HAL
open science

Sparsity with sign-coherent groups of variables via the cooperative-lasso

Julien Chiquet, Yves Grandvalet, Camille Charbonnier

To cite this version:

Julien Chiquet, Yves Grandvalet, Camille Charbonnier. Sparsity with sign-coherent groups of variables via the cooperative-lasso. Annals of Applied Statistics, 2012, 6 (2), pp.795-830. 10.1214/11-AOAS520 . hal-00707281

HAL Id: hal-00707281

https://hal.science/hal-00707281v1

Submitted on 12 Jun 2012

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

The Annals of Applied Statistics
2012, Vol. 6, No. 2, 795–830
DOI: 10.1214/11-AOAS520
© Institute of Mathematical Statistics, 2012

SPARSITY WITH SIGN-COHERENT GROUPS OF VARIABLES VIA THE COOPERATIVE-LASSO

BY JULIEN CHIQUET, YVES GRANDVALET[1] AND CAMILLE CHARBONNIER

CNRS UMR 8071 & Université d'Évry and Université de Technologie de Compiègne—CNRS UMR 6599 Heudiasyc

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 nonnegative, nonpositive or null parameters. To tackle this problem, we propose 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 and sparse group-Lasso comply with our theoretical results, showing consistent improvements in support recovery for sign-coherent groups. We finally propose two examples illustrating the wide applicability of the cooperative-Lasso: first to the processing of ordinal variables, where the penalty acts as a monotonicity prior; second 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.

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 nonpositive, nonnegative or null parameters. All such groups will be referred to as *sign-coherent*.

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

$$Y = X\boldsymbol{\beta}^\star + \varepsilon = \sum_{k=1}^{K} \sum_{j \in \mathcal{G}_k} X_j \beta_j^\star + \varepsilon, \tag{1}$$

Received March 2011; revised September 2011.

[1]Supported in part by the PASCAL2 Network of Excellence, the European ICT FP7 Grant 247022—MASH and the French National Research Agency (ANR) Grant ClasSel ANR-08-EMER-002.

Key words and phrases. Penalization, sparsity, grouped variables, ordinal variables, continuous variables, sign-coherence, microarray analysis.

where Y is a continuous response variable, $X = (X_1, \dots, X_p)$ is a vector of p predictor variables, $\boldsymbol{\beta}^\star$ is the vector of unknown parameters and ε is a zero-mean Gaussian error variable with variance σ^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 $\boldsymbol{\beta}^\star$ has few nonzero 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 $\boldsymbol{\beta}^\star$ is based on training data, consisting of a vector $\mathbf{y} = (y_1, \dots, y_n)^\intercal$ for responses and a $n \times p$ design matrix $\mathbf{X}$ whose jth column contains $\mathbf{x}_j = (x_j^1, \dots, x_j^n)^\intercal$, 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 ℓ_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 later developed by Yuan and Lin (2006), uses the group structure to define a shrinkage estimator of the form

$$\hat{\boldsymbol{\beta}}^{\text{group}} = \underset{\boldsymbol{\beta} \in \mathbb{R}^p}{\arg\min} \left\{ \frac{1}{2} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda \sum_{k=1}^{K} w_k \|\boldsymbol{\beta}_{\mathcal{G}_k}\| \right\}, \tag{2}$$

where $\mathcal{G}_k$ is the subset of indices defining the kth 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 $w_k = \sqrt{p_k}$, where p_k is the cardinality of $\mathcal{G}_k$ 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 λ which are large enough. Note that when we assign one group to each predictor, we recover the original Lasso [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, van de Geer and Bühlmann (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. Here, we build on the subdifferential calculus approach originally proposed by Osborne, Presnell and Turlach (2000) for the Lasso, whose active set algorithm has been adapted to the group-Lasso [Roth and Fischer (2008)].

Compared to the group-Lasso, this paper deals with a stronger 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, nonpositive or nonnega-

tive. This desideratum arises often when the groups gather redundant or consonant variables (a usual outcome when groups are defined from clusters of correlated variables). To perform this sign-coherent grouped variable selection, we propose a novel penalty that we call the cooperative-Lasso, in short the *coop-Lasso*.

The coop-Lasso is amenable to the selection of patterns that cannot be achieved with the group-Lasso. This ability, which can be observed for finite samples, also leads to consistency results under the mildest assumptions. Indeed, the consistency results for the group-Lasso assume that the set of nonzero coefficients of $\boldsymbol{\beta}^\star$ is an exact union of groups [Bach (2008); Nardi and Rinaldo (2008)], while exact support recovery may be achieved with coop-Lasso when some zero coefficients belong to a group having either positive or negative coefficients. For example, with groups $\mathcal{G}_1 = \{1, 2\}$ and $\mathcal{G}_2 = \{3, 4, 5\}$, the support of $\boldsymbol{\beta}^\star = (-1, 1, 0, 1, 1)^\intercal$ may be recovered with the coop-Lasso, but not with the group-Lasso, which may then deteriorate the performances of the Lasso [Huang and Zhang (2010)]. Friedman, Hastie and Tibshirani (2010) propose to overcome this restriction by adding an ℓ_1 penalty to the objective function in (2), in the vein of the hierarchical penalties of Zhao, Rocha and Yu (2009). The new term provides additional flexibility but demands an additional tuning parameter, while our approach takes a different stance by assuming sign-coherence, with the benefit of requiring a single tuning parameter.

