Handling correlated and repeated measurements with the smoothed multivariate square-root Lasso

Quentin Bertrand, Mathurin Massias, Alexandre Gramfort, Joseph Salmon

Handling correlated and repeated measurements with the smoothed Multivariate square-root Lasso

Quentin Bertrand[1*] **Mathurin Massias**[1*] **Alexandre Gramfort**[1] **Joseph Salmon**[2]
[1]INRIA, Université Paris Saclay, Palaiseau, France
[2]IMAG, Univ. Montpellier, CNRS, Montpellier, France
[*]These authors contributed equally.

Abstract

Sparsity promoting norms are frequently used in high dimensional regression. A limitation of such Lasso-type estimators is that the optimal regularization parameter depends on the unknown noise level. Estimators such as the concomitant Lasso address this dependence by jointly estimating the noise level and the regression coefficients. Additionally, in many applications, the data is obtained by averaging multiple measurements: this reduces the noise variance, but it dramatically reduces sample sizes and prevents refined noise modeling. In this work, we propose a concomitant estimator that can cope with complex noise structure by using non-averaged measurements. The resulting optimization problem is convex and amenable, thanks to smoothing theory, to state-of-the-art optimization techniques that leverage the sparsity of the solutions. Practical benefits are demonstrated on toy datasets, realistic simulated data and real neuroimaging data.

1 Introduction

In many statistical applications, the number of parameters p is much larger than the number of observations n. A popular approach to tackle linear regression problems in such scenarios is to consider convex ℓ_1-type penalties, as popularized by Tibshirani (1996). The use of these penalties relies on a regularization parameter λ trading data fidelity versus sparsity. Unfortunately, Bickel et al. (2009) showed that, in the case of homoscedastic Gaussian noise, the optimal λ is proportional to the standard deviation of the noise – referred to as *noise level*. Because the latter is rarely known in practice, one can jointly estimate the noise level and the regression coefficients, following pioneering work on concomitant estimation (Huber and Dutter, 1974; Huber, 1981). Adaptations to sparse regression (Owen, 2007) have been analyzed under the names of Square root Lasso (Belloni et al., 2011) or Scaled Lasso (Sun and Zhang, 2012).

In the aforementioned contributions, the noise parameter is a single scalar, the variance. Yet, in various applied settings, mixing data of different nature or from different sources is customary to increase the number of observations. This often leads to heteroscedasticity[1]: the data may be contaminated with non-white noise (see the statistical analysis of Daye et al. 2012; Wagener and Dette 2012; Kolar and Sharpnack 2012; Dalalyan et al. 2013 for non-uniform noise levels). Heteroscedasticity occurs for magneto-electroencephalographic (M/EEG) data, where observations come from three types of sensors (gradiometers, magnetometers and electrodes), leading to very different amplitudes and noise covariance matrices. To address this problem, estimators based on non-convex optimization problems were proposed (Lee and Liu, 2012) and analyzed for sub-Gaussian covariance matrices (Chen and Banerjee, 2017) through penalized Maximum Likelihood Estimation (MLE). Other estimators (Rothman et al., 2010; Rai et al., 2012) assume that the inverse of the covariance (the *precision*

[1]The term heteroscedastic may differ across communities, in this work it means correlated Gaussian noise.

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

matrix) is sparse, but the underlying optimization problems remain non-convex. A convex approach to heteroscedastic regression, the Smooth Generalized Concomitant Lasso (SGCL) was proposed by Massias et al. (2018a). Relying on smoothing techniques (Nesterov, 2005; Beck and Teboulle, 2012), the SGCL jointly estimates the regression coefficients and the noise *co-standard deviation matrix* (*i.e.,* the square root of the noise covariance matrix). However, in certain applications, such as M/EEG, the number of parameters in the co-standard deviation matrix ($\approx 10^4$) is typically equal to the number of observations, making it statistically hard to estimate accurately.

When observations are contaminated with a strong noise and the signal-to-noise ratio (SNR) is too low, a natural idea, if possible, is to repeat the measurements and make an average. Indeed, under the assumption that the signal of interest is corrupted by additive independent realizations of noise, averaging multiple measurements reduces the noise variance by the number of repetitions. Popular estimators for M/EEG usually discard the individual observations, therefore relying on homoscedastic noise models (Ou et al., 2009; Gramfort et al., 2013).

In this work we propose Concomitant Lasso with Repetitions (CLaR), an estimator that is

- designed to exploit all available measurements collected during repetitions of experiments,
- defined as the solution of a *convex* minimization problem, handled efficiently by proximal block coordinate descent techniques,
- built thanks to an *explicit* connection with nuclear norm smoothing[2],
- shown (through extensive benchmarks *w.r.t.* existing estimators) to leverage experimental repetitions to improve support identification
- available as open source code to reproduce all the experiments.

In Section 2, we recall the framework of concomitant estimation, and introduce CLaR. In Section 3, we detail the properties of CLaR, and derive an algorithm to solve it. Finally, Section 4 is dedicated to experimental results.

2 Heteroscedastic concomitant estimation

Probabilistic model Let r be the number of repetitions of the experiment. The r observation matrices are denoted $Y^{(1)}, \dots, Y^{(r)} \in \mathbb{R}^{n \times q}$ with n the number of sensors/samples and q the number of tasks/time samples. The mean over the repetitions of the observation matrices is written $\bar{Y} = \frac{1}{r}\sum_{l=1}^{r} Y^{(l)}$. Let $X \in \mathbb{R}^{n \times p}$ be the design (or gain) matrix, with p features stored column-wise: $X = [X_{:1}|\dots|X_{:p}]$, where for a matrix $A \in \mathbb{R}^{m \times n}$ its j^{th} column (*resp.* row) is denoted $A_{:j} \in \mathbb{R}^{m \times 1}$ (*resp.* $A_{j:} \in \mathbb{R}^{1 \times n}$. The matrix $\mathrm{B}^* \in \mathbb{R}^{p \times q}$ contains the coefficients of the linear regression model. Each measurement follows the following model:

$$\forall l \in [r], \quad Y^{(l)} = X\mathrm{B}^* + S^*\mathrm{E}^{(l)} \ , \tag{1}$$

where the entries of $\mathrm{E}^{(l)}$ are i.i.d. samples from standard normal distributions, the $\mathrm{E}^{(l)}$'s are independent, and $S^* \in \mathcal{S}^n_{++}$ is the co-standard deviation matrix, and $\mathcal{S}^n_{++}$ (*resp.* $\mathcal{S}^n_+$) stands for the set of positive (*resp.* semi-definite positive) matrices. Note that even if the observations $Y^{(1)}, \dots, Y^{(r)}$ differ because of the noise $\mathrm{E}^{(1)}, \dots, \mathrm{E}^{(r)}$, B^* and the noise structure S^* are shared across repetitions.

This is a natural assumption for stable physical systems observed with sensor or background noise.

Notation We write $\|\cdot\|$ (*resp.* $\langle \cdot, \cdot \rangle$) for the Euclidean norm (*resp.* inner product) on vectors and matrices, $\|\cdot\|_{p_1}$ for the ℓ_{p_1} norm, for any $p_1 \in [1, \infty)$. For a matrix $\mathrm{B} \in \mathbb{R}^{p \times q}$, $\|\mathrm{B}\|_{2,1} = \sum_{j=1}^{p} \|\mathrm{B}_{j:}\|$ (*resp.* $\|\mathrm{B}\|_{2,\infty} = \max_{j \in [p]} \|\mathrm{B}_{j:}\|$), and for any $p_1 \in [1, \infty]$, we write $\|\mathrm{B}\|_{\mathscr{S}, p_1}$ for the Schatten p_1-norm (*i.e.,* the ℓ_{p_1} norm of the singular values of B). The unit ℓ_{p_1} ball is written $\mathcal{B}_{p_1}$, $p_1 \in [1, \infty)$. For S_1 and $S_2 \in \mathcal{S}^n_+$, $S_1 \succeq S_2$ if $S_1 - S_2 \in \mathcal{S}^n_+$. When we write $S_1 \succeq S_2$ from now on we implicitly assume that both matrix belong to $\mathcal{S}^n_+$. For a square matrix $A \in \mathbb{R}^{n \times n}$, $\mathrm{Tr}(A)$ represents the trace of A and $\|A\|_S = \sqrt{\mathrm{Tr}(A^\top S A)}$ is the Mahalanobis norm induced by $S \in \mathcal{S}^n_{++}$. For $a, b \in \mathbb{R}$, we denote $(a)_+ = \max(a, 0)$, $a \vee b = \max(a, b)$ and $a \wedge b = \min(a, b)$. The block soft-thresholding operator at level $\tau > 0$, is denoted $\mathrm{BST}(\cdot, \tau)$, and reads for any vector x, $\mathrm{BST}(x, \tau) = (1 - \tau / \|x\|)_+ x$. The identity matrix of size $n \times n$ is denoted Id_n, and $[r]$ is the set of integers from 1 to r.

[2]Other Schatten norms are treated in Appendix A.2.

2.1 The proposed CLaR estimator

To leverage the multiple repetitions while taking into account the noise structure, we introduce the Concomitant Lasso with Repetitions:

Definition 1 (CLaR)**.** CLaR estimates the parameters of Model (1) by solving:

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

where $\lambda > 0$ controls the sparsity of $\hat{\mathrm{B}}^{\mathrm{CLaR}}$ and $\underline{\sigma} > 0$ controls the smallest eigenvalue of $\hat{\mathrm{S}}^{\mathrm{CLaR}}$.

2.2 Connections with concomitant Lasso on averaged data

In low SNR settings, a standard way to deal with strong noise is to use the averaged observation $\bar{Y} \in \mathbb{R}^{n\times q}$ instead of the raw observations. The associated model reads:

$$\bar{Y} = X\mathrm{B}^* + \tilde{S}^*\tilde{\mathrm{E}}\ , \tag{3}$$

with $\tilde{S}^* \triangleq S^*/\sqrt{r}$ and $\tilde{\mathrm{E}}$ has $i.i.d.$ entries drawn from a standard normal distribution. The SNR[3] is multiplied by $\sqrt{r}$, yet the number of samples goes from rnq to nq, making it statistically difficult to estimate the $\mathcal{O}(n^2)$ parameters of S^*. CLaR generalizes the Smoothed Generalized Concomitant Lasso (Massias et al., 2018a), which has the drawback of only targeting averaged observations:

Definition 2 (SGCL, Massias et al. 2018a)**.** SGCL estimates the parameters of Model (3), by solving:

$$(\hat{\mathrm{B}}^{\mathrm{SGCL}}, \hat{\mathrm{S}}^{\mathrm{SGCL}}) \in \underset{\substack{\mathrm{B}\in\mathbb{R}^{p\times q}\\ \tilde{S}\succeq\underline{\sigma}/\sqrt{r}\,\mathrm{Id}_n}}{\arg\min}\, \tilde{f}(\mathrm{B}, \tilde{S}) + \lambda \left\|\mathrm{B}\right\|_{2,1}, \text{ with } \tilde{f}(\mathrm{B}, \tilde{S}) \triangleq \frac{\left\|\bar{Y} - X\mathrm{B}\right\|_{\tilde{S}^{-1}}^2}{2nq} + \frac{\mathrm{Tr}(\tilde{S})}{2n}. \tag{4}$$

Remark 3. Note that $\hat{\mathrm{S}}^{\mathrm{CLaR}}$ estimates S^*, while $\hat{\mathrm{S}}^{\mathrm{SGCL}}$ estimates $\tilde{S}^* = S^*/\sqrt{r}$. Since we impose the constraint $\hat{\mathrm{S}}^{\mathrm{CLaR}} \succeq \underline{\sigma}\,\mathrm{Id}_n$, we rescale the constraint so that $\hat{\mathrm{S}}^{\mathrm{SGCL}} \succeq \underline{\sigma}/\sqrt{r}\,\mathrm{Id}_n$ in (4) for future comparisons. Also note that CLaR and SGCL are the same when $r = 1$ and $Y^{(1)} = \bar{Y}$.

The justification for CLaR is the following: if the quadratic loss $\left\|Y - X\mathrm{B}\right\|^2$ were used, the parameters of Model (1) could be estimated by using either $\left\|\bar{Y} - X\mathrm{B}\right\|^2$ or $\frac{1}{r}\sum\left\|Y^{(l)} - X\mathrm{B}\right\|^2$ as a data-fitting term. Yet, both alternatives yield the same solutions as the two terms are equal up to constants. Hence, the quadratic loss does not leverage the multiple repetitions and ignores the noise structure. On the contrary, the more refined data-fitting term of CLaR allows to take into account the individual repetitions, leading to improved performance in applications.

3 Results and properties of CLaR

We start this part by introducing some elements of smoothing theory (Nesterov, 2005; Beck and Teboulle, 2012) that sheds some light on the origin of the data-fitting term introduced earlier.

3.1 Smoothing of the nuclear norm

Let us analyze the data-fitting term of CLaR, by connecting it to the Schatten 1-norm. We derive a formula for the smoothing of the this norm (Proposition 4), which paves the way for a more general smoothing theory for matrix variables (see Appendix A). Let us define the following smoothing function:

$$\omega_{\underline{\sigma}}(\cdot) \triangleq \frac{1}{2}\left(\left\|\cdot\right\|^2 + n\right)\underline{\sigma}\ , \tag{5}$$

and the inf-convolution of functions f_1 and f_2, defined as $f_1 \,\square\, f_2(y) \triangleq \inf_x f_1(x) + f_2(y - x)$.

The next propositions are key to our framework and show the connection between the SGCL, CLaR and the Schatten 1-norm:

[3] See the definition we consider in Eq. (14).

Proposition 4 (Proof in Appendix A.3)**.** The $\omega_{\underline{\sigma}}$-smoothing of the Schatten-1 norm, *i.e.,* the function $\|\cdot\|_{\mathscr{S},1} \square \omega_{\underline{\sigma}} : \mathbb{R}^{n\times q} \mapsto \mathbb{R}$, is the solution of the following smooth optimization problem:

$$(\|\cdot\|_{\mathscr{S},1} \square \omega_{\underline{\sigma}})(Z) = \min_{S \succeq \underline{\sigma}\, \mathrm{Id}_n} \tfrac{1}{2} \|Z\|^2_{S^{-1}} + \tfrac{1}{2} \operatorname{Tr}(S) \ . \tag{6}$$

Definition 5 (Clipped Square Root)**.** For $\Sigma \in \mathcal{S}^n_+$ with spectral decomposition $\Sigma = U \operatorname{diag}(\gamma_1, \dots, \gamma_n) U^\top$ (U is orthogonal), let us define the *Clipped Square Root* operator:

$$\operatorname{ClSqrt}(\Sigma, \underline{\sigma}) = U \operatorname{diag}(\sqrt{\gamma_1} \vee \underline{\sigma}, \dots, \sqrt{\gamma_n} \vee \underline{\sigma}) U^\top \ . \tag{7}$$

Proposition 6 (proof in Appendix B.1)**.** Any solution of the CLaR Problem (2), $(\hat{\mathrm{B}}, \hat{S}) = (\hat{\mathrm{B}}^{\mathrm{CLaR}}, \hat{S}^{\mathrm{CLaR}})$ is also a solution of:

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

$$\hat{S} = \operatorname{ClSqrt}\left(\tfrac{1}{r} Z Z^\top, \underline{\sigma}\right) \ , \text{ where } Z = [Z^{(1)} | \dots | Z^{(r)}] \text{ and } Z^{(l)} = \tfrac{Y^{(l)} - X\mathrm{B}}{\sqrt{q}} \ .$$

Properties similar to Proposition 6 can be traced back to van de Geer (2016, Lemma 3.4, p. 37), where the following was used to prove oracle inequalities for the multivariate square-root Lasso[4]: if $ZZ^\top \succ 0$,

$$\|Z\|_{\mathscr{S},1} = \min_{S\in\mathcal{S}^n_{++}} \frac{1}{2} \|Z\|^2_{S^{-1}} + \frac{1}{2} \operatorname{Tr}(S) \ . \tag{8}$$

In other words Proposition 6 generalizes van de Geer (2016, Lemma 3.4, p. 37) for all matrices Z, getting rid of the condition $ZZ^\top \succ 0$. In the present contribution, the problem formulation in Proposition 4 is motivated by computational aspects, as it helps to address the combined non-differentiability of the data-fitting term $\|\cdot\|_{\mathscr{S},1}$ and the penalty $\|\cdot\|_{2,1}$ term. Other alternatives to exploit the multiple repetitions without simply averaging them, would consist in investigating other Schatten p_1-norms:

$$\operatorname*{arg\,min}_{\mathrm{B}\in\mathbb{R}^{p\times q}} \tfrac{1}{\sqrt{rq}} \|[Y^{(1)} - X\mathrm{B} | \dots | Y^{(r)} - X\mathrm{B}]\|_{\mathscr{S},p_1} + \lambda n \|\mathrm{B}\|_{2,1} \ , \tag{9}$$

Without smoothing, problems of the form given in Eq. (9) have the drawback of having two non-differentiable terms, and calling for primal-dual algorithms (Chambolle and Pock, 2011) with costly proximal operators. Even if the non-smooth Schatten 1-norm is replaced by the formula in (8), numerical challenges remain: S can approach 0 arbitrarily, hence, the gradient *w.r.t.* S of the data-fitting term is not Lipschitz over the optimization domain. A similar problem was raised for the concomitant Lasso by Ndiaye et al. (2017) who used smoothing techniques to address it. Here we replaced the nuclear norm ($p_1 = 1$) by its smoothed version $\|\cdot\|_{\mathscr{S},p_1} \square \omega_{\underline{\sigma}}$. Similar results for the Schatten 2-norm and Schatten ∞-norm are provided in the Appendix (Propositions 21 and 22).

3.2 Algorithmic details: convexity, (block) coordinate descent, parameters influence

We detail the principal results needed to solve Problem (2) numerically, leading to the implementation proposed in Algorithm 1. We first recall useful results for alternate minimization of convex composite problems.

Proposition 7 (Proof in Appendix B.2)**.** CLaR is jointly convex in (B, S). Moreover, f is convex and smooth on the feasible set, and $\|\cdot\|_{2,1}$ is convex and separable in $\mathrm{B}_{j:}$'s, thus minimizing the objective alternatively in S and in $\mathrm{B}_{j:}$'s (see Algorithm 1) converges to a global minimum.

Hence, for our alternate minimization implemenation, we only need to consider solving problems with B or S fixed, which we detail in the next propositions.

