Sparse high dimensional linear regression in the presence of heteroscedastic noise: applications to the M/EEG inverse problem

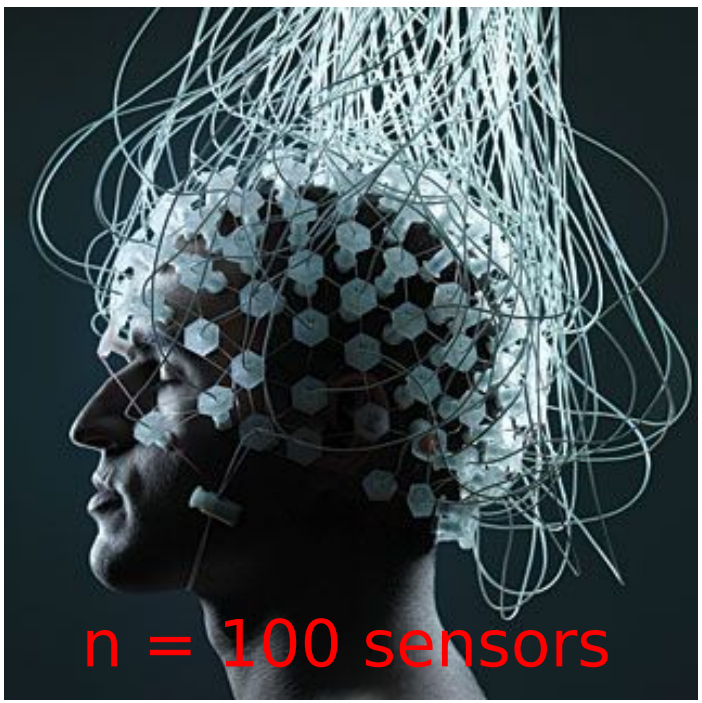
PGMO PhD award

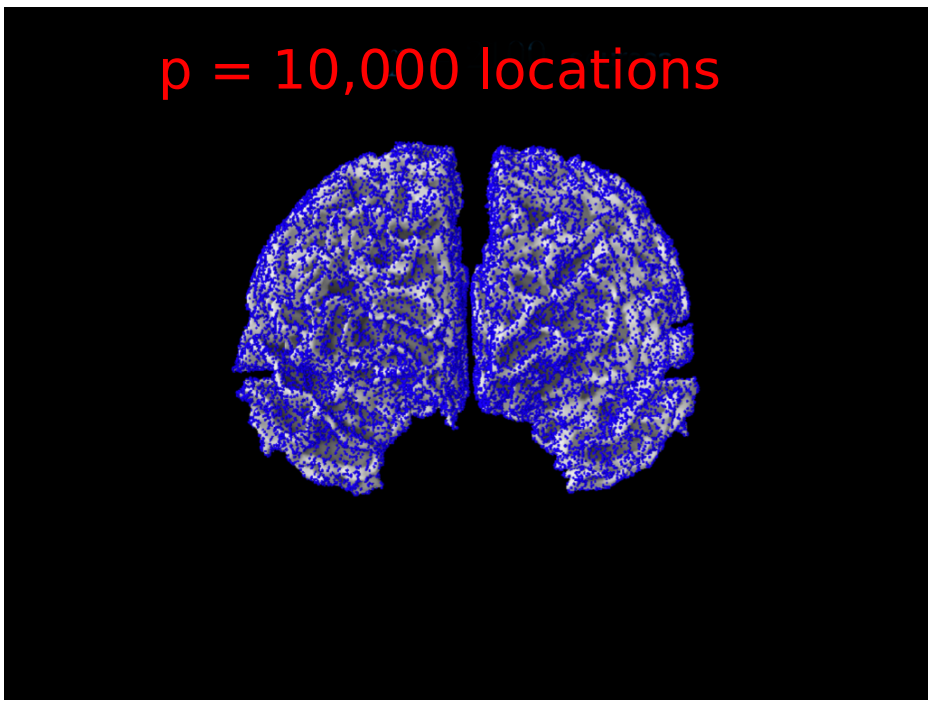
Mathurin Massias (Università di Genova)

prepared at INRIA and Télécom Paris
under the supervision of A. Gramfort and J. Salmon

The M/EEG inverse problem

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



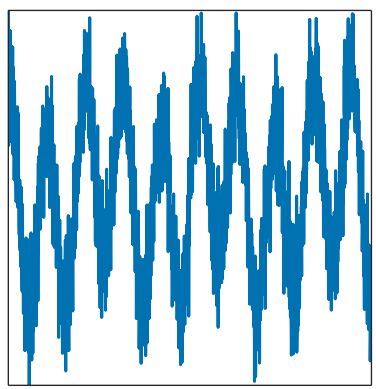


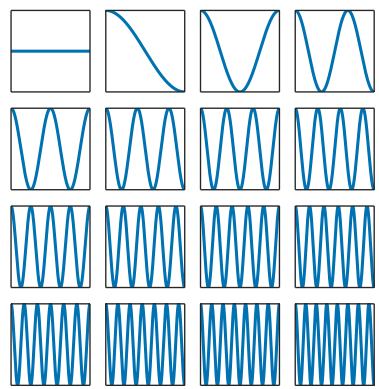
$n \ll p$: ill-posed problem!

Sparsity everywhere

Signals can often be represented combining few atoms/features:

- Fourier decomposition for sounds





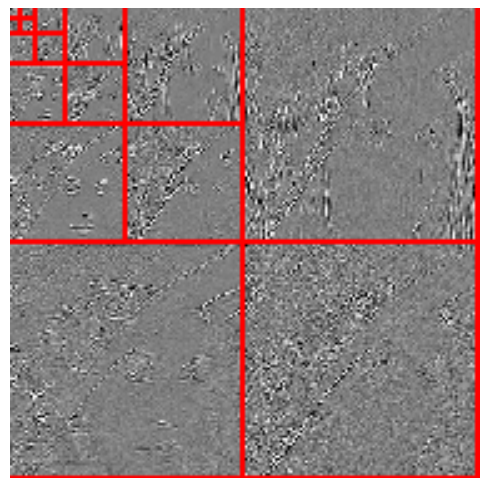
[1] I. Daubechies. *Ten lectures on wavelets*. SIAM, 1992.

[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]



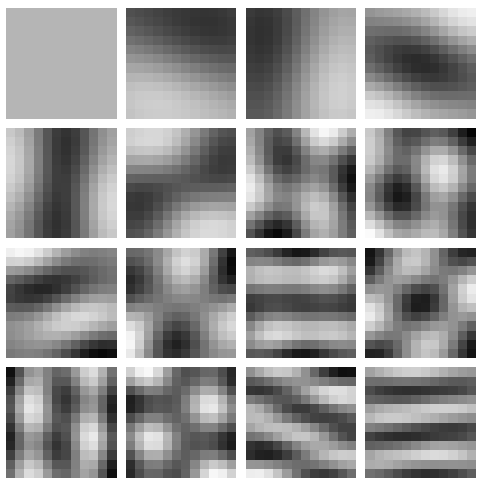
[1] I. Daubechies. *Ten lectures on wavelets*. SIAM, 1992.

