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

Quentin Bertrand, Mathurin Massias, Alexandre Gramfort, Joseph Salmon

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

Quentin Bertrand [*1] **Mathurin Massias** [*1] **Alexandre Gramfort** [1] **Joseph Salmon** [2]

Abstract

Sparsity promoting norms are frequently used in high dimensional regression. A limitation of Lasso-type estimators is that the regularization parameter depends on the noise level which varies between datasets and experiments. Estimators such as the concomitant Lasso address this dependence by jointly estimating the noise level and the regression coefficients. As sample sizes are often limited in high dimensional regimes, simplified heteroscedastic models are customary. However, in many experimental applications, data is obtained by averaging multiple measurements. This helps reducing the noise variance, yet it dramatically reduces sample sizes, preventing refined noise modeling. In this work, we propose an estimator that can cope with complex heteroscedastic noise structures by using non-averaged measurements and a concomitant formulation. The resulting optimization problem is convex, so thanks to smoothing theory, it is amenable to state-of-the-art proximal coordinate descent techniques that can leverage the expected sparsity of the solutions. Practical benefits are demonstrated on simulations and on neuroimaging applications.

1. Introduction

In many important 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 high dimension 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.

[*] These authors contributed equally. [1]INRIA, Université Paris-Saclay [2]IMAG,Univ Montpellier, CNRS, Montpellier, France. Correspondence to: Quentin Bertrand <firstname.lastname@inria.fr>, Mathurin Massias <firstname.lastname@inria.fr>.

February 7, 2019

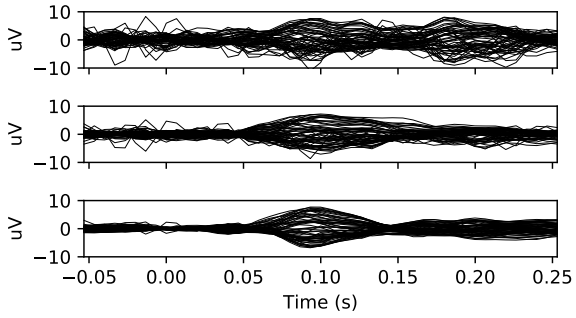

Figure 1: Amplitude $\bar{Y}$ of $n = 59$ EEG signals, averaged across $r = 5$ (top), $r = 10$ (middle), and $r = 50$ (bottom) repetitions. As the number of averaged repetitions increases, the noise is reduced and the measurements become smoother, revealing the brain response around 0.1 s.

Unfortunately, statistical analysis reveals that the optimal λ should be proportional to the noise level (Bickel et al., 2009), which is rarely known in practice. To tackle this issue, one can jointly estimate the noise level and the regression coefficients. Such a concomitant estimation (Huber and Dutter, 1974; Huber, 1981) has recently been adapted for sparse regression by Owen (2007) and analyzed under several names such as Square root Lasso (Belloni et al., 2011) or scaled Lasso (Sun and Zhang, 2012).

In these latter works, the noise parameter consists of a single variance parameter. However, in various applied settings, mixing data of different natures or coming from different sources is customary to increase the number of observations. This often leads to heteroscedasticity: the data may be contaminated with non-uniform noise levels (differing across features or samples). This is the case for magneto-electroencephalographic (M/EEG) data, where observations come from three different types of sensors (gradiometers, magnetometers and electrodes), leading to very different amplitudes, noise levels and noise covariance matrices. Attempts to cope with heteroscedasticity were analyzed in this context by Daye et al. (2012); Wagener and Dette (2012); Kolar and Sharpnack (2012); Dalalyan et al. (2013). Moreover, fast algorithms relying on smoothing techniques from the optimization commu-

nity (Nesterov, 2005; Beck and Teboulle, 2012) have been extended to heteroscedastic regression in a multi-task setting, through the Smooth Generalized Concomitant Lasso (SGCL, Massias et al. (2018)). The SGCL is designed to jointly estimate the regression coefficients and the noise *co-standard deviation matrix*[1]. However, in certain applications, such as with M/EEG data, the number of parameters in the co-standard deviation matrix ($\approx 10^4$) is typically equal to the number of observations, making it statistically impossible to estimate accurately.

When observations are contaminated with a strong noise and the signal-to-noise ratio (SNR) is too low, provided measurements can be repeated, a natural idea is to average them. Indeed, under the assumption that the signal of interest is corrupted by some additive independant noise realizations, averaging different measurements divides the noise variance by the number of repetitions. This is classically done in experimental sciences from chemistry, to physics or neuroscience as it generally allows to visually inspect signals. This effect is illustrated in Figure 1 for electroencephalography (EEG) data. By averaging from 5 to 50 repetitions of the electrical response of the brain to a stimulus one can reveal a so-called evoked brain response around 100 ms after stimulation. It is usually this type of averaged data which is plugged into optimization solvers, hence discarding individual observations that could be used to better characterize the noise and improve the statistical estimation (Gramfort et al., 2013; Ou et al., 2009).

In this work, we propose the Concomitant Lasso with Repetitions (CLaR), an estimator designed to exploit all available measurements collected during repetitions of experiments. The proposed concomitant formulation of the optimization problem derived has two strong benefits: first, the noise covariance is an explicit parameter of the model, on which it is easy to add structural constraints (*e.g.,* block diagonality) and second, smoothing theory leads to a cost function that can be minimized using efficient proximal coordinate descent techniques. By estimating the regression coefficients and the noise structure, this estimator demonstrates improvements in suppport identification compared to estimators using averaged data or assuming homoscedastic noise.

In Section 2, we recall the framework of concomitant estimation, and introduce our estimator. 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

Notation For a matrix $A \in \mathbb{R}^{m\times n}$ its j^{th} column (*resp.* j^{th} row) is denoted $A_{:j} \in \mathbb{R}^{m\times 1}$ (*resp.* $A_{j:} \in \mathbb{R}^{1\times n}$). Let r be the number of repetitions of the experiment. The r observations matrices are denoted $Y^{(1)}, \dots, Y^{(r)} \in \mathbb{R}^{n\times q}$ with n being the number of sensors/samples and q corresponds, for example, to a number of tasks or a number of time samples. The mean over the repetitions of the observations 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 matrix, with p features stored column-wise: $X = [X_{:1}, \dots, X_{:p}]$. The matrix $\mathrm{B} \in \mathbb{R}^{p\times q}$ contains the coefficients of the linear regression model. We write $\|\cdot\|$ (*resp.* $\langle\cdot,\cdot\rangle$) for the standard 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.* $\|\mathrm{B}\|_{2,\infty} = \max_{j\in[p]} \|\mathrm{B}_{j:}\|$) is its row-wise norm, and for any $p_1 \in [0,\infty]$, we write $\|\mathrm{B}\|_{s,p_1}$ for the Schatten p_1-norm (*i.e.,* the ℓ_{p_1} norm of the singular values of B). The notation $\mathcal{S}_+^n$ (*resp.* $\mathcal{S}_{++}^n$) stands for the set of positive semi-definite matrices (*resp.* positive definite matrices). For S_1 and $S_2 \in \mathcal{S}_+^n$, $S_1 \succeq S_2$ if $S_1 - S_2 \in \mathcal{S}_+^n$, and $S_1 \succeq \underline{\sigma}$ if $S_1 \succeq \underline{\sigma}\,\mathrm{Id}_n$. 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, $\mathrm{BST}(x, \tau) = (1 - \tau/\|x\|)_+ x$.

For $p_1 \in [1,\infty)$, let us write $\mathcal{B}_{p_1}$ for the associated unit ℓ_{p_1} ball. The identity matrix of size $n \times n$ is denoted Id_n, and $[r]$ is the set of integers from 1 to r.

2.1. Model and proposed estimator

We consider observations where r repetitions of the same experiment are performed, leading to measurements $Y^{(1)} \in \mathbb{R}^{n\times q}, \dots, Y^{(r)} \in \mathbb{R}^{n\times q}$. We assume each measurement follows the same linear 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.* standard normal distributions, the $E^{(l)}$s are independent and $S^* \in \mathcal{S}_{++}^n$ is the *co-standard deviation matrix*, *i.e.,* the square root of the noise covariance matrix.[2] Note that even if the observations $Y^{(1)}, \dots, Y^{(r)}$ are different because of the noise $\mathrm{E}^{(1)}, \dots, \mathrm{E}^{(r)}$, the true parameter B^* and the noise structure S^* are shared across repetitions.

