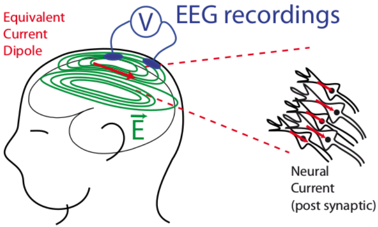
# Generalized Concomitant Multi-Task Lasso for sparse multimodal regression

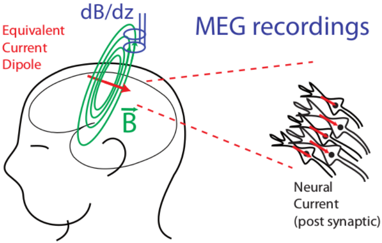
**Joseph Salmon**
http://josephsalmon.eu
IMAG, Univ Montpellier, CNRS
Montpellier, France

Joint work with:
**Mathurin Massias** (INRIA, Parietal Team)
**Olivier Fercoq** (Télécom ParisTech)
**Alexandre Gramfort** (INRIA, Parietal Team)

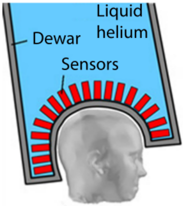
# M/EEG inverse problem for brain imaging

- sensors: magneto- and electro-encephalogram measurements during a cognitive experiment
- sources: brain locations



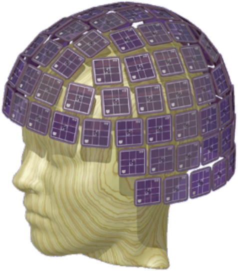

First EEG
recordings
in 1929
by H. Berger



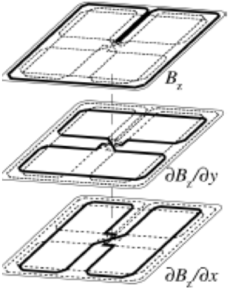

Hôpital La Timone
Marseille, France

# MEG elements: magnometers and gradiometers

Device

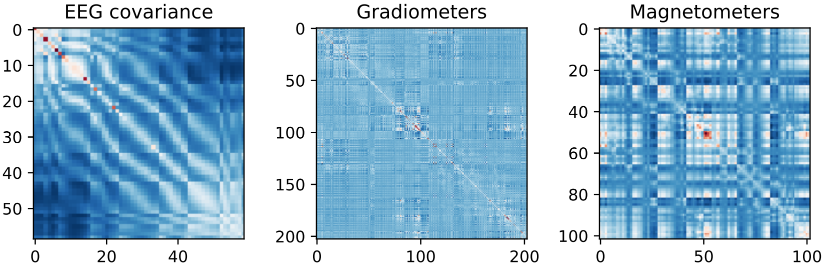


Sensors



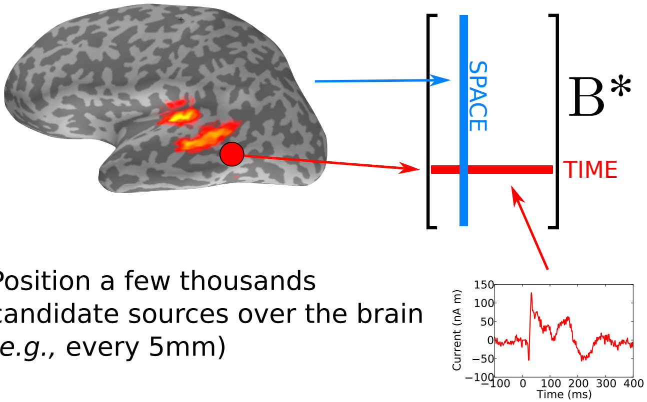

Detail of a sensor

# Noise is different for EEG / MEG (magnometers and gradiometers)




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

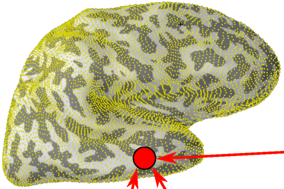
# Source modeling




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

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

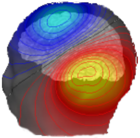
# Design matrix - Forward operator

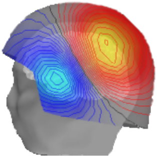


$$X = \left[\begin{array}{c} X_{\mathrm{EEG}} \\ \hdashline X_{\mathrm{MEG}} \end{array}\right] \in \mathbb{R}^{n \times p}$$

