Optimizational and statistical contributions to the $\ell_{2,1}$-regularized M/EEG inverse problem

PhD defense

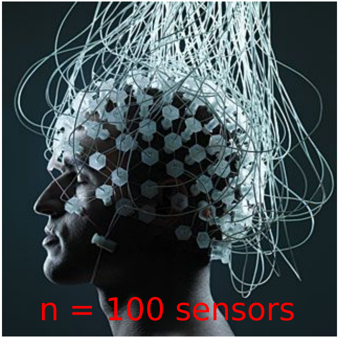
Mathurin Massias (INRIA)

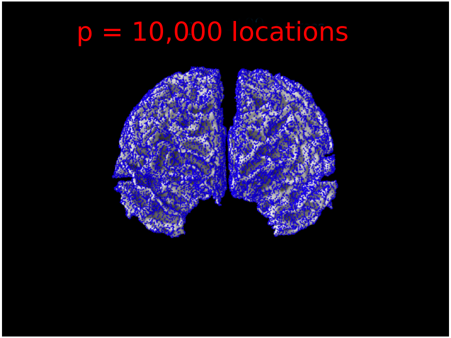
- G. Peyré (Rapporteur)
- M. Schmidt (Rapporteur)
- N. Pustelnik (Examinatrice)
- O. Fercoq (Examinateur)
- J. Mairal (Examinateur)
- J. Salmon (Directeur)
- A. Gramfort (Co-directeur)

Down by the pool

The M/EEG inverse problem

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



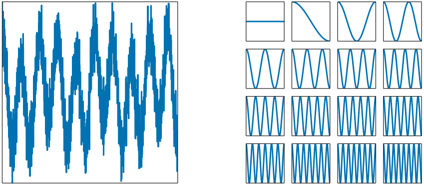


$n \ll p$: ill-posed problem!

Sparsity everywhere

Signals can often be represented combining few atoms/features:

- Fourier decomposition for sounds



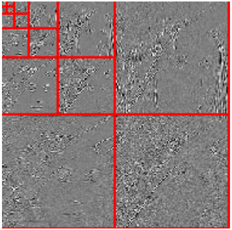
[1] I. Daubechies. *Ten lectures on wavelets*. SIAM, 1992.

[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]

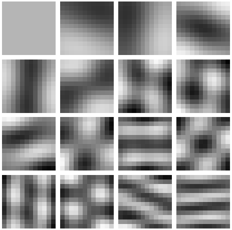
[1] I. Daubechies. *Ten lectures on wavelets*. SIAM, 1992.

[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]



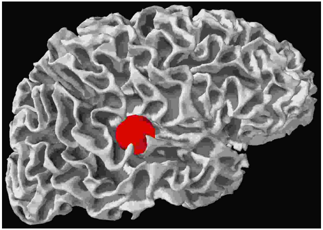
[1] I. Daubechies. *Ten lectures on wavelets*. SIAM, 1992.

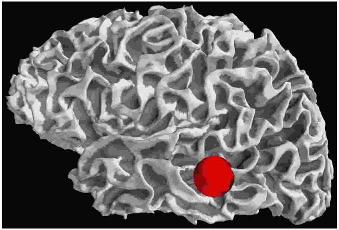
[2] B. A. Olshausen and D. J. Field. "Sparse coding with an overcomplete basis set: A strategy employed by V1?" In: *Vision research* (1997).

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



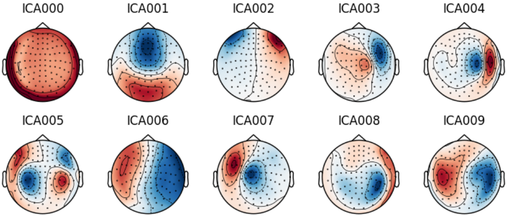


[1] I. Daubechies. *Ten lectures on wavelets*. SIAM, 1992.

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

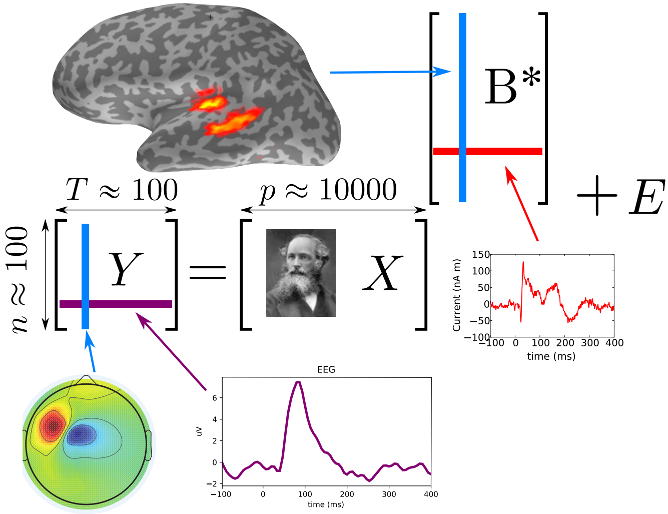
Justification for dipolarity assumption

- 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]

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

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

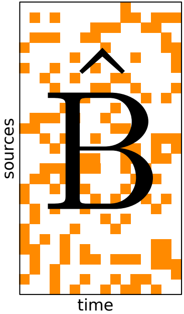
[4] R. Tibshirani. "Regression Shrinkage and Selection via the Lasso". In: *J. R. Stat. Soc. Ser. B Stat. Methodol.* 58.1 (1996), pp. 267–288.

[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 ✗

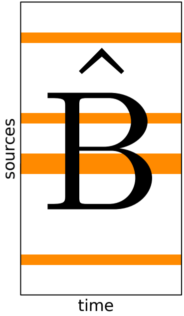
Lasso penalty

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

[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]

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



Sparse support: group structure ✓

Group-Lasso penalty

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

where $\mathrm{B}_{j:}$ the 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.