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

Definition 1 (Concomitant Lasso with Repetitions, CLaR)**.** CLaR estimates the parameters of (1) by solving

$$(\hat{\mathrm{B}}^{\mathrm{CLaR}}, \hat{\mathrm{S}}^{\mathrm{CLaR}}) \in \underset{\substack{\mathrm{B}\in\mathbb{R}^{p\times q}\\ S\succeq\underline{\sigma}}}{\arg\min}\, f(\mathrm{B}, S) + \lambda \|\mathrm{B}\|_{2,1} \ , \tag{2}$$

[1] *i.e.,* the square root of the noise covariance matrix

[2] since we impose $S^* \in \mathcal{S}_{++}^n$, it is uniquely defined

where $f(\mathrm{B}, S) := \frac{1}{2nqr} \sum_1^r \left\|Y^{(l)} - X\mathrm{B}\right\|_{S^{-1}}^2 + \frac{1}{2n} \operatorname{Tr}(S)$, $\lambda > 0$ 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 previous estimator

In low SNR settings, the 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 $\tilde{S}^* := \frac{1}{\sqrt{r}} S^*$ and $\tilde{\mathrm{E}}$ has $i.i.d.$ entries drawn from a standard normal distribution. The SNR of the average is multiplied by $\sqrt{r}$, yet the number of available samples to characterize the noise goes from rq to q. This explains why averaging is not sufficient to estimate complex heteroscedastic noise models, *i.e.,* when S^* is far from a scalar matrix. Our introduced CLaR is a generalization of the Smoothed Generalized Concomitant Lasso (Massias et al., 2018) which only targets averaged observations:

Definition 2 (Smoothed Generalized Concomitant Lasso, SGCL)**.** SGCL estimates the parameters of model (3), by solving:

$$(\hat{\mathrm{B}}^{\mathrm{SGCL}}, \hat{\mathrm{S}}^{\mathrm{SGCL}}) \in \underset{\substack{\mathrm{B} \in \mathbb{R}^{p \times q} \\ \tilde{S} \succeq \underline{\sigma}/\sqrt{r}}}{\arg\min} \tilde{f}(\mathrm{B}, \tilde{S}) + \lambda \left\|\mathrm{B}\right\|_{2,1} \ , \tag{4}$$

with $\tilde{f}(\mathrm{B}, \tilde{S}) := \frac{1}{2nq} \|\bar{Y} - X\mathrm{B}\|_{\tilde{S}^{-1}}^2 + \frac{1}{2n} \operatorname{Tr}(\tilde{S})$.

Remark 1. 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 $\hat{\mathrm{S}}^{\mathrm{CLaR}} \succeq \underline{\sigma}$, we adapt the scaling so that $\hat{\mathrm{S}}^{\mathrm{SGCL}} \succeq \underline{\sigma}/\sqrt{r}$ in (4) for future comparisons.

Remark 2. SGCL is a particular case of CLaR: for $r = 1$ and $Y^{(1)} = \bar{Y}$, the solution of CLaR is the same as the one of SGCL.

The justification for the introduction of CLaR is the following: using the quadratic loss $\|Y - X\mathrm{B}\|^2$, the parameters of model (1) can be estimated by using either $\|\bar{Y} - X\mathrm{B}\|^2$ or $\frac{1}{r}\sum \|Y^{(l)} - X\mathrm{B}\|^2$ as a data-fitting term. It turns out that in such a case, the two alternatives yield the same solutions (as the two terms are equal up to constants in B). Hence, apart from averaging, the quadratic loss does not leverage the multiple repetitions, and ignores the noise structure. Using the data-fitting term of CLaR allows to incorporate this structure.

2.3. Smoothing of the nuclear norm

In this section we shed some light on the properties of CLaR, especially with respect to smoothing theory (Nesterov, 2005; Beck and Teboulle, 2012). The following proposition relates the data-fitting term used in CLaR and SGCL showing the link with the smoothing of the Schatten 1-norm (*a.k.a.* the trace norm or the nuclear norm). For that, we introduce the following smoothing function:

$$\omega_{\underline{\sigma}}(\cdot) = \tfrac{\underline{\sigma}}{2} \left\|\cdot\right\|_F^2 + \underline{\sigma} \tfrac{n \wedge q}{2}, \tag{5}$$

and the inf-convolution of functions f and g, defined as $f \,\square\, g(y) = \inf_x f(x) + g(y - x)$.

Proposition 1. The $\omega_{\underline{\sigma}}$-smoothing of the Schatten-1 norm, *i.e.,* the function $\left\|\cdot\right\|_{s,1} \,\square\, \omega_{\underline{\sigma}} : \mathbb{R}^{n \times q} \mapsto \mathbb{R}$, is the solution of the following (smooth) optimization problem:

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

Proof of Proposition 1 is in Appendix A.3.

Definition 3 (Clipped Square Root)**.** For $S \in \mathcal{S}_+^n$, let us define the *Clipped Square Root* operator:

$$\mathrm{ClSqrt}(S, \underline{\sigma}) = U \operatorname{diag}(\sqrt{\gamma_1} \vee \underline{\sigma}, \dots, \sqrt{s}_n \vee \underline{\sigma}) U^\top \ , \tag{7}$$

where $S = U \operatorname{diag}(\gamma_1, \dots, \gamma_n) U^\top$ and U is orthogonal.

Remark 3. Note that $\mathrm{ClSqrt}(S, \underline{\sigma})$ is the projection of the square root of S onto the affine cone $\{S \in \mathcal{S}_{++}^n : S \succeq \underline{\sigma}\}$.

We can now state explicitly the connection between the SGCL, CLaR and the Schatten 1-norm.

Proposition 2 (Smoothing properties of CLaR)**.** The solution of the CLaR problem in Equation (2), $(\hat{\mathrm{B}}, \hat{S}) = (\hat{\mathrm{B}}^{\mathrm{CLaR}}, \hat{S}^{\mathrm{CLaR}})$ is a solution of:

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

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

Proof of Proposition 2 can be found in Appendix B.1.

Remark 4. Following Remark 1, similar properties can be obtained for $\hat{\mathrm{B}}^{\mathrm{SGCL}}$ and $\hat{S}^{\mathrm{SGCL}}$ letting $r = 1$, and substituting $\bar{Z} := \frac{\bar{Y} - X\mathrm{B}}{\sqrt{q}}$ to Z in the former proposition.

Ideas similar to the ones of Propositions 1 and 2 can be traced to van de Geer (2016, Lemma 3.4, p. 37), where the following formulation was introduced to prove oracle inequalities for the multivariate square-root Lasso[3]:

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

In the present contribution, the problem formulation in Proposition 1 is motivated by computational aspects, as it helps to address the combined non-differentiability of the data-fitting term $\left\|\cdot\right\|_{s,1}$ and the penalty $\left\|\cdot\right\|_{2,1}$ term.

[3] defined as the solution of Equation (9) with $p_1 = 1$

Other alternatives to exploit the multiple repetitions without simply averaging them, would consist in investigating other Schatten p_1-norms:

$$\arg\min_{\mathrm{B}} \tfrac{1}{\sqrt{rq}} \|[Y^{(1)} - X\mathrm{B}| \dots |Y^{(r)} - X\mathrm{B}]\|_{s,p_1} + \lambda n \|\mathrm{B}\|_{2,1} \quad , \tag{9}$$

Without smoothing, problems of the form given in Equation (9) have the drawback of having no smooth term, and solvers have to resort to proximal splitting algorithms (*e.g.,* the ones by Douglas and Rachford (1956) or Chambolle and Pock (2011)) that can handle the sum of two non-smooth components.

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 by Ndiaye et al. (2017) for the concomitant Lasso and leads to the introduction of smoothing techniques to address it.

Here we replaced the Schatten norm with $p_1 = 1$ by its smoothed version $\|\cdot\|_{s,p_1} \square \omega_{\underline{\sigma}}$, for some smooth function $\omega_{\underline{\sigma}}$. Results for other Schatten p_1-norms are provided in the Appendix; see Proposition 11 (*resp.* Proposition 12) for the case of the Schatten 2-norm (*resp.* Schatten ∞-norm).

