Generalized Concomitant Multi-Task Lasso for Sparse Multimodal Regression

Mathurin Massias, Olivier Fercoq, Alexandre Gramfort, Joseph Salmon

Generalized Concomitant Multi-Task Lasso for Sparse Multimodal Regression

Mathurin Massias
Inria
Université Paris-Saclay

Olivier Fercoq
LTCI, Télécom ParisTech
Université Paris-Saclay

Alexandre Gramfort
Inria
Université Paris-Saclay

Joseph Salmon
LTCI, Télécom ParisTech
Université Paris-Saclay

Abstract

In high dimension, it is customary to consider Lasso-type estimators to enforce sparsity. For standard Lasso theory to hold, the regularization parameter should be proportional to the noise level, which is often unknown in practice. A remedy is to consider estimators such as the Concomitant Lasso, which jointly optimize over the regression coefficients and the noise level. However, when data from different sources are pooled to increase sample size, noise levels differ and new dedicated estimators are needed. We provide new statistical and computational solutions to perform heteroscedastic regression, with an emphasis on brain imaging with magneto- and electroencephalography (M/EEG). When instantiated to de-correlated noise, our framework leads to an efficient algorithm whose computational cost is not higher than for the Lasso, but addresses more complex noise structures. Experiments demonstrate improved prediction and support identification with correct estimation of noise levels.

1 Introduction

In the context of regression, when the number of predictors largely exceeds the number of observations, sparse estimators provide interpretable and memory efficient models. Following the seminal work on the Lasso/Basis pursuit [Tibshirani, 1996, Chen and Donoho, 1995], a popular route to sparsity is to use convex ℓ_1-type penalties. Lasso-type estimators rely on a regularization parameter λ trading data-fitting versus sparsity, which requires careful tuning. Statistical analysis of the Lasso estimator states that λ should be proportional to the noise level[1] [Bickel et al., 2009], though the latter is rarely known in practice. To address this issue, it has been proposed to jointly estimate the noise level along with the regression coefficients. A notable approach is via joint penalized maximum likelihood after a change of variable to avoid minimization of a non-convex function [Städler et al., 2010]. Another approach, the Concomitant Lasso [Owen, 2007] (inspired by Huber [1981]), and equivalent to the Square-root/Scaled Lasso [Belloni et al., 2011, Sun and Zhang, 2012]) includes noise level estimation by modifying the Lasso objective function. This estimator, which reaches optimal statistical rates for sparse regression [Belloni et al., 2011, Sun and Zhang, 2012], makes the regularization parameter independent of the noise level. From a practical point of view, it is well-suited for high dimension settings [Reid et al., 2016], and current solvers [Ndiaye et al., 2017] make its computation as fast as for the Lasso. While first attempts used second order cone programming solvers [Belloni et al., 2011], *e.g.,* TFOCS [Becker et al., 2011], recent ones rely on coordinate descent algorithms [Tseng, 2001, Friedman et al., 2007] and safe screening rules [El Ghaoui et al., 2012, Fercoq et al., 2015].

In various applied contexts it is customary to pool observations from different sources or devices, to increase sample size and boost statistical power. Yet, this leads to datasets with heteroscedastic noise. Heteroscedasticity, to be opposed to homoscedasticity, is a common statistical phenomenon occurring when observations are contaminated with non-uniform noise levels [Engle, 1982, Carroll and Ruppert, 1988]. This is for example the case of magneto- and electroencephalography (M/EEG) data, usually recorded from three types of sensors (gradiometers, magnetometers and electrodes), each having different signal and noise amplitudes.

Proceedings of the 21^{st} International Conference on Artificial Intelligence and Statistics (AISTATS) 2018, Lanzarote, Spain. PMLR: Volume 84.

[1] with Gaussian noise, the noise level stands for the standard deviation

Several statistical contributions have tried to address heteroscedastic models in high dimensional regression. Most works have relied on an exponential representation of the variance (the log-variance being modeled as a linear combination of the features), leading to non-convex objective functions. Solvers considered for such approaches require alternate minimization [Kolar and Sharpnack, 2012], possibly in an iterative fashion [Daye et al., 2012], a notable difference with a jointly convex formulation, for which one can control global optimality with duality gap certificates as proposed here. Similarly, Wagener and Dette [2012] estimate the variance with a preliminary adaptive Lasso step, and correct the data-fitting term in a second step.

Here, we propose the multi-task Smoothed Generalized Concomitant Lasso (SGCL), an estimator that can handle data from different origins in a high dimensional sparse regression model by jointly estimating the regression coefficients and the noise levels of each modality or data source. Contrary to other heteroscedastic Lasso estimators such as ScHeDs (a second order cone program) [Dalalyan et al., 2013], its computational cost is comparable to the Lasso, as it can benefit from coordinate descent solvers [Tseng, 2001, Friedman et al., 2007] and other standard speed-ups for the Lasso (safe rules [El Ghaoui et al., 2012], strong rules [Tibshirani et al., 2012], etc.). This model also leads to a parameterization of the problem with one single scalar λ, independent of the multiple noise levels present in heterogeneous data.

Our manuscript is organized as follows. In Section 2, after reminding the necessary background of the Concomitant Lasso estimator, we introduce our general framework. We derive in Section 3 the necessary mathematical results to obtain an efficient solver based on coordinate descent. We then present in Section 4 a lightweight version of it, adapted to more specific noise models. Finally, in Section 5 we provide empirical evidence using simulations with known ground truth that our model yields better support recovery and prediction than homoscedastic estimators. On the problem of source localization with real M/EEG recordings, we show that the proposed model leads to consistent estimators of the noise standard deviations for each modality, hence learning from the data the right balance between modalities with high or low SNRs.