Lasso-type problems

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

Lasso-type problems

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

sparse Log. reg. $\quad \underset{\beta \in \mathbb{R}^p}{\arg\min} \sum_{i=1}^{n} \log\left(1 + \exp(-y_i \beta^\top x_i)\right) + \lambda\|\beta\|_1$

Lasso-type problems

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

sparse Log. reg. $\quad \underset{\beta \in \mathbb{R}^p}{\arg\min} \sum_{i=1}^{n} \log\left(1 + \exp(-y_i \beta^\top x_i)\right) + \lambda\|\beta\|_1$

Multi-task Lasso $\quad \underset{\mathrm{B} \in \mathbb{R}^{p \times T}}{\arg\min} \frac{1}{2}\|Y - X\mathrm{B}\|^2 + \lambda\|\mathrm{B}\|_{2,1}$

Table of Contents

Duality for the Lasso

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

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

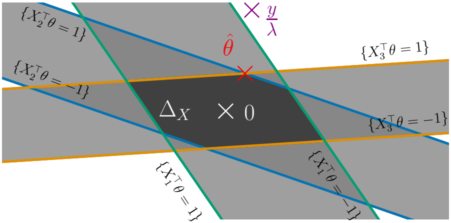
$\Delta_X = \left\{\theta \in \mathbb{R}^n : \forall j \in [p], \ |X_j^\top \theta| \leq 1\right\}$: **dual feasible set**

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**



$n = 2, p = 3$

Solving the Lasso in ML: cyclic CD[8,9]

To minimize: $$\mathcal{P}(\beta) = \frac{1}{2}\|y - \sum_{j=1}^{p} X_j \beta_j\|^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

$\beta^{(0)} = \mathbf{0}_p \in \mathbb{R}^p$

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

[8] J. Friedman, T. J. Hastie, and R. Tibshirani. "Regularization paths for generalized linear models via coordinate descent". In: *J. Stat. Softw.* 33.1 (2010), p. 1.

[9] R.-E. Fan et al. "LIBLINEAR: A library for large linear classification". In: *JMLR* 9 (2008), pp. 1871–1874.

Solving the Lasso in ML: cyclic CD[8,9]

To minimize: $\mathcal{P}(\beta) = \frac{1}{2}\|y - \sum_{j=1}^{p} X_j \beta_j\|^2 + \lambda \sum_{j=1}^{p} |\beta_j|$

$\beta^{(0)} = \mathbf{0}_p \in \mathbb{R}^p$

for $t = 1, \dots$ **do**

$\quad \beta_1^{(t)} \approx \arg\min_{\beta_1 \in \mathbb{R}} \mathcal{P}(\beta_1, \beta_2^{(t-1)}, \beta_3^{(t-1)}, \dots, \beta_{p-1}^{(t-1)}, \beta_p^{(t-1)})$

---

[8] J. Friedman, T. J. Hastie, and R. Tibshirani. "Regularization paths for generalized linear models via coordinate descent". In: *J. Stat. Softw.* 33.1 (2010), p. 1.

[9] R.-E. Fan et al. "LIBLINEAR: A library for large linear classification". In: *JMLR* 9 (2008), pp. 1871–1874.

Solving the Lasso in ML: cyclic CD[8,9]

To minimize: $\mathcal{P}(\beta) = \frac{1}{2}\|y - \sum_{j=1}^{p} X_j \beta_j\|^2 + \lambda \sum_{j=1}^{p} |\beta_j|$

$\beta^{(0)} = \mathbf{0}_p \in \mathbb{R}^p$

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

$\beta_1^{(t)} \approx \arg\min_{\beta_1 \in \mathbb{R}} \mathcal{P}(\beta_1, \beta_2^{(t-1)}, \beta_3^{(t-1)}, \ldots, \beta_{p-1}^{(t-1)}, \beta_p^{(t-1)})$

$\beta_2^{(t)} \approx \arg\min_{\beta_2 \in \mathbb{R}} \mathcal{P}(\beta_1^{(t)}, \beta_2, \beta_3^{(t-1)}, \ldots, \beta_{p-1}^{(t-1)}, \beta_p^{(t-1)})$

---

[8] J. Friedman, T. J. Hastie, and R. Tibshirani. "Regularization paths for generalized linear models via coordinate descent". In: *J. Stat. Softw.* 33.1 (2010), p. 1.

[9] R.-E. Fan et al. "LIBLINEAR: A library for large linear classification". In: *JMLR* 9 (2008), pp. 1871–1874.

Solving the Lasso in ML: cyclic CD[8,9]

To minimize: $\mathcal{P}(\beta) = \frac{1}{2}\|y - \sum_{j=1}^{p} X_j\beta_j\|^2 + \lambda \sum_{j=1}^{p} |\beta_j|$

$\beta^{(0)} = \mathbf{0}_p \in \mathbb{R}^p$

for $t = 1, \dots$ **do**

$$\left.\begin{array}{l}
\beta_1^{(t)} \approx \underset{\beta_1 \in \mathbb{R}}{\arg\min}\, \mathcal{P}(\beta_1, \beta_2^{(t-1)}, \beta_3^{(t-1)}, \dots, \beta_{p-1}^{(t-1)}, \beta_p^{(t-1)}) \\
\beta_2^{(t)} \approx \underset{\beta_2 \in \mathbb{R}}{\arg\min}\, \mathcal{P}(\beta_1^{(t)}, \beta_2, \beta_3^{(t-1)}, \dots, \beta_{p-1}^{(t-1)}, \beta_p^{(t-1)}) \\
\vdots \\
\beta_p^{(t)} \approx \underset{\beta_p \in \mathbb{R}}{\arg\min}\, \mathcal{P}(\beta_1^{(t)}, \beta_2^{(t)}, \beta_3^{(t)}, \dots, \beta_{p-1}^{(t)}, \beta_p)
\end{array}\right\} \text{1 epoch}$$

---

[8] J. Friedman, T. J. Hastie, and R. Tibshirani. “Regularization paths for generalized linear models via coordinate descent”. In: *J. Stat. Softw.* 33.1 (2010), p. 1.

[9] R.-E. Fan et al. “LIBLINEAR: A library for large linear classification”. In: *JMLR* 9 (2008), pp. 1871–1874.

Solving the Lasso in ML: cyclic CD[8,9]

To minimize: $$\mathcal{P}(\beta) = \frac{1}{2}\|y - \sum_{j=1}^p X_j \beta_j\|^2 + \lambda \sum_{j=1}^p |\beta_j|$$

$\beta^{(0)} = \mathbf{0}_p \in \mathbb{R}^p$

for $t = 1, \dots$ **do**

$$\left.\begin{aligned}
\beta_1^{(t)} &\approx \mathop{\arg\min}_{\beta_1 \in \mathbb{R}} \mathcal{P}(\beta_1, \beta_2^{(t-1)}, \beta_3^{(t-1)}, \dots, \beta_{p-1}^{(t-1)}, \beta_p^{(t-1)}) \\
\beta_2^{(t)} &\approx \mathop{\arg\min}_{\beta_2 \in \mathbb{R}} \mathcal{P}(\beta_1^{(t)}, \beta_2, \beta_3^{(t-1)}, \dots, \beta_{p-1}^{(t-1)}, \beta_p^{(t-1)}) \\
&\vdots \\
\beta_p^{(t)} &\approx \mathop{\arg\min}_{\beta_p \in \mathbb{R}} \mathcal{P}(\beta_1^{(t)}, \beta_2^{(t)}, \beta_3^{(t)}, \dots, \beta_{p-1}^{(t)}, \beta_p)
\end{aligned}\right\} \text{1 epoch}$$

When do we stop?

