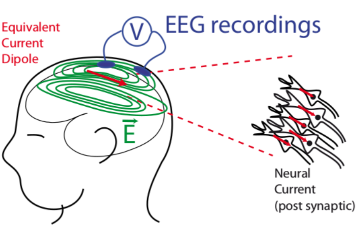
Concomitant Lasso with Repetitions (CLaR): beyond averaging multiple realizations of heteroscedastic noise

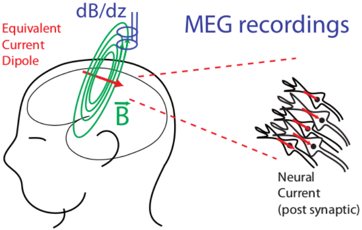
Quentin Bertrand

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

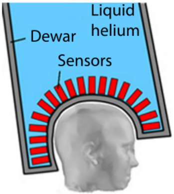
M/EEG inverse problem for brain imaging

- sensors: magneto- and electro-encephalogram measurements during a cognitive experiment
- sources: brain locations
- application to epilepsy treatment, brain aging detection, anesthesia problem



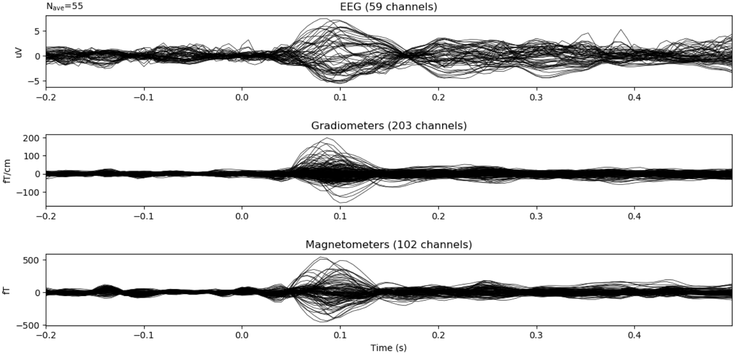


First EEG recordings in 1929 by H. Berger



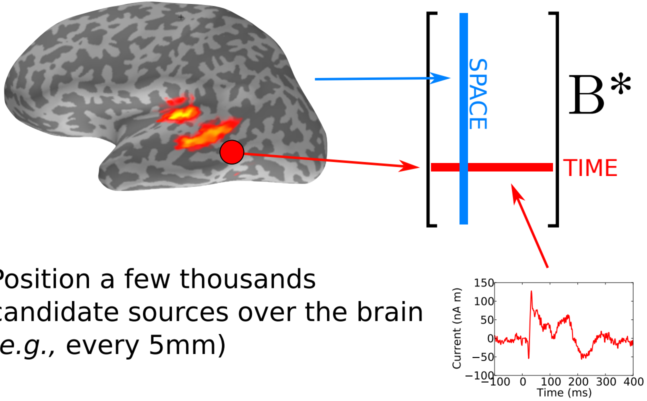
Hôpital La Timone
Marseille, France

M/EEG data



Nave=55
EEG (59 channels)
uV
5
0
−5
−0.2
−0.1
0.0
0.1
0.2
0.3
0.4
Gradiometers (203 channels)
fT/cm
200
100
0
−100
Magnetometers (102 channels)
fT
500
0
−500
Time (s)

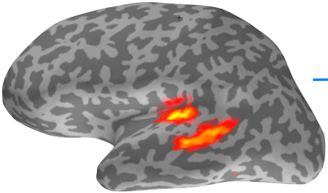
Source modeling (discretization with voxels)