[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]



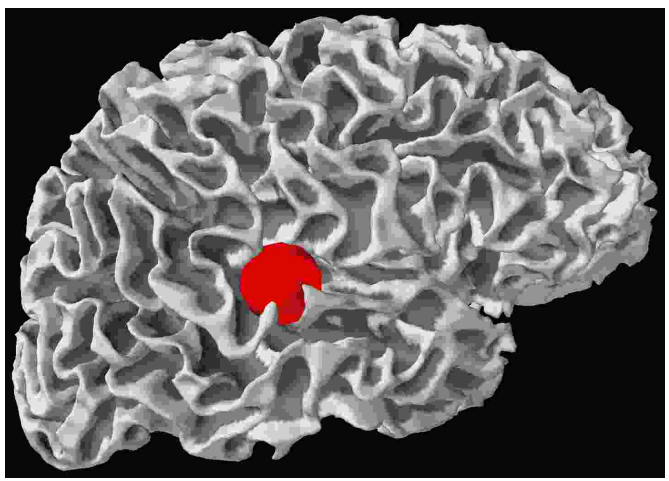
[1] I. Daubechies. *Ten lectures on wavelets*. SIAM, 1992.

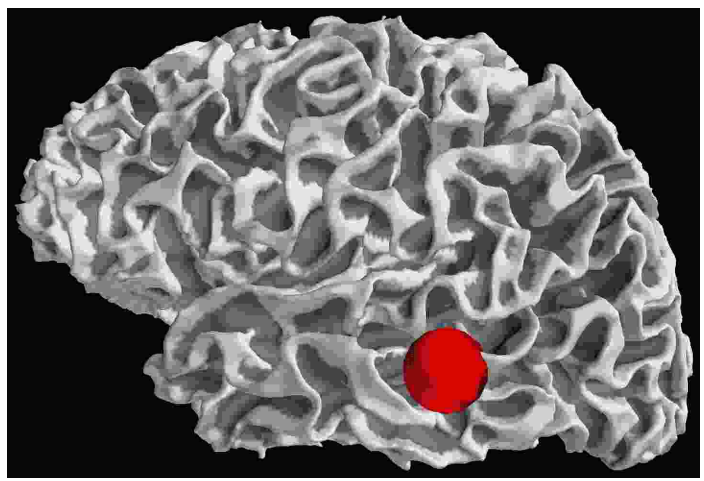
[2] B. A. Olshausen and D. J. Field. "Sparse coding with an overcomplete basis set: A strategy employed by V1?" In: *Vision research* (1997).

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

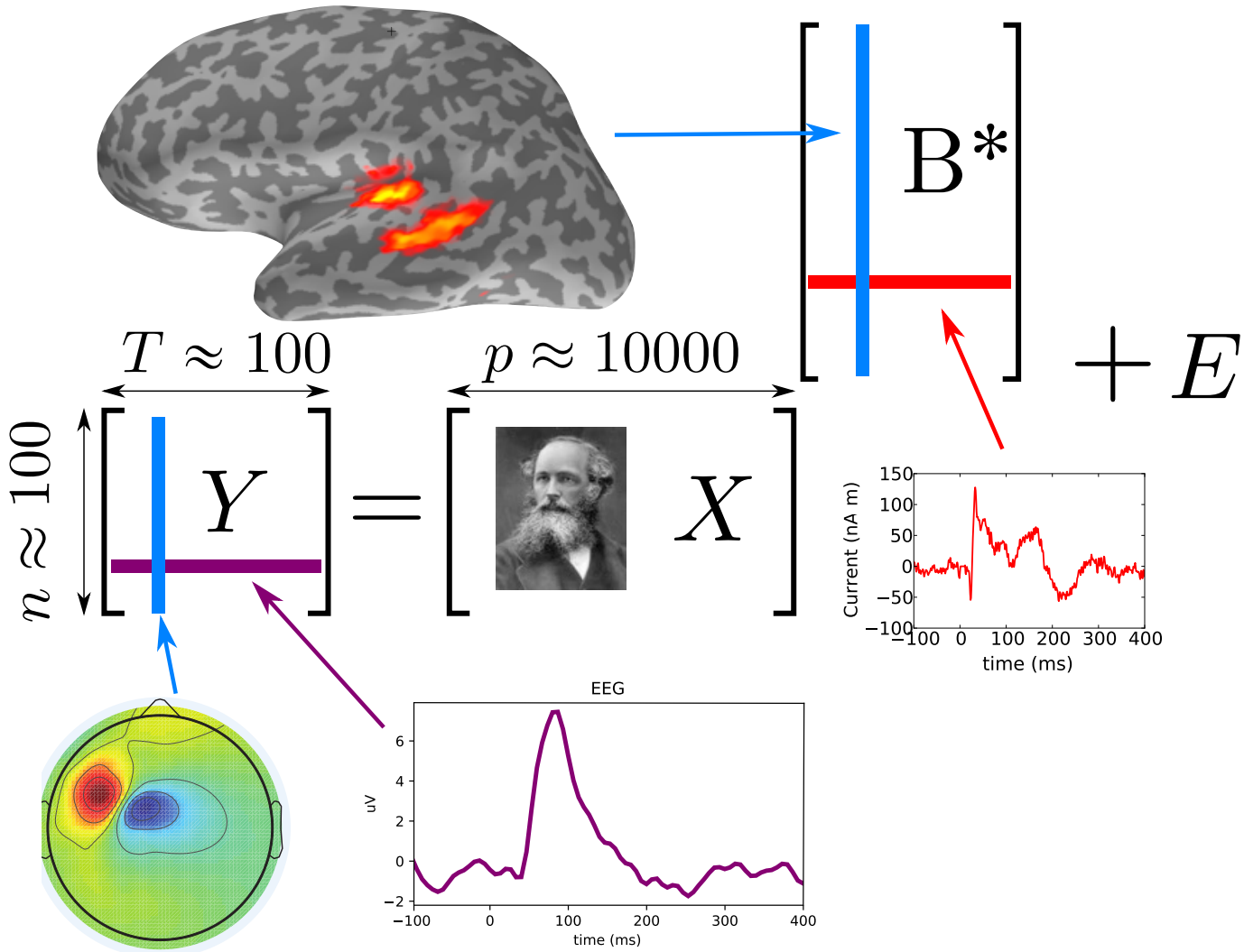




[1] I. Daubechies. *Ten lectures on wavelets*. SIAM, 1992.

