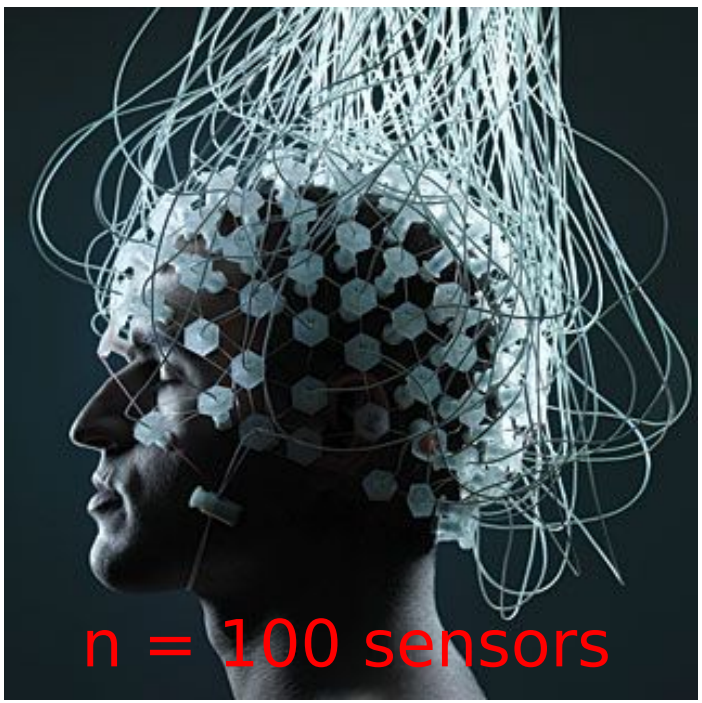
The smoothed multivariate square-root Lasso
(optimizational and statistical handling of correlated noise)

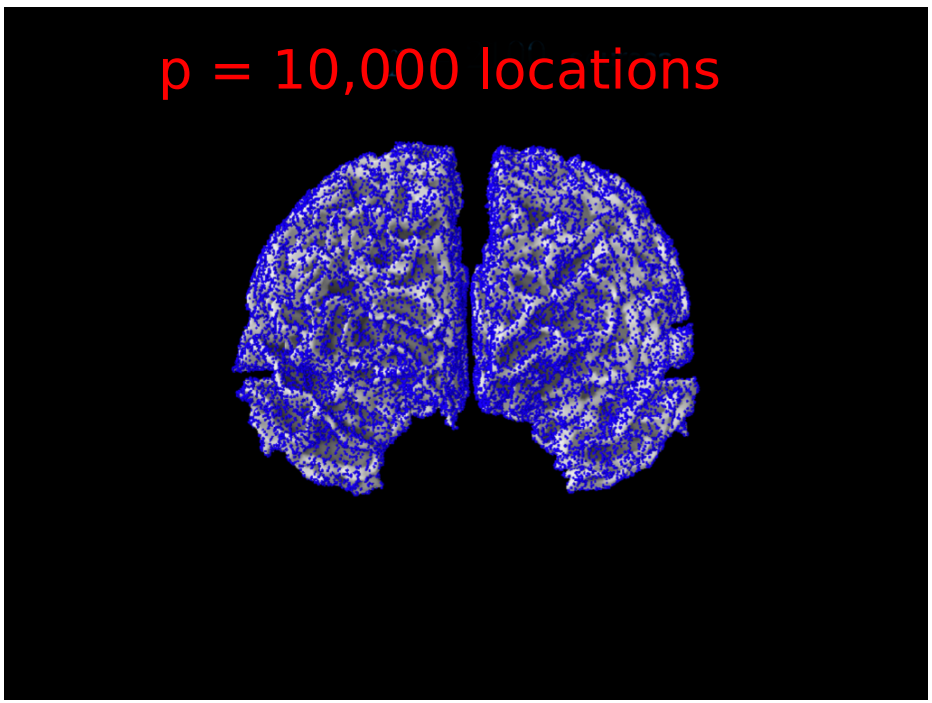
Mathurin Massias
https://mathurinm.github.io

Series of works with:
Quentin Bertrand (INRIA)
Olivier Fercoq (Institut Polytechnique de Paris)
Alexandre Gramfort (INRIA)
Joseph Salmon (IMAG, Univ. Montpellier, CNRS)

The M/EEG inverse problem

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



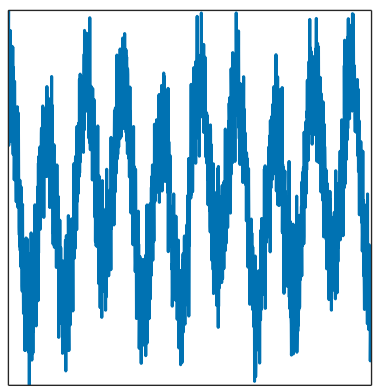


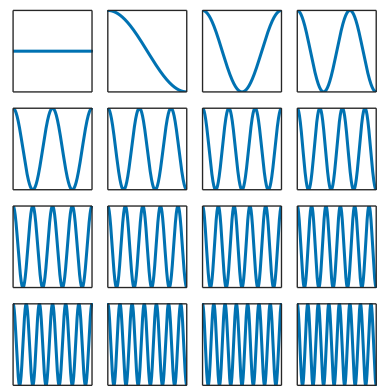
$n \ll p$: ill-posed problem

Sparsity everywhere

Signals can often be represented combining few atoms/features:

