Support recovery and sup-norm convergence rates for sparse pivotal estimation

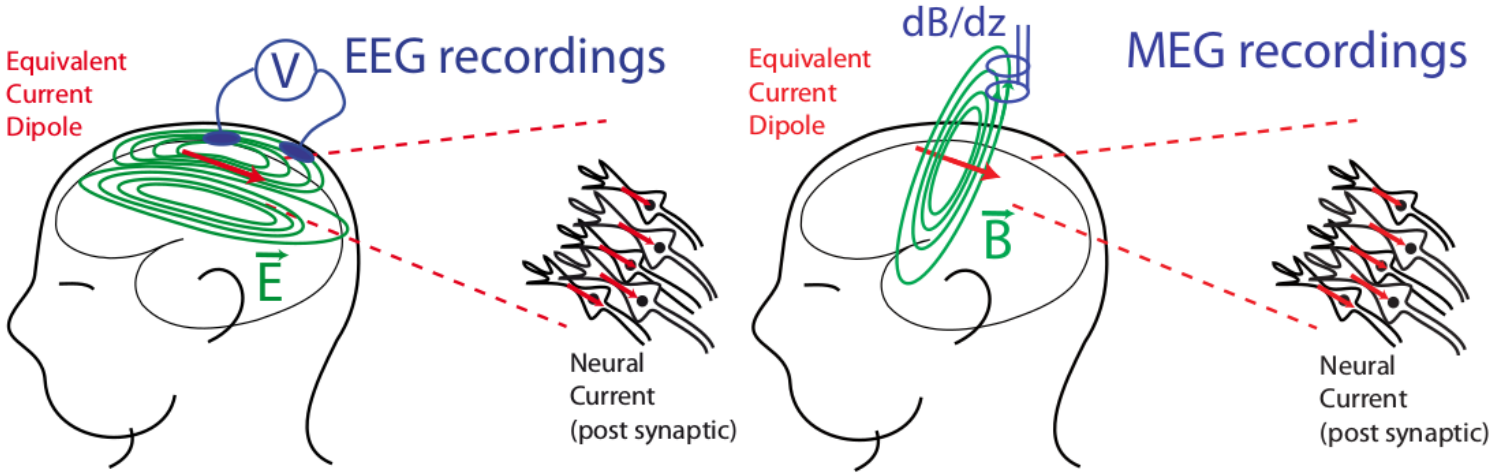
Quentin Bertrand

Joint work with:
Mathurin Massias (University of Genova)
Alexandre Gramfort (INRIA)
Joseph Salmon (IMAG, Univ Montpellier, CNRS)

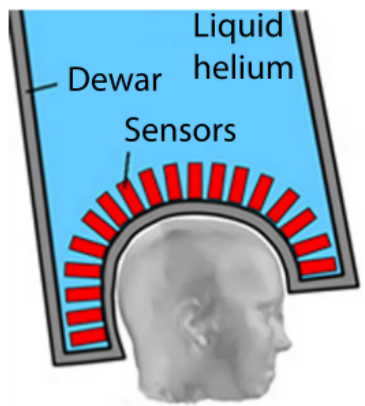
Table of Contents

M/EEG inverse problem for brain imaging

- sensors: electric and magnetic fields during a cognitive task
- goal: which parts of the brain are responsible for the signals?
- applications: epilepsy treatment, brain aging, anesthesia risks

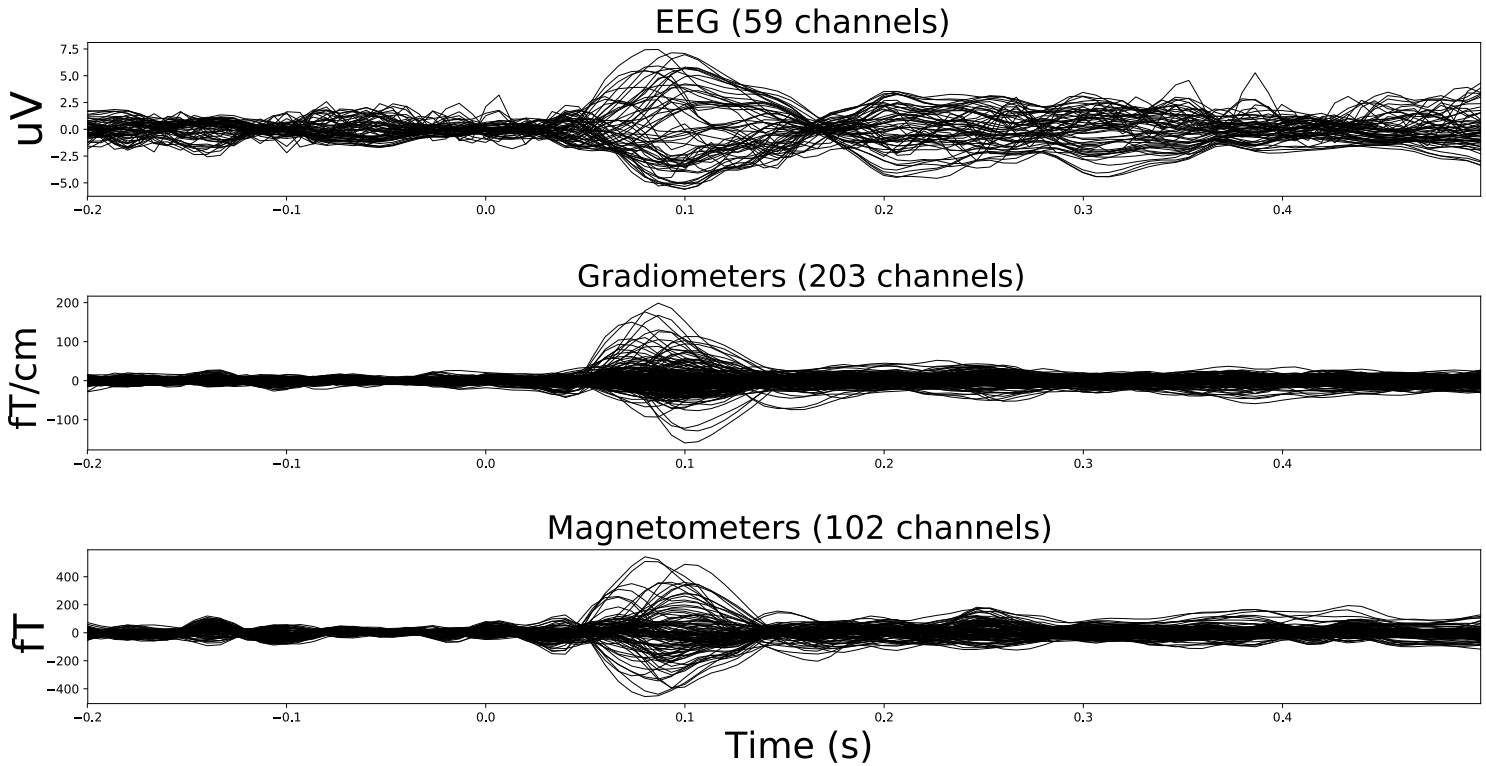


First EEG recordings in 1929 by H. Berger



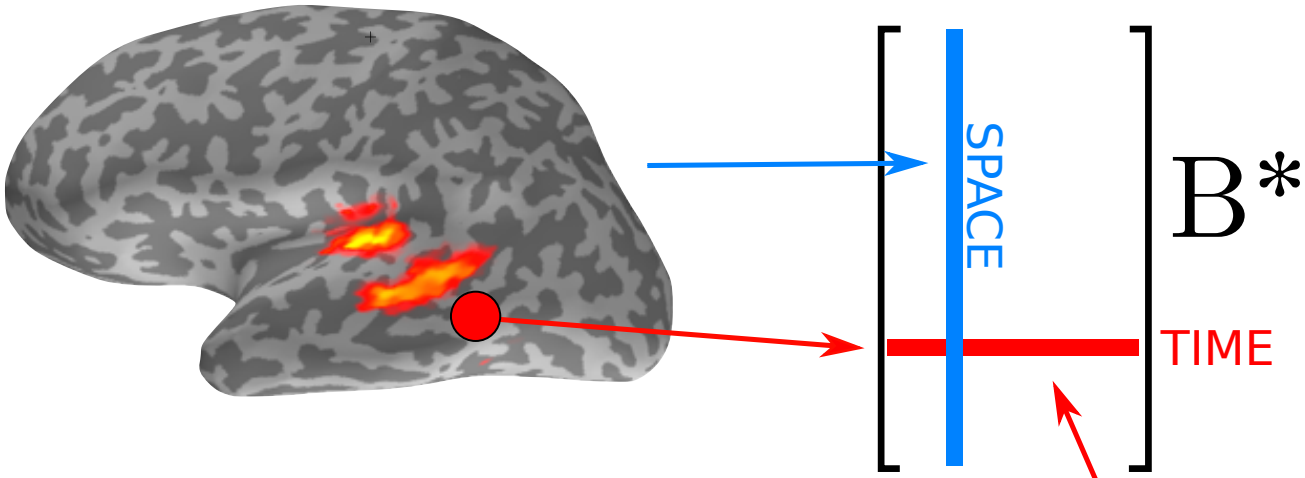
Hôpital La Timone
Marseille, France

M/EEG data

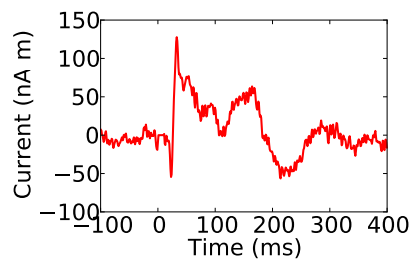


- 3 different types of sensor

Source modeling (discretization with voxels)

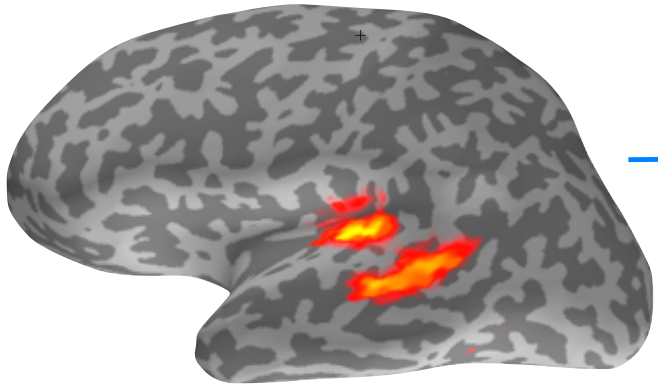


Position a few thousands candidate sources over the brain (*e.g.*, every 5mm)



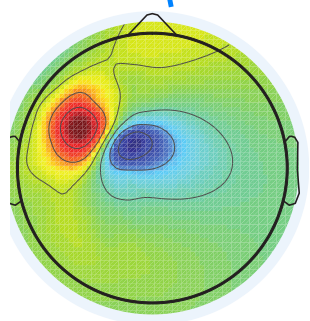
$$\mathrm{B}^* \in \mathbb{R}^{p \times T}$$

