

# Sparsity with sign-coherent groups of variables via the cooperative-Lasso

Julien Chiquet

Laboratoire Statistique et Génome  
Université d'Évry-Val-d'Essonne, UMR CNRS 8071 – USC INRA

Yves Grandvalet

Université de Technologie de Compiègne & CNRS, Heudiasyc UMR 6599, France

Camille Charbonnier

Laboratoire Statistique et Génome  
Université d'Évry-Val-d'Essonne, UMR CNRS 8071 – USC INRA

**Abstract.** We consider the problems of estimation and selection of parameters endowed with a known group structure, when the groups are assumed to be *sign-coherent*, that is, gathering either non-negative, non-positive or null parameters. To tackle this problem we propose a new penalty that we call the *cooperative-Lasso* penalty. We derive the optimality conditions defining the cooperative-Lasso estimate for generalized linear models and propose an efficient active set algorithm suited to high-dimensional problems. We study the asymptotic consistency of the estimator in the linear regression setup and derive its irrepresentable conditions, which are milder than the ones of the group-Lasso regarding the matching of groups with the sparsity pattern of the true parameters. We also address the problem of model selection in linear regression by deriving an approximation of the degrees of freedom of the cooperative-Lasso estimator. Simulations comparing the proposed estimator to the group-Lasso comply with our theoretical results, showing consistent improvements in support recovery for sign-coherent groups. We finally propose an approach widely applicable to the processing of genomic data, where the set of differentially expressed probes is enriched by incorporating all the probes of the microarray that are related to the corresponding genes. In an application to the estimation of chemotherapy pathologic response in breast cancer, the cooperative-Lasso demonstrates much better performances than its competitors.

*Keywords:* Penalization, Sparsity, Grouped variables, Sign-coherence, Microarray analysis

## 1. Introduction

This paper addresses the problems of estimation and inference of parameters when a group structure among parameters is known. We propose a new penalty for the case where the groups are assumed to gather either non-positive, non-negative or null parameters. All such groups will be referred to as *sign-coherent*.

As the main motivating example, we consider the linear regression model

$$Y = X\beta^* + \varepsilon = \sum_{k=1}^K \sum_{j \in \mathcal{G}_k} X_j \beta_j^* + \varepsilon, \quad (1)$$

where  $Y$  is a continuous response variable,  $X = (X_1, \dots, X_p)$  is a vector of  $p$  predictor variables,  $\beta^*$  is the vector of unknown parameters and  $\varepsilon$  is a zero-mean Gaussian error variable with variance  $\sigma^2$ . The set of indexes  $\{1, \dots, p\}$  is partitioned into  $K$  groups  $\{\mathcal{G}_k\}_{k=1}^K$  corresponding

to predictors and parameters. We will assume throughout this paper that  $\beta^*$  has few non-zero coefficients, with sparsity and sign patterns governed by the groups  $\mathcal{G}_k$ , that is, groups being likely to gather either positive, negative or null parameters.

The estimation and inference of  $\beta^*$  is based on training data, consisting of a vector  $\mathbf{y} = (y_1, \dots, y_n)^\top$  for the response variable and a  $n \times p$  design matrix  $\mathbf{X}$  whose  $j$ th column contains  $\mathbf{x}_j = (x_j^1, \dots, x_j^n)^\top$ , the  $n$  observations for variable  $X_j$ . For clarity, we assume that both  $\mathbf{y}$  and  $\{\mathbf{x}_j\}_{j=1, \dots, p}$  are centered so as to eliminate the intercept from fitting criteria.

Penalization methods that build on the  $\ell_1$ -norm, referred to as *Lasso* procedures (Least Absolute Shrinkage and Selection Operator), are now widely used to tackle simultaneously variable estimation and selection in sparse problems. Among these, the group-Lasso, independently proposed by Grandvalet and Canu (1999) and Bakin (1999) and developed in Yuan and Lin (2006), uses the group structure to define a shrinkage estimator of the form

$$\hat{\beta}^{\text{group}} = \arg \min_{\beta \in \mathbb{R}^p} \left\{ \frac{1}{2} \|\mathbf{y} - \mathbf{X}\beta\|^2 + \lambda \sum_{k=1}^K w_k \|\beta_{\mathcal{G}_k}\| \right\}, \quad (2)$$

where  $\mathcal{G}_k$  is the subset of indices defining the  $k$ th group of variables and  $\|\cdot\|$  is the Euclidean norm. The tuning parameter  $\lambda \geq 0$  controls the overall amount of penalty and weights  $w_k > 0$  adapt the level of penalty within a given group. Typically, one sets  $w_k = \sqrt{p_k}$  in order to adjust shrinkage according to group sizes. The penalizer in (2) is known to induce sparsity at the group level, setting a whole group of parameters to zero for values of  $\lambda$  which are large enough. Note that when we assign one group to each predictor, we recover the original Lasso of Tibshirani (1996).

The algorithms for finding the group-Lasso estimator have considerably improved recently. Foygel and Drton (2010) develop a block-wise algorithm, where each group of coefficients is updated at a time, using a single line search that provides the exact optimal value for one group, considering all other coefficients fixed. Meier et al. (2008) depart from linear regression in Problem (2) by studying group-Lasso penalties for logistic regression. Their block-coordinate descent method is applicable to generalized linear models. In the same context, Roth and Fischer (2008) propose a very efficient subset algorithm based on subdifferential calculus. They also formulate conditions for the uniqueness of group-Lasso solutions.

The consistency of the group-Lasso estimator has been extensively studied by Bach (2008) and Nardi and Rinaldo (2008). Their theorems assume that the set of non-zero coefficients of  $\beta^*$  is an exact union of groups. This means that not a single zero coefficient should belong to a group having some non-zero coefficients. Huang and Zhang (2010) introduce the concept of strong group sparsity, which measures how well a given group structure summarizes the sparsity pattern of  $\beta^*$ . Besides establishing conditions where the group-Lasso outperforms the classical Lasso, they show that an incorrect group structure, which cannot partition the zero and non-zero coefficients of  $\beta^*$ , has detrimental effects compared to the Lasso. Friedman et al. (2010a) propose to overcome this restriction by adding an  $\ell_1$  penalty to the objective function in (2). The new term induces sparsity at the component level to complement the sparsity induced at the group level, yet it requires an additional tuning parameter.

As stated at the beginning of this section, this paper deals with a strongest assumption regarding the group structure. Groups should not only reveal the sparsity pattern, but they should also be relevant for sign patterns: all coefficients within a group should be sign-coherent, that is, they should either be null, non-positive or non-negative. This desideratum arises often when the groups gather redundant or consonant variables (a usual outcome when groups are defined from clusters of correlated variables). It turns out that our stronger requisite leads to more ubiquitous consistency conditions: our assumptions are more flexible.

To enhance the group-Lasso with sign-coherent variable selection, letting the possibility to set some variables to zero within a group, we propose a novel penalty that we call the cooperative-Lasso, in short the *coop-Lasso*. Our approach does not require any additional parameter and can benefit from the algorithmic advances that were developed for the group-Lasso. Here, we build upon the subdifferential calculus approach originally proposed by Osborne et al. (2000) for the Lasso, which has been adapted by Roth and Fischer (2008) for the group-Lasso. Our approach is amenable to sample selection patterns that cannot be achieved with the group-Lasso. It thus leads to consistency results under mildest assumptions, where exact support recovery may be achieved when some zero coefficients belong to a group having either positive or negative coefficients otherwise. Note that we already applied this type of penalizer for the joint inference of several network structures (Chiquet et al., 2011). The present article examines the coop-Lasso in greater depth, with a theoretical analysis of selection consistency, new insights on shrinkage effects and proposals regarding model selection issues. New experimental results with simulated and real data also demonstrate the benefits of the coop-Lasso in regression and classification problems.

The rest of the paper is organized as follows: Section 2 presents the coop-Lasso penalty, with geometrical comparisons with the group-Lasso. We derive the optimality conditions which are the basis for a subset algorithm. The latter is easily extended to generalized linear models such as logistic regression. We also study the orthonormal design case that links the coop-Lasso estimator to the ordinary least squares estimator. Consistency results and the associated irrepresentable conditions are given in Section 3, supported by a simulation study. In Section 4, we address the problem of model selection by deriving an approximation of the degrees of freedom that can be used in the Bayesian Information Criterion (BIC) and Akaike Information Criterion (AIC). Section 5 is dedicated to numerical studies that are reminiscent of Breiman (1996)'s proposal. We provide quantitative results demonstrating the benefits of the coop-Lasso in terms of sign-coherence and prediction error. Section 6 investigates an application to the estimation of chemotherapy pathologic response in breast cancer, the cooperative-Lasso demonstrating much better performances than its competitors.

All proofs are postponed to the Appendix Section.

## 2. Cooperative-Lasso

### 2.1. Definitions and optimality conditions

*Group-norm and coop-norm.* We define a group structure by setting a partition of the index set  $\mathcal{I} = \{1, \dots, p\}$ , that is,

$$\mathcal{I} = \bigcup_{k=1}^K \mathcal{G}_k, \text{ with } \mathcal{G}_k \cap \mathcal{G}_\ell = \emptyset \text{ for } k \neq \ell .$$

Let  $\mathbf{v} = (v_1, \dots, v_p)^\top \in \mathbb{R}^p$  and  $p_k$  denote the cardinality of group  $k$ . We define  $\mathbf{v}_{\mathcal{G}_k} \in \mathbb{R}^{p_k}$  as the vector  $(v_j)_{j \in \mathcal{G}_k}$ . For the chosen groups  $\{\mathcal{G}_k\}_{k=1}^K$ , the group-Lasso norm reads

$$\|\mathbf{v}\|_{\text{group}} = \sum_{k=1}^K w_k \|\mathbf{v}_{\mathcal{G}_k}\|, \quad (3)$$

where  $w_k > 0$  are fixed parameters enabling to adapt the amount of penalty for each group.

Let  $\mathbf{v}^+ = (v_1^+, \dots, v_p^+)^\top$  and  $\mathbf{v}^- = (v_1^-, \dots, v_p^-)^\top$  be the componentwise positive and negative part of  $\mathbf{v}$ , that is,  $v_j^+ = \max(0, v_j)$  and  $v_j^- = \max(0, -v_j)$  respectively. We call *coop-norm* of  $\mathbf{v}$

the sum of group-norms on  $\mathbf{v}^+$  and  $\mathbf{v}^-$

$$\|\mathbf{v}\|_{\text{coop}} = \|\mathbf{v}^+\|_{\text{group}} + \|\mathbf{v}^-\|_{\text{group}} = \sum_{k=1}^K w_k (\|\mathbf{v}_{\mathcal{G}_k}^+\| + \|\mathbf{v}_{\mathcal{G}_k}^-\|) ,$$

which is clearly a norm on  $\mathbb{R}^p$ .

The coop-Lasso estimate of  $\boldsymbol{\beta}^*$  as defined in (1) is

$$\hat{\boldsymbol{\beta}}^{\text{coop}} = \arg \min_{\boldsymbol{\beta} \in \mathbb{R}^p} L(\boldsymbol{\beta}), \quad \text{with } L(\boldsymbol{\beta}) = \frac{1}{2} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda \|\boldsymbol{\beta}\|_{\text{coop}} , \quad (4)$$

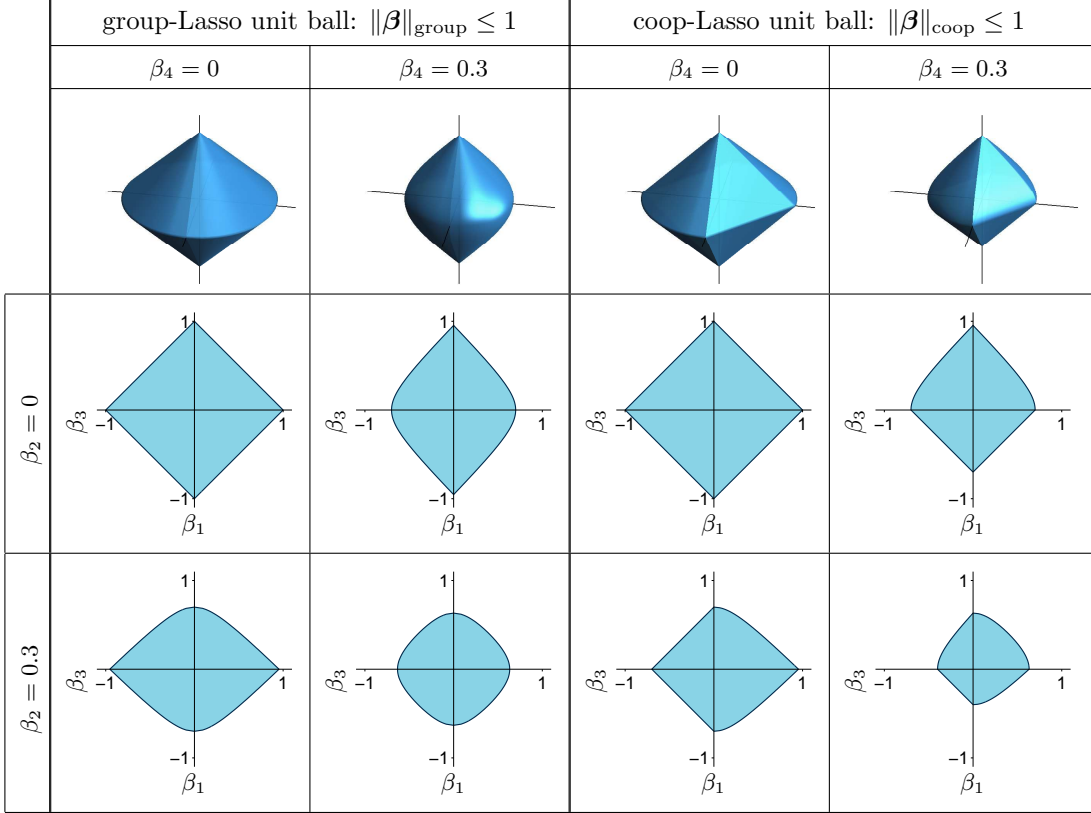
where  $\lambda \geq 0$  is a tuning parameter common to all groups. Appropriate choices for  $\lambda$  will be discussed in Sections 4 and 3 dealing with model selection and consistency respectively.