3. Properties of CLaR

We detail the principal results needed to solve Problem (2) numerically in this section, leading to the efficient implementation proposed in Algorithm 1. We first recall some technical results from alternate minimization to optimize composite problems.

3.1. Alternate minimization

Proposition 3. CLaR is jointly convex in B and S so minimizing the objective alternatively in S and in B (see Algorithm 1) converges to a global minimum.

Proof.

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

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

$(Z, \Sigma) \mapsto \operatorname{Tr} Z^\top \Sigma^{-1} Z$ is jointly convex over $\mathbb{R}^{n\times q} \times \mathbb{S}_{++}^n$, 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 $f + \lambda g$ is jointly convex in (B, S). Moreover the constraint set is convex and thus solving CLaR is a convex problem. □

As the next proposition shows, optimizing over S with B being fixed involves the Clipped Square Root operator of Definition 3.

Proposition 4 (Minimization in S)**.** Let $B \in \mathbb{R}^{n\times q}$ be fixed. The minimization of $f(\mathrm{B}, S)$ *w.r.t.* S with the constraint $S \succeq \underline{\sigma}$ admits the closed-form solution:

$$S = \mathrm{ClSqrt}\left(\frac{1}{r} \sum_{l=1}^r Z^{(l)} Z^{(l)\top}, \underline{\sigma} \right) \quad , \tag{10}$$

with $Z^{(l)} = \frac{1}{\sqrt{q}}(Y^{(l)} - X\mathrm{B})$.

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

$$\arg\min_{S \succeq \underline{\sigma}} \tfrac{1}{2} \|Z\|_{S^{-1}}^2 + \tfrac{1}{2} \operatorname{Tr}(S) \quad . \tag{11}$$

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

We can now state the update of the B term in the alternate minimization process. Here, because of the row-wise separability of $\|\cdot\|_{2,1}$, we use block updates over rows, whose closed-form are provided by the next proposition:

Proposition 5. 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:

$$\mathrm{B}_{j:} = \mathrm{BST}\left(\mathrm{B}_{j:} + \frac{X_{:j}^\top S^{-1}(Y - X\mathrm{B})}{\|X_{:j}\|_{S^{-1}}^2}, \frac{\lambda nq}{\|X_{:j}\|_{S^{-1}}^2} \right) \quad . \tag{12}$$

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 can be computed:

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

Hence, proximal block-coordinate descent converges (Tseng and Yun, 2009), and the update reads like Equation (12). The closed-form formula arises since the smooth part of the objective is quadratic and isotropic *w.r.t.* $\mathrm{B}_{j:}$. □

3.2. Critical parameter, duality gap and stopping criterion

As for the Lasso we show that there is a critical parameter, *i.e.,* there exists $\lambda_{\max} \geq 0$ such that whenever λ is greater than this value, the coefficients recovered vanish:

Proposition 6 (Critical regularization parameter)**.** For the CLaR estimator we have the following property: with $S_{\max} = \mathrm{ClSqrt}(\frac{1}{qr}\sum_{l=1}^{r} Y^{(l)}Y^{(l)\top}, \underline{\sigma})$,

$$\hat{\mathrm{B}} = 0, \quad \forall \lambda \geq \lambda_{\max} := \frac{1}{nq}\left\|X^\top S_{\max}^{-1}\bar{Y}\right\|_{2,\infty} \quad . \tag{13}$$

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

Fermat's rules states that

$$\begin{aligned}\hat{\mathrm{B}} = 0 &\Leftrightarrow 0 \in \partial\left(f(\cdot, S_{\max}) + \lambda\|\cdot\|_{2,1}(0)\right)\\ &\Leftrightarrow -\nabla f(\cdot, S_{\max}) \in \lambda\mathcal{B}_{\|\cdot\|_{2,\infty}}\\ &\Leftrightarrow \frac{1}{nq}\left\|X^\top S_{\max}^{-1}\bar{Y}\right\|_{2,\infty} := \lambda_{\max} \leq \lambda \quad . \end{aligned} \tag{14}$$

□

To ensure the convergence of our algorithm, we use the duality gap as a stopping criterion (as it guarantees a targeted sub-optimality level). For its computation we therefore need to explicitely state the dual optimization problem of Problem (2):

Proposition 7. With $\hat{\Theta} = (\hat{\Theta}^{(1)}, \dots, \hat{\Theta}^{(r)})$, the dual formulation of Problem (2) is

$$\begin{aligned}\hat{\Theta} = \underset{(\Theta^{(1)},\dots,\Theta^{(r)})\in\Delta_{X,\lambda}}{\arg\max} &\frac{\underline{\sigma}}{2}\left(1 - \frac{qn\lambda^2}{r}\sum_{l=1}^{r}\mathrm{Tr}\,\Theta^{(l)}\Theta^{(l)\top}\right)\\ &+ \frac{\lambda}{r}\sum_{l=1}^{r}\left\langle\Theta^{(l)}, Y^{(l)}\right\rangle \quad , \end{aligned} \tag{15}$$

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

$$\begin{aligned}\Delta_{X,\lambda} = \Big\{&(\Theta^{(1)}, \dots, \Theta^{(r)}) \in (\mathbb{R}^{n\times q})^r :\\ &\left\|X^\top\bar{\Theta}\right\|_{2,\infty} \leq 1, \left\|\sum_{l=1}^{r}\Theta^{(l)}\Theta^{(l)\top}\right\|_2 \leq \frac{r}{\lambda^2 n^2 q}\Big\} \quad . \end{aligned} \tag{16}$$

Proof of Proposition 7 is in Appendix B.2.

In Algorithm 1 the dual point Θ at iteration t is obtained through the relations $\Theta^{(l)} = \frac{1}{nq\lambda}(Y^{(l)} - X\mathrm{B})$ (with B the current primal iterate), and then projected on $\Delta_{X,\lambda}$.

Remark 5. Once the quantity $\mathrm{cov}_Y := \frac{1}{r}\sum_1^r Y^{(l)}Y^{(l)\top}$ has been pre-computed, the cost of updating S in CLaR does not depend on r, hence is the same as working with averaged data. Indeed, R being defined as $R = [Y^{(1)} - X\mathrm{B}|\dots|Y^{(r)} - X\mathrm{B}]$, computing $RR^\top = \mathrm{RRT}(\mathrm{cov}_Y, Y, X, \mathrm{B}) := r\mathrm{cov}_Y + r(X\mathrm{B})(X\mathrm{B})^\top - r\bar{Y}^\top(X\mathrm{B}) - r(X\mathrm{B})^\top\bar{Y}$ can be done in $\mathcal{O}(qn^2)$ (details are in Appendix B.3).

Algorithm 1 ALTERNATE MIN. FOR CLAR

input : $X, \bar{Y}, \underline{\sigma}, \lambda, f, T$
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 $= 1 \pmod f$ **then** `// noise update`
    $RR^\top = \mathrm{RRT}(\mathrm{cov}_Y, Y, X, \mathrm{B})$ `// Rk. (5)`
    $S \leftarrow \mathrm{ClSqrt}(\frac{1}{qr}RR^\top, \underline{\sigma})$ `// 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`
    $\bar{R} \leftarrow \bar{R} + X_{:j}\mathrm{B}_{j:}$
    $\mathrm{B}_{j:} \leftarrow \mathrm{BST}\Big(\frac{X_{:j}^\top S^{-1}\bar{R}}{L_j}, \frac{\lambda nq}{L_j}\Big)$
    $\bar{R} \leftarrow \bar{R} - X_{:j}\mathrm{B}_{j:}$
return B, S

3.3. Statistical comparison

In this section, 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 2).