Proposition 8 (Minimization in S; proof in Appendix B.3)**.** Let $B \in \mathbb{R}^{n\times q}$ be fixed. The minimization of $f(\mathrm{B}, S)$ *w.r.t.* S with the constraint $S \succeq \underline{\sigma}\, \mathrm{Id}_n$ admits the closed-form solution:

$$S = \operatorname{ClSqrt}\left(\frac{1}{r} \sum_{l=1}^{r} Z^{(l)} Z^{(l)\top}, \underline{\sigma} \right) \ , \quad \text{with } Z^{(l)} = \tfrac{1}{\sqrt{q}} (Y^{(l)} - X\mathrm{B}) \ . \tag{10}$$

[4]Defined as the solution of Equation (9) with $p_1 = 1$.

Algorithm 1 ALTERNATE MINIMIZATION FOR CLAR

input : $X, \bar{Y}, \underline{\sigma}, \lambda, T_{S \text{ update}}, T$
init : $\mathrm{B} = 0_{p,q}$, $S^{-1} = \underline{\sigma}^{-1} \mathrm{Id}_n$, $\bar{R} = \bar{Y}$, $\mathrm{cov}_Y = \frac{1}{r}\sum_{l=1}^{r} Y^{(l)} Y^{(l)\top}$ `// precomputed`
for iter $= 1, \dots, T$ **do**
  if iter $= 1 \pmod{T_{S \text{ update}}}$ **then** `// noise update`
    $RR^\top = \mathrm{RRT}(\mathrm{cov}_Y, Y, X, \mathrm{B})$ `// Eq. (13)`
    $S \leftarrow \mathrm{ClSqrt}(\frac{1}{qr} RR^\top, \underline{\sigma})$ `// Eq. (10)`
    for $j = 1, \dots, p$ **do** $L_j = X_{:j}^\top S^{-1} X_{:j}$
  for $j = 1, \dots, p$ **do** `// coef. update`
    $\bar{R} \leftarrow \bar{R} + X_{:j}\mathrm{B}_{j:}$; $\mathrm{B}_{j:} \leftarrow \mathrm{BST}\left(\frac{X_{:j}^\top S^{-1} \bar{R}}{L_j}, \frac{\lambda nq}{L_j}\right)$; $\bar{R} \leftarrow \bar{R} - X_{:j}\mathrm{B}_{j:}$
return B, S

Proposition 9 (Proof in Appendix B.4)**.** For a fixed $S \in \mathcal{S}_{++}^n$, each step of the block minimization of $f(\cdot, S) + \lambda \|\cdot\|_{2,1}$ in the j^{th} line of B admits a closed-form solution:

$$\mathrm{B}_{j:} = \mathrm{BST}\left(\mathrm{B}_{j:} + \frac{X_{:j}^\top S^{-1}(\bar{Y} - X\mathrm{B})}{\|X_{:j}\|_{S^{-1}}^2}, \frac{\lambda nq}{\|X_{:j}\|_{S^{-1}}^2}\right) \quad . \tag{11}$$

Critical parameter. There exists $\lambda_{\max} \geq 0$ such that whenever $\lambda \geq \lambda_{\max}$, the estimated coefficients vanish. This $\lambda_{\max}$ helps calibrating roughly λ in practice by choosing it as a fraction of $\lambda_{\max}$:

Proposition 10 (Critical regularization parameter; proof in Appendix B.5.)**.** For the CLaR estimator we have: with $S_{\max} \triangleq \mathrm{ClSqrt}\left(\frac{1}{qr}\sum_{l=1}^{r} Y^{(l)} Y^{(l)\top}, \underline{\sigma}\right)$,

$$\forall \lambda \geq \lambda_{\max} \triangleq \frac{1}{nq} \left\| X^\top S_{\max}^{-1} \bar{Y} \right\|_{2,\infty}, \quad \hat{\mathrm{B}}^{\mathrm{CLaR}} = 0 \quad . \tag{12}$$

Convex formulation benefits. Thanks to the convex formulation, convergence of Algorithm 1 can be ensured using the duality gap as a stopping criterion (as it guarantees a targeted sub-optimality level). To compute the duality gap, we derive the dual of Problem (2) in Proposition 24. In addition, convexity allows to leverage acceleration methods such as working sets strategies (Fan and Lv, 2008; Tibshirani et al., 2012; Johnson and Guestrin, 2015; Massias et al., 2018b) or safe screening rules (El Ghaoui et al., 2012; Fercoq et al., 2015) while retaining theoretical guarantees of convergence. Such techniques are trickier to adapt in the non-convex case (see Appendix C), as they could change the local minima reached.

Choice of $\underline{\sigma}$. Although $\underline{\sigma}$ has a smoothing interpretation, from a practical point of view it is an hyperparameter to set. As in Massias et al. (2018a), $\underline{\sigma}$ is always chosen as follows: $\underline{\sigma} = \|Y\| / (1000 \times nq)$. In practice, the experimental results were little affected by the choice of $\underline{\sigma}$.

Remark 11. Once $\mathrm{cov}_Y \triangleq \frac{1}{r}\sum_1^r Y^{(l)} Y^{(l)\top}$ is pre-computed, the cost of updating S does not depend on r, *i.e.,* is the same as working with averaged data. Indeed, with $R = [Y^{(1)} - X\mathrm{B} | \dots | Y^{(r)} - X\mathrm{B}]$, the following computation can be done in $\mathcal{O}(qn^2)$ (details are in Appendix B.7).

$$RR^\top = \mathrm{RRT}(\mathrm{cov}_Y, Y, X, \mathrm{B}) \triangleq r\mathrm{cov}_Y + r(X\mathrm{B})(X\mathrm{B})^\top - r\bar{Y}^\top (X\mathrm{B}) - r(X\mathrm{B})^\top \bar{Y} \quad . \tag{13}$$

Statistical properties showing the advantages of using CLaR (over SGCL) can be found in Appendix B.8. In particular the covariance estimation is improved.

4 Experiments

Our Python code (with Numba compilation Lam et al. 2015) is released as an open source package: `https://github.com/QB3/CLaR`. We compare CLaR to other estimators: SGCL (Massias et al., 2018a), an $\ell_{2,1}$ version of MLE (Chen and Banerjee, 2017; Lee and Liu, 2012) ($\ell_{2,1}$-MLE), a version of the $\ell_{2,1}$-MLE with multiple repetitions ($\ell_{2,1}$-MLER), an $\ell_{2,1}$ penalized version of MRCE (Rothman et al., 2010) with repetitions ($\ell_{2,1}$-MRCER) and the Multi-Task Lasso (MTL, Obozinski et al. 2010). The cost of an epoch of block coordinate descent is summarized in Table 1 in Appendix C.4 for each algorithm[5]. All competitors are detailed in Appendix C.

[5]The cost of computing the duality gap is also provided whenever available.

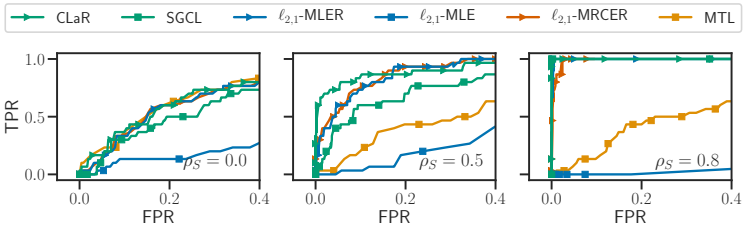


Figure 1 – *Influence of noise structure.* ROC curves of support recovery ($\rho_X = 0.6$, SNR $= 0.03$, $r = 20$) for different ρ_S values.

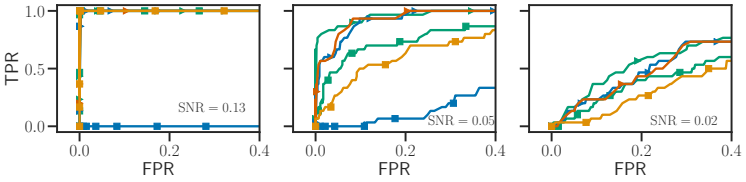


Figure 2 – *Influence of SNR.* ROC curves of support recovery ($\rho_X = 0.6$, $\rho_S = 0.4$, $r = 20$) for different SNR values.

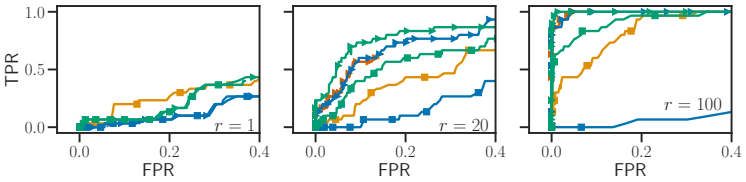


Figure 3 – *Influence of the number of repetitions.* ROC curves of support recovery ($\rho_X = 0.6$, SNR $= 0.03$, $\rho_S = 0.4$) for different r values.

Synthetic data Here we demonstrate the ability of our estimator to recover the support *i.e.,* the ability to identify the predictive features. There are $n = 150$ observations, $p = 500$ features, $q = 100$ tasks. The design X is random with Toeplitz-correlated features with parameter $\rho_X = 0.6$ (correlation between $X_{:i}$ and $X_{:j}$ is $\rho_X^{|i-j|}$), and its columns have unit Euclidean norm. The true coefficient B^* has 30 non-zeros rows whose entries are independent and normally centered distributed. S^* is a Toeplitz matrix with parameter ρ_S. The SNR is fixed and constant across all repetitions

$$\text{SNR} \triangleq \|X\mathrm{B}^*\|/\sqrt{r}\|X\mathrm{B}^* - \bar{Y}\| \ . \tag{14}$$

For Figures 1 to 3, the figure of merit is the ROC curve, *i.e.,* the true positive rate (TPR) against the false positive rate (FPR). For each estimator, the ROC curve is obtained by varying the value of the regularization parameter λ on a geometric grid of 160 points, from $\lambda_{\max}$ (specific to each algorithm) to $\lambda_{\min}$, the latter being estimator specific and chosen to obtain a FPR larger than 0.4.

Influence of noise structure. Figure 1 represents the ROC curves for different values of ρ_S. As ρ_S increases, the noise becomes more and more heteroscedastic. From left to right, the performance of heteroscedastic solvers (CLaR, SGCL, $\ell_{2,1}$-MRCER, $\ell_{2,1}$-MRCE, $\ell_{2,1}$-MLER) increases as they are designed to exploit correlations in the noise, while the performance of MTL decreases, as its homoscedastic model becomes less and less valid.

Influence of SNR. On Figure 2 we can see that when the SNR is high (left), all estimators (except $\ell_{2,1}$-MLE) reach the (0, 1) point. This means that for each algorithm (except $\ell_{2,1}$-MLE), there exists a λ such that the estimated support is exactly the true one. However, when the SNR decreases (middle), the performance of SGCL and MTL starts to drop, while that of CLaR, $\ell_{2,1}$-MLER and $\ell_{2,1}$-MRCER remains stable (CLaR performing better), highlighting their capacity to leverage multiple repetitions of measurements to handle the noise structure. Finally, when the SNR is too low (right), all algorithms perform poorly, but CLaR, $\ell_{2,1}$-MLER and $\ell_{2,1}$-MRCER still performs better.

Influence of the number of repetitions. Figure 3 shows ROC curves of all compared approaches for different r, starting from $r = 1$ (left) to 100 (right). Even with $r = 20$ (middle) CLaR outperforms the other estimators, and when $r = 100$ CLaR can better leverage the large number of repetitions.

Realistic data We now evaluate the estimators on realistic magneto- and electroencephalography (M/EEG) data. The M/EEG recordings measure the electrical potential and magnetic fields induced by the active neurons. Data are time series of length q with n sensors and p sources mapping to locations in the brain. Because the propagation of the electromagnetic fields is driven by the linear Maxwell equations, one can assume that the relation between the measurements $Y^{(1)}, \ldots, Y^{(r)}$ and the amplitudes of sources in the brain B^* is linear.

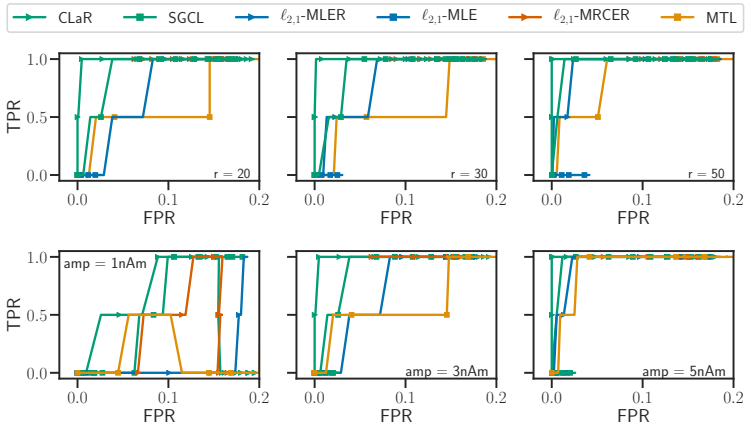


Figure 4 – *Influence of the number of repetitions.* ROC curves with empirical X and S and simulated B^* ($\mathrm{amp} = 2\,\mathrm{nA.m}$), for different number of repetitions.

Figure 5 – *Amplitude influence.* ROC curves with empirical X and S and simulated B^* ($r = 50$), for different amplitudes of the signal.

The M/EEG inverse problem consists in identifying B^*. Because of the limited number of sensors (a few hundreds in practice), as well as the physics of the problem, the M/EEG inverse problem is severely ill-posed and needs to be regularized. Moreover, the experiments being usually short (less than 1 s.) and focused on specific cognitive functions, the number of active sources is expected to be small, *i.e.,* B^* is assumed to be row-sparse. This plausible biological assumption motivates the framework of Section 2 (Ou et al., 2009).

Dataset. We use the *sample* dataset from MNE (Gramfort et al., 2014). The experimental conditions are here auditory stimulations in the right or left ears, leading to two main foci of activations in bilateral auditory cortices (*i.e.,* 2 non-zeros rows for B^*). For this experiment, we keep only the gradiometer magnetic channels. After removing one channel corrupted by artifacts, this leads to $n = 203$ signals. The length of the temporal series is $q = 100$, and the data contains $r = 50$ repetitions. We choose a source space of size $p = 1281$ which corresponds to about 1 cm distance between neighboring sources. The orientation is fixed, and normal to the cortical mantle.

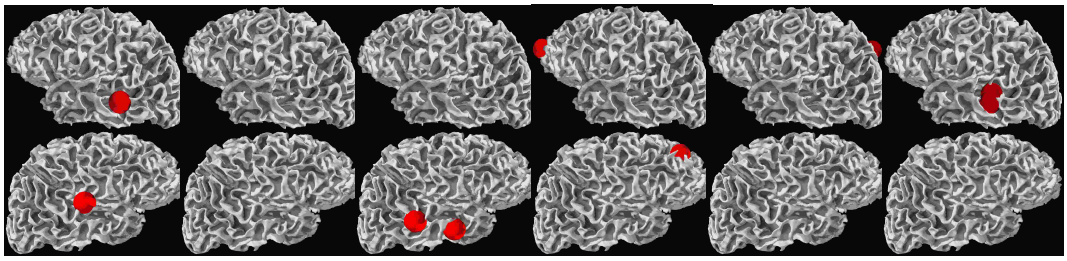
Realistic MEG data simulations. We use here true empirical values for X and S by solving Maxwell equations and taking an empirical co-standard deviation matrix. To generate realistic MEG data we simulate neural responses B^* with 2 non-zeros rows corresponding to areas known to be related to auditory processing (Brodmann area 22). Each non-zero row of B^* is chosen as a sinusoidal signal with realistic frequency (5 Hz) and amplitude ($\mathrm{amp} \sim 1 - 10$ nAm). We finally simulate r MEG signals $Y^{(l)} = X\mathrm{B}^* + S^*\mathrm{E}^{(l)}$, $\mathrm{E}^{(l)}$ being matrices with i.i.d. normal entries.

The signals being contaminated with correlated noise, if one wants to use homoscedastic solvers it is necessary to whiten the data first (and thus to have an estimation of the covariance matrix, the later often being unknown). In this experiment we demonstrate that without this whitening process, the homoscedastic solver MTL fails, as well as solvers which does not take in account the repetitions: SGCL and $\ell_{2,1}$-MLE. In this scenario CLaR, $\ell_{2,1}$-MLER and $\ell_{2,1}$-MRCER do succeed in recovering the sources, CLaR leading to the best results. As for the synthetic data, Figures 4 and 5 are obtained by varying the estimator-specific regularization parameter λ from $\lambda_{\max}$ to $\lambda_{\min}$ on a geometric grid.

Amplitude influence. Figure 5 shows ROC curves for different values of the amplitude of the signal. When the amplitude is high (right), all the algorithms perform well, however when the amplitude decreases (middle) only CLaR leads to good results, almost hitting the $(0, 1)$ corner. When the amplitude gets lower (left) all algorithms perform worse, CLaR still yielding the best results.

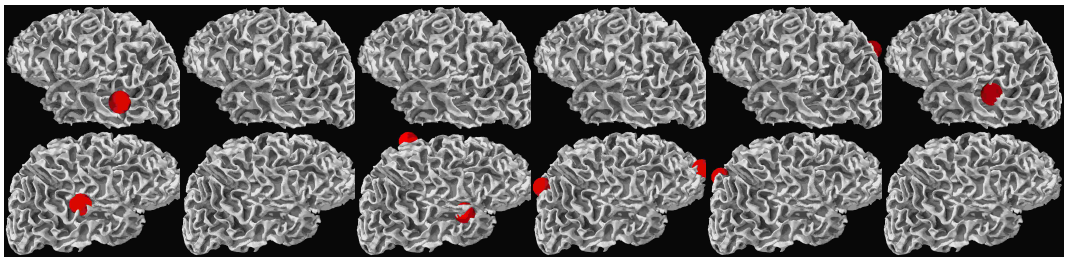
Influence of the number of repetitions. Figure 4 shows ROC curves for different number of repetitions r. When the number of repetitions is high (right, $r = 50$), the algorithms taking into account all the repetitions (CLaR, $\ell_{2,1}$-MLER, $\ell_{2,1}$-MRCER) perform best, almost hitting the $(0, 1)$ corner, whereas the algorithms which do not take into account all the repetitions ($\ell_{2,1}$-MLE, MTL, SGCL) perform poorly. As soon as the number of repetitions decreases (middle and left) the performances of all the algorithms except CLaR start dropping severely. CLaR is once again the algorithm taking the most advantage of the number of repetitions.

Real data As before, we use the *sample* dataset from MNE, keeping only the magnetometer magnetic channels ($n = 102$ signals). We choose a source space of size $p = 7498$ (about 5 mm



(a) CLaR (b) SGCL (c) $\ell_{2,1}$-MLER (d) $\ell_{2,1}$-MLE (e) $\ell_{2,1}$-MRCER (f) MTL

Figure 6 – *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 .



