Generalized Concomitant Multi-Task Lasso for sparse multimodal regression

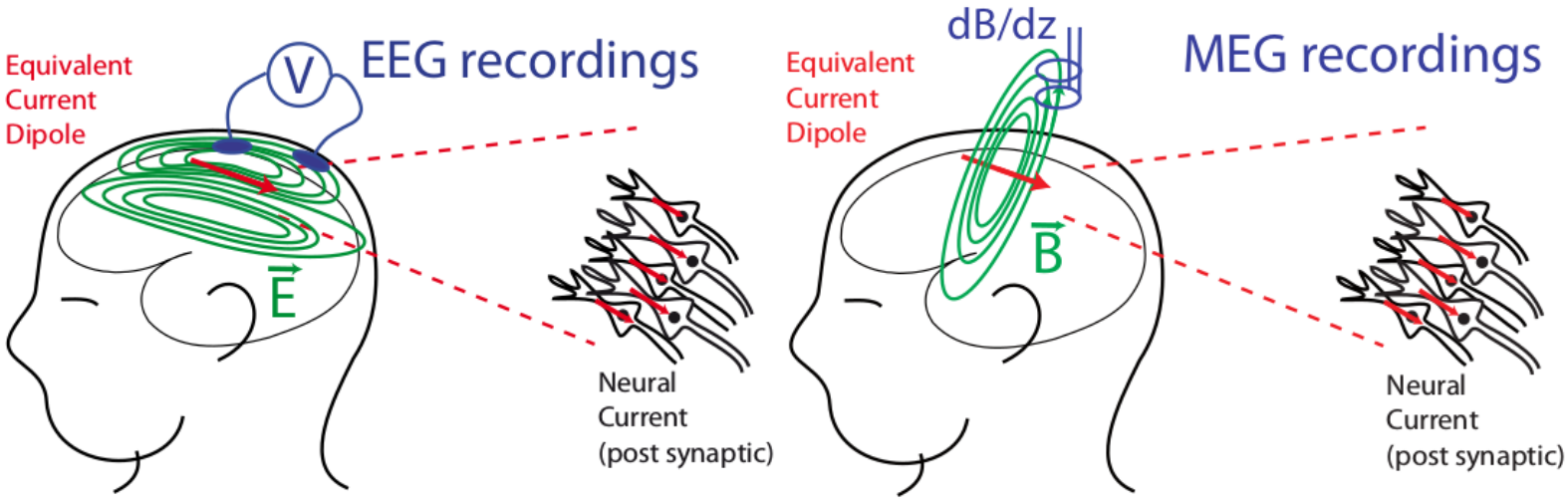
Mathurin Massias
https://mathurinm.github.io
INRIA Saclay

Joint work with:
Olivier Fercoq (Télécom ParisTech)
Alexandre Gramfort (INRIA Saclay)
Joseph Salmon (Télécom ParisTech)