Section 6 describes two applications where sign-coherence is a sensible assumption. The first one considers ordered categorical data, which are common in regression and classification. The coop-Lasso can then be used to induce a monotonic response to the ordered levels of a covariate, without translating each level of the categorical variable into a prescribed quantitative value. The second application describes the situation where redundancy in measurements causes sign-coherence to be expected. Similar behaviors should be observed when features have been grouped by a clustering algorithm such as average linkage hierarchical clustering, which are nowadays routinely used for grouping genes in microarray data analysis [Eisen et al. (1998); Park, Hastie and Tibshirani (2007); Ma, Song and Huang (2007)].

Finally, in numerous problems of multiple inference, the sign-coherence assumption is also reasonable: when predicting closely related responses (e.g., regressing male and female life expectancy against economic and social variables) or when analyzing multilevel data (e.g., predicting academic achievement against individual factors across schools), the set of coefficients associated to a predictor (resp., for all response variables or all data clusters) forms a group that can often be considered as sign-coherent because effects can be assumed to be qualitatively similar. Along these lines, we successfully applied the coop-Lasso penalizer for the joint inference of several network structures [Chiquet, Grandvalet and Ambroise (2011)].

The rest of the paper is organized as follows: Section 2 presents the coop-Lasso penalty, with the derivation of the optimality conditions which are the basis for an

active set algorithm. Consistency results and the associated irrepresentable conditions are given in Section 3. In Section 4 we derive an approximation of the degrees of freedom that can be used in the Bayesian Information Criterion (BIC) and the Akaike Information Criterion (AIC) for model selection. Section 5 is dedicated to simulations assessing the performances of the coop-Lasso in terms of sparsity pattern recovery, parameters estimation and robustness. Section 6 considers real data sets, with ordinal and continuous covariates. Note that all proofs are postponed until the Appendix.

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,

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

Let $\mathbf{v} = (v_1, \dots, v_p)^\intercal \in \mathbb{R}^p$ 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

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

where $w_k > 0$ are fixed parameters enabling to adapt the amount of penalty for each group. Likewise, the sparse group-Lasso norm [Friedman, Hastie and Tibshirani (2010)] is defined as a convex combination of the group-Lasso and the ℓ_1 norms:

$$\|\mathbf{v}\|_{\text{sgl}} = \alpha \|\mathbf{v}\|_{\text{group}} + (1 - \alpha) \|\mathbf{v}\|_1, \tag{4}$$

where α is meant to be a tuning parameter, but may be fixed to $1/2$ [Friedman, Hastie and Tibshirani (2010); Zhou et al. (2010)]. We will always set it to this default value in what follows.

Let $\mathbf{v}^+ = (v_1^+, \dots, v_p^+)^\intercal$ and $\mathbf{v}^- = (v_1^-, \dots, v_p^-)^\intercal$ 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 $\boldsymbol{\beta}^\star$ as defined in (1) is

$$\hat{\boldsymbol{\beta}}^{\text{coop}} = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\arg\min}\, L(\boldsymbol{\beta}) \qquad \text{with } L(\boldsymbol{\beta}) = \frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda\|\boldsymbol{\beta}\|_{\text{coop}}, \tag{5}$$

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

Illustrations of the group, sparse group and coop norms are given in Figure 1 for a vector $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3, \beta_4)^\intercal$ with two groups $\mathcal{G}_1 = \{1, 2\}$ and $\mathcal{G}_2 = \{3, 4\}$. We represent several views of the unit ball for each of these norms. For the coop-norm, this ball represents the set of feasible solutions for an optimization problem equivalent to (5), where the sum of squared residuals is minimized under unitary constraints on $\|\boldsymbol{\beta}\|_{\text{coop}}$. The same interpretation holds for the group and sparse group norms, provided the sum of squared residuals is minimized under unitary constraints on $\|\boldsymbol{\beta}\|_{\text{group}}$ and $\|\boldsymbol{\beta}\|_{\text{sgl}}$, 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 row illustrates that when β_4 is null its group companion β_3 may also be exactly zero (corners on the boundary at $\beta_3 = 0$), while the second row shows that this event is improbable when β_4 differs from zero (smooth boundary at $\beta_3 = 0$). The second and third columns display the same type of relationships within $\mathcal{G}_1$ between β_2 and β_1, which are expected due to the symmetries of the unit ball. The last column displays ℓ_2 balls, which characterize the within-groups feasibility subsets, showing that once a group is activated, all its members will be nonzero.