(a) CLaR (b) SGCL (c) $\ell_{2,1}$-MLER (d) $\ell_{2,1}$-MLE (e) $\ell_{2,1}$-MRCER (f) MTL

Figure 7 – *Real data, right auditory stimulations* ($n = 102$, $q = 7498$, $q = 76$, $r = 33$) Sources found in the left hemisphere (top) and the right hemisphere (bottom) after right auditory stimulations.

between neighboring sources). The orientation is fixed, and normal to the cortical mantle. As for realistic data, X is the empirical design matrix, but this time we use the empirical measurements $Y^{(1)}, \dots, Y^{(r)}$. The experiment are left or right auditory stimulations, extensive results for right auditory stimulations (*resp.* visual stimulations) can be found in Appendix D.3 (*resp.* Appendix D.4 and D.5). As two sources are expected (one in each hemisphere, in bilateral auditory cortices), we vary λ by dichotomy between $\lambda_{\max}$ (returning 0 sources) and a $\lambda_{\min}$ (returning more than 2 sources), until finding a λ giving exactly 2 sources. Results are provided in Figures 6 and 7. Running times of each algorithm are of the same order of magnitude and can be found in Appendix D.2.

Comments on Figure 6, left auditory stimulations. Sources found by the algorithms are represented by red spheres. SGCL, $\ell_{2,1}$-MLE and $\ell_{2,1}$-MRCER completely fail, finding sources that are not in the auditory cortices at all (SGCL sources are deep, thus not in the auditory cortices, and cannot be seen). MTL and $\ell_{2,1}$-MLER do find sources in auditory cortices, but only in one hemisphere (left for MTL and right for $\ell_{2,1}$-MLER). CLaR is the only one that finds one source in each hemisphere in the auditory cortices as expected.

Comments on Figure 7, right auditory stimulations. In this experiment we only keep $r = 33$ repetitions (out of 65 available) and it can be seen that only CLaR finds correct sources, MTL finds sources only in one hemisphere and all the other algorithms do find sources that are not in the auditory cortices. This highlights the robustness of CLaR, even with a limited number of repetitions, confirming previous experiments (see Figure 3).

Conclusion This work introduces CLaR, a sparse estimator for multitask regression. It is designed to handle heteroscedastic noise in the context of repeated observations, a standard framework in applied sciences such as neuroimaging. The resulting optimization problem can be solved efficiently with state-of-the-art convex solvers, and the algorithmic cost is the same as for single repetition data. The theory of smoothing connects CLaR to the Schatten 1-Lasso in a principled manner, which opens the way to the use of more sophisticated datafitting terms. The benefits of CLaR for support recovery in heteroscedastic context were extensively evaluated against a large number of competitors, both on simulations and on empirical MEG data.

Acknowledgments This work was funded by ERC Starting Grant SLAB ERC-YStG-676943.

References

A. Argyriou, T. Evgeniou, and M. Pontil. Convex multi-task feature learning. *Machine Learning*, 73 (3):243–272, 2008.

F. Bach, R. Jenatton, J. Mairal, and G. Obozinski. Convex optimization with sparsity-inducing norms. *Foundations and Trends in Machine Learning*, 4(1):1–106, 2012.

A. Beck. *First-Order Methods in Optimization*, volume 25. SIAM, 2017.

A. Beck and M. Teboulle. Smoothing and first order methods: A unified framework. *SIAM J. Optim.*, 22(2):557–580, 2012.

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

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

S. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge University Press, 2004.

A. Chambolle and T. Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *J. Math. Imaging Vis.*, 40(1):120–145, 2011.

S. Chen and A. Banerjee. Alternating estimation for structured high-dimensional multi-response models. In *NIPS*, pages 2838–2848, 2017.

A. S. Dalalyan, M. Hebiri, K. Meziani, and J. Salmon. Learning heteroscedastic models by convex programming under group sparsity. In *ICML*, 2013.

J. Daye, J. Chen, and H. Li. High-dimensional heteroscedastic regression with an application to eQTL data analysis. *Biometrics*, 68(1):316–326, 2012.

L. El Ghaoui, V. Viallon, and T. Rabbani. Safe feature elimination in sparse supervised learning. *J. Pacific Optim.*, 8(4):667–698, 2012.

J. Fan and J. Lv. Sure independence screening for ultrahigh dimensional feature space. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 70(5):849–911, 2008.

O. Fercoq, A. Gramfort, and J. Salmon. Mind the duality gap: safer rules for the lasso. In *ICML*, pages 333–342, 2015.

J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

A. Gramfort, D. Strohmeier, J. Haueisen, M. S. Hämäläinen, and M. Kowalski. Time-frequency mixed-norm estimates: Sparse M/EEG imaging with non-stationary source activations. *NeuroImage*, 70: 410–422, 2013.

A. Gramfort, M. Luessi, E. Larson, D. A. Engemann, D. Strohmeier, C. Brodbeck, L. Parkkonen, and M. S. Hämäläinen. MNE software for processing MEG and EEG data. *NeuroImage*, 86:446–460, 2014.

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

P. J. Huber and R. Dutter. Numerical solution of robust regression problems. In *Compstat 1974 (Proc. Sympos. Computational Statist., Univ. Vienna, Vienna, 1974)*, pages 165–172. Physica Verlag, Vienna, 1974.

T. B. Johnson and C. Guestrin. Blitz: A principled meta-algorithm for scaling sparse optimization. In *ICML*, pages 1171–1179, 2015.

M. Kolar and J. Sharpnack. Variance function estimation in high-dimensions. In *ICML*, pages 1447–1454, 2012.

S. K. Lam, A. Pitrou, and S. Seibert. Numba: A LLVM-based Python JIT Compiler. In *Proceedings of the Second Workshop on the LLVM Compiler Infrastructure in HPC*, pages 1–6. ACM, 2015.

W. Lee and Y. Liu. Simultaneous multiple response regression and inverse covariance matrix estimation via penalized Gaussian maximum likelihood. *Journal of multivariate analysis*, 111: 241–255, 2012.

M. Massias, O. Fercoq, A. Gramfort, and J. Salmon. Generalized concomitant multi-task lasso for sparse multimodal regression. In *AISTATS*, volume 84, pages 998–1007, 2018a.

M. Massias, A. Gramfort, and J. Salmon. Celer: a Fast Solver for the Lasso with Dual Extrapolation. In *ICML*, 2018b.

E. Ndiaye, O. Fercoq, A. Gramfort, V. Leclère, and J. Salmon. Efficient smoothed concomitant lasso estimation for high dimensional regression. *Journal of Physics: Conference Series*, 904(1), 2017.

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

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.

W. Ou, M. Hämaläinen, and P. Golland. A distributed spatio-temporal EEG/MEG inverse solver. *NeuroImage*, 44(3):932–946, Feb 2009.

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

N. Parikh, S. Boyd, E. Chu, B. Peleato, and J. Eckstein. Proximal algorithms. *Foundations and Trends in Machine Learning*, 1(3):1–108, 2013.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.*, 12:2825–2830, 2011.

P. Rai, A. Kumar, and H. Daume. Simultaneously leveraging output and task structures for multiple-output regression. In *NIPS*, pages 3185–3193, 2012.

A. J. Rothman, E. Levina, and J. Zhu. Sparse multivariate regression with covariance estimation. *Journal of Computational and Graphical Statistics*, 19(4):947–962, 2010.

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

R. Tibshirani. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 58(1):267–288, 1996.

R. Tibshirani, J. Bien, J. Friedman, T. J. Hastie, N. Simon, J. Taylor, and R. J. Tibshirani. Strong rules for discarding predictors in lasso-type problems. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 74(2): 245–266, 2012.

P. Tseng. Convergence of a block coordinate descent method for nondifferentiable minimization. *J. Optim. Theory Appl.*, 109(3):475–494, 2001.

P. Tseng and S. Yun. Block-coordinate gradient descent method for linearly constrained nonsmooth separable optimization. *J. Optim. Theory Appl.*, 140(3):513, 2009.

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

J. Wagener and H. Dette. Bridge estimators and the adaptive Lasso under heteroscedasticity. *Math. Methods Statist.*, 21:109–126, 2012.

A Smoothing theory for convex optimization

We start this section by introducing some additional useful notation, in particular the Fenchel conjugate[6].

Notation Let $d \in \mathbb{N}$, and let $\mathcal{C}$ be a subset of $\mathbb{R}^d$. We write $\iota_{\mathcal{C}}$ for the indicator function of the set $\mathcal{C}$, *i.e.,* $\iota_{\mathcal{C}}(x) = 0$ if $x \in \mathcal{C}$ and $\iota_{\mathcal{C}}(x) = +\infty$ otherwise, and $\Pi_{\mathcal{C}}$ for the projection on the (closed and convex) set $\mathcal{C}$. The Fenchel conjugate of a function $h : \mathbb{R}^d \mapsto \mathbb{R}$ is written h^* and is defined for any $y \in \mathbb{R}^d$, by $h^*(y) = \sup_{x \in \mathbb{R}^d} \langle x, y \rangle - h(x)$.

For $p_1 \in [1, \infty)$, let us write $\mathcal{B}_{\mathscr{S},p_1}$ for the Schatten-p_1 unit ball, and $\|\cdot\|_{p_1}$ for the standard ℓ_{p_1}-norm in $\mathbb{R}^d$.

A.1 Basic properties of inf-convolution

Proposition 12. Let $h : \mathbb{R}^d \to \mathbb{R}$ be a closed proper convex function and let $\omega : \mathbb{R}^d \to \mathbb{R}$ be a convex function with Lipschitz gradient. Let $\omega_{\underline{\sigma}} \triangleq \underline{\sigma}\omega\left(\frac{\cdot}{\underline{\sigma}}\right)$. Then, the following holds (see Parikh et al. 2013, p. 136):

$$h^{**} = h \ , \tag{15}$$
$$(h \,\square\, \omega_{\underline{\sigma}})^* = h^* + \omega_{\underline{\sigma}}^* \ , \tag{16}$$
$$\omega_{\underline{\sigma}}^* = \underline{\sigma}\omega^* \ , \tag{17}$$
$$\|\cdot\|_{p_1}^* = \iota_{\mathcal{B}_{p_1^*}} , \text{ where } \frac{1}{p_1} + \frac{1}{p_1^*} = 1 \ , \tag{18}$$
$$(h + \delta)^* = h^* - \delta, \quad \forall \delta \in \mathbb{R}^d \ , \tag{19}$$
$$\left(\tfrac{1}{2}\|\cdot\|^2\right)^* = \tfrac{1}{2}\|\cdot\|^2 \ . \tag{20}$$

From Equations (17), (19) and (20) it follows that

$$\omega(\cdot) = \tfrac{1}{2}\|\cdot\|^2 + \tfrac{1}{2} \implies \omega_{\underline{\sigma}}^* = \tfrac{\sigma}{2}\|\cdot\|^2 - \tfrac{\sigma}{2} \ . \tag{21}$$

A.2 Smoothing of Schatten norms

In all this section, the variable is a matrix $Z \in \mathbb{R}^{n \times q}$.

Lemma 13. Let $c \in \mathbb{R}, p_1 \in [1, \infty]$. Let $p_1^* \in [1, \infty]$ be the Hölder conjugate of p_1, $\frac{1}{p_1} + \frac{1}{p_1^*} = 1$. For the choice $\omega(\cdot) = \frac{1}{2}\|\cdot\|^2 + c$, the following holds true:

$$\left(\|\cdot\|_{\mathscr{S},p_1} \,\square\, \omega_{\underline{\sigma}}\right)(Z) = \frac{1}{2\underline{\sigma}}\|Z\|^2 + c\underline{\sigma} - \frac{\sigma}{2}\left\|\Pi_{\mathcal{B}_{\mathscr{S},p_1^*}}\left(\frac{Z}{\underline{\sigma}}\right) - \frac{Z}{\underline{\sigma}}\right\|^2 \ .$$

Proof.

$$\begin{aligned}
\left(\|\cdot\|_{\mathscr{S},p_1} \,\square\, \omega_{\underline{\sigma}}\right)(Z) &= \left(\|\cdot\|_{\mathscr{S},p_1} \,\square\, \omega_{\underline{\sigma}}\right)^{**}(Z) && \text{(using Eq. (15))} \\
&= \left(\|\cdot\|_{\mathscr{S},p_1}^* + \omega_{\underline{\sigma}}^*\right)^*(Z) && \text{(using Eq. (16))} \\
&= \left(\iota_{\mathcal{B}_{\mathscr{S},p_1^*}} + \tfrac{\sigma}{2}\|\cdot\|^2 - c\underline{\sigma}\right)^*(Z) && \text{(using Eqs. (18) and (21))} \\
&= \left(\tfrac{\sigma}{2}\|\cdot\|^2 + \iota_{\mathcal{B}_{\mathscr{S},p_1^*}}\right)^*(Z) + c\underline{\sigma} && \text{(using Eq. (19))} \ .
\end{aligned} \tag{22}$$

[6] Sometimes this is also referred to as the Fenchel transform.

We can now compute the last Fenchel transform remaining:

$$\begin{aligned}
\left(\frac{\sigma}{2}\left\|\cdot\right\|^2 + \iota_{\mathcal{B}_{\mathscr{S},p_1^*}}\right)^*(Z) &= \sup_{U\in\mathbb{R}^{n\times q}} \left(\langle U, Z\rangle - \frac{\sigma}{2}\left\|U\right\|^2 - \iota_{\mathcal{B}_{\mathscr{S},p_1^*}}(U)\right) \\
&= \sup_{U\in\mathcal{B}_{\mathscr{S},p_1^*}} \left(\langle U, Z\rangle - \frac{\sigma}{2}\left\|U\right\|^2\right) \\
&= -\inf_{U\in\mathcal{B}_{\mathscr{S},p_1^*}} \left(\frac{\sigma}{2}\left\|U\right\|^2 - \langle U, Z\rangle\right) \\
&= -\underline{\sigma}\cdot\inf_{U\in\mathcal{B}_{\mathscr{S},p_1^*}} \left(\tfrac{1}{2}\left\|U\right\|^2 - \left\langle U, \tfrac{Z}{\underline{\sigma}}\right\rangle\right) \\
&= -\underline{\sigma}\cdot\inf_{U\in\mathcal{B}_{\mathscr{S},p_1^*}} \left(\tfrac{1}{2}\left\|U - \tfrac{Z}{\underline{\sigma}}\right\|^2 - \tfrac{1}{2\underline{\sigma}^2}\left\|Z\right\|^2\right) \\
&= \tfrac{1}{2\underline{\sigma}}\left\|Z\right\|^2 - \tfrac{\sigma}{2}\cdot\inf_{U\in\mathcal{B}_{\mathscr{S},p_1^*}} \left(\left\|U - \tfrac{Z}{\underline{\sigma}}\right\|^2\right) \\
&= \tfrac{1}{2\underline{\sigma}}\left\|Z\right\|^2 - \tfrac{\sigma}{2}\left\|\Pi_{\mathcal{B}_{\mathscr{S},p_1^*}}\left(\tfrac{Z}{\underline{\sigma}}\right) - \tfrac{Z}{\underline{\sigma}}\right\|^2 . \qquad (23)
\end{aligned}$$

The result follows by combining Eqs. (22) and (23).

□

A.3 Schatten 1-norm (nuclear/trace norm), proof of Proposition 4

Let us first recall/prove some preliminary lemmas.

A.3.1 Preliminary lemmas

First we need the formula of the projection of a matrix onto the Schatten infinity ball:

Lemma 14 (Projection onto $\mathcal{B}_{\mathscr{S},\infty}$, Beck 2017, Example 7.31, p. 194)**.** Let $Z \in \mathbb{R}^{n\times q}$, let $Z = V\,\mathrm{diag}(\gamma_1,\ldots,\gamma_{n\wedge q})W^T$ be the singular value decomposition of Z, then:

$$\Pi_{\|\cdot\|_{\mathscr{S},\infty}}(Z) = V\,\mathrm{diag}(\gamma_1\wedge 1,\ldots,\gamma_{n\wedge q}\wedge 1)W^T \ . \qquad (24)$$

Then we need to link the value of the primal to the singular values of $ZZ^\top$:

Lemma 15 (Value of the primal)**.** Let $\gamma_1,\ldots,\gamma_{n\wedge q}$ be the singular value decomposition of Z, then:

i)

$$\min_{S\succeq\underline{\sigma}\,\mathrm{Id}_n} \frac{1}{2}\,\mathrm{Tr}[Z^\top S^{-1}Z] + \frac{1}{2}\,\mathrm{Tr}(S) = \frac{1}{2}\sum_{i=1}^{n\wedge q}\frac{\gamma_i^2}{\gamma_i\vee\underline{\sigma}} + \frac{1}{2}\sum_{i=1}^{n\wedge q}\gamma_i\vee\underline{\sigma} + \frac{1}{2}(n-n\wedge q)\underline{\sigma} \ , \qquad (25)$$

ii)

$$\min_{S\succeq\underline{\sigma}\,\mathrm{Id}_q} \frac{1}{2}\,\mathrm{Tr}[ZS^{-1}Z^\top] + \frac{1}{2}\,\mathrm{Tr}(S) = \frac{1}{2}\sum_{i=1}^{n\wedge q}\frac{\gamma_i^2}{\gamma_i\vee\underline{\sigma}} + \frac{1}{2}\sum_{i=1}^{n\wedge q}\gamma_i\vee\underline{\sigma} + \frac{1}{2}(q-n\wedge q)\underline{\sigma} \ . \qquad (26)$$

Proof of Lemma 15 i). The minimum in the left hand side is attained in $\hat{S} = U \operatorname{diag}(\gamma_1 \vee \underline{\sigma}, \ldots, \gamma_{n\wedge q} \vee \underline{\sigma}, \underline{\sigma}, \ldots, \underline{\sigma}) U^\top$ (see Massias et al. 2018a, see Prop.2, p.3).