Illustrations of the group-norm and coop-norm are given in Figure 1 for a vector  $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3, \beta_4)^\top$  with two groups  $\mathcal{G}_1 = \{1, 2\}$  and  $\mathcal{G}_2 = \{3, 4\}$ . We represent several views of the unit balls for the group-norm and the coop-norm, that is,  $\|\boldsymbol{\beta}\|_{\text{group}} \leq 1$  and  $\|\boldsymbol{\beta}\|_{\text{coop}} \leq 1$ . These balls represent the set of feasible solutions for an optimization problem equivalent to (2) and (4), where the sum of squared residuals is minimized under unitary constraints on  $\|\boldsymbol{\beta}\|_{\text{group}}$  and  $\|\boldsymbol{\beta}\|_{\text{coop}}$  respectively.

These plots provide some insight into the sparsity pattern that originates from the penalties, since sparsity is related to the singularities of the boundary of the feasible set. First, consider the group-Lasso: the first column illustrates that when  $\beta_4$  is null its group companion  $\beta_3$  may also be exactly zero (corners on the boundary at  $\beta_3 = 0$ ), while the second column shows that this event is improbable when  $\beta_4$  differs from zero (smooth boundary at  $\beta_3 = 0$ ). The second and third rows display the same type of relationships within  $\mathcal{G}_1$  between  $\beta_2$  and  $\beta_1$ , which are expected due to the symmetries of the unit ball. Now, consider the coop-norm: we see that there are additional discontinuities resulting in new edges on the 3-D plots. As before, we observe that when  $\beta_4$  is null  $\beta_3$  may also be exactly zero, but in addition we may also have  $\beta_1$  or  $\beta_2$  exactly null. Accordingly, in the second and third rows we see that we may have  $\beta_3$  null when  $\beta_4$  is non-zero. These new edges result in some new opportunities to zero some coefficients when the group-Lasso would have allowed a solution with opposite signs within a group. The second crucial difference with the group-Lasso is the loss of the axial symmetry when some variables are non-zero: decoupling the positive and negative parts of the regression coefficients favors solutions where signs match within a group. The unit coop-norm ball covers identical volumes in the positive and negative orthant (where it is identical to the group-norm), but it is shrunk in the other ones, where there are some sign mismatches. The coop-Lasso penalty is more stringent than the group-Lasso penalty.

Before stating the optimality conditions for Problem (4), we introduce notation related to the sparsity pattern of parameters, which will be required to express the necessary and sufficient condition for optimality. First, we recall that the unknown vector of parameters  $\boldsymbol{\beta}^*$  is typically sparse; its support is denoted  $\mathcal{S} = \{j, \beta_j^* \neq 0\}$  and  $\mathcal{S}^c = \{j, \beta_j^* = 0\}$  is the complementary set of true zeros. Once the problem has been supplied with a group structure, we define  $\mathcal{S}_k = \mathcal{S} \cap \mathcal{G}_k$  and  $\mathcal{S}_k^c = \mathcal{S}^c \cap \mathcal{G}_k$  as the sets of relevant, respectively irrelevant, predictors within group  $k$ , for all  $k = 1, \dots, K$ . Similar notations  $\mathcal{S}(\boldsymbol{\beta})$ ,  $\mathcal{S}_k(\boldsymbol{\beta})$  and  $\mathcal{S}_k^c(\boldsymbol{\beta})$  are defined for an arbitrary vector  $\boldsymbol{\beta} \in \mathbb{R}^p$ . Furthermore, for clarity and brevity, we introduce the functions  $\{\varphi_j\}_{j=1}^p$ , which return the componentwise positive or negative part of a vector according to the sign of its  $j$ th element:

$$\forall k \in \{1, \dots, K\}, \forall j \in \mathcal{G}_k, \forall \mathbf{v} \in \mathbb{R}^{p_k}, \quad \varphi_j(\mathbf{v}) = (\text{sign}(v_j)\mathbf{v})^+ = \begin{cases} \mathbf{0} & \text{if } v_j = 0 , \\ \mathbf{v}^+ & \text{if } v_j > 0 , \\ \mathbf{v}^- & \text{if } v_j < 0 . \end{cases} \quad (5)$$



**Figure 1.** Feasible set for the group-Lasso and coop-Lasso penalties. Top row: cuts through  $(\beta_1, \beta_2, \beta_3)$  at  $\beta_4 = 0$  and  $\beta_4 = 0.3$ :  $(\beta_1, \beta_2)$  span the horizontal plane and  $\beta_3$  is on the vertical axis; bottom rows: cuts through  $(\beta_1, \beta_3)$  at various values of  $(\beta_2, \beta_4)$ .

*Optimality conditions.* The objective function  $L$  in (4) is continuous and coercive, thus Problem (4) admits at least one minimum. If  $\mathbf{X}$  has rank  $p$ , then the minimum is unique since  $L$  is strictly convex. Furthermore,  $L$  is smooth, except at some locations with zero coefficients, due to the singularities of the coop-norm. Since  $L$  is convex, a necessary and sufficient condition for the optimality of  $\beta$  is that the null vector  $\mathbf{0}$  belongs to the subdifferential of  $L$  whose expression is provided in the following lemma.

LEMMA 1. For all  $\beta \in \mathbb{R}^p$ , the subdifferential of the objective function of Problem (4) is

$$\partial_{\beta} L(\beta) = \{\mathbf{v} \in \mathbb{R}^p : \mathbf{v} = \mathbf{X}^{\top}(\mathbf{X}\beta - \mathbf{y}) + \lambda\boldsymbol{\theta}\} , \quad (6)$$

where  $\boldsymbol{\theta} \in \mathbb{R}^p$  is any vector belonging to the subdifferential of the coop-norm, that is:

$$\forall k \in \{1, \dots, K\}, \forall j \in \mathcal{S}_k(\beta), \quad \theta_j = \frac{w_k \beta_j}{\|\varphi_j(\beta_{\mathcal{G}_k})\|} , \quad (7a)$$

$$\forall k \in \{1, \dots, K\}, \forall j \in \mathcal{S}_k^c(\beta), \quad \|\varphi_j(\boldsymbol{\theta}_{\mathcal{G}_k})\| \leq w_k . \quad (7b)$$

The following optimality conditions, which result directly from Lemma 1, are an essential building block of the algorithm we propose to compute the coop-Lasso estimate. They also

provide an important basis for showing the consistency results.

**THEOREM 1.** *Problem (4) admits at least one solution, which is unique if  $\mathbf{X}$  has rank  $p$ . All critical points  $\boldsymbol{\beta}$  of the objective function  $L$  verifying the following conditions are global minima:*

$$\forall k \in \{1, \dots, K\}, \forall j \in \mathcal{S}_k(\boldsymbol{\beta}), \quad \mathbf{x}_j^\top (\mathbf{X}\boldsymbol{\beta} - \mathbf{y}) + \frac{\lambda \omega_k \beta_j}{\|\boldsymbol{\varphi}_j(\boldsymbol{\beta}_{\mathcal{G}_k})\|} = 0, \quad (8a)$$

$$\forall k \in \{1, \dots, K\}, \forall j \in \mathcal{S}_k^c(\boldsymbol{\beta}), \quad \|\boldsymbol{\varphi}_j((\mathbf{X}_{\mathcal{G}_k})^\top (\mathbf{X}\boldsymbol{\beta} - \mathbf{y}))\| \leq \lambda \omega_k, \quad (8b)$$

where  $\mathbf{X}_{\mathcal{G}_k}$  is the submatrix of  $\mathbf{X}$  with all rows and columns indexed by  $\mathcal{G}_k$ .

Note here an important distinction compared to the group-Lasso, where the optimality conditions are expressed solely according to the groups  $\mathcal{G}_k$  (see e.g. Roth and Fischer, 2008). Hence, while the sparsity pattern of the solution is strongly constrained by the predefined group structure in the group-Lasso, deviations from this structure are possible for the coop-Lasso, since the groups  $\mathcal{S}_k$  and  $\mathcal{S}_k^c$ , though being also defined from  $\mathcal{G}_k$ , are adapted to the penalty parameter  $\lambda$  and to the data  $(\mathbf{X}, \mathbf{y})$ . The asymptotic analysis of Section 3 confirms that exact support recovery is possible even when the support of  $\boldsymbol{\beta}^*$  cannot be expressed as a simple union of groups, provided the groups intersecting the true support are sign-coherent.

## 2.2. Algorithm

The efficient approaches developed for the Lasso take advantage of the sparsity of the solution by solving a series of small linear systems, whose sizes are incrementally increased/decreased (Osborne et al., 2000). This approach was pursued for the group-Lasso (Roth and Fischer, 2008) and we proposed an algorithm in the same vein for the coop-Lasso in the framework of multiple network inference (Chiquet et al., 2011). We provide here a more detailed description of the latter in the specific context of linear regression.

The algorithm starts from a sparse initial guess, say  $\boldsymbol{\beta} = 0$ , and iterates two steps:

- (a) the first step solves Problem (4) with respect to  $\boldsymbol{\beta}_{\mathcal{A}}$ , the subset of “active” variables, currently identified as being non-zero. At this stage the current feasible set is restricted to the orthants where the gradient of the coop-norm has no discontinuities : the optimization problem is thus smooth. Some variables may then be declared inactive if the current optimal  $\boldsymbol{\beta}_{\mathcal{A}}$  reaches the boundary of the current feasible set.
- (b) the second step assesses the completeness of the set  $\mathcal{A}$ , by checking the optimality conditions with respect to inactive variables. We add the group that violates most these conditions. When no such violation exists, the current solution is optimal.

These two steps outline the algorithm, which is detailed in more technical terms in Algorithm 1. The principle is readily applied to any generalized linear model by simply defining the appropriate objective function  $L$ . In our current implementation (a pre-release of our R-package `scoop` is available at <http://stat.genopole.cnrs.fr/logiciels/scoop>) the linear and logistic regression models are implemented using Broyden–Fletcher–Goldfarb–Shanno (BFGS) quasi-

Newton updates with box constraints to solve the smooth optimization problem in Step a.