Now, consider the sparse group-norm: the combination of the group and Lasso penalties has uniformly shrunk the feasible set toward the Lasso ℓ_1 unit ball, thus creating new edges that provide a chance to zero any parameter in any situation, with an elastic-net-like penalty [Zou and Hastie (2005)] within and between groups. The comparison of the last two columns illustrates that the differentiation between the within-group and between group penalties is less marked than for the group-Lasso.

Finally, consider the coop-norm: compared to the group-norm, there are also additional discontinuities resulting in new edges on the 3-D plots. While the sparse group-Lasso edges where created by a uniform shrinking toward the ℓ_1 unit ball, the coop-Lasso new edges result from slicing the group-Lasso unit ball, depriving sign-incoherent orthants from some of the group-Lasso feasible solutions ($\|\boldsymbol{\beta}\|_{\text{coop}} > \|\boldsymbol{\beta}\|_{\text{group}}$ in these regions). Note that, in general, there are less new edges than with the sparse group-Lasso, since the new opportunities to zero some coefficients are limited to the case where the group-Lasso would have allowed a solution with opposite signs within a group. The crucial difference with the group and sparse group-Lasso is the loss of the axial symmetry when some

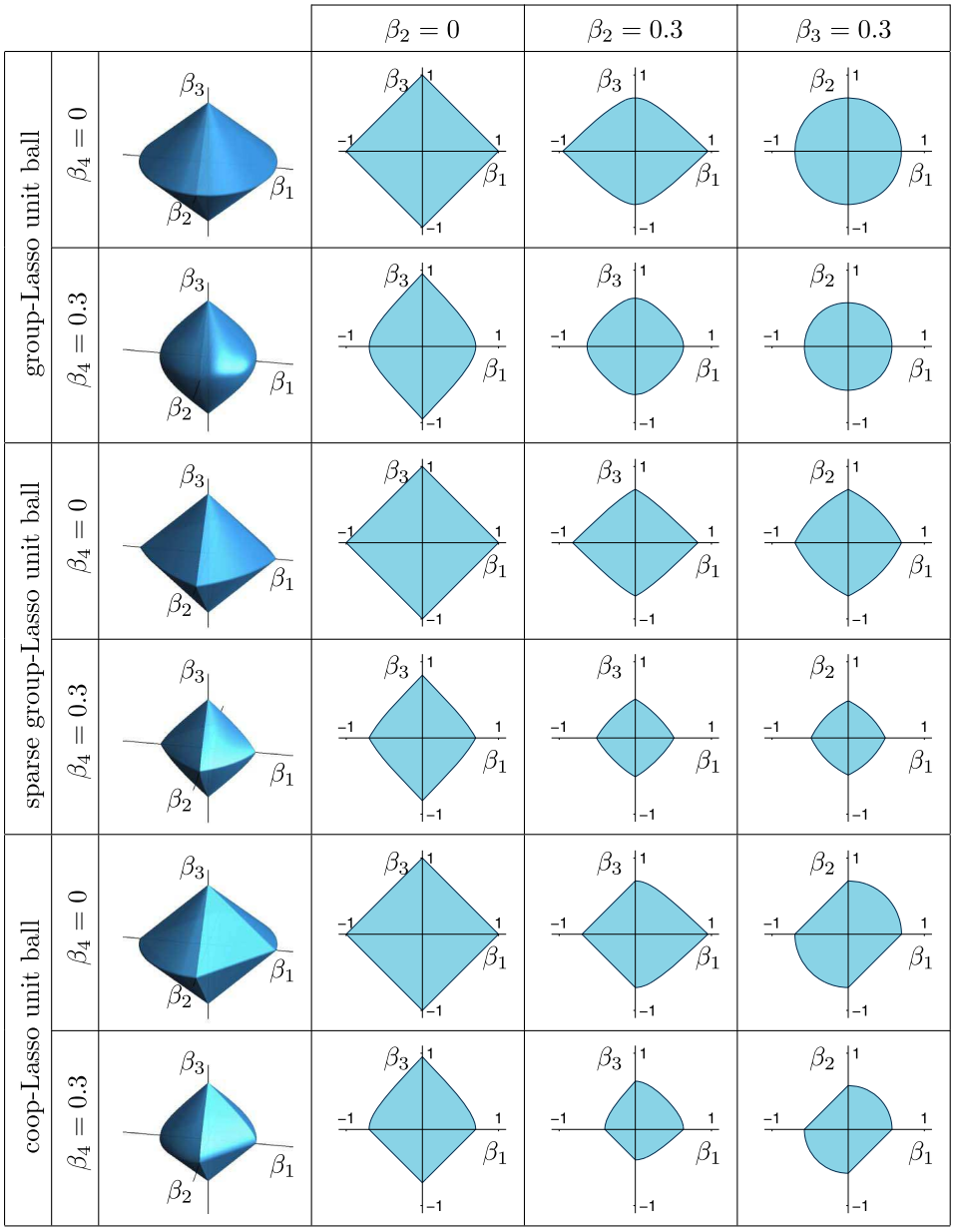

FIG. 1. *Feasible sets for the coop-Lasso, group-Lasso and sparse group-Lasso penalties. First column: cuts through* $(\beta_1, \beta_2, \beta_3)$ *at* $\beta_4 = 0$ *and* $\beta_4 = 0.3$: (β_1, β_2) *span the horizontal plane and* β_3 *is on the vertical axis; second and third columns: cuts through* (β_1, β_3) *at various values of* (β_2, β_4); *last column: cuts through* (β_1, β_2) *at various values of* (β_3, β_4).