- Fourier decomposition for sounds





---

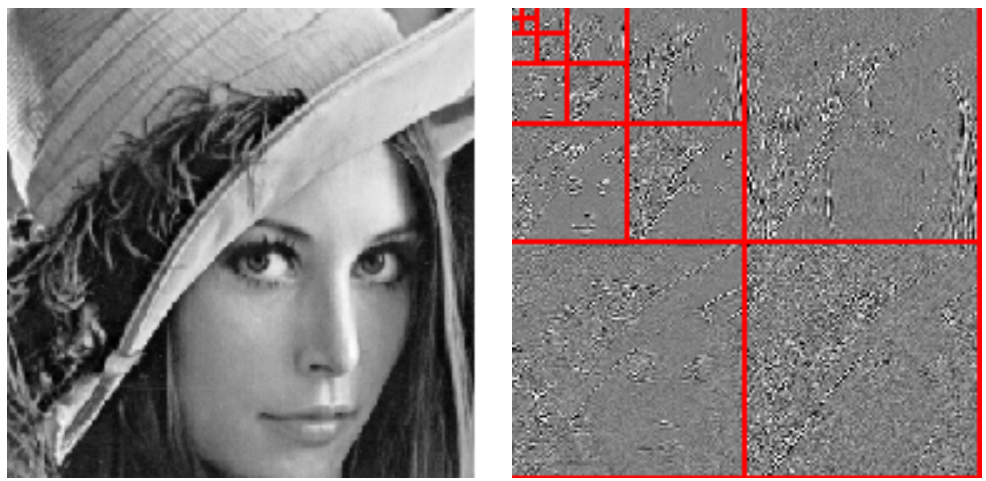
(1) I. Daubechies. *Ten lectures on wavelets*. SIAM, 1992.

(2) 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)[1]



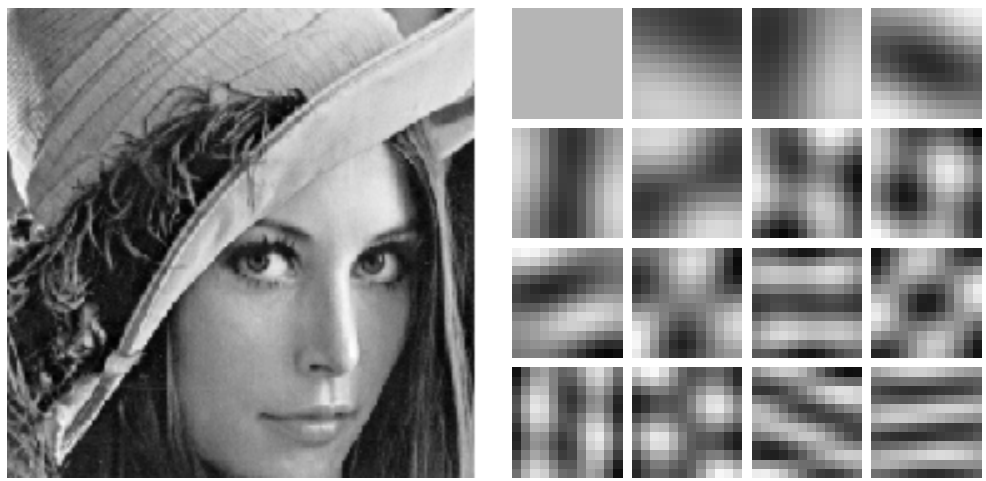
[1] I. Daubechies. *Ten lectures on wavelets*. SIAM, 1992.

[2] 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)[1]
- Dictionary learning for images (2000's)[2]



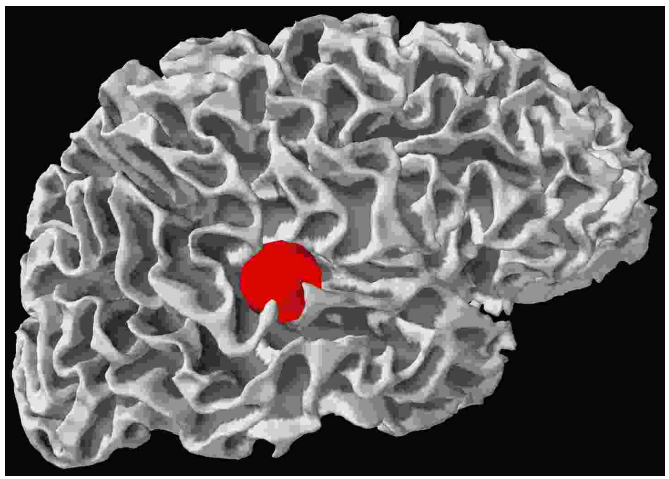
[1] I. Daubechies. *Ten lectures on wavelets*. SIAM, 1992.

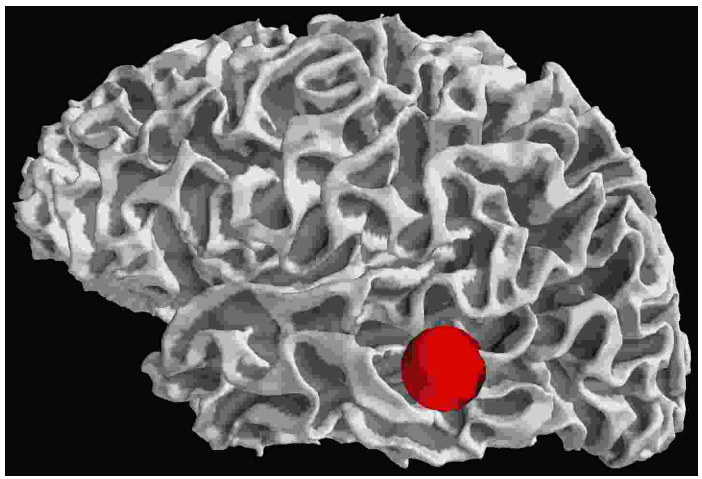
[2] B. A. Olshausen and D. J. Field. "Sparse coding with an overcomplete basis set: A strategy employed by V1?" In: *Vision research* (1997).