---

**Algorithm 1:** Coop-Lasso Fitting Algorithm

---

**Init.** Start from a feasible  $\boldsymbol{\beta} \leftarrow \boldsymbol{\beta}^0$

$$\mathcal{A}_+ \leftarrow \{j \in \mathcal{G}_k : \|\boldsymbol{\beta}_{\mathcal{G}_k}^+\| > 0, k = 1, \dots, K\}$$

$$\mathcal{A}_- \leftarrow \{j \in \mathcal{G}_k : \|\boldsymbol{\beta}_{\mathcal{G}_k}^-\| > 0, k = 1, \dots, K\}$$

**Step a** On  $\mathcal{A} \leftarrow \mathcal{A}_+ \cup \mathcal{A}_-$ , find a solution to the smooth problem

$$\boldsymbol{\beta}_{\mathcal{A}} \leftarrow \arg \min_{\mathbf{v} \in \mathbb{R}^{|\mathcal{A}|}} \frac{1}{2} \|\mathbf{y} - \mathbf{X}_{\mathcal{A}} \mathbf{v}\|^2 + \lambda \|\mathbf{v}\|_{\text{coop}} \quad \text{s.t.} \quad \begin{cases} v_j \geq 0 & \text{if } j \in \mathcal{A}_+ \cap \overline{\mathcal{A}}_- \\ v_j \leq 0 & \text{if } j \in \mathcal{A}_- \cap \overline{\mathcal{A}}_+ \end{cases}$$

Identify groups inactivated during optimization

$$\mathcal{A}_+ \leftarrow \mathcal{A}_+ \setminus \{j \in \mathcal{G}_k \subseteq \mathcal{A}_+ : \|\boldsymbol{\beta}_{\mathcal{G}_k}^+\| = 0 \text{ and } \min_{\mathbf{v} \in \partial_{\boldsymbol{\beta}_{\mathcal{G}_k}^+} L(\boldsymbol{\beta})} \|\mathbf{v}^-\| = 0, k = 1, \dots, K\}$$

$$\mathcal{A}_- \leftarrow \mathcal{A}_- \setminus \{j \in \mathcal{G}_k \subseteq \mathcal{A}_- : \|\boldsymbol{\beta}_{\mathcal{G}_k}^-\| = 0 \text{ and } \min_{\mathbf{v} \in \partial_{\boldsymbol{\beta}_{\mathcal{G}_k}^-} L(\boldsymbol{\beta})} \|\mathbf{v}^+\| = 0, k = 1, \dots, K\}$$

**Step b** Identify the greatest violation of optimality conditions:

$$g_+^k \leftarrow \min_{\mathbf{v} \in \partial_{\boldsymbol{\beta}_{\mathcal{G}_k}^+} L(\boldsymbol{\beta})} \|\mathbf{v}^+\|, \quad q \leftarrow \arg \max_k g_+^k$$

$$g_-^k \leftarrow \min_{\mathbf{v} \in \partial_{\boldsymbol{\beta}_{\mathcal{G}_k}^-} L(\boldsymbol{\beta})} \|\mathbf{v}^-\|, \quad r \leftarrow \arg \max_k g_-^k$$

**if**  $\max(g_+^q, g_-^r) = 0$  **then**

| Stop and return  $\boldsymbol{\beta}$ , which is optimal

**else**

| **if**  $g_+^q > g_-^r$  **then**  $\mathcal{A}_- \leftarrow \mathcal{A}_- \cup \mathcal{G}_q$  **else**  $\mathcal{A}_+ \leftarrow \mathcal{A}_+ \cup \mathcal{G}_r$

| Repeat Steps 1 and 2 until convergence

---

Finally, note that to compute a series of solutions along the regularization path for Problem (4) we simply choose a series of penalties  $\lambda^1 = \lambda_{\max} > \dots > \lambda^l > \dots > \lambda^L = \lambda_{\min} \geq 0$  such that  $\hat{\boldsymbol{\beta}}^{\text{coop}}(\lambda_{\max}) = \mathbf{0}$ , that is

$$\lambda_{\max} = \max_{k \in \{1, \dots, K\}} \max_{j \in \mathcal{G}_k} \frac{1}{w_k} \|\boldsymbol{\varphi}_j(\mathbf{X}^\top \mathbf{y})\|.$$

We then use the usual warm start strategy, where the feasible initial guess for  $\hat{\boldsymbol{\beta}}^{\text{coop}}(\lambda^l)$ , the coop-Lasso estimate with penalty parameter  $\lambda^l$ , is initialized with  $\hat{\boldsymbol{\beta}}^{\text{coop}}(\lambda^{l-1})$ .

### 2.3. Orthonormal design case

The orthonormal design case, where  $\mathbf{X}^\top \mathbf{X} = \mathbf{I}_p$ , has been providing useful insights for penalization techniques regarding the effects of shrinkage. Indeed, in this particular case, most usual shrinkage estimators can be expressed in closed-form as functions of the ordinary least squares (OLS) estimate. These expressions pave the way for the derivation of approximations of the degrees of freedom (Tibshirani, 1996; Yuan and Lin, 2006, and Section 4), which may be convenient for model selection in the absence of exact formulae.

In the orthonormal setting, for any  $\beta_j$ , we have  $\mathbf{x}_j^\top (\mathbf{X} \boldsymbol{\beta} - \mathbf{y}) = \beta_j - \hat{\beta}_j^{\text{ols}}$ . The optimality

conditions (8a) and (8b) can then be written as

$$\forall k \in \{1, \dots, K\}, \forall j \in \mathcal{G}_k, \hat{\beta}_j^{\text{coop}} = \left(1 - \frac{\lambda w_k}{\|\varphi_j(\hat{\beta}_{\mathcal{G}_k}^{\text{ols}})\|}\right)^+ \hat{\beta}_j^{\text{ols}}. \quad (9)$$

For reference, we recall the solution to the group-Lasso (Yuan and Lin, 2006) in the same condition

$$\forall k \in \{1, \dots, K\}, \forall j \in \mathcal{G}_k, \hat{\beta}_j^{\text{group}} = \left(1 - \frac{\lambda w_k}{\|\hat{\beta}_{\mathcal{G}_k}^{\text{ols}}\|}\right)^+ \hat{\beta}_j^{\text{ols}}, \quad (10)$$

while the Lasso solution (Tibshirani, 1996) is

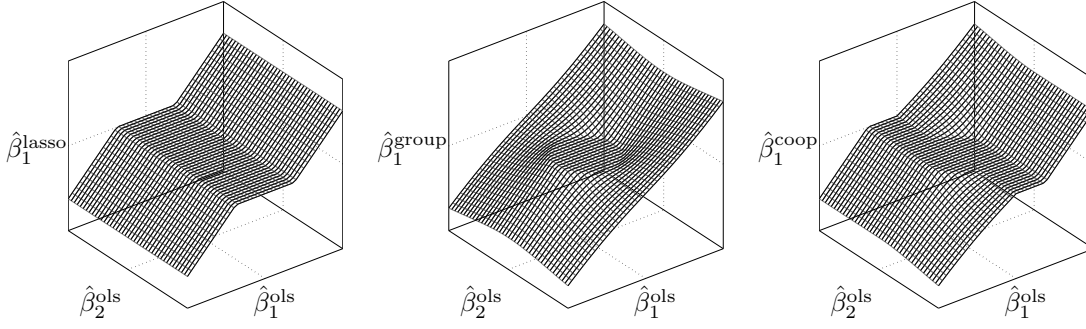
$$\forall j \in \{1, \dots, p\}, \hat{\beta}_j^{\text{lasso}} = \left(1 - \frac{\lambda w_k}{|\hat{\beta}_j^{\text{ols}}|}\right)^+ \hat{\beta}_j^{\text{ols}}. \quad (11)$$

Equations (9)–(11) reveal strong commonalities. First, the coefficients of these shrinkage estimators are of the sign of the OLS estimates. Second, the small coefficients belonging to a norm-specific neighborhood of zero are shrunk to zero and the amount of shrinkage is inversely proportional to this norm elsewhere. Finally, by grouping the terms corresponding to one group in Equations (9)–(10), a uniform translation effect, analogous to the one observed for the Lasso, comes into view:

$$\begin{aligned} \forall k \in \{1, \dots, K\}, \forall j \in \mathcal{G}_k, \|\varphi_j(\hat{\beta}_{\mathcal{G}_k}^{\text{coop}})\| &= \left(\|\varphi_j(\hat{\beta}_{\mathcal{G}_k}^{\text{ols}})\| - \lambda w_k\right)^+, \\ \forall k \in \{1, \dots, K\}, \|\hat{\beta}_{\mathcal{G}_k}^{\text{group}}\| &= \left(\|\hat{\beta}_{\mathcal{G}_k}^{\text{ols}}\| - \lambda w_k\right)^+, \\ \forall j \in \{1, \dots, p\}, |\hat{\beta}_j^{\text{lasso}}| &= \left(|\hat{\beta}_j^{\text{ols}}| - \lambda w_k\right)^+. \end{aligned} \quad (12)$$

The group-Lasso (10) differs primarily from the Lasso (11) owing to the common penalty  $\lambda w_k / \|\hat{\beta}_{\mathcal{G}_k}^{\text{ols}}\|$  for all the coefficients belonging to group  $k$ . The magnitude of shrinkage is determined by all within-group OLS coefficients, and thus radically different from a ridge regression penalty in this regard. For the coop-Lasso estimator (9), two penalties possibly apply to group  $k$ , for the positive and the negative OLS coefficients respectively. If all within-group OLS coefficients are of same sign, coop-Lasso is identical to group-Lasso; if some signs disagree, the magnitude of the penalty only depends on the within-group OLS coefficients with identical sign. In the extreme case where exactly one OLS coefficient is positive/negative, the coop-penalty is identical to a Lasso penalty on this coefficient.

Figure 2 provides a visual representation of Equations (9)–(11) for a group with two components, say  $\mathcal{G}_k = \{1, 2\}$ . We plot  $\hat{\beta}_1^{\text{lasso}}, \hat{\beta}_1^{\text{group}}$  and  $\hat{\beta}_1^{\text{coop}}$  as functions of  $(\hat{\beta}_1^{\text{ols}}, \hat{\beta}_2^{\text{ols}})$ . On the left hand side, the Lasso translates the  $\hat{\beta}_1^{\text{ols}}$  coefficient towards zero, eventually truncating them at zero, regardless of  $\hat{\beta}_2^{\text{ols}}$ : there is no interaction between coefficients. The group-Lasso, at the center, has a non-linear shrinking behavior (quite different from the Lasso or ridge penalties in this respect) and sets  $\hat{\beta}_1^{\text{group}}$  to zero within a Euclidean ball centered at zero. In the right hand side, the coop-Lasso appears as a cross-breed of Lasso and group-Lasso: it is identical to the group-Lasso in the positive and negative quadrants but identical to the Lasso when the signs of the OLS coefficients mismatch. For groups with more than two components, intermediate solutions



**Figure 2.** Lasso, group-Lasso and coop-Lasso coefficient estimate, for a group with 2 elements  $\mathcal{G}_k = \{1, 2\}$ , as a function of the OLS coefficients.

would be possible. This in-between status of the coop-Lasso, which allows for some flexibility with respect to the predefined group structure, will be confirmed in the following consistency analysis.

### 3. Consistency

Beyond its sanity-check value, a consistency analysis brings along an appreciation of the strengths and limitations of an estimation scheme. Here we concentrate on the estimation of the support of the parameter vector, that is, the position its zero entries. Our proof technique is drawn from the previous works on the Lasso (Yuan and Lin, 2007) and the group-Lasso (Bach, 2008).

In this type of analysis, some assumptions on the joint distribution of  $(X, Y)$  are required to guarantee the convergence of empirical covariances. For the sake of simplicity and coherence, we keep assuming that data are centered so that we have zero mean random variables and  $\Psi = \mathbb{E}[XX^\top]$  is the covariance matrix of  $X$ .

(A1)  $X$  and  $Y$  have finite fourth order moments  $\mathbb{E}[\|X\|^4] < \infty$ ,  $\mathbb{E}[Y^4] < \infty$ .

(A2) The covariance matrix  $\Psi = \mathbb{E}[XX^\top] \in \mathbb{R}^{p \times p}$  is invertible.

In addition to these standard technical assumptions, we need a more specific one, substantially avoiding situations where the coop-Lasso will almost never recover the true support.

(A3) All sign-incoherent groups are either included in or excluded from the true support:  $\forall k \in \{1, \dots, K\}$ , if  $\|(\beta_{\mathcal{G}_k}^*)^+\| > 0$  and  $\|(\beta_{\mathcal{G}_k}^*)^-\| > 0$ , then  $\forall j \in \mathcal{G}_k$ ,  $\beta_j^* \neq 0$ .

Note that this latter assumption is less stringent than the one required for the group-Lasso since it does not require that each group of variables should either be included in or excluded from the support. For the coop-Lasso, sign-coherent groups may intersect the support.

We now introduce suitable variants of the strong irrepresentable conditions, which provide a control on the spurious relationships that may arise from confounding variables and are thus necessary to guarantee support recovery. Here these conditions result in two assumptions, a general one, on the magnitude of correlations between relevant and irrelevant variables, and a more specific one for groups which intersect the support, on the sign of correlations. These conditions will be expressed in a compact vectorial form using the diagonal weighting matrix  $\mathbf{D}(\beta)$  such that,

$$\forall k \in \{1, \dots, K\}, \forall j \in \mathcal{S}_k(\beta), (\mathbf{D}(\beta))_{jj} = w_k \|\varphi_j(\beta_{\mathcal{G}_k})\|^{-1}. \quad (13)$$

(A4) For every group  $\mathcal{G}_k$  including at least one null coefficient (that is, such that  $\beta_j^* = 0$  for some  $j \in \mathcal{G}_k$  or equivalently  $\mathcal{S}_k^c \neq \emptyset$ ), there exists  $\eta > 0$  such that

$$\frac{1}{w_k} \max(\|(\Psi_{\mathcal{S}_k^c \mathcal{S}} \Psi_{\mathcal{S} \mathcal{S}}^{-1} \mathbf{D}(\beta_{\mathcal{S}}^*) \beta_{\mathcal{S}}^*)^+\|, \|(\Psi_{\mathcal{S}_k^c \mathcal{S}} \Psi_{\mathcal{S} \mathcal{S}}^{-1} \mathbf{D}(\beta_{\mathcal{S}}^*) \beta_{\mathcal{S}}^*)^-\|) \leq 1 - \eta, \quad (14)$$

where  $\Psi_{\mathcal{S}\mathcal{T}}$  is the submatrix of  $\Psi$  with lines and columns respectively indexed by  $\mathcal{S}$  and  $\mathcal{T}$ .