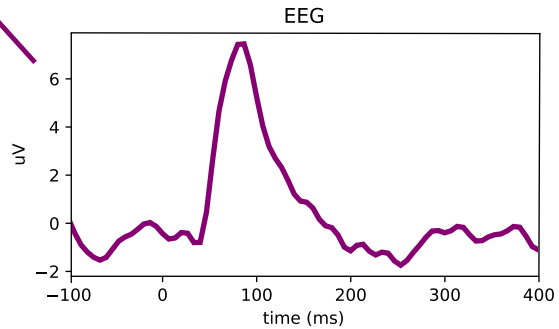
The M/EEG inverse problem: modeling

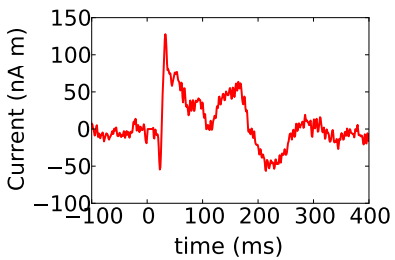


$$T \approx 100 \qquad p \approx 10000$$

$$n \approx 100 \qquad Y = X \qquad B^* \qquad + E$$

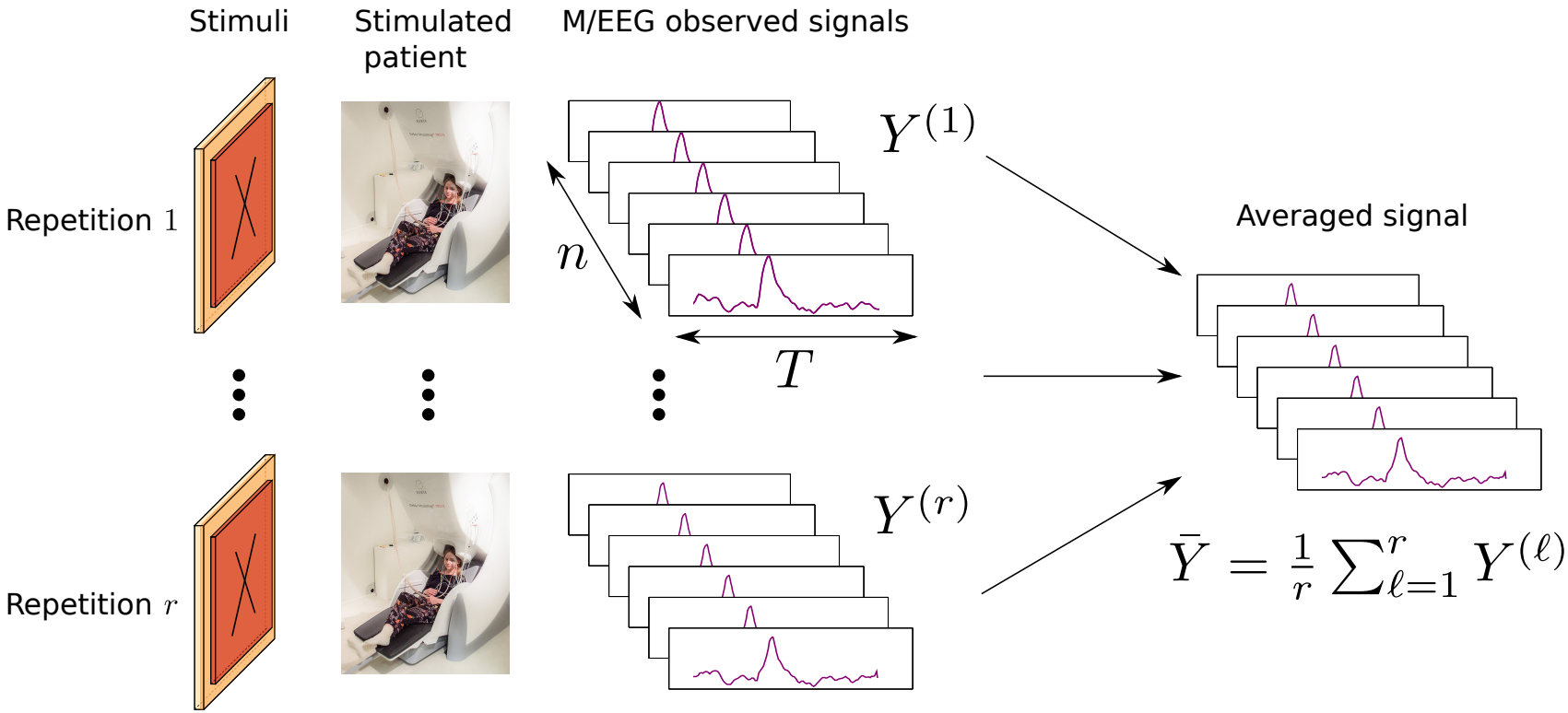






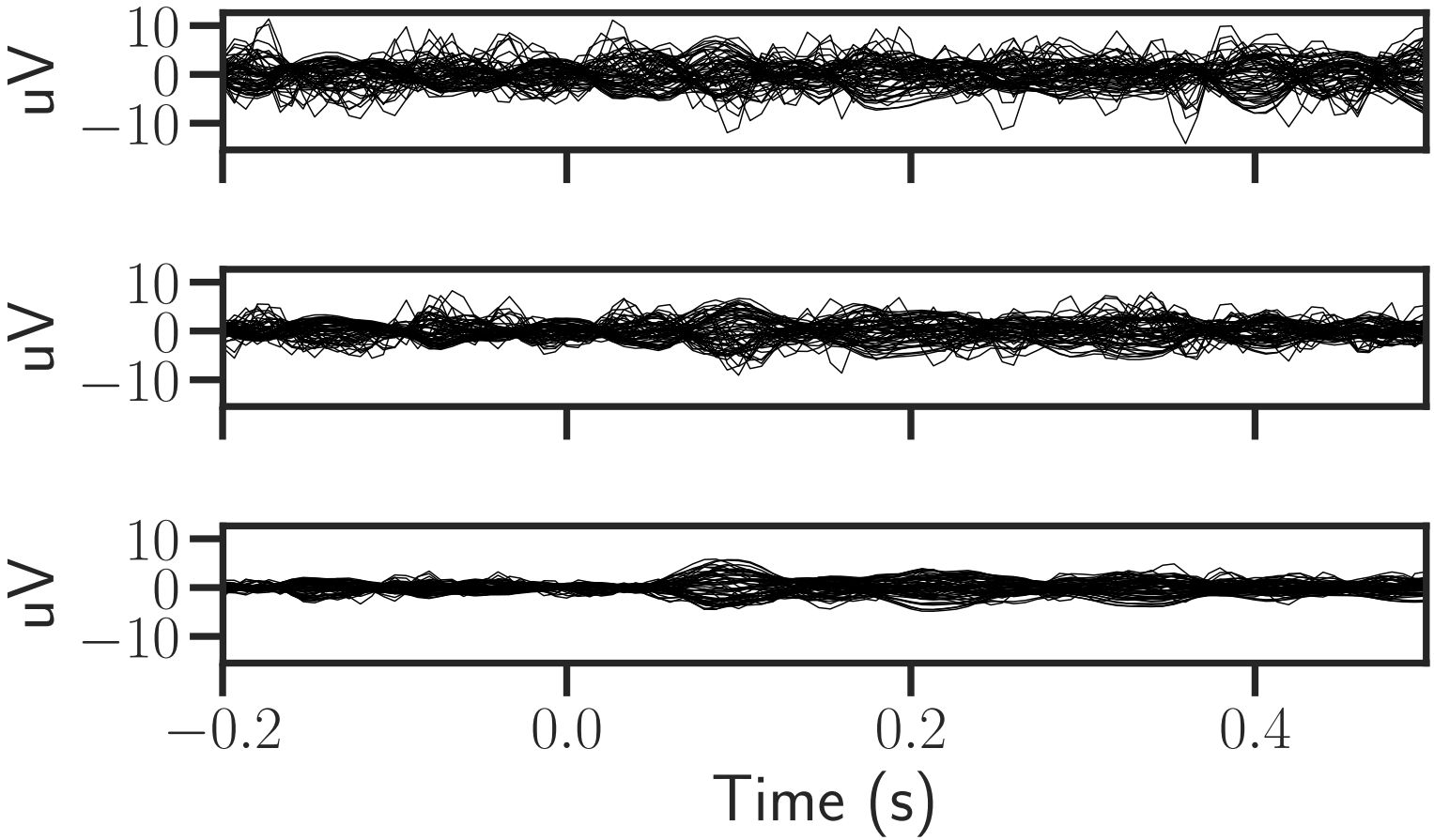
$$n << p$$

Very noisy data: must repeat recordings



Very noisy data: must repeat recordings

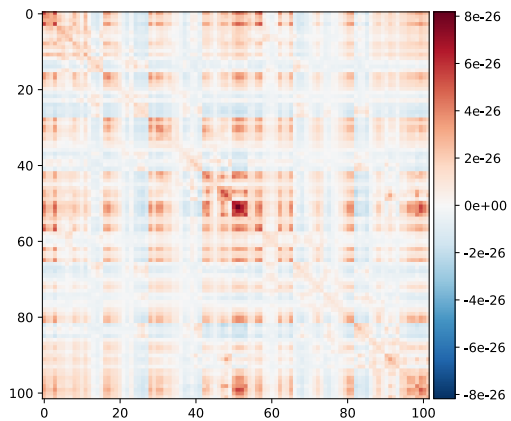
- average of 5 (top) / 10 (middle) / 50 (bottom) repetitions



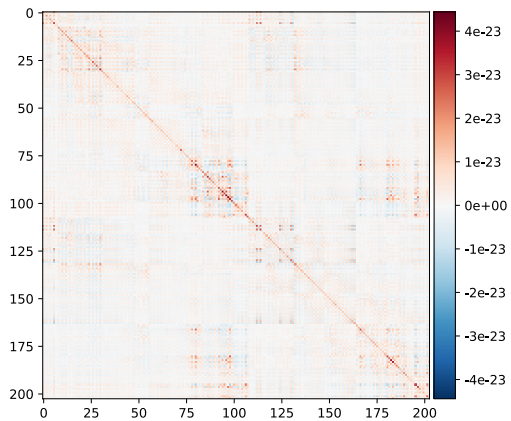
Noise covariance for each type of sensor

$$Y = X\mathrm{B} + \mathrm{E}, \quad \mathrm{E} \sim \mathcal{N}(0, \Sigma) \tag{1}$$

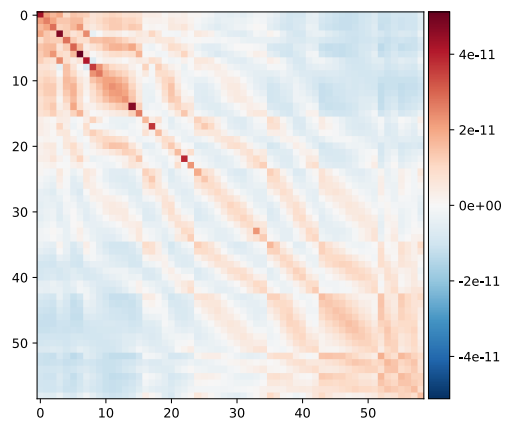
Magnetometers



Gradiometers



EEG



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

A Multi-Task framework

Multi-Task regression notation:

- n observations (*e.g.*, number of sensors)
- T tasks (*e.g.*, temporal information)
- p features
- r number of repetitions
- $Y^{(1)}, \ldots, Y^{(r)} \in \mathbb{R}^{n \times T}$ observation matrices; $\bar{Y} = \frac{1}{r} \sum_l Y^{(l)}$
- $X \in \mathbb{R}^{n \times p}$ design matrix (known)

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

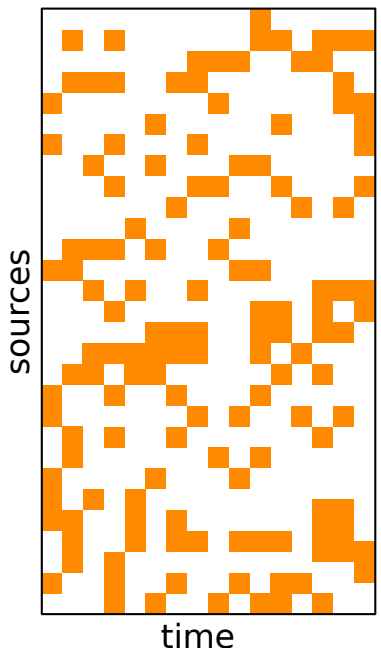
where

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

Multi-Task penalties[1]

Popular convex penalties considered:

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



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

Sparse support: no structure

Penalty: **Lasso type**

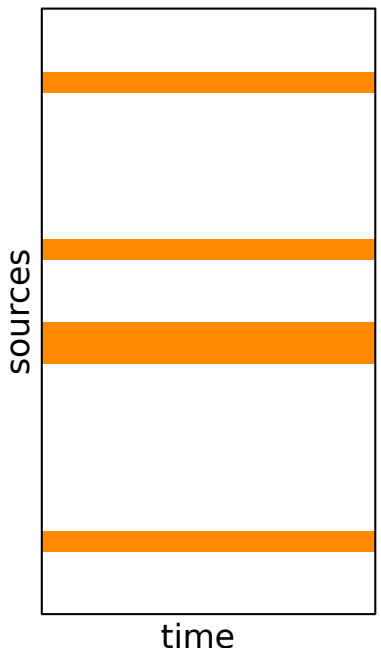
$$\Omega(\mathrm{B}) = \|\mathrm{B}\|_1 = \sum_{j=1}^{p} \sum_{k=1}^{T} |\mathrm{B}_{j,k}|$$

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

Multi-Task penalties[1]

Popular convex penalties considered: Multi-Task Lasso (MTL)

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



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

Sparse support: group structure ✓

Penalty: **Group-Lasso type**

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

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

[1] 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.