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

Mathematical model: linear regression



Lasso[3,4]: the "modern least-squares"[5]

$$\hat{\beta} \in \underset{\beta \in \mathbb{R}^p}{\arg\min} \underbrace{\frac{1}{2} \|y - X\beta\|^2 + \lambda \|\beta\|_1}_{\mathcal{P}(\beta)}$$

- $y \in \mathbb{R}^n$: observations
- $X = [X_1 | \dots | X_p] \in \mathbb{R}^{n \times p}$: design matrix
- **sparsity**: for λ large enough, $\|\hat{\beta}\|_0 \ll p$

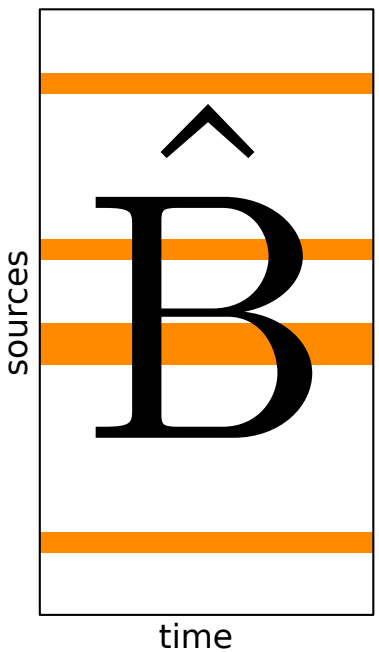
[3] R. Tibshirani. "Regression Shrinkage and Selection via the Lasso". In: *J. R. Stat. Soc. Ser. B Stat. Methodol.* 58.1 (1996), pp. 267–288.

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

[5] 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.

Group-sparsity across time[6]

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



Sparse support: group structure

Group-Lasso penalty

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

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

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

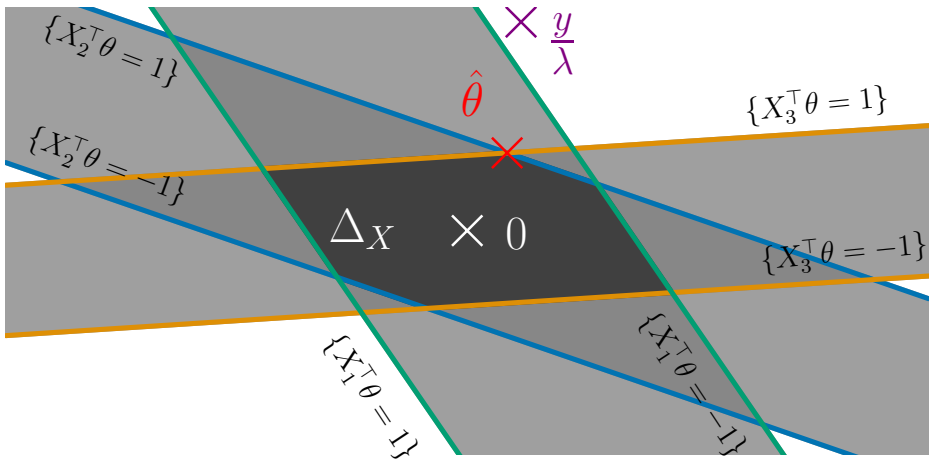
Table of Contents

Duality for the Lasso

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

dual $$\hat{\theta} = \arg\max_{\theta \in \Delta_X} \underbrace{\frac{1}{2} \|y\|^2 - \frac{\lambda^2}{2} \|y/\lambda - \theta\|^2}_{\mathcal{D}(\theta)}$$

$\Delta_X = \left\{\theta \in \mathbb{R}^n \ : \ \forall j \in [p], \ |X_j^\top \theta| \leq 1\right\}$: **dual feasible set**

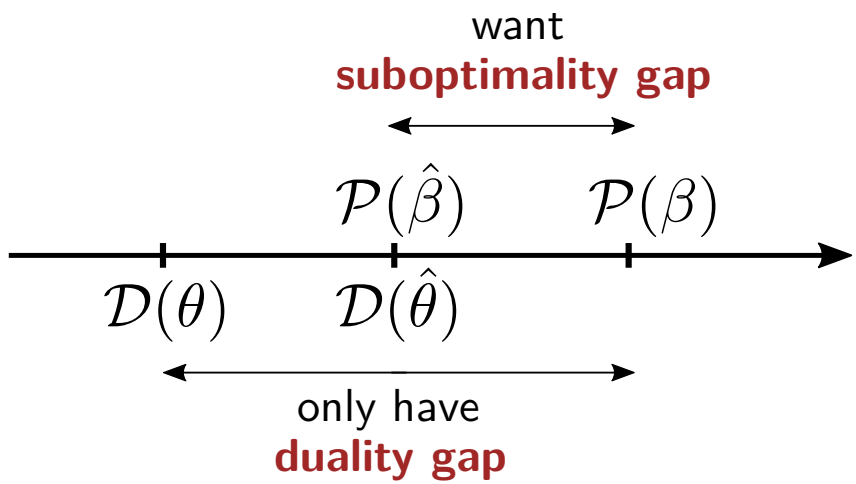


$n = 2, p = 3$

Duality gap as a stopping criterion

For any primal-dual pair $\beta \in \mathbb{R}^p, \theta \in \Delta_X$:

$$\mathcal{P}(\beta) \geq \mathcal{P}(\hat{\beta}) = \mathcal{D}(\hat{\theta}) \geq \mathcal{D}(\theta)$$



Which dual point?

Primal-dual link at optimum:

$$\boxed{\hat{\theta} = (y - X\hat{\beta})/\lambda}$$

Standard approach[7]: at iteration t, corresponding to primal $\beta^{(t)}$ and **residuals** $r^{(t)} \triangleq y - X\beta^{(t)}$, take

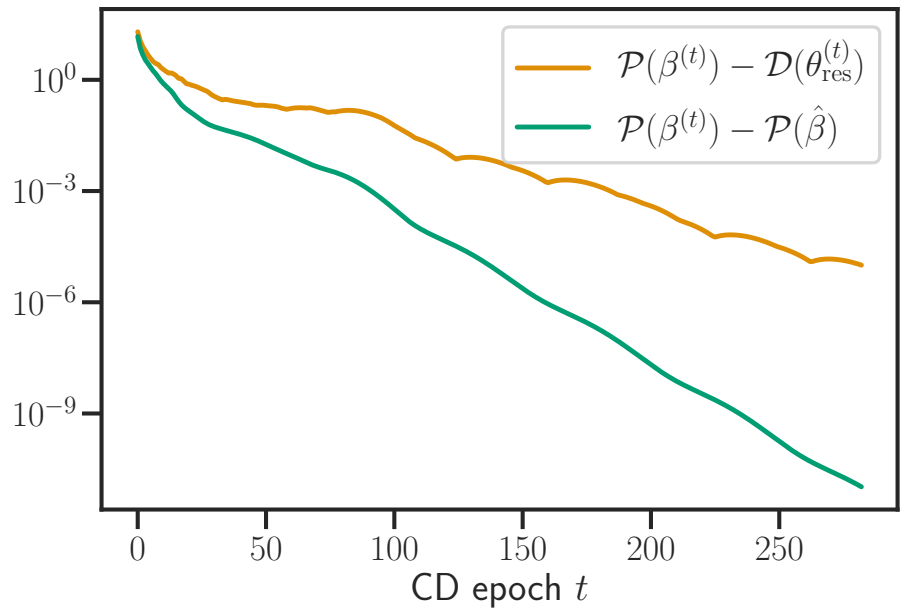
$$\theta = \theta^{(t)}_{\text{res}} \triangleq r^{(t)} / \max(\lambda, \|X^\top r^{(t)}\|_\infty)$$