Sparsity everywhere

Signals can often be represented combining few atoms/features:

- Fourier decomposition for sounds
- Wavelets for images (1990's)[1]
- Dictionary learning for images (2000's)[2]
- Here we assume that measurements are explained by a few active brain sources



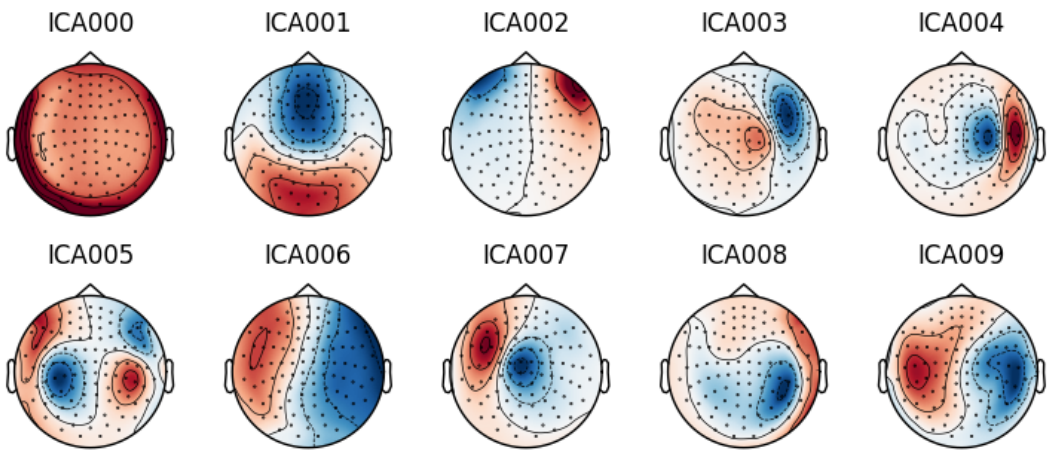


[1] I. Daubechies. *Ten lectures on wavelets*. SIAM, 1992.

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

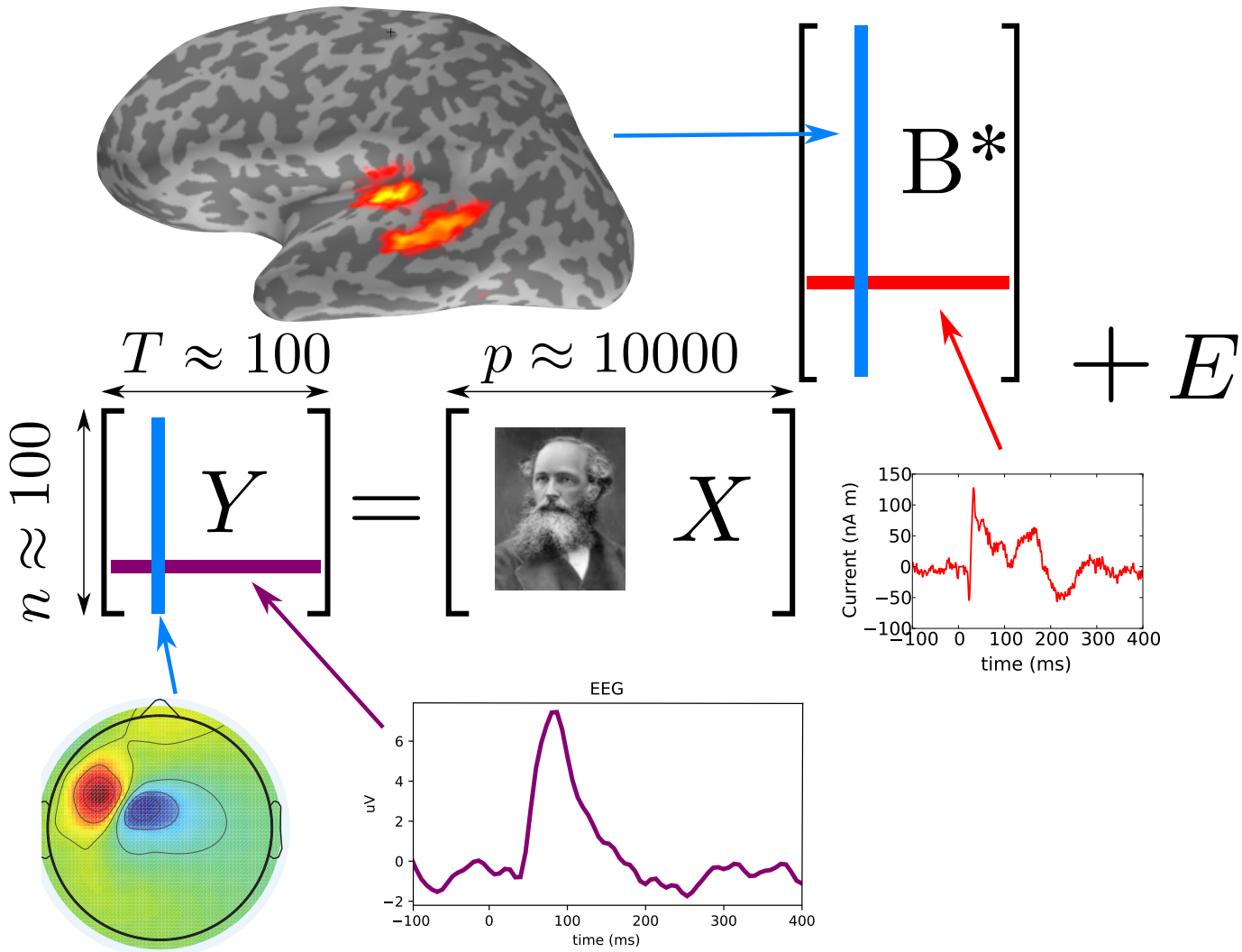
Justification for dipolarity assumption

- short duration
- simple cognitive task
- repetitions of experiment average out other sources
- ICA recovers dipolar patterns,[3] well modelled by focal sources:



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

Mathematical model: linear regression



Lasso[4],[5]: the "modern least-squares"[6]

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

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

---

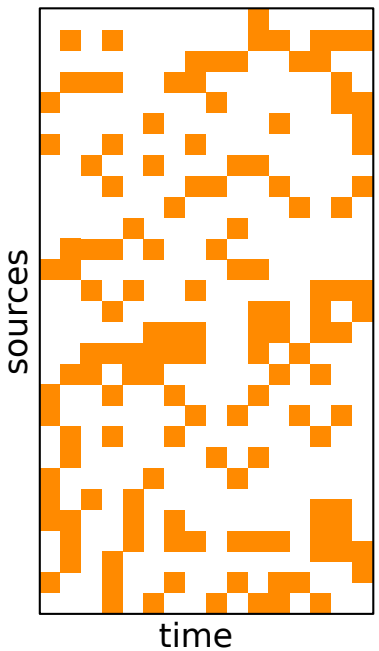
[4] R. Tibshirani. "Regression Shrinkage and Selection via the Lasso". In: *J. R. Stat. Soc. Ser. B Stat. Methodol.* 58.1 (1996), pp. 267–288.