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

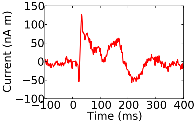
$$B^* \in \mathbb{R}^{p \times q}$$

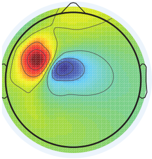
The M/EEG inverse problem: modeling

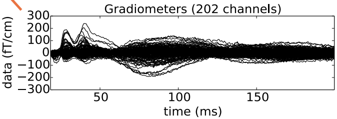


$q \approx 100$ $\quad$ $p \approx 10000$ $\quad$ $n \approx 100$

$$Y = X \, B^* + E$$

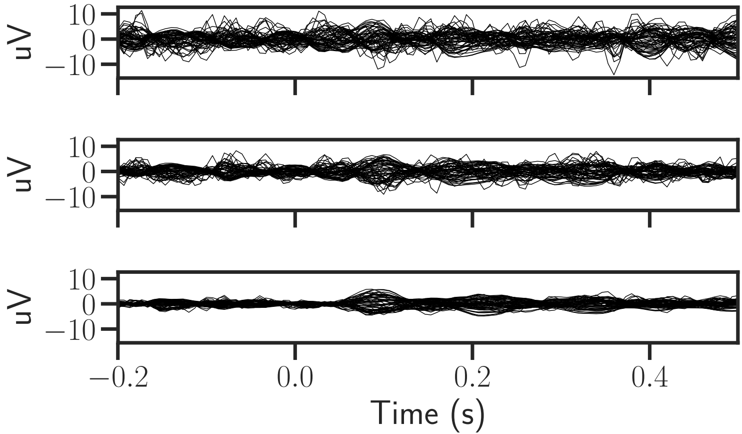




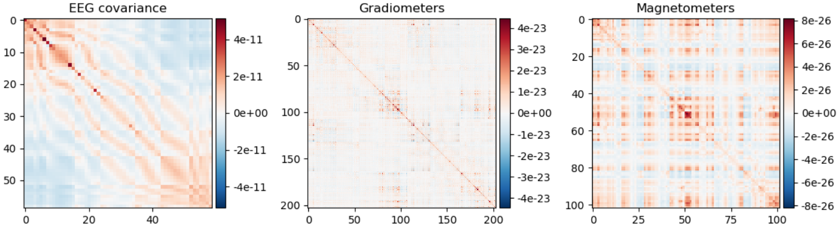


Multiple repetitions structure:

- $r = 5$ repetitions (top)
- $r = 10$ repetitions (middle)
- $r = 50$ repetitions (bottom)



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



- 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; $\bar{Y} = \frac{1}{r} \sum_l Y^{(l)}$
- $X \in \mathbb{R}^{n \times p}$ forward matrix

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

where

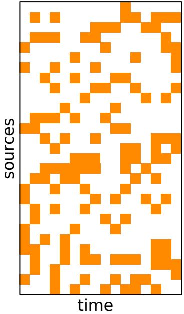
- $\mathrm{B}^* \in \mathbb{R}^{p \times q}$: true source activity matrix (**unknown**)
- $S \in \mathbb{S}^n_{++}$ co-standard deviation matrix[1] (**unknown**)
- $\mathrm{E}^{(1)}, \dots, \mathrm{E}^{(r)} \in \mathbb{R}^{n \times q}$: white Gaussian noise

[1] $S \succeq \underline{\sigma}$ means $S - \underline{\sigma}$ is Semi-Definite Positive

Multi-tasks penalties[2]

Popular convex penalties considered:

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



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

Sparse support: no structure

Penalty: **Lasso type**

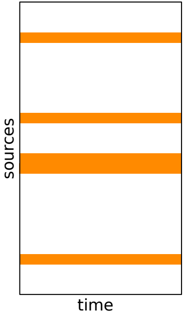
$$\Omega(\mathrm{B}) = \|\mathrm{B}\|_1 = \sum_{j=1}^{p} \sum_{k=1}^{q} |\mathrm{B}_{j,k}|$$

[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)

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



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

[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 data-fitting term

- Classical multi-tasks 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 ?**

Multi-tasks data-fitting term

- Classical multi-tasks 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-tasks data-fitting term

- Classical multi-tasks 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:

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

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

Multi-tasks data-fitting term

- Classical multi-tasks 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 to take into account the correlated noise

Multi-tasks data-fitting term

- Classical multi-tasks 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 to take into account the correlated noise
- Need for another data-fitting term !

Multi-tasks data-fitting term

- Classical multi-tasks 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 to take into account the correlated noise
- Need for another data-fitting term !

The Smoothed Concomitant Lasso[3]

Recall of A. Gramfort talk: in the iid case.
Idea: replacing

- $\|\cdot\|_F^2$
- by $\|\cdot\|_F \,\square\, \underline{\sigma}\,\omega\left(\frac{\cdot}{\underline{\sigma}}\right)(Z) = \min_{\sigma \geq \underline{\sigma}} \left(\frac{\|Z\|_F^2}{2\sigma} + \frac{\sigma}{2}\right)$

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

- λ^* does not depend on the noise level anymore
- efficient block coordinate descent solvers
- generalization to correlated gaussian noise ?

[3] 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[3]

Recall of A. Gramfort talk: in the iid case.
Idea: replacing

- $\|\cdot\|_F^2$
- by $\|\cdot\|_F \,\square\, \underline{\sigma}\,\omega\left(\frac{\cdot}{\underline{\sigma}}\right)(Z) = \min_{\sigma \geq \underline{\sigma}} \left(\frac{\|Z\|_F^2}{2\sigma} + \frac{\sigma}{2}\right)$

$$\boxed{(\hat{\mathrm{B}}^{(\lambda)}, \hat{\sigma}^{(\lambda)}) \in \underset{\mathrm{B} \in \mathbb{R}^p, \sigma \geq \underline{\sigma}}{\arg\min} \frac{\left\|\bar{Y} - X\mathrm{B}\right\|_F^2}{2n\sigma} + \frac{\sigma}{2} + \lambda \|\beta\|_1}$$

- λ^* does not depend on the noise level anymore
- efficient block coordinate descent solvers
- generalization to correlated gaussian noise ?

[3] 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 !