Summary

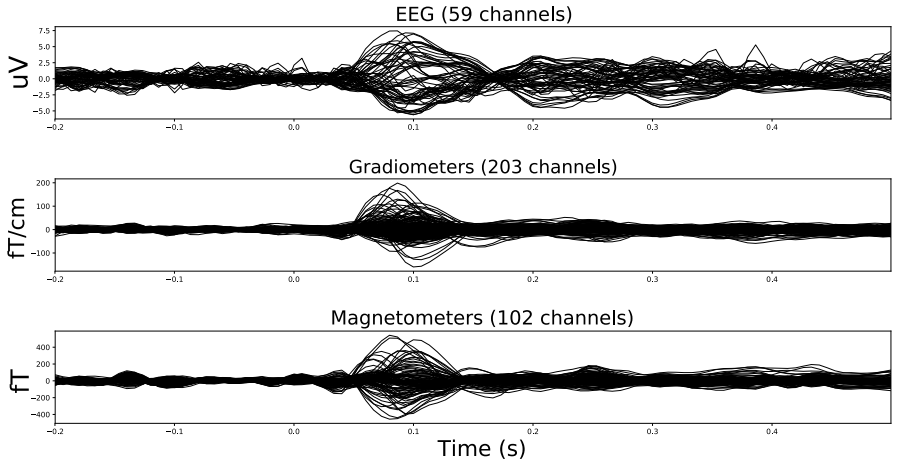


Figure: What you have.

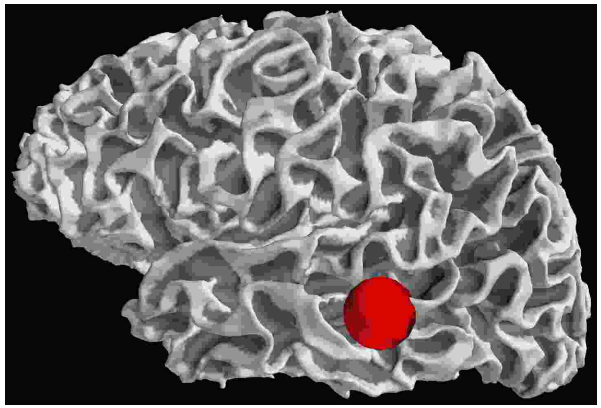


Figure: What you want.

This is typically done using optimization based estimators:

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

Table of Contents

Reminder on the Lasso theory[2] (i.i.d. case, Single-Task, $y = X\beta + \sigma^* \varepsilon$)

$$\hat{\beta} \in \arg\min_{\beta \in \mathbb{R}^p} \frac{1}{2n} \|y - X\beta\|_2^2 + \lambda \|\beta\|_1$$

Theorem

- i.i.d. Gaussian noise
- $+$ X satisfying the "Restricted Eigenvalue" property
- $+$ $\lambda = 2\sigma_* \sqrt{\frac{2 \log(p/\delta)}{n}}$
- $\Longrightarrow$ with probability $1 - \delta$:

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

BUT σ_* is unknown in practice !

[2] 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.

Reminder on the Lasso theory[2] (i.i.d. case, Single-Task, $y = X\beta + \sigma^* \varepsilon$)

$$\hat{\beta} \in \arg\min_{\beta \in \mathbb{R}^p} \frac{1}{2n} \|y - X\beta\|_2^2 + \lambda \|\beta\|_1$$

Theorem

- i.i.d. Gaussian noise
- $+ X$ satisfying the "Restricted Eigenvalue" property
- $+ \lambda = 2\sigma_* \sqrt{\frac{2 \log(p/\delta)}{n}}$
- $\Longrightarrow$ with probability $1 - \delta$:

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

BUT σ_* is unknown in practice !

[2] 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.

Reminder on the square-root Lasso[3][4][5] (i.i.d. case, Single-Task)

$$\hat{\beta} \in \arg\min_{\beta \in \mathbb{R}^p} \frac{1}{\sqrt{n}} \|y - X\beta\|_2 + \lambda \|\beta\|_1$$

Theorem

- i.i.d. Gaussian noise
- $+ X$ satisfying the "Restricted Eigenvalue" property
- $+ \lambda = 2\sqrt{\frac{2\log(p/\delta)}{n}}$
- $\Longrightarrow$ with high probability:

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

λ does not depend on σ_* anymore!

[3] 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.

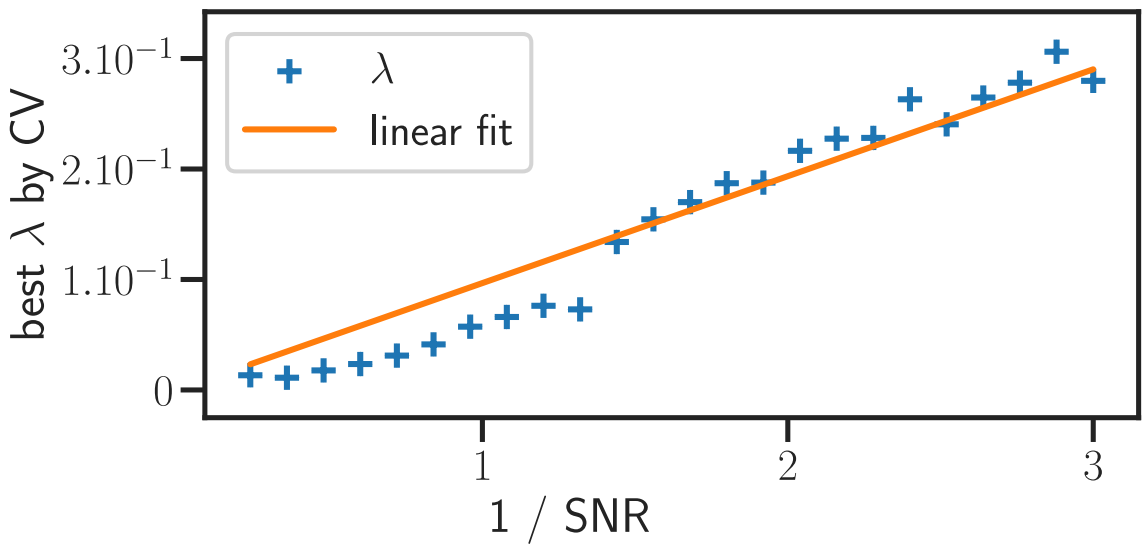
[4] T. Sun and C.-H. Zhang. "Scaled sparse linear regression". In: *Biometrika* 99.4 (2012), pp. 879–898.

[5] C. Giraud. *Introduction to high-dimensional statistics*. Vol. 138. CRC Press, 2014.

Example

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

Confirmed in practice:

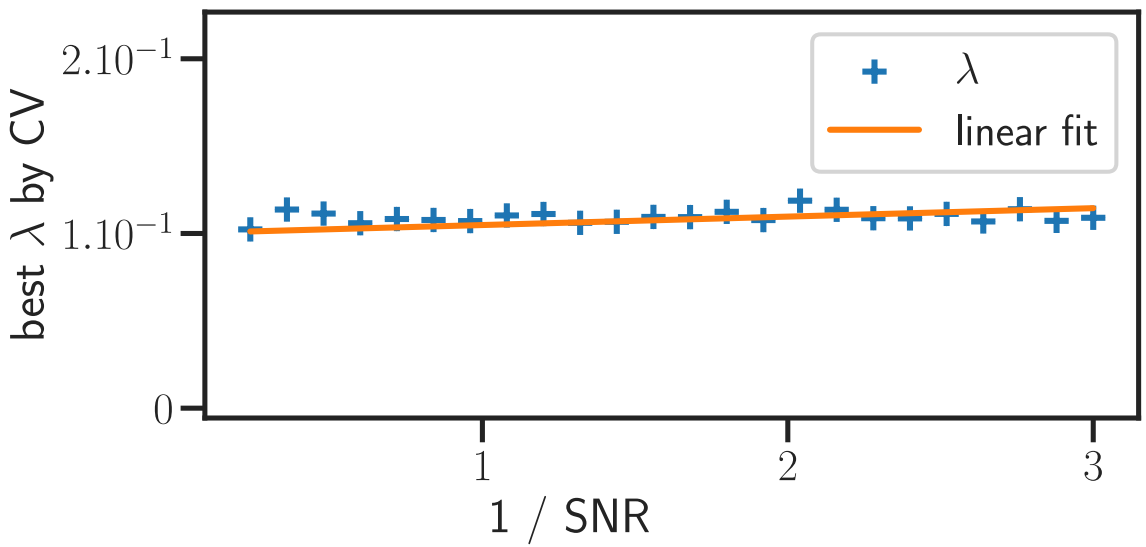


Lasso

Example

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

Confirmed in practice:



Square-root Lasso

The Smoothed Concomitant Lasso[6] (i.i.d. case, Single-Task)