2 Concomitant estimators

Notation For any integer $d \in \mathbb{N}$, we denote by $[d]$ the set $\{1, \dots, d\}$. Our observation matrix is $Y \in \mathbb{R}^{n \times q}$ with n the number of samples, with q the number of tasks and the design matrix $X = [X_1, \dots, X_p] \in \mathbb{R}^{n \times p}$ has p explanatory variables or features, stored column-wise. The standard Euclidean norm (resp. inner product) on vectors or matrices is written $\|\cdot\|$ (resp. $\langle \cdot, \cdot \rangle$), the ℓ_1 norm $\|\cdot\|_1$, the ℓ_∞ norm $\|\cdot\|_\infty$, and the matrix transposition of a matrix Q is denoted by $Q^\top$. For $\mathrm{B} \in \mathbb{R}^{p \times q}$, its j^{th} row is $\mathrm{B}_{j,:}$ and $\|\mathrm{B}\|_{2,1} = \sum_{j=1}^p \|\mathrm{B}_{j,:}\|$ (resp. $\|\mathrm{B}\|_{2,\infty} = \max_{j \in [p]} \|\mathrm{B}_{j,:}\|$) is its row-wise $\ell_{2,1}$ (resp. $\ell_{2,\infty}$) norm. For matrices, $\|\cdot\|_2$ is the spectral norm. For real numbers a and b, $a \vee b$ stands for the maximum of a and b, and $(a)_+ = a \vee 0$.

We denote $\mathrm{BST}(\cdot, \tau)$ the block soft-thresholding operator at level $\tau > 0$, *i.e.,* $\mathrm{BST}(x, \tau) = (1 - \tau / \|x\|)_+ \cdot x$ for any $x \in \mathbb{R}^d$ (with the convention $\frac{0}{0} = 1$).

The sub-gradient of a convex function $f : \mathbb{R}^d \to \mathbb{R}$ at x is defined as $\partial f(x) = \{z \in \mathbb{R}^d : \forall y \in \mathbb{R}^d, f(y) - f(x) \geq \langle z, y - x \rangle\}$. We denote by ι_C the indicator function of a set C, defined as $\iota_C(x) = 0$ if $x \in C$ and $\iota_C(x) = +\infty$ if $x \notin C$. The identity matrix of size $n \times n$ is denoted by I_n (or I, when there is no dimension ambiguity).

We write $\mathbb{S}^n$ for the set of symmetric matrices and $\mathbb{S}^n_+$ (resp. $\mathbb{S}^n_{++}$) for the set of positive semi-definite matrices (resp. positive definite matrices). For two matrices S_1 and S_2 we write $S_1 \succeq S_2$ (resp. $S_1 \succ S_2$) for $S_1 - S_2 \in \mathbb{S}^n_+$ (resp. $S_1 - S_2 \in \mathbb{S}^n_{++}$). The symbol Tr denotes the trace operator, and $\|A\|_S = \sqrt{\mathrm{Tr}\, A^\top S A}$ is the Mahalanobis norm induced by $S \in \mathbb{S}^n_{++}$. For more compact notation, for $\underline{\sigma} > 0$ we denote $\underline{\Sigma} = \underline{\sigma}\, \mathrm{I}_n$.

As much as possible, we denote vectors with lower case letters and matrices with upper case ones.

2.1 Reminder on Concomitant Lasso

Let us first recall the Concomitant Lasso estimator, following the vector formulation ($y \in \mathbb{R}^n$) proposed in Owen [2007], Sun and Zhang [2012].

Definition 1. For $\lambda > 0$, the Concomitant Lasso coefficient and standard deviation estimators are defined as solutions of the optimization problem

$$\operatorname*{arg\,min}_{\beta \in \mathbb{R}^p, \sigma > 0} \frac{\|y - X\beta\|^2}{2n\sigma} + \frac{\sigma}{2} + \lambda \|\beta\|_1 . \tag{1}$$

To avoid numerical issues when σ approaches 0, it was proposed in Ndiaye et al. [2017] to add a constraint on σ in the objective function. Following the terminology introduced in Nesterov [2005], this was coined the Smoothed Concomitant Lasso.

Definition 2. For $\underline{\sigma} > 0$ and $\lambda > 0$, the Smoothed Concomitant Lasso estimator $\hat{\beta}$ and its associated standard deviation estimator $\hat{\sigma}$ are defined as

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

2.2 General problem formulation

Motivated by our application to the M/EEG inverse problem (see Section 5.3, we present all the results in a multi-task setting. These results are still valid for single task problems ($q = 1$), in which case the formulas and algorithms are simpler (see Appendix B).

We now extend the Smoothed Concomitant Lasso to more general noise models, and present some properties obtained thanks to convexity and duality. As a warning, in our formulation the matrix $\Sigma^* \in \mathbb{S}^n_{++}$ is the co-standard-deviation matrix (the square-root of the covariance matrix), in contrast with the standard Gaussian noise model notation. The model reads:

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

where entries of E are independent, centered and normally distributed.

Definition 3. For $\underline{\sigma} > 0$ and $\lambda > 0$ (recall that we denote $\underline{\Sigma} = \underline{\sigma}\,\mathrm{I}_n$), we define the multi-task Smoothed Generalized Concomitant Lasso (multi-task SGCL) estimator $\hat{\mathrm{B}}$ and its associated co-standard-deviation matrix $\hat{\Sigma}$ as the solutions of the optimization problem

$$(\hat{\mathrm{B}}, \hat{\Sigma}) \in \underset{\mathrm{B}\in\mathbb{R}^{p\times q}, \Sigma\in\mathbb{S}^n_{++}, \Sigma\succeq\underline{\Sigma}}{\arg\min} \mathcal{P}^{(\lambda)}(\mathrm{B}, \Sigma) \ , \tag{4}$$

with $\mathcal{P}^{(\lambda)}(\mathrm{B}, \Sigma) = \dfrac{\|Y - X\mathrm{B}\|^2_{\Sigma^{-1}}}{2nq} + \dfrac{\mathrm{Tr}(\Sigma)}{2n} + \lambda \|\mathrm{B}\|_{2,1}$.

Remark 1. Concomitant estimators such as the Smoothed Concomitant Lasso rely on perspective functions. A general framework for optimization with similar functions is provided in Combettes and Müller [2016], and applies for instance to other (potentially non-convex) alternative noise estimators, *e.g.,* TREX [Lederer and Müller, 2015]. Yet, we are not aware of a matrix perspective theory as we propose in the present work to handle anisotropic noise.