$$
\begin{aligned}
\min_{S \succeq \underline{\sigma} \operatorname{Id}_n} \frac{1}{2} \operatorname{Tr}[Z^\top S^{-1} Z] + \frac{1}{2} \operatorname{Tr}(S) &= \frac{1}{2} \operatorname{Tr}[Z^\top \hat{S}^{-1} Z] + \frac{1}{2} \operatorname{Tr}(\hat{S}) \quad \text{, with} \\
&= \frac{1}{2} \operatorname{Tr}[\hat{S}^{-1} Z Z^\top] + \frac{1}{2} \operatorname{Tr}(\hat{S}) \quad \text{, with } \hat{S} = U \operatorname{diag}(\gamma_1 \vee \underline{\sigma}, \ldots, \gamma_{n\wedge q} \vee \underline{\sigma}, \underline{\sigma}, \ldots, \underline{\sigma}) U^\top \\
&= \frac{1}{2} \operatorname{Tr}[U \operatorname{diag}(\gamma_1^2/\gamma_1 \vee \underline{\sigma}, \ldots, \gamma_{n\wedge q}^2/\gamma_{n\wedge q} \vee \underline{\sigma}, 0, \ldots, 0) U^\top] \\
&+ \frac{1}{2} \operatorname{Tr}[U \operatorname{diag}(\gamma_1 \vee \underline{\sigma}, \ldots, \gamma_{n\wedge q} \vee \underline{\sigma}, \underline{\sigma}, \ldots, \underline{\sigma}) U^\top] \\
&= \frac{1}{2} \sum_{i=1}^{n\wedge q} \frac{\gamma_i^2}{\gamma_i \vee \underline{\sigma}} + \frac{1}{2} \sum_{i=1}^{n\wedge q} \gamma_i \vee \underline{\sigma} + \frac{1}{2} \sum_{n\wedge q+1}^{n} \underline{\sigma} \\
&= \frac{1}{2} \sum_{i=1}^{n\wedge q} \frac{\gamma_i^2}{\gamma_i \vee \underline{\sigma}} + \frac{1}{2} \sum_{i=1}^{n\wedge q} \gamma_i \vee \underline{\sigma} + \frac{1}{2} (n - n \wedge q) \underline{\sigma} \ . \qquad (27)
\end{aligned}
$$

This completes the proof of Lemma 15 *i)*. Equation (26) is obtained by symmetry. □

A.3.2 Main result: an explicit variational formula for the inf-convolution smoothing of the nuclear norm

We now recall the main result that we claim to prove:

Proposition 4 (Proof in Appendix A.3)**.** The $\omega_{\underline{\sigma}}$-smoothing of the Schatten-1 norm, *i.e.,* the function $\|\cdot\|_{\mathscr{S},1} \square \omega_{\underline{\sigma}} : \mathbb{R}^{n\times q} \mapsto \mathbb{R}$, is the solution of the following smooth optimization problem:

$$
(\|\cdot\|_{\mathscr{S},1} \square \omega_{\underline{\sigma}})(Z) = \min_{S \succeq \underline{\sigma} \operatorname{Id}_n} \tfrac{1}{2} \|Z\|_{S^{-1}}^2 + \tfrac{1}{2} \operatorname{Tr}(S) \ . \qquad (6)
$$

Proof. Let $V \operatorname{diag}(\gamma_1, \ldots, \gamma_{n\wedge q}) W^\top$ be the singular values decomposition of Z. We remind that $\Pi_{\mathcal{B}_{\mathscr{S},\infty}}$, the projection over $\mathcal{B}_{\mathscr{S},\infty}$, is given by (see Beck 2017, Example 7.31, p. 194):

$$
\begin{aligned}
\Pi_{\mathcal{B}_{\mathscr{S},\infty}} \left(\frac{Z}{\underline{\sigma}} \right) &= V \operatorname{diag} \left(\Pi_{\mathcal{B}_\infty} \left(\tfrac{\gamma_1}{\underline{\sigma}}, \ldots, \tfrac{\gamma_{n\wedge q}}{\underline{\sigma}} \right) \right) W^\top \\
&= V \operatorname{diag} \left(\tfrac{\gamma_1}{\underline{\sigma}} \wedge 1, \ldots, \tfrac{\gamma_{n\wedge q}}{\underline{\sigma}} \wedge 1 \right) W^\top \ , \qquad (28)
\end{aligned}
$$

where we used that the (vectorial) projection over $\mathcal{B}_\infty$ is given coordinate-wise by $\left(\Pi_{\mathcal{B}_\infty}(\gamma_i)\right)_i = (\gamma_i \wedge 1)_i$. Then we have,

$$\begin{aligned}
\left\|\Pi_{\mathcal{B}_{\mathscr{S},\infty}}\left(\tfrac{Z}{\underline{\sigma}}\right) - \tfrac{Z}{\underline{\sigma}}\right\|^2 &\overset{Eq.\ (28)}{=} \left\|V \operatorname{diag}\left(\tfrac{\gamma_1}{\underline{\sigma}} \wedge 1 - \tfrac{\gamma_1}{\underline{\sigma}}, \dots, \tfrac{\gamma_{n\wedge q}}{\underline{\sigma}} \wedge 1 - \tfrac{\gamma_{n\wedge q}}{\underline{\sigma}}\right) W^\top\right\|^2 \\
&= \sum_{i=1}^{n\wedge q}\left(\tfrac{\gamma_i}{\underline{\sigma}} \wedge 1 - \tfrac{\gamma_i}{\underline{\sigma}}\right)^2 \\
&= \frac{1}{\underline{\sigma}^2}\sum_{i=1}^{n\wedge q}(\gamma_i \wedge \underline{\sigma} - \gamma_i)^2 \\
&= \frac{1}{\underline{\sigma}^2}\sum_{\gamma_i>\underline{\sigma}}(\gamma_i \wedge \underline{\sigma} - \gamma_i)^2 \\
&= \frac{1}{\underline{\sigma}^2}\sum_{\gamma_i>\underline{\sigma}}(\underline{\sigma} - \gamma_i)^2 \\
&= \frac{1}{\underline{\sigma}^2}\sum_{\gamma_i>\underline{\sigma}}\left(\underline{\sigma}^2 + \gamma_i^2 - 2\underline{\sigma}\gamma_i\right) \\
&= \sum_{\gamma_i>\underline{\sigma}} 1 + \frac{1}{\underline{\sigma}^2}\sum_{\gamma_i>\underline{\sigma}}\gamma_i^2 - 2\frac{1}{\underline{\sigma}}\sum_{\gamma_i>\underline{\sigma}}\gamma_i \qquad \text{leading to} \\
-\frac{\underline{\sigma}}{2}\left\|\Pi_{\mathcal{B}_{\mathscr{S},\infty}}\left(\tfrac{Z}{\underline{\sigma}}\right) - \tfrac{Z}{\underline{\sigma}}\right\|^2 &= -\frac{\underline{\sigma}}{2}\sum_{\gamma_i>\underline{\sigma}} 1 - \frac{1}{2\underline{\sigma}}\sum_{\gamma_i>\underline{\sigma}}\gamma_i^2 + \sum_{\gamma_i>\underline{\sigma}}\gamma_i \ . \qquad (29)
\end{aligned}$$

By combining Lemma 13 and Eq. (29) with $p_1^* = \infty, c \in \mathbb{R}$, it follows:

$$\begin{aligned}
\left(\|\cdot\|_{\mathscr{S},1} \square\, \omega_{\underline{\sigma}}\right)(Z) &= \frac{1}{2\underline{\sigma}}\sum_{i=1}^{n}\gamma_i^2 + c\underline{\sigma} - \frac{\underline{\sigma}}{2}\sum_{\gamma_i>\underline{\sigma}} 1 - \frac{1}{2\underline{\sigma}}\sum_{\gamma_i>\underline{\sigma}}\gamma_i^2 + \sum_{\gamma_i>\underline{\sigma}}\gamma_i \\
&= \frac{1}{2\underline{\sigma}}\sum_{\gamma_i\le\underline{\sigma}}^{n}\gamma_i^2 + c\underline{\sigma} - \frac{\underline{\sigma}}{2}\sum_{\gamma_i>\underline{\sigma}} 1 + \sum_{\gamma_i>\underline{\sigma}}\gamma_i \qquad \text{by grouping the } \gamma_i \text{ terms} \\
&= \frac{1}{2\underline{\sigma}}\sum_{\gamma_i\le\underline{\sigma}}^{n}\gamma_i^2 + \sum_{\gamma_i>\underline{\sigma}}\gamma_i - \frac{\underline{\sigma}}{2}\sum_{\gamma_i>\underline{\sigma}} 1 + c\underline{\sigma} \qquad \text{by reordering.} \qquad (30)
\end{aligned}$$

The goal is now to link the optimization problem to the right-hand side of Equation (30). Let $ZZ^\top = U^T \operatorname{diag}(\underbrace{\gamma_1, \dots, \gamma_{n\wedge q}, 0, \dots, 0}_{\in\mathbb{R}^n})U$ be the eigenvalues decomposition of $ZZ^\top$.

$$\begin{aligned}
\min_{S\succeq\underline{\sigma}\mathrm{Id}_n}\frac{1}{2}\operatorname{Tr}[Z^\top S^{-1}Z] + \frac{1}{2}\operatorname{Tr}(S) &= \frac{1}{2}\sum_{i=1}^{n\wedge q}\frac{\gamma_i^2}{\gamma_i \vee \underline{\sigma}} + \frac{1}{2}\sum_{i=1}^{n\wedge q}\gamma_i \vee \underline{\sigma} + \frac{1}{2}(n - n\wedge q)\underline{\sigma} \quad \text{(using Lemma 15)} \\
&= \frac{1}{2\underline{\sigma}}\sum_{\gamma_i\le\underline{\sigma}}\gamma_i^2 + \frac{1}{2}\sum_{\gamma_i>\underline{\sigma}}\gamma_i + \frac{1}{2}(n - n\wedge q)\underline{\sigma} + \frac{1}{2}\sum_{\gamma_i>\underline{\sigma}}\gamma_i + (n - n\wedge q)\underline{\sigma} \\
&= \frac{1}{2\underline{\sigma}}\sum_{\gamma_i\le\underline{\sigma}}\gamma_i^2 + \sum_{\gamma_i>\underline{\sigma}}\gamma_i + \frac{\underline{\sigma}}{2}\sum_{\gamma_i\le\underline{\sigma}} 1 + \frac{1}{2}(n - n\wedge q)\underline{\sigma} \\
&= \frac{1}{2\underline{\sigma}}\sum_{\gamma_i\le\underline{\sigma}}\gamma_i^2 + \sum_{\gamma_i>\underline{\sigma}}\gamma_i + \frac{\underline{\sigma}}{2}(n\wedge q - \sum_{\gamma_i>\underline{\sigma}} 1) + \frac{1}{2}(n - n\wedge q)\underline{\sigma} \\
&= \frac{1}{2\underline{\sigma}}\sum_{\gamma_i\le\underline{\sigma}}\gamma_i^2 + \sum_{\gamma_i>\underline{\sigma}}\gamma_i - \frac{\underline{\sigma}}{2}\sum_{\gamma_i>\underline{\sigma}} 1 + \underbrace{\frac{\underline{\sigma}}{2}n\wedge q + \frac{1}{2}(n - n\wedge q)\underline{\sigma}}_{\frac{\underline{\sigma}}{2}n} \\
&= \frac{1}{2\underline{\sigma}}\sum_{\gamma_i\le\underline{\sigma}}\gamma_i^2 + \sum_{\gamma_i>\underline{\sigma}}\gamma_i - \frac{\underline{\sigma}}{2}\sum_{\gamma_i>\underline{\sigma}} 1 + \frac{\underline{\sigma}}{2}n \ , \qquad (31)
\end{aligned}$$

and identifying Equations (30) and (31) leads to the result for $c = \frac{n}{2}$. □

A.4 Properties of the proposed smoothing for the nuclear norm

First let us recall the defintion of smoothable function and μ-smooth approximation of Beck and Teboulle (2012, Def. 2.1, p.560):

Definition 16 (Smoothable functions, μ-smooth approximation)**.** Let $g : \mathbb{E} \leftarrow]-\infty, +\infty]$ be a closed and proper convex function, and let $E \subseteq \mathrm{dom}\,(g)$ be a close convex set. The function g is called (α, δ, K)-*smoothable* on E if there exists δ_1, δ_2 satisfying $\delta_1 + \delta_2 = \delta > 0$ such that for every μ there exists a continuously differentiable convex function $g_\mu : \mathbb{E} \leftarrow]-\infty, +\infty[$ such that the following hold:

i) $g(x) - \delta_1\mu \leq g_\mu(x) \leq g(x) + \delta_2\mu$ for every $x \in E$.

ii) The function g_μ has a Lipschitz constant which is less than or equal to $K + \frac{\alpha}{\mu}$, *i.e.,* that there exists $K \leq 0, \alpha > 0$, such that:

$$\|\nabla g_\mu(x) - \nabla g_\mu(y)\|^* \leq \left(K + \frac{\alpha}{\mu}\right) \|x - y\| \text{ for every } x, y \in E \ . \tag{32}$$

The function g is called a μ-*smooth* approximation of g with parameters (α, δ, K).

The nuclear norm $\|\cdot\|_{\mathscr{S},1}$ is non-smooth in 0. One can construct a smooth approximation of the nuclear norm based on the following variational formula, if $ZZ^\top \succ 0$:

$$\|Z\|_{\mathscr{S},1} = \min_{S \succ 0} \frac{1}{2} \mathrm{Tr}[Z^\top S^{-1} Z] + \frac{1}{2} \mathrm{Tr}(S) \ , \tag{33}$$

see van de Geer (2016, Lemma 3.4, p. 37). When $ZZ^\top \not\succ 0$, one can approximate $\|\cdot\|_{\mathscr{S},1}$ with

$$\min_{S \succeq \underline{\sigma}\,\mathrm{Id}} \frac{1}{2} \mathrm{Tr}[Z^\top S^{-1} Z] + \frac{1}{2} \mathrm{Tr}(S) = \|\cdot\|_{\mathscr{S},1} \square\, \omega_{\underline{\sigma}} \ , \tag{34}$$

as shown in Appendix A.3. It can be shown that this approximation of the nuclear norm is close to nuclear norm. For formally, with Beck and Teboulle (2012, Def. 2.1, p.560) definition of μ-smooth approximation one can prove that:

Proposition 17. $\|\cdot\|_{\mathscr{S},1} \square\, \omega_{\underline{\sigma}}$ is a $\underline{\sigma}$-smooth approximation of $\|\cdot\|_{\mathscr{S},1}$ with parameters $(1, \frac{n}{2}, 0)$. More precisely: $\|\cdot\|_{\mathscr{S},1} \square\, \omega_{\underline{\sigma}}$ has a gradient $\underline{\sigma}$-Lipschitz and

$$0 \leq \|\cdot\|_{\mathscr{S},1} \square\, \omega_{\underline{\sigma}} - \|\cdot\|_{\mathscr{S},1} = \frac{\underline{\sigma}}{2} \sum_{\gamma_i \leq \underline{\sigma}} \left(1 - \frac{\gamma_i}{\underline{\sigma}}\right)^2 \leq \frac{\underline{\sigma}}{2} n \ . \tag{35}$$

Proof. Since ω is 1-smooth, Beck and Teboulle (2012, Thm. 4.1, p. 567) shows that $\|\cdot\|_{\mathscr{S},1} \square\, \omega_{\underline{\sigma}}$ is $\underline{\sigma}$-smooth.

Let $Z \in \mathbb{R}^{n \times q}$ and let $\gamma_1, \ldots, \gamma_{n \wedge q}$ be its singular value decomposition:

$$\begin{aligned}
\left(\|\cdot\|_{\mathscr{S},1} \square\, \omega_{\underline{\sigma}}\right)(Z) - \|Z\|_{\mathscr{S},1} &= \frac{1}{2} \sum_{i=1}^{n \wedge q} \frac{\gamma_i^2}{\gamma_i \vee \underline{\sigma}} + \frac{1}{2} \sum_{i=1}^{n \wedge q} \gamma_i \vee \underline{\sigma} + \frac{1}{2} \sum_{n \wedge q + 1}^{n} \underline{\sigma} - \sum_{i=1}^{n \wedge q} \gamma_i \\
&= \frac{1}{2} \sum_{i=1}^{n \wedge q} \left(\frac{\gamma_i^2}{\gamma_i \vee \underline{\sigma}} + \gamma_i \vee \underline{\sigma} - 2\gamma_i\right) + \frac{1}{2} \sum_{n \wedge q + 1}^{n} \underline{\sigma} \\
&= \frac{1}{2} \sum_{\gamma_i \leq \underline{\sigma}} \left(\frac{\gamma_i^2}{\gamma_i \vee \underline{\sigma}} + \gamma_i \vee \underline{\sigma} - 2\gamma_i\right) + \frac{1}{2} \sum_{n \wedge q + 1}^{n} \underline{\sigma} \\
&= \frac{1}{2} \sum_{\gamma_i \leq \underline{\sigma}} \left(\frac{\gamma_i^2}{\underline{\sigma}} + \underline{\sigma} - 2\gamma_i\right) + \frac{1}{2} \sum_{n \wedge q + 1}^{n} \underline{\sigma} \\
&= \frac{1}{2} \sum_{\gamma_i \leq \underline{\sigma}} \frac{(\gamma_i - \underline{\sigma})^2}{\underline{\sigma}} + \frac{1}{2}(n - n \wedge q)\underline{\sigma}
\end{aligned} \tag{36}$$

Hence,

$$0 \leq \left(\|\cdot\|_{\mathscr{S},1} \square \omega_{\underline{\sigma}}\right)(Z) - \|Z\|_{\mathscr{S},1} = \frac{1}{2}\sum_{\gamma_i \leq \underline{\sigma}} \frac{(\gamma_i - \underline{\sigma})^2}{\underline{\sigma}} + \frac{1}{2}(n - n \wedge q)\underline{\sigma} \leq \frac{\underline{\sigma}}{2} n \tag{37}$$

Moreover this bound is attained when $Z = 0$. □

A.5 Comparison with another smoothing of the nuclear norm

Another regularization was proposed in Argyriou et al. (2008); Bach et al. (2012, p. 62):

$$\min_{S \succ 0} \underbrace{\frac{1}{2}\operatorname{Tr}[Z^\top S^{-1} Z] + \frac{1}{2}\operatorname{Tr}(S) + \frac{\underline{\sigma}^2}{2}\operatorname{Tr}(S^{-1})}_{h(S^{-1})} \ . \tag{38}$$

By putting the gradient of the objective function in Equation (38) to zero it follows that:

$$0 = \nabla h(\hat{S}^{-1}) = ZZ^\top - \hat{S}^2 + \underline{\sigma}^2 \operatorname{Id} \ , \tag{39}$$

leading to :

$$\hat{S} = (ZZ^\top + \underline{\sigma}^2 \operatorname{Id})^{\frac{1}{2}} \ . \tag{40}$$

Let $\gamma_1, \ldots, \gamma_{n \wedge q}$ be the singular value decomposition of Z:

$$\begin{aligned}
\frac{1}{2}\operatorname{Tr}[Z^\top \hat{S}^{-1} Z] + \frac{1}{2}\operatorname{Tr}(\hat{S}) + \frac{\underline{\sigma}^2}{2}\operatorname{Tr}(\hat{S}^{-1}) &= \frac{1}{2}\sum_{i=1}^{n}\left(\frac{\gamma_i^2}{\sqrt{\gamma_i^2 + \underline{\sigma}^2}} + \sqrt{\gamma_i^2 + \underline{\sigma}^2} + \frac{\underline{\sigma}^2}{\sqrt{\gamma_i^2 + \underline{\sigma}^2}}\right) \\
&= \frac{1}{2}\sum_{i=1}^{n}\left(\frac{\gamma_i^2 + \gamma_i^2 + \underline{\sigma}^2 + \underline{\sigma}^2}{\sqrt{\gamma_i^2 + \underline{\sigma}^2}}\right) \\
&= \sum_{i=1}^{n}\sqrt{\gamma_i^2 + \underline{\sigma}^2} \ .
\end{aligned} \tag{41}$$