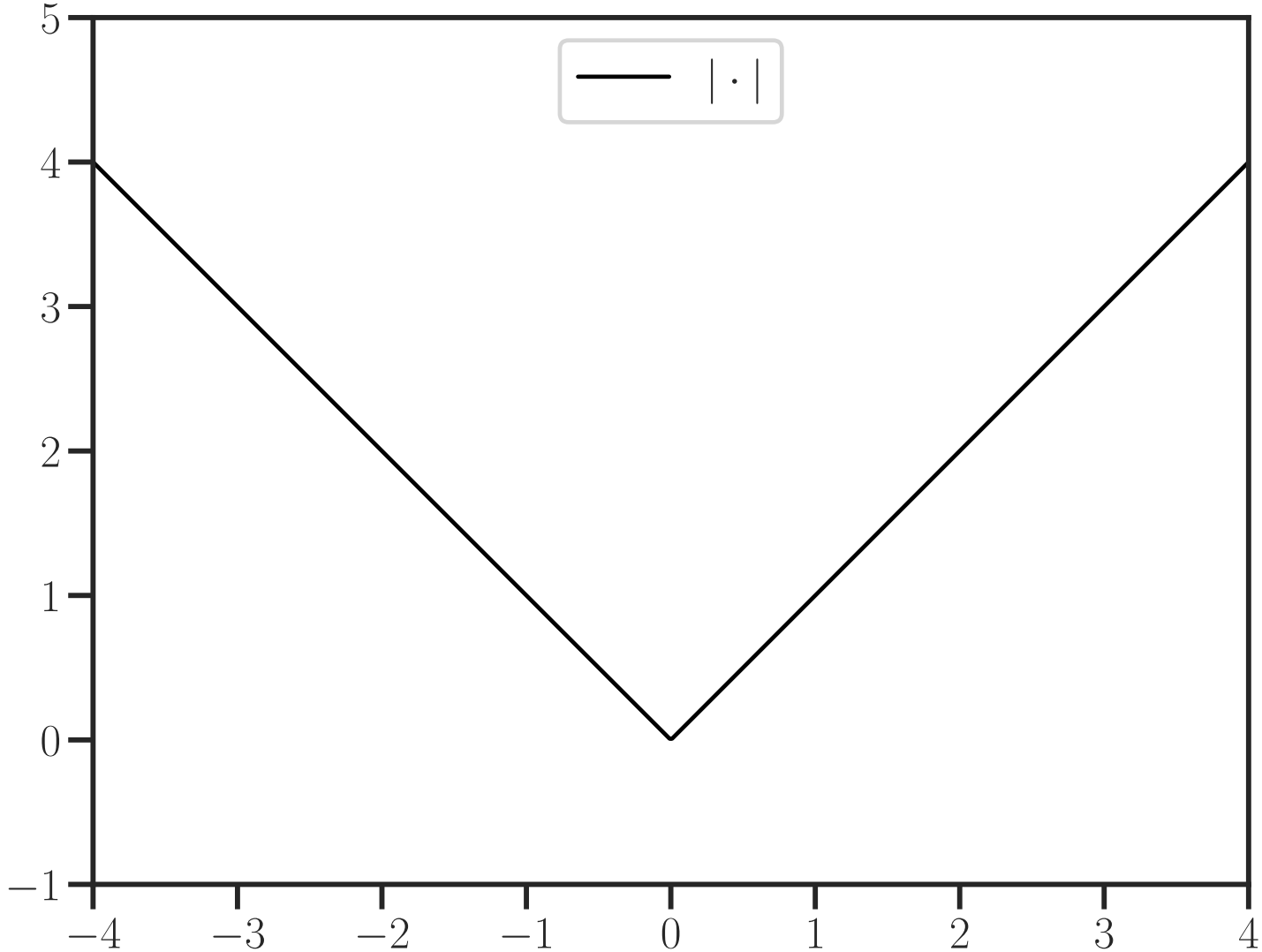
$$\hat{\beta}^{(\lambda)} \in \arg\min_{\beta \in \mathbb{R}^p} \frac{1}{\sqrt{n}} \underbrace{\|y - X\beta\|_2}_{\text{non-smooth}} + \lambda \underbrace{\|\beta\|_1}_{\text{non-smooth}}$$

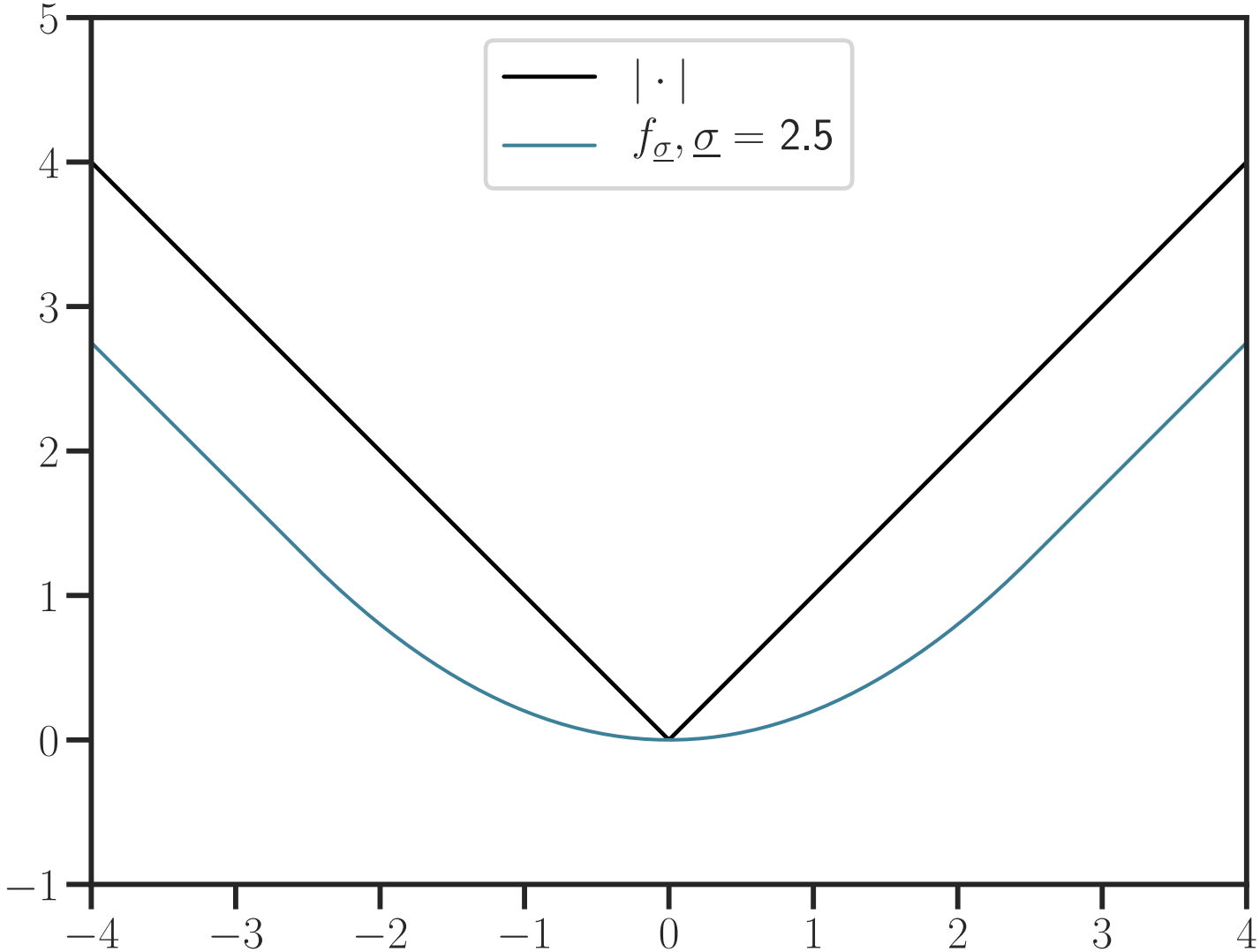
Idea: replacing $\|\cdot\|_2$ by $\underbrace{\|\cdot\|_2 \square \underline{\sigma}\, \omega\left(\frac{\cdot}{\underline{\sigma}}\right)}_{\text{smooth}}(z) = \min_{\sigma \geq \underline{\sigma}} \left(\frac{\|z\|_2^2}{2\sigma} + \frac{\sigma}{2}\right)$

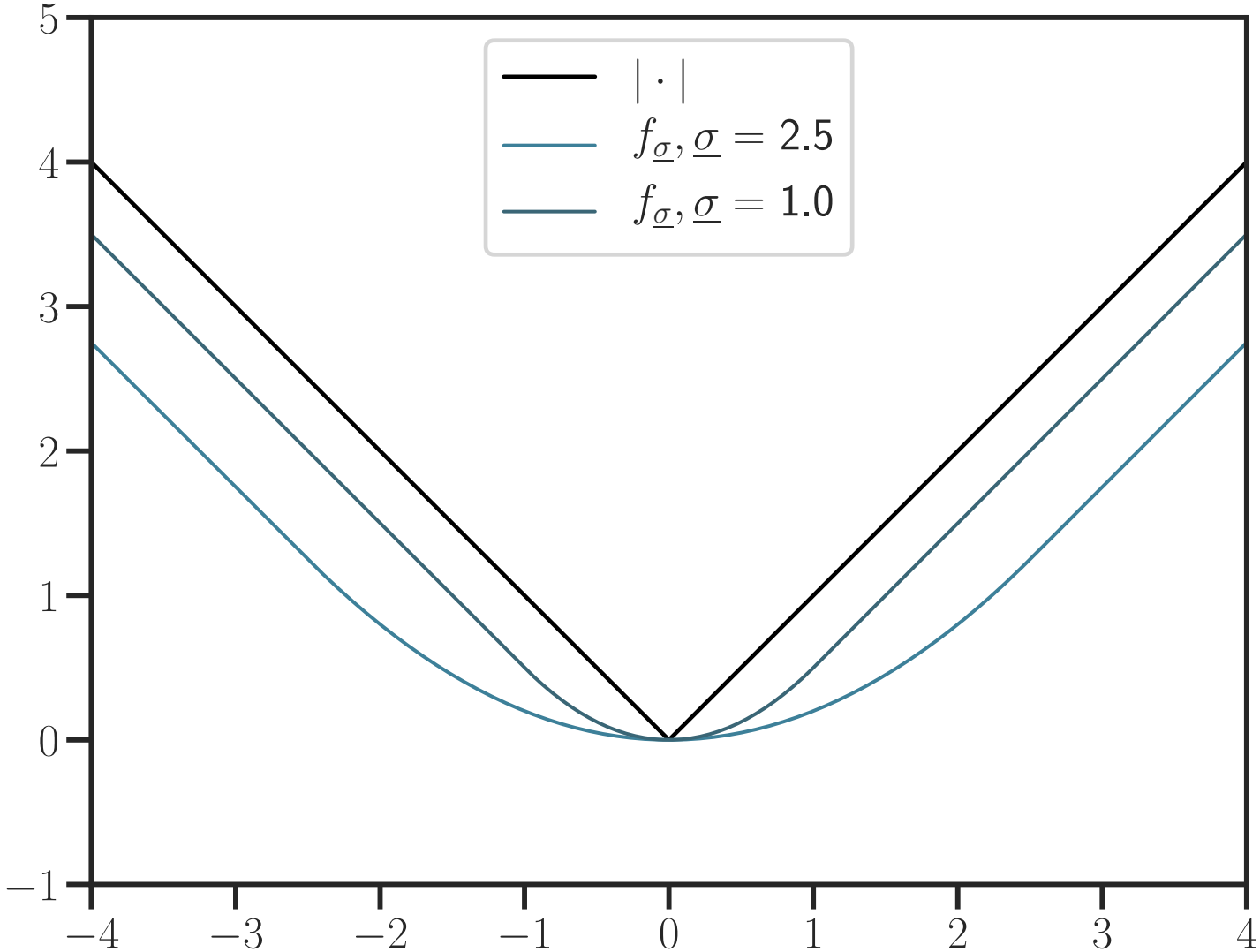
$$\boxed{(\hat{\beta}^{(\lambda)}, \hat{\sigma}^{(\lambda)}) \in \arg\min_{\beta \in \mathbb{R}^p, \sigma \geq \underline{\sigma}} \frac{\|y - X\beta\|_2^2}{2n\sigma} + \frac{\sigma}{2} + \lambda \|\beta\|_1}$$