variables are nonzero: decoupling the positive and negative parts of the regression coefficients favors solutions where signs match within a group. Slicing of the unit group-norm ball does not affect the positive and negative orthants, but large areas corresponding to sign mismatches have been peeled off, as best seen on the last column, which also illustrates the strong differentiation between within-group and between-group penalties.

Before stating the optimality conditions for problem (5), we introduce some 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 $\boldsymbol{\beta}^\star$ is typically sparse; its support is denoted $\mathcal{S} = \{j, \boldsymbol{\beta}^\star_j \neq 0\}$ and $\mathcal{S}^c = \{j, \boldsymbol{\beta}^\star_j = 0\}$ 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 $\mathcal{S}^c_k = \mathcal{S}^c \cap \mathcal{G}_k$ as the sets of relevant, respectively irrelevant, predictors within group k, for all $k = 1, \dots, K$. Similar notation $\mathcal{S}(\boldsymbol{\beta})$, $\mathcal{S}_k(\boldsymbol{\beta})$ and $\mathcal{S}^c_k(\boldsymbol{\beta})$ is defined for an arbitrary vector $\boldsymbol{\beta} \in \mathbb{R}^p$. Furthermore, for clarity and brevity, we introduce the functions $\{\boldsymbol{\varphi}_j\}_{j=1}^p$, which return the componentwise positive or negative part of a vector according to the sign of its jth element, that is, $\forall k \in \{1, \dots, K\}, \forall j \in \mathcal{G}_k, \forall \mathbf{v} \in \mathbb{R}^{p_k}$,

$$\boldsymbol{\varphi}_j(\mathbf{v}) = (\operatorname{sign}(v_j)\mathbf{v})^+ = \begin{cases} \mathbf{0}, & \text{if } v_j = 0, \\ \mathbf{v}^+, & \text{if } v_j > 0, \\ \mathbf{v}^-, & \text{if } v_j < 0. \end{cases} \tag{6}$$

Optimality conditions. The objective function L in (5) is continuous and coercive, thus problem (5) 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 $\boldsymbol{\beta}$ 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 $\boldsymbol{\beta} \in \mathbb{R}^p$, the subdifferential of the objective function of problem (5) is*

$$\partial_{\boldsymbol{\beta}} L(\boldsymbol{\beta}) = \{\mathbf{v} \in \mathbb{R}^p : \mathbf{v} = \mathbf{X}^\intercal(\mathbf{X}\boldsymbol{\beta} - \mathbf{y}) + \lambda\boldsymbol{\theta}\}, \tag{7}$$

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

$$\forall k \in \{1, \dots, K\}, \forall j \in \mathcal{S}_k(\boldsymbol{\beta}) \qquad \theta_j = \frac{w_k \beta_j}{\|\boldsymbol{\varphi}_j(\boldsymbol{\beta}_{\mathcal{G}_k})\|}, \tag{8a}$$

$$\forall k \in \{1, \dots, K\}, \forall j \in \mathcal{S}^c_k(\boldsymbol{\beta}) \qquad \|\boldsymbol{\varphi}_j(\boldsymbol{\theta}_{\mathcal{G}_k})\| \leq w_k. \tag{8b}$$

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* (5) *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*:

$$\text{(9a)} \qquad \forall k \in \{1, \dots, K\}, \forall j \in \mathcal{S}_k(\boldsymbol{\beta}) \qquad \mathbf{x}_j^\intercal(\mathbf{X}\boldsymbol{\beta} - \mathbf{y}) + \frac{\lambda w_k \beta_j}{\|\boldsymbol{\varphi}_j(\boldsymbol{\beta}_{\mathcal{G}_k})\|} = 0,$$

$$\text{(9b)} \qquad \forall k \in \{1, \dots, K\}, \forall j \in \mathcal{S}_k^c(\boldsymbol{\beta}) \qquad \left\|\boldsymbol{\varphi}_j\big((\mathbf{X}_{\cdot\mathcal{G}_k})^\intercal(\mathbf{X}\boldsymbol{\beta} - \mathbf{y})\big)\right\| \le \lambda w_k,$$

where $\mathbf{X}_{\cdot\mathcal{G}_k}$ *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. The asymptotic analysis of Section 3 confirms that exact support recovery is possible even when the support of $\boldsymbol{\beta}^\star$ 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, Presnell and Turlach (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, Grandvalet and Ambroise (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:

1. The first step solves problem (5) with respect to $\boldsymbol{\beta}_{\mathcal{A}}$, the subset of "active" variables, currently identified as being nonzero. 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. One or more variables may then be declared inactive if the current optimal $\boldsymbol{\beta}_{\mathcal{A}}$ reaches the boundary of the current feasible set.
2. The second step assesses the completeness of the set $\mathcal{A}$, by checking the optimality conditions with respect to inactive variables. We add a group that violates these conditions. In our implementation, we pick the one that most violates the optimality condition, since this strategy has been observed to require few changes in the active set. 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:

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

$$\boldsymbol{\beta}_{\mathcal{A}} \leftarrow \arg\min_{\mathbf{v} \in \mathbb{R}^{|\mathcal{A}|}} \frac{1}{2}\|\mathbf{y} - \mathbf{X}_{.\mathcal{A}}\mathbf{v}\|^2 + \lambda\|\mathbf{v}\|_{\text{coop}}$$

$$\text{s.t.} \begin{cases} v_j \geq 0, & \text{if } j \in \mathcal{A}_+ \cap \mathcal{A}_-^c, \\ v_j \leq 0, & \text{if } j \in \mathcal{A}_- \cap \mathcal{A}_+^c, \end{cases}$$

where $\mathcal{A}_-^c$ and $\mathcal{A}_+^c$ are the complementary sets of $\mathcal{A}_-$ and $\mathcal{A}_+$, respectively. Identify groups inactivated during optimization

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

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

Step 2 Identify the greatest violation of optimality conditions:

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

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

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

//stat.genopole.cnrs.fr/logiciels/scoop) the linear and logistic regression models are implemented using either Broyden–Fletcher–Goldfarb–Shanno (BFGS) quasi-Newton updates with box constraints, or proximal methods [Beck and Teboulle (2009)] to solve the smooth optimization problem in Step 1.

Finally, note that to compute a series of solutions along the regularization path for problem (5), we simply choose a series of penalties $\lambda^1 = \lambda_{\max} > \cdots > \lambda^l > \cdots > \lambda^L = \lambda_{\min} \geq 0$ such that $\hat{\boldsymbol{\beta}}^{\text{coop}}(\lambda_{\max}) = \mathbf{0}$, that is,

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

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 λ^l, is initialized with $\hat{\boldsymbol{\beta}}^{\text{coop}}(\lambda^{l-1})$.

2.3. *Orthonormal design case.* The orthonormal design case, where $\mathbf{X}^\intercal\mathbf{X} = \mathbf{I}_p$, 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 β_j, we have $\mathbf{x}_j^\intercal(\mathbf{X}\boldsymbol{\beta} - \mathbf{y}) = \beta_j - \hat{\beta}_j^{\text{ols}}$. The optimality conditions (9a) and (9b) can then be written as

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

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

$$\forall k \in \{1,\dots,K\}, \forall j \in \mathcal{G}_k \qquad \hat{\beta}_j^{\text{group}} = \left(1 - \frac{\lambda w_k}{\|\hat{\boldsymbol{\beta}}_{\mathcal{G}_k}^{\text{ols}}\|}\right)^+ \hat{\beta}_j^{\text{ols}}, \tag{11}$$

while the Lasso solution [Tibshirani (1996)] is

$$\forall j \in \{1,\dots,p\} \qquad \hat{\beta}_j^{\text{lasso}} = \left(1 - \frac{\lambda}{|\hat{\beta}_j^{\text{ols}}|}\right)^+ \hat{\beta}_j^{\text{ols}}. \tag{12}$$

Equations (10)–(12) reveal strong commonalities. First, the coefficients of these shrinkage estimators are of the sign of the OLS estimates. Second, the norm used in the penalty defines a region where small OLS coefficients are shrunk to zero, while large ones are shrunk inversely proportional to this norm. Finally, by grouping the terms corresponding to one group in equations (10)–(11), a uniform translation effect, analogous to the one observed for the Lasso, comes into view:

$$\begin{aligned} &\forall k \in \{1,\dots,K\}, \forall j \in \mathcal{G}_k && \|\boldsymbol{\varphi}_j(\hat{\boldsymbol{\beta}}_{\mathcal{G}_k}^{\text{coop}})\| = \big(\|\boldsymbol{\varphi}_j(\hat{\boldsymbol{\beta}}_{\mathcal{G}_k}^{\text{ols}})\| - \lambda w_k\big)^+, \\ &\forall k \in \{1,\dots,K\} && \|\hat{\boldsymbol{\beta}}_{\mathcal{G}_k}^{\text{group}}\| = (\|\hat{\boldsymbol{\beta}}_{\mathcal{G}_k}^{\text{ols}}\| - \lambda w_k)^+, \\ &\forall j \in \{1,\dots,p\} && |\hat{\beta}_j^{\text{lasso}}| = (|\hat{\beta}_j^{\text{ols}}| - \lambda w_k)^+. \end{aligned} \tag{13}$$

The group-Lasso (11) differs primarily from the Lasso (12) owing to the common penalty $\lambda w_k / \|\hat{\boldsymbol{\beta}}^{\text{ols}}_{\mathcal{G}_k}\|$ for all the coefficients belonging to group k. The magnitude of shrinkage is determined by all within-group OLS coefficients, and is thus radically different from a ridge regression penalty in this regard. For the coop-Lasso estimator (10), two penalties possibly apply to group k, for the positive and the negative OLS coefficients, respectively. If all within-group OLS coefficients are of the 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 an 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.