EEG:
Forward field of the electrodes

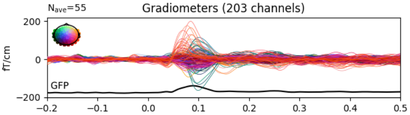
MEG:
Forward field of sensor

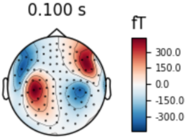


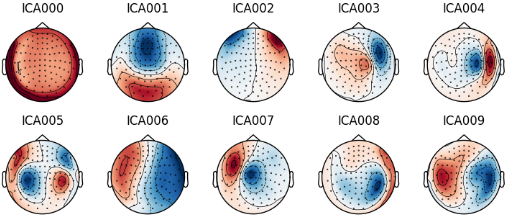


$X$ : gain matrix / forward operator obtained by Maxwell's equations

# Sparsity assumption: cortical sources produce dipolar patterns well modelled by focal sources



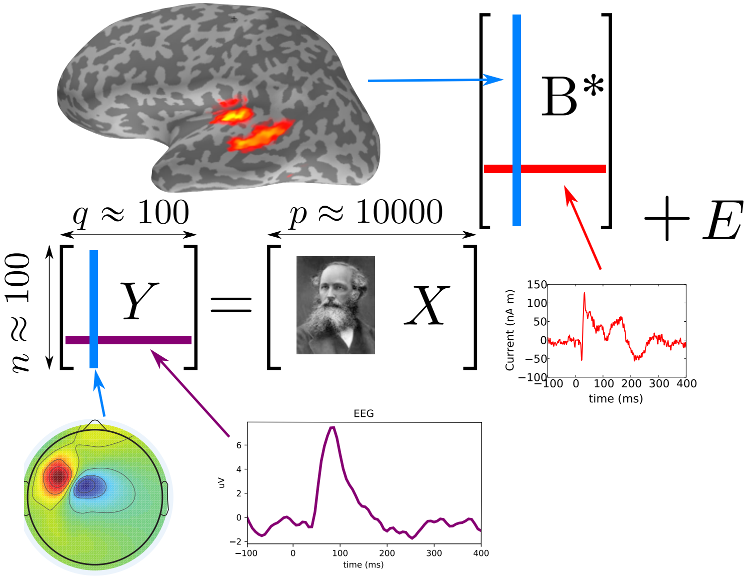




ICA : Blind source separation recovers dipolar patterns(1)

http://martinos.org/mne/stable/auto_tutorials/plot_visualize_evoked.html
http://martinos.org/mne/stable/auto_tutorials/plot_artifacts_correction_ica.html

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

# The M/EEG inverse problem: modeling



# A multi-task framework

Multi-task regression:

- $n$ observations (*e.g.*, number of sensors)
- $q$ tasks (*e.g.*, temporal information)
- $p$ features
- $Y \in \mathbb{R}^{n \times q}$ observation matrix
- $X \in \mathbb{R}^{n \times p}$ forward matrix

$$\boxed{Y = X\mathrm{B}^* + E}$$

where

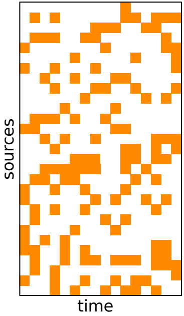
- $\mathrm{B}^* \in \mathbb{R}^{p \times q}$ : true source activity matrix
- $E \in \mathbb{R}^{n \times q}$ : additive white Gaussian noise (for simplicity)

Notation remark: capital letters refer to matrices

# Multi-tasks penalties[2]

Popular convex penalties considered:

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




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

Sparse support: no structure

Penalty: **Lasso type**

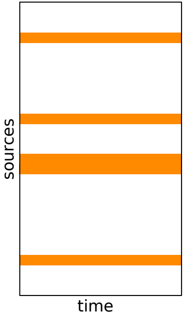
$$\Omega(\mathrm{B}) = \|\mathrm{B}\|_1 = \sum_{j=1}^{p} \sum_{k=1}^{q} |\mathrm{B}_{j,k}|$$

[2] 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-tasks penalties[2]

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

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




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

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 $\mathrm{B}$

[2] 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

# Step back on the Lasso case ($q = 1$)

Sparse Gaussian model: $y = X\beta^* + \sigma_* \varepsilon$

- $y \in \mathbb{R}^n$: observation
- $X \in \mathbb{R}^{n \times p}$: design matrix
- $\beta^* \in \mathbb{R}^p$: signal to recover; unknown
- $\|\beta^*\|_0 = s^*$: sparsity level (small w.r.t. $p$); $s^*$ unknown
- $\varepsilon \sim \mathcal{N}(0, \sigma_*^2 \operatorname{Id}_n)$; $\sigma_*$ unknown

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

# Lasso theory[3],[4]

**Theorem**

For Gaussian noise model and $X$ satisfying the "Restricted Eigenvalue" property, for $\lambda = 2\sigma_* \sqrt{\frac{2\log(p/\delta)}{n}}$, then

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

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

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

Rem: $\kappa_{s^*}^2$ controls the conditioning of $X$ when extracting the $s^*$columns of $X$ associated to the true support

**BUT** $\sigma_*$ is unknown in practice !

[3] 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.

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

# Lasso theory[3],[4]

**Theorem**

For Gaussian noise model and $X$ satisfying the "Restricted Eigenvalue" property, for $\lambda = 2\sigma_* \sqrt{\frac{2\log(p/\delta)}{n}}$, then

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

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

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

Rem: $\kappa_{s^*}^2$ controls the conditioning of $X$ when extracting the $s^*$columns of $X$ associated to the true support

**BUT** $\sigma_*$ is unknown in practice !

(3) 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.

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

# Joint estimation of $\beta$ and $\sigma$

How to calibrate (theoretically) $\lambda$ when $\sigma_*$ is unknown?

Intuitive idea: initialize $\lambda$

- run Lasso with $\lambda$; get $\beta$
- estimate $\sigma$, *e.g.*, with residual $\sigma \leftarrow \frac{\|y - X\beta\|}{\sqrt{n}}$
- re-scale $\lambda \propto \sigma$, and run Lasso with it
- iterate (until convergence)

*N.B.*: exactly the Concomitant Lasso[5] / Scaled-Lasso[6] implementation

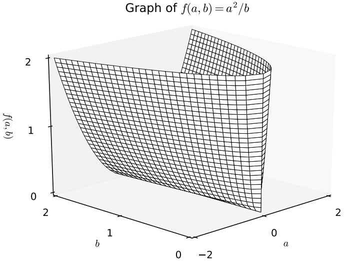
[5] A. B. Owen. "A robust hybrid of lasso and ridge regression". In: *Contemporary Mathematics* 443 (2007), pp. 59–72.