Proposition 18. $Z \mapsto \min_{S \succ 0} \frac{1}{2}\operatorname{Tr}[Z^\top S^{-1} Z] + \frac{1}{2}\operatorname{Tr}(S) + \frac{\underline{\sigma}^2}{2}\operatorname{Tr}(S^{-1})$ is a $\underline{\sigma}$-smooth approximation of $\|\cdot\|_{\mathscr{S},1}$ with parameters $(1, n, 0)$. More precicely: $Z \mapsto \min_{S \succ 0} \frac{1}{2}\operatorname{Tr}[Z^\top S^{-1} Z] + \frac{1}{2}\operatorname{Tr}(S) + \frac{\underline{\sigma}^2}{2}\operatorname{Tr}(S^{-1})$ has a gradient $\underline{\sigma}$-Lipschitz and

$$0 \leq \min_{S \succ 0} \frac{1}{2}\operatorname{Tr}[Z^\top S^{-1} Z] + \frac{1}{2}\operatorname{Tr}(S) + \frac{\underline{\sigma}^2}{2}\operatorname{Tr}(S^{-1}) - \|Z\|_{\mathscr{S},1} = \underline{\sigma}\sum_i \frac{1}{\sqrt{1 + \frac{\gamma_i^2}{\underline{\sigma}^2}} + \frac{\gamma_i}{\underline{\sigma}}} \leq \underline{\sigma} n \ . \tag{42}$$

Proof. $\sum_{i=1}^{n \wedge q} \sqrt{\gamma_i^2 + \underline{\sigma}^2}$ is a $\underline{\sigma}$-smooth approximation of $\sum_{i=1}^{n \wedge q} \sqrt{\gamma_i^2} = \|Z\|_{\mathscr{S},1}$, see Beck and Teboulle (2012, Example 4.6).

$$\begin{aligned}
\min_{S \succ 0} \frac{1}{2}\operatorname{Tr}[Z^\top S^{-1} Z] + \frac{1}{2}\operatorname{Tr}(S) + \frac{\underline{\sigma}^2}{2}\operatorname{Tr}(S^{-1}) - \|Z\|_{\mathscr{S},1} &= \sum_{i=1}^{n}\left(\sqrt{\gamma_i^2 + \underline{\sigma}^2} - \gamma_i\right) \\
&= \sum_{i=1}^{n}\frac{\underline{\sigma}^2}{\sqrt{\gamma_i^2 + \underline{\sigma}^2} + \gamma_i} \\
&= \underline{\sigma}\sum_{i=1}^{n}\frac{1}{\sqrt{1 + \frac{\gamma_i^2}{\underline{\sigma}^2}} + \frac{\gamma_i}{\underline{\sigma}}} \qquad (43) \\
&\leq \underline{\sigma} n \ . \qquad (44)
\end{aligned}$$

Moreover this bound is attained when $Z = 0$. □

It can be seen that with a fixed Lipschitz constant, the proposed smoothing is (at least) a twice better approximation. This can be quantify event more precisely:

Proposition 19.

$$0 \leq \underbrace{\left(\|\cdot\|_{\mathscr{S},1} \,\square\, \omega_{\underline{\sigma}}\right)(Z) - \|Z\|_{\mathscr{S},1}}_{\mathrm{Err}_1(Z)} \leq \frac{1}{2}\left(\underbrace{\min_{S\succ 0} \frac{1}{2}\operatorname{Tr}[Z^\top S^{-1} Z] + \frac{1}{2}\operatorname{Tr}(S) + \frac{\underline{\sigma}^2}{2}\operatorname{Tr}(S^{-1}) - \|Z\|_{\mathscr{S},1}}_{\mathrm{Err}_2(Z)}\right) . \tag{45}$$

More precisely

$$\frac{1}{2}\mathrm{Err}_2(Z) - \mathrm{Err}_1(Z) = \frac{\underline{\sigma}}{2}\sum_{\gamma_i \geq \sigma} \underbrace{\left(\sqrt{1+\frac{\gamma_i^2}{\underline{\sigma}^2}} - \frac{\gamma_i}{\underline{\sigma}}\right)}_{\geq 0} + \frac{\underline{\sigma}}{2}\sum_{\gamma_i < \sigma} \underbrace{\left(\frac{1}{\sqrt{1+\frac{\gamma_i^2}{\underline{\sigma}^2}} + \frac{\gamma_i}{\underline{\sigma}}} - (1+\frac{\gamma_i}{\underline{\sigma}})^2\right)}_{\geq 0} , \tag{46}$$

which means that for a fixed smoothing constant $\underline{\sigma}$, our smoothing is at least twice uniformly better. Moreover our smoothing can be much better, in particular when a lot a singular values are around $\underline{\sigma}$.

Proof. Using the formulas of Err_1 (Equation (36)) and Err_2 (Equation (43)), Equation (46) is direct. In Equation (46) the positivity of the first sum is trivial, the positivity of the second can be obtain with an easy function study. □

A.6 Schatten 1-norm (nuclear/trace norm) with repetitions

Let $Z^{(1)}, \dots, Z^{(r)}$ be matrices in $\mathbb{R}^{n\times q}$, then we define $Z \in \mathbb{R}^{n\times qr}$ by $Z = [Z^{(1)}|\dots|Z^{(r)}]$.

Proposition 20. For the choice $\omega(Z) = \frac{1}{2}\|Z\|^2 + \frac{n\wedge qr}{2}$, then the following holds true:

$$\left(\|\cdot\|_{\mathscr{S},1} \,\square\, \omega_{\underline{\sigma}}(\cdot)\right)(Z) = \min_{S\succeq \underline{\sigma}\,\mathrm{Id}_n} \frac{1}{2}\sum_{l=1}^{r}\operatorname{Tr}\left(Z^{(l)\top} S^{-1} Z^{(l)}\right) + \frac{1}{2}\operatorname{Tr}(S) \ . \tag{47}$$

Proof. The result is a direct application of Proposition 4, with $Z = [Z^{(1)}|\dots|Z^{(r)}]$. It suffices to notice that $\operatorname{Tr} Z^\top S^{-1} Z = \sum_{l=1}^{r}\operatorname{Tr}\left(Z^{(l)\top} S^{-1} Z^{(l)}\right)$. □

A.7 Schatten 2-norm (Frobenius norm)

Proposition 21. For the choice $\omega(\cdot) = \frac{1}{2}\|\cdot\|^2 + \frac{1}{2}$, and for $Z \in \mathbb{R}^{n\times q}$ then the following holds true:

$$\left(\|\cdot\| \,\square\, \omega_{\underline{\sigma}}\right)(Z) = \min_{\sigma\geq\underline{\sigma}}\left(\frac{1}{2\sigma}\|Z\|^2 + \frac{\sigma}{2}\right) = \begin{cases} \frac{\|Z\|^2}{2\underline{\sigma}} + \frac{\underline{\sigma}}{2} \ , & \text{if } \|Z\| \leq \underline{\sigma} \ , \\ \|Z\| \ , & \text{if } \|Z\| > \underline{\sigma} \ . \end{cases} \tag{48}$$

Proof. Let us recall that $\|\cdot\| = \|\cdot\|_{\mathscr{S},2}$. Therefore

$$\Pi_{\mathcal{B}_{\mathscr{S},2}}\left(\frac{Z}{\underline{\sigma}}\right) = \begin{cases} 0 \ , & \text{if } \|Z\| \leq \underline{\sigma} \ , \\ \frac{Z}{\|Z\|} \ , & \text{if } \|Z\| > \underline{\sigma} \ . \end{cases} \tag{49}$$

By combining Equation (49) and Lemma 13 with $p_1 = p_1^* = 2$, and $c = \frac{1}{2}$, the later yields

$$\left(\|\cdot\| \,\square\, \omega_{\underline{\sigma}}\right)(Z) = \begin{cases} \frac{1}{2\underline{\sigma}}\|Z\|^2 + \frac{\underline{\sigma}}{2} \ , & \text{if } \|Z\| \leq \underline{\sigma} \ , \\ \|Z\| \ , & \text{if } \|Z\| > \underline{\sigma} \ . \end{cases}$$

□

A.8 Schatten infinity-norm (spectral norm)

Proposition 22. For the choice $\omega(\cdot) = \frac{1}{2}\|\cdot\|^2 + \frac{1}{2}$ and for $Z \in \mathbb{R}^{n\times q}$, then the following holds true:

$$\left(\|\cdot\|_{\mathscr{S},\infty} \square\, \omega_{\underline{\sigma}}\right)(Z) = \begin{cases} \frac{1}{2\underline{\sigma}}\|Z\|^2 + \frac{\underline{\sigma}}{2} \ , & \text{if } \|Z\|_{\mathscr{S},1} \leq 1 \ , \\ \frac{\underline{\sigma}}{2}\sum_{i=1}^{n\wedge q}\left(\frac{\gamma_i^2}{\underline{\sigma}^2} - \nu^2\right)_+ + \frac{\underline{\sigma}}{2} \ , & \text{if } \|Z\|_{\mathscr{S},1} > 1 \ , \end{cases}$$

where $\nu \geq 0$ is defined by the implicit equation

$$\left\|\left(\mathrm{ST}\left(\frac{\gamma_1}{\underline{\sigma}}, \nu\right), \ldots, \mathrm{ST}\left(\frac{\gamma_{n\wedge q}}{\underline{\sigma}}, \nu\right)\right)\right\|_1 = 1 \ . \tag{50}$$

Proof. We remind that $\Pi_{\mathcal{B}_{\mathscr{S},\infty}}$, the projection over $\mathcal{B}_{\mathscr{S},\infty}$, is given by Beck (2017, Example 7.31, p. 192):

$$\Pi_{\mathcal{B}_{\mathscr{S},1}}\left(\frac{Z}{\underline{\sigma}}\right) = \begin{cases} \frac{Z}{\underline{\sigma}} \ , & \text{if } \|Z\|_{\mathscr{S},1} \leq \underline{\sigma} \ , \\ V\,\mathrm{diag}(\mathrm{ST}(\frac{\gamma_i}{\underline{\sigma}}, \nu))W^\top \ , & \text{if } \|Z\|_{\mathscr{S},1} > \underline{\sigma} \ , \end{cases} \tag{51}$$

γ being defined by the implicit equation

$$\left\|\left(\mathrm{ST}\left(\frac{\gamma_1}{\underline{\sigma}}, \nu\right), \ldots, \mathrm{ST}\left(\frac{\gamma_{n\wedge q}}{\underline{\sigma}}, \nu\right)\right)\right\|_1 = 1 \ . \tag{52}$$

By combining Equation (51) and Lemma 13 (with $p_1^* = \infty, c = \frac{1}{2}$) it follows that

$$(\|\cdot\| \,\square\, \omega_{\underline{\sigma}})(Z) = \begin{cases} \frac{1}{2\underline{\sigma}}\|Z\|^2 + \frac{\underline{\sigma}}{2} \ , & \text{if } \|Z\|_{\mathscr{S},1} \leq \underline{\sigma} \ , \\ \frac{1}{2\underline{\sigma}}\|Z\|^2 + \frac{\underline{\sigma}}{2} - \frac{\underline{\sigma}}{2}\left\|\Pi_{\mathcal{B}_{\mathscr{S},1}}\left(\frac{Z}{\underline{\sigma}}\right) - \frac{Z}{\underline{\sigma}}\right\|^2 \ , & \text{if } \|Z\|_{\mathscr{S},1} > \underline{\sigma} \ . \end{cases} \tag{53}$$

Let us compute $\left\|\Pi_{\mathcal{B}_{\mathscr{S},1}}\left(\frac{Z}{\underline{\sigma}}\right) - \frac{Z}{\underline{\sigma}}\right\|^2$. If $\|Z\|_{\mathscr{S},1} > \underline{\sigma}$ we have

$$\begin{aligned} \left\|\Pi_{\mathcal{B}_{\mathscr{S},1}}\left(\frac{Z}{\underline{\sigma}}\right) - \frac{Z}{\underline{\sigma}}\right\|^2 &= \frac{1}{\underline{\sigma}^2}\left\|V\,\mathrm{diag}((\gamma_i - \nu\underline{\sigma})_+ - \gamma_i)W^\top\right\| \qquad \text{(using Equation (51))} \\ &= \frac{1}{\underline{\sigma}^2}\sum_{i=1}^{n\wedge q}\left((\gamma_i - \nu\underline{\sigma})_+ - \gamma_i\right)^2 \\ &= \frac{1}{\underline{\sigma}^2}\Big(\sum_{\gamma_i \geq \nu\underline{\sigma}}^{n\wedge q}\nu^2\underline{\sigma}^2 + \sum_{\gamma_i < \nu\underline{\sigma}}^{n\wedge q}\gamma_i^2\Big) \ . \end{aligned} \tag{54}$$

By plugging Equation (54) into Equation (53) it follows, that if $\|Z\|_{\mathscr{S},1} > \underline{\sigma}$:

$$\begin{aligned} \left(\|\cdot\| \,\square\, \omega_{\underline{\sigma}}\right)(Z) &= \frac{1}{2\underline{\sigma}}\sum_{i=1}^{n\wedge q}\gamma_i^2 + \frac{\underline{\sigma}}{2} - \frac{1}{2\underline{\sigma}}\sum_{\gamma_i \geq \nu\underline{\sigma}}^{n\wedge q}\nu^2\underline{\sigma}^2 - \frac{1}{2\underline{\sigma}}\sum_{\gamma_i < \nu\underline{\sigma}}^{n\wedge q}\gamma_i^2 \\ &= \frac{1}{2\underline{\sigma}}\sum_{\gamma_i \geq \nu\underline{\sigma}}^{n\wedge q}\left(\gamma_i^2 - \nu^2\underline{\sigma}^2\right) + \frac{\underline{\sigma}}{2} \\ &= \frac{\underline{\sigma}}{2}\sum_{i=1}^{n\wedge q}\left(\frac{\gamma_i^2}{\underline{\sigma}^2} - \nu^2\right)_+ + \frac{\underline{\sigma}}{2} \ . \end{aligned} \tag{55}$$

Proposition 22 follows by plugging Equation (55) for the case $\|Z\|_{\mathscr{S},1} > \underline{\sigma}$, and the fact that when $\|Z\|_{\mathscr{S},1} \leq \underline{\sigma}$ the result is straightforward. □

Remark 23. Since $\nu \mapsto \left\|\left(\mathrm{ST}\left(\frac{\gamma_1}{\underline{\sigma}}, \nu\right), \ldots, \mathrm{ST}\left(\frac{\gamma_{n\wedge q}}{\underline{\sigma}}, \nu\right)\right)\right\|_1$ is decreasing and piecewise linear, the solution of Equation (50) can be computed exactly in $\mathcal{O}(n \wedge q \ \log(n \wedge q))$ operations.

B Proofs CLaR

B.1 Proof of Proposition 6

Proposition 6 (proof in Appendix B.1)**.** Any solution of the CLaR Problem (2), $(\hat{\mathrm{B}}, \hat{S}) = (\hat{\mathrm{B}}^{\mathrm{CLaR}}, \hat{S}^{\mathrm{CLaR}})$ is also a solution of:

$$\hat{\mathrm{B}} = \operatorname*{arg\,min}_{\mathrm{B}\in\mathbb{R}^{p\times q}} \left(\|\cdot\|_{\mathscr{S},1} \,\square\, \omega_{\underline{\sigma}} \right)(Z) + \lambda n \|\mathrm{B}\|_{2,1}$$
$$\hat{S} = \mathrm{ClSqrt}\left(\tfrac{1}{r} ZZ^\top, \underline{\sigma}\right) \ , \text{ where } Z = [Z^{(1)}|\dots|Z^{(r)}] \text{ and } Z^{(l)} = \tfrac{Y^{(l)} - X\mathrm{B}}{\sqrt{q}} \ .$$

Proof. Proposition 6 follows from Appendix A.6 by choosing $Z = \frac{1}{\sqrt{rq}}[Y^{(1)} - X\mathrm{B}, \dots, Y^{(r)} - X\mathrm{B}]$ and by taking the $\arg\min$ over B. □

B.2 Proof of Proposition 7

Proposition 7 (Proof in Appendix B.2)**.** CLaR is jointly convex in (B, S). Moreover, f is convex and smooth on the feasible set, and $\|\cdot\|_{2,1}$ is convex and separable in $\mathrm{B}_{j:}$'s, thus minimizing the objective alternatively in S and in $\mathrm{B}_{j:}$'s (see Algorithm 1) converges to a global minimum.

Proof.

$$f(\mathrm{B}, S) = \frac{1}{2nqr} \sum_1^r \left\| Y^{(l)} - X\mathrm{B} \right\|_{S^{-1}}^2 + \frac{1}{2n} \mathrm{Tr}(S) = \mathrm{Tr}(Z^T S^{-1} Z) + \frac{1}{2n} \mathrm{Tr}(S) \ ,$$

with $Z = \frac{1}{\sqrt{2nqr}}[Y^{(1)} - X\mathrm{B}|\dots|Y^{(r)} - X\mathrm{B}]$.

First note that the (joint) function $(Z, \Sigma) \mapsto \mathrm{Tr}\, Z^\top \Sigma^{-1} Z$ is jointly convex over $\mathbb{R}^{n\times q} \times \mathcal{S}_{++}^n$, see Boyd and Vandenberghe (2004, Example 3.4). This means that f is jointly convex in (Z, S), moreover $\mathrm{B} \mapsto \frac{1}{\sqrt{2nqr}}[Y^{(1)} - X\mathrm{B}|\dots|Y^{(r)} - X\mathrm{B}]$ is linear in B, thus f is jointly convex in (B, S), meaning that $(\mathrm{B}, S) \to f + \lambda \|\cdot\|_{2,1}$ is jointly convex in (B, S) . Moreover the constraint set is convex and thus solving CLaR is a convex problem.