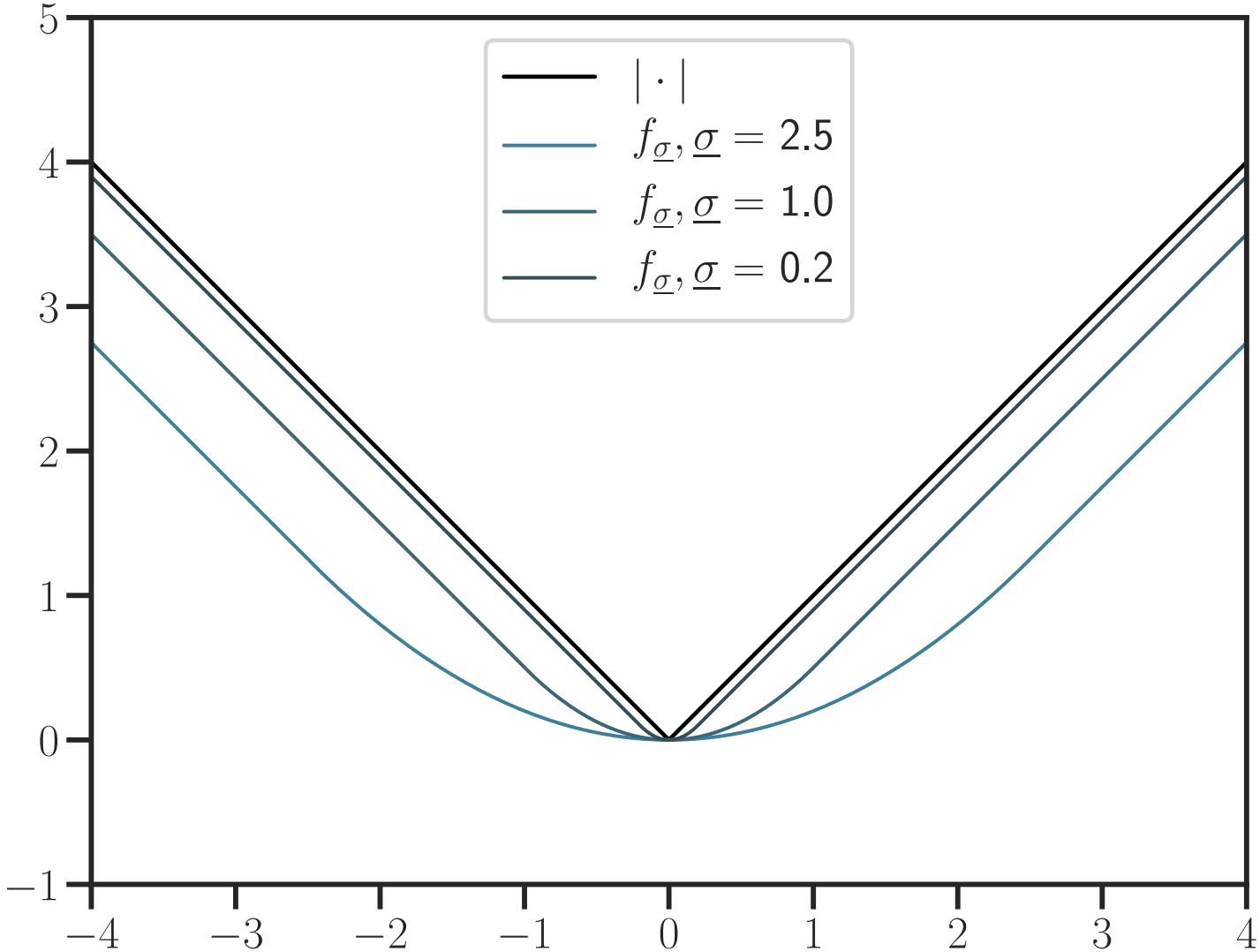
- jointly convex: alternate minimization

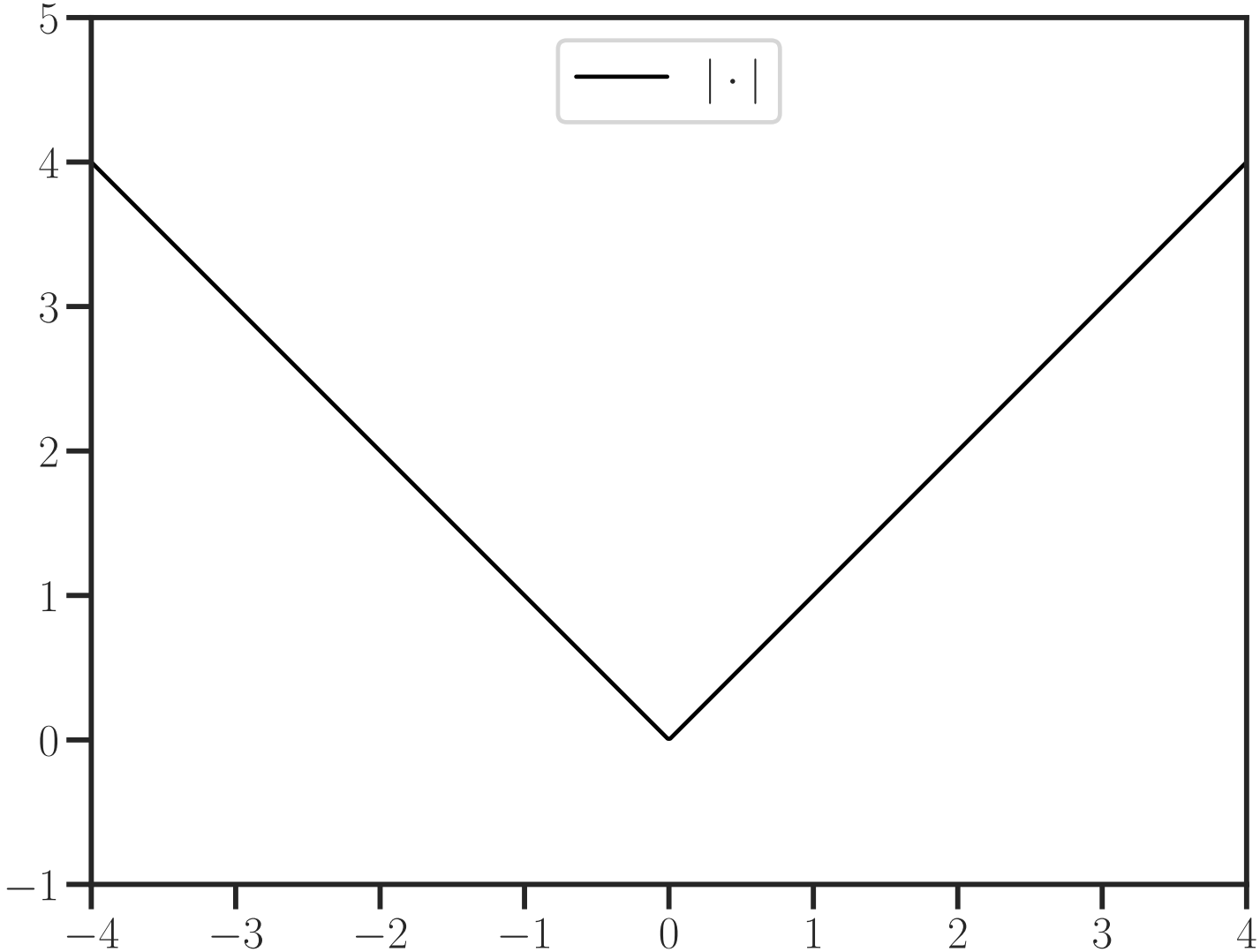
[6] E. Ndiaye et al. "Efficient Smoothed Concomitant Lasso Estimation for High Dimensional Regression". In: *Journal of Physics: Conference Series* 904.1 (2017), p. 012006.