- converges to $\hat{\theta}$ (provided $\beta^{(t)}$ converges to $\hat{\beta}$)
- costs like 1 epoch, $\mathcal{O}(np)$
  $\hookrightarrow$ rule of thumb: compute $\theta^{(t)}_{\text{res}}$ and dgap every 10 iterations

[7] J. Mairal. "Sparse coding for machine learning, image processing and computer vision". PhD thesis. École normale supérieure de Cachan, 2010.

Residuals rescaling: conservative bound

$$\theta_{\text{res}}^{(t)} = r^{(t)} / \max(\lambda, \|X^\top r^{(t)}\|_\infty)$$



Leukemia dataset: $p = 7129, \quad n = 72, \quad \lambda = \lambda_{\max}/10$

$\hookrightarrow$ do better by exploiting structure in iterates

---

$\lambda_{\max} = \|X^\top y\|_\infty$ is the smallest λ leading to $\hat{\beta} = 0$

VAR structure in residuals

Proposition[8]

Under a non degeneracy condition, proximal gradient and coordinate descent achieve sign id.: $\operatorname{sign} \beta_j^{(t)} = \operatorname{sign} \hat{\beta}_j$.
Then, Lasso residuals are Vector AutoRegressive (**VAR**):

$$r^{(t+1)} = A r^{(t)} + b$$

$\hookrightarrow$ we "only" need to fit a VAR to infer $\lim_{t \to \infty} r^{(t)} = \lambda \hat{\theta}$

We do not know when the sign is identified

cheap solution $\hookrightarrow$ **Anderson acceleration/extrapolation**

[8] M. Massias, A. Gramfort, and J. Salmon. "Celer: a fast solver for the Lasso with dual extrapolation". In: *ICML*. 2018, pp. 3321–3330.

Simple example: extrapolation in 1D

1D autoregressive process:

$$x^{(t)} = ax^{(t-1)} + b \underset{t\to\infty}{\to} x^*$$

we have

$$x^{(t)} - x^* = a(x^{(t-1)} - x^*)$$
$$x^{(t-1)} - x^* = a(x^{(t-2)} - x^*)$$

"Aitken's Δ^2": 2 unknowns, so 2 eqs or 3 points $x^{(t)}, x^{(t-1)}, x^{(t-2)}$ are enough to find x^*![9]

[9] A. Aitken. "On Bernoulli's numerical solution of algebraic equations". In: *Proceedings of the Royal Society of Edinburgh* 46 (1926), pp. 289–305.

Generalization[10] to VAR $r^{(t)} \in \mathbb{R}^n$

- fix $K = 5$ (small)
- keep track of K past residuals $r^{(t)}, \ldots, r^{(t+1-K)}$
- $U^{(t)} = \left[r^{(t+1-K)} - r^{(t-K)}, \ldots, r^{(t)} - r^{(t-1)}\right] \in \mathbb{R}^{n \times K}$
- solve $(U^{(t)})^\top U^{(t)} z = \mathbf{1}_K$
- $c = z / z^\top \mathbf{1}_K$

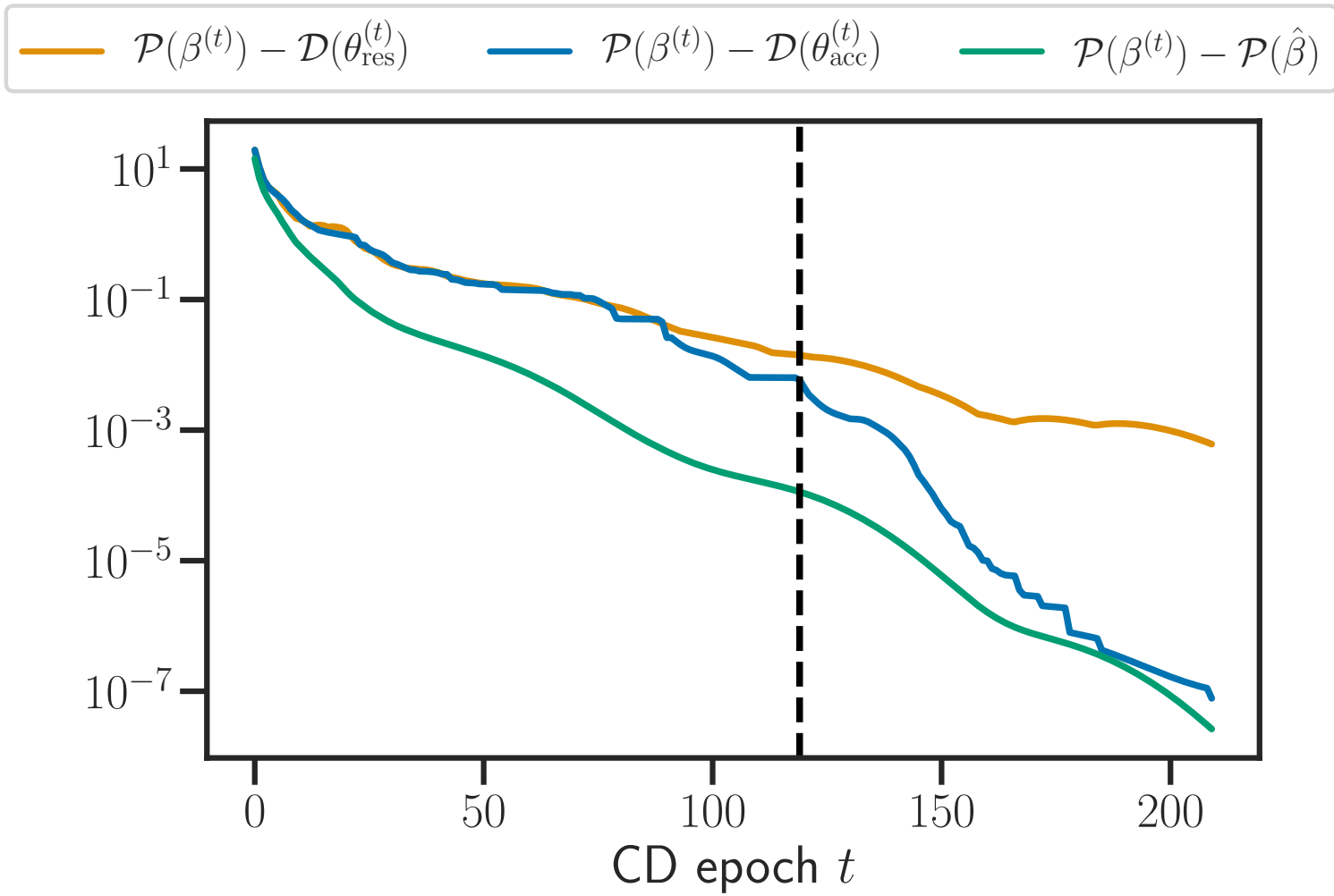
$$r_{\text{accel}}^{(t)} \triangleq \sum_{k=1}^{K} c_k r^{(t+1-k)}$$