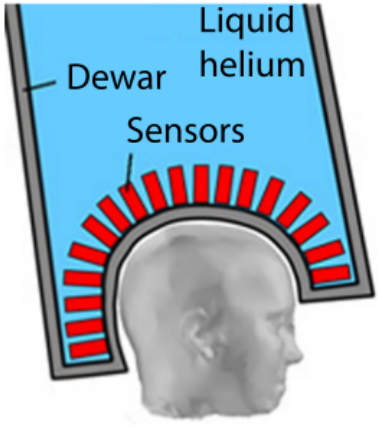
Table of Contents

"One" motivation: M/EEG inverse problem

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

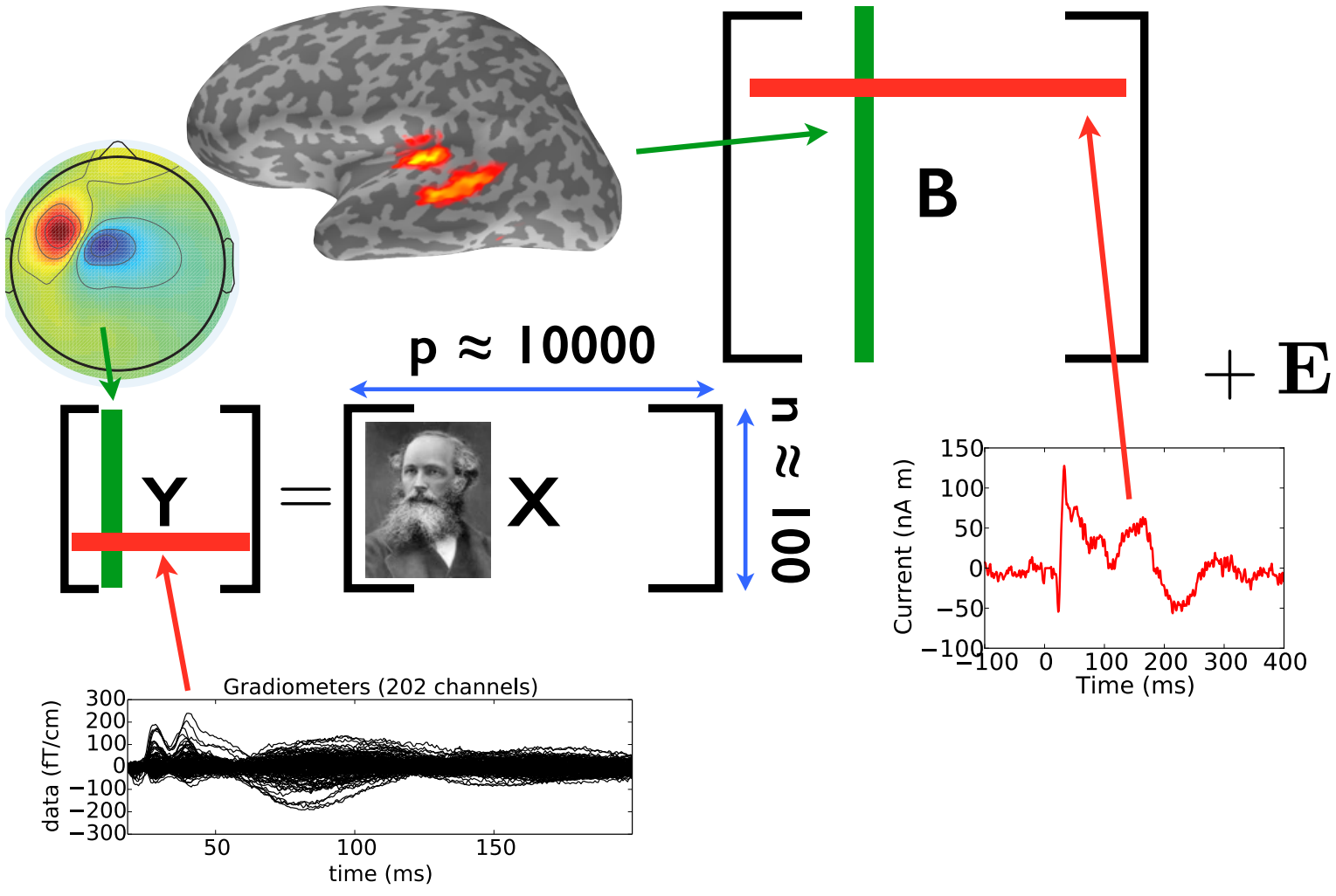


First EEG recordings in 1929 by H. Berger



Hôpital La Timone
Marseille, France

The M/EEG inverse problem: modelisation



The M/EEG inverse problem: modelisation

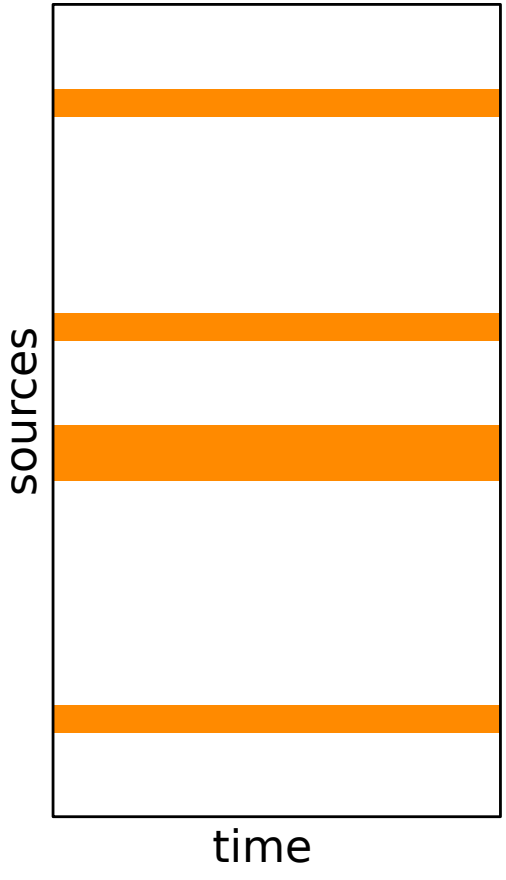
Multi-task regression:

- n observations
- q tasks
- p features
- $Y \in \mathbb{R}^{n \times q}$ observation matrix
- $X \in \mathbb{R}^{n \times p}$ forward matrix

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

where
$\mathrm{B}^* \in \mathbb{R}^{p \times q}$ is the true source activity matrix
$\mathrm{E} \in \mathbb{R}^{n \times q}$ is an additive white noise.

$\ell_{2,1}$ regularization



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

Penalty: Group-Lasso ($\ell_{2,1}$)

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

(B_j: j^{th} row of B)

$\rightarrow$ we solve:

$$\hat{\mathrm{B}} \in \underset{\mathrm{B} \in \mathbb{R}^{p \times q}}{\arg\min} \frac{1}{2} \|Y - X\mathrm{B}\|^2 + \lambda \|\mathrm{B}\|_{2,1}$$

a.k.a. Multiple Measurement Vector (MMV) in signal processing or multi-task Lasso in ML [Obozinski et al., 2010]

Table of Contents

Choice of λ and noise level

- The noise E is assumed white gaussian, with i.i.d. entries $\mathrm{E}_{i,t} \sim \mathcal{N}(0, \sigma^2)$
- For optimal performance of the Lasso, λ should be proportional to the noise level ($\lambda \propto \sigma\sqrt{\frac{\log p}{n}}$) [Bickel et al., 2009]
- Yet σ is unknown in practice !

Joint estimation of β and σ

(illustrated on Lasso for simplicity: model is $y = X\beta^* + \varepsilon$)

Intuitive idea:

- run Lasso with some λ, get $\hat{\beta}$
- estimate σ with residuals: $\sigma = \|y - X\hat{\beta}\|/\sqrt{n}$
- relaunch Lasso with $\lambda \propto \sigma$
- etc.

Note: this is the original implementation proposed for the Scaled-Lasso [Sun and Zhang, 2012]

Concomitant Lasso

Concomitant Lasso [Owen, 2007] (inspired by Huber [1981]):

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

also equivalent to the Square-root/Scaled Lasso [Belloni et al., 2011, Sun and Zhang, 2012].

Note: note that $\frac{\sigma}{2}$ acts as a penalty over the noise

Solving the Concomitant Lasso

A jointly convex formulation (w.r.t. both β and σ):

$$\underset{\beta \in \mathbb{R}^p, \sigma > 0}{\arg\min} \frac{\|y - X\beta\|^2}{2n\sigma} + \frac{\sigma}{2} + \lambda \|\beta\|_1$$

- β update: smooth + separable function, make CD steps [Friedman et al., 2007].

Solving the Concomitant Lasso

A jointly convex formulation (w.r.t. both β and σ):

$$\underset{\beta \in \mathbb{R}^p, \sigma > 0}{\arg\min} \frac{\|y - X\beta\|^2}{2n\sigma} + \frac{\sigma}{2} + \lambda \|\beta\|_1$$

- β update: smooth + separable function, make CD steps [Friedman et al., 2007].
- σ update: 1D optimization problem, closed-form $\sigma = \|y - X\beta\| / \sqrt{n}$.

Solving the Concomitant Lasso

A jointly convex formulation (w.r.t. both β and σ):

$$\underset{\beta \in \mathbb{R}^p, \sigma > 0}{\arg\min} \frac{\|y - X\beta\|^2}{2n\sigma} + \frac{\sigma}{2} + \lambda \|\beta\|_1$$

- β update: smooth + separable function, make CD steps [Friedman et al., 2007].
- σ update: 1D optimization problem, closed-form $\sigma = \|y - X\beta\| / \sqrt{n}$.