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 some more interesting statistical properties with respect to one obtained by 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:

$$\hat{\Sigma}^{\mathrm{CLaR}} := \frac{1}{qr}\sum_{l=1}^{r}(Y^{(l)} - X\hat{\mathrm{B}})(Y^{(l)} - X\hat{\mathrm{B}})^\top \quad , \tag{17}$$

$$\hat{\Sigma}^{\mathrm{SGCL}} := \frac{1}{qr}\Big(\sum_{l=1}^{r}Y^{(l)} - X\hat{\mathrm{B}}\Big)\Big(\sum_{l=1}^{r}Y^{(l)} - X\hat{\mathrm{B}}\Big)^\top, \tag{18}$$

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

Proposition 8. 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 (17) and (18), then one can check

Table 1: Algorithms cost

	CD epoch	gap computation
CLaR	$\mathcal{O}(\frac{n^3+qn^2}{f} + pn^2 + pnq)$	$\mathcal{O}(rnq + p)$
SGCL	$\mathcal{O}(\frac{n^3+qn^2}{f} + pn^2 + pnq)$	$\mathcal{O}(nq + p)$
MTLR	$\mathcal{O}(npqr)$	$\mathcal{O}(rnq + p)$
MTL	$\mathcal{O}(npq)$	$\mathcal{O}(nq + p)$

that

$$\mathbb{E}(\hat{\Sigma}^{\mathrm{CLaR}}) = \mathbb{E}(\hat{\Sigma}^{\mathrm{SGCL}}) = \Sigma^* \ , \tag{19}$$

$$\mathrm{cov}(\hat{\Sigma}^{\mathrm{CLaR}}) = \frac{1}{r}\mathrm{cov}(\hat{\Sigma}^{\mathrm{SGCL}}) \ . \tag{20}$$

Proof of Proposition 8 can be found in Appendix B.4

Proposition 8 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[4], 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.

4. Experiments

The code is released as an open source package and can be found here `https://github.com/QB3/CLaR`. The implementation is in Python, with Numba compilation (Lam et al., 2015) to increase the speed of the algorithm.

Comparison with other estimators We compare CLaR to other estimators: SGCL, the Multi-Task Lasso (MTL, Obozinski et al. 2010) and a version of the MTL with repetitions: MTLR. Recall that CLaR solves Problem (2) and that SGCL solves Problem (4); we also remind the definition of MTL and MTLR:

$$\hat{\mathrm{B}}^{\mathrm{MTL}} \in \underset{\mathrm{B}\in\mathbb{R}^{p\times q}}{\arg\min} \frac{1}{2nq} \left\|\bar{Y} - X\mathrm{B}\right\|^2 + \lambda \left\|\mathrm{B}\right\|_{2,1} \ . \tag{21}$$

With $Y = [Y^{(1)}|\dots|Y^{(r)}] \in \mathbb{R}^{n\times rq}$, MTLR first solves:

$$\hat{\mathrm{B}}^{C} \in \underset{\mathrm{B}\in\mathbb{R}^{p\times rq}}{\arg\min} \frac{1}{2nq} \|Y - X\mathrm{B}\|^2 + \lambda \left\|\mathrm{B}\right\|_{2,1} \ , \tag{22}$$

then the estimator is defined as $\hat{\mathrm{B}}^{\mathrm{MTLR}} = \frac{1}{r}\sum_{l=1}^{r} \hat{\mathrm{B}}^{(l)}$, with $\hat{\mathrm{B}}^{C} = [\hat{\mathrm{B}}^{(1)}|\dots|\hat{\mathrm{B}}^{(r)}]$.

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

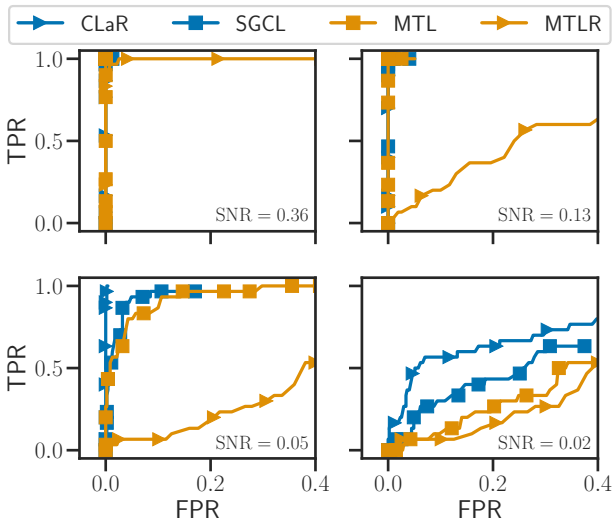

Figure 2: ROC curves of true support recovery for the CLaR, SGCL, MTL and MTLR with $\rho_X = 0.6$, $\rho_S = 0.6$, $r = 50$ for different values of SNR.

The cost of an epoch of coordinate descent and the cost of computing the gap for each for each algorithm are summarized in Table 1.

4.1. Support recovery

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

$$\mathrm{SNR} := \|X\mathrm{B}^*\|/\|X\mathrm{B}^* - Y^{(l)}\| \ . \tag{23}$$

For Figures 2 to 5, fhe figure of merit is the ROC curve, *i.e.,* the true positive rate against the false positive rate. For the four estimators, the ROC curve is obtained by varying the value of the regularization parameter λ on a geometric grid of 160 points, starting from $\lambda_{\max}$ (specific to each algorithm) to $\lambda_{\min}$, the latest being also specific and chosen to obtain large enough false positive rates.

SNR influence On Figure 2 we can see that when the SNR is high (top left), all curves reach the (0, 1) point. This means that for each algorithm, there exists a λ such that the estimated support is exactly the true one. However, when the SNR decreases (top right, bottom left), the performance of SGCL, MTL and MTLR starts to drop, while

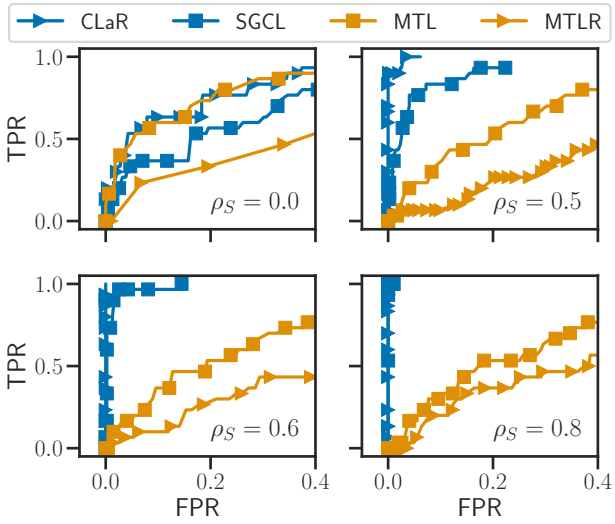

Figure 3: ROC curves of true support recovery for the CLaR, SGCL, MTL and MTLR with $\rho_X = 0.6$, SNR $= 0.03$, $r = 50$ for different values of ρ_S.

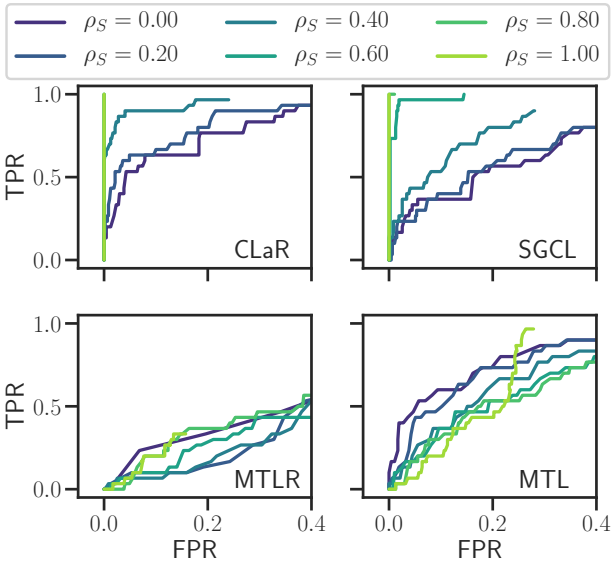

Figure 4: ROC curves of true support recovery for the CLaR, SGCL, MTL and MTLR with $\rho_X = 0.6$, SNR $= 0.03$, $r = 50$ for different values of ρ_S.

that of CLaR remains stable. Finally, when the SNR is too low (bottom right), all algorithms perform poorly, but CLaR still performs better. This highlights the capacity of CLaR to leverage multiple repetitions of measurements to finely estimate the heteroscedastic noise.