Note that such a simple analytical formulation is not available for the sparse group-Lasso estimate $\hat{\boldsymbol{\beta}}^{\text{sgl}}$, but an expression can be obtained by chaining two simple shrinkage operations. Introducing an intermediate solution $\tilde{\boldsymbol{\beta}}^{\text{sgl}}$, we have, $\forall k \in \{1, \dots, K\}$ and $\forall j \in \mathcal{G}_k$,

$$\hat{\beta}^{\text{sgl}}_j = \left(1 - \frac{\lambda(1-\alpha) w_k}{\|\tilde{\boldsymbol{\beta}}^{\text{sgl}}_{\mathcal{G}_k}\|}\right)^+ \tilde{\beta}^{\text{sgl}}_j \qquad \text{where } \tilde{\beta}^{\text{sgl}}_j = \left(1 - \frac{\lambda\alpha}{|\hat{\beta}^{\text{ols}}_j|}\right)^+ \hat{\beta}^{\text{ols}}_j. \tag{14}$$

The intermediate solution $\tilde{\boldsymbol{\beta}}^{\text{sgl}}$ is the Lasso estimator with penalty parameter $\lambda\alpha$, which acts as the OLS estimate for a group-Lasso of parameter $\lambda(1-\alpha)$.

Figure 2 provides a visual representation of equations (10)–(12) and (14) for a group with two components, say, $\mathcal{G}_k = \{1, 2\}$. We plot $\hat{\beta}^{\text{lasso}}_1$, $\hat{\beta}^{\text{group}}_1$, $\hat{\beta}^{\text{sgl}}_1$ and $\hat{\beta}^{\text{coop}}_1$ as functions of $(\hat{\beta}^{\text{ols}}_1, \hat{\beta}^{\text{ols}}_2)$. Top-left, the Lasso translates the $\hat{\beta}^{\text{ols}}_1$ coefficient toward zero, eventually truncating them at zero, regardless of $\hat{\beta}^{\text{ols}}_2$: there is no interaction between coefficients. The group-Lasso, top-right, has a nonlinear shrinking behavior (quite different from the Lasso or ridge penalties in this respect) and sets $\hat{\beta}^{\text{group}}_1$ to zero within a Euclidean ball centered at zero. The sparse group-Lasso, bottom-left, is a hybrid of Lasso and group-Lasso, whose shrinking behavior lies between its two ancestors. Bottom-right, the coop-Lasso appears as another form of cross-breed, identical to the group-Lasso in the positive and negative quadrants, and identical to the Lasso when the signs of the OLS coefficients mismatch. For groups with more than two components, intermediate solutions would be possible. This behavior is shown to allow for some flexibility with respect to the predefined group structure 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 of 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

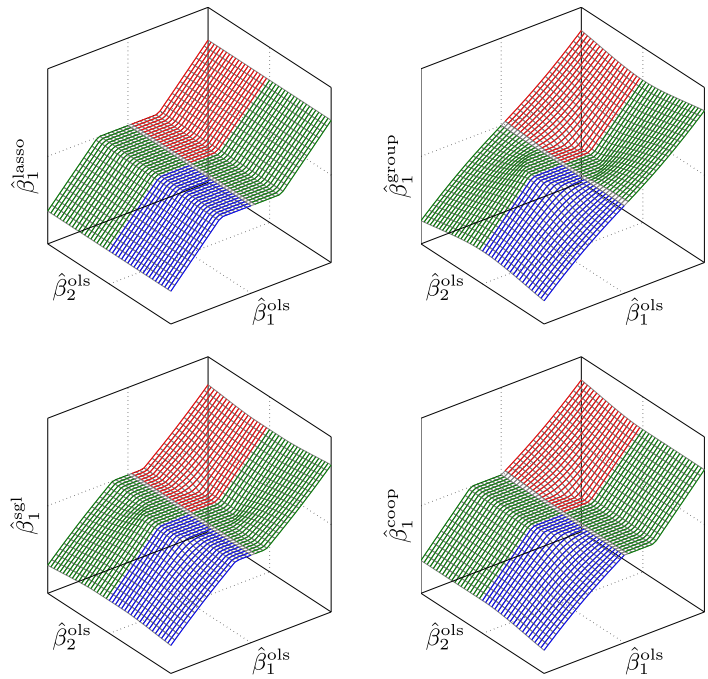

FIG. 2. *Lasso, group, sparse group and coop Lasso coefficient estimates, for a group with* 2 *elements* $\mathcal{G}_k = \{1, 2\}$, *as a function of the OLS coefficients. The colors emphasize the positive and negative quadrants of the* $(\hat{\beta}_1^{\text{ols}}, \hat{\beta}_2^{\text{ols}})$ *plane, with red and blue, respectively.*

simplicity and coherence, we keep assuming that data are centered so that we have zero mean random variables and $\boldsymbol{\Psi} = \mathbb{E}[XX^\intercal]$ is the covariance matrix of X:

(A1) X and Y have finite 4th order moments $\mathbb{E}[\|X\|^4] < \infty$, $\mathbb{E}[Y^4] < \infty$.
(A2) The covariance matrix $\boldsymbol{\Psi} = \mathbb{E}[XX^\intercal] \in \mathbb{R}^{p \times p}$ 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 included in the true support: $\forall k \in \{1, \dots, K\}$, if $\|(\boldsymbol{\beta}^\star_{\mathcal{G}_k})^+\| > 0$ and $\|(\boldsymbol{\beta}^\star_{\mathcal{G}_k})^-\| > 0$, then $\forall j \in \mathcal{G}_k$, $\boldsymbol{\beta}^\star_j \neq 0$.

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.

The spurious relationships that may arise from confounding variables are controlled by the so-called strong irrepresentable condition, which guarantees support recovery for the Lasso [Yuan and Lin (2007)] and the group-Lasso [Bach (2008)].

We now introduce suitable variants of these conditions for the coop-Lasso. They 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 $\mathbf{D}(\boldsymbol{\beta})$ such that,

$$\forall k \in \{1, \dots, K\}, \forall j \in \mathcal{S}_k(\boldsymbol{\beta}) \qquad (\mathbf{D}(\boldsymbol{\beta}))_{jj} = w_k \|\boldsymbol{\varphi}_j(\boldsymbol{\beta}_{\mathcal{G}_k})\|^{-1}. \tag{15}$$

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

$$\frac{1}{w_k} \max(\|(\boldsymbol{\Psi}_{\mathcal{S}_k^c\mathcal{S}} \boldsymbol{\Psi}_{\mathcal{S}\mathcal{S}}^{-1} \mathbf{D}(\boldsymbol{\beta}_{\mathcal{S}}^\star)\boldsymbol{\beta}_{\mathcal{S}}^\star)^+\|, \|(\boldsymbol{\Psi}_{\mathcal{S}_k^c\mathcal{S}} \boldsymbol{\Psi}_{\mathcal{S}\mathcal{S}}^{-1} \mathbf{D}(\boldsymbol{\beta}_{\mathcal{S}}^\star)\boldsymbol{\beta}_{\mathcal{S}}^\star)^-\|) \leq 1 - \eta, \tag{16}$$

where $\boldsymbol{\Psi}_{\mathcal{S}\mathcal{T}}$ is the submatrix of $\boldsymbol{\Psi}$ 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, letting ν_k be the sign of these coefficients [$\nu_k = 1$ if $\|(\boldsymbol{\beta}_{\mathcal{G}_k}^\star)^+\| > 0$ and $\nu_k = -1$ if $\|(\boldsymbol{\beta}_{\mathcal{G}_k}^\star)^-\| > 0$], the following inequalities should hold:

$$\nu_k \boldsymbol{\Psi}_{\mathcal{S}_k^c\mathcal{S}} \boldsymbol{\Psi}_{\mathcal{S}\mathcal{S}}^{-1} \mathbf{D}(\boldsymbol{\beta}_{\mathcal{S}}^\star)\boldsymbol{\beta}_{\mathcal{S}}^\star \preceq \mathbf{0}, \tag{17}$$

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 (16), except that the elements of the weighting matrix $\mathbf{D}$ are $w_k \|\boldsymbol{\beta}_{\mathcal{G}_k}\|^{-1}$ and that the ℓ_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 (5) to allow a direct comparison with the known similar results previously stated for the Lasso and group-Lasso [Yuan and Lin (2007); Bach (2008)]:

$$\hat{\boldsymbol{\beta}}_n^{\text{coop}} = \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \frac{1}{2n} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda_n \|\boldsymbol{\beta}\|_{\text{coop}}, \tag{18}$$

where $\lambda_n = \lambda/n$.

THEOREM 2. *If assumptions* (A1)–(A5) *are satisfied*, *the coop-Lasso estimator is asymptotically unbiased and has the property of exact support recovery*:

$$\hat{\boldsymbol{\beta}}_n^{\text{coop}} \xrightarrow{P} \boldsymbol{\beta}^\star \quad \textit{and} \quad \mathbb{P}\big(\mathcal{S}(\hat{\boldsymbol{\beta}}_n^{\text{coop}}) = \mathcal{S}\big) \to 1, \tag{19}$$

for every sequence λ_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 (16) is weaker. Indeed, the norm in (16) is dominated by the ℓ_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 comprising sign-incoherent ones, 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 $\boldsymbol{\beta}^\star = (1, 1, -1, -1, 0, 0, 0, 0)$, 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 $\boldsymbol{\Psi}$ is chosen so that the irrepresentable conditions hold for the coop-Lasso, but not for the group-Lasso, which, we recall, are more demanding for the current situation, with sign-coherent groups. The random error ε follows a centered Gaussian distribution with standard deviation $\sigma = 0.1$, inducing 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 1000 samples of size $n = 20$ from the described model, and computed the corresponding 1000 regularization paths for the group-Lasso, sparse group-Lasso and coop-Lasso. Figure 3 reports the 50% coverage intervals (lower and upper quartiles) along the regularization paths. In this setup, the sparse group-Lasso behaves as the group-Lasso, leading to nearly identical graphs. Estimation is difficult in this small sample problem ($n = 20$, $p = 8$), and the two versions of the group-Lasso, which first select the wrong covariates, never reach the situation where they would have a decisive advantage upon OLS, while the coop-Lasso