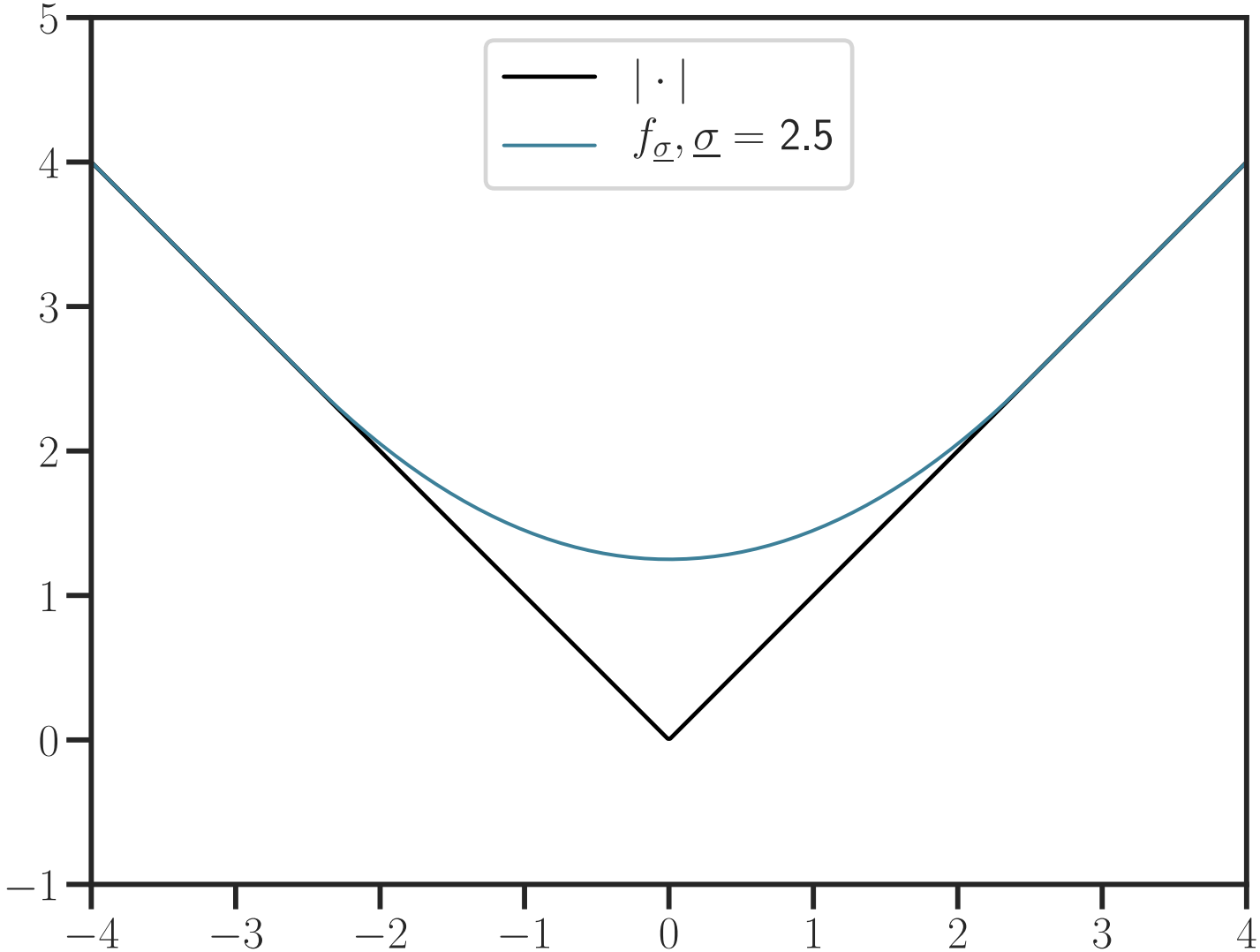

|·|
5
4
3
2
1
0
−1
−4
−3
−2
−1
0
1
2
3
4

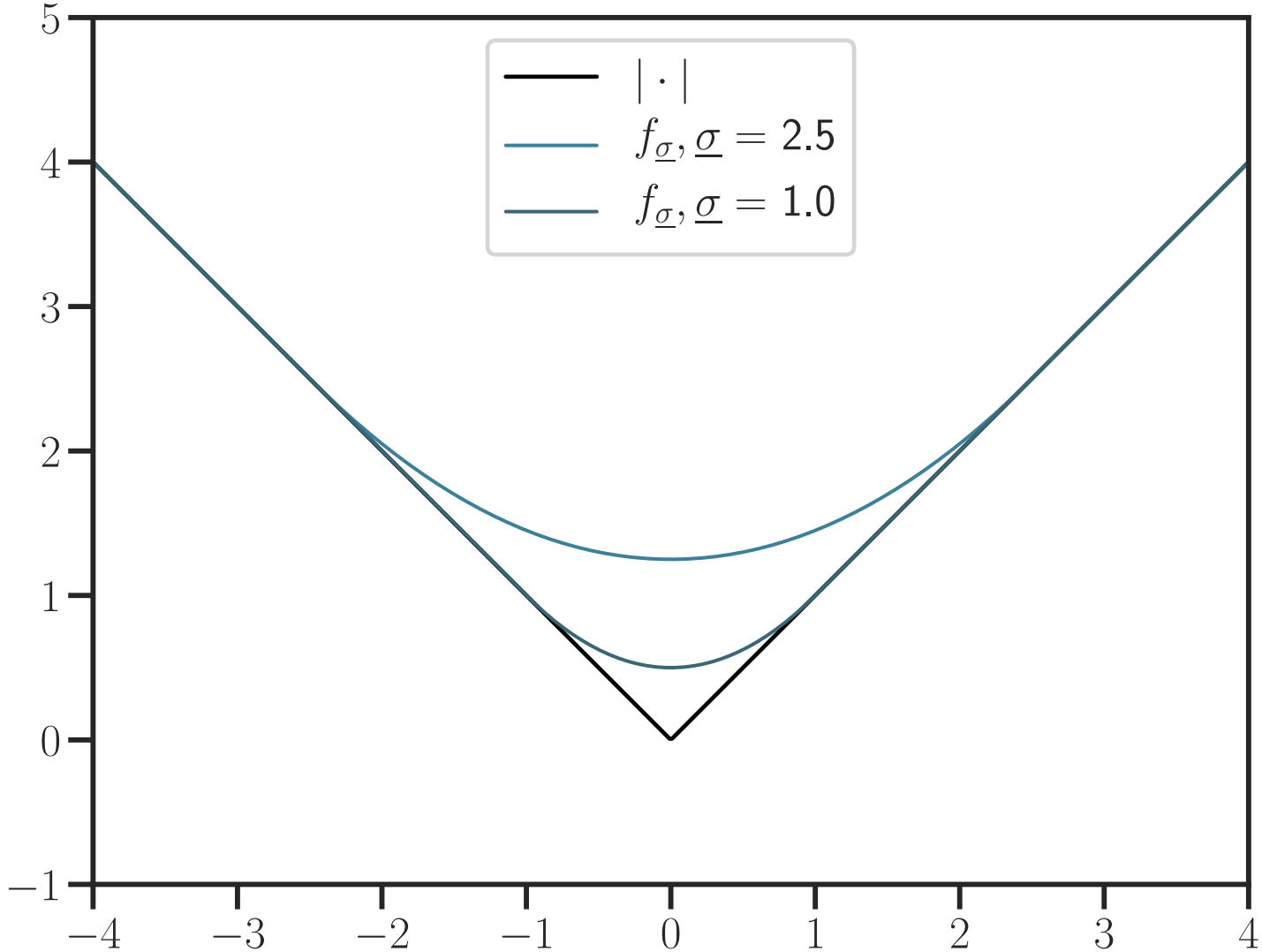

$|\cdot|$
$f_{\underline{\sigma}}, \underline{\sigma} = 2.5$
5
4
3
2
1
0
−1
−4
−3
−2
−1
0
1
2
3
4

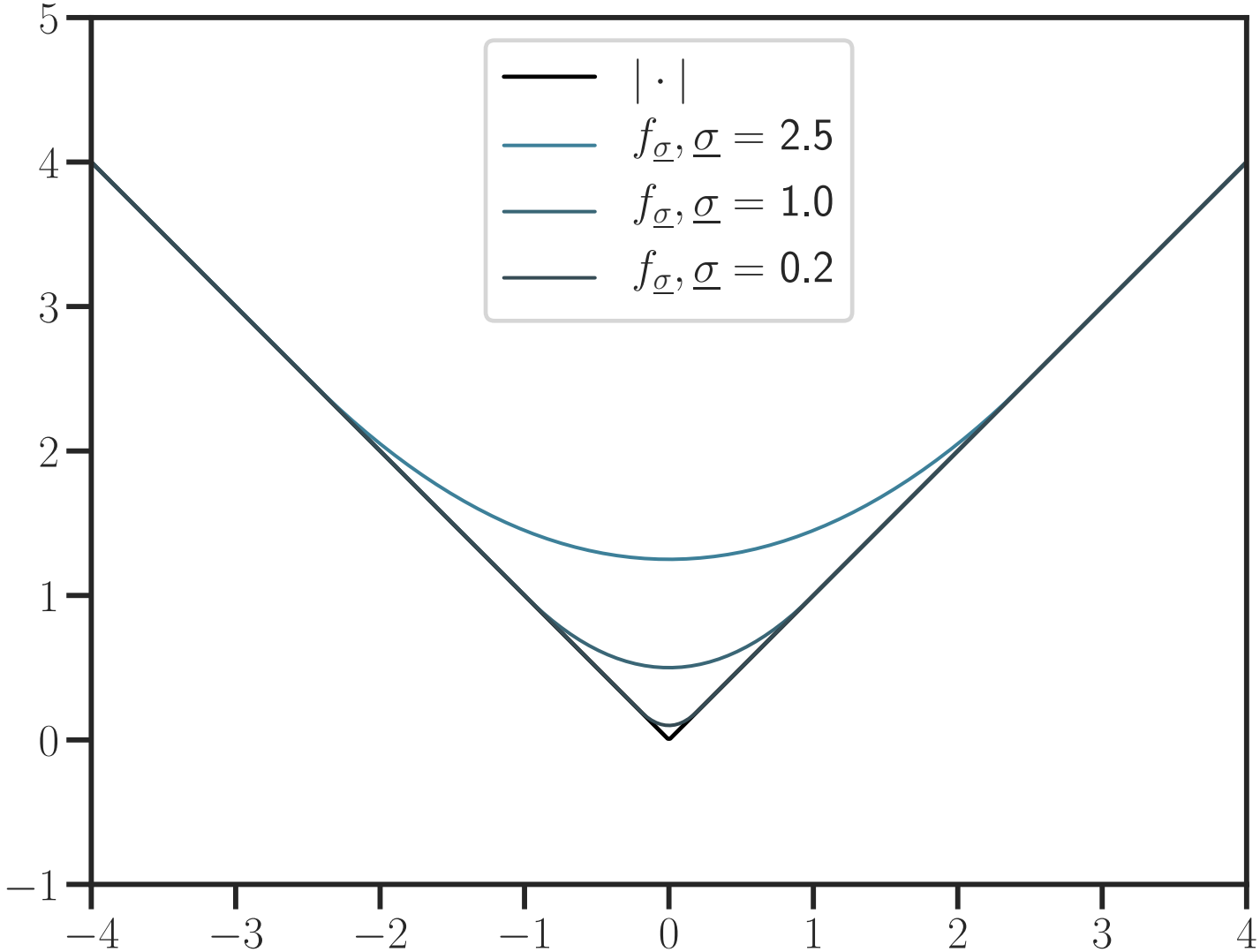

$|\cdot|$
$f_{\underline{\sigma}}, \underline{\sigma} = 2.5$
$f_{\underline{\sigma}}, \underline{\sigma} = 1.0$
5
4
3
2
1
0
−1
−4
−3
−2
−1
0
1
2
3
4


$|\cdot|$
$f_{\underline{\sigma}}, \underline{\sigma} = 2.5$
$f_{\underline{\sigma}}, \underline{\sigma} = 1.0$
$f_{\underline{\sigma}}, \underline{\sigma} = 0.2$
5
4
3
2
1
0
−1
−4
−3
−2
−1
0
1
2
3
4


|·|
5
4
3
2
1
0
−1
−4
−3
−2
−1
0
1
2
3
4


$|\cdot|$
$f_{\underline{\sigma}}, \underline{\sigma} = 2.5$
5
4
3
2
1
0
−1
−4
−3
−2
−1
0
1
2
3
4


$|\cdot|$
$f_{\underline{\sigma}}, \underline{\sigma} = 2.5$
$f_{\underline{\sigma}}, \underline{\sigma} = 1.0$
5
4
3
2
1
0
−1
−4
−3
−2
−1
0
1
2
3
4


$|\cdot|$
$f_{\underline{\sigma}}, \underline{\sigma} = 2.5$
$f_{\underline{\sigma}}, \underline{\sigma} = 1.0$
$f_{\underline{\sigma}}, \underline{\sigma} = 0.2$
5
4
3
2
1
0
−1
−4
−3
−2
−1
0
1
2
3
4

Solving the Smooth Concomitant Lasso

$$(\hat{\beta}^{(\lambda)}, \hat{\sigma}^{(\lambda)}) \in \underset{\beta \in \mathbb{R}^p, \sigma \geq \underline{\sigma}}{\arg\min} \frac{\|y - X\beta\|_2^2}{2n\sigma} + \frac{\sigma}{2} + \lambda \|\beta\|_1$$

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

Alternate iteratively:

- Fix σ: do one epoch of Lasso coordinate descent on β_js: For $j \in 1 \dots p$

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

Solving the Smooth Concomitant Lasso

$$(\hat{\beta}^{(\lambda)}, \hat{\sigma}^{(\lambda)}) \in \underset{\beta \in \mathbb{R}^p, \sigma \geq \underline{\sigma}}{\arg\min} \frac{\|y - X\beta\|_2^2}{2n\sigma} + \frac{\sigma}{2} + \lambda \|\beta\|_1$$

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

Alternate iteratively:

- Fix σ: do one epoch of Lasso coordinate descent on β_js: For $j \in 1 \dots p$

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

- Fix β: closed form solution to update σ

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

Solving the Smooth Concomitant Lasso

$$(\hat{\beta}^{(\lambda)}, \hat{\sigma}^{(\lambda)}) \in \underset{\beta \in \mathbb{R}^p, \sigma \geq \underline{\sigma}}{\arg\min} \frac{\|y - X\beta\|_2^2}{2n\sigma} + \frac{\sigma}{2} + \lambda \|\beta\|_1$$

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

Alternate iteratively:

- Fix σ: do one epoch of Lasso coordinate descent on β_js: For $j \in 1 \dots p$

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

- Fix β: closed form solution to update σ

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

The Smoothed Concomitant Lasso (i.i.d. case, Single-Task)

$$(\hat{\beta}^{(\lambda)}, \hat{\sigma}^{(\lambda)}) \in \underset{\beta \in \mathbb{R}^p, \sigma \geq \underline{\sigma}}{\arg\min} \frac{\|y - X\beta\|_2^2}{2n\sigma} + \frac{\sigma}{2} + \lambda \|\beta\|_1$$

Questions:

- What is the impact of the smoothing on statistical properties? (sup-norm convergence? support recovery?)
- How to choose $\underline{\sigma}$?
- Can this estimator (with unknown noise level σ^*) generalize for correlated noise (with unknown covariance matrix Σ^*)?