Proposition 1. *Problem (4) is jointly convex.*

Proof. The constraint set is convex and the matrix-fractional function $(Z, \Sigma) \mapsto \mathrm{Tr}\, Z^\top \Sigma^{-1} Z$ is jointly convex over $\mathbb{R}^{n\times q} \times \mathbb{S}^n_{++}$, *cf.* Boyd and Vandenberghe [2004, Example 3.4]. □

As for $\underline{\sigma}$ in the Smoothed Concomitant Lasso, the constraint $\Sigma \succeq \underline{\Sigma}$ acts as a regularizer in the dual, and it is introduced for numerical stability. In practice, the value of $\underline{\Sigma} = \underline{\sigma}\,\mathrm{I}_n$ can be set as follows:

- If prior information on the minimal noise level present in the data is available, $\underline{\sigma}$ can be set as this bound. Indeed, if $\hat{\Sigma} \succ \underline{\Sigma}$, then the constraint $\Sigma \succeq \underline{\Sigma}$ is not active and the solution to (4) is a solution of the non-smoothed problem with $\underline{\Sigma} = 0$.
- Without prior information on the noise level, one can use a proportion of the initial estimation of the noise standard deviation $\underline{\sigma} = 10^{-\alpha}\|Y\|/\sqrt{nq}$, with for example $\alpha \in \{2, 3\}$.

3 Properties of the multi-task SGCL

3.1 Optimization

Theorem 1. *The dual formulation of the multi-task Smoothed Generalized Concomitant Lasso reads*

$$\hat{\Theta} = \underset{\Theta\in\Delta_{X,\lambda}}{\arg\max} \underbrace{\langle Y, \lambda\Theta\rangle + \underline{\sigma}\left(\frac{1}{2} - \frac{nq\lambda^2}{2}\|\Theta\|^2\right)}_{\mathcal{D}^{(\lambda,\underline{\Sigma})}(\Theta)} \ , \tag{5}$$

for $\Delta_{X,\lambda} = \left\{\Theta \in \mathbb{R}^{n\times q} : \|X^\top\Theta\|_{2,\infty} \le 1, \|\Theta\|_2 \le \frac{1}{\lambda n\sqrt{q}}\right\}$.

The link between $\hat{\mathrm{B}}$ *and* $\hat{\Sigma}$ *is detailed in Proposition 2. We also have the link-equation between primal and dual solutions:*

$$\hat{\Theta} = \frac{1}{nq\lambda}\hat{\Sigma}^{-1}(Y - X\hat{\mathrm{B}}) \ , \tag{6}$$

and the sub-differential inclusion:

$$X^\top\hat{\Sigma}^{-1}(Y - X\hat{\mathrm{B}}) \in nq\lambda\partial\|\cdot\|_{2,1}(\hat{\mathrm{B}}) \ . \tag{7}$$

The proof of these results is in Appendix A.1.

Remark 2. The link equation provides a natural way to construct a dual feasible point from any pair (B, Σ). Since at convergence Equation (6) holds, we can choose as a dual point $\Theta = \Sigma^{-1}(Y - X\mathrm{B})/\alpha$ where $\alpha = \|X^\top\Sigma^{-1}(Y - X\mathrm{B})\|_{2,\infty} \vee \lambda n\sqrt{q}\|\Sigma^{-1}(Y - X\mathrm{B})\|_2$ is a scalar chosen to make Θ dual feasible.

Proposition 2. *The solution of*

$$\Sigma_0 \in \underset{\Sigma\in\mathbb{S}^n_{++}, \Sigma\succeq\underline{\Sigma}}{\arg\min} \frac{1}{2nq}\|Y - X\mathrm{B}\|^2_{\Sigma^{-1}} + \frac{1}{2n}\mathrm{Tr}(\Sigma) \ , \tag{8}$$

is given by

$$\Sigma_0 = \Psi(Z, \underline{\sigma}) := U\,\mathrm{diag}(\mu_1, \ldots, \mu_r, \underline{\sigma}, \ldots, \underline{\sigma})U^\top \ , \tag{9}$$

where $Z = \frac{1}{\sqrt{q}}(Y - X\mathrm{B})$, $U\,\mathrm{diag}(\lambda_1, \ldots, \lambda_r, 0, \ldots, 0)U^\top$ *is an eigenvalue decomposition of* $ZZ^\top$, *(*i.e., *r is the rank of $ZZ^\top$, $\lambda_1 \ge \cdots \ge \lambda_r > 0$ and $UU^\top = \mathrm{I}_n$), and for $i \in [r]$, $\mu_i = \sqrt{\lambda_i} \vee \underline{\sigma}$.*

The proof is included in Appendix A.2.

Remark 3. Formula (9) makes it straightforward to compute Σ^{-1} and $\mathrm{Tr}\,\Sigma$, which we rather store than Σ for computational efficiency in Algorithm 1.

At every update of Σ, it is also beneficial to precompute $\Sigma^{-1}X$ and $\Sigma^{-1}R$: maintaining $\Sigma^{-1}R$ rather than R avoids multiplication by Σ^{-1} at every BCD step.

Remark 4. Similarly to the Concomitant Lasso, and contrary to the Lasso, the multi-task SGCL is equivariant under scaling of the response, in the following sense: consider the transformation

$$Y' = \alpha Y,\ \mathrm{B}' = \alpha \mathrm{B},\ \Sigma' = \alpha \Sigma, \quad (\alpha > 0)\ ,$$

which leaves the model (3) invariant. Then one can check that the solutions of (4) are multiplied by the same factor: $\hat{\mathrm{B}}' = \alpha \hat{\mathrm{B}}$ and $\hat{\Sigma}' = \alpha \hat{\Sigma}$

As for the Lasso, the null vector is optimal for the multi-task Smoothed Generalized Concomitant Lasso problem when the regularization parameter becomes too large. We refer to the smallest λ leading to a null solution as the *critical parameter* and denote it $\lambda_{\max}$. Its computation is detailed in the next proposition.