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

# Concomitant Lasso

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

- $\frac{\sigma}{2}$ : penalty on noise level, roots in robust estimation[(7), (8)]
- jointly convex program: $(a, b) \mapsto a^2/b$ is convex




[(7)] P. J. Huber 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.

[(8)] P. J. Huber. *Robust Statistics*. John Wiley & Sons Inc., 1981.

# Concomitant performance

**Theorem**[9],[10]

For Gaussian noise model and $X$ satisfying the "Restricted Eigenvalue" property and $\lambda = 2\sqrt{\frac{2\log(p/\delta)}{n}}$, then

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

with high probability, where $\hat{\beta}^{(\lambda)}$ is a Concomitant Lasso solution

Rem: provide same rate as Lasso, **without knowing** $\sigma_*$

Rem: $\lambda$ has no dimension, but calibration still needed in practice...

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

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

# Link with $\sqrt{\text{Lasso}}$[11]

- Independently, $\sqrt{\text{Lasso}}$ analyzed to get "$\sigma$ free" choice of $\lambda$

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

- Connections with Concomitant Lasso:
  $\left(\hat{\beta}^{(\lambda)}_{\sqrt{\text{Lasso}}}, \hat{\sigma}^{(\lambda)}_{\sqrt{\text{Lasso}}}\right)$ is solution of the Concomitant Lasso when

$$\hat{\sigma}^{(\lambda)}_{\sqrt{\text{Lasso}}} = \frac{\left\| y - X\hat{\beta}^{(\lambda)}_{\sqrt{\text{Lasso}}} \right\|}{\sqrt{n}} \neq 0$$

Rem: non-smooth data fitting term with non-smooth regularization

[11] 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.

# The Smoothed Concomitant Lasso[13]

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

- useful for optimization with small $\lambda$
- with prior information on the minimal noise level, one can set $\underline{\sigma}$ as this bound (recovers Concomitant Lasso)
- setting $\underline{\sigma} = \epsilon$, smoothing theory asserts that $\frac{\epsilon}{2}$-solutions for the smoothed problem provide $\epsilon$-solutions for the $\sqrt{\mathrm{Lasso}}$[12]

---

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

[13] E. Ndiaye et al. "Efficient Smoothed Concomitant Lasso Estimation for High Dimensional Regression". In: *NCMIP*. 2017.

# Smoothing aparté[14],[15]

Motivation: smooth a non-smooth function $f$ to ease optimization

Smoothing: for $\mu > 0$, a "smoothed" version of $f$ is $f_\mu$

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

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

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

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

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

# Smoothing aparté[14],[15]

Motivation: smooth a non-smooth function $f$ to ease optimization

Smoothing: for $\mu > 0$, a "smoothed" version of $f$ is $f_\mu$

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

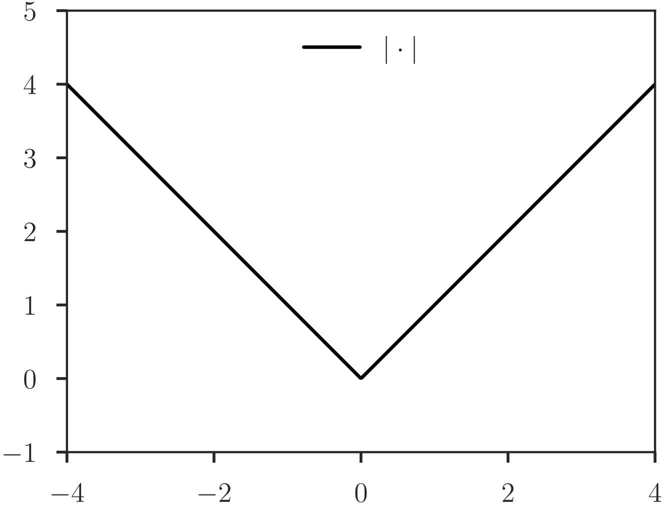
- $\omega$ is a predefined smooth function (s.t. $\nabla\omega$ is Lipschitz)

| Kernel smoothing analogy: | Fourier: $\mathcal{F}(f)$ | Fenchel/Legendre: $f^*$ |
|---|---|---|
| | **convolution**: $\star$ | **inf-convolution**: $\Box$ |
| | $\mathcal{F}(f \star g) = \mathcal{F}(f) \cdot \mathcal{F}(g)$ | $(f\Box g)^* = f^* + g^*$ |
| | Gaussian : $\mathcal{F}(g) = g$ | $\omega = \frac{\lVert\cdot\rVert^2}{2}$ : $\quad \omega^* = \omega$ |
| | $f_h = \frac{1}{h} g\left(\frac{\cdot}{h}\right) \star f$ | $f_\mu = \mu\omega\left(\frac{\cdot}{\mu}\right)\Box f$ |