The function f is convex and smooth on the feasible set and $\|\cdot\|_{2,1}$ is convex in B and separable in $\mathrm{B}_{j:}$'s, thus (see Tseng 2001; Tseng and Yun 2009) $f + \lambda \|\cdot\|_{2,1}$ can be minimized through coordinate descent in S and the $\mathrm{B}_{j:}$'s (on the feasible set). □

B.3 Proof of Proposition 8

Proposition 8 (Minimization in S; proof in Appendix B.3)**.** Let $B \in \mathbb{R}^{n\times q}$ be fixed. The minimization of $f(\mathrm{B}, S)$ *w.r.t.* S with the constraint $S \succeq \underline{\sigma}\, \mathrm{Id}_n$ admits the closed-form solution:

$$S = \mathrm{ClSqrt}\left(\frac{1}{r} \sum_{l=1}^r Z^{(l)} Z^{(l)\top}, \underline{\sigma} \right) \ , \quad \text{with } Z^{(l)} = \tfrac{1}{\sqrt{q}}(Y^{(l)} - X\mathrm{B}) \ . \tag{10}$$

Proof. Minimizing $f(\mathrm{B}, \cdot)$ amounts to solving

$$\operatorname*{arg\,min}_{S\succeq\underline{\sigma}\,\mathrm{Id}_n} \tfrac{1}{2} \|Z\|_{S^{-1}}^2 + \tfrac{1}{2} \mathrm{Tr}(S) \ , \quad \text{with } Z = \frac{1}{\sqrt{r}}[Z^{(1)}|\dots|Z^{(l)}] \ . \tag{56}$$

The solution is $\mathrm{ClSqrt}\left(ZZ^\top, \underline{\sigma}\right)$ (see Massias et al. 2018a, Appendix A2), and $ZZ^\top = \frac{1}{r}\sum_{l=1}^r Z^{(l)} Z^{(l)\top}$. □

B.4 Proof of Proposition 9

Proposition 9 (Proof in Appendix B.4)**.** For a fixed $S \in \mathcal{S}^n_{++}$, each step of the block minimization of $f(\cdot, S) + \lambda \|\cdot\|_{2,1}$ in the j^{th} line of B admits a closed-form solution:

$$\mathrm{B}_{j:} = \mathrm{BST}\left(\mathrm{B}_{j:} + \frac{X_{:j}^\top S^{-1}(\bar{Y} - X\mathrm{B})}{\|X_{:j}\|^2_{S^{-1}}}, \frac{\lambda nq}{\|X_{:j}\|^2_{S^{-1}}}\right) \quad . \tag{11}$$

Proof. The function to minimize is the sum of a smooth term $f(\cdot, S)$ and a non-smooth but separable term, $\|\cdot\|_{2,1}$, whose proximal operator [7] can be computed:

- f is $\|X_{:j}\|^2_{S^{-1}}/nq$-smooth with respect to $\mathrm{B}_{j:}$, with partial gradient $\nabla_j f(\cdot, S) = -\frac{1}{nq} X_{:j}^\top S^{-1}(\bar{Y} - X\mathrm{B})$,
- $\|\mathrm{B}\|_{2,1} = \sum_{j=1}^p \|\mathrm{B}_{j:}\|$ is row-wise separable over B, with $\mathrm{prox}_{\lambda nq/\|X_{:j}\|^2_{S^{-1}}, \|\cdot\|}(\cdot) = \mathrm{BST}\left(\cdot, \frac{\lambda nq}{\|X_{:j}\|^2_{S^{-1}}}\right)$.

Hence, proximal block-coordinate descent converges (Tseng and Yun, 2009), and the update are given by Equation (11). The closed-form formula arises since the smooth part of the objective is quadratic and isotropic *w.r.t.* $\mathrm{B}_{j:}$. □

B.5 Proof of $\lambda_{\max}$ CLaR

Proof. First notice that if $\hat{\mathrm{B}} = 0$, then $\hat{S} = \mathrm{ClSqrt}\left(\frac{1}{qr}\sum_{l=1}^r Y^{(l)} Y^{(l)\top}, \underline{\sigma}\right) \triangleq S_{\max}$.

Fermat's rules states that

$$\begin{aligned}
\hat{\mathrm{B}} = 0 &\Leftrightarrow 0 \in \partial\big(f(\cdot, S_{\max}) + \lambda\|\cdot\|_{2,1}\big)(0) \\
&\Leftrightarrow -\nabla f(\cdot, S_{\max}) \in \lambda \mathcal{B}_{\|\cdot\|_{2,\infty}} \\
&\Leftrightarrow \frac{1}{nq}\left\|X^\top S^{-1}_{\max} \bar{Y}\right\|_{2,\infty} \triangleq \lambda_{\max} \leq \lambda \ .
\end{aligned} \tag{57}$$

□

B.6 Proof of dual formulation

Proposition 24. With $\hat{\Theta} = (\hat{\Theta}^{(1)}, \dots, \hat{\Theta}^{(r)})$, the dual formulation of Problem (2) is

$$\hat{\Theta} = \underset{(\Theta^{(1)}, \dots, \Theta^{(r)}) \in \Delta_{X,\lambda}}{\arg\max} \frac{\underline{\sigma}}{2}\left(1 - \frac{qn\lambda^2}{r}\sum_{l=1}^r \mathrm{Tr}\,\Theta^{(l)}\Theta^{(l)\top}\right) + \frac{\lambda}{r}\sum_{l=1}^r \left\langle \Theta^{(l)}, Y^{(l)} \right\rangle \quad ,$$

with $\bar{\Theta} = \frac{1}{r}\sum_1^r \Theta^{(l)}$ and

$$\Delta_{X,\lambda} = \left\{(\Theta^{(1)}, \dots, \Theta^{(r)}) \in (\mathbb{R}^{n\times q})^r : \left\|X^\top \bar{\Theta}\right\|_{2,\infty} \leq 1, \left\|\sum_{l=1}^r \Theta^{(l)}\Theta^{(l)\top}\right\|_2 \leq \frac{r}{\lambda^2 n^2 q}\right\} \ .$$

In Algorithm 1 the dual point Θ at iteration t is obtained through a residual rescaling similar to the way the dual point is created,*i.e.,* $\Theta^{(l)} = \frac{1}{nq\lambda}(Y^{(l)} - X\mathrm{B})$ (with B the current primal iterate); then the dual point hence created is projected on $\Delta_{X,\lambda}$.

Proof. Let the primal optimum be

$$p^* \triangleq \min_{\substack{\mathrm{B}\in\mathbb{R}^{p\times q} \\ S \succeq \underline{\sigma}\,\mathrm{Id}_n}} \frac{1}{2nqr}\sum_{l=1}^r \left\|Y^{(l)} - X\mathrm{B}\right\|^2_{S^{-1}} + \frac{1}{2n}\mathrm{Tr}(S) + \lambda\|\mathrm{B}\|_{2,1}$$

[7] As a reminder, for a scalar $t \in \mathbb{R}$, the proximal operator of a function $h : \mathbb{R}^d \mapsto \mathbb{R}$ can be defined for any $x_0 \in \mathbb{R}^d$ by $\mathrm{prox}_{t,h}(x_0) = \arg\min_{x\in\mathbb{R}^d} \frac{1}{2t}\|x - x_0\|^2 + h(x)$.

Then

$$p^* = \min_{\substack{\mathrm{B}\in\mathbb{R}^{p\times q}\\ R^{(l)}=Y^{(l)}-X\mathrm{B},\ \forall l\in[r]\\ S\succeq\underline{\sigma}\,\mathrm{Id}_n}} \frac{1}{2nqr}\sum_{l=1}^{r}\left\|R^{(l)}\right\|_{S^{-1}}^2 + \frac{1}{2n}\operatorname{Tr}(S) + \lambda\left\|\mathrm{B}\right\|_{2,1}$$

$$= \min_{\substack{\mathrm{B}\in\mathbb{R}^{p\times q}\\ R^{(1)},\dots,R^{(r)}\in\mathbb{R}^{n\times q}\\ S\succeq\underline{\sigma}\,\mathrm{Id}_n}} \max_{\Theta^{(1)},\dots,\Theta^{(r)}\in\mathbb{R}^{n\times q}} \frac{1}{2nqr}\sum_{l=1}^{r}\left\|R^{(l)}\right\|_{S^{-1}}^2 + \frac{1}{2n}\operatorname{Tr}(S)$$

$$+\lambda\left\|\mathrm{B}\right\|_{2,1} + \frac{\lambda}{r}\sum_{l=1}^{r}\left\langle\Theta^{(l)}, Y^{(l)}-X\mathrm{B}-R^{(l)}\right\rangle \ .$$

Since Slater's conditions are met $\min$ and $\max$ can be inverted:

$$p^* = \max_{\Theta^{(1)},\dots,\Theta^{(r)}\in\mathbb{R}^{n\times q}} \min_{\substack{\mathrm{B}\in\mathbb{R}^{p\times q}\\ R^{(1)},\dots,R^{(r)}\in\mathbb{R}^{n\times q}\\ S\succeq\underline{\sigma}\,\mathrm{Id}_n}} \frac{1}{2nqr}\sum_{l=1}^{r}\left\|R^{(l)}\right\|_{S^{-1}}^2 + \frac{1}{2n}\operatorname{Tr}(S) \tag{58}$$

$$+\lambda\left\|\mathrm{B}\right\|_{2,1} + \frac{\lambda}{r}\sum_{l=1}^{r}\left\langle\Theta^{(l)}, Y^{(l)}-X\mathrm{B}-R^{(l)}\right\rangle$$

$$= \max_{\Theta^{(1)},\dots,\Theta^{(r)}\in\mathbb{R}^{n\times q}} \left(\min_{S\succeq\underline{\sigma}\,\mathrm{Id}_n}\frac{1}{r}\sum_{l=1}^{r}\min_{R^{(l)}\in\mathbb{R}^{n\times q}}\left(\frac{\left\|R^{(l)}\right\|_{S^{-1}}^2}{2nq} - \left\langle\Theta^{(l)}, R^{(l)}\right\rangle\right) + \frac{1}{2n}\operatorname{Tr}(S)\right.$$

$$\left. +\lambda\min_{\mathrm{B}\in\mathbb{R}^{p\times q}}\left(\left\|\mathrm{B}\right\|_{2,1} - \left\langle\bar{\Theta}, X\mathrm{B}\right\rangle\right) + \frac{\lambda}{r}\sum_{l=1}^{r}\left\langle\Theta^{(l)}, Y^{(l)}\right\rangle\right) \ . \tag{59}$$

Morover we have

$$\min_{R^{(l)}\in\mathbb{R}^{n\times q}}\left(\frac{\left\|R^{(l)}\right\|_{S^{-1}}^2}{2nq} - \left\langle\Theta^{(l)}, R^{(l)}\right\rangle\right) = -\frac{nq\lambda^2}{2}\left\langle\Theta^{(l)}{\Theta^{(l)}}^\top, S\right\rangle$$

and

$$\min_{\mathrm{B}\in\mathbb{R}^{p\times q}}\left(\left\|\mathrm{B}\right\|_{2,1} - \langle\bar{\Theta}, X\mathrm{B}\rangle\right) = -\max\left(\langle X^\top\bar{\Theta}, \mathrm{B}\rangle - \left\|\mathrm{B}\right\|_{2,1}\right) = -\iota_{\mathcal{B}_{2,\infty}}(X^\top\bar{\Theta}) \ .$$

This leads to:

$$
\begin{aligned}
d^* = & \max_{\Theta^{(1)},\dots,\Theta^{(r)}\in\mathbb{R}^{n\times q}} \min_{S\succeq\underline{\sigma}\,\mathrm{Id}_n} -\frac{1}{r}\sum_{l=1}^{r}\tfrac{nq\lambda^2}{2}\left\langle \Theta^l{\Theta^l}^\top, S\right\rangle - \lambda\iota_{\mathcal{B}_{2,\infty}}(X^\top\bar{\Theta}) + \frac{\mathrm{Tr}(S)}{2n} \\
& \qquad + \frac{\lambda}{r}\sum_{l=1}^{r}\langle\Theta^{(l)}, Y^{(l)}\rangle \\
= & \max_{\Theta^{(1)},\dots,\Theta^{(r)}\in\mathbb{R}^{n\times q}} \frac{1}{2n}\min_{S\succeq\underline{\sigma}\,\mathrm{Id}_n}\left(\left\langle \mathrm{Id}_n, S\right\rangle - \frac{qn^2\lambda^2}{r}\sum_{l=1}^{r}\left\langle \Theta^{(l)}{\Theta^{(l)}}^\top, S\right\rangle\right) - \lambda\iota_{\mathcal{B}_{2,\infty}}(X^\top\bar{\Theta}) \\
& \qquad + \frac{\lambda}{r}\sum_{l=1}^{r}\langle\Theta^{l}, Y^{(l)}\rangle \\
= & \max_{\Theta^{(1)},\dots,\Theta^{(r)}\in\mathbb{R}^{n\times q}} \frac{1}{2n}\min_{S\succeq\underline{\sigma}\,\mathrm{Id}_n}\left\langle \mathrm{Id}_n - \tfrac{qn^2\lambda^2}{r}\sum_{l=1}^{r}\Theta^{(l)}{\Theta^{(l)}}^\top S\right\rangle - \lambda\iota_{\mathcal{B}_{2,\infty}}(X^\top\bar{\Theta}) \\
& \qquad + \frac{\lambda}{r}\sum_{l=1}^{r}\langle\Theta^{l}, Y^{(l)}\rangle \ . \qquad (60)
\end{aligned}
$$

$$
\begin{aligned}
& \min_{S\succeq\underline{\sigma}\,\mathrm{Id}_n}\left\langle \mathrm{Id}_n - \tfrac{qn^2\lambda^2}{r}\sum_{l=1}^{r}\Theta^{(l)}{\Theta^{(l)}}^\top, S\right\rangle \\
& \qquad = \begin{cases} \left\langle \mathrm{Id}_n - \frac{qn^2\lambda^2}{r}\sum_{l=1}^{r}\Theta^{(l)}{\Theta^{(l)}}^\top, \underline{\sigma}\right\rangle \ , & \text{if } \mathrm{Id}_n \succeq \frac{qn^2\lambda^2}{r}\sum_{l=1}^{r}\Theta^{(l)}{\Theta^{(l)}}^\top \ , \\ -\infty \ , & \text{otherwise.} \end{cases} \qquad (61)
\end{aligned}
$$

It follows that the dual problem of CLaR is

$$
\max_{(\Theta^{(1)},\dots,\Theta^{(r)})\in\Delta_{X,\lambda}} \frac{\underline{\sigma}}{2}\left(1 - \frac{qn\lambda^2}{r}\sum_{l=1}^{r}\mathrm{Tr}\,\Theta^{(l)}{\Theta^{(l)}}^\top\right) + \frac{\lambda}{r}\sum_{l=1}^{r}\left\langle \Theta^{(l)}, Y^{(l)}\right\rangle \ , \qquad (62)
$$

where $\Delta_{X,\lambda} \triangleq \left\{(\Theta^{(1)},\dots,\Theta^{(r)})\in\mathbb{R}^{n\times q\times r} : \left\|X^\top\bar{\Theta}\right\|_{2,\infty}\le 1, \left\|\sum_{l=1}^{r}\Theta^l{\Theta^l}^\top\right\| \le \frac{r}{\lambda^2 n^2 q}\right\}$. □

B.7 Proof of Remark 11

Remark 11. Once $\mathrm{cov}_Y \triangleq \frac{1}{r}\sum_1^r Y^{(l)}Y^{(l)\top}$ is pre-computed, the cost of updating S does not depend on r, *i.e.,* is the same as working with averaged data. Indeed, with $R = [Y^{(1)} - X\mathrm{B}|\dots|Y^{(r)} - X\mathrm{B}]$, the following computation can be done in $\mathcal{O}(qn^2)$ (details are in Appendix B.7).

$$
RR^\top = \mathrm{RRT}(\mathrm{cov}_Y, Y, X, \mathrm{B}) \triangleq r\mathrm{cov}_Y + r(X\mathrm{B})(X\mathrm{B})^\top - r\bar{Y}^\top(X\mathrm{B}) - r(X\mathrm{B})^\top\bar{Y} \ . \qquad (13)
$$

Proof.

$$
\begin{aligned}
RR^\top &= \sum_{l=1}^{r} R^{(l)}R^{(l)\top} \\
&= \sum_{l=1}^{r}(Y^{(l)} - X\mathrm{B})(Y^{(l)} - X\mathrm{B})^\top \\
&= \sum_{l=1}^{r} Y^{(l)}Y^{(l)\top} - \sum_{1}^{r} Y^{(l)}(X\mathrm{B})^\top - \sum_{1}^{r} X\mathrm{B}Y^{(l)\top} + rX\mathrm{B}(X\mathrm{B})^\top \\
&= r\mathrm{cov}_Y - r\bar{Y}^\top X\mathrm{B} - r(X\mathrm{B})^\top\bar{Y} + rX\mathrm{B}(X\mathrm{B})^\top \qquad (63)
\end{aligned}
$$

□

B.8 Statistical comparison

In this subsection, we show the statistical interest of using all repetitions of the experiments instead of using a mere averaging as SGCL would do (remind that the later is equivalent to CLaR with $r = 1$ and $Y^{(1)} = \bar{Y}$, see Remark 3).

Let us introduce Σ^*, the true covariance matrix of the noise (*i.e.,* $\Sigma^* = S^{*2}$ with our notation). In SGCL and CLaR alternate minimization consists in a succession of estimations of B^* and Σ^* (more precisely $S = \mathrm{ClSqrt}(\Sigma, \underline{\sigma})$ is estimated along the process). In this section we explain why the estimation of Σ^* provided by CLaR has better statistical properties than that of SGCL. For that, we can compare the estimates of Σ^* one would obtain provided that the true parameter B^* is known by both SGCL and CLaR. In such "ideal" scenario, the associated estimators of Σ^* could be written:

$$\hat{\Sigma}^{\mathrm{CLaR}} \triangleq \frac{1}{qr}\sum_{l=1}^{r}(Y^{(l)} - X\hat{\mathrm{B}})(Y^{(l)} - X\hat{\mathrm{B}})^\top \ , \tag{64}$$

$$\hat{\Sigma}^{\mathrm{SGCL}} \triangleq \frac{1}{qr}\Big(\sum_{l=1}^{r}Y^{(l)} - X\hat{\mathrm{B}}\Big)\Big(\sum_{l=1}^{r}Y^{(l)} - X\hat{\mathrm{B}}\Big)^\top, \tag{65}$$

with $\hat{\mathrm{B}} = \mathrm{B}^*$, and satisfy the following properties:

Proposition 25. Provided that the true signal is known, and that the covariance estimator $\hat{\Sigma}^{\mathrm{CLaR}}$ and $\hat{\Sigma}^{\mathrm{SGCL}}$ are defined thanks to Equations (64) and (65), then one can check that

$$\mathbb{E}(\hat{\Sigma}^{\mathrm{CLaR}}) = \mathbb{E}(\hat{\Sigma}^{\mathrm{SGCL}}) = \Sigma^* \ , \tag{66}$$

$$\mathrm{cov}(\hat{\Sigma}^{\mathrm{CLaR}}) = \frac{1}{r}\mathrm{cov}(\hat{\Sigma}^{\mathrm{SGCL}}) \ . \tag{67}$$

Proposition 25 states that $\hat{\Sigma}^{\mathrm{CLaR}}$ and $\hat{\Sigma}^{\mathrm{SGCL}}$ are unbiased estimators of Σ^* but our newly introduced CLaR, improves the estimation of the covariance structure by a factor r, the number of repetitions performed.