Solving the Concomitant Lasso

A jointly convex formulation (w.r.t. both β and σ):

$$\underset{\beta \in \mathbb{R}^p, \sigma > 0}{\arg\min} \frac{\|y - X\beta\|^2}{2n\sigma} + \frac{\sigma}{2} + \lambda \|\beta\|_1$$

- β update: smooth + separable function, make CD steps [Friedman et al., 2007].
- σ update: 1D optimization problem, closed-form $\sigma = \|y - X\beta\| / \sqrt{n}$.

But what if we hit $y = X\beta$?

Smoothed Concomitant Lasso

Smoothed Concomitant Lasso [Ndiaye et al., 2017]:

$$\underset{\beta \in \mathbb{R}^p, \sigma > \underline{\sigma}}{\arg\min} \; \frac{\|y - X\beta\|^2}{2n\sigma} + \frac{\sigma}{2} + \lambda \|\beta\|_1$$

Smoothed Concomitant Lasso

Smoothed Concomitant Lasso [Ndiaye et al., 2017]:

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

Called *smoothed* following the terminology of Nesterov [2005], because it amounts to smoothing in the dual:

$$\hat{\theta} = \underset{\theta \in \Delta_X}{\arg\max} \langle y, \lambda\theta \rangle + \underline{\sigma} \left(\frac{1}{2} - \frac{n\lambda^2}{2} \|\theta\|^2 \right)$$

for $\Delta_X = \left\{ \theta \in \mathbb{R}^n : \|X^\top \theta\|_\infty \leq 1, \|\theta\| \leq \frac{1}{\lambda n} \right\}$.

$\rightarrow$ state-of-the-art solvers as fast as for the Lasso.

Table of Contents

What about more complex noise models?

What can we do if the noise is not white?

Smoothed Generalized Concomitant Lasso (SGCL):

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

with $\|Z\|_A = \mathrm{Tr}\, Z^\top A Z$ and $\underline{\Sigma} = \underline{\sigma}\, \mathrm{Id}$.

What about more complex noise models?

What can we do if the noise is not white?

Smoothed Generalized Concomitant Lasso (SGCL):

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

with $\|Z\|_A = \mathrm{Tr}\, Z^\top A Z$ and $\underline{\Sigma} = \underline{\sigma}\, \mathrm{Id}$.

In the case $\Sigma = \sigma\, \mathrm{Id}$, we recover the Smoothed Concomitant.

What about more complex noise models?

What can we do if the noise is not white?

Smoothed Generalized Concomitant Lasso (SGCL):

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

with $\|Z\|_A = \mathrm{Tr}\, Z^\top A Z$ and $\underline{\Sigma} = \underline{\sigma}\, \mathrm{Id}$.

In the case $\Sigma = \sigma\, \mathrm{Id}$, we recover the Smoothed Concomitant.

Note: the noise penalty is now on the sum of the eigenvalues of Σ

Solving the SGCL

Jointly convex formulation, we can still use alternate minimization and use the duality gap as stopping criterion.

Σ fixed: smooth + ℓ_1-type, BCD works

Solving the SGCL

Jointly convex formulation, we can still use alternate minimization and use the duality gap as stopping criterion.

B fixed: with the current residuals $R = Y - X\mathrm{B}$, the problem is:

$$\underset{\Sigma \in \mathbb{S}_{++}^n, \Sigma \succeq \underline{\Sigma}}{\arg\min} \ \frac{1}{2nq} \operatorname{Tr}[R^\top \Sigma^{-1} R] + \frac{1}{2n} \operatorname{Tr}(\Sigma) \ .$$

Closed-form solution: if $U^\top \operatorname{diag}(s_1, \ldots, s_n) U$ is the SVD of $RR^\top$:

$$\underset{\Sigma}{\arg\min} = U^\top \operatorname{diag}(\max(\underline{\sigma}, \sqrt{s_1}), \ldots, \max(\underline{\sigma}, \sqrt{s_n})) U \ .$$

Alternate minimization

Algorithm: ALTERNATE MIN. FOR MULTI-TASK SGCL

input : $X, Y, \underline{\Sigma}, \lambda, f, T$

init : $\mathrm{B} = 0_{p,q}$, $\Sigma^{-1} = \underline{\Sigma}^{-1}$, $R = Y$

for $\mathrm{iter} = 1, \ldots, T$ **do**

  if $\mathrm{iter} = 1 \pmod f$ **then**

    $\Sigma \leftarrow \Psi(R, \underline{\Sigma})$ // closed-form sol. of minimization in Σ

    for $j = 1, \ldots, p$ **do**

      $L_j = X_j^\top \Sigma^{-1} X_j$ // Lipschitz constants

  for $j = 1, \ldots, p$ **do**

    $R \leftarrow R + X_j \mathrm{B}_j$ // partial residual update

    $\mathrm{B}_j \leftarrow \mathrm{BST}\Big(\frac{X_j^\top \Sigma^{-1} R}{L_j}, \frac{\lambda n q}{L_j}\Big)$ // coef. update