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

[6] 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.

Sparsity inducing penalties[7]

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



Sparse support: no structure ✗

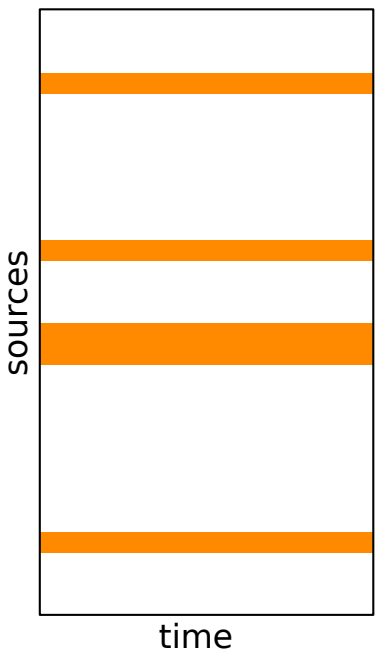
Lasso penalty

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

[7] 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.

Sparsity inducing penalties[7]

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



Sparse support: group structure ✓

Group-Lasso penalty

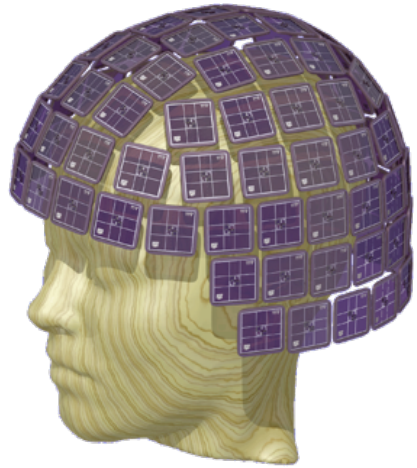
$$\|\mathrm{B}\|_{2,1} \triangleq \sum_{j=1}^{p} \|\mathrm{B}_{j:}\|_2$$

where $\mathrm{B}_{j:} = j$-th row of B

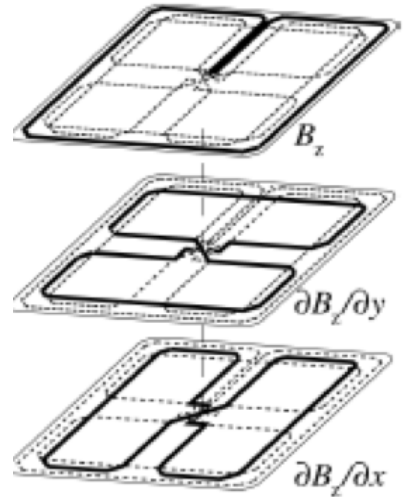
[7] G. Obozinski, B. Taskar, and M. I. Jordan. "Joint covariate selection and joint subspace selection for multiple classification problems". In: *Statistics and Computing* 20.2 (2010), pp. 231–252.