Proposition 3. *For the multi-task SGCL estimator we have the following property: with* $\Sigma_{\max} = \Psi(Y, \underline{\sigma})$,

$$\hat{\mathrm{B}} = 0, \quad \forall \lambda \geq \lambda_{\max} := \frac{1}{nq} \|X^\top \Sigma_{\max}^{-1} Y\|_{2,\infty}\ . \quad (10)$$

Proof. Fermat's rule for (4) states that

$$\begin{aligned}(0, \hat{\Sigma}) \in & \underset{\mathrm{B} \in \mathbb{R}^{p \times q}, \Sigma \in \mathbb{S}^n_{++}, \Sigma \succeq \underline{\Sigma}}{\arg\min} \mathcal{P}(\mathrm{B}, \Sigma) \\ & \Leftrightarrow 0 \in \left\{ -\tfrac{1}{nq} X^\top \hat{\Sigma}^{-1} Y \right\} + \lambda \mathcal{B}_{2,\infty} \\ & \Leftrightarrow \tfrac{1}{nq} \left\| X^\top \hat{\Sigma}^{-1} Y \right\|_{2,\infty} \leq \lambda.\end{aligned}$$

Thus, $\lambda_{\max} = \frac{1}{nq} \|X^\top \hat{\Sigma}^{-1} Y\|_{2,\infty}$ is the critical parameter. Then, notice that for $\hat{\mathrm{B}} = 0$ one has

$$\hat{\Sigma} = \Psi(Y, \underline{\sigma}) = \underset{\Sigma \in \mathbb{S}^n_{++}, \Sigma \succeq \underline{\Sigma}}{\arg\min} \frac{1}{2nq} \|Y\|^2_{\Sigma^{-1}} + \frac{1}{2n} \operatorname{Tr}(\Sigma)\ .$$

□

3.2 Algorithm

Since the multi-task SGCL formulation is jointly convex, one can rely on alternate minimization to find a solution. Moreover, the formulation has the appealing property that when Σ is fixed, the convex problem in B is a standard "smooth + ℓ_1-type" problem. This can be solved easily using standard block coordinate descent (BCD) algorithm. Alternatively, when B is fixed, the minimization in Σ has the closed-form solution of Proposition 2. The Σ update being more costly than the B update, one can perform it every f BCD epochs (*i.e.,* if $f = 10$, every ten passes over the p rows $\mathrm{B}_{j,:}$). This minimization scheme is summarized in Algorithm 1, and details of the updates formulas are given in Appendix A.3.

Algorithm 1: ALTERNATE MIN. FOR MULTI-TASK SGCL

input : $X, Y, \underline{\Sigma}, \lambda, f, T$
init : $\mathrm{B} = 0_{p,q}$, $\Sigma^{-1} = \underline{\Sigma}^{-1}$, $R = Y$
for iter $= 1, \dots, T$ **do**
  if iter $= 1 \pmod f$ **then**
    $\Sigma \leftarrow \Psi(R, \underline{\Sigma})$ `// update of Proposition 2`
    for $j = 1, \dots, p$ **do**
      $L_j = X_j^\top \Sigma^{-1} X_j$
  for $j = 1, \dots, p$ **do**
    $R \leftarrow R + X_j \mathrm{B}_j$ `// partial residual update`
    $\mathrm{B}_j \leftarrow \mathrm{BST}\left(\frac{X_j^\top \Sigma^{-1} R}{L_j}, \frac{\lambda nq}{L_j}\right)$ `// coef. update`
    $R \leftarrow R - X_j \mathrm{B}_j$ `// residual update`
return B, Σ

Since strong duality holds for (4), we use the duality gap as a stopping criterion for the convergence. Every f epochs of BCD as presented in Algorithm 1, we compute a dual point Θ (see Remark 2), evaluate the duality gap, and stop if it is lower than $\epsilon = 10^{-6} / \|Y\|$. The pair (B, Σ) obtained when the duality gap goes below ϵ is then guaranteed to be an ϵ-solution of (4).

3.3 Statistical limitations

We provided the most general framework to adapt the Concomitant Lasso to multi-task and non scalar covariances. However, in its general formulation the multi-task Smoothed Generalized Concomitant Lasso has an obvious drawback: in practice estimating Σ^* requires to fit $n(n-1)/2$ parameters with only nq observations, which is problematic if q is not large enough. Hence, additional regularization might be needed to provide an accurate estimator of Σ^*, *e.g.,* following the direction proposed in Ledoit and Wolf [2004].

Nevertheless, in common practical scenarios Σ^* can be assumed to have a more regular structure. In the following, we address in details a special case of heteroscedastic models, where the noise is de-correlated and has a known block-wise structure.

4 Block homoscedastic model

Motivated by the specificities of the M/EEG inverse problem (see Section 5.3), but more generally by supervised learning problems where data come from an identified, finite set of sources, we propose a specification of (3) when more assumptions can be made about the noise. Indeed, when the observations come from K different sources or K types of sensors (in the M/EEG case: magnetometers, gradiometers and electrodes), Σ can be estimated in a simplified way. Assuming inde-

pendent noise among data sources or sensor types, we propose a variant of (3), called the *block homoscedastic* model. In this model, Σ^* is constrained to be diagonal, the diagonal being constant over known blocks.

Formally, if the k-th group of sensors is composed of n_k sensors ($\sum_1^K n_k = n$), with design matrix $X^k \in \mathbb{R}^{n_k \times p}$, observation matrix $Y^k \in \mathbb{R}^{n_k \times q}$ and noise level $\sigma_k^* > 0$, the block homoscedastic model is a combination of K homoscedastic models: $\forall k \in [K], \quad Y^k = X^k \mathrm{B}^* + \sigma_k^* \mathrm{E}^k$, with the entries of E^k independently sampled from $\mathcal{N}(0,1)$. In the following we denote

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

$$\Sigma^* = \begin{pmatrix} \sigma_1^* \mathrm{I}_{n_1} & & 0 \\ & \ddots & \\ 0 & & \sigma_K^* \mathrm{I}_{n_K} \end{pmatrix} \in \mathbb{S}^n_{++}.$$