Statistical properties[7]

$$(\hat{\beta}^{(\lambda)}, \hat{\sigma}^{(\lambda)}) \in \underset{\beta \in \mathbb{R}^p, \sigma \geq \underline{\sigma}}{\arg\min} \frac{\|y - X\beta\|_2^2}{2n\sigma} + \frac{\sigma}{2} + \lambda \|\beta\|_1$$

Proposition

- i.i.d. Gaussian noise
- $+ X$ satisfying the "mutual incoherence" property
- $+ \lambda \sim \frac{\sqrt{\log p}}{\sqrt{n}}$
- $+ \underline{\sigma} \leq \sigma_* / \sqrt{2}$
- $\Longrightarrow$ with high probability:

$$\|\beta^* - \hat{\beta}\|_\infty \leq C \sigma_* \sqrt{\frac{\log p}{n}}$$

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

SCL support recovery perforance as a function of λ and $\underline{\sigma}$

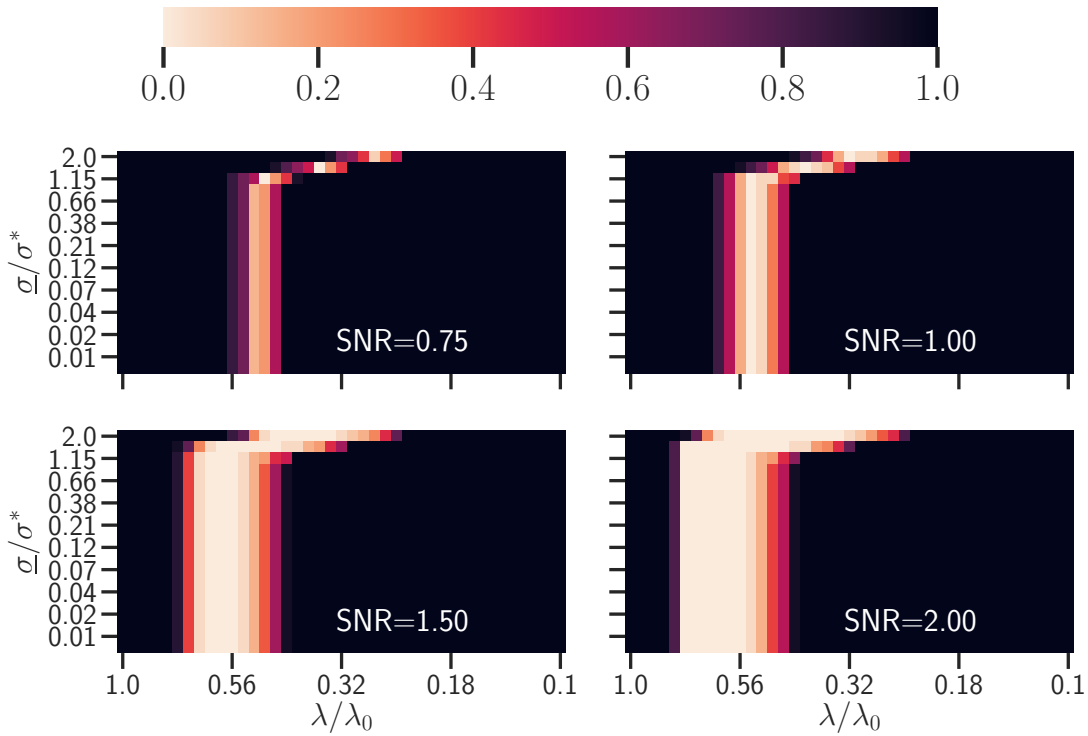


Figure: (Synthetic data, $n = 50$, $p = 1000$) Hard recovery loss for different values of SNR for the SCL.

The Smoothed Concomitant Lasso (i.i.d. case, Single-Task)

$$(\hat{\beta}^{(\lambda)}, \hat{\sigma}^{(\lambda)}) \in \underset{\beta \in \mathbb{R}^p, \sigma \geq \underline{\sigma}}{\arg\min} \frac{\|y - X\beta\|_2^2}{2n\sigma} + \frac{\sigma}{2} + \lambda \|\beta\|_1$$

Questions:

- What is the impact of the smoothing on statistical properties? (sup-norm convergence? support recovery?) ✓
- How to choose $\underline{\sigma}$? ✓
- Can this estimator (with unknown noise level σ^*) generalize for correlated noise (with unknown covariance matrix Σ^*)?

Table of Contents

Smoothing other norms

More interesting: S. van de Geer introduced the pivotal *multivariate* $\sqrt{\text{Lasso}}$[8]:

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

hard to solve ,[9] statistical analysis makes stringent assumptions

Smoothing the datafit makes optim. and stats easier!

---

[8] S. van de Geer and B. Stucky. "χ 2-confidence sets in high-dimensional regression". In: *Statistical analysis for high-dimensional data*. Springer, 2016, pp. 279–306.

[9] A. J. Molstad. "Insights and algorithms for the multivariate square-root lasso". In: *arXiv preprint arXiv:1909.05041* (2019).

Smoothing the nuclear norm[10]

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

$$\|\cdot\|_* \square \left(\frac{1}{2\underline{\sigma}} \|\cdot\|^2 + \frac{n}{2} \right)(Z) = \sum_i \mathsf{huber}_{\underline{\sigma}}(\gamma_i)$$

$$= \min_{S \succeq \underline{\sigma}} \left(\frac{1}{2} \|Z\|_{S^{-1}}^2 + \frac{1}{2} \operatorname{Tr}(S) \right)$$

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

---

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

Generalization ? Yes !
(correlated Gaussian noise, Multi-Task)

SGCL[11]:

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

Benefits

- jointly convex formulation

Drawbacks:

- Statistically: estimate a matrix of size n^2.
  $\mathcal{O}(n^2)$ parameters to estimate for S only nT observations

[11] M. Massias et al. "Generalized Concomitant Multi-Task Lasso for Sparse Multimodal Regression". In: *AISTATS*. vol. 84. 2018, pp. 998–1007.

Generalization ? Yes !
(correlated Gaussian noise, Multi-Task)

SGCL[11]:

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

Benefits

- jointly convex formulation

Drawbacks:

- Statistically: estimate a matrix of size n^2.
  $\mathcal{O}(n^2)$ parameters to estimate for S only nT observations

[11] M. Massias et al. "Generalized Concomitant Multi-Task Lasso for Sparse Multimodal Regression". In: *AISTATS*. vol. 84. 2018, pp. 998–1007.

Statistical properties[12]

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

Proposition

- i.i.d. Gaussian noise
- $+$ X satisfying the "mutual incoherence" property
- $+$ $\lambda \sim \frac{\sqrt{\log p}}{T\sqrt{n}}$
- $+$ $\underline{\sigma} \leq \sigma_*/\sqrt{2}$
- $\Longrightarrow$ with high probability: at least $1 - \cdots - 2ne^{-cT/n}$

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

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

SGCL support recovery perforance as a function of λ and $\underline{\sigma}$

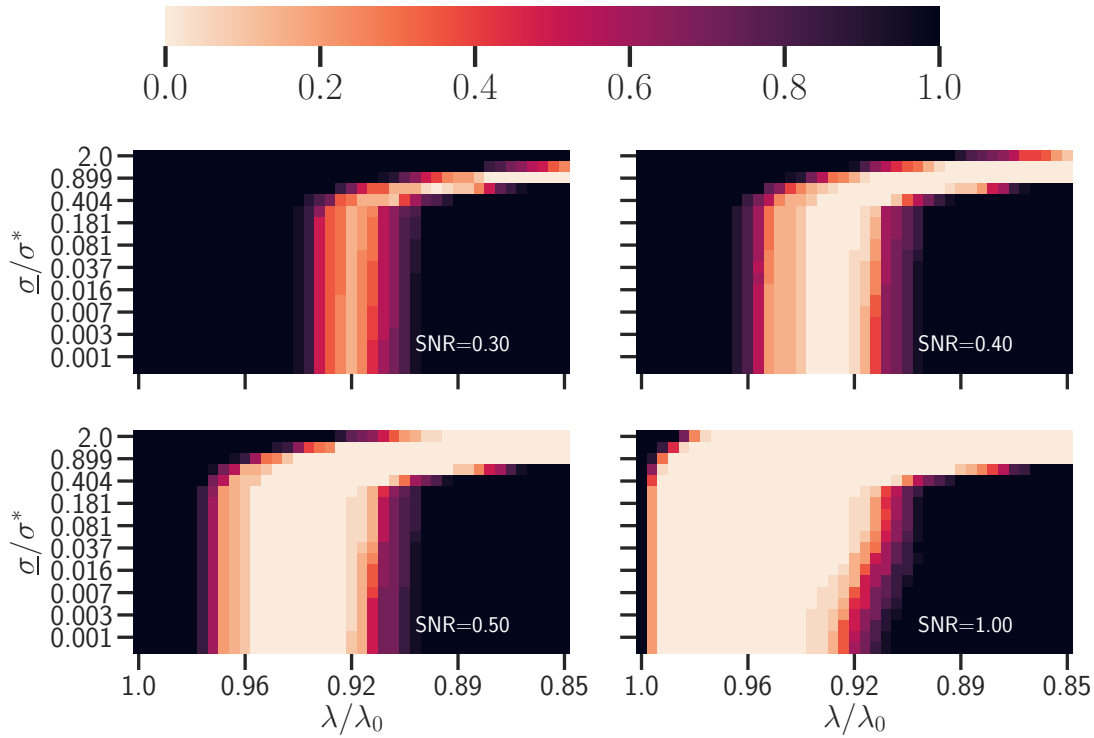


Figure: (Synthetic data, $n = 50$, $p = 1000$, $T = 20$) Hard recovery loss for different values of SNR for the SGCL.

Can take advantage of the repetitions? Yes!

Smoothed Generalized Concomitant Lasso (SGCL)[13]:

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

Concomitant Lasso with Repetitions (CLaR)[14]:

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

---

[13] M. Massias et al. "Generalized Concomitant Multi-Task Lasso for Sparse Multimodal Regression". In: *AISTATS*. vol. 84. 2018, pp. 998–1007.

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

SGCL and CLaR computations: B update

Jointly convex: alternate minimization converges

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

(Block) Coordinate Descent (Soft-Threshold.) : update B row-wise

Possible refinements:

- (Gap) safe screening rules[15]
- Strong rules[16]
- Active sets methods[17] etc.

---

[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] R. Tibshirani et al. "Strong rules for discarding predictors in lasso-type problems". In: *J. R. Stat. Soc. Ser. B Stat. Methodol.* 74.2 (2012), pp. 245–266.

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

SGCL and CLaR computations: B update

Jointly convex: alternate minimization converges

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

(Block) Coordinate Descent (Soft-Threshold.) : update B row-wise

Possible refinements:

- (Gap) safe screening rules[15]
- Strong rules[16]
- Active sets methods[17] etc.

---

[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] R. Tibshirani et al. "Strong rules for discarding predictors in lasso-type problems". In: *J. R. Stat. Soc. Ser. B Stat. Methodol.* 74.2 (2012), pp. 245–266.