M/EEG specifity #1: combined measurements

Device



Sensors

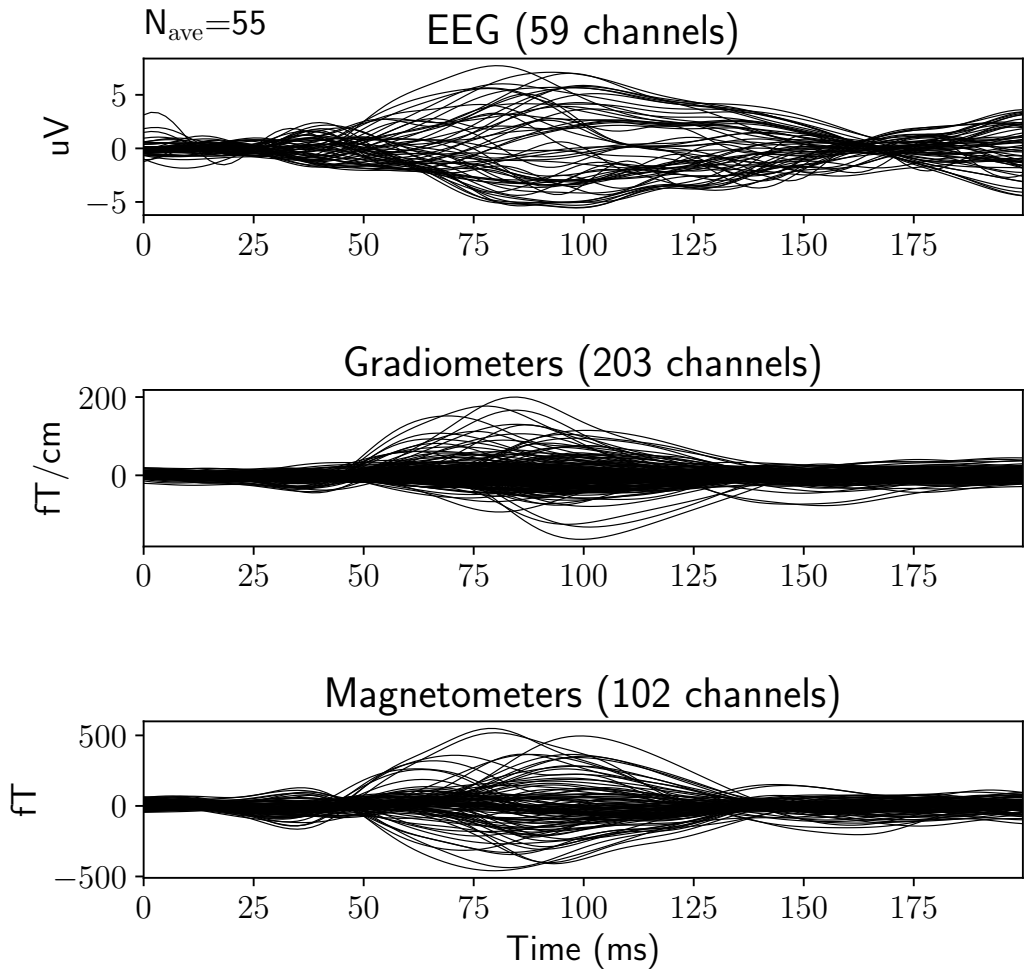


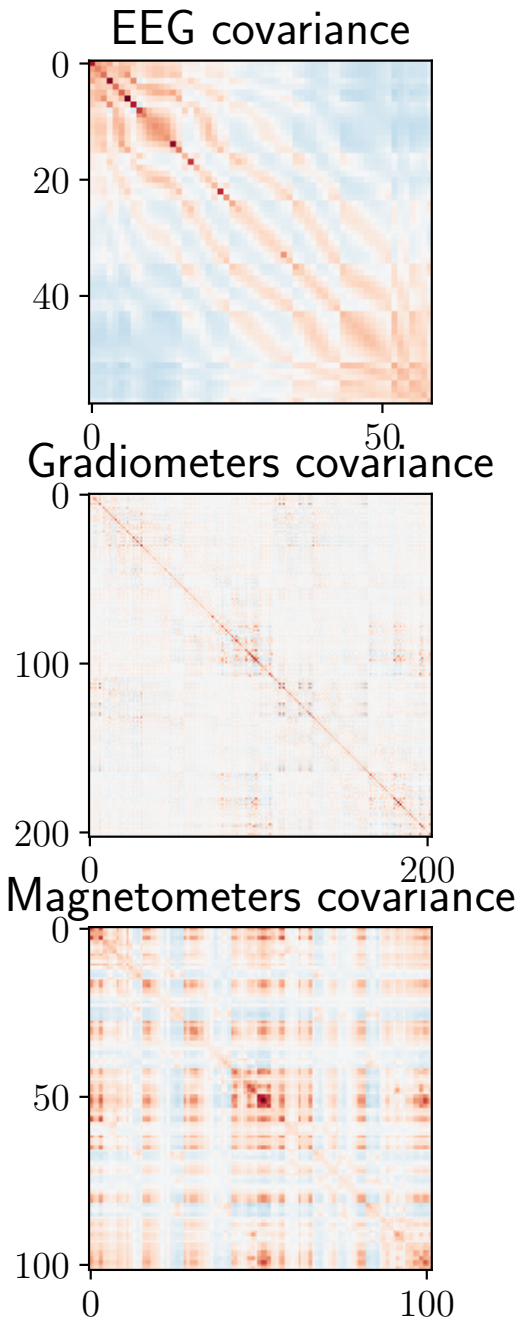
Sensor detail

Structure of Y and X:

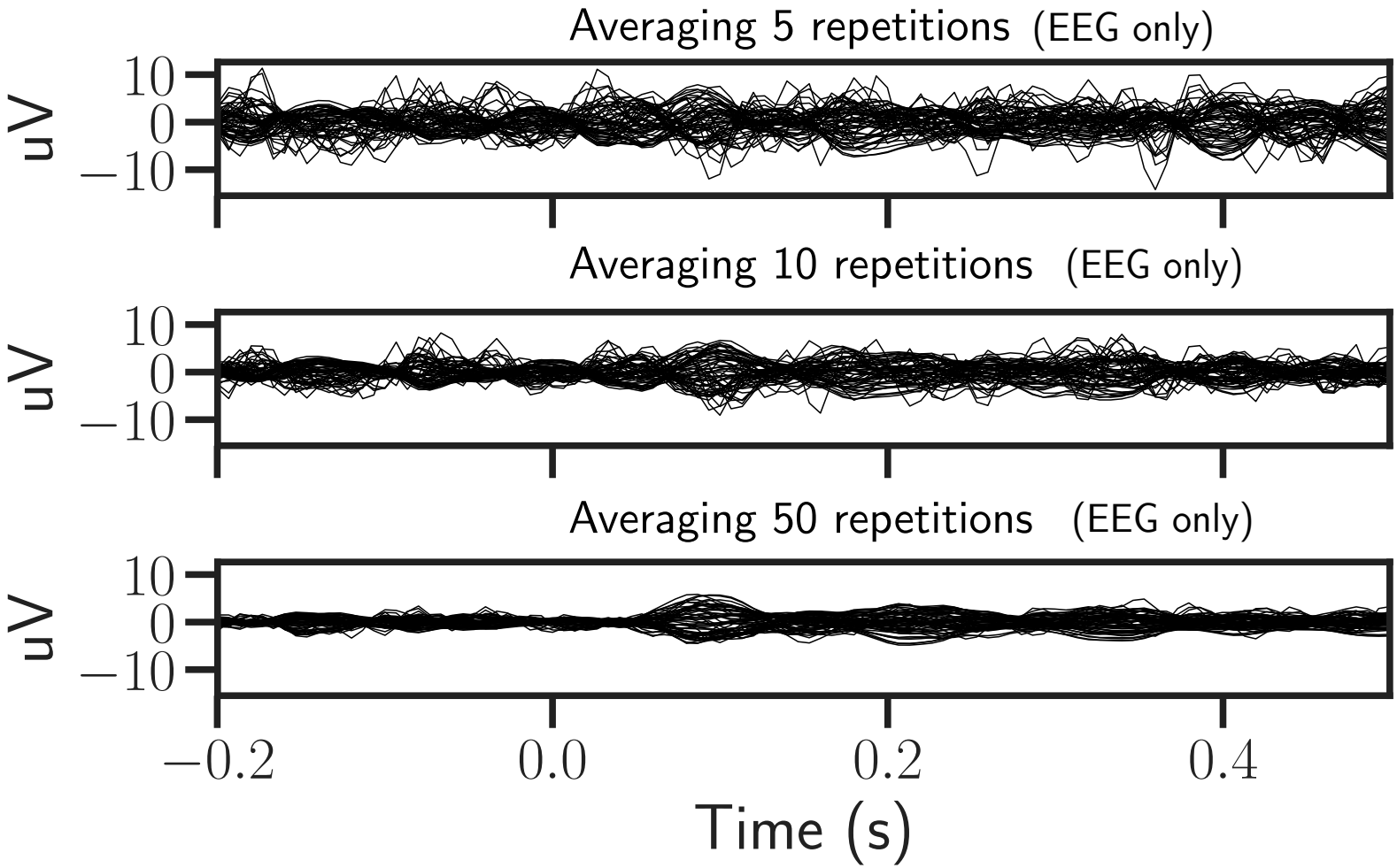
$$\begin{pmatrix} Y_{\text{EEG}} \\ \hdashline Y_{\text{grad}} \\ \hdashline Y_{\text{mag}} \end{pmatrix} \qquad \begin{pmatrix} X_{\text{EEG}} \\ \hdashline X_{\text{grad}} \\ \hdashline X_{\text{mag}} \end{pmatrix}$$