Following this model, we call multi-task Smoothed Block Homoscedastic Concomitant Lasso (multi-task SBHCL) the estimator similar to (4) with the additional constraint that Σ is a diagonal matrix $\mathrm{diag}(\sigma_1 \mathrm{I}_{n_1}, \dots, \sigma_K \mathrm{I}_{n_K})$, with K constraints $\sigma_k \geq \underline{\sigma}_k$:

$$\underset{\substack{\mathrm{B} \in \mathbb{R}^{p \times q}, \\ \sigma_1, \dots, \sigma_K \in \mathbb{R}^K_{++} \\ \sigma_k \geq \underline{\sigma}_k, \forall k \in [K]}}{\arg\min} \; \sum_{k=1}^K \left(\frac{\left\| Y^k - X^k \mathrm{B} \right\|^2}{2nq\sigma_k} + \frac{n_k \sigma_k}{2n} \right) + \lambda \left\| \mathrm{B} \right\|_{2,1} \,. \tag{11}$$

Since (11) does not admit a closed-form solution, we also propose an iterative solver (see Section 3.2), along with a stopping condition based on the duality gap, which is derived for this problem in Appendix A.4.

- When the constraints on the σ_k's are not saturated at optimality, formulation (11) has an equivalent square-root Lasso [Belloni et al., 2011] formulation: $\arg\min_{\mathrm{B} \in \mathbb{R}^{p \times q}} \frac{1}{nq} \sum_{k=1}^K \sqrt{n_k} \left\| Y^k - X^k \mathrm{B} \right\| + \lambda \left\| \mathrm{B} \right\|_{2,1}$.
- To fix the values of the lower bounds on the noise levels σ_k, we use an arbitrary proportion of the initial estimation of the noise variances per block *i.e.*, $\underline{\Sigma} = 10^{-\alpha} \, \mathrm{diag}(\|Y^1\|/\sqrt{n_1 q} \, \mathrm{I}_{n_1}, \dots, \|Y^K\|/\sqrt{n_K q} \, \mathrm{I}_{n_K})$. $\alpha = 3$ is used in the experiments.

The equivalents of Theorem 1, Proposition 2 and Proposition 3 for the multi-task SBHCL are:

Theorem 2. *The dual formulation of* (11) *is*

$$\hat{\Theta} = \underset{\Theta \in \Delta'_{X,\lambda}}{\arg\max} \langle Y, \lambda \Theta \rangle + \sum_{k=1}^K \frac{\underline{\sigma}_k}{2} \left(\frac{n_k}{n} - nq\lambda^2 \|\Theta^k\|^2 \right) \,,$$

where $\Delta'_{X,\lambda}$ *is defined by*

$$\Delta'_{X,\lambda} = \Big\{ \Theta \in \mathbb{R}^{n \times q} : \\ \|X^\top \Theta\|_{2,\infty} \leq 1, \forall k \in [K], \|\Theta^k\| \leq \tfrac{\sqrt{n_k}}{n\lambda\sqrt{q}} \Big\}.$$

Proposition 4. *When optimizing* (11) *with* $\hat{\mathrm{B}}$ *being fixed, then* $\hat{\Sigma} = \mathrm{diag}(\hat{\sigma}_1 \mathrm{I}_{n_1}, \dots, \hat{\sigma}_K \mathrm{I}_{n_K})$, *with residuals* $R^k = Y^k - X^k \hat{\mathrm{B}}$ *and* $\hat{\sigma}_k = \underline{\sigma}_k \vee (\|R^k\|/\sqrt{n_k q})$.

Proposition 5. *For the multi-task SBHCL the critical parameter is* $\lambda_{\max} := \frac{1}{nq} \|X^\top \Sigma_{\max}^{-1} Y\|_{2,\infty}$ *where* $\Sigma_{\max} = \mathrm{diag}(\sigma_1^{\max} \mathrm{I}_{n_1}, \dots, \sigma_K^{\max} \mathrm{I}_{n_K})$ *and* $\forall k \in [K], \sigma_k^{\max} = \underline{\sigma}_k \vee (\|Y^k\| / \sqrt{n_k q})$.

The proofs are similar to those of Proposition 2 and Proposition 3 and delayed to Appendix A.4.

The strategy of Algorithm 1 can also be applied to the multi-task SBHCL. Because of the special form of Σ, the computations are lighter and the standard deviations σ_k's can be updated at each coordinate descent update. Indeed, updating all the σ_k's may seem costly, since a naive implementation requires to recompute all the residual norms $\|R^k\|$, where $R^k = Y^k - X^k \mathrm{B}$, which is $\mathcal{O}(nq)$. However, it is possible to store the values of $\|R^k\|^2$ and update them at each B_j update with a $\mathcal{O}(kq)$ cost. Indeed, if we denote $\tilde{\mathrm{B}}_j$ and $\tilde{R}^k$ the values before the update, we have:

$$R^k = \tilde{R}^k + X_j^{k\,\top} (\tilde{\mathrm{B}}_j - \mathrm{B}_j)$$
$$\left\| R^k \right\|^2 = \|\tilde{R}^k\|^2 + 2 \, \mathrm{Tr}[(\tilde{\mathrm{B}}_j - \mathrm{B}_j) \tilde{R}^{k\,\top} X_j^k] \\ + \|\tilde{\mathrm{B}}_j - \mathrm{B}_j\|^2 L_{j,k}$$

and all the quantities $\tilde{R}^{k\,\top} X_j^k$ are already computed for the soft-thresholding step. As $k \leq n$, this makes the cost of one B_j update of Algorithm 2 $\mathcal{O}(nq)$, the same cost as for the $\ell_{2,1}$ regularized Lasso, *a.k.a.* multi-task Lasso (MTL) [Obozinski et al., 2010].

5 Experiments

To demonstrate the benefits of handling non-homoscedastic noise, we now present experiments using both simulations and real M/EEG data. First, we show that taking into account multiple noise levels improves both prediction performance and support identification. We then illustrate on M/EEG data that the estimates of the noise standard deviations using multi-task SBHCL match the expected behavior when increasing the SNR of the data. We also demonstrate empirically the benefit of our proposed multi-task SBHCL to reduce the variance of the estimation. The implementation is done in Python/Cython and is available at https://github.com/mathurinm/SHCL.