    $R \leftarrow R - X_j \mathrm{B}_j$ // residual update

return B, Σ

Complexity? OK if we store $\Sigma^{-1}X$, and $\Sigma^{-1}R$ instead of R.

Main drawbacks

- Σ update OK for M/EEG because $n = 300$, but $\mathcal{O}(n^3)$ SVD problematic otherwise.
- $\mathcal{O}(n^2)$ parameters to infer for Σ, with nq observations: works only for $q \gtrsim 10n$.

Table of Contents

Block Homoscedastic model

In our case we record 3 different types of signals:

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

$\neq$ physical natures, different noise levels

Observations are divided into 3 blocks $\&$ the partition is known.

Block Homoscedastic model

K groups of observations:

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

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

For each block, homoscedastic model with white noise:

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

and entries of E^k i.i.d. $\sim \mathcal{N}(0, 1)$

Smoothed Block Homoscedastic Concomitant (SBHCL)

Plugging a diagonal Σ (constant over consecutive blocks) into the general model yields:

Block Homoscedastic Concomitant:

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

$\rightarrow \Sigma$ down from $\frac{n(n-1)}{2}$ parameters (hopeless without more structure) to K.

Solving the SBHCL

- (block) coordinate descent steps remain the same
- computing $\Sigma^{-1}R$ for the BCD is easier

Solving the SBHCL

- (block) coordinate descent steps remain the same
- computing $\Sigma^{-1}R$ for the BCD is easier
- σ_k's updates are simple and can even be performed at each B_j update (as for the concomitant)

Solving the SBHCL

- (block) coordinate descent steps remain the same
- computing $\Sigma^{-1}R$ for the BCD is easier
- σ_k's updates are simple and can even be performed at each B_j update (as for the concomitant)

Alternate minimization

Algorithm: ALTERNATE MIN. FOR MULTI-TASK SBHCL

input : $X^1, \dots, X^K, Y^1, \dots, Y^K, \underline{\sigma}_1, \dots, \underline{\sigma}_K, \lambda, T$

init : $\mathrm{B} = 0_{p,q}$, $\forall k \in [K], \sigma_k = \|Y^k\|/\sqrt{n_k q}$, $R^k = Y^k$,
$\forall k \in [K], \forall j \in [p], L_{k,j} = \|X_j^k\|_2^2$

for iter $= 1, \dots, T$ **do**

  for $j = 1, \dots, p$ **do**

    for $k = 1, \dots, K$ **do**

      $R^k \leftarrow R^k + X_j^k \mathrm{B}_j$ // residual update

    $\mathrm{B}_j \leftarrow \mathrm{BST}\Big(\sum_{k=1}^{K} \frac{X_j^{k\top} R^k}{\sigma_k}, \lambda n q\Big) \Big/ \sum_{k=1}^{K} \frac{L_{k,j}}{\sigma_k}$

    for $k = 1, \dots, K$ **do**

      $R^k \leftarrow R^k - X_j^k \mathrm{B}_j$ // residual update

      $\sigma_k \leftarrow \underline{\sigma}_k \vee \frac{\|R^k\|}{\sqrt{n_k q}}$ // smart std dev update