---

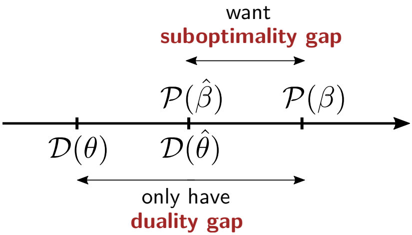
[8] J. Friedman, T. J. Hastie, and R. Tibshirani. "Regularization paths for generalized linear models via coordinate descent". In: *J. Stat. Softw.* 33.1 (2010), p. 1.

[9] R.-E. Fan et al. "LIBLINEAR: A library for large linear classification". In: *JMLR* 9 (2008), pp. 1871–1874.

Duality gap as a stopping criterion

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

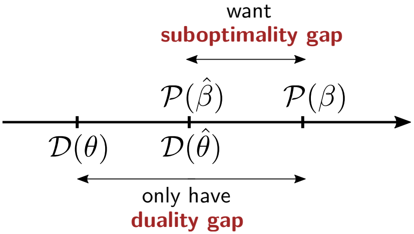
$$\mathcal{P}(\beta) \geq \mathcal{P}(\hat{\beta}) = \mathcal{D}(\hat{\theta}) \geq \mathcal{D}(\theta)$$



Duality gap as a stopping criterion

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

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



$$\forall \beta, (\exists \theta \in \Delta_X, \, \mathrm{dgap}(\beta, \theta) \leq \epsilon) \Rightarrow \mathcal{P}(\beta) - \mathcal{P}(\hat{\beta}) \leq \epsilon$$

β is an ϵ-solution whenever $\mathrm{dgap}(\beta, \theta) \leq \epsilon$

Which dual point?

Primal-dual link at optimum:

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

Standard approach[10]: at epoch t, corresponding to primal $\beta^{(t)}$ and **residuals** $r^{(t)} \triangleq y - X\beta^{(t)}$, take

$$\theta = \theta_{\text{res}}^{(t)} \triangleq r^{(t)}/\lambda$$

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

Which dual point?

Primal-dual link at optimum:

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

Standard approach[10]: at epoch t, corresponding to primal $\beta^{(t)}$ and **residuals** $r^{(t)} \triangleq y - X\beta^{(t)}$, take

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

"residuals rescaling"

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

Which dual point?

Primal-dual link at optimum:

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

Standard approach[10]: at epoch t, corresponding to primal $\beta^{(t)}$ and **residuals** $r^{(t)} \triangleq y - X\beta^{(t)}$, take

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

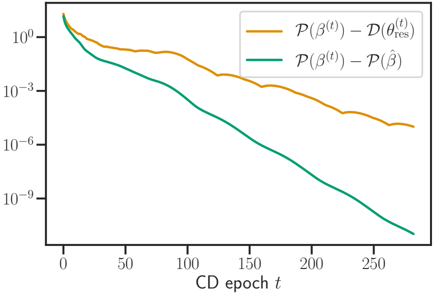
"residuals rescaling"

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

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

Residuals rescaling: conservative bound

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



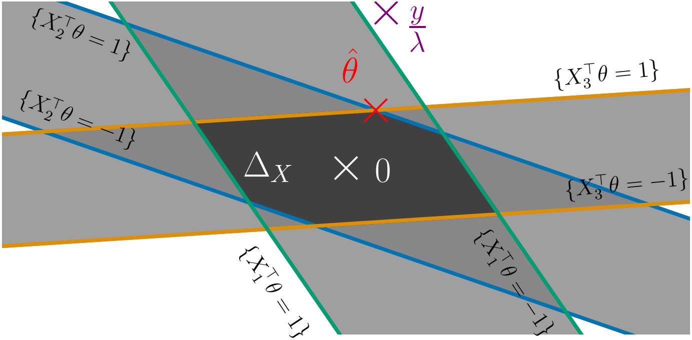
Leukemia dataset: $p = 7129, \quad n = 72, \quad \lambda = \lambda_{\max}/10$

$\hookrightarrow$ do better by exploiting previous iterates

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

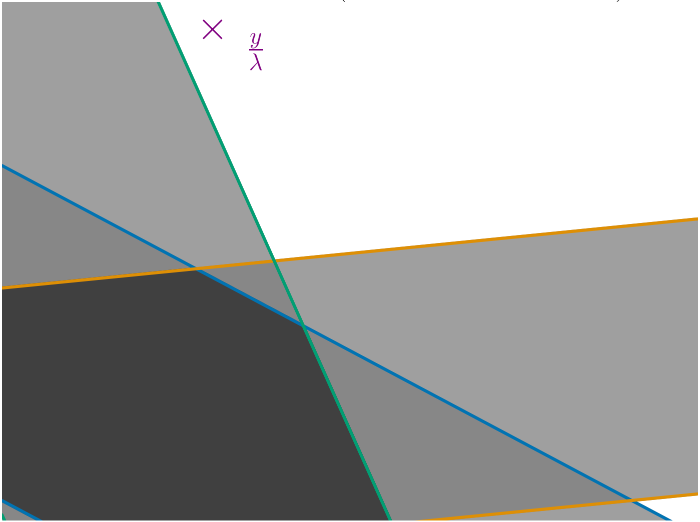
Residuals regularity, sign identification

$$n = 2, p = 3, \hat{\beta} = (0, -0.6, 1.3)$$



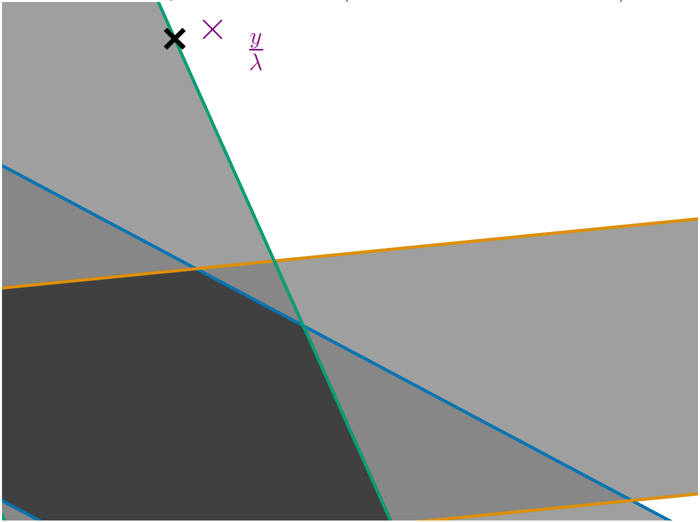
Residuals regularity, sign identification

