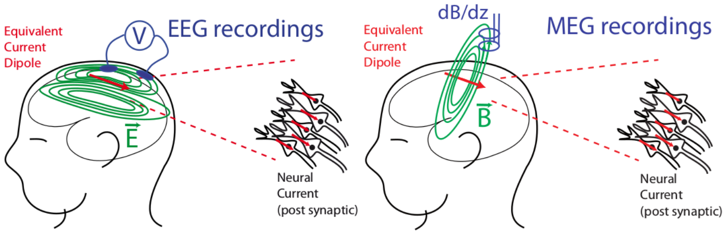
Concomitant Lasso with Repetitions (CLaR): beyond averaging multiple realizations of correlated noise

Quentin Bertrand

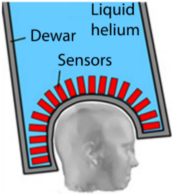
Joint work with:
Mathurin Massias (INRIA)
Alexandre Gramfort (INRIA)
Joseph Salmon (IMAG, Univ Montpellier, CNRS)

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

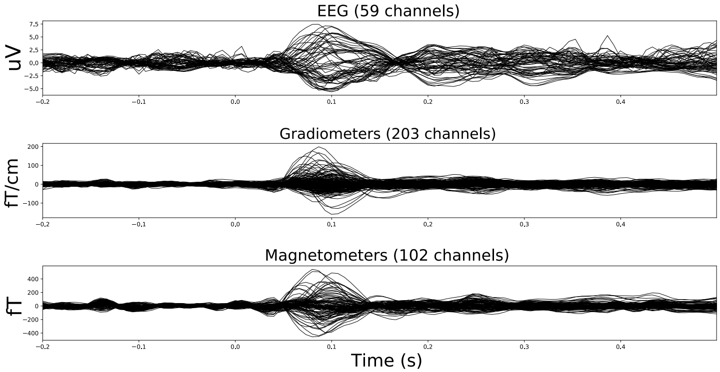


First EEG recordings in 1929 by H. Berger



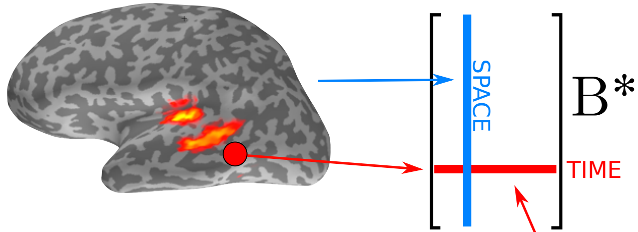
Hôpital La Timone Marseille, France

M/EEG data

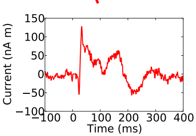


- 3 different types of sensor

Source modeling (discretization with voxels)

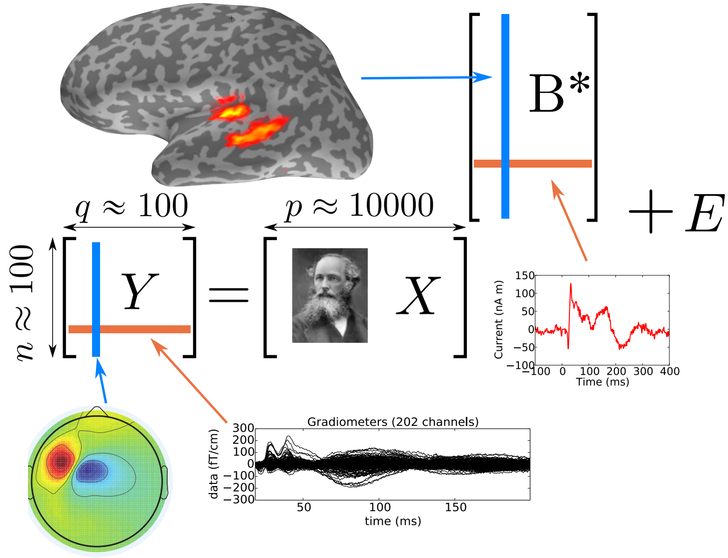


Position a few thousands candidate sources over the brain (*e.g.,* every 5mm)