(A5) For every group  $\mathcal{G}_k$  intersecting the support and including either positive or negative coefficients, let  $\nu_k$  be the sign of these coefficients ( $\nu_k = 1$  if  $\|(\beta_{\mathcal{G}_k}^*)^+\| > 0$  and  $\nu_k = -1$  if  $\|(\beta_{\mathcal{G}_k}^*)^-\| > 0$ ), the following inequalities should hold:

$$\nu_k \Psi_{\mathcal{S}_k^c \mathcal{S}} \Psi_{\mathcal{S} \mathcal{S}}^{-1} \mathbf{D}(\beta_{\mathcal{S}}^*) \beta_{\mathcal{S}}^* \preceq \mathbf{0}, \quad (15)$$

where  $\preceq$  denotes componentwise inequality.

Note that the irrepresentable condition for the group-Lasso only considers correlations between groups included and excluded from the support. It is otherwise similar to (14), except that the elements of the weighting matrix  $\mathbf{D}$  are  $w_k \|\beta_{\mathcal{G}_k}\|^{-1}$  and that the  $\ell_2$  norm replaces  $\max(\|(\cdot)^+\|, \|(\cdot)^-\|)$ .

We now have all the components for stating the coop-Lasso consistency theorem, which will consider the following normalized (equivalent) form of the optimization problem (4) to allow a direct comparison with the known similar results previously stated for the Lasso and group-Lasso (Yuan and Lin, 2007; Bach, 2008):

$$\hat{\beta}^{\text{coop}} = \arg \min_{\beta \in \mathbb{R}^p} \frac{1}{2} \|\mathbf{y} - \mathbf{X}\beta\|_n^2 + \lambda_n \|\beta\|_{\text{coop}}, \quad (16)$$

where  $\|\cdot\|_n$  denotes the empirical norm and  $\lambda_n = \lambda/n$ .

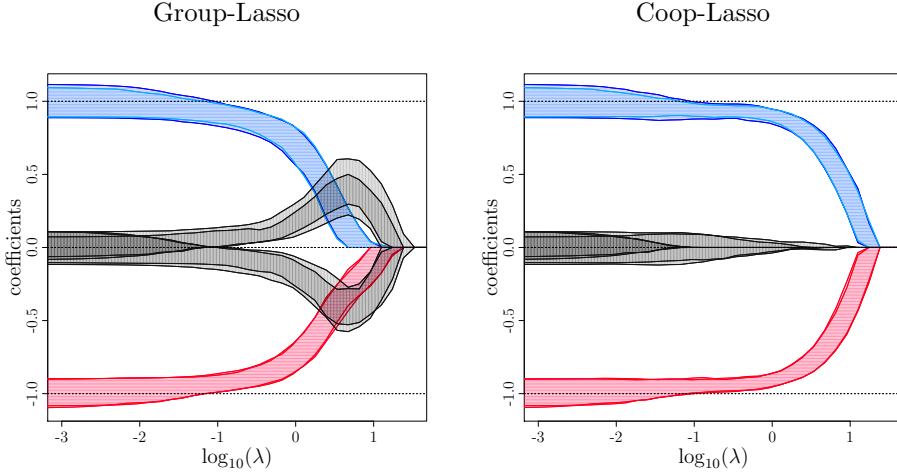
**THEOREM 2.** *If assumptions (A1-5) are satisfied, the coop-Lasso estimator is asymptotically unbiased and has the property of exact support recovery:*

$$\hat{\beta}^{\text{coop}} \xrightarrow{P} \beta^* \quad \text{and} \quad \mathbb{P}(\mathcal{S}(\hat{\beta}^{\text{coop}}) = \mathcal{S}) \rightarrow 1, \quad (17)$$

for every sequence  $\lambda_n$  such that  $\lambda_n = \lambda_0 n^{-\gamma}$ ,  $\gamma \in (0, 1/2)$ .

Compared to the group-Lasso, the consistency of support recovery for the coop-Lasso differs primarily regarding possible intersection (besides inclusion and exclusion) between groups and support. This additional flexibility applies to every sign-coherent group. Even if the support is the union of groups, when all groups are sign-coherent, the coop-Lasso has still an edge on group-Lasso since the irrepresentable condition (14) is weaker. Indeed, the norm in (14) is dominated by the  $\ell_2$  norm used for the group-Lasso. The next paragraph illustrates that this difference can have remarkable outcomes. Finally, when the support is the union of groups that are not sign-coherent, there is no systematic advantage in favor of one or the other method. While the norm used by the coop-Lasso is dominated by the norm used by the group-Lasso, the weighting matrix  $\mathbf{D}$  has smaller entries for the latter.

*Illustration.* We generate data from the regression model (1), with  $\beta^* = (1, 1, -1, -1, 0, 0, 0, 0)$ , equipped with the group structure  $\{\mathcal{G}_k\}_{k=1}^4 = \{\{1, 2\}, \{3, 4\}, \{5, 6\}, \{7, 8\}\}$ . The vector  $X$  is generated as a centered Gaussian random vector whose covariance matrix  $\Psi$  is chosen so that



**Figure 3.** 50% coverage intervals for the estimated coefficients along the regularization path for the group-Lasso (left) and coop-Lasso (right): horizontal stripes correspond to the relevant variables ( $\beta_j^* = \pm 1$ ) and vertical stripes to the irrelevant ones ( $\beta_j^* = 0$ ).

the irreprentable conditions hold for the coop-Lasso, but not for group-Lasso, which, we recall, are more demanding for the current situation, with sign-coherent groups. The random error  $\varepsilon$  follows a centered Gaussian distribution with standard deviation  $\sigma = 0.1$ , inducing to a very high signal to noise ratio ( $R^2 = 0.99$  on average), so that asymptotics provide a realistic view of the finite sample situation.

We generated 100 samples of size  $n = 20$  from the described model, computed the corresponding 100 regularization paths for the group-Lasso and the coop-Lasso. Figure 3 reports the 50% coverage intervals (lower and upper quartiles) along the regularization path for each method. Clearly, group-Lasso first selects the wrong covariates and never reaches the situation where it would have an obvious advantage upon the ordinary least squares estimator, while the coop-Lasso immediately selects the right covariates, whose coefficients steadily dominate the irrelevant ones.

#### 4. Model selection

Model selection amounts here to choose the penalization parameter  $\lambda$ , which restricts the size of the estimate  $\hat{\beta}(\lambda)$ . Trial values  $\{\lambda_{\min}, \dots, \lambda_{\max}\}$  define the set of models we have to choose from along the regularization path. The process aims at picking the model with minimum prediction error, or the one closest to the model from which data have been generated, assuming the model is correct, that is, Equation (1) holds. Here “closest” is typically measured by a distance between  $\hat{\beta}$  and  $\beta^*$ , either based on the value of the coefficients or on their support (true model selection), and sometimes also on the sign correctness of each non-zero entry.

Among the prerequisite for the selection process to be valid, the previous consistency analysis comes up with suitable orders of magnitude for the penalty parameter  $\lambda$ . However, this order of magnitude is not a proper value to be plugged in (4) and the practice is to use data driven approaches for selecting an appropriate penalty parameter.

Cross-validation is a recommended option (Hesterberg et al., 2008) when looking for the model minimizing the prediction error, but it is slow and not well suited to select the model closest to the true one. Analytical criteria provide a faster way to perform model selection and,

though the information criteria AIC and BIC rely on asymptotic derivations, they often offer good practical performances. The BIC and AIC criteria for the Lasso (Zou et al., 2007) and group-Lasso (Yuan and Lin, 2006) have been defined through the effective degrees of freedom:

$$\text{AIC}(\lambda) = \frac{\|\mathbf{y} - \hat{\mathbf{y}}(\lambda)\|^2}{\sigma^2} + 2\text{df}(\lambda) , \quad (18)$$

$$\text{BIC}(\lambda) = \frac{\|\mathbf{y} - \hat{\mathbf{y}}(\lambda)\|^2}{\sigma^2} + \log(n)\text{df}(\lambda) , \quad (19)$$

where  $\hat{\mathbf{y}}(\lambda) = \mathbf{X}\hat{\boldsymbol{\beta}}(\lambda)$  is the vector of predicted values for (4) with penalty parameter  $\lambda$ ,  $\sigma^2$  is the variance of the zero-mean Gaussian error variable  $\varepsilon$  in (1) and  $\text{df}(\lambda)$  is the number of degrees of freedom of the selected model. Assuming that Equation (1) holds and a differentiability condition on the mapping  $\hat{\mathbf{y}}(\lambda)$ , Efron (2004), using Stein's theory of unbiased risk estimate (Stein, 1981), shows that

$$\text{df}(\lambda) \doteq \frac{1}{\sigma^2} \sum_{i=1}^n \text{cov}(\hat{y}_i(\lambda), y_i) = \mathbb{E} \left[ \text{tr} \left( \frac{\partial \hat{\mathbf{y}}(\lambda)}{\partial \mathbf{y}} \right) \right] , \quad (20)$$

where the expectation is taken with respect to  $\mathbf{y}$  or equivalently to the noise  $\varepsilon$ . Yuan and Lin (2006) proposed an approximation of the trace term in the right-hand side of (20), which is used to estimate  $\text{df}(\lambda)$  for the group-Lasso:

$$\tilde{\text{df}}_{\text{group}}(\lambda) = \sum_{k=1}^K \mathbf{1} \left( \left\| \hat{\boldsymbol{\beta}}_{\mathcal{G}_k}^{\text{group}}(\lambda) \right\| > 0 \right) \left( 1 + \frac{\left\| \hat{\boldsymbol{\beta}}_{\mathcal{G}_k}^{\text{group}}(\lambda) \right\|}{\left\| \boldsymbol{\beta}_{\mathcal{G}_k}^{\text{ols}} \right\|} (p_k - 1) \right) , \quad (21)$$

where  $\mathbf{1}(\cdot)$  is the indicator function and  $p_k$  is the number of elements in  $\mathcal{G}_k$ . For orthonormal design matrices, (21) is an unbiased estimate of the true degrees of freedom of the group-Lasso and Yuan and Lin (2006) suggest that this approximation is relevant in more general settings, by reporting that “the performance of this approximate  $C_p$ -criterion [directly derived from (21)] is generally comparable with that of fivefold cross-validation and is sometimes better”.

This approximation of  $\text{df}(\lambda)$  relies on the OLS estimate and is hence limited to setups where the latter exists and is unique. In particular, the sample size should be larger than the number of predictors ( $n \geq p$ ). To overcome this restriction, we suggest a more general approximation to the degrees of freedom, based on the ridge estimator

$$\hat{\boldsymbol{\beta}}^{\text{ridge}}(\gamma) = (\mathbf{X}^T \mathbf{X} + \gamma \mathbf{I})^{-1} \mathbf{X}^T \mathbf{y} , \quad (22)$$

which can be computed even for small sample sizes ( $n < p$ ).

**PROPOSITION 1.** *Consider the coop-Lasso estimator  $\hat{\boldsymbol{\beta}}^{\text{coop}}(\lambda)$  defined by (4). Assuming that data are generated according to model (1), and that  $\mathbf{X}$  is orthonormal, the following expression of  $\tilde{\text{df}}_{\text{coop}}(\lambda)$  is an unbiased estimate of  $\text{df}(\lambda)$  defined in (20) for the coop-Lasso fit*

$$\begin{aligned} \tilde{\text{df}}_{\text{coop}}(\lambda) = & \sum_{k=1}^K \mathbf{1} \left( \left\| \left( \hat{\boldsymbol{\beta}}_{\mathcal{G}_k}^{\text{coop}}(\lambda) \right)^+ \right\| > 0 \right) \left( 1 + \frac{p_+^k - 1}{1 + \gamma} \frac{\left\| \left( \hat{\boldsymbol{\beta}}_{\mathcal{G}_k}^{\text{coop}}(\lambda) \right)^+ \right\|}{\left\| \left( \hat{\boldsymbol{\beta}}_{\mathcal{G}_k}^{\text{ridge}}(\gamma) \right)^+ \right\|} \right) \\ & + \mathbf{1} \left( \left\| \left( \hat{\boldsymbol{\beta}}_{\mathcal{G}_k}^{\text{coop}}(\lambda) \right)^- \right\| > 0 \right) \left( 1 + \frac{p_-^k - 1}{1 + \gamma} \frac{\left\| \left( \hat{\boldsymbol{\beta}}_{\mathcal{G}_k}^{\text{coop}}(\lambda) \right)^- \right\|}{\left\| \left( \hat{\boldsymbol{\beta}}_{\mathcal{G}_k}^{\text{ridge}}(\gamma) \right)^- \right\|} \right) , \quad (23) \end{aligned}$$

where  $p_+^k$  and  $p_-^k$  are respectively the number of positive and negative entries in  $\hat{\beta}_{\mathcal{G}_k}^{\text{ridge}}(\gamma)$ .

Proposition 1 raises a practical issue regarding the choice of a good reference  $\hat{\beta}^{\text{ridge}}(\gamma)$ . In our numerous simulations (most of which are not reported here), we did not observe a high sensitivity to  $\gamma$ , though high values degrade performances. When  $\mathbf{X}$  is full rank we use  $\gamma = 0$  (the OLS estimate) and correspondingly a vanishing  $\gamma$  (the Moore-Penrose solution) when  $\mathbf{X}$  is of smaller rank. More refined strategies are left for future works.