Noise structure influence Figure 3 represents the ROC curves of CLaR, SGCL, MTL and MTLR for different values of ρ_S. As ρ_S increases, the noise becomes less and less heteroscedastic: from top left to bottom right, the performance of CLaR and SGCL increases as they are designed to exploit correlations in the noise, while the performance of MTL and MTLR decreases, as their homoscedastic model becomes less and less valid.

The same idea is presented in a different manner in Figure 4 which shows the cases where CLaR performs well and poorly. For instance when the structure of the correlation matrix of the noise is close to identity, CLaR performs poorly. This a limitation of CLaR (and SGCL) as if the noise turns out to be homoscedastic, CLaR just adds $n^2/2$ model-useless parameters that are likely to fit the noise.

Influence of the number of repetitions Figure 5 shows ROC curves of CLaR, SGCL, MTL and MTLR for different values of r, starting from $r = 1$, where CLaR and SGCL are equivalent (as well as MTL and MTLR), to $r = 100$. It can be seen that even with $r = 20$ CLaR outperforms the other estimators, and that with $r = 100$ CLaR benefits more than any of the other algorithms from the large number of repetitions.

4.2. Real data

We now evaluate our estimators on real magneto and electroencephalography (M/EEG) data. The M/EEG recordings measure the electrical potential and magnetic field induced by the active neurons. Data are time series of length q with n sensors and p sources (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 $Y^{(1)}, \dots, Y^{(r)}$ and the amplitudes of sources in the brain B^* is linear. The M/EEG inverse problem consists in identifying B^*. Because of the limited number of sensors (a few hundred in practice), as well as the physics of the problem, the M/EEG inverse problem is severely ill-posed and needs regularization that provides plausible biological solutions. Because the experiments are usually short (less than 1 s.) and focused on specific cogntive functions, the number of active sources is expected to be small, *i.e.,* B^* is assumed to be row-sparse. Thus the M/EEG inverse problem fits the framework of Section 2.

We use the *sample* dataset from MNE (Gramfort et al., 2014). The experimental conditions are here auditory stimulations in the right or left ears or visual stimulations in the right or left visual fields, leading to different active locations in the brain (*i.e.,* different B^* for each type of stimulation). For this experiment, we keep only the gradiometer magnetic channels, and we reject one of them due to strong artifacts. This leads to 203 gradiometers, *i.e.,* $n = 203$ signals. The length of the temporal series is $q = 30$, and the data contains $r = 50$ repetitions. We choose a source

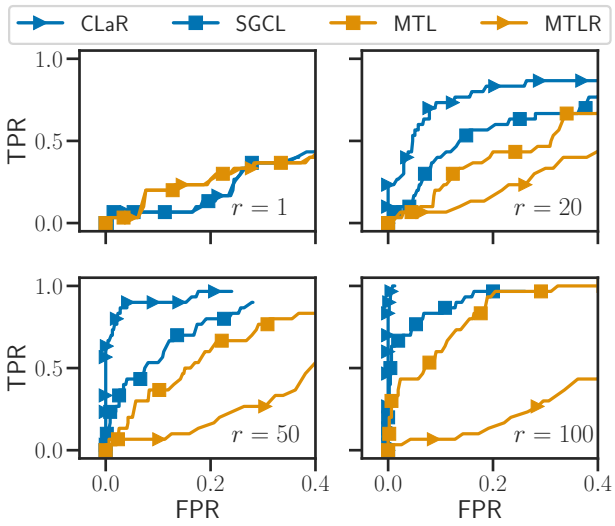

Figure 5: ROC curves of true support recovery for the CLaR, SGCL, MTL and MTLR with $\rho_X = 0.6$, $\rho_S = 0.4$, SNR $= 0.03$, for different values of r.

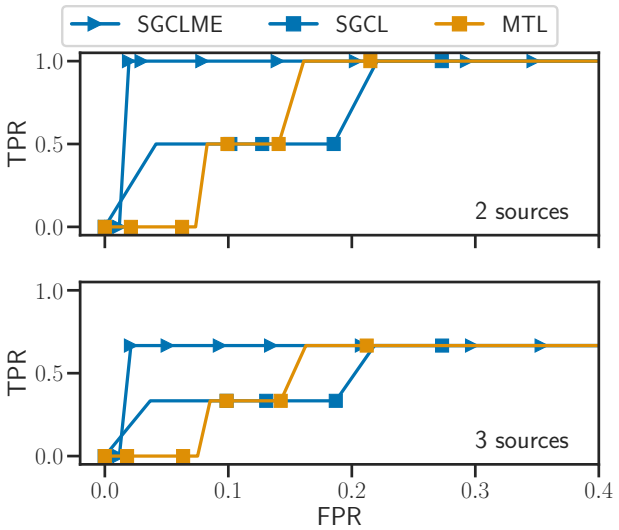

Figure 6: ROC curves of true support recovery for (top) 2 sources: one left auditory and one right auditory source, and for (bottom) 3 sources : one left auditory, one right auditory and one left visual source for the CLaR, SGCL and MTL with real M/EEG design and noise.

space of size $p = 1281$ (oct-3 resolution). The orientation is fixed, and normal to the cortical mantle.

To generate the semi-real data we use the real design matrix X and the real co-standard deviation matrix S, estimated on pre-stimulation data. We then generate plausible MEG signals with MNE. 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 poor or not known). In this experiment we demonstrate that without this whitening process, the homoscedastic solver MTL fails, whereas CLaR succeeds (we dropped here MTLR since it performed poorly on synthetic data and was way to slow to converge).

However, it would not make sense to apply our solver directly on the design matrix X. Here we describe the preprocessing applied to X and Y, using information from X only. First we pre-whiten the data Y and the design matrix X by multiplying from the left by the whitener matrix $W_0 = \mathrm{diag}((\|X_{i:}\|)_{i\in[n]})$. Then, we rescale each column of the design matrix X to have a (Euclidean) norm of 1. Finally, as in Section 4.1, Figure 6 is obtained by varying the estimator-specific regularization parameter λ from $\lambda_{\max}$ to $\lambda_{\min}$ on a geometric grid.

Results of this experiment are in Figure 6. With (top) 2 active sources in the brain, 1 auditory left and 1 auditory right (*i.e.*, $\|\mathrm{B}\|^*_{2,1} = 2$), the MTL estimator performs poorly and does not recover the full support before reaching a false positive rate (FPR) of 0.18. It makes sense since the MTL is not designed to cope with heteroscedastic noise, and the data is not whitened in this experiment. SGCL also performs poorly, because of its statistical limitations: the number of observations used to estimate the covariance matrix in SGCL is too low (here $n \times q \approx 6 \times 10^3$ observations to estimate $n^2/2 \approx 2 \times 10^4$ parameters). CLaR performs better and is able to recover the true sources with a lower FPR. It is worth noting that in CLaR $r \times q \times n \approx 3 \times 10^5$ observations are used to estimate the $n^2/2 \approx 2 \times 10^4$ parameters of the covariance matrix.

Figure 6 (bottom) shows an experiment with 3 active sources, 1 auditory left, 1 auditory right and 1 visual left, CLaR can recover 2 sources among 3 with a low FPR, whereas SGCL and CLaR do not identify $2/3$ the sources before at least a FPR of 0.15.

Conclusion

This work introduces CLaR, a sparse regression estimator 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 standard tools, 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 for the smoothing of other Schatten norms. The benefits of CLaR for support recovery in heteroscedastic context were extensively investigated both on simulations and on real data.

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

References

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.

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.

J. Douglas and H. H. Rachford. On the numerical solution of heat conduction problems in two and three space variables. *Transactions of the American mathematical Society*, 82(2):421–439, 1956.

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.

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.

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, 2018.

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.

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.

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.

Notation For a set $\mathcal{C} \subset \mathbb{R}^{p\times q}$ 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 convex) set $\mathcal{C}$. For $p_1 \in [1, \infty)$, let us write $\mathcal{B}_{s,p_1}$ for the Schatten-p_1 unit ball.

A. Smoothing

A.1. Basic properties of inf-convolution

Proposition 9. 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 $\omega_{\underline{\sigma}} := \underline{\sigma}\omega\left(\frac{\cdot}{\underline{\sigma}}\right)$.