$$\boxed{\theta_{\text{accel}}^{(t)} \triangleq r_{\text{accel}}^{(t)} / \max(\lambda, \|X^\top r_{\text{accel}}^{(t)}\|_\infty)}$$

Cost: $\mathcal{O}(K^3 + K^2 n + np)$

[10] D. Scieur, A. d'Aspremont, and F. Bach. "Regularized Nonlinear Acceleration". In: *NeurIPS*. 2016, pp. 712–720.

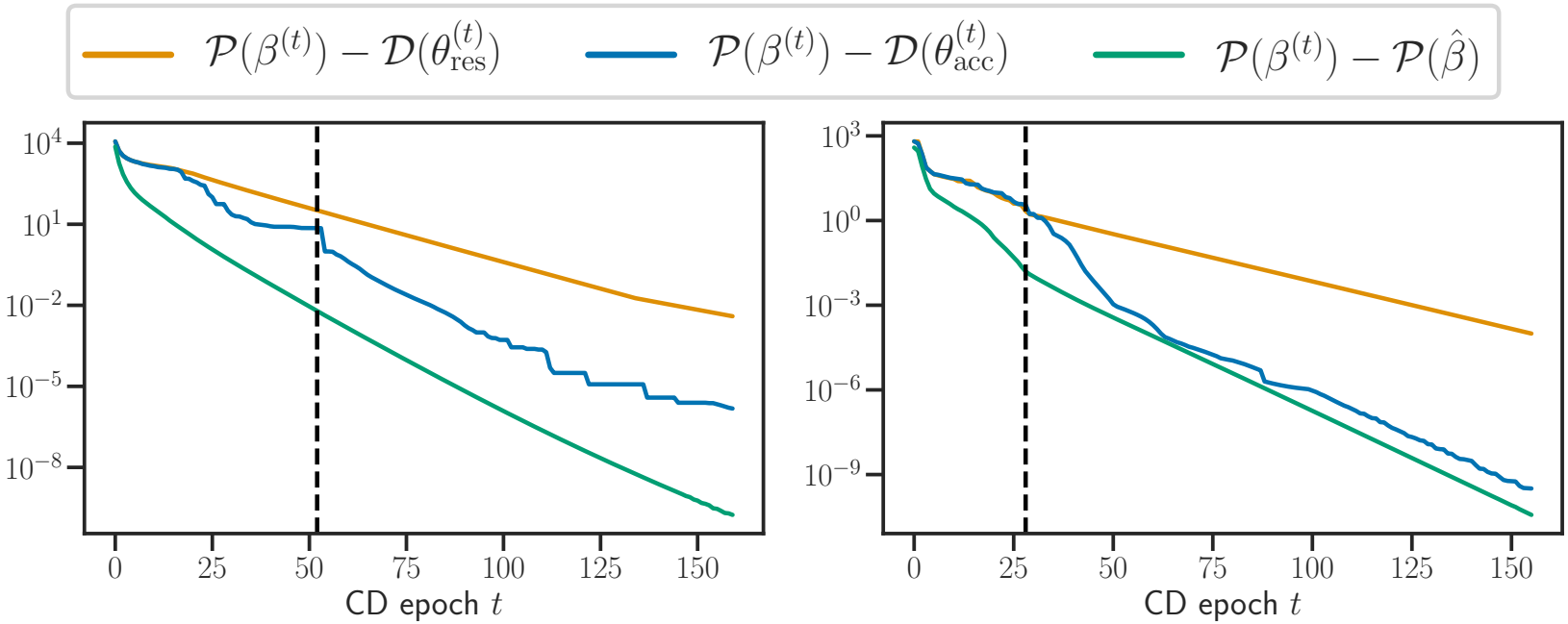
Dual extrapolation for the Lasso



Leukemia dataset: $p = 7129, \quad n = 72, \quad \lambda = \lambda_{\max}/10$

Applicability to other models

We showed *asymptotic* VAR structure, also exploitable[11]



logreg, *rcv1* dataset:

$p = 20\text{k}, \quad n = 20\text{k}$

$\lambda = \lambda_{\max}/20 \quad (\|\hat{\beta}\|_0 = 395)$

MTL, real **MEG** data:

$p = 7498, \quad n = 305$

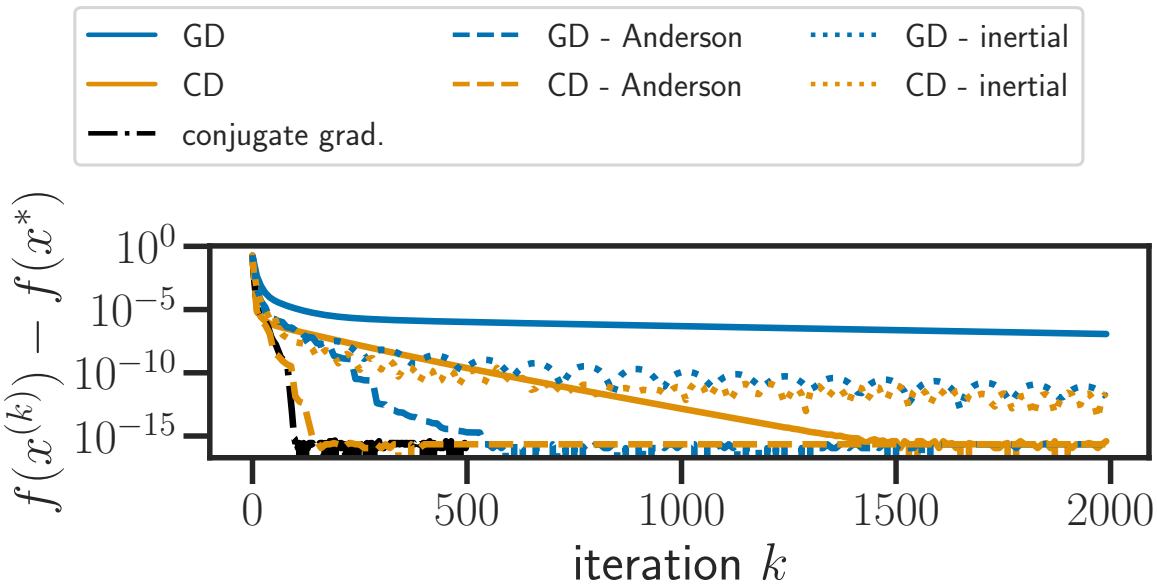
$\lambda = \lambda_{\max}/10 \quad (\|\hat{\mathrm{B}}\|_{2,0} = 45)$