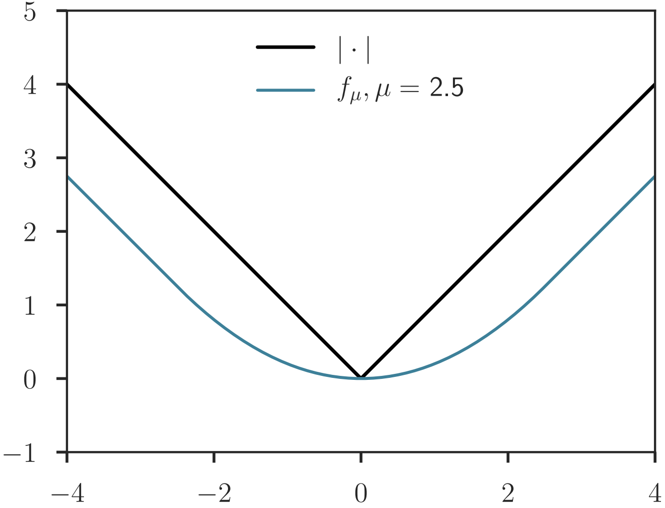
[14] Y. Nesterov. "Smooth minimization of non-smooth functions". In: *Math. Program.* 103.1 (2005), pp. 127–152.

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

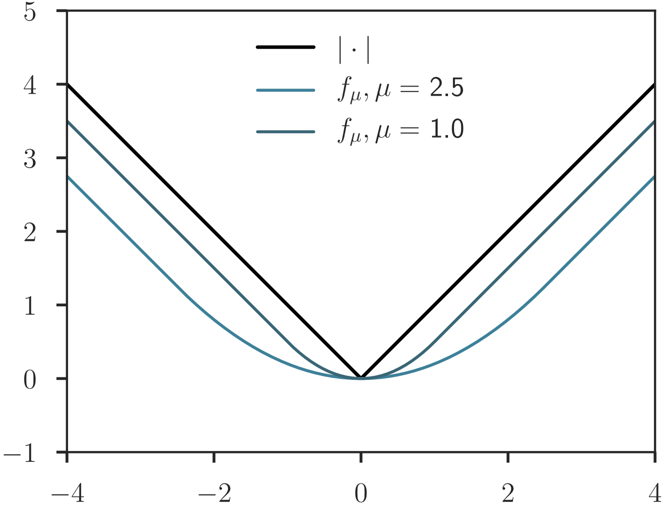
# Huber function: $\omega(t) = \frac{t^2}{2}$



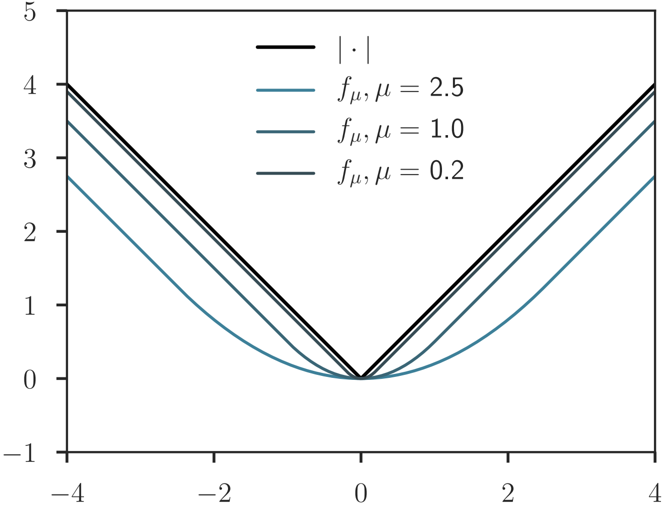
# Huber function: $\omega(t) = \frac{t^2}{2}$



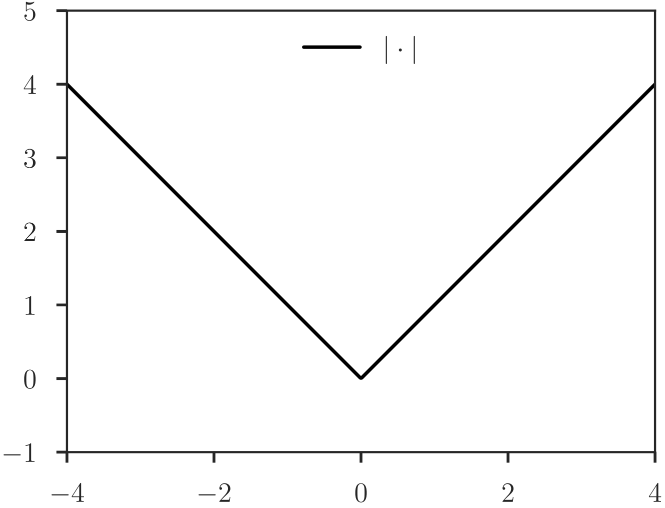
# Huber function: $\omega(t) = \frac{t^2}{2}$



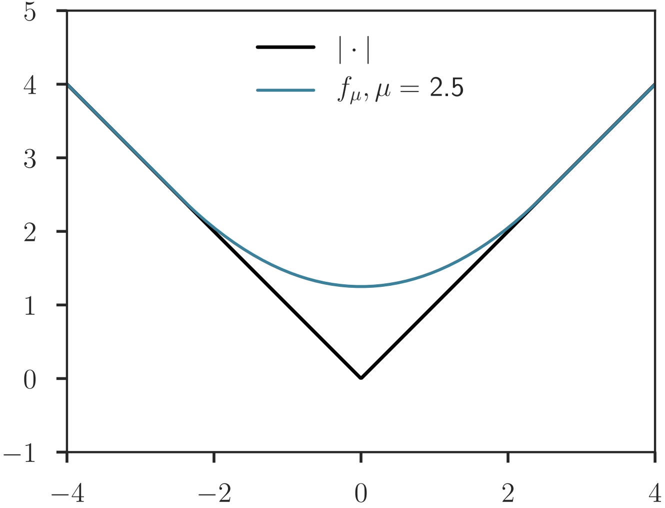
# Huber function: $\omega(t) = \frac{t^2}{2}$