Section 5 illustrates that, even in non-orthonormal settings, plugging expression (23) for the degrees of freedom  $\text{df}(\lambda)$  of the coop-Lasso in BIC (19) or AIC (18) provides sensible model selection criteria. As expected, BIC, which is more stringent than AIC, is better at retrieving the sparsity pattern of  $\beta^*$ , while AIC is slightly better regarding prediction error.

## 5. Simulation study

We report here experimental results in the regression setup, with the linear regression model (1). Our simulation protocol is inspired from the one proposed by Breiman (1995, 1996) to test the nonnegative garrote estimator, which inspired the Lasso.

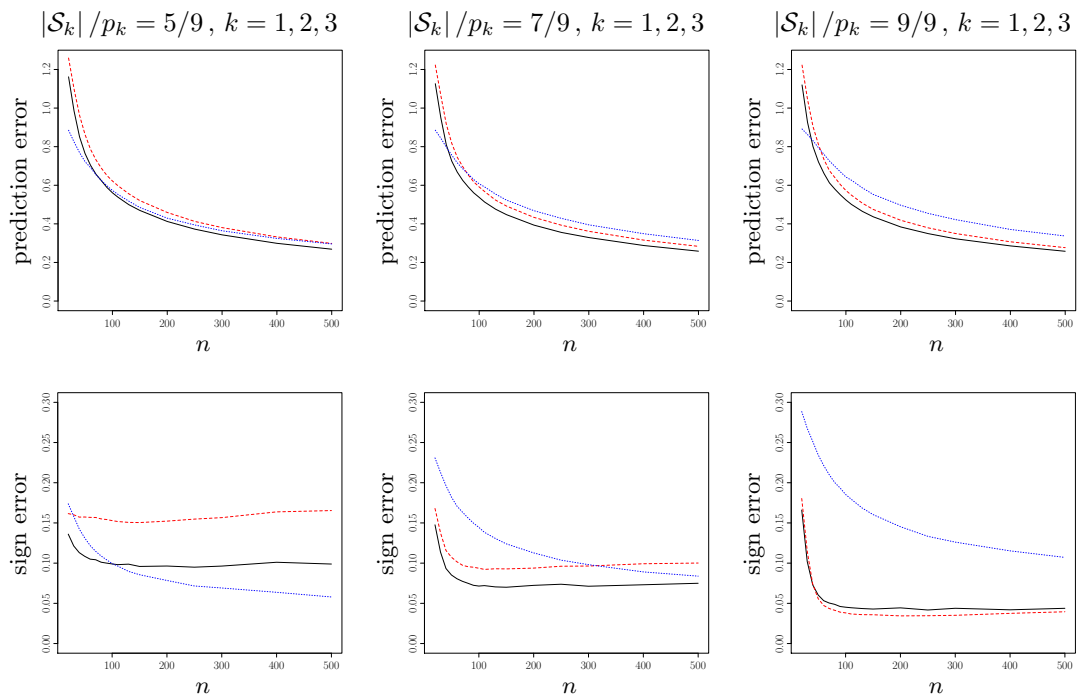
### 5.1. Data generation

The structure of  $\beta^* \in \mathbb{R}^p$  is controlled through sparsity at coefficient and group levels. Here we have  $p = 90$ , forming  $K = 10$  groups of identical size,  $p_k = 9$ . All groups of parameters follow the same wave pattern: for  $j \in \{1, \dots, 9\}$ ,  $(\beta_{\mathcal{G}_k}^*)_j \propto \nu_k \left( (h - |5 - j|)^+ \right)^2$ , where  $\nu_k \in \{0, 1\}$  is a switch at the group level and  $h \in \{1, 2, 3, 4, 5\}$  governs the wave width, that is, the within-group sparsity, with respectively  $|\mathcal{S}_k| \in \{1, 3, 5, 7, 9\}$  non-zero coefficients in each group included in the support. The covariates are drawn from a multivariate normal distribution  $X \sim \mathcal{N}(\mathbf{0}, \Psi)$  with covariances  $\Psi_{ij} = \rho^{|i-j|}$ , where  $\rho \in [-1, 1]$ . Finally, the response is corrupted by an error variable  $\varepsilon \sim \mathcal{N}(0, 1)$  and the magnitude of the vector of parameters  $\beta^*$  is chosen to have an  $R^2$  around 0.75.

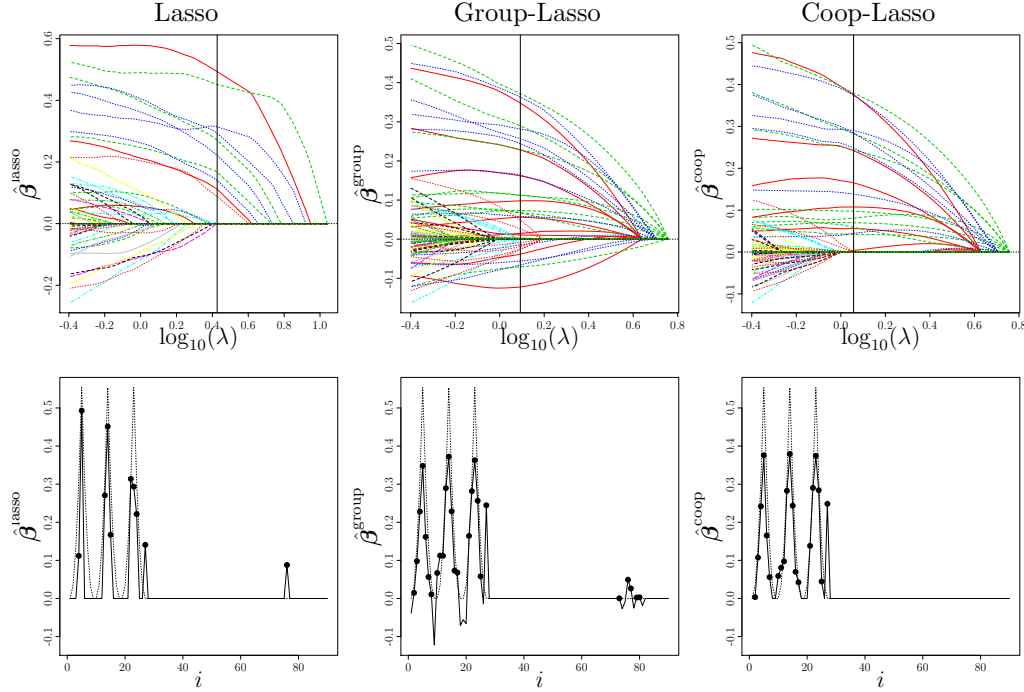
Note that the covariance of the covariates is purposely disconnected from the group structure. This setting may either be considered as unfair to the group methods, or equally adverse for all Lasso-type estimators, in the sense that none of their support recovery conditions are fulfilled when  $\rho \neq 0$ . Situations more or less advantageous for group methods are then produced thanks to the parameter  $h$ , which determines how the support of  $\beta^*$  matches the group structure.

### 5.2. Results

Figure 4 reports the average performances of the Lasso, group-Lasso and coop-Lasso, on 1000 independent draws, with model selection carried out by BIC for all estimators. The BIC for the Lasso is computed through (19), replacing the degrees of freedom by Zou et al.'s proposal (namely the number of non-zero entries in  $\hat{\beta}^{\text{lasso}}(\lambda)$ ). We display the mean squared prediction error and the support recovery (more precisely recovery of the sign of true parameters) when the true support comprises 3 groups out of 9 and a moderate correlation level  $\rho = 0.4$ . We did not observe a crucial role of these two parameters regarding the relative merits of the three methods. Regarding this point, the number of within-group non-zero coefficients (with  $h \in \{3, 4, 5\}$  corresponding to  $|\mathcal{S}_k| \in \{5, 7, 9\}$ ) and the number of training samples  $n$  play a more important role.



**Figure 4.** Mean prediction error (top) and mean sign error (bottom) for the Lasso (dotted blue), group-Lasso (dashed red) and coop-Lasso (solid black) according to sample size for three within-group sparsity patterns.



**Figure 5.** Regularization path (top) and parameter recovery (bottom) with BIC-based selection for the Lasso, group-Lasso, and coop-Lasso for a training set of size  $n = 120$  drawn from a model with  $|\mathcal{S}_k| = 7$  for  $k = 1, 2, 3$ . The true signal is the thin dotted line for each proposal.

All estimators perform about equally well in prediction accuracy. For small sample sizes, the Lasso has a slight but systematic advantage on the group estimators. A deeper analysis of these results (not shown) reveals that this difference is due to the model selection procedure, with BIC incurring less degradation compared to the best model for the Lasso. Regarding support recovery, the group methods perform systematically significantly better for small sample sizes. Then the Lasso catches up as the sample size grows, eventually gaining on group methods when groups only match loosely the support. The coop-Lasso ranks first or a close second in all experimental conditions with a rapid settlement towards its asymptotic behavior. It thus appears as the method of choice regarding inference issues when groups conform to the sign-coherence assumption. Note that, according to this performance measure, BIC performs very well, incurring a very small model selection loss (not shown) compared to the best one in all situations.

Finally, Figure 5 displays the Lasso, group-Lasso, and coop-Lasso regularization paths for a training set of size  $n = 120$  drawn from the model with  $|\mathcal{S}_k| = 7$  in the 3 first groups (still with  $\rho = 0.4$ ). The regularization paths display some interesting features, the behavior of the coop-Lasso being qualitatively intermediate between the Lasso and the group-Lasso. As expected, the non-zero coefficients appear one by one for the Lasso, groupwise for the group-lasso and the coop-Lasso, with the latter keeping some of these coefficients to zero until some irrelevant groups are activated. We observe a steady rise of the magnitude of the more important coefficients for all methods. For the small and zero coefficients belonging to the groups included in the support, the group lasso path is the wiggliest while the Lasso is the stablest. Most zero coefficients belonging to the groups excluded from the support are clearly identified by all methods, appearing together

at the end of the path at the smallest penalties. Those coefficients are correctly identified as zeros in a wider range for the coop-Lasso, whose BIC model (horizontal line) is the only one to exclude irrelevant groups. The solution selected by BIC is displayed below each path. The true signal is represented as a thin dotted line. The lasso estimate includes a non-zero coefficient from the irrelevant groups and is besides quite conservative, excluding many non-zero parameters from its support. One irrelevant group is improperly activated by the group-Lasso and none by the coop-Lasso; the group-Lasso is the only method which estimates negative coefficients, while the picture is clearer and more faithful for the coop-Lasso.

## 6. Robust microarray gene selection

Most studies on response to chemotherapy have considered breast cancer as a single homogeneous entity. However, it is a complex disease whose strong heterogeneity should not be overlooked. The dataset proposed by Hess et al. (2006) consists in gene expression profiling of patients treated with chemotherapy prior to surgery, classified as presenting either a pathologic complete response (pCR) or a residual disease (not-pCR). It records the signal of 22 269 probes<sup>†</sup> examining the human genome, each probe being related to a unique gene. Following Jeanmougin et al. (2011), we restrict our analysis to the basal tumors: for this particular subtype of breast cancer, clinical and pathologic features are homogeneous in the data set, whereas the response to chemotherapy is balanced, with 15 tumors being labeled pCR and 14 not-pCR. This setup is thus propitious to the statistical analysis of response to chemotherapy from the sole activity of genes.

### 6.1. Methodology

The microarray data have been collected to infer genes that predict the response to preoperative chemotherapy. The usual processing of this type of data is based on probe measurements that are related to genes in the final interpretation of the statistical analysis. Here we would like to take a different stance, by associating the gene entity to the statistical inference process.

As a matter of fact, we typically observe that some probes related to the very same gene have different behaviors. Requiring a consensus at the gene level supports biological coherence, thus exercising caution in an inference process where statistically plausible explanations are numerous, due to the noisy probe signals and to the cumbersome  $n \ll p$  setup (here  $n = 29$  and  $p = 22\,269$ ). Since the probes related to a given gene relate to sequences that are predominantly cooperating, the sign-coherence assumed by the coop-Lasso is particularly appropriate to improve robustness to the measurement noise and to encourage biologically plausible solutions.

Our protocol includes a preselection of probes that facilitates the analysis for the non-adaptive penalization methods compared here, and also provides an assessment of the benefits of adding seemingly less relevant probes into the statistical analysis. We proceed as follows:

- select a restricted number  $d$  of probes from classical differential analysis, where probes are sorted by decreasing p-values;
- determine the genes associated to these  $d$  probes, retrieve all the probes related to these genes, and select the corresponding  $p$  probes,  $p \geq d$ , regardless of their signal;
- fit a model with group penalties where groups are defined by genes.

<sup>†</sup>Actually, the data set reports the average signal in probe sets, which are a collection of probes designed to interrogate a given sequence. In this paper, the term “probe” designates Affymetrix probe sets to avoid confusion with the group structure that will be considered at a higher level.

Note that the binary response requires an appropriate model, such as logistic regression. The coop-Lasso fitting algorithm, which is easily adaptable to generalized linear models, follows exactly the structure provided in Algorithm 1, where the adequate likelihood function replaces the sum of square residuals in Step a.