$$t = 0, \qquad \beta = (0.000, 0.000, 0.000)$$



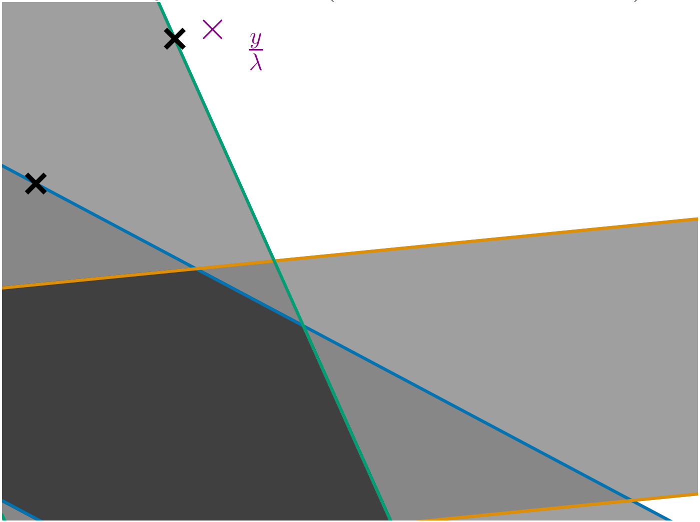
Residuals regularity, sign identification

$t = 1,\ j = 1, \beta = (0.217, 0.000, 0.000)$



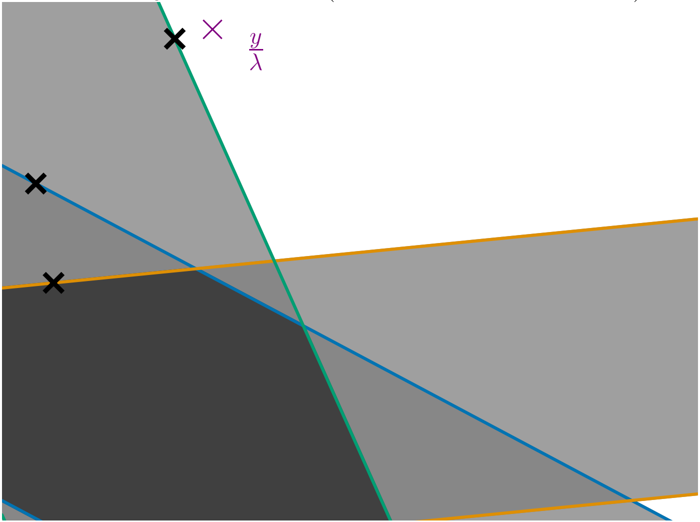
Residuals regularity, sign identification

$t = 1,\ j = 2, \beta = (0.217, -1.306, 0.000)$



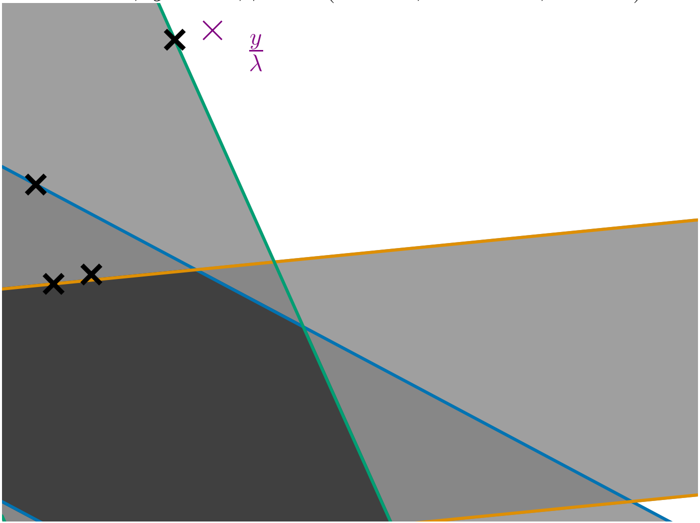
Residuals regularity, sign identification

$t = 1,\ j = 3, \beta = (0.217, -1.306, 0.735)$



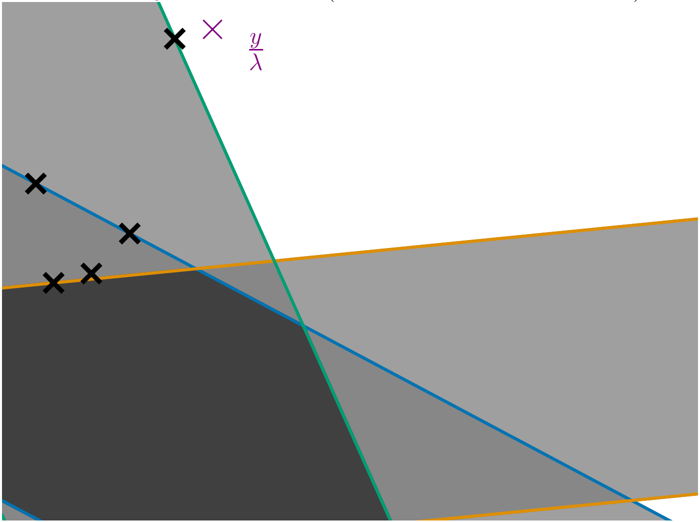
Residuals regularity, sign identification

$t = 2,\ j = 1, \beta = (0.000, -1.306, 0.735)$



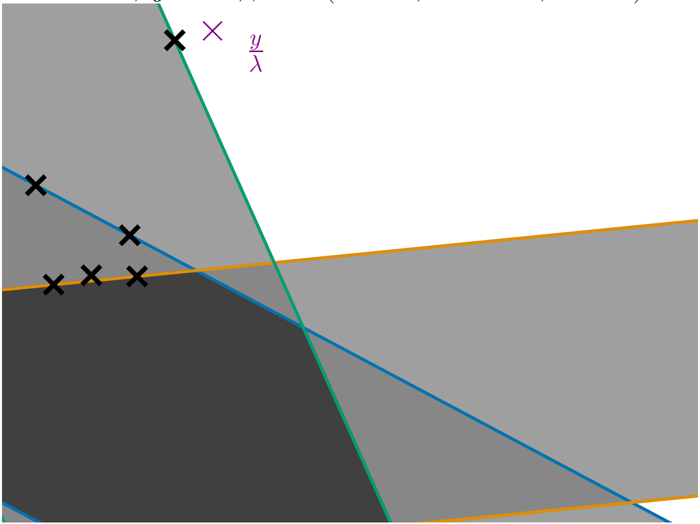
Residuals regularity, sign identification