Python software: https://github.com/mathurinm/celer
(sklearn-compatible, reproducible benchmarks, documentation)

[11] M. Massias et al. "Dual extrapolation for sparse Generalized Linear Models". In: *J. Mach. Learn. Res.* (2020).

Additional contributions

- Improve screening rules and working sets with **Celer**
- Anderson acceleration directly **in the primal** (more general)[12]



Ordinary Least Squares problem

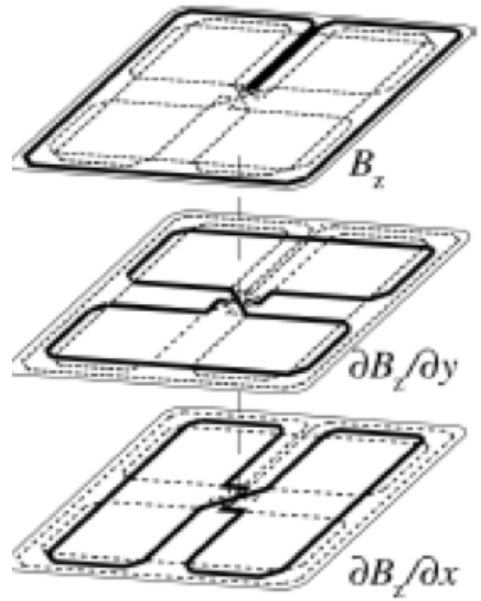
[12] Q. Bertrand and M. Massias. *Anderson acceleration of coordinate descent*. 2020.