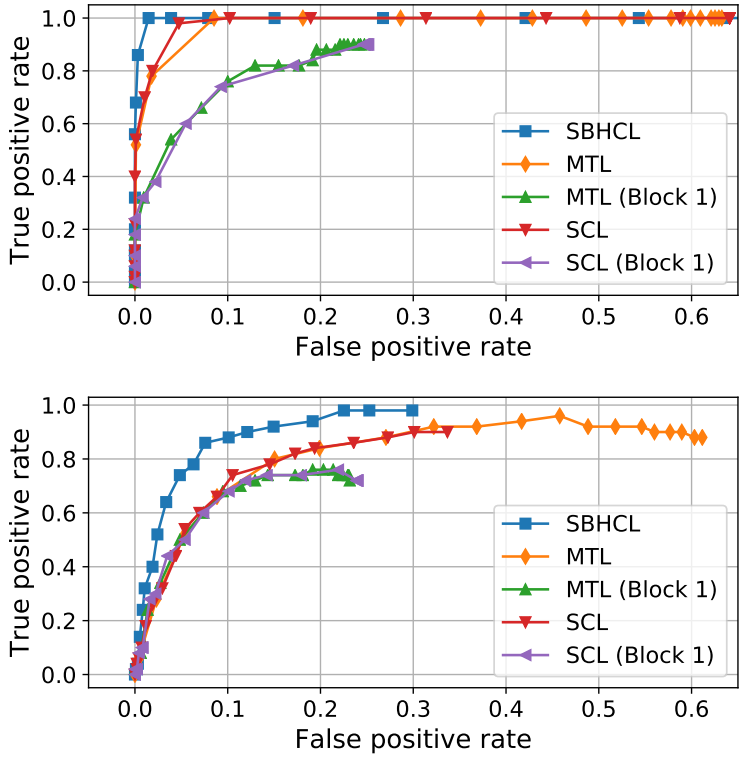
return $\mathrm{B}, \sigma_1, \dots, \sigma_k$

In practice

Design:

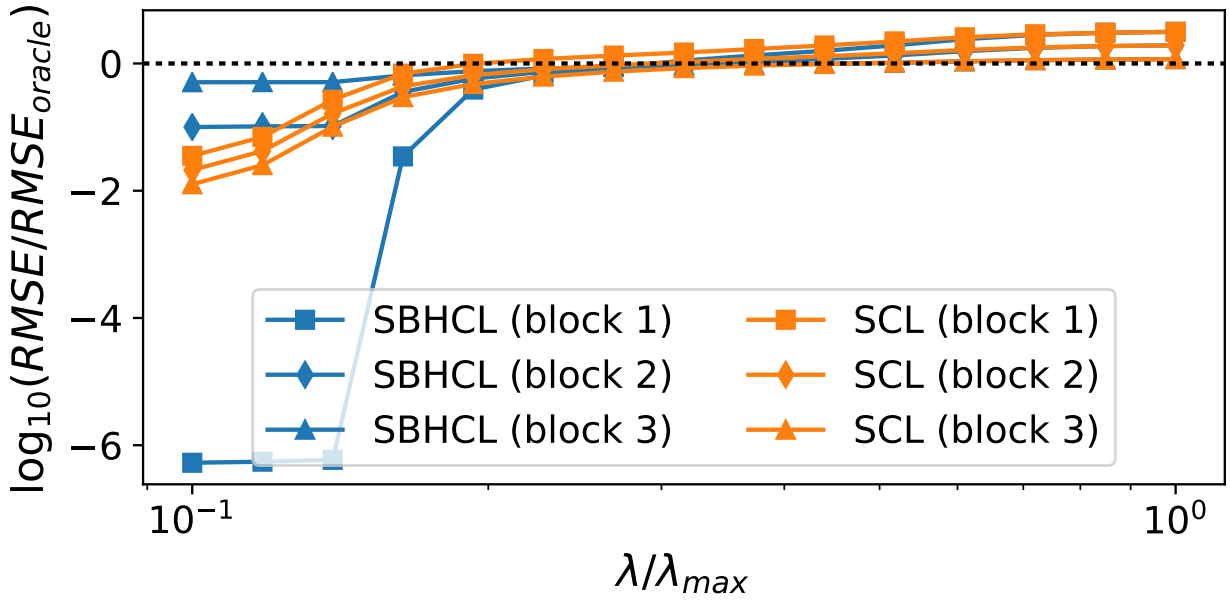
- $(n, p, q) = 300, 1000, 100$
- X Toeplitz-correlated: $Cov(X_i, X_j) = \rho^{|i-j|}$, $\rho \in]0, 1[$
- 3 blocks with noise in ratio 1, 2, 5