3 sensor types $\Longrightarrow$ 3 noise structures





M/EEG specificity #2: averaging repetitions of experiment



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
- $Y^{(1)}, \dots, Y^{(r)} \in \mathbb{R}^{n \times T}$ 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[8] (**unknown**)
- $\mathrm{E}^{(1)}, \dots, \mathrm{E}^{(r)} \in \mathbb{R}^{n \times T}$: white Gaussian noise

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

Data-fitting term

- Classical estimator: use averaged[(9)] signal $\bar{Y}$

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

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

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

[(9)] & whitened, say using baseline data

Data-fitting term

- Classical estimator: use averaged[9] signal $\bar{Y}$

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

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

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

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

[9] & whitened, say using baseline data

Data-fitting term

- Classical estimator: use averaged[(9)] signal $\bar{Y}$

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

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

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

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

$\hookrightarrow$ investigate other datafits

[(9)] & whitened, say using baseline data

Lasso and optimal λ [10],[11]

Theorem

For $y = X\beta^* + \sigma_* \varepsilon$, ε standard Gaussian 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.

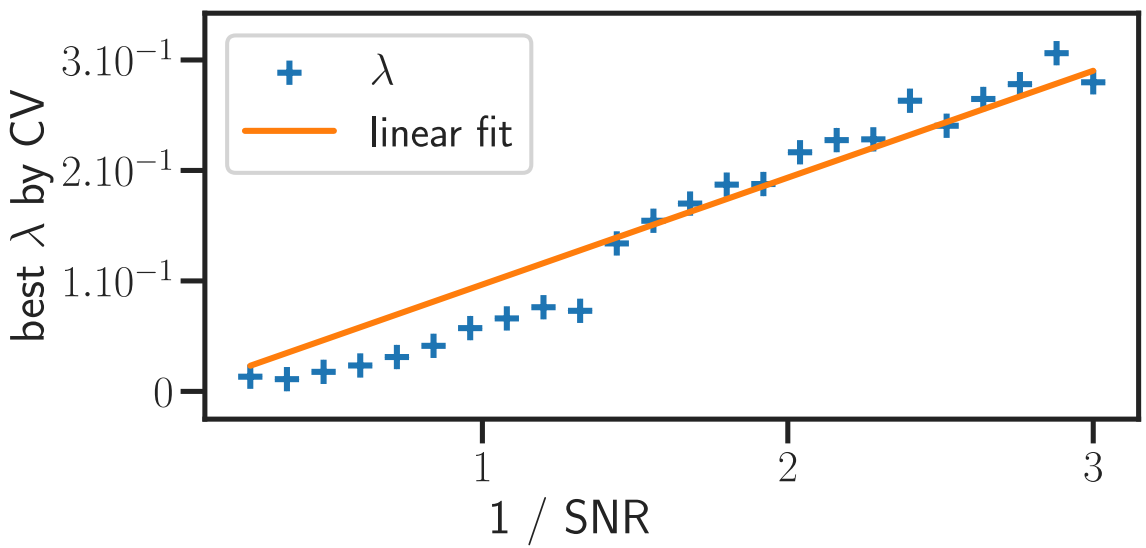
[11] A. S. Dalalyan, M. Hebiri, and J. Lederer. "On the Prediction Performance of the Lasso". In: *Bernoulli* 23.1 (2017), pp. 552–581.

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

Same guarantees, but optimal λ independent of σ_* (*pivotality*):

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

Confirmed in practice:



Lasso

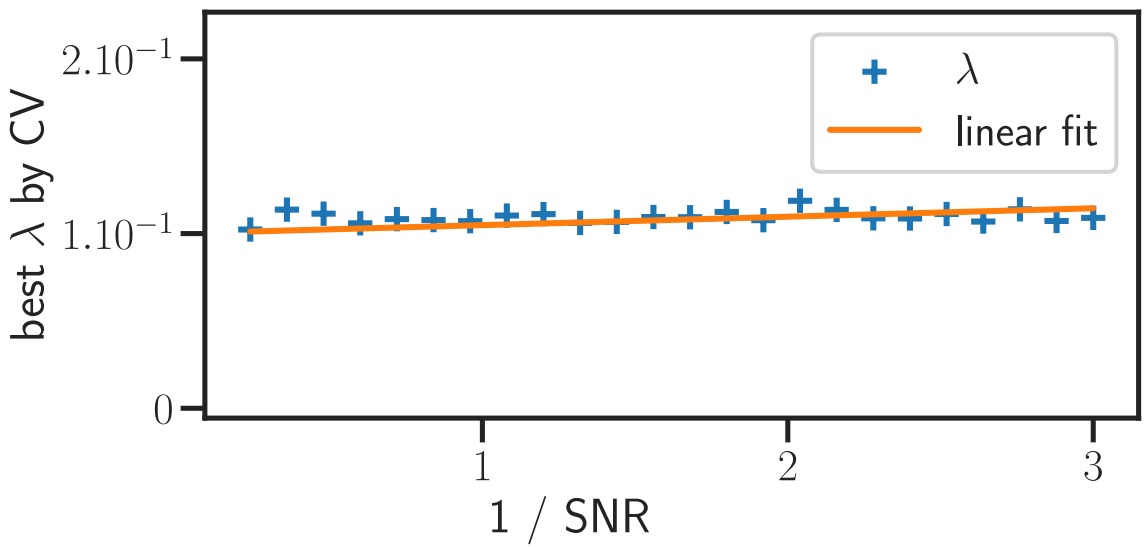
[12] A. Belloni, V. Chernozhukov, and L. Wang. "Square-root Lasso: pivotal recovery of sparse signals via conic programming". In: *Biometrika* 98.4 (2011), pp. 791–806.

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

Same guarantees, but optimal λ independent of σ_* (*pivotality*):

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

Confirmed in practice:



Square-root Lasso

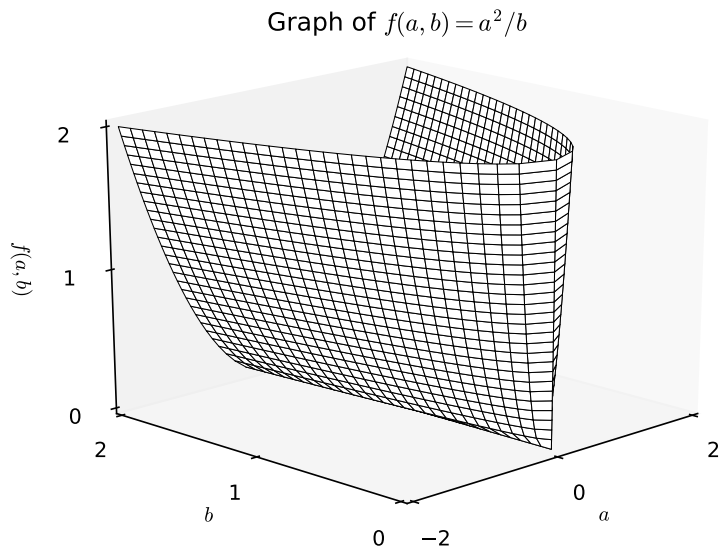
[12] A. Belloni, V. Chernozhukov, and L. Wang. "Square-root Lasso: pivotal recovery of sparse signals via conic programming". In: *Biometrika* 98.4 (2011), pp. 791–806.

Unhappy optimizer

$\sqrt{\text{Lasso}}$ is 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**, smooth + separable: solvable by alternate min. in β and σ



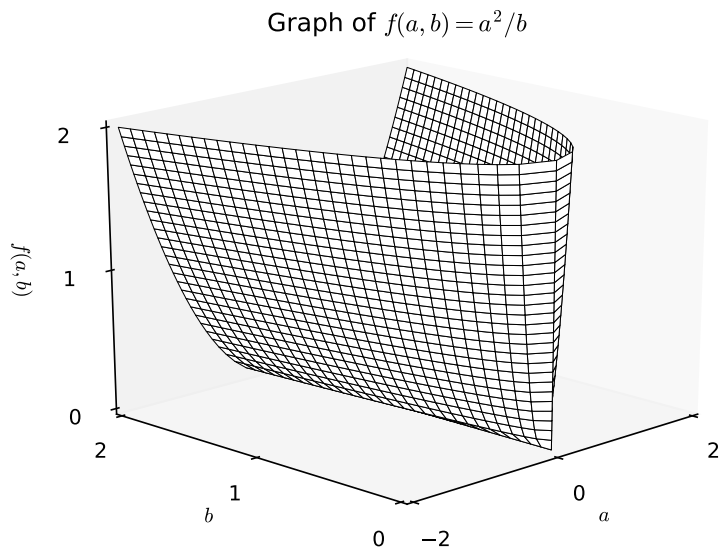
[13] A. B. Owen. "A robust hybrid of lasso and ridge regression". In: *Contemporary Mathematics* 443 (2007), pp. 59–72.

Unhappy optimizer

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

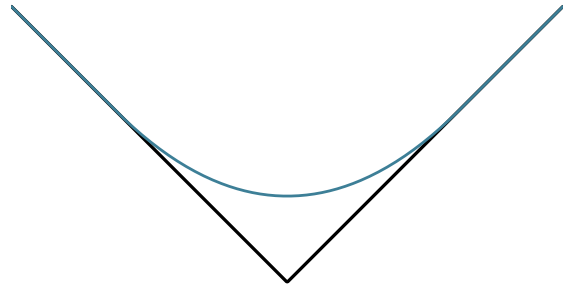
$$(\hat{\beta}, \hat{\sigma}) \in \underset{\beta \in \mathbb{R}^p, \sigma \geq \underline{\sigma}}{\arg\min} \frac{\|y - X\beta\|^2}{2n\sigma} + \frac{\sigma}{2} + \lambda \|\beta\|_1$$

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



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

Concomitant origin: smoothing the $\sqrt{\text{Lasso}}$[16]



"**Huberization**": replace $\|\cdot\|$ by a smooth approximation:

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

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

Leads to the Smoothed[14],[15] Concomitant Lasso formulation:

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

---

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

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

[16] 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 other norms

Smoothing Frobenius norm yields a trivial gen. of conco Lasso

More interesting: S. van de Geer introduced the pivotal *multivariate* $\sqrt{\text{Lasso}}$[17]:

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

hard to solve, statistical analysis makes stringent assumptions

Smoothing the datafit makes optim. and stats easier!