Algorithm 2: ALTERNATE MIN. FOR MULTI-TASK SBHCL

input : $X^1, \dots, X^K, Y^1, \dots, Y^K, \underline{\sigma}_1, \dots, \underline{\sigma}_K, \lambda, T$
init : $\mathrm{B} = 0_{p,q}$,
$\forall k \in [K], \sigma_k = \|Y^k\|/\sqrt{n_k q}, R^k = Y^k$,
$\forall k \in [K], \forall j \in [p], L_{k,j} = \|X_j^k\|_2^2$
for iter $= 1, \dots, T$ **do**
  for $j = 1, \dots, p$ **do**
    for $k = 1, \dots, K$ **do**
      $R^k \leftarrow R^k + X_j^k \mathrm{B}_j$ // residual update
    $\mathrm{B}_j \leftarrow \mathrm{BST}\left(\sum_{k=1}^K \frac{X_j^{k\top} R^k}{\sigma_k}, \lambda n q\right) / \sum_{k=1}^K \frac{L_{k,j}}{\sigma_k}$
    // soft-thresholding
    for $k = 1, \dots, K$ **do**
      $R^k \leftarrow R^k - X_j^k \mathrm{B}_j$ // residual update
      $\sigma_k \leftarrow \underline{\sigma}_k \vee \frac{\|R^k\|}{\sqrt{n_k q}}$ // std dev update
return $\mathrm{B}, \sigma_1, \dots, \sigma_k$

We consider the case where the block structure of the noise is known by the practitioner. Therefore, all experiments use the block homoscedastic setting. Note that this is relevant with the M/EEG framework where the variability of the noise is due to different data acquisitions sensors that are known.

5.1 Prediction performance

We first study the impact of the multi-task SBHCL on prediction performance, evaluated on left-out data.

The experiment setup is as follows. There are $n = 300$ observations, $p = 1,000$ features and $q = 100$ tasks. The design X is random with Toeplitz-correlated features with parameter $\rho = 0.7$ (correlation between features i and j is $\rho^{|i-j|}$). The true coefficient matrix B^* has 20 non-zero rows, whose entries are independently and normally (centered and reduced) distributed. We simulate data coming from $K = 3$ sources (each one containing 100 observations) whose respective noise levels are σ^*, $2\sigma^\star$ and $5\sigma^\star$. The standard deviation $\sigma^\star$ is chosen so that the signal-to-noise ratio

$$\mathrm{SNR} := \|Y\|/\|X\mathrm{B}^*\| = 1 \ .$$

The two estimators are trained for λ varying on a logarithmic grid of 15 values between the critical parameter[2] $\lambda_{\max}$ and $\lambda_{\max}/10$. The training set contains 150 samples ($n_1 = n_2 = n_3 = 50$ of each data source) and the test set consists of the remaining 150.

Figure 1 shows prediction performance for the Smooth Concomitant Lasso (SCL), which estimates a single noise level for all blocks, and the multi-task SBHCL. Since each block has a different noise level, for each block k and each estimator, we report the Root Mean Squared error (RMSE, $\|Y^k - X^k\hat{\mathrm{B}}\|/\sqrt{qn_k}$) normalized by the oracle RMSE ($\|Y^k - X^k\mathrm{B}^*\|/\sqrt{qn_k}$). After taking the log, zero value means a perfect estimation, a positive value means under-fitting of the block, while a negative value corresponds to over-fitting. Figure 1 reports normalized RMSE values on both the training and the test data.

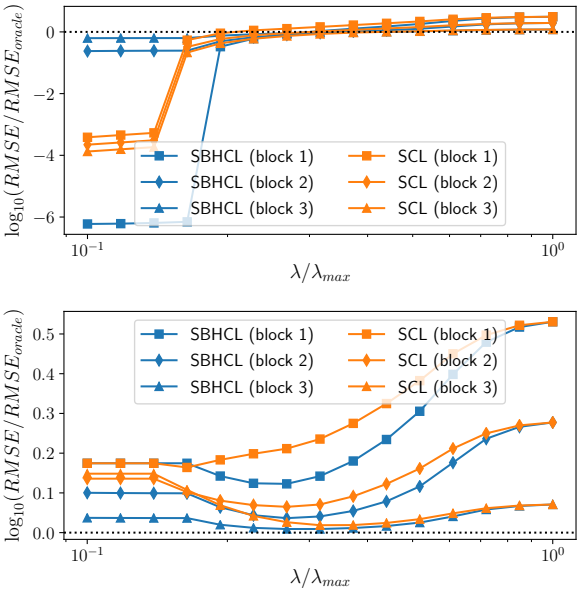

Figure 1: RMSE normalized by oracle RMSE, per block, for the multi-task SBHCL and Smooth Concomitant Lasso (SCL), on training (top) and testing (bottom) set, for various values of λ. The flexibility of the block homoscedastic model enables the multi-task Smoothed Block Homoscedastic Concomitant Lasso to reach a lower RMSE on every block of the test set.

As it can be observed, the RMSE for the multi-task SBHCL is lower on every block of the test set, meaning that it has better prediction performance. By attributing a higher noise standard deviation to the noisiest block (block 3), the multi-task SBHCL is able to down-weight the impact of these samples on the estimation, while still benefiting from it.

