Statistical properties for i.i.d. case[24]

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

Proposition

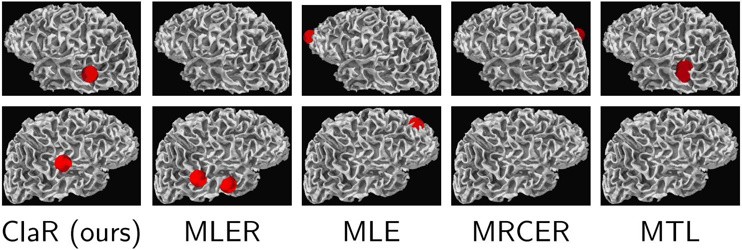
- i.i.d. Gaussian noise
- X satisfying the "mutual incoherence" property
- $\lambda \propto \frac{\sqrt{\log p}}{T\sqrt{n}}$ (independent of σ_*)
- $c_1 \underline{\sigma} \leq \sigma_* \leq c_2 \bar{\sigma}$

$\Longrightarrow$ with probability at least $1 - ne^{-cT/n}$

$$\frac{1}{T}\|\mathrm{B}^* - \hat{\mathrm{B}}\|_{2,\infty} \leq C\sigma_* \frac{1}{T}\sqrt{\frac{\log p}{n}}$$

[24] M. Massias et al. "Support recovery and sup-norm convergence rates for sparse pivotal regression". In: *AISTATS*. 2020.

Real data experiments



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

Links

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

A. Gramfort

- Papers: arXiv / personal webpage[25],[26],[27]
- CLaR Python code https://github.com/QB3/CLaR

[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.

[27] M. Massias et al. "Support recovery and sup-norm convergence rates for sparse pivotal regression". In: *AISTATS*. 2020.

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.
- 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. and D. L. Donoho. "Atomic decomposition by basis pursuit". In: *SPIE*. 1995.
- Dalalyan, A. S., M. Hebiri, and J. Lederer. "On the Prediction Performance of the Lasso". In: *Bernoulli* 23.1 (2017), pp. 552–581.
- Daubechies, I. *Ten lectures on wavelets*. SIAM, 1992.
- Delorme, A. et al. "Independent EEG sources are dipolar". In: *PloS one* 7.2 (2012), e30135.
- Massias, M. et al. "Generalized concomitant multi-task Lasso for sparse multimodal regression". In: *AISTATS*. Vol. 84. 2018, pp. 998–1007.
- Massias, M. et al. "Support recovery and sup-norm convergence rates for sparse pivotal regression". In: *AISTATS*. 2020.
- Ndiaye, E. et al. "Efficient Smoothed Concomitant Lasso Estimation for High Dimensional Regression". In: *Journal of Physics: Conference Series* 904.1 (2017), p. 012006.

References III

- Nesterov, Y. "Smooth minimization of non-smooth functions". In: *M. Prog.* 103.1 (2005), pp. 127–152.
- – ."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.
- Olshausen, B. A. and D. J. Field. "Sparse coding with an overcomplete basis set: A strategy employed by V1?" In: *Vision research* (1997).
- 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. "Regression Shrinkage and Selection via the Lasso". In: *J. R. Stat. Soc. Ser. B Stat. Methodol.* 58.1 (1996), pp. 267–288.
- van de Geer, S. *Estimation and testing under sparsity*. École d'Été de Probabilités de Saint-Flour. 2016.

Statistical assumptions

Gaussian noise: the entries $\mathrm{E}_{i,j}$ are i.i.d. $\mathcal{N}(0, {\sigma_*}^2)$ random variables.

Mutual incoherence: The *Gram matrix* $\Psi \triangleq \frac{1}{n} X^\top X$ satisfies

$$\Psi_{jj} = 1 \ , \ \text{and} \ \max_{j' \neq j} \left| \Psi_{jj'} \right| \leq \tfrac{1}{7\alpha s}, \forall j \in [p] \ ,$$

for some integer $s \geq 1$ and some constant $\alpha > 1$.

Residuals bound: For the multivariate square-root Lasso, $\hat{\mathrm{E}}^\top \hat{\mathrm{E}}$ is invertible, and there exists η such that

$$\|(\tfrac{1}{T}\hat{\mathrm{E}}^\top \hat{\mathrm{E}})^{\frac{1}{2}}\|_2 \leq C\sigma^*$$

Smoothing parameter value: $\underline{\sigma}$, $\bar{\sigma}$ and η verify: $\underline{\sigma} \leq \frac{\sigma^*}{\sqrt{2}}$ and $\bar{\sigma} = (2 + \eta)\sigma^*$ with $\eta \geq 1$.

Competitors

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

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

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

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

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

- MRCER has an additional term $\mu \left\| \Sigma^{-1} \right\|$ *w.r.t.* $\ell_{2,1}$-MLER