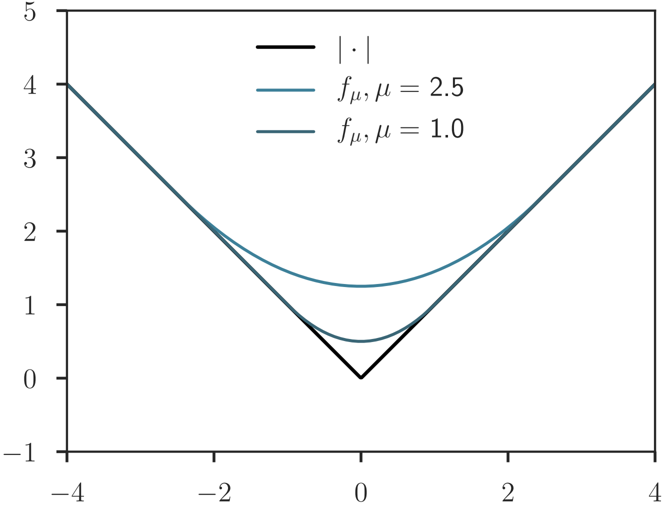
## Huber function (bis): $\omega(t) = \frac{t^2}{2} + \frac{1}{2}$



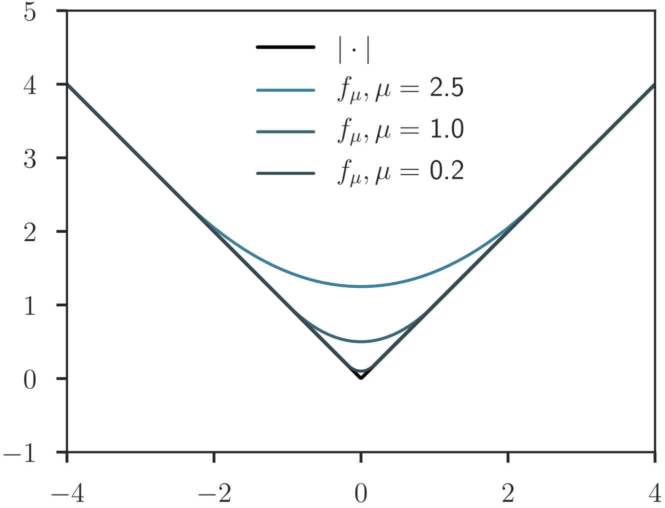
# Huber function (bis): $\omega(t) = \frac{t^2}{2} + \frac{1}{2}$



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



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



# Huberization of the $\sqrt{\text{Lasso}}$

"**Huberization**": $f(z) = \frac{\|z\|}{\sqrt{n}}$, $\mu = \underline{\sigma}$, $\omega(z) = \frac{\|z\|^2}{2} + \frac{1}{2}$

$$
\begin{aligned}
f_{\underline{\sigma}}(z) = \underline{\sigma}\omega\left(\frac{\cdot}{\underline{\sigma}}\right) \square f(z) &= \begin{cases} \frac{\|z\|^2}{2n\,\underline{\sigma}} + \frac{\underline{\sigma}}{2}, & \text{if } \frac{\|z\|}{\sqrt{n}} \leq \underline{\sigma} \\ \frac{\|z\|}{\sqrt{n}}, & \text{if } \frac{\|z\|}{\sqrt{n}} > \underline{\sigma} \end{cases} \\
&= \min_{\sigma \geq \underline{\sigma}} \left( \frac{\|z\|^2}{2n\sigma} + \frac{\sigma}{2} \right)
\end{aligned}
$$

Leads to the Smoothed Concomitant Lasso formulation:

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

# Solving the Smooth Concomitant Lasso

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

**Jointly convex** formulation : can be optimized by alternate minimization w.r.t. $\beta$ and $\sigma$ (gradient Lipschitz)

Alternate iteratively:

- Fix $\sigma$: (approximatively) solve a Lasso problem to update $\beta$

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

# Solving the Smooth Concomitant Lasso

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

**Jointly convex** formulation : can be optimized by alternate minimization w.r.t. $\beta$ and $\sigma$ (gradient Lipschitz)

Alternate iteratively:

- Fix $\sigma$: (approximatively) solve a Lasso problem to update $\beta$
$$\hat{\beta} \in \underset{\beta \in \mathbb{R}^p}{\arg\min} \frac{\|y - X\beta\|^2}{2n} + \lambda\sigma \|\beta\|_1 \quad \text{(Lasso step)}$$

- Fix $\beta$: closed form solution to update $\sigma$
$$\hat{\sigma} = \max\left(\frac{\|y - X\beta\|}{\sqrt{n}}, \underline{\sigma}\right) \quad \text{(Noise estimation step)}$$

# Solving the Smooth Concomitant Lasso

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

**Jointly convex** formulation : can be optimized by alternate minimization w.r.t. $\beta$ and $\sigma$ (gradient Lipschitz)

<u>Alternate iteratively</u>:

- Fix $\sigma$: (approximatively) solve a Lasso problem to update $\beta$

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

- Fix $\beta$: closed form solution to update $\sigma$

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

# Table of Contents

# Back to multi-task : $Y = X\mathrm{B}^* + E$

General case: $Y \in \mathbb{R}^{n\times q}$, $\mathrm{B} \in \mathbb{R}^{p\times q}$, and the noise $E \in \mathbb{R}^{n\times q}$ might have some structure evolving along the $n$ samples (sensors)

[16] S. van de Geer. *Estimation and testing under sparsity*. Vol. 2159. Lecture Notes in Mathematics. Lecture notes from the 45th Probability Summer School held in Saint-Four, 2015, École d'Été de Probabilités de Saint-Flour. Springer, 2016, pp. xiii+274.