While the 3 normalized RMSE have similar behaviors on the test set for the SCL, for low values of λ, the multi-task SBHCL overfits more on the least noisy block. However this does not result in degraded prediction performance on the test set, neither for this block nor for others, and the prediction is even better on the noisiest block. Indeed, the SCL overfits more on the the noisiest block, which has a greater impact on prediction (as overfitting on noiseless data would lead to

[2] Note that $\lambda_{\max}$ is model specific

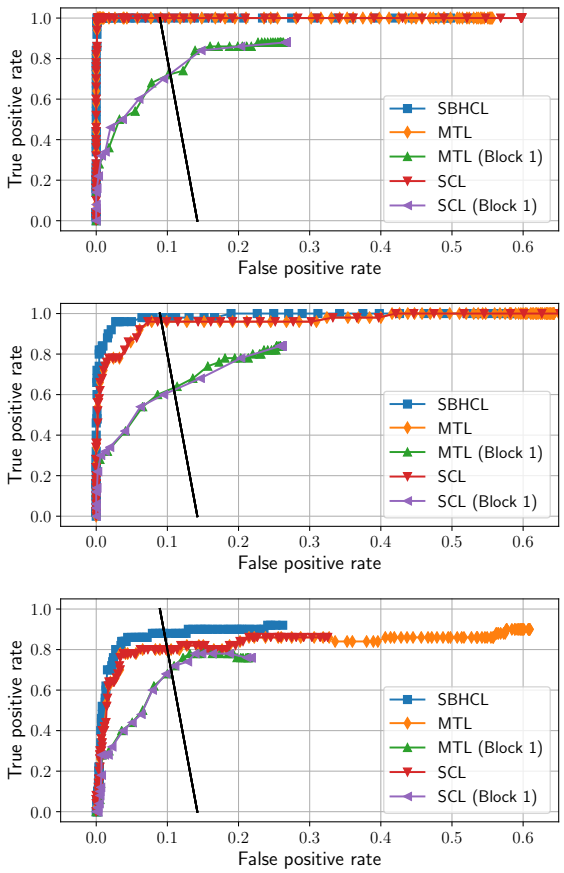

Figure 2: ROC curves of true support recovery for the SBHCL, the MTL and SCL on all blocks, and the MTL and SCL on the least noisy block. The black curve marks the limit of supports of size $0.9n$. Top: SNR $= 5$, $\rho = 0.1$, middle: SNR$=1$, $\rho = 0.1$, bottom: SNR $= 1$, $\rho = 0.9$.

perfect parameter inference). When the regularization parameter becomes too low, taking into account different noise levels allows our estimator to limit the impact of overfitting by favoring the most reliable source. This experiments shows that our formulation is appealing for parameter selection, as the best left-out prediction is obtained for similar values of λ.

5.2 Support recovery

In this experiment, we demonstrate the superior performance of the multi-task SBHCL for support recovery, *i.e.,* its ability to correctly identify the predictive features. The experimental setup is the same as in Section 5.1, except that the support of B^* is of size 50. We also vary $\rho \in \{0.1, 0.9\}$ and the SNR $\in \{1, 5\}$ (additional results are included in Appendix C).

The five estimators compared on Figure 2 are the multi-task SBHCL, the SCL, the MTL, and also the MTL and the SCL trained on the least noisy block (*i.e.,* the most favorable block). Following the empirical evaluation from [Bühlmann and Mandozzi, 2014], the figure of merit is the ROC curve, *i.e.,* the true positive rate as a function of the false positive rate. The curve is obtained by varying the value of λ (lower values leading to larger predicted support and therefore potentially more false positives).

Table 1: Mean values of pAUC for the main estimators, across ten simulations of X and Y.

SNR	1	1	5
ρ	0.1	0.9	0.1
SBHCL	0.92 ± 0.12	0.86 ± 0.05	0.98 ± 0.02
MTL	0.79 ± 0.08	0.71 ± 0.07	0.99 ± 0.00
MTL (block 1)	0.44 ± 0.04	0.48 ± 0.05	0.48 ± 0.04

We can see that when the SNR is sufficiently high (top graph with $SNR = 5$), the multi-task SBHCL, the SCL and the MTL successfully recover the true support, while the MTL or SCL trained on the least noisy block with only one third of the data fails. However, when the SNR is lower (middle graph with SNR $= 1$), the multi-task SBHCL still achieves almost perfect support identification, while the performance of the MTL and SCL decreases. The performance is naturally even worse when using only one block of samples. Finally, when the features are more correlated ($\rho = 0.9$) and the conditioning of X is degraded, the multi-task SBHCL, despite not perfectly recovering the true support, still has superior performance. Note also that unsurprisingly the MTL and the SCL lead to almost perfectly the same ROC curves as both estimators (if $\underline{\sigma}$ is small enough) have the same solution path. Any difference between SCL and MTL in our graph is due to the choice of a discrete set of λ values.

To study the stability of Figure 2, we repeat the simulation 10 times. Since the curves are not guaranteed to reach TPR $= 1$, it is not possible to use AUC as a scalar figure of merit. As we are usually interested in sparse estimators when the recovered support is small, we follow Bühlmann and Mandozzi [2014, Fig. 1-4], and limit the study to estimated supports of size inferior to $0.9n$ (*i.e.,* the part to the left of the black curve). This leads to the use of pAUC or "0.9–performance": the area under the ROC curve, but restricted to the left of the black line, and normalized by its maximal value. The mean pAUCs for 10 repetitions, for all estimators in the different settings are in Table 1.

5.3 Results on joint M/EEG real data

We now evaluate our estimator on magneto- and electroencephalography (M/EEG) data. The data consists

of M/EEG recordings, which measure the electric potential and magnetic field induced by active neurons. Data are time-series so that n corresponds to the number of sensors and q to the number of consecutive time instants in the data. Thanks to their high temporal resolution, M/EEG help to elucidate where and precisely when cognitive processes happen in the brain [Baillet, 2017]. The so-called M/EEG inverse problem, which consists in identifying active brain regions, can be cast as a high-dimensional sparse regression problem. Because of the limited number of sensors, as well as the physics of the problem, this problem is severely ill-posed, and regularization is needed to provide solutions which are both biologically plausible and robust to measurement noise [Wipf et al., 2008, Haufe et al., 2008, Gramfort et al., 2013]. As foci of neural activity are observed from a distance by M/EEG and since only a small number of brain regions are involved in a cognitive task during a short time interval, it is common to employ sparsity-promoting regularizations. Amongst these, the ℓ_1/ℓ_2 penalty has been successfully applied to the M/EEG inverse problem in either time [Ou et al., 2009] or frequency domain [Gramfort et al., 2013].

The experimental condition considered is a monaural auditory stimulation in the right ear of the subject. The same subject undergoes the same stimulation 61 times, and the M/EEG measurements are recorded from 0.2 s before to 0.5 s after the stimulus. The data (from the MNE software [Gramfort et al., 2014]) thus contains 61 repetitions (*trials*) of this stimulation.