## 6.2. Experimental setup

We select the first  $d = 10$  most differentiated probes, as identified by the analysis of Jeanmougin et al. (2011), on the 22 269 probes for the  $n = 29$  patients with basal tumor. These 10 probes correspond to 10 genes, themselves associated to  $p = 27$  probes on the microarray as a whole. All signals are normalized to have a unitary within-class variance, and 10 groups of probes are formed, with 1 to 6 elements per group (see Table 1). Then, we compare the following methods:

- logistic Lasso on the  $d = 10$  most differentiated probes (hereafter **probes**),
- logistic Lasso on the  $p = 27$  probes (with no group effect, hereafter **lasso**),
- logistic group-Lasso on the  $p = 27$  probes (with group effect, hereafter **group**),
- logistic coop-Lasso on the  $p = 27$  probes (with signed group effect, hereafter **coop**).

Logistic Lasso is implemented by the R-package `glmnet` (Friedman et al., 2010b), while logistic group-Lasso and coop-Lasso estimation is performed with our code (available at <http://stat.genopole.cnrs.fr/logiciels/scoop>).

Well-motivated analytical model selection criteria are not available today for Lasso-type penalties beyond the regression setup. Here model selection is carried out by 5-fold cross-validation based on the binomial deviance. Let  $\kappa : \{1, \dots, n\} \rightarrow \{1, \dots, K\}$  be a random partition of the  $n$  observations in  $K$  blocks of approximately same size, where  $\kappa(i) = k$  if observation  $i$  belongs to the  $k$ th block, we compute

$$\text{CV}(\lambda; \kappa) = \frac{1}{n} \sum_{i=1}^n D\left(y_i, \hat{y}_i^{-\kappa(i)}(\lambda)\right), \text{ where } D(y, \hat{y}) = 2 \left( (1-y) \log \frac{1-y}{1-\hat{y}} + y \log \frac{y}{\hat{y}} \right), \quad (24)$$

and where  $\hat{y}_i^{-\kappa(i)}(\lambda)$  is the estimated response for the  $i$ th observation for the model indexed by  $\lambda$  and adjusted by removing the observations belonging to block of observation  $i$ , that is with identical  $\kappa(i)$  value. The CV error (24) is computed for each method (**probes**, **lasso**, **group** and **coop**) using the same block partition  $\kappa$ . Finally, the CV score is averaged on  $M = 10$  distinct runs, differing by the mapping  $\kappa_m$ ,  $m = 1, \dots, M$ . This averaging produces the score  $\overline{\text{CV}}$ , whose variability due to the random choice of  $\kappa$  is reduced.

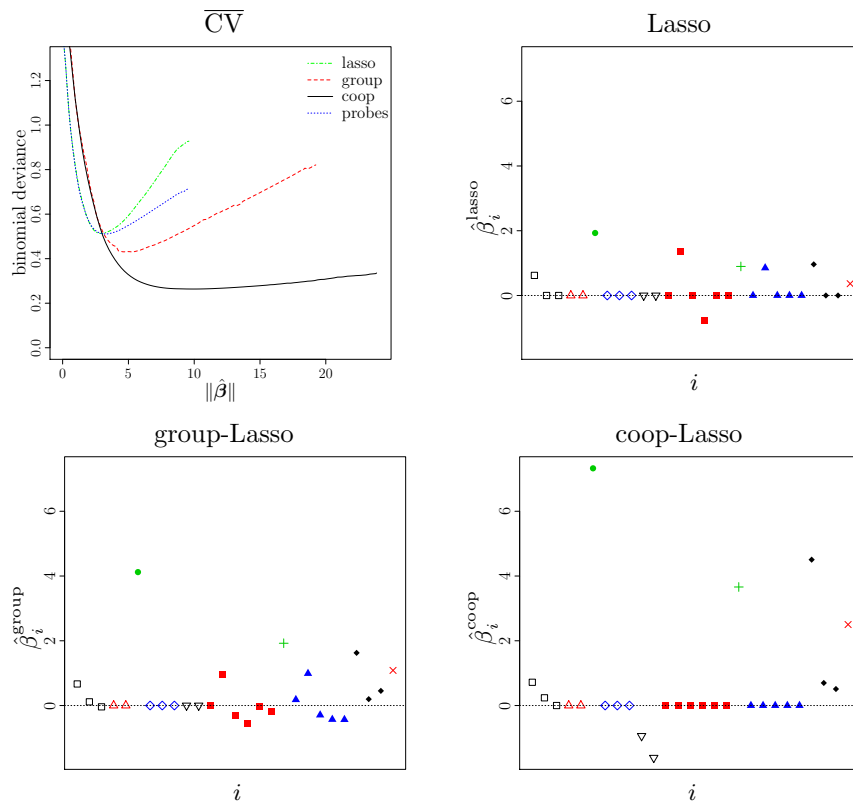
## 6.3. Results

The average cross-validation score is represented for each estimation method in Figure 6. The performances of the lasso on the original set of  $d$  probes (**probes**) and on the expanded set of  $p$  probes (**lasso**) are alike, with very similar parameter coefficients (not shown for **probes**). This suggests that our variable enrichment proposal is useless when the group structure is not taken into account. In contrast, the group-Lasso and to a greater extent the coop-Lasso, significantly improve the  $\overline{\text{CV}}$  score, indicating that some relevant information is extracted from the additional variables when supplied with the apposite group structure.

The bad performances of the two Lasso variants, and to a lesser extent of the group-Lasso, could be due to model selection instability problems. If  $\lambda^*$  were oversensitive to the training

**Table 1.** Genes corresponding to the probes selected by differential analysis, size of groups of probes, and  $\ell_2$ -norm of each group of parameters for each estimate.

$\mathcal{G}_k$ (gene)	$p_k$	symbol	probes	lasso	group	coop
FRMD4B	3	□	0.38	0.62	0.68	0.75
RNPS1	2	△	0	0	0	0
PHLDA3	1	●	1.82	1.93	4.12	7.32
TBC1D22A	3	◇	0	0	0	0
ECE1	2	▽	0.89	0	0	1.87
LZTS1	6	■	1.34	1.57	1.15	0
RPP38	1	+	0.95	0.90	1.92	3.66
GTSE1	5	▲	0.88	0.85	1.21	0
PAK4	3	◆	1.68	0.96	1.70	4.58
CHST10	1	×	0.79	0.36	1.08	2.50



**Figure 6.** Cross-validation deviance score (top left) and regression coefficients attached to each probe for lasso, group and coop.

**Table 2.** Best average CV score  $\overline{\text{CV}}(\lambda^*)$  and averaged best CV score  $\overline{\text{CV}}^*$ .

	probes	lasso	group	coop
$\overline{\text{CV}}(\lambda^*)$	0.511	0.513	0.430	0.263
$\overline{\text{CV}}^*$	0.474	0.499	0.372	0.194

sample, the averaging of CV scores could erase the minima. This possibility is ruled out by the results reported in Table 2, where we report the optimist  $\overline{\text{CV}}^*$  score, corresponding to the averaged best CV scores:

$$\overline{\text{CV}}^* = \frac{1}{M} \sum_{m=1}^M \min_{\lambda} \text{CV}(\lambda; \kappa_m) . \quad (25)$$

Table 2 supports that the performances of the coop-Lasso are not simply due to a stabler model selection, since the averaged minima for the Lasso and the group-Lasso are way above the minimum averaged score for the coop-Lasso.

For each procedure, the final parameter is estimated by adjustment to the whole data set using  $\lambda^* = \arg \min_{\lambda} \overline{\text{CV}}(\lambda)$ . The regression coefficients are represented in Figure 6 for **lasso**, **group** and **coop**: the coefficient attached to each probe is illustrated by a marker whose symbol stands for the corresponding gene (see designation in Table 1). The profile of the group-Lasso coefficients resembles the Lasso, with some additional scatter among the probes related to a given gene. The coop-Lasso is sparser at the gene level, with two groups (LZTS1 ■ and GTSE1 ▲) that were estimated to be sign-incoherent by the group-Lasso now switched off, and another group (ECE1 ▼) activated. According to this estimate, all groups are sign-coherent. The  $\overline{\text{CV}}$  score supports that our initial assumption is relevant for the problem at hand.

## 7. Discussion

The *coop-Lasso* is a variant of the *group-Lasso* that was originally proposed in the context of multi-task learning, more precisely for inferring related networks with Gaussian Graphical Models (Chiquet et al., 2011). Here we develop its analysis in the linear regression setup and demonstrate its value for prediction and inference with generalized linear models. Along this paper we provide an implementation of the fitting algorithm in the R package **scoop**, which makes this new penalized estimate publicly available for linear and logistic regression (the *coop-Lasso* for multiple network inference is also available in the R package **simone**).

The coop-Lasso differs from the group-Lasso by the assumption that the group structure is *sign-coherent*, namely that groups gather either non-positive, non-negative or null parameters. Though encouraging sign-consistent groups, the coop-Lasso penalty enables to recover various within-group sign patterns (positive, negative, null, non-positive, non-negative, non-null). This flexibility greatly reduces the incentive to drive within-group sparsity with an additional parameter (Friedman et al., 2010a) that later leads to an unwieldy model selection step.

Under suitable irrepresentable conditions, the proposed penalty leads to consistent model selection, even when the true sparsity pattern does not match the group structure. When the groups are sign-coherent the coop-Lasso compares favorably to the group-Lasso, recovering the true support under mildest assumptions.

We present an approximation of the effective degrees of freedom of the coop-Lasso which, once plugged into AIC or BIC, provides a fast way to select the tuning parameter in the linear regression setup. We provide empirical results demonstrating the capabilities of the coop-Lasso in terms of prediction and parameter selection, with BIC performing very well regarding support recovery even for small sample sizes. Finally, the application to genomic data opens a vast

potential field of great practical interest for this type of penalty, both in terms of prediction and interpretability. Our forthcoming investigations will aim at substantiating this ambition by conducting large scale experiments in this application domain.

## Acknowledgments

We would like to thank Marie Jeanmougin for her helpful comments on the breast cancer data set and for sharing her differential analysis on the subset of basal tumors. We also thank Catherine Matias for her careful reading of the manuscript and Christophe Ambroise for fruitful discussions.

Yves Grandvalet was partially supported by the PASCAL2 Network of Excellence, the European ICT FP7 under grant No 247022 - MASH, and the French National Research Agency (ANR) under grant ClasSel ANR-08-EMER-002.

## A. Proofs

### A.1. Proof of Lemma 1

Let us use  $\mathcal{T}_k$  as a shorthand for  $\mathcal{S}_k(\beta)$ , Chiquet et al. (2011) show that the subdifferential  $\theta$  obey the following conditions:

$$\max \left( \|\theta_{\mathcal{G}_k}^+\|, \|\theta_{\mathcal{G}_k}^-\| \right) \leq w_k \quad \text{if } \beta_{\mathcal{G}_k} = \mathbf{0} , \quad (26a)$$

$$\theta_{\mathcal{T}_k} = w_k \beta_{\mathcal{T}_k} \|\beta_{\mathcal{T}_k}\|^{-1}, \|\theta_{\mathcal{T}_k}^-\| \leq w_k, \|\theta_{\mathcal{T}_k}^+\| = 0 \quad \text{if } \|\beta_{\mathcal{G}_k}^+\| > 0, \|\beta_{\mathcal{G}_k}^-\| = 0 , \quad (26b)$$

$$\theta_{\mathcal{T}_k} = w_k \beta_{\mathcal{T}_k} \|\beta_{\mathcal{T}_k}\|^{-1}, \|\theta_{\mathcal{T}_k}^+\| \leq w_k, \|\theta_{\mathcal{T}_k}^-\| = 0 \quad \text{if } \|\beta_{\mathcal{G}_k}^-\| > 0, \|\beta_{\mathcal{G}_k}^+\| = 0 , \quad (26c)$$

$$\forall j \in \mathcal{G}_k, \theta_j = w_k \beta_j |\text{sign}(\beta_j)|^{-1}, \quad \text{if } \|\beta_{\mathcal{G}_k}^-\| > 0, \|\beta_{\mathcal{G}_k}^+\| > 0 . \quad (26d)$$

We thus simply have to prove the equivalence of conditions (7) and (26) for all  $\beta_{\mathcal{G}_k}$  values.

For  $\beta_{\mathcal{G}_k} = \mathbf{0}$ , (7) reads

$$\|\theta_{\mathcal{G}_k}^+\| \leq w_k \text{ and } \|\theta_{\mathcal{G}_k}^-\| \leq w_k , \quad (27)$$

which is equivalent to (26a).