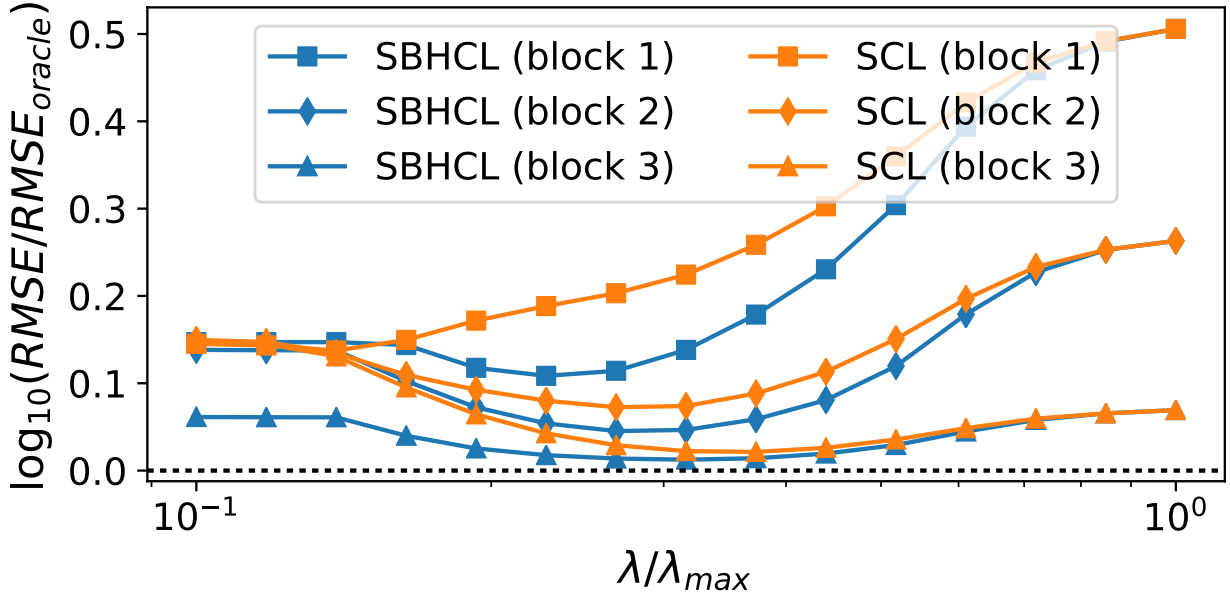
Support recovery



ROC curves of true support recovery: for SBHCL, MTL and SCL on all blocks, and the MTL and SCL on the least noisy block. Top: SNR=1, $\rho = 0.1$, Bottom: $\mathrm{SNR} = 1$, $\rho = 0.9$.

Prediction performance





Take home message

- more general noise models: possible to estimate full covariance if there are enough tasks
- if more noise structure is known (*e.g.*, block homoscedastic model): not more costly than the multi-task Lasso (MTL)
- taking into account multiple noise levels helps: both for prediction and support identification
- using additional (though noisier) data helps!

Python code is available at https://github.com/mathurinm/SHCL.

This work was funded by ERC Starting Grant SLAB ERC-YStG-676943.
Slides powered by MooseTeX.

References I

G. Obozinski, B. Taskar, and M. I. Jordan. Joint covariate selection and joint subspace selection for multiple classification problems. *Statistics and Computing*, 20(2):231–252, 2010.

P. J. Bickel, Y. Ritov, and A. B. Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *Ann. Statist.*, 37(4):1705–1732, 2009.

T. Sun and C.-H. Zhang. Scaled sparse linear regression. *Biometrika*, 99(4): 879–898, 2012.

A. B. Owen. A robust hybrid of lasso and ridge regression. *Contemporary Mathematics*, 443:59–72, 2007.

P. J. Huber. *Robust Statistics*. John Wiley & Sons Inc., 1981.

A. Belloni, V. Chernozhukov, and L. Wang. Square-root Lasso: pivotal recovery of sparse signals via conic programming. *Biometrika*, 98(4):791–806, 2011.

J. Friedman, T. J. Hastie, H. Höfling, and R. Tibshirani. Pathwise coordinate optimization. *Ann. Appl. Stat.*, 1(2):302–332, 2007.

E. Ndiaye, O. Fercoq, A. Gramfort, V. Leclère, and J. Salmon. Efficient smoothed concomitant Lasso estimation for high dimensional regression. In *NCMIP*, 2017.

Y. Nesterov. Smooth minimization of non-smooth functions. *Math. Program.*, 103(1):127–152, 2005.