$t = 2,\ j = 2, \beta = (0.000, -0.945, 0.735)$



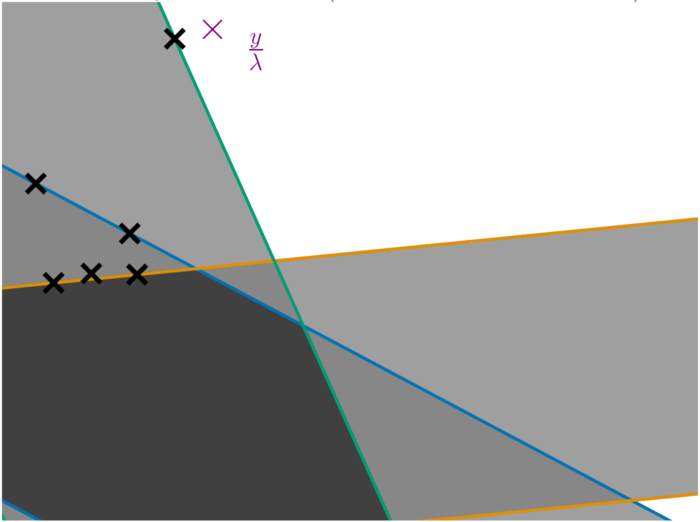
Residuals regularity, sign identification

$t = 2,\ j = 3, \beta = (0.000, -0.945, 1.039)$



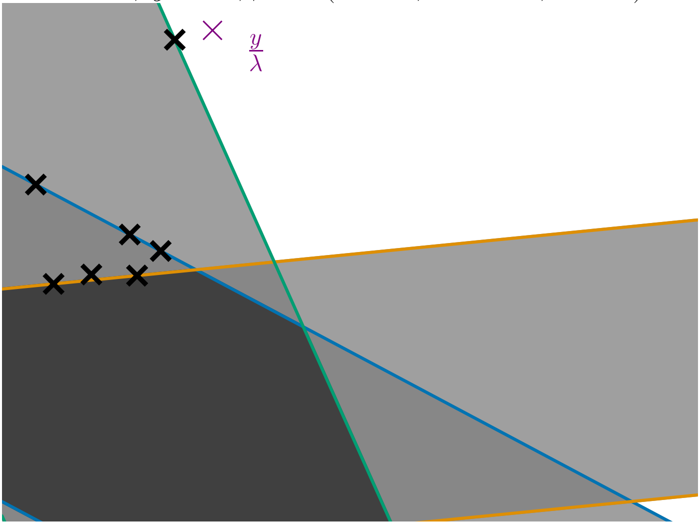
Residuals regularity, sign identification

$t = 3,\ j = 1, \beta = (0.000, -0.945, 1.039)$



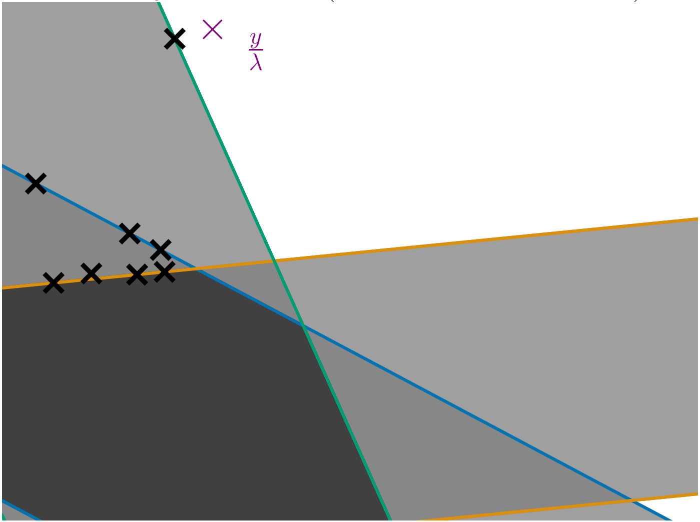
Residuals regularity, sign identification