For  $\beta_{\mathcal{G}_k} \neq \mathbf{0}$ , the equalities for  $\theta_{\mathcal{T}_k}$  in (26b)–(26d) are equivalent to (7a), thus setting the equivalence between (7) and (26) for all non-zero coefficients. For  $\beta_{\mathcal{T}_k}^c$ , let us consider the case (26b), where all non-zero parameters within group  $k$  are positive. The first equation of (26b) implies that  $\|\theta_{\mathcal{T}_k}^+\| = w_k$  and  $\|\theta_{\mathcal{T}_k}^-\| = 0$ . Hence,  $\|\theta_{\mathcal{T}_k}^-\| \leq w_k$  and  $\|\theta_{\mathcal{T}_k}^+\| = 0$  imply (27), so that (26b) implies (7). The contraposition is also easy to check. From (7a), when all coefficients are positive, we have that  $\|\theta_{\mathcal{T}_k}^-\| = 0$  and  $\|\theta_{\mathcal{T}_k}^+\| = w_k$ . Then, this implies that (7b) reads

$$\|\theta_{\mathcal{T}_k}^-\| \leq w_k \text{ and } \|\theta_{\mathcal{T}_k}^+\| = 0 ,$$

which defines  $\theta_{\mathcal{T}_k}^c$  in (26b). The proof is similar for (26c) where all non-zero parameters within group  $k$  are positive.

### A.2. Proof of Proposition 1

We assume here that  $\mathbf{X}^\top \mathbf{X} = \mathbf{I}_p$ . We introduce the ridge estimator in the computation of the trace in Equation (20), through the chain rule, yielding an unbiased estimate of df:

$$\tilde{\text{df}}_{\text{coop}}(\lambda) = \text{tr} \left( \frac{\partial \hat{\mathbf{y}}(\lambda)}{\partial \mathbf{y}} \right) = \text{tr} \left( \frac{\partial \mathbf{X}^\top \hat{\boldsymbol{\beta}}^{\text{coop}}(\lambda)}{\partial \hat{\boldsymbol{\beta}}^{\text{ridge}}(\gamma)} \frac{\partial \hat{\boldsymbol{\beta}}^{\text{ridge}}(\gamma)}{\partial \mathbf{y}} \right) = \frac{1}{1+\gamma} \sum_{k=1}^K \sum_{j \in \mathcal{G}_k} \frac{\partial \hat{\beta}_j^{\text{coop}}(\lambda)}{\partial \hat{\beta}_j^{\text{ridge}}(\gamma)},$$

where the last equation derives from the definition (22) of the ridge estimator with regularization parameter  $\gamma$ . Then, the expression of the coop-Lasso as a function of the ridge regression estimate is simply obtained from Equation (9), using that, in the orthonormal case, we have  $\hat{\boldsymbol{\beta}}^{\text{ols}} = (1+\gamma)\hat{\boldsymbol{\beta}}^{\text{ridge}}(\gamma)$ . Dropping the reference to  $\lambda$  and  $\gamma$  that are obvious from the context, we have

$$\forall k \in \{1, \dots, K\}, \forall j \in \mathcal{G}_k, \hat{\beta}_j^{\text{coop}} = \left( 1 - \frac{\lambda w_k}{(1+\gamma) \|\boldsymbol{\varphi}_j(\hat{\boldsymbol{\beta}}_{\mathcal{G}_k}^{\text{ridge}})\|} \right)^+ (1+\gamma) \hat{\beta}_j^{\text{ridge}}. \quad (28)$$

Then, for  $j \in \mathcal{G}_k$ , routine differentiation gives

$$\frac{1}{1+\gamma} \frac{\partial \hat{\beta}_j^{\text{coop}}}{\partial \hat{\beta}_j^{\text{ridge}}} = \mathbf{1} \left( \|\hat{\beta}_j^{\text{coop}}\| > 0 \right) \left( 1 - \frac{\lambda w_k}{(1+\gamma)} \left( \frac{1}{\|\boldsymbol{\varphi}_j(\hat{\boldsymbol{\beta}}_{\mathcal{G}_k}^{\text{ridge}})\|} - \frac{(\hat{\beta}_j^{\text{ridge}})^2}{\|\boldsymbol{\varphi}_j(\hat{\boldsymbol{\beta}}_{\mathcal{G}_k}^{\text{ridge}})\|^3} \right) \right).$$

The summation over the positive and negative elements of  $\mathcal{G}_k$  reduces to two terms

$$\begin{aligned} \frac{1}{1+\gamma} \sum_{j \in \mathcal{G}_k} \frac{\partial \hat{\beta}_j^{\text{coop}}}{\partial \hat{\beta}_j^{\text{ridge}}} &= \mathbf{1} \left( \|(\hat{\boldsymbol{\beta}}_{\mathcal{G}_k}^{\text{coop}})^+\| > 0 \right) \left( p_+^k - \frac{\lambda w_k}{1+\gamma} \frac{(p_+^k - 1)}{\|(\hat{\boldsymbol{\beta}}_{\mathcal{G}_k}^{\text{ridge}})^+\|} \right) \\ &\quad + \mathbf{1} \left( \|(\hat{\boldsymbol{\beta}}_{\mathcal{G}_k}^{\text{coop}})^-\| > 0 \right) \left( p_-^k - \frac{\lambda w_k}{1+\gamma} \frac{(p_-^k - 1)}{\|(\hat{\boldsymbol{\beta}}_{\mathcal{G}_k}^{\text{ridge}})^-\|} \right) \\ &= \mathbf{1} \left( \|(\hat{\boldsymbol{\beta}}_{\mathcal{G}_k}^{\text{coop}})^+\| > 0 \right) + \left( 1 - \frac{\lambda w_k}{(1+\gamma) \|(\hat{\boldsymbol{\beta}}_{\mathcal{G}_k}^{\text{ridge}})^+\|} \right)^+ (p_+^k - 1) \\ &\quad + \mathbf{1} \left( \|(\hat{\boldsymbol{\beta}}_{\mathcal{G}_k}^{\text{coop}})^-\| > 0 \right) + \left( 1 - \frac{\lambda w_k}{(1+\gamma) \|(\hat{\boldsymbol{\beta}}_{\mathcal{G}_k}^{\text{ridge}})^-\|} \right)^+ (p_-^k - 1). \end{aligned}$$

From (28), we have

$$\forall k \in \{1, \dots, K\}, \forall j \in \mathcal{G}_k, \left( 1 - \frac{\lambda w_k}{(1+\gamma) \|\boldsymbol{\varphi}_j(\hat{\boldsymbol{\beta}}_{\mathcal{G}_k}^{\text{ridge}})\|} \right)^+ = \frac{1}{1+\gamma} \frac{\|\boldsymbol{\varphi}_j(\hat{\boldsymbol{\beta}}_{\mathcal{G}_k}^{\text{coop}})\|}{\|\boldsymbol{\varphi}_j(\hat{\boldsymbol{\beta}}_{\mathcal{G}_k}^{\text{ridge}})\|},$$

which is used twice to simplify the previous expression. Summing over all groups concludes the proof.

**A.3. Proof of Theorem 2**

Our asymptotic results are established on the scaled Problem (16). We then follow the three steps proof technique proposed by Yuan and Lin (2007) for the Lasso and also applied by Bach (2008) for the group-Lasso:

- (a) restrict the estimation problem to the true support;
- (b) complete this estimate by 0 outside the true support;
- (c) prove that this artificial estimate satisfies optimality conditions for the original coop-Lasso problem with probability tending to 1.

Then, under (A2), the solution is unique, leading to the conclusion that the coop-Lasso estimator is equal to this artificial estimate with probability tending to 1, which ends the proof. Note however a slight yet important difference along the discussion: since we authorize divergences between the group structure  $\{\mathcal{G}_k\}_{k=1}^K$  and the true support  $\mathcal{S}$ , the irrerepresentable conditions (A4-5) for the coop-Lasso cannot be expressed simply in terms of coop-norms (as it is done with the group-norm in Bach, 2008). We will see that this does not impede the development of the proof.

As a first step, we prove two simple lemmas. Lemma 2 states that the coop-Lasso estimate, restricted on the true support  $\mathcal{S}$ , is consistent when  $\lambda_n \rightarrow 0$ . Lemma 3 provides the basis for the inequalities (14) and (15) that express our irrerepresentable conditions.

LEMMA 2. *Assuming (A1-3), let  $\tilde{\beta}_{\mathcal{S}}$  be the unique minimizer of the regression problem restricted to the true support  $\mathcal{S}$ :*

$$\tilde{\beta}_{\mathcal{S}} = \arg \min_{\mathbf{v} \in \mathbb{R}^{|\mathcal{S}|}} \frac{1}{2} \|\mathbf{y} - \mathbf{X}_{\cdot \mathcal{S}} \mathbf{v}\|_n^2 + \lambda_n \sum_{k: \mathcal{S}_k \neq \emptyset} w_k (\|\mathbf{v}_{\mathcal{S}_k}^+\| + \|\mathbf{v}_{\mathcal{S}_k}^-\|) ,$$

If  $\lambda_n \rightarrow 0$ , then  $\tilde{\beta}_{\mathcal{S}} \xrightarrow{P} \beta_{\mathcal{S}}^*$ .

PROOF. This lemma stems from standard results of M-estimation (Van der Vaart, 1998). Let  $\boldsymbol{\varepsilon} = \mathbf{y} - \mathbf{X} \boldsymbol{\beta}^*$ , and write  $\boldsymbol{\Psi}^n = \mathbf{X}^\top \mathbf{X} / n$ . If  $\lambda_n \rightarrow 0$ , then under (A1-2), for any  $\mathbf{v} \in \mathbb{R}^{|\mathcal{S}|}$

$$\begin{aligned} Z_n(\mathbf{v}) &= \frac{1}{2} \|\mathbf{y} - \mathbf{X}_{\cdot \mathcal{S}} \mathbf{v}\|_n^2 + \lambda_n \sum_{k: \mathcal{S}_k \neq \emptyset} w_k (\|\mathbf{v}_{\mathcal{S}_k}^+\| + \|\mathbf{v}_{\mathcal{S}_k}^-\|) \\ &= \frac{1}{2} (\boldsymbol{\beta}_{\mathcal{S}}^* - \mathbf{v})^\top \boldsymbol{\Psi}_{\mathcal{S}\mathcal{S}}^n (\boldsymbol{\beta}_{\mathcal{S}}^* - \mathbf{v}) - \frac{1}{n} \boldsymbol{\varepsilon}^\top \mathbf{X}_{\cdot \mathcal{S}} (\boldsymbol{\beta}_{\mathcal{S}}^* - \mathbf{v}) + \frac{\boldsymbol{\varepsilon}^\top \boldsymbol{\varepsilon}}{2n} + \lambda_n \sum_{k: \mathcal{S}_k \neq \emptyset} w_k (\|\mathbf{v}_{\mathcal{S}_k}^+\| + \|\mathbf{v}_{\mathcal{S}_k}^-\|) \end{aligned}$$

tends in probability to

$$Z(\mathbf{v}) = \frac{1}{2} (\boldsymbol{\beta}_{\mathcal{S}}^* - \mathbf{v})^\top \boldsymbol{\Psi}_{\mathcal{S}\mathcal{S}} (\boldsymbol{\beta}_{\mathcal{S}}^* - \mathbf{v}) + \frac{1}{2} \sigma^2 .$$

It follows from the strict convexity of  $Z_n$  that  $\arg \min Z_n(\mathbf{v}) \xrightarrow{P} \arg \min Z(\mathbf{v}) = \boldsymbol{\beta}_{\mathcal{S}}^*$  (Knight and Fu, 2000), which ends the proof. □

LEMMA 3. *Consider a sequence of random variables  $S_n$  such that  $S_n \xrightarrow{P} S$ . Suppose there exists  $\delta > 0$  such that for a given norm  $\mu$  the limit  $S$  is bounded away from 1:*

$$\mu(S) \leq 1 - \delta .$$

Then,

$$\mathbb{P}(\mu(S_n) \leq 1) \rightarrow 1 .$$

PROOF. By triangular inequality and thanks to the constraint on  $\mu(S)$ :

$$\mathbb{P}(\mu(S_n) \leq 1) \geq \mathbb{P}(\mu(S_n - S) \leq 1 - \mu(S)) \geq \mathbb{P}(\mu(S_n - S) \leq \delta) ,$$

Convergence in probability of  $S_n$  to  $S$  concludes the proof:

$$\mathbb{P}(\mu(S_n - S) \leq \delta) \rightarrow 1 , \quad \text{therefore} \quad \mathbb{P}(\mu(S_n) \leq 1) \rightarrow 1 .$$

□

Let us consider the full vector  $\tilde{\beta}$  with coefficients  $\tilde{\beta}_S$  defined as in Lemma 2 and other coefficients null,  $\tilde{\beta}_{S^c} = \mathbf{0}$ . We now proceed to the last step of the proof of Theorem 2, by proving that  $\tilde{\beta}$  satisfies the coop-Lasso optimality conditions with probability tending to 1 under the additional conditions (A4-5). The final conclusion then results from the uniqueness of the coop-Lasso estimator.