Empirically[8], we have also observed that $\hat{\Sigma}^{\mathrm{CLaR}}$ has larger eigenvalues than $\hat{\Sigma}^{\mathrm{SGCL}}$, leading to a less biased estimation of S^* after clipping the singular values.

Let us recall that

$$\Sigma^{\mathrm{SGCL}} = \frac{1}{qr}\left(\sum_{l=1}^{r}R^{(l)}\right)\left(\sum_{l=1}^{r}R^{(l)}\right)^\top \qquad \text{and} \qquad \Sigma^{\mathrm{CLaR}} = \frac{1}{qr}\sum_{l=1}^{r}R^{(l)}R^{(l)\top} \ . \tag{68}$$

Proof of Equation (66)

Proof. If $\mathrm{B} = \mathrm{B}^*$, $R^{(l)} = S^*\mathrm{E}^{(l)}$, where $\mathrm{E}^{(l)}$ are random matrices with normal i.i.d. entries, and the result trivially follows. □

Proof of Equation (67)

Proof. If $\hat{\mathrm{B}} = \mathrm{B}^*$, $Y^{(l)} - X\hat{\mathrm{B}} = S^*\mathrm{E}^{(l)}$, where the $\mathrm{E}^{(l)}$'s are random matrices with normal i.i.d. entries.

Now, on the one hand :

$$\hat{\Sigma}^{\mathrm{SGCL}} = \frac{1}{qr}\left(\sum_{l=1}^{r}S^*\mathrm{E}^{(l)}\right)\left(\sum_{l=1}^{r}S^*\mathrm{E}^{(l)}\right)^\top \ .$$

Since $\frac{1}{\sqrt{r}}\sum_{l=1}^{r}S^*\mathrm{E}^{(l)} \underset{law}{\sim} S^*\mathrm{E}$ it follows that

$$\hat{\Sigma}^{\mathrm{SGCL}} \underset{law}{\sim} \frac{1}{q}S^*\mathrm{E}(S^*\mathrm{E})^\top,$$

$$\mathrm{cov}(\hat{\Sigma}^{\mathrm{SGCL}}) = \frac{1}{q^2}\mathrm{cov}(S^*\mathrm{E}(S^*\mathrm{E})^\top) \ .$$

[8] In that case we plug $\hat{\mathrm{B}} = \hat{\mathrm{B}}^{\mathrm{CLaR}}$ (resp. $\hat{\mathrm{B}} = \hat{\mathrm{B}}^{\mathrm{CLaR}}$) in Proposition 25.

On the other hand:

$$\hat{\Sigma}^{\mathrm{CLaR}} = \frac{1}{qr}\sum_{l=1}^{r} S^*\mathrm{E}^{(l)}(S^*\mathrm{E}^{(l)})^\top \quad .$$

Since the $\mathrm{E}^{(l)}$'s are independent it follows that

$$\begin{aligned}\mathrm{cov}(\hat{\Sigma}^{\mathrm{CLaR}}) &= \frac{1}{r^2q^2}\sum_{l=1}^{r}\mathrm{cov}\left(S^*\mathrm{E}^{(l)}(S^*\mathrm{E}^{(l)})^\top\right) = \frac{1}{r^2q^2}\sum_{l=1}^{r}\mathrm{cov}\left(S^*\mathrm{E}(S^*\mathrm{E})^\top\right) \\ &= \frac{1}{rq^2}\mathrm{cov}\left(S^*\mathrm{E}(S^*\mathrm{E})^\top\right) = \frac{1}{r}\mathrm{cov}\left(\hat{\Sigma}^{\mathrm{SGCL}}\right) \quad .\end{aligned}$$

□

C Alternative estimators

We compare CLaR to several estimators: SGCL (Massias et al., 2018a), the (smoothed) $\ell_{2,1}$-Maximum Likelihood ($\ell_{2,1}$-MLE), and a version of the $\ell_{2,1}$-MLE with multiple repetitions ($\ell_{2,1}$-MLER), an $\ell_{2,1}$ penalized version of the Multivariate Regression with Covariance Estimation (Rothman et al., 2010) ($\ell_{2,1}$-MRCE), an $\ell_{2,1}$ penalized version of $\ell_{2,1}$-MRCE with repetitions ($\ell_{2,1}$-MRCER) and the Multi-Task Lasso (Obozinski et al. 2010, MTL). The cost of an epoch of block coordinate descent and the cost of computing the duality gap for each algorithm are summarized in Table 1. The updates of each algorithms are summarized in Table 2.

CLaR solves Problem (2) and SGCL solves Equation (4), let us introduce the definition of the alternative estimation procedures.

C.1 Multi-Task Lasso (MTL)

The MTL (Obozinski et al., 2010) is the classical estimator used when the additive noise is supposed to be homoscedastic (with no correlation). MTL is obtained by solving:

$$\hat{\mathrm{B}}^{\mathrm{MTL}} \in \underset{\mathrm{B}\in\mathbb{R}^{p\times q}}{\arg\min}\ \frac{1}{2nq}\left\|\bar{Y} - X\mathrm{B}\right\|^2 + \lambda\left\|\mathrm{B}\right\|_{2,1} \quad . \tag{69}$$

Remark 26. It can be seen that trying to use all the repetitions in the MTL leads to MTL itself because $\left\|\bar{Y} - X\mathrm{B}\right\|^2 = \frac{1}{r}\sum_l \left\|Y^{(l)} - X\mathrm{B}\right\|^2$.

C.2 $\ell_{2,1}$-Maximum Likelihood ($\ell_{2,1}$-MLE)

Here we study a penalized Maximum Likelihood Estimator (Chen and Banerjee, 2017) ($\ell_{2,1}$-MLE). When minimizing $\ell_{2,1}$-Maximum Likelihood the natural parameters of the problem are the regression coefficients B and the precision matrix Σ^{-1}. Since real M/EEG covariance matrices are not full rank, one has to be algorithmically careful when Σ becomes singular. To avoid such numerical errors and to be consistent with the smoothed estimator proposed in the paper (CLaR), let us define the (smoothed) $\ell_{2,1}$-MLE as following:

$$(\hat{\mathrm{B}}^{\ell_{2,1}-\mathrm{MLE}}, \hat{\Sigma}^{\ell_{2,1}-\mathrm{MLE}}) \in \underset{\substack{\mathrm{B}\in\mathbb{R}^{p\times q}\\ \Sigma\succeq\underline{\sigma}^2/r^2}}{\arg\min}\ \left\|\bar{Y} - X\mathrm{B}\right\|^2_{\Sigma^{-1}} - \log\det(\Sigma^{-1}) + \lambda\left\|\mathrm{B}\right\|_{2,1} \quad , \tag{70}$$

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

$$(\hat{\mathrm{B}}^{\ell_{2,1}\mathrm{MLER}}, \hat{\Sigma}^{\ell_{2,1}\mathrm{MLER}}) \in \underset{\substack{\mathrm{B}\in\mathbb{R}^{p\times q}\\ \Sigma\succeq\underline{\sigma}^2}}{\arg\min}\ \sum_1^r \left\|Y^{(l)} - X\mathrm{B}\right\|^2_{\Sigma^{-1}} - \log\det(\Sigma^{-1}) + \lambda\left\|\mathrm{B}\right\|_{2,1} \quad . \tag{71}$$

Problems (70) and (71) are not convex because the objective functions are not convex in $(\mathrm{B}, \Sigma^{-1})$, however they are biconvex, *i.e.,* convex in B and convex in Σ^{-1}. Alternate minimization can be used to solve Problems (70) and (71), but without guarantees to converge toward a global minimum.

Minimization in $\mathrm{B}_{j:}$ As for CLaR and SGCL the updates in $\mathrm{B}_{j:}$'s for $\ell_{2,1}$-MLE and $\ell_{2,1}$-MLER clearly read:

$$\mathrm{B}_{j:} = \mathrm{BST}\left(\mathrm{B}_{j:} + \frac{X_{:j}^\top \Sigma^{-1}(Y - X\mathrm{B})}{\|X_{:j}\|_{S^{-1}}^2}, \frac{\lambda n q}{\|X_{:j}\|_{S^{-1}}^2}\right) \ . \tag{72}$$

Minimization in Σ^{-1}: for $\ell_{2,1}$-MLE (*resp.* for $\ell_{2,1}$-MLER) the update in Σ reads

$$\Sigma = \mathrm{Cl}(\Sigma^{\mathrm{EMP}}, \underline{\sigma}^2) \quad (\textit{resp.}\ \Sigma = \mathrm{Cl}(\Sigma^{\mathrm{EMP},r}, \underline{\sigma}^2)) \ , \tag{73}$$

with $\Sigma^{\mathrm{EMP}} \triangleq \frac{1}{q}(\bar{Y} - X\mathrm{B})(\bar{Y} - X\mathrm{B})^\top$ (*resp.* $\Sigma^{\mathrm{EMP},r} \triangleq \frac{1}{rq}\sum_{l=1}^r (Y^{(l)} - X\mathrm{B})(Y^{(l)} - X\mathrm{B})^\top$)

Let us prove the last result. Minimizing Problem (70) in Σ^{-1} amounts to solving

$$\hat{\Sigma}^{-1} \in \underset{0 \prec \Sigma^{-1} \preceq 1/\underline{\sigma}^2}{\arg\min} \quad \left\langle \Sigma^{\mathrm{EMP}}, \Sigma^{-1} \right\rangle - \log\det(\Sigma^{-1}) \ . \tag{74}$$

Theorem 27. Let $\Sigma^{\mathrm{EMP}} = U\,\mathrm{diag}(\sigma_i^2)U^\top$ be an eigenvalue decomposition of Σ^{EMP}, a solution to Problem (74) is given by:

$$\hat{\Sigma}^{-1} = U\,\mathrm{diag}\left(\frac{1}{\sigma_i^2 \vee \underline{\sigma}^2}\right)U^\top \tag{75}$$

Theorem 27 is very intuitive, the solution of the smoothed optimization problem (74) is the solution of the non-smoothed problem, where the eigenvalues of the solution have lifted to satisfy the constraint. Let us proove this result.

Proof. The KKT conditions of Problem (74) for conic programming (see Boyd and Vandenberghe 2004, p. 267) state that the optimum in the primal $\hat{\Sigma}^{-1}$ and the optimum in the dual $\hat{\Gamma}$ should satisfy:

$$\Sigma^{\mathrm{EMP}} - \hat{\Sigma} + \hat{\Gamma} = 0 \ , \qquad \hat{\Gamma}^\top(\hat{\Sigma}^{-1} - \frac{1}{\underline{\sigma}^2}\,\mathrm{Id}_n) = 0 \ ,$$

$$\hat{\Gamma} \in \mathcal{S}_+^n \ , \qquad 0 \prec \hat{\Sigma}^{-1} \preceq \frac{1}{\underline{\sigma}^2} \ .$$

Since Problem (74) is convex these conditions are also sufficient. Let us propose a primal-dual point $(\hat{\Sigma}^{-1}, \hat{\Gamma})$ satisfying the KKT conditions. Let $\Sigma^{\mathrm{EMP}} = U\,\mathrm{diag}(\sigma_i^2)U^\top$ be an eigenvalue decomposition of Σ^{EMP}, one can check that

$$\hat{\Sigma}^{-1} = U\,\mathrm{diag}\left(\frac{1}{\sigma_i^2 \vee \underline{\sigma}^2}\right)U^\top \ ,$$

$$\hat{\Gamma} = U\,\mathrm{diag}(\sigma_i^2 \vee \underline{\sigma}^2 - \sigma_i^2)U^\top \ .$$

verify the KKT conditions, leading to the desired result. □

C.3 Multivariate Regression with Covariance Estimation (MRCE)

MRCE (Rothman et al., 2010) jointly estimates the regression coefficients (assumed to be sparse) and the precision matrix (*i.e.,* the inverse of the covariance matrix), which is supposed to be sparse as well. Originally in Rothman et al. (2010) the sparsity enforcing term on the regression coefficients was an ℓ_1-norm, which is not well suited for our problem, that is why in Appendix C.3.2we introduce an $\ell_{2,1}$ penalized version of MRCE: $\ell_{2,1}$-MRCE.

C.3.1 Multivariate Regression with Covariance Estimation

$\ell_{2,1}$-MRCE if defined as the solution of the following optimization problem:

$$(\hat{\mathrm{B}}^{\mathrm{MRCE}}, \hat{\Sigma}^{\mathrm{MRCE}}) \in \underset{\substack{\mathrm{B}\in\mathbb{R}^{p\times q} \\ \Sigma^{-1}\succ 0}}{\arg\min} \ \left\|\bar{Y} - X\mathrm{B}\right\|_{\Sigma^{-1}}^2 - \log\det(\Sigma^{-1}) + \lambda\left\|\mathrm{B}\right\|_1 + \mu\left\|\Sigma^{-1}\right\|_1 \ . \tag{76}$$

Problem (76) is not convex, but can be solved heuristically (see Rothman et al. 2010 for details) by coordinate descent doing soft-tresholdings for the udpdates in $\mathrm{B}_{j:}$'s and solving a Graphical Lasso (Friedman et al., 2008) for the update in Σ^{-1}. The ℓ_1-norm being not well suited for our problem, we introduce an $\ell_{2,1}$ version of MRCE.

C.3.2 Multivariate Regression with Covariance Estimation with $l_{2,1}$-norm ($\ell_{2,1}$-MRCE)

The ℓ_1-norm penalization on the regression penalization B being not well suited for our problem, one can think to an $\ell_{2,1}$-penalized version of MRCE defined as follow:

$$(\hat{\mathrm{B}}^{\ell_{2,1}\mathrm{MRCE}}, \hat{\Sigma}^{\ell_{2,1}\mathrm{MRCE}}) \in \underset{\substack{\mathrm{B}\in\mathbb{R}^{p\times q}\\ \Sigma^{-1}\succ 0}}{\arg\min} \left\|\bar{Y} - X\mathrm{B}\right\|^2_{\Sigma^{-1}} - \log\det(\Sigma^{-1}) + \lambda \left\|\mathrm{B}\right\|_{2,1} + \mu \left\|\Sigma^{-1}\right\|_1 \quad . \tag{77}$$

In order to combine $\ell_{2,1}$-MRCE to take take advantage of all the repetitions, one can think of the following estimator:

$$(\hat{\mathrm{B}}^{\ell_{2,1}\mathrm{MRCER}}, \hat{\Sigma}^{\ell_{2,1}\mathrm{MRCER}}) \in \underset{\substack{\mathrm{B}\in\mathbb{R}^{p\times q}\\ \Sigma^{-1}\succeq 0}}{\arg\min} \sum_1^r \left\|Y^{(l)} - X\mathrm{B}\right\|^2_{\Sigma^{-1}} - \log\det(\Sigma^{-1}) + \lambda \left\|\mathrm{B}\right\|_{2,1} + \mu \left\|\Sigma^{-1}\right\|_1 \quad . \tag{78}$$

As for Appendix C.3.1, Problem (77) (*resp.* (78)) can be heuristically solved through coordinate descent.

Update in $\mathrm{B}_{j:}$ It is the same as $\ell_{2,1}$-MLE and $\ell_{2,1}$-MLER:

$$\mathrm{B}_{j:} = \mathrm{BST}\left(\mathrm{B}_{j:} + \frac{X_{:j}^\top \Sigma^{-1}(Y - X\mathrm{B})}{\|X_{:j}\|^2_{S^{-1}}}, \frac{\lambda nq}{\|X_{:j}\|^2_{S^{-1}}}\right) \quad . \tag{79}$$

Update in Σ^{-1} Minimizing (77) in Σ^{-1} amounts to solve:

$$\mathrm{glasso}(\Sigma, \mu) \triangleq \underset{\Sigma^{-1}\succ 0}{\arg\min} \left\langle \Sigma^{\mathrm{EMP}}, \Sigma^{-1}\right\rangle - \log\det(\Sigma^{-1}) + \mu \left\|\Sigma^{-1}\right\|_1 \quad . \tag{80}$$

This is a well known and well studied problem (Friedman et al., 2008) that can be solved through coordinate descent. For ourselves we used the `scikit-learn` (Pedregosa et al., 2011) implementation of the Graphical Lasso. Note that applying the Graphical Lasso on very ill conditioned empirical covariance matrix such as Σ^{EMP} is very long. We thus only considered $\ell_{2,1}$-MRCER were the Graphical Lasso is applied on $\Sigma^{\mathrm{EMP},r}$.

C.4 Algorithms summary

Each estimator, proposed or compared to is based on an optimization problem to solve. Each optimization problem is solve with block coordinate descent, whether there is theoretical guarantees for it to converge toward a global minimum (for convex formulations, CLaR, SGCL and MTL), or not (for non-convex formulations, $\ell_{2,1}$-MLE, $\ell_{2,1}$-MLER, $\ell_{2,1}$-MRCER). The cost for the updates for each algorithm can be found in Table 1. The formula for the updates in $\mathrm{B}_{j:}$'s and S/Σ for each algorithm can be found in Table 2.

Let $T_{S\ \mathrm{update}}$ be the number of updates of B for one update of S or Σ.

Table 1 – Algorithms cost in time summary

	CD epoch cost	convex	dual gap cost
CLaR	$\mathcal{O}(\frac{n^3+qn^2}{T_{S\ \mathrm{update}}} + pn^2 + pnq)$	yes	$\mathcal{O}(rnq + p)$
SGCL	$\mathcal{O}(\frac{n^3+qn^2}{T_{S\ \mathrm{update}}} + pn^2 + pnq)$	yes	$\mathcal{O}(nq + p)$
$\ell_{2,1}$-MLER	$\mathcal{O}(\frac{n^3+qn^2}{T_{S\ \mathrm{update}}} + pn^2 + pnq)$	no	not convex
$\ell_{2,1}$-MLE	$\mathcal{O}(\frac{n^3+qn^2}{T_{S\ \mathrm{update}}} + pn^2 + pnq)$	no	not convex
$\ell_{2,1}$-MRCER	$\mathcal{O}(\frac{\mathcal{O}(\mathrm{glasso})}{T_{S\ \mathrm{update}}} + pn^2 + pnq)$	no	not convex
MTL	$\mathcal{O}(npq)$	yes	$\mathcal{O}(nq + p)$

Recalling that $\Sigma^{\mathrm{EMP}} \triangleq \frac{1}{q}(\bar{Y} - X\mathrm{B})(\bar{Y} - X\mathrm{B})^\top$ and $\Sigma^{\mathrm{EMP},r} \triangleq \frac{1}{rq}\sum_{l=1}^r (Y^{(l)} - X\mathrm{B})(Y^{(l)} - X\mathrm{B})^\top$, a summary of the updates in S/Σ and $\mathrm{B}_{j:}$'s for each algorithm is given in Table 2.