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

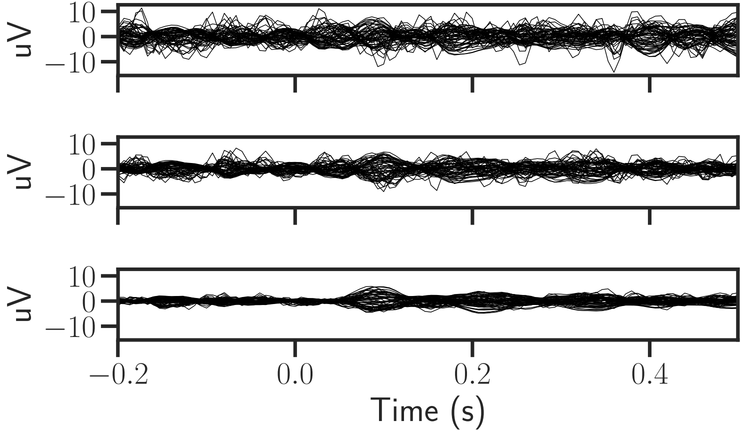
The M/EEG inverse problem: modeling



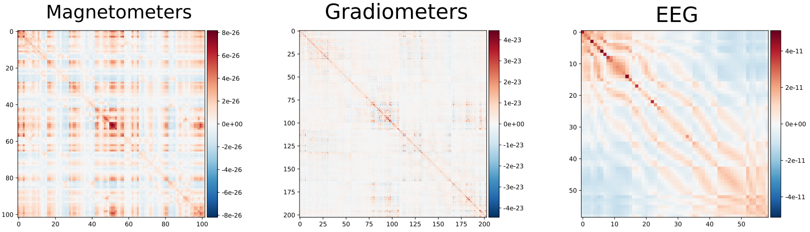
$$n << p$$

Very noisy data: must repeat recordings

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



Noise covariance for each type of sensor



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

A Multi-Task framework

Multi-Task regression notation:

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

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

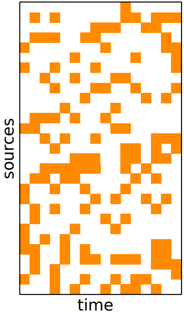
where

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

Multi-Task penalties[1]

Popular convex penalties considered:

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



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

Sparse support: no structure

Penalty: **Lasso type**

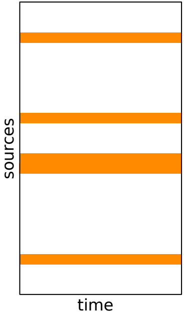
$$\Omega(\mathrm{B}) = \|\mathrm{B}\|_1 = \sum_{j=1}^{p} \sum_{k=1}^{q} |\mathrm{B}_{j,k}|$$

[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)

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



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 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.

Multi-Task data-fitting term

- Classical Multi-Task estimator: use averaged signal

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

- **How to take advantage of the number of repetitions?**

Multi-Task data-fitting term

- Classical Multi-Task estimator: use averaged signal

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

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

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

Multi-Task data-fitting term

- Classical Multi-Task estimator: use averaged signal

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

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

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

- It's a fail! $\hat{\mathrm{B}}^{\text{repet}} = \hat{\mathrm{B}}$ (because of data-fitting loss $\|\cdot\|_F^2$)

Multi-Task data-fitting term

- Classical Multi-Task estimator: use averaged signal

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

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

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

- It's a fail! $\hat{\mathrm{B}}^{\text{repet}} = \hat{\mathrm{B}}$ (because of data-fitting loss $\|\cdot\|_F^2$)
- Moreover $\|\cdot\|_F^2$ is not designed for correlated noise

Multi-Task data-fitting term

- Classical Multi-Task estimator: use averaged signal

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

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

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

- It's a fail! $\hat{\mathrm{B}}^{\text{repet}} = \hat{\mathrm{B}}$ (because of data-fitting loss $\|\cdot\|_F^2$)
- Moreover $\|\cdot\|_F^2$ is not designed for correlated noise
- Need another data-fitting term!

Multi-Task data-fitting term

- Classical Multi-Task estimator: use averaged signal
$$\hat{\mathrm{B}} \in \operatorname*{arg\,min}_{\mathrm{B} \in \mathbb{R}^{p \times q}} \left(\frac{1}{2nq} \left\| \bar{Y} - X\mathrm{B} \right\|_F^2 + \lambda \Omega(\mathrm{B}) \right)$$
- **How to take advantage of the number of repetitions?**
- Intuitive estimator:
$$\hat{\mathrm{B}}^{\text{repet}} \in \operatorname*{arg\,min}_{\mathrm{B} \in \mathbb{R}^{p \times q}} \left(\frac{1}{2nqr} \sum_{l=1}^{r} \left\| Y^{(l)} - X\mathrm{B} \right\|_F^2 + \lambda \Omega(\mathrm{B}) \right)$$
- It's a fail! $\hat{\mathrm{B}}^{\text{repet}} = \hat{\mathrm{B}}$ (because of data-fitting loss $\|\cdot\|_F^2$)
- Moreover $\|\cdot\|_F^2$ is not designed for correlated noise
- Need another data-fitting term!

Reminder on the Lasso theory[2][3] (i.i.d. case, Single-Task)

Theorem

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

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

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.

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

Reminder on the Lasso theory[2][3] (i.i.d. case, Single-Task)

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 $1-\delta$:

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

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.

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

Reminder on the Square root Lasso[4][5][6] (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:

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

λ does not depend on σ_* anymore!

(4) 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.

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

(6) C. Giraud. *Introduction to high-dimensional statistics*. Vol. 138. CRC Press, 2014.

The Smoothed Concomitant Lasso[7] (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

Question: can this estimator (with unknown σ^*) generalize for correlated Gaussian noise?

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

The Smoothed Concomitant Lasso[7] (i.i.d. case, Single-Task)

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

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

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

- jointly convex: alternate minimization

Question: can this estimator (with unknown σ^*) generalize for correlated Gaussian noise?