The following holds (see Parikh et al. (2013, p. 136)):

$$h^{**} = h \ , \tag{24}$$
$$(h \,\square\, \omega_{\underline{\sigma}})^* = h^* + \omega_{\underline{\sigma}}^* \ , \tag{25}$$
$$\omega_{\underline{\sigma}}^* = \underline{\sigma}\omega^* \ , \tag{26}$$
$$\|\cdot\|_p^* = \iota_{\mathcal{B}_q}, \text{ where } \frac{1}{p} + \frac{1}{q} = 1 \ , \tag{27}$$
$$(h + \delta)^* = h^* - \delta \quad \forall \delta \in \mathbb{R}^d \ , \tag{28}$$
$$\left(\frac{1}{2}\|\cdot\|^2\right)^* = \frac{1}{2}\|\cdot\|^2 \ . \tag{29}$$

From Equations (26), (28) and (29) it follows that

$$\omega(\cdot) = \frac{1}{2}\|\cdot\|^2 + \frac{1}{2} \implies \omega_{\underline{\sigma}}^* = \frac{\sigma}{2}\|\cdot\|^2 - \frac{\sigma}{2} \ . \tag{30}$$

A.2. Smoothing of Schatten norms

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

Lemma 1. 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 $\omega(\cdot) = \frac{1}{2}\|\cdot\|_F^2 + c$, the following holds true:

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

Proof.

$$\begin{aligned}
(\|\cdot\|_{s,p_1} \,\square\, \omega_{\underline{\sigma}})(Z) &= (\|\cdot\|_{s,p_1} \,\square\, \omega_{\underline{\sigma}})^{**}(Z) && \text{(using Equation (24))} \\
&= \left(\|\cdot\|_{s,p_1}^* + \omega_{\underline{\sigma}}^*\right)^*(Z) && \text{(using Equation (25))} \\
&= \left(\iota_{\mathcal{B}_{s,p_1'}} + \frac{\sigma}{2}\|\cdot\|^2 - c\underline{\sigma}\right)^*(Z) && \text{(using Eqs. (27) and (30))} \\
&= \left(\frac{\sigma}{2}\|\cdot\|^2 + \iota_{\mathcal{B}_{s,p_1'}}\right)^*(Z) + c\underline{\sigma} && \text{(using Eq. (28))} \ .
\end{aligned} \tag{31}$$

We can now compute the last Fenchel transform remaining:

$$\begin{aligned}
\left(\frac{\sigma}{2}\left\|\cdot\right\|_F^2 + \iota_{\mathcal{B}_{s,p_1'}}\right)^*(Z) &= \sup_{U\in\mathbb{R}^{n\times q}} \left(\langle U, Z\rangle - \frac{\sigma}{2}\left\|U\right\|_F^2 - \iota_{\mathcal{B}_{s,p_1'}}(U)\right) \\
&= \sup_{U\in\mathcal{B}_{s,p_1'}} \left(\langle U, Z\rangle - \frac{\sigma}{2}\left\|U\right\|_F^2\right) \\
&= -\inf_{U\in\mathcal{B}_{s,p_1'}} \left(\frac{\sigma}{2}\left\|U\right\|_F^2 - \langle U, Z\rangle\right) \\
&= -\underline{\sigma}\cdot\inf_{U\in\mathcal{B}_{s,p_1'}} \left(\frac{1}{2}\left\|U\right\|_F^2 - \left\langle U, \frac{Z}{\underline{\sigma}}\right\rangle\right) \\
&= -\underline{\sigma}\cdot\inf_{U\in\mathcal{B}_{s,p_1'}} \left(\frac{1}{2}\left\|U - \frac{Z}{\underline{\sigma}}\right\|_F^2 - \frac{1}{2\underline{\sigma}^2}\left\|Z\right\|_F^2\right) \\
&= \frac{1}{2\underline{\sigma}}\left\|Z\right\|_F^2 - \frac{\sigma}{2}\cdot\inf_{U\in\mathcal{B}_{s,p_1'}} \left(\left\|U - \frac{Z}{\underline{\sigma}}\right\|_F^2\right) \\
&= \frac{1}{2\underline{\sigma}}\left\|Z\right\|_F^2 - \frac{\sigma}{2}\left\|\Pi_{\mathcal{B}_{s,p_1'}}\left(\frac{Z}{\underline{\sigma}}\right) - \frac{Z}{\underline{\sigma}}\right\|_F^2 \quad . \qquad (32)
\end{aligned}$$

The result follows by combining Equations (31) and (32).

□

A.3. Schatten 1-norm (nuclear/trace norm)

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

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

Proof. Let $V\operatorname{diag}(\gamma_1, \ldots, \gamma_{n\wedge q})W^\top$ be the singular values decomposition of Z. We remind that $\Pi_{\mathcal{B}_{s,\infty}}$, the projection over $\mathcal{B}_{s,\infty}$, is given by (see (Beck, 2017, Example 7.31, p. 192)):

$$\begin{aligned}
\Pi_{\mathcal{B}_{s,\infty}}\left(\frac{Z}{\underline{\sigma}}\right) &= V\operatorname{diag}\left(\Pi_{\mathcal{B}_\infty}\left(\frac{\gamma_1}{\underline{\sigma}}, \ldots, \frac{\gamma_{n\wedge q}}{\underline{\sigma}}\right)\right)W^\top \\
&= V\operatorname{diag}\left(\frac{\gamma_1}{\underline{\sigma}}\wedge 1, \ldots, \frac{\gamma_{n\wedge q}}{\underline{\sigma}}\wedge 1\right)W^\top \quad , \qquad (33)
\end{aligned}$$

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,

$$\begin{aligned}
\left\|\Pi_{\mathcal{B}s,\infty}\left(\frac{Z}{\underline{\sigma}}\right) - \frac{Z}{\underline{\sigma}}\right\|_F^2 &= \left\|V\operatorname{diag}\left(\frac{\gamma_1}{\underline{\sigma}}\wedge 1 - \frac{\gamma_1}{\underline{\sigma}}, \ldots, \frac{\gamma_{n\wedge q}}{\underline{\sigma}}\wedge 1 - \frac{\gamma_{n\wedge q}}{\underline{\sigma}}\right)W^\top\right\|_F^2 \qquad \text{(using Equation (33))} \\
&= \sum_{i=1}^{n\wedge q}\left(\frac{\gamma_i}{\underline{\sigma}}\wedge 1 - \frac{\gamma_i}{\underline{\sigma}}\right)^2 \\
&= \frac{1}{\underline{\sigma}^2}\sum_{i=1}^{n\wedge q}\left(\gamma_i\wedge\underline{\sigma} - \gamma_i\right)^2 \quad . \qquad (34)
\end{aligned}$$

By combining Equation (32) and Lemma 1 with $p_1' = \infty, c = \frac{n\wedge q}{2}$, the later yields

$$\left(\left\|\cdot\right\|_{s,1} \square\, \omega_{\underline{\sigma}}\right)(Z) = (n\wedge q)\frac{\underline{\sigma}}{2} + \frac{1}{2\underline{\sigma}}\sum_{\gamma_i\le\underline{\sigma}}\gamma_i^2 - \frac{1}{2}\sum_{\gamma_i\ge\underline{\sigma}}\underline{\sigma} + \sum_{\gamma_i\ge\underline{\sigma}}\gamma_i \quad . \qquad (35)$$