SGCL[4]:

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

Benefits

- jointly convex formulation (=nuclear norm smoothing)
- efficient block coordinate descent solvers

Drawbacks:

- Statistically: $\mathcal{O}(n^2)$ parameters to estimate for S only nq observations
- Computationally: S update cost is $\mathcal{O}(n^3)$ slow in general (SVD computation)

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

Generalization ? Yes !

SGCL[4]:

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

Benefits

- jointly convex formulation (=nuclear norm smoothing)
- efficient block coordinate descent solvers

Drawbacks:

- Statistically: $\mathcal{O}(n^2)$ parameters to estimate for S only nq observations
- Computationally: S update cost is $\mathcal{O}(n^3)$ slow in general (SVD computation)

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

Can deal with repetitions ? Yes !

CLaR[5]:

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

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

[5] Q. Bertrand et al. "Concomitant Lasso with Repetitions (CLaR): beyond averaging multiple realizations of heteroscedastic noise". In: *arXiv preprint arXiv:1902.02509* (2019).

Proposition

Datafit of CLaR[6]

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

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

- justification for the estimator introduced heuristically
- generalization of van de Geer[7]

(6) Q. Bertrand et al. "Concomitant Lasso with Repetitions (CLaR): beyond averaging multiple realizations of heteroscedastic noise". In: *arXiv preprint arXiv:1902.02509* (2019).

(7) 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.

Competitors

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

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

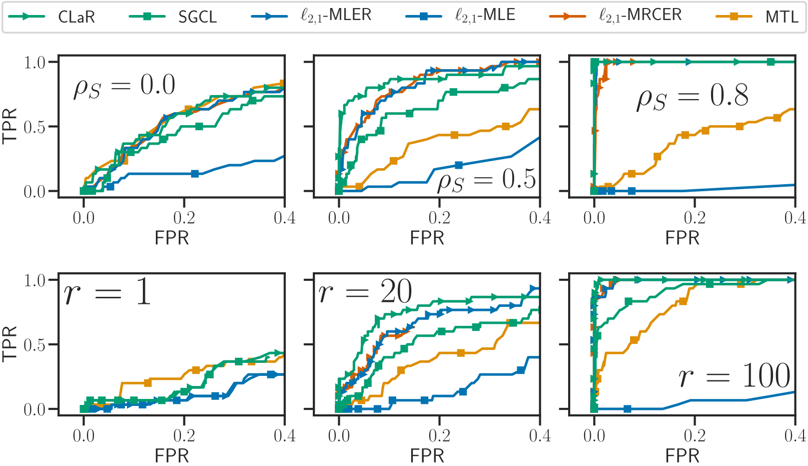
- and its repetitions version ($\ell_{2,1}$-MLER):

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

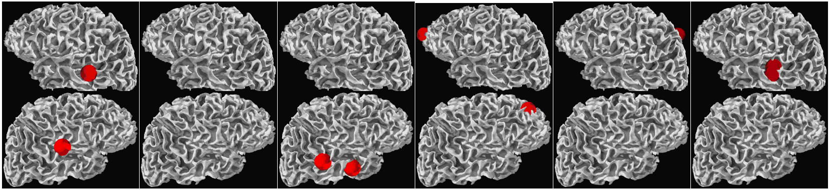
- $\ell_{2,1}$-MLE and $\ell_{2,1}$-MLER are bi-convex but not jointly convex

Simulated scenarios

- $n = 150$, $p = 500$, $q = 100$
- X Toeplitz-correlated: $\mathrm{Cov}(X_i, X_j) = \rho^{|i-j|}$, $\rho_X \in]0, 1[$
- S Toeplitz matrix: $S_{i,j} = \rho^{|i-j|}$, $\rho_S \in]0, 1[$



Real data



(a) CLaR (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 .

- deep sources for SGCL and $\ell_{2,1}$-MRCER not visible

Conclusion and perspectives

- New insights for handling correlated noise in multi-task
- Handling refined noise structure benefits:
  improve support identification (and prediction)

Conclusion and perspectives

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

Conclusion and perspectives

- New insights for handling correlated noise in multi-task
- Handling refined noise structure benefits:
  improve support identification (and prediction)
- Numerical cost equivalent to classical Multi-Task Lasso for "simple" noise structure (*e.g.*, block homoscedastic)
- Future work: non-convex penalties, statistical analysis, etc.

Conclusion and perspectives

- New insights for handling correlated noise in multi-task
- Handling refined noise structure benefits:
  improve support identification (and prediction)
- Numerical cost equivalent to classical Multi-Task Lasso for "simple" noise structure (*e.g.*, block homoscedastic)
- Future work: non-convex penalties, statistical analysis, etc.

Merci!

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

A. Gramfort

- Paper: arXiv[8], [9]
- Python code online for CLaR https://github.com/QB3/CLaR

---

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

[9] Q. Bertrand et al. "Concomitant Lasso with Repetitions (CLaR): beyond averaging multiple realizations of heteroscedastic noise". In: *arXiv preprint arXiv:1902.02509* (2019).

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**