Table of Contents

MEG sensors: magnetometers and gradiometers

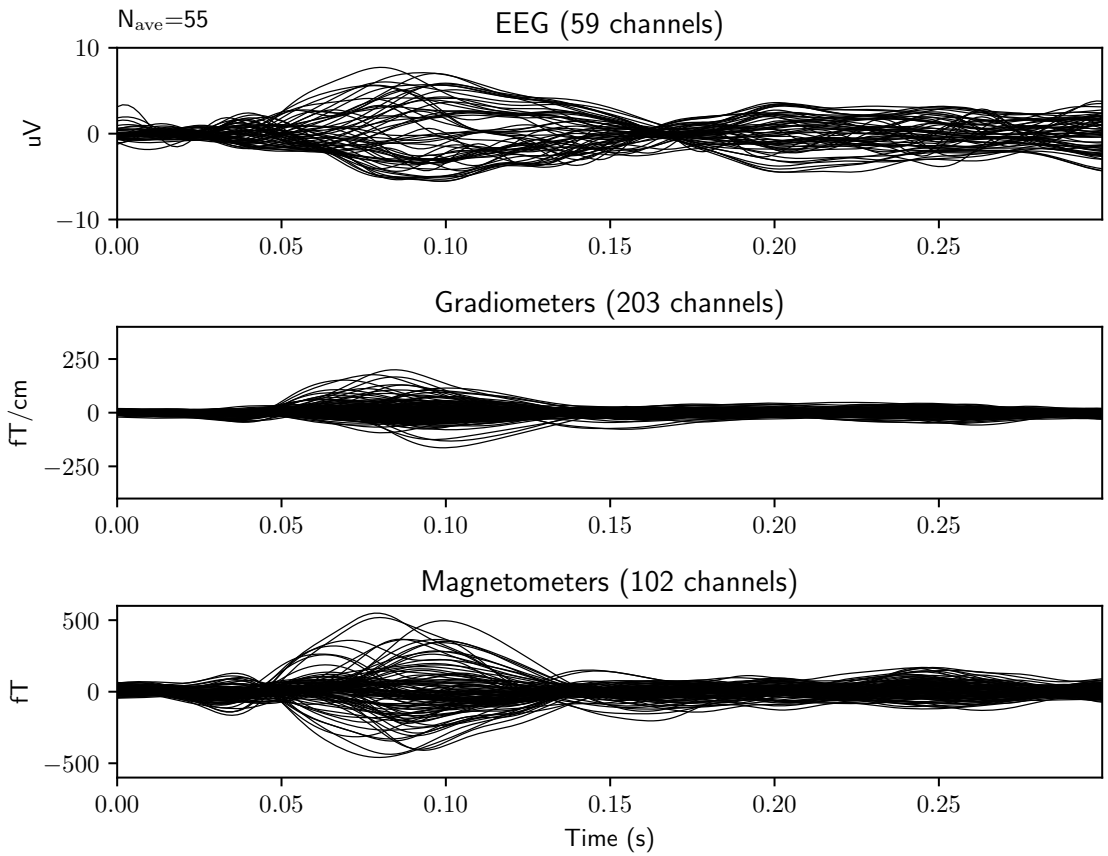
Device

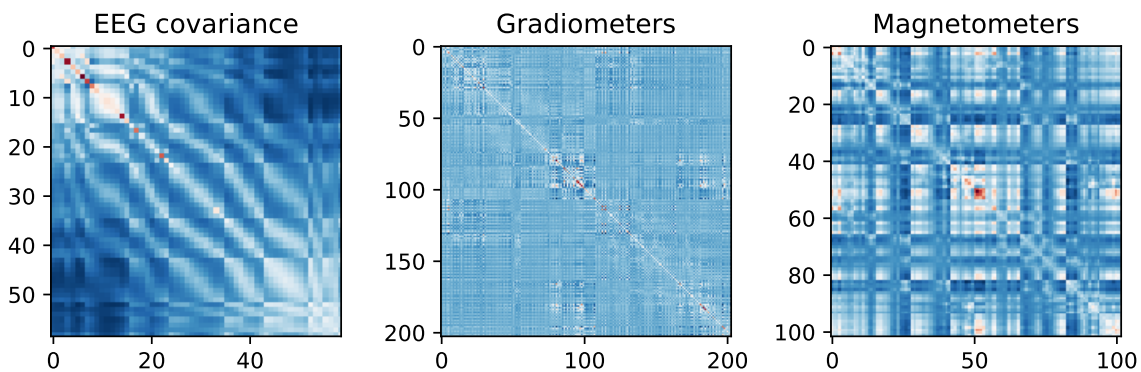
Sensors



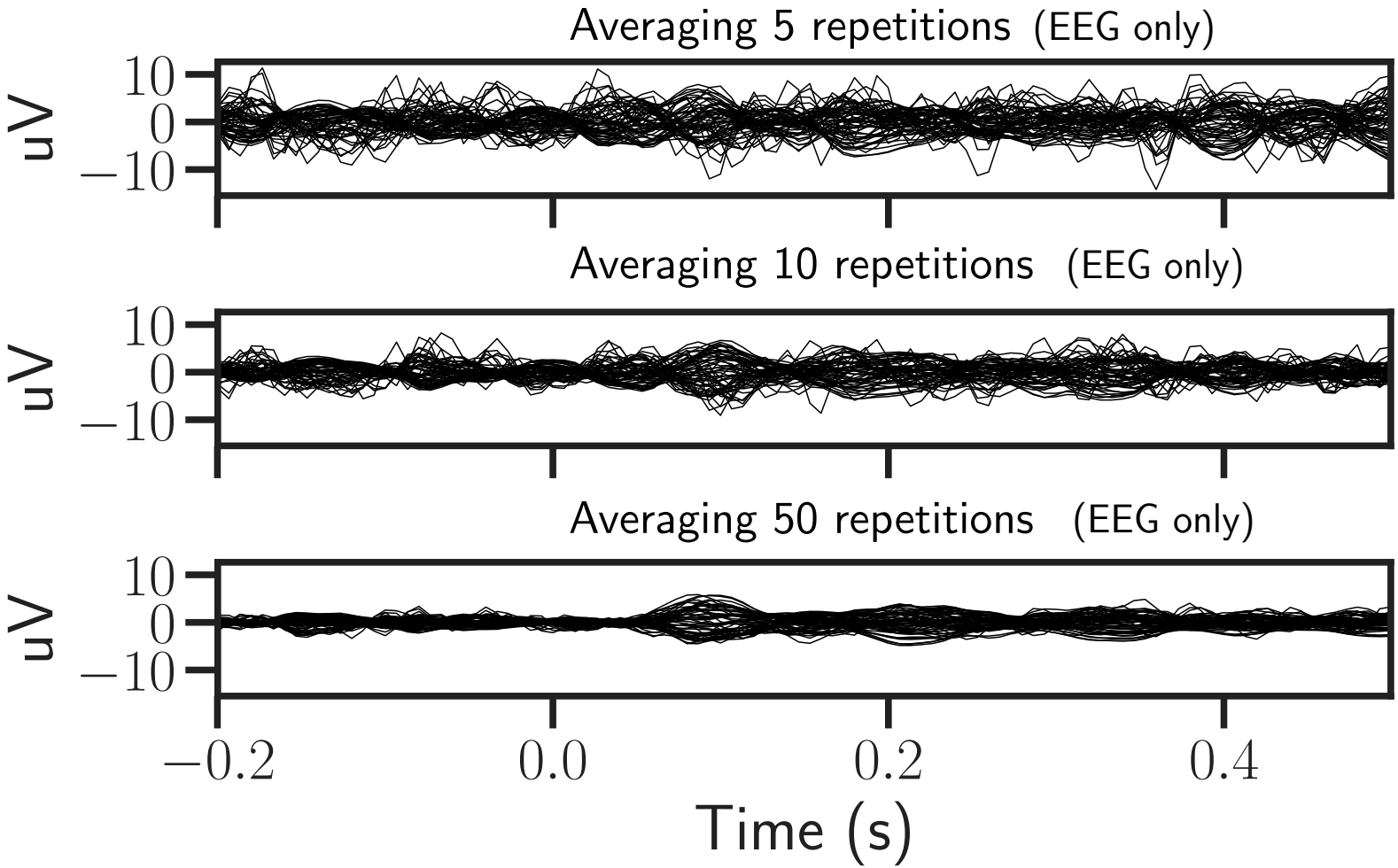
Detail of a sensor

3 sensor types $\Longrightarrow$ 3 noise structures





Low SNR: averaging repetitions of experiment



Our complete stats model for M/EEG

- r repetitions of the same experiment
- $Y^{(1)}, \ldots, Y^{(r)} \in \mathbb{R}^{n \times T}$ observation matrices; $\bar{Y} = \frac{1}{r} \sum_l Y^{(l)}$

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

- $S_* \in \mathbb{S}^n_{++}$ co-standard deviation matrix (**unknown**)
- $\mathrm{E}^{(1)}, \ldots, \mathrm{E}^{(r)} \in \mathbb{R}^{n \times T}$: white Gaussian noise

Data-fitting term

- M/EEG standard: use whitened, averaged signal

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

- **Double goal**: take advantage of the number of repetitions, address correlated noise
- trying to average the datafits with squared Frobenius norm yields same solution as working with $\bar{Y}$

Lasso and optimal λ[13,14]

Theorem

For $y = X\beta^* + \sigma_* \varepsilon$, 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

BUT σ_* is unknown in practice !

[13] 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.

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

Pivotality: the $\sqrt{\text{Lasso}}$[16]

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

- has an optimal λ independent of σ_*
- but slow to optimize $\hookrightarrow$ use smoothed *Concomitant* Lasso[15] formulation:

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

same solutions when $\|y - X\hat{\beta}_{\sqrt{\text{Lasso}}}\| \geq \sqrt{n}\underline{\sigma}$

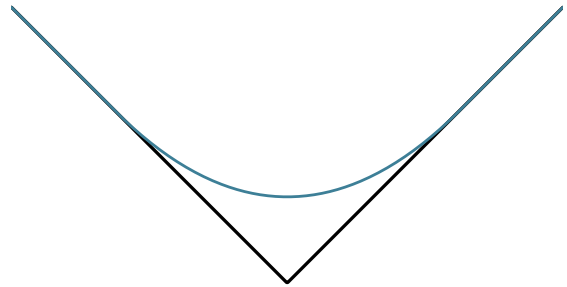
---

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

[16] 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.

Concomitant origin: smoothing the $\sqrt{\text{Lasso}}$[19]

"**Huberization**": replace $\|y - X\beta\|$ by a smooth approximation:



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

Leads to the Smoothed[17, 18] Concomitant Lasso formulation:

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

---

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

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

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

Multitask generalization

To address correlated noise, we had introduced a Smooth Generalized Concomitant Lasso (SGCL):[20]

$$\boxed{\underset{\substack{\mathrm{B}\in\mathbb{R}^{p\times T}\\ S\succ\underline{\sigma}\,\mathrm{Id}_n}}{\arg\min}\ \frac{1}{2nT}\|Y-X\mathrm{B}\|_{S^{-1}}^2+\frac{1}{2n}\operatorname{Tr}(S)+\lambda\|\mathrm{B}\|_{2,1}}$$

S. van de Geer introduced the pivotal multivariate $\sqrt{\text{Lasso}}$[21]:

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

SGCL turns out to be a *smoothed multivariate square-root Lasso*!