Moreover it can be noticed that

$$\begin{aligned}\min_{\substack{S\in\mathbb{S}^n_{++}\\ S\succeq\underline{\sigma}\,\mathrm{Id}_n}} \frac{1}{2}\operatorname{Tr}[Z^\top S^{-1}Z]+\frac{1}{2}\operatorname{Tr}(S) &= \frac{1}{2}\sum_{i=1}^{n\wedge q}\frac{\gamma_i^2}{\gamma_i\vee\underline{\sigma}}+\frac{1}{2}\sum_{i=1}^{n\wedge q}\gamma_i\vee\underline{\sigma}\\ &= \frac{1}{2\underline{\sigma}}\sum_{\gamma_i\le\underline{\sigma}}\gamma_i^2+\frac{1}{2}\sum_{\gamma_i\ge\underline{\sigma}}\gamma_i+\frac{1}{2}\sum_{\gamma_i\le\underline{\sigma}}\underline{\sigma}+\frac{1}{2}\sum_{\gamma_i\ge\underline{\sigma}}\gamma_i\\ &= \frac{1}{2\underline{\sigma}}\sum_{\gamma_i\le\underline{\sigma}}\gamma_i^2+\sum_{\gamma_i\ge\underline{\sigma}}\gamma_i+\frac{1}{2}\sum_{\gamma_i\le\underline{\sigma}}\underline{\sigma}+(n\wedge q)\frac{\underline{\sigma}}{2}-(n\wedge q)\frac{\underline{\sigma}}{2}\\ &= \frac{1}{2\underline{\sigma}}\sum_{\gamma_i\le\underline{\sigma}}\gamma_i^2+\sum_{\gamma_i\ge\underline{\sigma}}\gamma_i+(n\wedge q)\frac{\underline{\sigma}}{2}-\frac{1}{2}\sum_{\gamma_i\ge\underline{\sigma}}\underline{\sigma}\ , \end{aligned}\tag{36}$$

and identifying Equation (36) and Equation (35) leads to the result. □

A.4. 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 10. For the choice $\omega(Z)=\frac{1}{2}\|Z\|_F^2+\frac{n\wedge qr}{2}$, then the following holds true:

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

Proof. The result is a direct application of Proposition 1, 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.5. Schatten 2-norm (Frobenius norm)

Proposition 11. For the choice $\omega(\cdot)=\frac{1}{2}\|\cdot\|_F^2+\frac{1}{2}$, and for $Z\in\mathbb{R}^{n\times q}$ then the following holds true:

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

Proof. Let us recall that $\|\cdot\|_F=\|\cdot\|_{s,2}$, then

$$\Pi_{\mathcal{B}_{s,2}}\left(\frac{Z}{\underline{\sigma}}\right)=\begin{cases}0, & \text{if } \|Z\|_F\le\underline{\sigma}\\ \frac{Z}{\|Z\|_F}, & \text{if } \|Z\|_F>\underline{\sigma}\end{cases}\ .\tag{37}$$

By combining Equation (37) and Lemma 1 with $p_1'=2, c=\frac{1}{2}$, the later yields

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

□

A.6. Schatten infinity-norm (spectral norm)

Proposition 12. For the choice $\omega(\cdot)=\frac{1}{2}\|\cdot\|_F^2+\frac{1}{2}$ and for $Z\in\mathbb{R}^{n\times q}$, then the following holds true:

$$(\|\cdot\|_{s,\infty}\,\square\,\omega_{\underline{\sigma}})(Z)=\begin{cases}\frac{1}{2\underline{\sigma}}\|Z\|_F^2+\frac{\underline{\sigma}}{2}, & \text{if } \|Z\|_1\le 1\\ \frac{\underline{\sigma}}{2}\sum_{i=1}^{n\wedge q}\left(\frac{\gamma_i^2}{\underline{\sigma}^2}-\gamma^2\right)_+\!+\frac{\underline{\sigma}}{2} & \text{if } \|Z\|_1>1\end{cases}\ ,$$

where $\gamma\ge 0$ is defined by the implicit equation

$$\left\|\left(\operatorname{ST}\left(\frac{\gamma_1}{\underline{\sigma}},\gamma\right),\dots,\operatorname{ST}\left(\frac{\gamma_{n\wedge q}}{\underline{\sigma}},\gamma\right)\right)\right\|_1=1\tag{38}$$

Proof. We remind that $\Pi_{\mathcal{B}_{s,\infty}}$, the projection over $B_{s,1}$, is given by Beck (2017, Example 7.31, p. 192):

$$\Pi_{\mathcal{B}_{s,1}}\left(\frac{Z}{\underline{\sigma}}\right) = \begin{cases} \frac{Z}{\underline{\sigma}}, & \text{if } \|Z\|_{s,1} \leq \underline{\sigma} \\ V \operatorname{diag}(\mathrm{ST}(\frac{\gamma_i}{\underline{\sigma}}, \gamma)) W^\top, & \text{if } \|Z\|_{s,1} > \underline{\sigma} \end{cases}, \tag{39}$$

γ being defined by the implicit equation

$$\left\| \left(\mathrm{ST}\left(\frac{\gamma_1}{\underline{\sigma}}, \gamma\right), \ldots, \mathrm{ST}\left(\frac{\gamma_{n\wedge q}}{\underline{\sigma}}, \gamma\right) \right) \right\|_{s,1} = 1 \ . \tag{40}$$

By combining Equation (37) and Lemma 1 (with $p_1' = \infty, c = \frac{1}{2}$) it follows that

$$(\|\cdot\|_F \square \omega_{\underline{\sigma}})(Z) = \begin{cases} \frac{1}{2\underline{\sigma}} \|Z\|_F^2 + \frac{\underline{\sigma}}{2}, & \text{if } \|Z\|_1 \leq \underline{\sigma} \\ \frac{1}{2\underline{\sigma}} \|Z\|_F^2 + \frac{\underline{\sigma}}{2} - \frac{\underline{\sigma}}{2} \left\| \Pi_{\mathcal{B}_{s,1}}\left(\frac{Z}{\underline{\sigma}}\right) - \frac{Z}{\underline{\sigma}} \right\|_F^2, & \text{if } \|Z\|_1 > \underline{\sigma} \end{cases} . \tag{41}$$

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

$$\begin{aligned} \left\| \Pi_{\mathcal{B}_{s,1}}\left(\frac{Z}{\underline{\sigma}}\right) - \frac{Z}{\underline{\sigma}} \right\|_F^2 &= \frac{1}{\underline{\sigma}^2} \left\| V \operatorname{diag}((\gamma_i - \gamma\underline{\sigma})_+ - \gamma_i) W^\top \right\| \qquad \text{(using Equation (39))} \\ &= \frac{1}{\underline{\sigma}^2} \sum_{i=1}^{n\wedge q} \left((\gamma_i - \gamma\underline{\sigma})_+ - \gamma_i \right)^2 \\ &= \frac{1}{\underline{\sigma}^2} \Big(\sum_{\gamma_i \geq \gamma\underline{\sigma}}^{n\wedge q} \gamma^2 \underline{\sigma}^2 + \sum_{\gamma_i < \gamma\underline{\sigma}}^{n\wedge q} \gamma_i^2 \Big) \ . \end{aligned} \tag{42}$$

By plugging Equation (42) into Equation (41) it follows, that if $\|Z\|_{s,1} > \underline{\sigma}$

$$\begin{aligned} (\|\cdot\|_F \square \omega_{\underline{\sigma}})(Z) &= \frac{1}{2\underline{\sigma}} \sum_{i=1}^{n\wedge q} \gamma_i^2 + \frac{\underline{\sigma}}{2} - \frac{1}{2\underline{\sigma}} \sum_{\gamma_i \geq \gamma\underline{\sigma}}^{n\wedge q} \gamma^2 \underline{\sigma}^2 - \frac{1}{2\underline{\sigma}} \sum_{\gamma_i < \gamma\underline{\sigma}}^{n\wedge q} \gamma_i^2 \\ &= \frac{1}{2\underline{\sigma}} \sum_{\gamma_i \geq \gamma\underline{\sigma}}^{n\wedge q} \left(\gamma_i^2 - \gamma^2 \underline{\sigma}^2 \right) + \frac{\underline{\sigma}}{2} \\ &= \frac{\underline{\sigma}}{2} \sum_{i=1}^{n\wedge q} \left(\frac{\gamma_i^2}{\underline{\sigma}^2} - \gamma^2 \right)_+ + \frac{\underline{\sigma}}{2} \ . \end{aligned} \tag{43}$$

Proposition 12 follows by plugging Equation (43) into Equation (41).