# Back to multi-task : $Y = X\mathrm{B}^* + E$

General case: $Y \in \mathbb{R}^{n \times q}$, $\mathrm{B} \in \mathbb{R}^{p \times q}$, and the noise $E \in \mathbb{R}^{n \times q}$ might have some structure evolving along the $n$ samples (sensors)

**Smoothed Generalized Concomitant Lasso** (SGCL):

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

with $\|R\|^2_{\Sigma^{-1}} := \mathrm{Tr}(R^\top \Sigma^{-1} R)$ (Mahalanobis distance)

[16] S. van de Geer. *Estimation and testing under sparsity*. Vol. 2159. Lecture Notes in Mathematics. Lecture notes from the 45th Probability Summer School held in Saint-Four, 2015, École d'Été de Probabilités de Saint-Flour. Springer, 2016, pp. xiii+274.

# Back to multi-task : $Y = X\mathrm{B}^* + E$

General case: $Y \in \mathbb{R}^{n \times q}$, $\mathrm{B} \in \mathbb{R}^{p \times q}$, and the noise $E \in \mathbb{R}^{n \times q}$ might have some structure evolving along the $n$ samples (sensors)

**Smoothed Generalized Concomitant Lasso** (SGCL):

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

with $\|R\|^2_{\Sigma^{-1}} := \mathrm{Tr}(R^\top \Sigma^{-1} R)$ (Mahalanobis distance)

- jointly convex formulation (=nuclear norm smoothing[16])
- noise penalty on the sum of the eigenvalues of $\Sigma$

Beware: $\Sigma$ not a covariance (co-standard deviation?)

[16] S. van de Geer. *Estimation and testing under sparsity*. Vol. 2159. Lecture Notes in Mathematics. Lecture notes from the 45th Probability Summer School held in Saint-Four, 2015, École d'Été de Probabilités de Saint-Flour. Springer, 2016, pp. xiii+274.

# SGCL and computation: $\mathrm{B}$ update

Jointly convex formulation: alternate minimization still converging

$\mathrm{B}$ **Update - $\Sigma$ fixed**:

"smooth + non-smooth" optimization; use Block Coordinate Descent (Iterative Block Soft-Thresholding) to update $\mathrm{B}$ row-wise

Possible refinements:

- (Gap) safe screening rules[17],[18]
- Stong rules[19]
- Active sets methods[20] etc.

---

[17] 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.

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

[19] 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.

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

# SGCL and computation: $\mathrm{B}$ update

Jointly convex formulation: alternate minimization still converging

$\mathrm{B}$ **Update - $\Sigma$ fixed**:

"smooth + non-smooth" optimization; use Block Coordinate Descent (Iterative Block Soft-Thresholding) to update $\mathrm{B}$ row-wise

Possible refinements:

- (Gap) safe screening rules[17],[18]
- Stong rules[19]
- Active sets methods[20] etc.

---

(17) 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.

(18) E. Ndiaye et al. "Gap Safe screening rules for sparsity enforcing penalties". In: *J. Mach. Learn. Res.* 18.128 (2017), pp. 1–33.

(19) 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.

(20) T. B. Johnson and C. Guestrin. "BLITZ: A Principled Meta-Algorithm for Scaling Sparse Optimization". In: *ICML*. 2015, pp. 1171–1179.

# SGCL and computation: $\Sigma$ update

$\Sigma$ **Update** - $\mathrm{B}$ **fixed**:

with $R = Y - X\mathrm{B}$ (**residuals**), the problem can be reformulated

$$\hat{\Sigma} = \operatorname*{arg\,min}_{\Sigma \in \mathbb{S}^n_{++}, \Sigma \succeq \underline{\sigma} \,\mathrm{Id}_n} \left( \frac{1}{2nq} \operatorname{Tr}[R^\top \Sigma^{-1} R] + \frac{1}{2n} \operatorname{Tr}(\Sigma) \right)$$

# SGCL and computation: $\Sigma$ update

$\Sigma$ **Update** - $\mathrm{B}$ **fixed**:

with $R = Y - X\mathrm{B}$ (**residuals**), the problem can be reformulated

$$\hat{\Sigma} = \operatorname*{arg\,min}_{\Sigma \in \mathbb{S}^n_{++}, \Sigma \succeq \underline{\sigma} \operatorname{Id}_n} \left( \frac{1}{2nq} \operatorname{Tr}[R^\top \Sigma^{-1} R] + \frac{1}{2n} \operatorname{Tr}(\Sigma) \right)$$

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

if $U^\top \operatorname{diag}(s_1, \dots, s_n) U$ is the spectral decomposition of $\frac{1}{q} R R^\top$:

$$\hat{\Sigma} = U^\top \operatorname{diag}(\max(\underline{\sigma}, \sqrt{s_1}), \dots, \max(\underline{\sigma}, \sqrt{s_n})) U$$