Comments on Table 2 The updates in S/Σ and $\mathrm{B}_{j:}$'s are given in Table 2. Although the updates may look similar, all the algorithms can lead to very different results, see Figures 6, 9, 11 and 13.

Table 2 – Algorithms updates summary

	update in $\mathrm{B}_{j:}$	update in S/Σ
CLaR	$\mathrm{B}_{j:} = \mathrm{BST}\left(\mathrm{B}_{j:} + \frac{X_{:j}^\top S^{-1}(Y - X\mathrm{B})}{\|X_{:j}\|_{S^{-1}}^2}, \frac{\lambda n q}{\|X_{:j}\|_{S^{-1}}^2}\right)$	$S = \mathrm{ClSqrt}(\Sigma^{\mathrm{EMP},r}, \underline{\sigma})$
SGCL	$\mathrm{B}_{j:} = \mathrm{BST}\left(\mathrm{B}_{j:} + \frac{X_{:j}^\top S^{-1}(Y - X\mathrm{B})}{\|X_{:j}\|_{S^{-1}}^2}, \frac{\lambda n q}{\|X_{:j}\|_{S^{-1}}^2}\right)$	$S = \mathrm{ClSqrt}(\Sigma^{\mathrm{EMP}}, \underline{\sigma})$
$\ell_{2,1}$-MLER	$\mathrm{B}_{j:} = \mathrm{BST}\left(\mathrm{B}_{j:} + \frac{X_{:j}^\top \Sigma^{-1}(Y - X\mathrm{B})}{\|X_{:j}\|_{\Sigma^{-1}}^2}, \frac{\lambda n q}{\|X_{:j}\|_{\Sigma^{-1}}^2}\right)$	$\Sigma = \mathrm{Cl}(\Sigma^{\mathrm{EMP},r}, \underline{\sigma}^2)$
$\ell_{2,1}$-MLE	$\mathrm{B}_{j:} = \mathrm{BST}\left(\mathrm{B}_{j:} + \frac{X_{:j}^\top \Sigma^{-1}(Y - X\mathrm{B})}{\|X_{:j}\|_{\Sigma^{-1}}^2}, \frac{\lambda n q}{\|X_{:j}\|_{\Sigma^{-1}}^2}\right)$	$\Sigma = \mathrm{Cl}(\Sigma^{\mathrm{EMP}}, \underline{\sigma}^2)$
$\ell_{2,1}$-MRCER	$\mathrm{B}_{j:} = \mathrm{BST}\left(\mathrm{B}_{j:} + \frac{X_{:j}^\top \Sigma^{-1}(Y - X\mathrm{B})}{\|X_{:j}\|_{\Sigma^{-1}}^2}, \frac{\lambda n q}{\|X_{:j}\|_{\Sigma^{-1}}^2}\right)$	$\Sigma = \mathrm{glasso}(\Sigma^{\mathrm{EMP},r}, \mu)$
MTL	$\mathrm{B}_{j:} = \mathrm{BST}\left(\mathrm{B}_{j:} + \frac{X_{:j}^\top (Y - X\mathrm{B})}{\|X_{:j}\|^2}, \frac{\lambda n q}{\|X_{:j}\|^2}\right)$	no update in S/Σ

D Supplementary experiments

In this section we describe the preprocessing pipeline used for the realistic and real data (see Appendix D.1). We then propose time comparison for all the algorithms (see Appendix D.2). And finally we expose supplementary experiments on real data (see Appendix D.3 to D.4).

D.1 Preprocessing steps for realistic and real data

When using multi-modal data without whitening, one has to rescale properly data, indeed data needs to have the same order of magnitude, otherwise some mode (for example EEG data) could be (almost) completely ignored by the optimization algorithm. The preprocessing pipeline used to rescale realistic data (Figures 4 and 5) and real data (Figures 6, 9, 11 and 13) is described in Algorithm 2.

Algorithm 2 PREPROCESSING STEPS FOR REALISTIC AND REAL DATA

input : $X, Y^{(1)}, \ldots, Y^{(r)}$
// rescale each line of X
for $i = 1, \ldots, n$ **do**
  for $l = 1, \ldots, r$ **do**
    $Y_{i,:}^{(l)} \leftarrow Y_{i,:}^{(l)} / \|X_{i,:}\|$
  $X_{i,:} \leftarrow X_{i,:} / \|X_{i,:}\|$
// rescale each column of X
for $j = 1, \ldots, q$ **do**
  $X_{:,j} \leftarrow X_{:,j} / \|X_{:,j}\|$
return $X, Y^{(1)}, \ldots, Y^{(r)}$

D.2 Time comparison

The goal of this experiment is to show that our algorithm (CLaR) is as costly as a Multi-Task Lasso or other competitors (in the M/EEG context, *i.e.,* n not too large). The time taken by each algorithm to produce Figure 6 (real data, left auditory stimulations) is given in Figure 8. In this experiment the tolerance is set to tol=10^{-3}, the safe stopping criterion is duality gap $<$ tol (only available for convex optimization problems). The heuristic stopping criterion is "if the objective do not decrease enough anymore then stop" *i.e.,* if objective$(\mathrm{B}^{(t)}, \Sigma^{(t)}) -$ objective$(\mathrm{B}^{(t+1)}, \Sigma^{(t+1)}) <$ tol$/10$ then stop. The safe stopping criterion is only available for CLaR, SGCL and MTL (it takes too much time *i.e.,* more than 10min for SGCL to have a duality gap under the fixed tol, so we remove it).

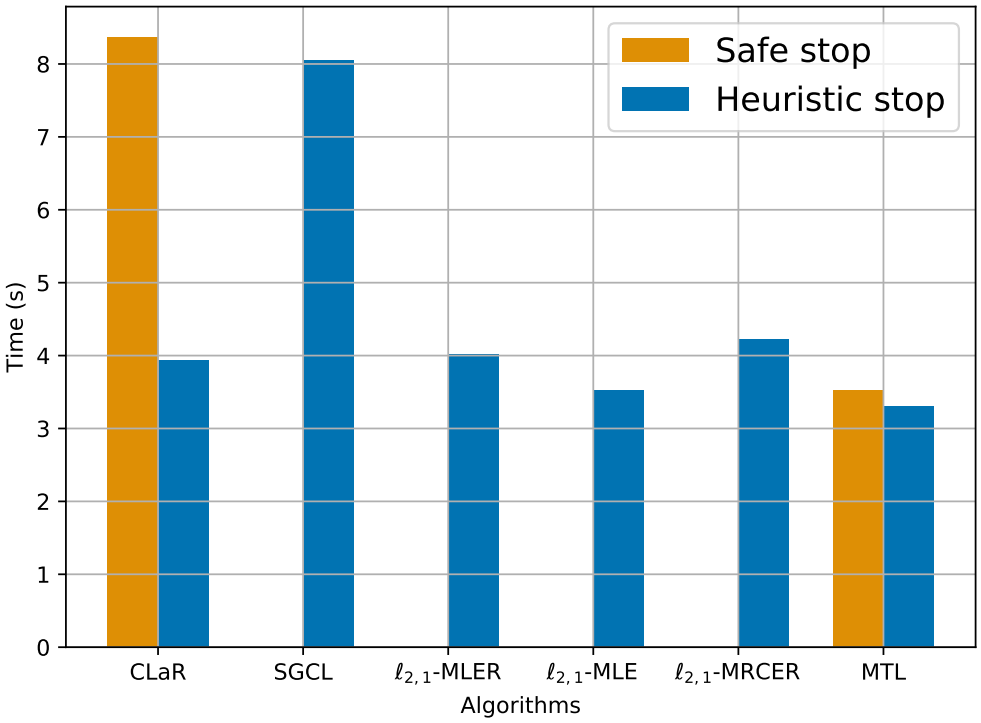
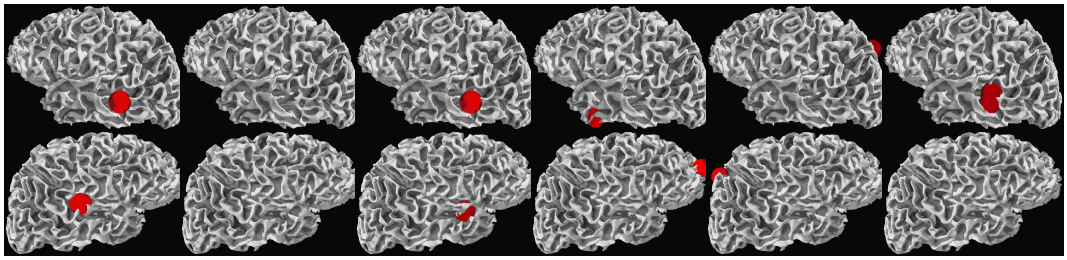


Figure 8 – *Time comparison, real data,* $n = 102$, $p = 7498$, $q = 54$, $r = 56$ Time for each algorithm to produce Figure 6.

Comment on Figure 8 Figure 8 shows that if we use the heuristic stopping criterion, CLaR is as fast the other algorithm. In addition CLaR has a safe stopping criterion which only take 2 to 3 more time than the heuristic one (less than 10sec).

D.3 Supplementary experiments on real data: right auditory stimulations

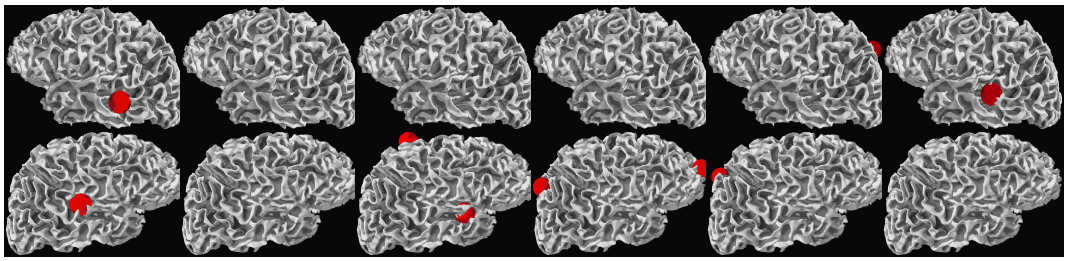


(a) CLaR (b) SGCL (c) $\ell_{2,1}$-MLER (d) $\ell_{2,1}$-MLE (e) $\ell_{2,1}$-MRCER (f) MTL

Figure 9 – *Real data (*$n = 102$, $q = 7498$, $q = 76$, $r = 65$*)* Sources found in the left hemisphere (top) and the right hemisphere (bottom) after right auditory stimulations.

Figures 9 and 10 show the solution given by each algorithm on real data after right auditory stimulations. As two sources are expected (one in each hemisphere, in bilateral auditory cortices), we vary λ by dichotomy between $\lambda_{\max}$ (returning 0 sources) and a $\lambda_{\min}$ (returning more than 2 sources), until finding a lambda giving exactly 2 sources. Figure 9 (*resp.* Figure 10) shows the solution given by the algorithms taking in account all the repetitions (*resp.* only half of the repetitions). When the number of repetitions is high (Figure 9) only CLaR and $\ell_{2,1}$-MLER find one source in each auditory cortices, MTL does find sources only in one hemisphere, all the other algorithms fail by finding sources not

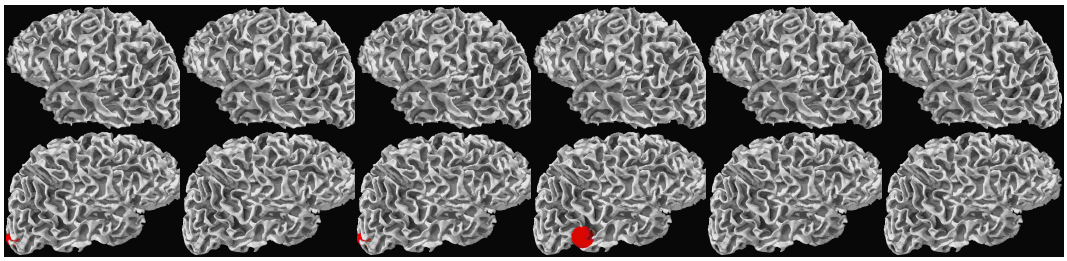
in the auditory cortices at all. Moreover when the number of repetitions is decreasing (Figure 10) $\ell_{2,1}$-MLER fails and only CLaR does find 2 sources, one in each hemisphere. Once again CLaR is more robust and performs better, even when the number of repetitions is low.



(a) CLaR (b) SGCL (c) $\ell_{2,1}$-MLER (d) $\ell_{2,1}$-MLE (e) $\ell_{2,1}$-MRCER (f) MTL

Figure 10 – *Real data ($n = 102$, $q = 7498$, $q = 76$, $r = 33$)* Sources found in the left hemisphere (top) and the right hemisphere (bottom) after right auditory stimulations.

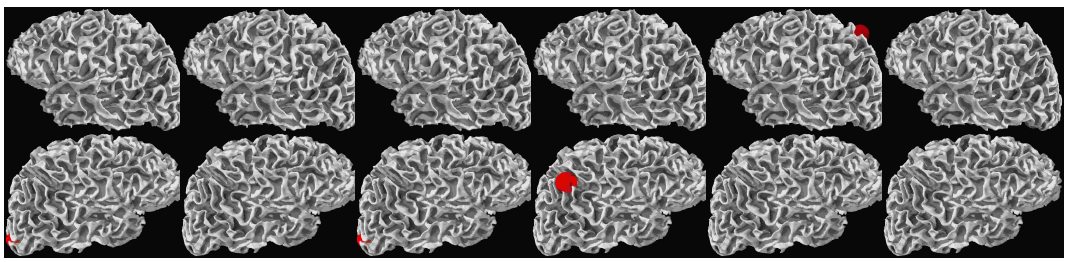
D.4 Supplementary experiments on real data: left visual stimulations



(a) CLaR (b) SGCL (c) $\ell_{2,1}$-MLER (d) $\ell_{2,1}$-MLE (e) $\ell_{2,1}$-MRCER (f) MTL

Figure 11 – *Real data ($n = 102$, $q = 7498$, $q = 48$, $r = 71$)* Sources found in the left hemisphere (top) and the right hemisphere (bottom) after left visual stimulations.

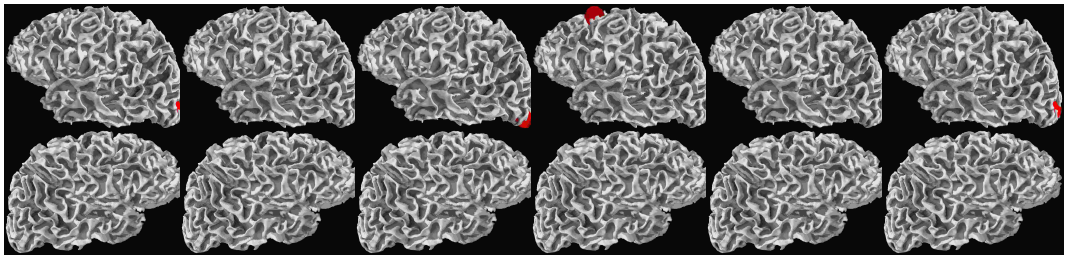
Figures 11 and 12 show the results for each algorithm after left visual stimulations. As one source is expected (in the right hemisphere), we vary λ by dichotomy between $\lambda_{\max}$ (returning 0 sources) and a $\lambda_{\min}$ (returning more than 1 sources), until finding a lambda giving exactly 1 source. When the number of repetitions is high (Figure 11) only CLaR and $\ell_{2,1}$-MLER do find a source in the visual cortex. When the number of repetitions decreases, CLaR and $\ell_{2,1}$-MLER still find one source in the visual cortex, other algorithms fail. This highlights this importance to take in account the repetitions.



(a) CLaR (b) SGCL (c) $\ell_{2,1}$-MLER (d) $\ell_{2,1}$-MLE (e) $\ell_{2,1}$-MRCER (f) MTL

Figure 12 – *Real data ($n = 102$, $q = 7498$, $q = 48$, $r = 36$)* Sources found in the left hemisphere (top) and the right hemisphere (bottom) after left visual stimulations.

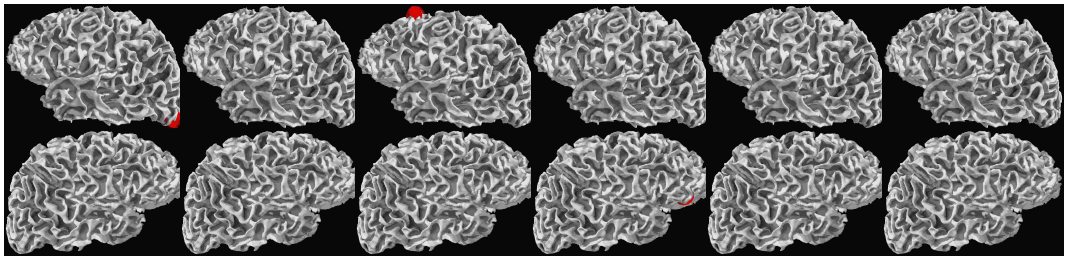
D.5 Supplementary experiments on real data: right visual stimulations



(a) CLaR (b) SGCL (c) $\ell_{2,1}$-MLER (d) $\ell_{2,1}$-MLE (e) $\ell_{2,1}$-MRCER (f) MTL

Figure 13 – *Real data ($n = 102$, $q = 7498$, $q = 48$, $r = 61$)* Sources found in the left hemisphere (top) and the right hemisphere (bottom) after right visual stimulations.

Figures 13 and 14 show the results for each algorithm after right visual stimulations. As one source is expected (in the left hemisphere), we vary λ by dichotomy between $\lambda_{\max}$ (returning 0 sources) and a $\lambda_{\min}$ (returning more than 1 sources), until finding a lambda giving exactly 1 source. When the number of repetitions is high (Figure 13) only CLaR, $\ell_{2,1}$-MLER and MTL do find a source in the visual cortex. When the number of repetitions decreases (Figure 14), only CLaR finds one source in the visual cortex, other algorithms fail. This highlights once again the robustness of CLaR, even with a limited number of repetitions.



(a) CLaR (b) SGCL (c) $\ell_{2,1}$-MLER (d) $\ell_{2,1}$-MLE (e) $\ell_{2,1}$-MRCER (f) MTL

Figure 14 – *Real data ($n = 102$, $q = 7498$, $q = 48$, $r = 31$)* Sources found in the left hemisphere (top) and the right hemisphere (bottom) after right visual stimulations.