$t = 3,\ j = 2, \beta = (0.000, -0.723, 1.039)$

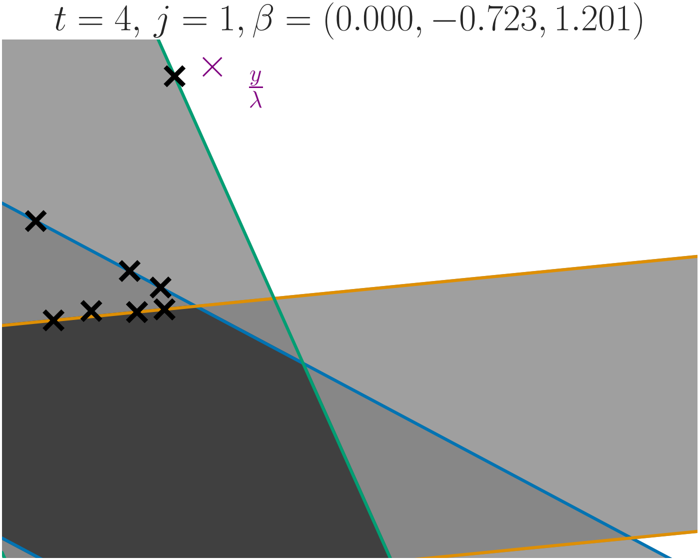


Residuals regularity, sign identification

$t = 3,\ j = 3, \beta = (0.000, -0.723, 1.201)$

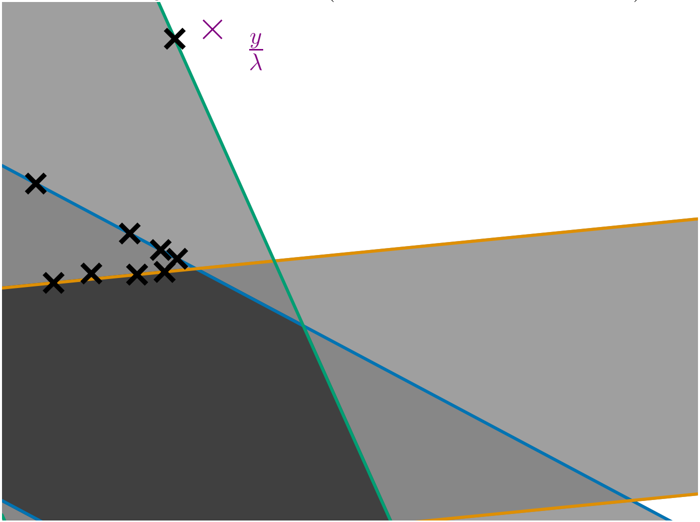


Residuals regularity, sign identification

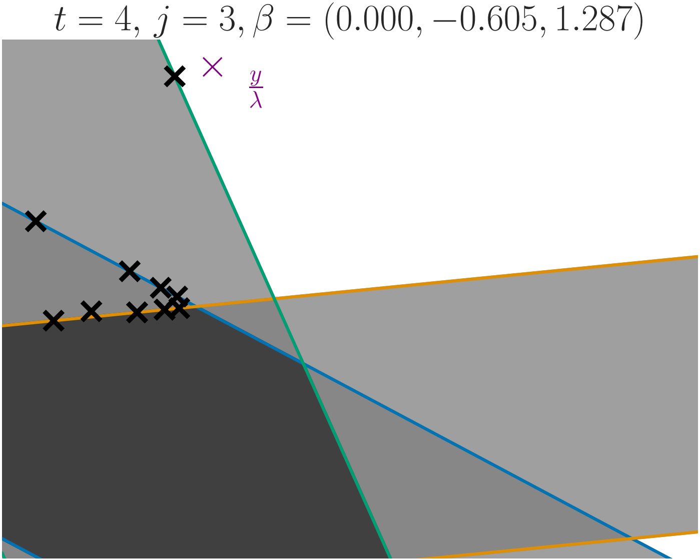


Residuals regularity, sign identification

$t = 4,\ j = 2, \beta = (0.000, -0.605, 1.201)$



Residuals regularity, sign identification



VAR regularity in residuals

Theorem[11]

Under uniqueness assumption, ISTA/CD achieves sign id.: $\operatorname{sign}\beta_j^{(t)} = \operatorname{sign}\hat{\beta}_j$. Then, Lasso residuals are Vector AutoRegressive (**VAR**):

$$\boxed{r^{(t+1)} = Ar^{(t)} + b}$$

$\hookrightarrow$ 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

Need a cheaper solution $\hookrightarrow$ **extrapolation**

[11] 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^*![12]

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

Generalization[13] to VAR $r^{(t)} \in \mathbb{R}^n$

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

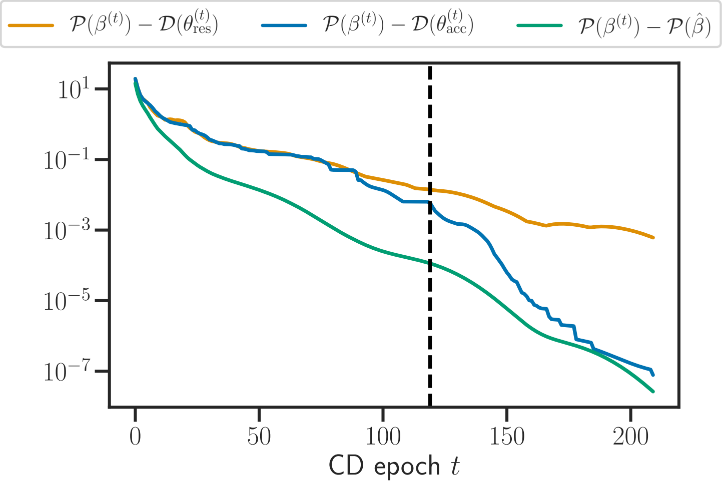
$$r_{\mathrm{accel}}^{(t)} \triangleq \sum_{k=1}^{K} c_k r^{(t+1-k)}$$

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

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

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

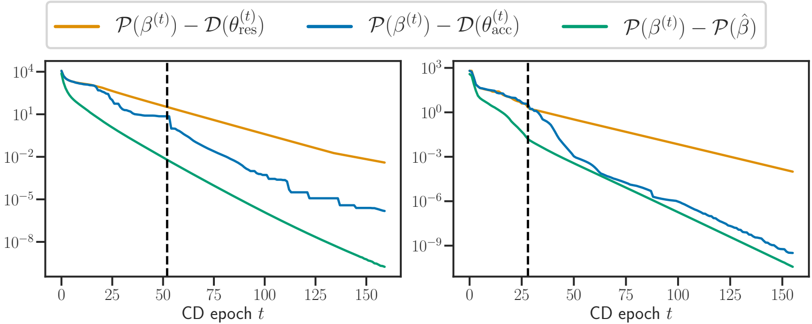
Dual extrapolation for the Lasso



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

Applicability to other models

We showed *asymptotic* VAR structure, also exploitable[14]



logreg, *rcv1* dataset:
$p = 20\text{k}, \quad n = 20\text{k}$
$\lambda = \lambda_{\max}/20 \ \ (\|\hat\beta\|_0 = 395)$

MTL, real **MEG** data:
$p = 7498, \quad n = 305$
$\lambda = \lambda_{\max}/10 \ \ (\|\hat{\mathrm{B}}\|_{2,0} = 45)$

[14] M. Massias et al. "Dual extrapolation for sparse Generalized Linear Models". In: *submission to JMLR* (2019).

Additional speed-ups

Two approaches:

- safe screening[15,16] (**backward approach**): remove feature j when it is certified that $\hat{\beta}_j = 0$
- working set[17] (**forward approach**): focus on j's for which it is very likely that $\hat{\beta}_j \neq 0$

---

[15] 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.

[16] A. Bonnefoy et al. "A dynamic screening principle for the lasso". In: *EUSIPCO*. 2014.

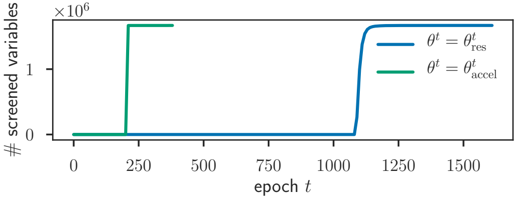
[17] T. B. Johnson and C. Guestrin. "Blitz: A Principled Meta-Algorithm for Scaling Sparse Optimization". In: *ICML*. 2015, pp. 1171–1179.