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

SGCL and CLaR computations: S update

S Update (B fixed):

For SGCL and CLaR the problem can be reformulated as

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

SGCL and CLaR computations: S update

S Update (B fixed):

For SGCL and CLaR the problem can be reformulated as

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

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

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

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

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

SGCL and CLaR computations: S update

S Update (B fixed):

For SGCL and CLaR the problem can be reformulated as

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

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

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

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

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

The Smoothed Concomitant Lasso (i.i.d. case, Single-Task)

$$(\hat{\beta}^{(\lambda)}, \hat{\sigma}^{(\lambda)}) \in \underset{\beta \in \mathbb{R}^p, \sigma \geq \underline{\sigma}}{\arg\min} \frac{\|y - X\beta\|_2^2}{2n\sigma} + \frac{\sigma}{2} + \lambda \|\beta\|_1$$

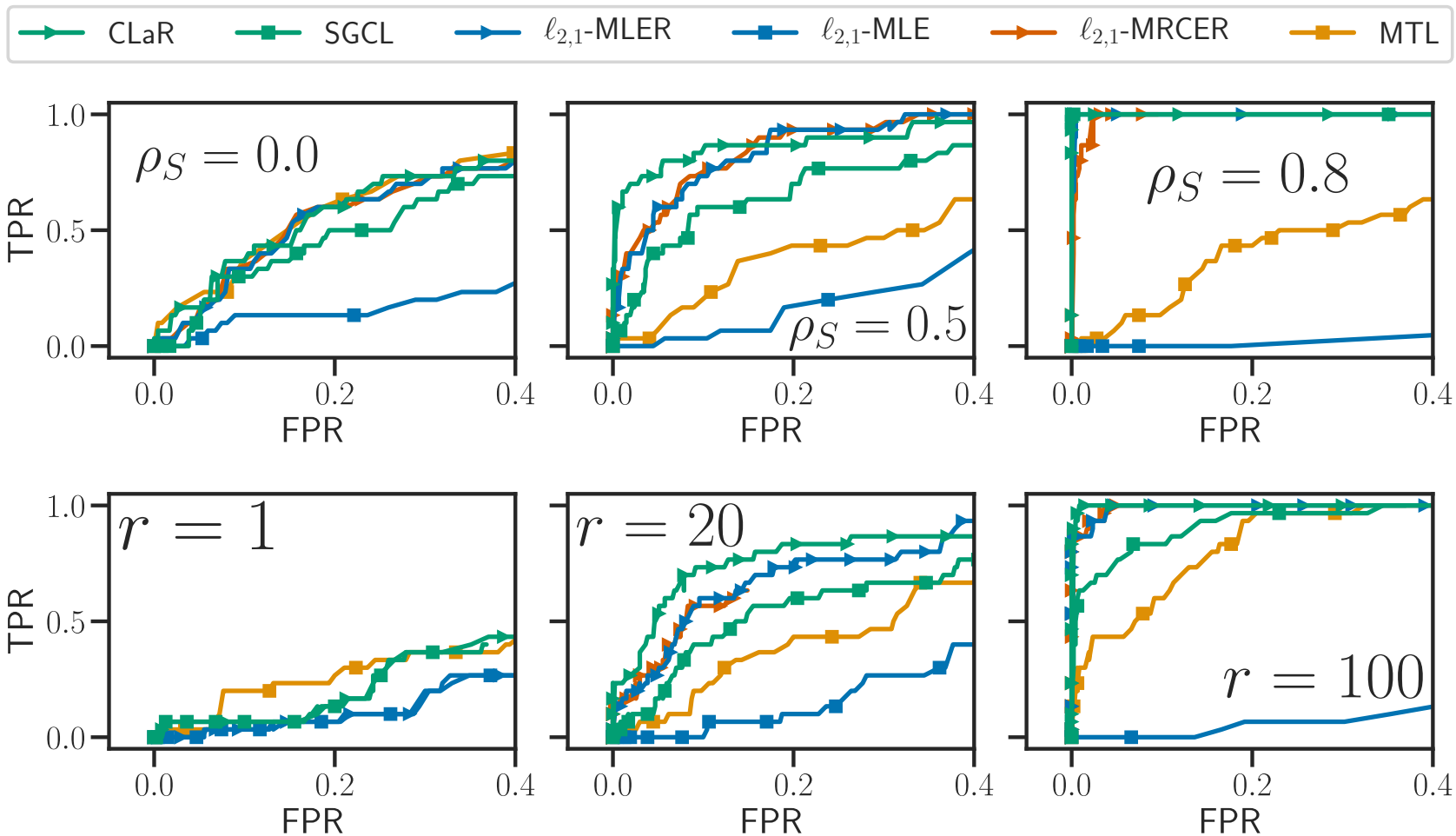
Questions:

- What is the impact of the smoothing on statistical properties? (sup-norm convergence? support recovery?) ✓
- How to choose $\underline{\sigma}$? ✓
- Can this estimator (with unknown noise level σ^*) generalize for correlated noise (with unknown covariance matrix Σ^*)? ✓

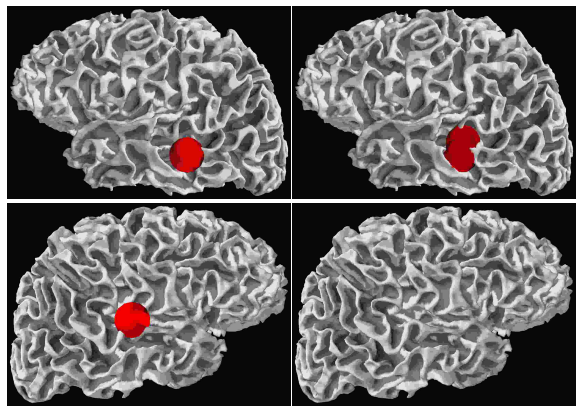
Table of Contents

Simulated scenarios

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



Real data

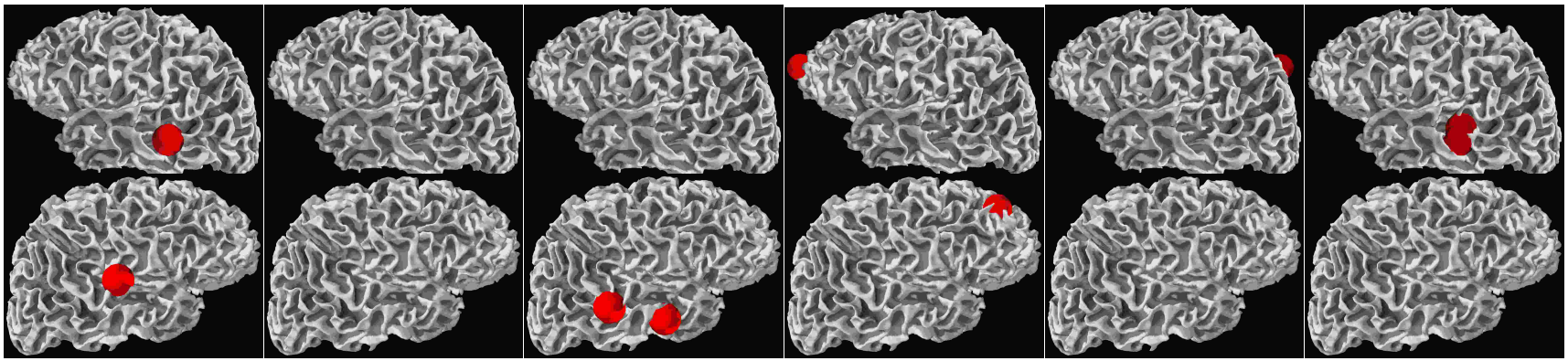


(a) ours (b) MTL

Figure: *Real data, left auditory stimulations* ($n = 102$, $p = 7498$, $q = 76$, $r = 63$) Sources found in the left hemisphere (top) and the right hemisphere (bottom) after left auditory stimulations.

- expected: 2 sources (one in each auditory cortex)
- λ chosen such that $\|\hat{\mathrm{B}}\|_{2,0} = 2$

Real data



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

Figure: *Real data, left auditory stimulations* ($n = 102$, $p = 7498$, $q = 76$, $r = 63$) Sources found in the left hemisphere (top) and the right hemisphere (bottom) after left auditory stimulations.

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

Conclusion and perspectives

- Statistical analysis of the smoothing for the *square-root Lasso* and the *multivariate square-root Lasso*
- Guidelines on how to set the smoothing parameter σ

Conclusion and perspectives

- Statistical analysis of the smoothing for the *square-root Lasso* and the *multivariate square-root Lasso*
- Guidelines on how to set the smoothing parameter $\underline{\sigma}$
- New estimator to handle correlated noise and repetitions in Multi-Task

Conclusion and perspectives

- Statistical analysis of the smoothing for the *square-root Lasso* and the *multivariate square-root Lasso*
- Guidelines on how to set the smoothing parameter $\underline{\sigma}$
- New estimator to handle correlated noise and repetitions in Multi-Task
- Improved support identification

Conclusion and perspectives

- Statistical analysis of the smoothing for the *square-root Lasso* and the *multivariate square-root Lasso*
- Guidelines on how to set the smoothing parameter $\underline{\sigma}$
- New estimator to handle correlated noise and repetitions in Multi-Task
- Improved support identification
- Future work: study the influence of $\underline{\sigma}$ from the optimization point of view, see if the analysis holds for the LAD-Lasso

Conclusion and perspectives

- Statistical analysis of the smoothing for the *square-root Lasso* and the *multivariate square-root Lasso*
- Guidelines on how to set the smoothing parameter $\underline{\sigma}$
- New estimator to handle correlated noise and repetitions in Multi-Task
- Improved support identification
- Future work: study the influence of $\underline{\sigma}$ from the optimization point of view, see if the analysis holds for the LAD-Lasso

Merci!

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

A. Gramfort

- Python code online https://github.com/QB3/CLaR
- Papers: arXiv[18], [19] [20]

[18] M. Massias et al. "Generalized Concomitant Multi-Task Lasso for Sparse Multimodal Regression". In: *AISTATS*. vol. 84. 2018, pp. 998–1007.

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

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

Competitors

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

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

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

$$(\hat{\mathrm{B}}, \hat{\Sigma}) \in \underset{\substack{\mathrm{B}\in\mathbb{R}^{p\times q} \\ \Sigma \succeq \underline{\sigma}^2}}{\arg\min} \sum_1^r \left\| Y^{(l)} - X\mathrm{B} \right\|_{\Sigma^{-1}}^2 - \log\det(\Sigma^{-1}) + \lambda \left\|\mathrm{B}\right\|_{2,1} \quad .$$

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

Smoothing of matrix norm

Huber-like formula for the Frobenius norm

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

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

What about other norms ?

Smoothing of matrix norm

Huber-like formula for the Frobenius norm

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

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

What about other norms ?

Huber-like formula for the nuclear/trace norm

$$\|\cdot\|_{s,1} \,\square\, \omega_{\underline{\sigma}}(Z) = \begin{cases} \frac{\|Z\|_F^2}{2\underline{\sigma}} + \frac{\underline{\sigma}}{2} n \wedge q, & \text{if } \|Z\|_\infty \leq \underline{\sigma} \\ \frac{1}{2\underline{\sigma}} \sum_i \gamma_i^2 - (\gamma_i \wedge \underline{\sigma} - \gamma_i)^2, & \text{if } \|Z\|_\infty > \underline{\sigma} \end{cases}$$

$$= \min_{S \succeq \underline{\sigma}} \tfrac{1}{2} \|Z\|_{S^{-1}}^2 + \tfrac{1}{2} \operatorname{Tr}(S)$$

γ_i: singular values of Z

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

Smoothing of matrix norm

Huber-like formula for the Frobenius norm

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

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

What about other norms ?

Huber-like formula for the nuclear/trace norm

$$\|\cdot\|_{s,1} \,\square\, \omega_{\underline{\sigma}}(Z) = \begin{cases} \frac{\|Z\|_F^2}{2\underline{\sigma}} + \frac{\underline{\sigma}}{2} n \wedge q, & \text{if } \|Z\|_\infty \leq \underline{\sigma} \\ \frac{1}{2\underline{\sigma}} \sum_i \gamma_i^2 - (\gamma_i \wedge \underline{\sigma} - \gamma_i)^2, & \text{if } \|Z\|_\infty > \underline{\sigma} \end{cases}$$

$$= \min_{S \succeq \underline{\sigma}} \tfrac{1}{2} \|Z\|_{S^{-1}}^2 + \tfrac{1}{2} \operatorname{Tr}(S)$$

γ_i: singular values of Z

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