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

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

SGCL[8]:

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

Benefits

- jointly convex formulation

Drawbacks:

- Statistically: $\mathcal{O}(n^2)$ parameters to estimate for S only nq observations

[8] 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[8]:

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

Benefits

- jointly convex formulation

Drawbacks:

- Statistically: $\mathcal{O}(n^2)$ parameters to estimate for S only nq observations

Question: can this estimator take advantage of the number of repetitions?

[8] 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[8]:

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

Benefits

- jointly convex formulation

Drawbacks:

- Statistically: $\mathcal{O}(n^2)$ parameters to estimate for S only nq observations

Question: can this estimator take advantage of the number of repetitions?

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

Can take advantage of repetitions? Yes!

CLaR[9]:

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

- Statistically: $\mathcal{O}(n^2)$ parameters to estimate for S with nqr observations (r = number of repetitions)

[9] **Bertrand_Massias_Gramfort_Salmon19**.

Proposition

Link with the Trace norm[10]

$$\hat{\mathrm{B}}^{\mathrm{CLaR}} = \underset{\mathrm{B} \in \mathbb{R}^{p \times q}}{\arg\min} \left(\|\cdot\|_{\mathrm{Tr}} \,\square\, \omega_{\underline{\sigma}} \right)(Z) + \lambda n \|\mathrm{B}\|_{2,1} \enspace .$$

where $Z = \frac{1}{\sqrt{q}}[Y^{(1)} - X\mathrm{B}| \dots |Y^{(r)} - X\mathrm{B}]$.

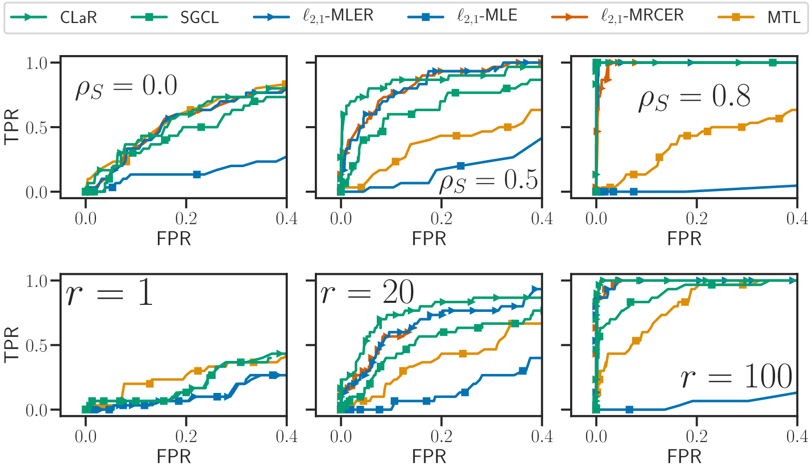
- justification for the estimator introduced heuristically
- generalization of van de Geer[11]

[10] **Bertrand_Massias_Gramfort_Salmon19**.

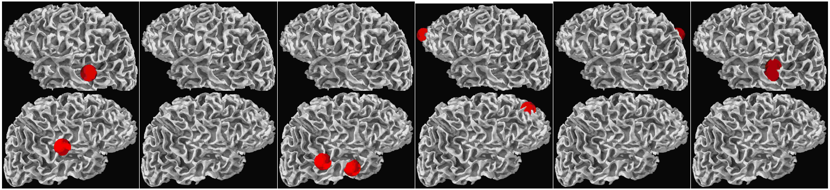
[11] S. van de Geer. *Estimation and testing under sparsity*. Vol. 2159. Lecture Notes in Mathematics. Lecture notes from the 45th Probability Summer School held in Saint-Four, 2015, École d'Été de Probabilités de Saint-Flour. Springer, 2016, pp. xiii+274.

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) CLaR (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

- New estimator to handle correlated noise and repetitions in Multi-Task
- Improved support identification

Conclusion and perspectives

- New estimator to handle correlated noise and repetitions in Multi-Task
- Improved support identification
- Numerical cost "similar" to classical Multi-Task Lasso

Conclusion and perspectives

- New estimator to handle correlated noise and repetitions in Multi-Task
- Improved support identification
- Numerical cost "similar" to classical Multi-Task Lasso
- Ongoing work: non-convex penalties, statistical analysis.

Conclusion and perspectives

- New estimator to handle correlated noise and repetitions in Multi-Task
- Improved support identification
- Numerical cost "similar" to classical Multi-Task Lasso
- Ongoing work: non-convex penalties, statistical analysis.

Merci!

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

A. Gramfort

- Python code online for CLaR https://github.com/QB3/CLaR
- Papers: arXiv[(12), (13)]

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

(13) **Bertrand_Massias_Gramfort_Salmon19**.

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

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

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

What about other norms ?

Smoothing of matrix norm

Huber-like formula for the Frobenius norm

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

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

What about other norms ?

Huber-like formula for the nuclear/trace norm

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

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

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

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

What about other norms ?

Huber-like formula for the nuclear/trace norm

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

γ_i: singular values of Z

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