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

Smoothing the nuclear norm[18]

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

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

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

---

[18] 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)[19]:

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

Concomitant Lasso with Repetitions (CLaR)[20]:

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

---

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

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

SGCL and CLaR: alternate updates

Alternate minimization converges

B **update (**S **fixed)**: standard MTL 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 involving clipped EVD of:

$$\frac{1}{T}(\bar{Y} - X\mathrm{B})(\bar{Y} - X\mathrm{B})^\top \text{ or } \frac{1}{rT} \sum_{l=1}^{r} (Y^{(l)} - X\mathrm{B})(Y^{(l)} - X\mathrm{B})^\top$$

Support recovery guarantees[21]

For the multivariate $\sqrt{\text{Lasso}}$/its smoothed version, with Gaussian noise, mutual incoherence (α) and bounded residuals/correct value of smoothing parameter (η).

Let $C = (1 + \frac{16}{7(\alpha-1)})$, $A \geq \sqrt{2}$, and $\lambda = \frac{2\sqrt{2}}{\sqrt{nq}}(1 + A\sqrt{(\log p)/q})$.

There exists $c \geq 1/64$ s.t. with proba $\geq 1 - p^{1-A^2/2} - 2ne^{-cq/n}$,

$$\frac{1}{q}\|\hat{\mathrm{B}} - \mathrm{B}^*\|_{2,\infty} \leq C(3+\eta)\lambda\sigma^*$$

Moreover if

$$\min_{j \in \mathcal{S}^*} \frac{1}{q}\|\mathrm{B}^*_{j:}\|_2 > 2C(3+\eta)\lambda\sigma^*$$

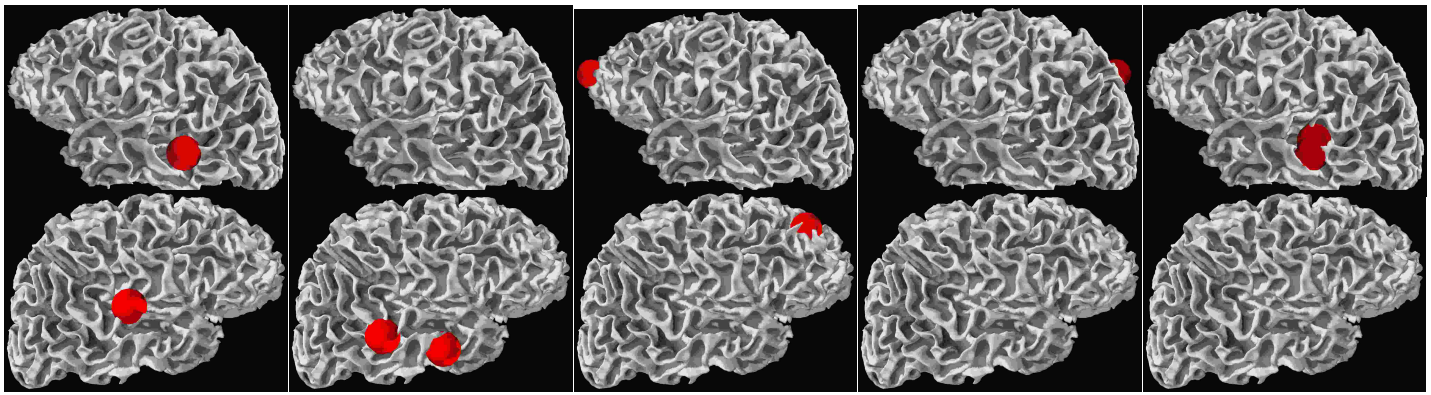
then with the same probability:

$$\mathcal{S}^* = \hat{\mathcal{S}} \triangleq \{j \in [p] : \frac{1}{q}\|\hat{\mathrm{B}}_{j:}\|_2 > C(3+\eta)\lambda\sigma^*\}$$

---

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

Real data experiments



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

Figure: *Left auditory stimulations* ($n = 102$, $p = 7498$, $T = 76$, $r = 63$)
Sources found in the left and right hemispheres

- 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

- Papers: arXiv / personal webpage[22],[23],[24]
- Python code online for CLaR https://github.com/QB3/CLaR

[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.

[24] 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.
- Tibshirani, R. "Regression Shrinkage and Selection via the Lasso". In: *J. R. Stat. Soc. Ser. B Stat. Methodol.* 58.1 (1996), pp. 267–288.

References IV

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

Statistical assumptions

Gaussian noise: the entries of $\mathrm{E}_1, \ldots, \mathrm{E}_n$ 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

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

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

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

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

Alternate minimization

Algorithm: ALTERNATE MIN.

input : $X, Y, \underline{\sigma}, \lambda, f$

init : $\mathrm{B} = 0_{p,q}$, $R = Y$

for iter $= 1, \ldots,$ **do**

  if iter $= 1 \pmod f$ **then**

    $S \leftarrow \mathsf{SpCl}(\frac{1}{T} R R^\top, \underline{\sigma})$ // closed-form sol. of minimization in S

    for $j = 1, \ldots, p$ **do**

      $L_j = X_{:j}^\top S^{-1} X_{:j}$ // Lipschitz constants

  for $j = 1, \ldots, p$ **do**

    $R \leftarrow R + X_{:j} \mathrm{B}_{j:}$ // partial residual update

    $\mathrm{B}_{j:} \leftarrow \mathrm{BST}\Big(X_{:j}^\top S^{-1} R / L_j, \lambda n T / L_j\Big)$ // coef. update

    $R \leftarrow R - X_{:j} \mathrm{B}_{j:}$ // residual update

return B, S

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

Smoothing aparté[25],[26]

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

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

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

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):

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

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