□

B. Proofs CLaR

B.1. Proof of Proposition 2

Proof. Proposition 2 follows from Appendix A.4 by choosing $Z = \frac{1}{\sqrt{rq}}[Y^{(1)} - X\mathrm{B}, \ldots, Y^{(r)} - X\mathrm{B}]$ and by taking the $\arg\min$ over B. □

B.2. Proof of Proposition 7

Proof. Let the primal optimum be

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

Then

$$\begin{aligned}
p^* &= \min_{\substack{\mathrm{B}\in\mathbb{R}^{p\times q}\\ R^{(l)}=Y^{(l)}-X\mathrm{B},\ \forall l\in[r]\\ S\succeq\underline{\sigma}}} \frac{1}{2nqr}\sum_{l=1}^{r}\left\|R^{(l)}\right\|_{S^{-1}}^2 + \frac{1}{2n}\operatorname{Tr}(S) + \lambda\left\|\mathrm{B}\right\|_{2,1}\\
&= \min_{\substack{\mathrm{B}\in\mathbb{R}^{p\times q}\\ R^{(1)},\dots,R^{(r)}\in\mathbb{R}^{n\times q}\\ S\succeq\underline{\sigma}}} \max_{\Theta^{(1)},\dots,\Theta^{(r)}\in\mathbb{R}^{n\times q}} \frac{1}{2nqr}\sum_{l=1}^{r}\left\|R^{(l)}\right\|_{S^{-1}}^2 + \frac{1}{2n}\operatorname{Tr}(S) + \lambda\left\|\mathrm{B}\right\|_{2,1} + \frac{\lambda}{r}\sum_{l=1}^{r}\left\langle\Theta^{(l)}, Y^{(l)}-X\mathrm{B}-R^{(l)}\right\rangle .
\end{aligned}$$

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

$$\begin{aligned}
p^* &= \max_{\Theta^{(1)},\dots,\Theta^{(r)}\in\mathbb{R}^{n\times q}} \min_{\substack{\mathrm{B}\in\mathbb{R}^{p\times q}\\ R^{(1)},\dots,R^{(r)}\in\mathbb{R}^{n\times q}\\ S\succeq\underline{\sigma}}} \frac{1}{2nqr}\sum_{l=1}^{r}\left\|R^{(l)}\right\|_{S^{-1}}^2 + \frac{1}{2n}\operatorname{Tr}(S) + \lambda\left\|\mathrm{B}\right\|_{2,1} + \frac{\lambda}{r}\sum_{l=1}^{r}\left\langle\Theta^{(l)}, Y^{(l)}-X\mathrm{B}-R^{(l)}\right\rangle\\
&= \max_{\Theta^{(1)},\dots,\Theta^{(r)}\in\mathbb{R}^{n\times q}} \left(\min_{S\succeq\underline{\sigma}}\frac{1}{r}\sum_{l=1}^{r}\min_{R^{(l)}\in\mathbb{R}^{n\times q}}\left(\frac{\left\|R^{(l)}\right\|_{S^{-1}}^2}{2nq} - \left\langle\Theta^{(l)}, R^{(l)}\right\rangle\right) + \lambda\min_{\mathrm{B}\in\mathbb{R}^{p\times q}}\left(\left\|\mathrm{B}\right\|_{2,1} - \left\langle\bar{\Theta}, X\mathrm{B}\right\rangle\right)\right.\\
&\qquad\left. + \frac{\operatorname{Tr}(S)}{2n} + \frac{\lambda}{r}\sum_{l=1}^{r}\left\langle\Theta^{(l)}, Y^{(l)}\right\rangle\right) . \qquad (44)
\end{aligned}$$

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:

$$\begin{aligned}
d^* &= \max_{\Theta^{(1)},\dots,\Theta^{(r)}\in\mathbb{R}^{n\times q}} \min_{S\succeq\underline{\sigma}} -\frac{1}{r}\sum_{l=1}^{r}\frac{nq\lambda^2}{2}\left\langle\Theta^l{\Theta^l}^\top, S\right\rangle - \lambda\iota_{\mathcal{B}_{2,\infty}}(X^\top\bar{\Theta}) + \frac{\operatorname{Tr}(S)}{2n} + \frac{\lambda}{r}\sum_{l=1}^{r}\langle\Theta^{(l)}, Y^{(l)}\rangle\\
&= \max_{\Theta^{(1)},\dots,\Theta^{(r)}\in\mathbb{R}^{n\times q}} \frac{1}{2n}\min_{S\succeq\underline{\sigma}}\left(\left\langle\mathrm{Id}_n, S\right\rangle - \frac{qn^2\lambda^2}{r}\sum_{l=1}^{r}\left\langle\Theta^{(l)}{\Theta^{(l)}}^\top, S\right\rangle\right) - \lambda\iota_{\mathcal{B}_{2,\infty}}(X^\top\bar{\Theta}) + \frac{\lambda}{r}\sum_{l=1}^{r}\langle\Theta^l, Y^{(l)}\rangle\\
&= \max_{\Theta^{(1)},\dots,\Theta^{(r)}\in\mathbb{R}^{n\times q}} \frac{1}{2n}\min_{S\succeq\underline{\sigma}}\left\langle\mathrm{Id}_n - \frac{qn^2\lambda^2}{r}\sum_{l=1}^{r}\Theta^{(l)}{\Theta^{(l)}}^\top S\right\rangle - \lambda\iota_{\mathcal{B}_{2,\infty}}(X^\top\bar{\Theta}) + \frac{\lambda}{r}\sum_{l=1}^{r}\langle\Theta^l, Y^{(l)}\rangle . \qquad (45)
\end{aligned}$$

$$\min_{S\succeq\underline{\sigma}}\left\langle\mathrm{Id}_n - \frac{qn^2\lambda^2}{r}\sum_{l=1}^{r}\Theta^{(l)}{\Theta^{(l)}}^\top, S\right\rangle = \begin{cases}\left\langle\mathrm{Id}_n - \frac{qn^2\lambda^2}{r}\sum_{l=1}^{r}\Theta^{(l)}{\Theta^{(l)}}^\top, \underline{\sigma}\right\rangle, & \text{if } \mathrm{Id}_n - \frac{qn^2\lambda^2}{r}\sum_{l=1}^{r}\Theta^{(l)}{\Theta^{(l)}}^\top \succeq 0\\ -\infty, & \text{otherwise.}\end{cases}$$

It follows that the dual problem of CLaR is

$$\max_{(\Theta^{(1)},\dots,\Theta^{(r)})\in\Delta_{X,\lambda}} \frac{\sigma}{2}\left(1-\frac{qn\lambda^2}{r}\sum_{l=1}^{r}\operatorname{Tr}\Theta^{(l)}\Theta^{(l)\top}\right)+\frac{\lambda}{r}\sum_{l=1}^{r}\left\langle\Theta^{(l)},Y^{(l)}\right\rangle \quad , \tag{46}$$

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

B.3. Proof of Remark 5

Proof.

$$\begin{aligned}
RR^\top &= \sum_{l=1}^{r} R^{(l)}R^{(l)\top}\\
&= \sum_{l=1}^{r}(Y^{(l)}-X\mathrm{B})(Y^{(l)}-X\mathrm{B})^\top\\
&= \sum_{l=1}^{r}Y^{(l)}Y^{(l)\top}-\sum_{1}^{r}Y^{(l)}(X\mathrm{B})^\top-\sum_{1}^{r}X\mathrm{B}Y^{(l)\top}+rX\mathrm{B}(X\mathrm{B})^\top\\
&= r\mathrm{cov}_Y-r\bar{Y}^\top X\mathrm{B}-r(X\mathrm{B})^\top\bar{Y}+rX\mathrm{B}(X\mathrm{B})^\top
\end{aligned} \tag{47}$$

□

B.4. Proof of Proposition 8

Let us recall that

$$\Sigma^{\mathrm{SGCL}}=\frac{1}{qr}(\sum_{l=1}^{r}R^{(l)})(\sum_{l=1}^{r}R^{(l)})^\top \quad ,$$

and

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

B.4.1. Proof of Equation (19)

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. □

B.4.2. Proof of Equation (20)

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 :

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

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

$$\begin{aligned}
\hat{\Sigma}^{\mathrm{SGCL}} &\underset{law}{\sim}\frac{1}{q}S^*\mathrm{E}(S^*\mathrm{E})^\top,\\
\mathrm{cov}(\hat{\Sigma}^{\mathrm{SGCL}}) &= \frac{1}{q^2}\mathrm{cov}(S^*\mathrm{E}(S^*\mathrm{E})^\top) \quad .
\end{aligned}$$

On the other hand:

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

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

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

□