Duality strikes again: gap safe screening

Gap safe screening rule[18]:

$$\boxed{\forall (\beta, \theta) \in \mathbb{R}^p \times \Delta_X, \quad |X_j^\top \theta| < 1 - \|X_j\| \sqrt{\frac{2}{\lambda^2} \mathsf{dgap}(\beta, \theta)} \Rightarrow \hat{\beta}_j = 0}$$

better dual point $\Rightarrow$ better gap safe screening

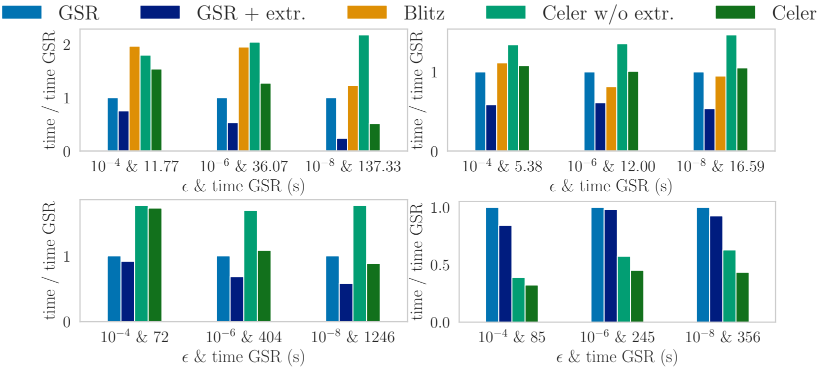


Finance dataset: $(p = 1.5 \times 10^6, n = 1.5 \times 10^4)$, $\lambda = \lambda_{\max}/5$

[18] E. Ndiaye et al. "Gap Safe screening rules for sparsity enforcing penalties". In: *JMLR* 18.128 (2017), pp. 1–33.

Celer: working sets with extrapolation & aggressive screening

Screening can be used aggressively to define WS[19], a **better dual point also helps**



rcv1 dataset, fine and coarse Lasso path (wrt Gap Safe Rules)
news20 dataset, fine and coarse Logreg path (wrt Gap Safe Rules)

[19] M. Massias, A. Gramfort, and J. Salmon. "From safe screening rules to working sets for faster Lasso-type solvers". In: *NIPS-OPT workshop*. 2017.

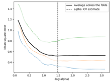
Contributions

- Improved the numerical efficiency of Lasso-type solvers.
  Duality used in stopping criterion & safe feature identification
  (asymptotic) VAR structure of $X\beta^{(t)} \hookrightarrow$ better dual

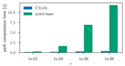
- https://github.com/mathurinm/celer:
  fast implementation, standard API, documented, easy to install and reproducible benchmarks:

Fork me on GitHub

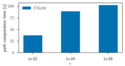
Examples Gallery¶



Run LassoCV for cross-validation on Leukemia



Lasso path computation on Leukemia dataset



Lasso path computation on Finance/log1p

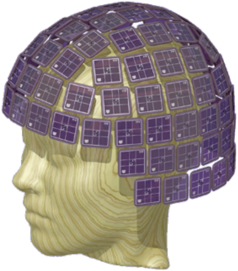
Table of Contents

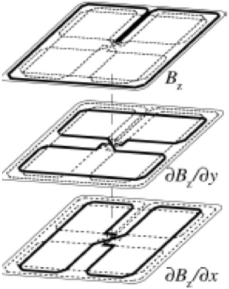
Solving Lasso-type problems, fast

Noise modeling and pivotality

MEG sensors: magnetometers and gradiometers

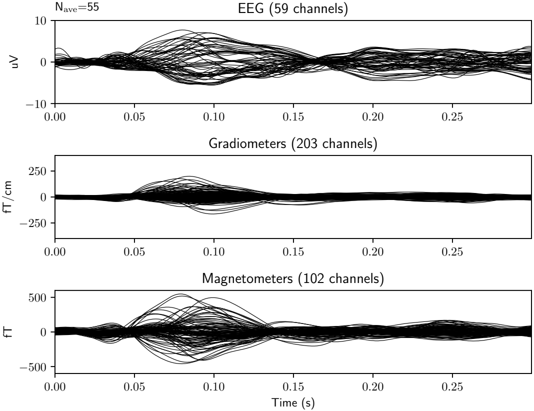

Device

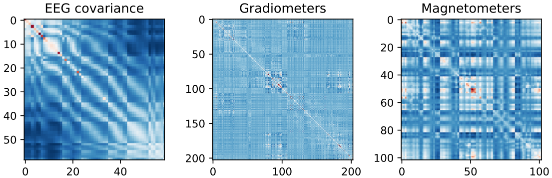

Sensors



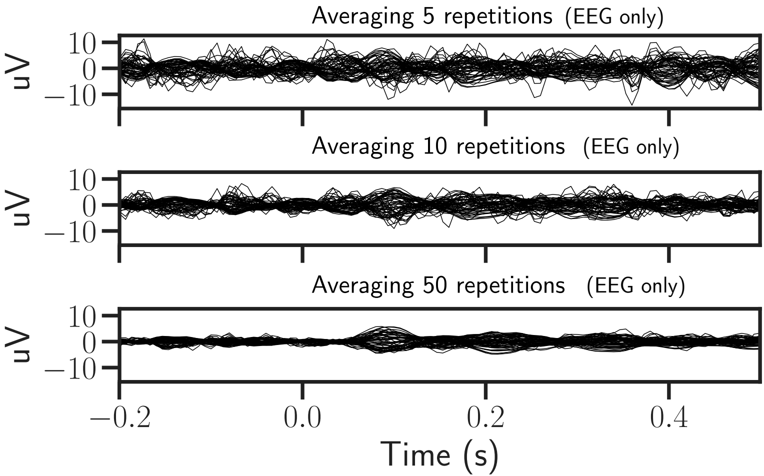
Detail of a sensor

3 sensor types $\Longrightarrow$ 3 noise structures





Low SNR: averaging repetitions of experiment



Our complete stats model for M/EEG

- n observations (number of sensors)
- T tasks (temporal information)
- p features (spatial description)
- r repetitions of the same 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

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

- $\mathrm{B}^* \in \mathbb{R}^{p \times T}$: true source activity matrix (**unknown**)
- $S_* \in \mathbb{S}^n_{++}$ co-standard deviation matrix (**unknown**)
- $\mathrm{E}^{(1)}, \dots, \mathrm{E}^{(r)} \in \mathbb{R}^{n \times T}$: 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
- need to go beyond squared Frobenius norm:

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

$\hookrightarrow$ yields the same $\hat{\mathrm{B}}$

Lasso and optimal λ[20,21]

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

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

BUT σ_* is unknown in practice !

[20] 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.