# Main drawbacks

- Statistically: $\mathcal{O}(n^2)$ parameters to infer for $\Sigma$, with only $nq$ observations (need $q$ large w.r.t. $n$)

- Computationally: $\Sigma$ update cost is $\mathcal{O}(n^3)$ too slow in general (SVD computation)
  Note: fine for MEG/EEG problems ($n \approx 300$)

# Main drawbacks

- Statistically: $\mathcal{O}(n^2)$ parameters to infer for $\Sigma$, with only $nq$ observations (need $q$ large w.r.t. $n$)

- Computationally: $\Sigma$ update cost is $\mathcal{O}(n^3)$ too slow in general (SVD computation)
  Note: fine for MEG/EEG problems ($n \approx 300$)

More structure needed on $\Sigma$!

# Table of Contents

# Block Homoscedastic model

In the MEG/EEG case : 3 different types of signals are recorded

- electrodes : measure the electric potentials
- magnetometers : measure the magnetic field
- gradiometers : measure the gradient of the magnetic field

$\neq$ physical natures $\Longrightarrow$ different noise levels

Key point: observations divided into 3 blocks (known partition)

# Block homoscedastic model

$K$ groups of observations ($K$ sensors modalities)

$$X = \begin{pmatrix} X^1 \\ \vdots \\ X^K \end{pmatrix}, \, Y = \begin{pmatrix} Y^1 \\ \vdots \\ Y^K \end{pmatrix}, \, E = \begin{pmatrix} E^1 \\ \vdots \\ E^K \end{pmatrix}$$

$$\Sigma^* = \operatorname{diag}(\sigma_1^* \operatorname{Id}_{n_1}, \dots, \sigma_K^* \operatorname{Id}_{n_K}) \text{ where } n = n_1 + \cdots + n_K$$

For each block, the entries $E_{i,j}^k \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$ (homoscedastic):

$$\boxed{Y^k = X^k \mathrm{B}^* + \sigma_k^* E^k}$$

**MEG/EEG case**: $K = 3$, corresponding to 3 physical signals

1. EEG
2. MEG magnetometers
3. MEG gradiometers

# Smoothed Block Homoscedastic Concomitant (SBHCL)

Additional constraints: $\Sigma$ piecewise constant **diagonal**, *i.e.*,

$$\Sigma = \operatorname{diag}(\sigma_1 \operatorname{Id}_{n_1}, \ldots, \sigma_K \operatorname{Id}_{n_K})$$

**Block Homoscedastic Concomitant**:

$$\mathop{\arg\min}_{\substack{\mathrm{B}\in\mathbb{R}^{p\times q},\\ \sigma_1,\ldots,\sigma_K\in\mathbb{R}^K_{++}\\ \sigma_k\geq\underline{\sigma}_k,\forall k\in[K]}} \sum_{k=1}^{K}\left(\frac{\|Y^k - X^k\mathrm{B}\|^2}{2nq\sigma_k} + \frac{n_k\sigma_k}{2n}\right) + \lambda\,\|\mathrm{B}\|_{2,1}$$

Benefit:

- number of parameters reduced $\frac{n(n+1)}{2} \to K$
- $\Sigma$ update: update/maintain the $\sigma_k$'s (storing residuals)

# Table of Contents

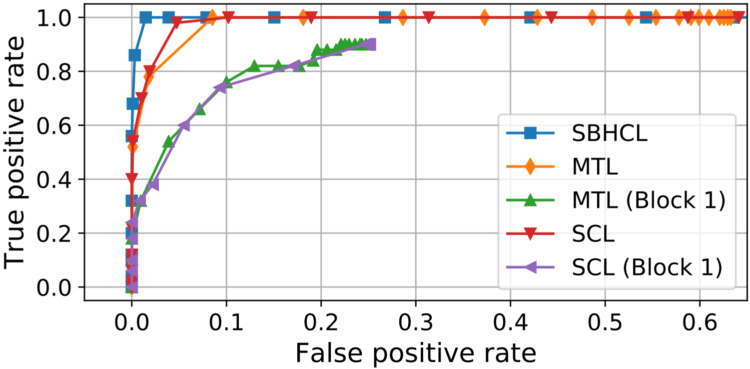
# Simulated scenario

Simulated block homoscedastic design:

- $n = 300$, with equals block sizes $n_1 = n_2 = n_3 = 100$
- $p = 1000$
- $q = 100$
- $X$ Toeplitz-correlated: $\mathrm{Cov}(X_i, X_j) = \rho^{|i-j|}$, $\rho \in ]0,1[$
- 3 blocks with standard deviation in ratio 1, 2, 5

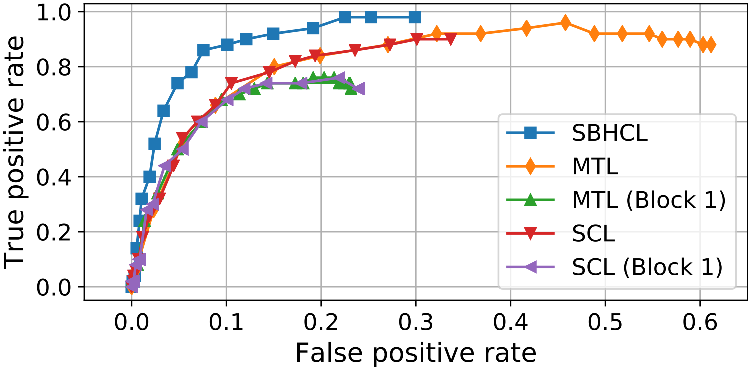
Rem: Block 1 has smallest standard deviation