Smoothing makes optimization and statistical analysis easy!

---

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

[21] S. van de Geer. *Estimation and testing under sparsity*. Lecture Notes in Mathematics. Springer, 2016.

Leveraging repetitions

Smoothed Generalized Concomitant Lasso (SGCL)[22]

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

Concomitant Lasso with Repetitions (CLaR)[23]

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

---

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

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

Support recovery guarantees

Proposition (smoothed multivariate $\sqrt{\text{Lasso}}$)[24]

Under classical assumptions.[25] For constants C, c and $A \geq \sqrt{2}$, if $\lambda = \frac{2\sqrt{2}}{\sqrt{nq}}(1 + A\sqrt{(\log p)/q})$, with high probability:

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

Moreover if

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

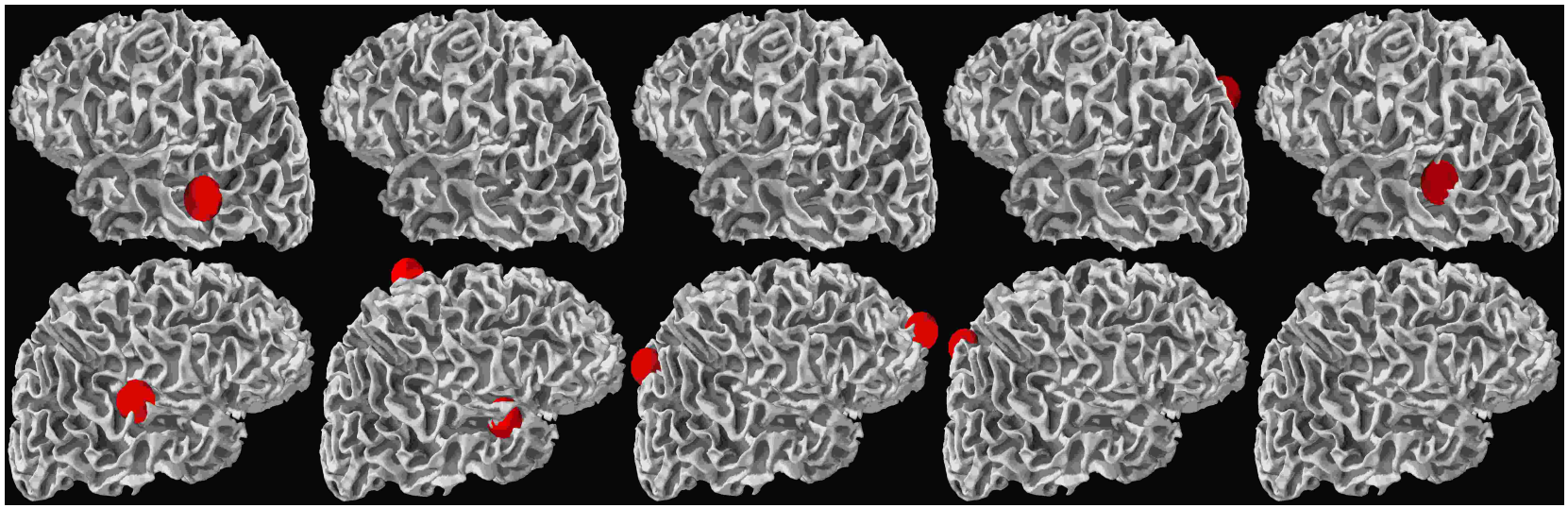
then with the same probability, the support is recovered

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

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

[25] Gaussian noise, mutual incoherence (α), correct value of smoothing parameter (η)

Dissemination



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

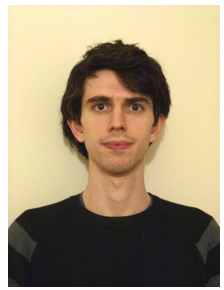
Figure: *Real data, sources found after right auditory stimulations.*

SGCL & CLaR package: `https://github.com/mathurinm/sgcl`

Thank you

Joseph, Alexandre, Olivier, Samuel, Quentin, Pierre, Thomas







- M. Massias, O. Fercoq, A. Gramfort, and J. Salmon. *Smoothed generalized concomitant Lasso for sparse multimodal regression. AISTATS*, 2018
- M. Massias, A. Gramfort, and J. Salmon. *Celer: a fast solver for the Lasso with dual extrapolation. ICML*, 2018
- Q. Bertrand*, M. Massias*, A. Gramfort, and J. Salmon. *Handling correlated and repeated measurements with the smoothed multivariate square-root Lasso. NeurIPS*, 2019
- M. Massias, S. Vaiter, A. Gramfort, and J. Salmon. *Dual extrapolation for sparse GLMs. to appear in JMLR*, 2020
- M. Massias*, Q. Bertrand*, A. Gramfort, and J. Salmon. *Support recovery and sup-norm convergence rates for sparse pivotal estimation. AISTATS*, 2020
- Q. Bertrand and M. Massias. *Anderson acceleration of coordinate descent. submitted*, 2020

References I

- Aitken, A. "On Bernoulli's numerical solution of algebraic equations". In: *Proceedings of the Royal Society of Edinburgh* 46 (1926), pp. 289–305.
- 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. and M. Massias. *Anderson acceleration of coordinate descent*. 2020.
- Bertrand, Q. et al. "Handling correlated and repeated measurements with the smoothed multivariate square-root Lasso". In: *NeurIPS*. 2019.

References II

- 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.
- 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.
- Mairal, J. "Sparse coding for machine learning, image processing and computer vision". PhD thesis. École normale supérieure de Cachan, 2010.

References III

- Massias, M., A. Gramfort, and J. Salmon. "Celer: a fast solver for the Lasso with dual extrapolation". In: *ICML*. 2018, pp. 3321–3330.
- Massias, M. et al. "Dual extrapolation for sparse Generalized Linear Models". In: *J. Mach. Learn. Res.* (2020).
- Massias, M. et al. "Generalized concomitant multi-task Lasso for sparse multimodal regression". In: *AISTATS*. 2018, pp. 998–1007.
- Massias, M. et al. "Support recovery and sup-norm convergence rates for sparse pivotal estimation.". 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.
- Nesterov, Y. "Smooth minimization of non-smooth functions". In: *Math. Program.* 103.1 (2005), pp. 127–152.

References IV

- 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: *Cont. Math.* 443 (2007), pp. 59–72.
- Scieur, D., A. d'Aspremont, and F. Bach. "Regularized Nonlinear Acceleration". In: *NeurIPS*. 2016, pp. 712–720.
- 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*. Lecture Notes in Mathematics. Springer, 2016.