First, consider optimality conditions with respect to  $\beta_S$ . As a result of Lemma 2, the probability that  $\tilde{\beta}_j \neq 0$  for every  $j \in S$  tends to 1. Thereby,  $\tilde{\beta}_S$  satisfies (8a) on the restriction of  $\mathbf{X}$  to covariates in  $S$  with probability tending to 1. As  $\tilde{\beta}_{S^c} = \mathbf{0}$ , then  $\mathbf{X}\tilde{\beta} = \mathbf{X}_{\cdot S}\tilde{\beta}_S$  and for every  $j \in S$ ,  $\|\varphi_j(\tilde{\beta}_{S_k})\| = \|\varphi_j(\tilde{\beta}_{G_k})\|$ , therefore  $\tilde{\beta}_S$  satisfies (8a) in the original problem with probability tending to 1.

Second,  $\tilde{\beta}_{S^c}$  should also verify the optimality conditions (8) with probability tending to 1. With assumption (A3), we only have to consider two cases that read:

- if group  $k$  is excluded from the support, one must have

$$\mathbb{P}\left(\frac{1}{\lambda_n w_k} \max\left(\|((\mathbf{X}_{\cdot S_k^c})^\top(\mathbf{X}\tilde{\beta} - \mathbf{y}))^+\|_n, \|((\mathbf{X}_{\cdot S_k^c})^\top(\mathbf{X}\tilde{\beta} - \mathbf{y}))^-\|_n\right) \leq 1\right) \rightarrow 1 ; \quad (29)$$

- if group  $k$  intersects the support, with either positive ( $\nu_k = 1$ ) or negative ( $\nu_k = -1$ ) coefficients, one must have

$$\mathbb{P}\left(\{\nu_k(\mathbf{X}_{\cdot S_k^c})^\top(\mathbf{X}\tilde{\beta} - \mathbf{y}) \succeq \mathbf{0}\} \cap \left\{\frac{1}{\lambda_n w_k} \|(\mathbf{X}_{\cdot S_k^c})^\top(\mathbf{X}\tilde{\beta} - \mathbf{y})\|_n \leq 1\right\}\right) \rightarrow 1 . \quad (30)$$

To prove (29) and (30), we study the asymptotics of  $(\mathbf{X}_{\cdot S_k^c})^\top(\mathbf{X}\tilde{\beta} - \mathbf{y})/n$  for any group such that  $S_k^c$  is not empty. As a consequence of the existence of the fourth order moments of the centered random variables  $X$  and  $Y$ , the multivariate central limit theorem applies, yielding:

$$\frac{\mathbf{X}^\top \mathbf{X}}{n} = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i^\top \mathbf{x}_i = \Psi + O_P(n^{-1/2}), \quad \frac{\mathbf{X}^\top \boldsymbol{\varepsilon}}{n} = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i \boldsymbol{\varepsilon}_i = O_P(n^{-1/2}) \quad (31)$$

Then, we derive from (31) and the definition of  $\tilde{\beta}$  that

$$\begin{aligned} \frac{1}{n}(\mathbf{X}_{\cdot S_k^c})^\top(\mathbf{X}\tilde{\beta} - \mathbf{y}) &= \frac{1}{n}(\mathbf{X}_{\cdot S_k^c})^\top \mathbf{X}(\tilde{\beta} - \beta^*) - \frac{1}{n}(\mathbf{X}_{\cdot S_k^c})^\top \boldsymbol{\varepsilon} \\ &= \frac{1}{n}(\mathbf{X}_{\cdot S_k^c})^\top \mathbf{X}_{\cdot S}(\tilde{\beta}_S - \beta_S^*) + O_P(n^{-1/2}) \\ &= \Psi_{S_k^c S}(\tilde{\beta}_S - \beta_S^*) + O_P(n^{-1/2}) . \end{aligned} \quad (32)$$

While the combination of (31) and optimality conditions (8a) on  $\tilde{\beta}_S$  leads to:

$$\Psi_{SS}(\tilde{\beta}_S - \beta_S^*) = -\lambda_n \mathbf{D}(\tilde{\beta}_S) \tilde{\beta}_S + O_P(n^{-1/2}) , \quad (33)$$

where  $\mathbf{D}(\cdot)$  is the weighting matrix (13). Put (32) and (33) together to finally obtain:

$$\frac{1}{n}(\mathbf{X}_{\cdot, \mathcal{S}_k^c})^\top (\mathbf{X}\tilde{\boldsymbol{\beta}} - \mathbf{y}) = -\lambda_n \boldsymbol{\Psi}_{\mathcal{S}_k^c} \boldsymbol{\Psi}_{\mathcal{S}}^{-1} \mathbf{D}(\tilde{\boldsymbol{\beta}}_{\mathcal{S}}) \tilde{\boldsymbol{\beta}}_{\mathcal{S}} + O_P(n^{-1/2}) . \quad (34)$$

Now, define for any  $k$  such that  $\mathcal{S}_k^c$  is not empty:

$$R_{k,n} = \frac{1}{w_k \lambda_n} \frac{1}{n} (\mathbf{X}_{\cdot, \mathcal{S}_k^c})^\top (\mathbf{X}\tilde{\boldsymbol{\beta}} - \mathbf{y}) \quad \text{and} \quad R_k = -\frac{1}{w_k} \boldsymbol{\Psi}_{\mathcal{S}_k^c} \boldsymbol{\Psi}_{\mathcal{S}}^{-1} \mathbf{D}(\boldsymbol{\beta}_{\mathcal{S}}^*) \boldsymbol{\beta}_{\mathcal{S}}^* ,$$

Limits (29) and (30) are expressed:

- if group  $k$  is excluded from the support, one must have

$$\mathbb{P} \left( \max \left( \|R_{k,n}^+\|, \|R_{k,n}^-\| \right) \leq 1 \right) \rightarrow 1 ;$$

- if group  $k$  intersects the support, with either positive ( $\nu_k = 1$ ) or negative ( $\nu_k = -1$ ) coefficients, one must have

$$\mathbb{P} \left( \{ \nu_k R_{k,n} \geq \mathbf{0} \} \cap \{ \|(\nu_k R_{k,n})^+\| \leq 1 \} \right) \rightarrow 1 .$$

Remark that, as a continuous function of  $\tilde{\boldsymbol{\beta}}_{\mathcal{S}}$ ,  $\mathbf{D}(\tilde{\boldsymbol{\beta}}_{\mathcal{S}}) \tilde{\boldsymbol{\beta}}_{\mathcal{S}}$  converges in probability to  $\mathbf{D}(\boldsymbol{\beta}_{\mathcal{S}}^*) \boldsymbol{\beta}_{\mathcal{S}}^*$ . Therefore, with a decrease rate for  $\lambda_n$  chosen such that  $n^{1/2} \lambda_n \rightarrow \infty$ , equation (34) implies

$$R_{k,n} \xrightarrow{P} R_k . \quad (35)$$

It now suffices to successively apply Lemma 3 to the appropriate vectors and norms to show that  $\tilde{\boldsymbol{\beta}}_{\mathcal{S}^c}$  satisfies (29) and (30):

- if group  $k$  is excluded from the support, (A4) assumes that there exists  $\eta > 0$ , such that

$$\max(\|R_k^+\|, \|R_k^-\|) \leq 1 - \eta ,$$

and Lemma 3 applied to  $\mu(u) = \max(\|u^+\|, \|u^-\|)$  provides

$$\mathbb{P}\{\max(\|R_{k,n}^+\|, \|R_{k,n}^-\|) \leq 1\} \rightarrow 1 .$$

- if group  $k$  intersects the support, with either positive ( $\nu_k = 1$ ) or negative ( $\nu_k = -1$ ) coefficients,

$$\begin{aligned} \mathbb{P} \left( \{ \|(\nu_k R_{k,n})^+\| \leq 1 \} \cap \{ \nu_k R_{k,n} \geq \mathbf{0} \} \right) &= 1 - \mathbb{P} \left( \{ \|(\nu_k R_{k,n})^+\| > 1 \} \cup \{ \nu_k R_{k,n} < \mathbf{0} \} \right) \\ &\geq 1 - \mathbb{P} \left( \|(\nu_k R_{k,n})^+\| > 1 \right) - \mathbb{P} \left( \nu_k R_{k,n} < \mathbf{0} \right) \\ &\geq 1 - \mathbb{P} \left( \max(\|R_{k,n}^+\|, \|R_{k,n}^-\|) > 1 \right) - \mathbb{P} \left( \nu_k R_{k,n} < \mathbf{0} \right) . \end{aligned}$$

As previously, the first probability in the sum tends to 0 because of (A4) and Lemma 3. The second probability tends to 0 from (A5) and of the convergence in probability of  $R_{k,n}$  to  $R_k$ . Therefore the overall probability tends to 1.

Denote by  $A_{k,n}$  these events on which coefficients in  $\mathcal{S}_k^c$  are set to 0. We just showed that individually for each group  $k$  with true null coefficients,  $P(A_{k,n}) \rightarrow 1$ . This implies that,

$$\mathbb{P} \left( \bigcup_{k: \mathcal{S}_k^c \neq \emptyset} A_{k,n}^c \right) \leq \sum_{k: \mathcal{S}_k^c \neq \emptyset} \mathbb{P} (A_{k,n}^c) \rightarrow 0,$$

which in turn concludes the proof:

$$\mathbb{P} \left( \bigcap_{k: S_k^c \neq \emptyset} A_{k,n} \right) \rightarrow 1.$$

□

## References

- Bach, F. (2008). Consistency of the group lasso and multiple kernel learning. *Journal of Machine Learning Research* 9, 1179–1225.
- Bakin, S. (1999). *Adaptive regression and model selection in data mining problems*. Ph. D. thesis, Australian National University, Canberra.
- Breiman, L. (1995). Better subset regression using the nonnegative garrote. *Technometrics* 37(4), 373–384.
- Breiman, L. (1996). Heuristics of instability and stabilization in model selection. *The Annals of Statistics* 24(6), 2350–2383.
- Chiquet, J., Y. Grandvalet, and C. Ambroise (2011). Inferring multiple graphical structures. *Statistic and Computing*.
- Efron, B. (2004). The estimation of prediction error: Covariance penalties and cross-validation (with discussion). *Journal of the American Statistical Association* 99, 619–642.
- Foygel, R. and M. Drton (2010). Exact block-wise optimization in group lasso for linear regression. *arxiv preprint*.
- Friedman, J., T. Hastie, and R. Tibshirani (2010a). A note on the group lasso and a sparse group lasso. *arxiv preprint*.
- Friedman, J., T. Hastie, and R. Tibshirani (2010b, 2). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software* 33(1), 1–22.
- Grandvalet, Y. and S. Canu (1999). Outcomes of the equivalence of adaptive ridge with least absolute shrinkage. In *Advances in Neural Information Processing Systems 11 (NIPS 1998)*, pp. 445–451.
- Hess, K., K. Anderson, W. Symmans, V. Valero, N. Ibrahim, J. Mejia, D. Booser, R. Theriault, U. Buzdar, P. Dempsey, R. Rouzier, N. Sneige, J. Ross, T. Vidaurre, H. Gómez, G. Hortobagyi, and L. Pustzai (2006). Pharmacogenomic predictor of sensitivity to preoperative chemotherapy with paclitaxel and fluorouracil, doxorubicin, and cyclophosphamide in breast cancer. *Journal of Clinical Oncology* 24(26), 4236–4244.
- Hesterberg, T., N. M. Choi, L. Meier, and C. Fraley (2008). Least angle and  $\ell_1$  penalized regression: A review. *Statistics Surveys* 2, 61–93.
- Huang, J. and T. Zhang (2010). The benefit of group sparsity. *arxiv preprint*.
- Jeanmougin, M., M. Guedj, and C. Ambroise (2011). Defining a robust biological prior from pathway analysis to drive network inference. *Journal de la Société Française de Statistique*.

- Knight, K. and W. Fu (2000). Asymptotics for lasso-type estimators. *Annals of Statistics* 28(5), 1356–1378.
- Meier, L., S. Van De Geer, and P. Bühlmann (2008). The group lasso for logistic regression. *Journal of the Royal Statistical Society, Series B* 70, 53–71.
- Nardi, Y. and A. Rinaldo (2008). On the asymptotic properties of the group lasso estimator for linear models. *Electronic Journal of Statistics* 2, 605–633.
- Osborne, M. R., B. Presnell, and B. A. Turlach (2000). On the LASSO and its dual. *J. Comput. Graph. Statist.* 9(2), 319–337.
- Roth, V. and B. Fischer (2008). The group-lasso for generalized linear models: Uniqueness of solutions and efficient algorithms. In *ICML '08: Proceedings of the 25th international conference on Machine Learning*, pp. 848–855.
- Stein, C. (1981). Estimation of the mean of a multivariate normal distribution. *The Annals of Statistics* 9, 1135–1151.
- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B Stat. Methodol.* 58(1), 267–288.
- Van der Vaart, A. (1998). *Asymptotic Statistics*. Cambridge University Press.
- Yuan, M. and Y. Lin (2006). Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. Ser. B Stat. Methodol.* 68(1), 49–67.
- Yuan, M. and Y. Lin (2007). On the non-negative garrotte estimator. *J. R. Stat. Soc. Ser. B Stat. Methodol.* 69(2), 143–161.
- Zou, H., T. Hastie, and R. Tibshirani (2007). On the "degrees of freedom" of the lasso. *The Annals of Statistics* 35(5), 2173–2192.