In the experimental setup we have 204 gradiometers, 102 magnetometers and 60 EEG electrodes. We have discarded one magnetometer and one electrode corrupted by strong artifacts. We therefore have $K = 3$ sensor types with $n_1 = 203$, $n_2 = 102$ and $n_3 = 59$ (so $n = 364$). X is obtained by numerically solving the M/EEG forward problem using $p = 1884$ candidate sources distributed over the cortical surface ($X \in \mathbb{R}^{364\times1884}$). The orientations of the dipoles are assumed known and normal to the cortical mantle.

The measurements for $q = 1$ (single time measurements) are selected 75 ms after the stimulus onset, and between 60 and 115 ms (resp. 70 and 102 ms) after the stimulus for $q = 34$ (resp $q = 20$). This time interval corresponds to the main cortical response to the auditory stimulation.

For a number t of repetitions of experiment (t ranging from 2 to 56), we create an observation matrix Y_t by averaging the first t trials. By doing so, the noise standard deviations of each block should be proportional to $1/\sqrt{t}$. We then run the multi-task SBHCL with fixed λ, equal to 3% of $\lambda_{\max}$. Figure 3 shows the noise standard deviation estimated by the multi-task SBHCL, when ran on a single time instant (single task), 20 and 34 time instants.

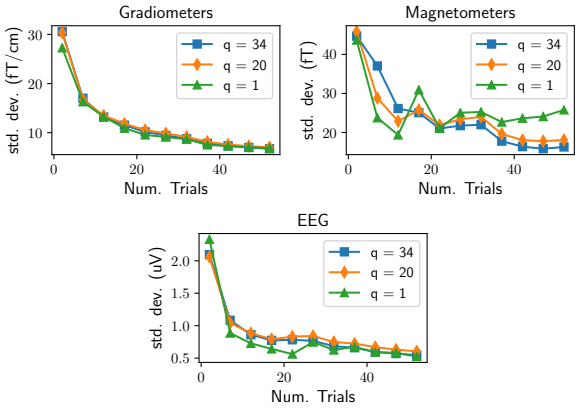

Figure 3: Noise standard deviation estimated on auditory data for $q = 1$, $q = 20$ and $q = 34$ time instants using the SBHCL estimator. Data consist of combined MEG gradiometers ($n_1 = 203$ sensors) and magnetometers ($n_2 = 102$ sensors), as well as EEG ($n_3 = 59$ sensors). We used $\lambda = 0.03\lambda_{\max}$.

We can see that the estimated values are plausible: they have the correct orders of magnitude, as well as the expected $1/\sqrt{t}$ decrease. We also see that taking more tasks into account leads to more stable noise estimation: for magnetometers, the curve is smoother for $q = 34$ than for $q = 20$ and $q = 1$. Indeed, using more tasks reduces the variance of the estimation.

6 Conclusion

This work proposes the multi-task Smoothed Generalized Concomitant Lasso, a new sparse regression estimator designed to deal with heterogeneous observations coming from different origins and corrupted by different levels of noise. Despite the joint estimation of the regression coefficients as well as the noise level, the problem considered is jointly convex, thus guaranteeing global convergence which one can check by duality gap certificates. The efficient BCD strategy we proposed leads to a computational complexity not higher than the one observed for a classic sparse regression model, while solving a fundamental practical problem. Indeed with the SBHCL, the regularization parameter is less sensitive to the noise level of each combined modality, making it easier to tune across experimental conditions and datasets. Finally, thanks to the flexibility of our model, better prediction performance and support recovery are achieved *w.r.t.* traditional homoscedastic estimators.

Acknowledgments

This work was funded by ERC Starting Grant SLAB ERC-YStG-676943 and by the chair Machine Learning for Big Data of Télécom ParisTech.

References

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

S. S. Chen and D. L. Donoho. Atomic decomposition by basis pursuit. In *SPIE*, 1995.

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

N. Städler, P. Bühlmann, and S. van de Geer. ℓ_1-penalization for mixture regression models. *TEST*, 19(2):209–256, 2010.

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

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

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

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

S. Reid, R. Tibshirani, and J. Friedman. A study of error variance estimation in lasso regression. *Stat. Sin.*, 26(1):35–67, 2016.

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

S. R. Becker, E. J. Candès, and M. C. Grant. Templates for convex cone problems with applications to sparse signal recovery. *Math. Program. Comput.*, 3(3):165–218, 2011.

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

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

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

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

R. Engle. Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica*, 50(4):987–1007, 1982.

R. J. Carroll and D. Ruppert. *Transformation and weighting in regression*, volume 30. CRC Press, 1988.

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

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

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

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.

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

P. L. Combettes and C. L. Müller. Perspective functions: Proximal calculus and applications in high-dimensional statistics. *J. Math. Anal. Appl.*, 2016.

J. Lederer and C. L. Müller. Don't fall for tuning parameters: Tuning-free variable selection in high dimensions with the TREX. In *AAAI*, pages 2729–2735, 2015.

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

O. Ledoit and M. Wolf. A well-conditioned estimator for large-dimensional covariance matrices. *J. Multivariate Anal.*, 88(2):365–411, 2004.

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, Apr 2010.

P. Bühlmann and J. Mandozzi. High-dimensional variable screening and bias in subsequent inference, with an empirical comparison. *Computational Statistics*, 29(3):407–430, Jun 2014.

S. Baillet. Magnetoencephalography for brain electrophysiology and imaging. *Nat. Neurosci.*, 20(3):327–339, 03 2017.

D. P. Wipf, J. P. Owen, H. Attias, K. Sekihara, and S. S. Nagarajan. Estimating the location and orientation of complex, correlated neural activity using MEG. In *NIPS*, pages 1777–1784. 2008.

S. Haufe, V. V. Nikulin, A. Ziehe, K.-R. Müller, and G. Nolte. Combining sparsity and rotational invariance in EEG/MEG source reconstruction. *NeuroImage*, 42(2):726–738, Aug. 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.

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

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.