[21] 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 $\sqrt{\text{Lasso}}$[23]

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

has an optimal λ independent of σ_*

hard to optimize $\hookrightarrow$ use *Concomitant* Lasso[22] formulation (introduced earlier!):

$$\boxed{(\hat{\beta}_{\text{conco}}, \hat{\sigma}_{\text{conco}}) \in \arg\min_{\beta \in \mathbb{R}^p, \sigma \geq 0} \frac{\|y - X\beta\|^2}{2n\sigma} + \frac{\sigma}{2} + \lambda \|\beta\|_1}$$

same solutions when $\|y - X\hat{\beta}_{\sqrt{\text{Lasso}}}\| \neq 0$

---

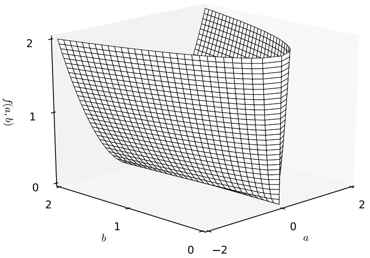
[22] A. B. Owen. "A robust hybrid of lasso and ridge regression". In: *Cont. Math.* 443 (2007), pp. 59–72.

[23] 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.

Easy to solve

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

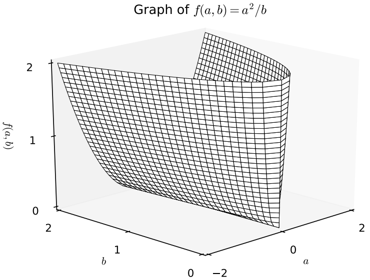
Jointly convex formulation, smooth + separable: optimized by alternate minimization w.r.t. β and σ



Easy to solve

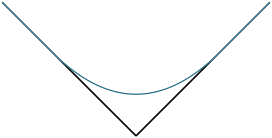
$$\boxed{(\hat{\beta}, \hat{\sigma}) \in \mathop{\arg\min}_{\beta \in \mathbb{R}^p, \sigma \geq \underline{\sigma}} \frac{\|y - X\beta\|^2}{2n\sigma} + \frac{\sigma}{2} + \lambda \|\beta\|_1}$$

Jointly convex formulation, smooth + separable: optimized by alternate minimization w.r.t. β and σ
Smooth Concomitant: gradient is Lipschitz!



Concomitant origin: smoothing the $\sqrt{\mathrm{Lasso}}$[26]

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



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

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

Leads to the Smoothed[24,25] Concomitant Lasso formulation:

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

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

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

[26] 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):[27]

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

S. van de Geer introduced the pivotal multivariate $\sqrt{\text{Lasso}}$[28]:

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

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

Smoothing makes optimization and statistical analysis easy!

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

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

Leveraging repetitions

Smoothed Generalized Concomitant Lasso (SGCL)[29]

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

Concomitant Lasso with Repetitions (CLaR)[30]

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

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

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

SGCL and CLaR computations: B update

Alternate minimization converges

B **update (S fixed)**: standard MTL optimization, off-the-shelf techniques and refinements of 1st part

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:

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

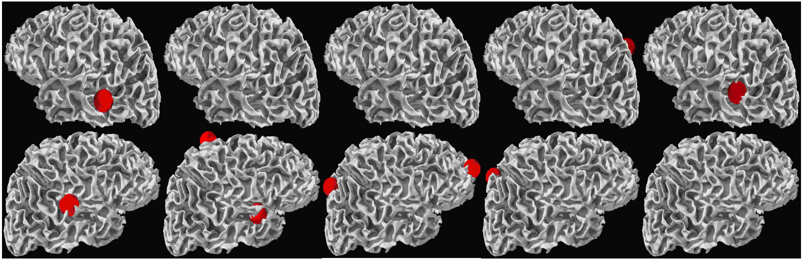
Leveraging repetitions

- Statistically: $\mathcal{O}(n^2)$ parameters to estimate for S
  - SGCL: only nT observations (need T large w.r.t. n)
  - CLaR: nTr observations
- Computationally: SVD for S costs $\mathcal{O}(n^3)$, high in general but fine for MEG/EEG problems ($n \approx 300$)

Rem: more structure can easily be incorporated to estimate S, *e.g.*, block diagonal, etc.

Dissemination

MLE: $$\underset{\substack{\mathrm{B}\in\mathbb{R}^{p\times T}\\ \Sigma\succ 0}}{\arg\min}\ \frac{1}{2}\|\bar{Y} - X\mathrm{B}\|^2_{\Sigma^{-1}} - \log\det(\Sigma^{-1}) + \lambda\|\mathrm{B}\|_{2,1}$$



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

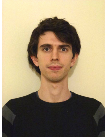
Figure: *Real data, sources found after right auditory stimulations.*

SGCL & CLaR, https://github.com/mathurinm/sgcl
Also implemented non convex time frequency solvers as a preliminary step to my algorithms in MNE

Thank you

Joseph, Alexandre, Olivier, Samuel, Quentin, Pierre A, 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. Soon resubmitted to JMLR*, 2019
- M. Massias*, Q. Bertrand*, A. Gramfort, and J. Salmon. *Support recovery and sup-norm convergence rates for sparse pivotal estimation. Under review for AISTATS*, 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. 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.

References II

- Bonnefoy, A. et al. "A dynamic screening principle for the lasso". In: *EUSIPCO*. 2014.
- 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.
- 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.

References III

- Fan, R.-E. et al. “LIBLINEAR: A library for large linear classification”. In: *JMLR* 9 (2008), pp. 1871–1874.
- Friedman, J., T. J. Hastie, and R. Tibshirani. “Regularization paths for generalized linear models via coordinate descent”. In: *J. Stat. Softw.* 33.1 (2010), p. 1.
- 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.
- Mairal, J. “Sparse coding for machine learning, image processing and computer vision”. PhD thesis. École normale supérieure de Cachan, 2010.

References IV

- Massias, M., A. Gramfort, and J. Salmon. "Celer: a fast solver for the Lasso with dual extrapolation". In: *ICML*. 2018, pp. 3321–3330.
- – ."From safe screening rules to working sets for faster Lasso-type solvers". In: *NIPS-OPT workshop*. 2017.
- Massias, M. et al. "Dual extrapolation for sparse Generalized Linear Models". In: *submission to JMLR* (2019).
- Massias, M. et al. "Generalized concomitant multi-task Lasso for sparse multimodal regression". In: *AISTATS*. 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.
- Ndiaye, E. et al. "Gap Safe screening rules for sparsity enforcing penalties". In: *JMLR* 18.128 (2017), pp. 1–33.

References V

- 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.
- 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.

References VI

- Tibshirani, R. J. "The lasso problem and uniqueness". In: *Electron. J. Stat.* 7 (2013), pp. 1456–1490.
- van de Geer, S. *Estimation and testing under sparsity*. Lecture Notes in Mathematics. Springer, 2016.