# Support recovery: ROC curve w.r.t. $\lambda$, $\rho = 0.1$



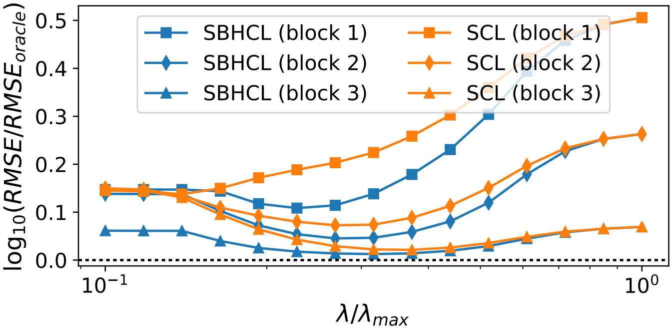

| | |
|---|---|
| SBHCL: | Smoothed Block Homoscedastic Concomitant |
| MTL: | Multi-Task Lasso |
| SCL: | Smooth Concomitant Lasso (single $\sigma$ for all blocks) |
| MTL (Block 1): | MTL on least noisy block |
| SCL (Block 1): | SCL on least noisy block |

# Support recovery: ROC curve w.r.t. $\lambda$, $\rho = 0.9$




| | |
|---|---|
| SBHCL: | Smoothed Block Homoscedastic Concomitant |
| MTL: | Multi-Task Lasso |
| SCL: | Smooth Concomitant Lasso (single $\sigma$ for all blocks) |
| MTL (Block 1): | MTL on least noisy block |
| SCL (Block 1): | SCL on least noisy block |

# Prediction error: RMSE curve w.r.t. $\lambda$, $\rho = 0.7$




RMSE (Root Mean Square Error) normalized by oracle RMSE, per block, for the multi-task SBHCL and SCL on testing set

Conclusion: align best $\lambda$'s for all modalities

# Conclusion and perspectives

- New insights for handling (structured) noise in multi-task
- Handling multiple noise levels: improve both prediction and support identification

# Conclusion and perspectives

- New insights for handling (structured) noise in multi-task
- Handling multiple noise levels: improve both prediction and support identification
- Numerical cost equivalent to classical Multi-Task Lasso for "simple" noise structure (*e.g.*, block homoscedastic)

# Conclusion and perspectives

- New insights for handling (structured) noise in multi-task
- Handling multiple noise levels: improve both prediction and support identification
- Numerical cost equivalent to classical Multi-Task Lasso for "simple" noise structure (*e.g.*, block homoscedastic)
- Future work: handling repetitions of neuro tasks, non-convex penalties, etc.

# Conclusion and perspectives

- New insights for handling (structured) noise in multi-task
- Handling multiple noise levels: improve both prediction and support identification
- Numerical cost equivalent to classical Multi-Task Lasso for "simple" noise structure (*e.g.*, block homoscedastic)
- Future work: handling repetitions of neuro tasks, non-convex penalties, etc.

# Merci!

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

A. Gramfort

- Paper: arXiv / personnal webpage[21]
- Python code online: https://github.com/mathurinm/SHCL

Powered with **MooseTeX**

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

# References I


- Beck, A. and M. Teboulle. "Smoothing and first order methods: A unified framework". In: *SIAM J. Optim.* 22.2 (2012), pp. 557–580.
- Belloni, A., V. Chernozhukov, and L. Wang. "Square-root Lasso: pivotal recovery of sparse signals via conic programming". In: *Biometrika* 98.4 (2011), pp. 791–806.
- 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.
- Dalalyan, A. S., M. Hebiri, and J. Lederer. "On the Prediction Performance of the Lasso". In: *Bernoulli* 23.1 (2017), pp. 552–581.
- Delorme, A. et al. "Independent EEG sources are dipolar". In: *PloS one* 7.2 (2012), e30135.

# References II


- 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.
- Giraud, C. *Introduction to high-dimensional statistics*. Vol. 138. CRC Press, 2014.
- 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.
- Massias, M. et al. "Generalized Concomitant Multi-Task Lasso for Sparse Multimodal Regression". In: *AISTATS*. Vol. 84. 2018, pp. 998–1007.

# References III


- Ndiaye, E. et al. “Efficient Smoothed Concomitant Lasso Estimation for High Dimensional Regression”. In: *NCMIP*. 2017.
- Ndiaye, E. et al. “Gap Safe screening rules for sparsity enforcing penalties”. In: *J. Mach. Learn. Res.* 18.128 (2017), pp. 1–33.
- 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.
- Owen, A. B. “A robust hybrid of lasso and ridge regression”. In: *Contemporary Mathematics* 443 (2007), pp. 59–72.
- Sun, T. and C.-H. Zhang. “Scaled sparse linear regression”. In: *Biometrika* 99.4 (2012), pp. 879–898.

# References IV


- Tibshirani, R. 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.
- van de Geer, S. *Estimation and testing under sparsity*. Vol. 2159. Lecture Notes in Mathematics. Lecture notes from the 45th Probability Summer School held in Saint-Four, 2015, École d'Été de Probabilités de Saint-Flour. Springer, 2016, pp. xiii+274.