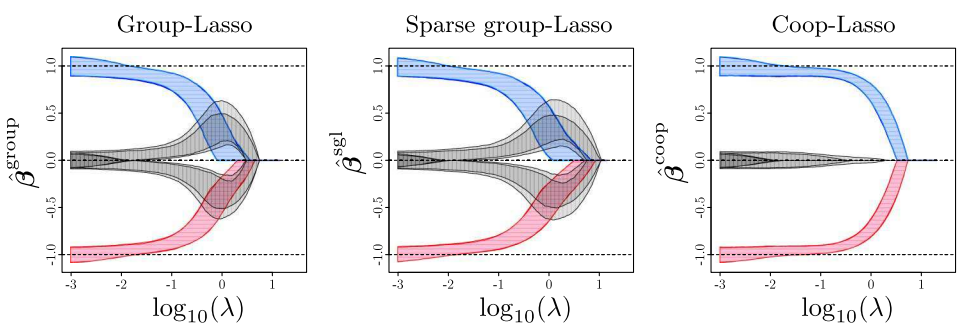

FIG. 3. 50% *coverage intervals for the group* (*left*), *sparse group* (*center*) *and* (*right*) *Lasso estimated coefficients along regularization paths*: *coefficients from the support of* $\boldsymbol{\beta}^\star$ *are marked by colored horizontal stripes and the other ones by gray vertical stripes*.

immediately selects the right covariates, whose coefficients steadily dominate the irrelevant ones. Model selection is also difficult, and the BIC criteria provided in Section 4 select often the OLS model (in about 10% and 50% of cases for the coop-Lasso and the group-Lasso, respectively). The average root mean square error on parameters is of order 10^{-1} for all methods, with a slight edge for the coop-Lasso. The sign error is much more contrasted: 31% for the coop-Lasso *vs.* 46% for the group-Lasso, not far better than the 50% of OLS.

4. Model selection. Model selection amounts here to choosing the penalization parameter λ, which restricts the size of the estimate $\hat{\boldsymbol{\beta}}(\lambda)$. 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 $\hat{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}^\star$, either based on the value of the coefficients or on their support (true model selection), and sometimes also on the sign correctness of each nonzero 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 λ. However, it does not provide a proper value to be plugged in (5) 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, Hastie and Tibshirani (2007)] and group-Lasso [Yuan and Lin (2006)] have been defined through the effective degrees of freedom:

$$\mathrm{AIC}(\lambda) = \frac{\|\mathbf{y} - \hat{\mathbf{y}}(\lambda)\|^2}{\sigma^2} + 2\,\mathrm{df}(\lambda), \tag{20}$$

$$\mathrm{BIC}(\lambda) = \frac{\|\mathbf{y} - \hat{\mathbf{y}}(\lambda)\|^2}{\sigma^2} + \log(n)\,\mathrm{df}(\lambda), \tag{21}$$

where $\hat{\mathbf{y}}(\lambda) = \mathbf{X}\hat{\boldsymbol{\beta}}(\lambda)$ is the vector of predicted values for (5) with penalty parameter λ, σ^2 is the variance of the zero-mean Gaussian error variable ε in (1) and $\mathrm{df}(\lambda)$ 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

$$\mathrm{df}(\lambda) \doteq \frac{1}{\sigma^2}\sum_{i=1}^{n} \mathrm{cov}(\hat{y}_i(\lambda), y_i) = \mathbb{E}\left[\mathrm{tr}\left(\frac{\partial \hat{\mathbf{y}}(\lambda)}{\partial \mathbf{y}}\right)\right], \tag{22}$$

where the expectation is taken with respect to $\mathbf{y}$ or, equivalently, to the noise $\boldsymbol{\varepsilon}$. Yuan and Lin (2006) proposed an approximation of the trace term in the right-hand side of (22), which is used to estimate $\mathrm{df}(\lambda)$ for the group-Lasso:

$$\widetilde{\mathrm{df}}_{\mathrm{group}}(\lambda) = \sum_{k=1}^{K} \mathbb{1}\big(\|\hat{\boldsymbol{\beta}}^{\mathrm{group}}_{\mathcal{G}_k}(\lambda)\| > 0\big)\bigg(1 + \frac{\|\hat{\boldsymbol{\beta}}^{\mathrm{group}}_{\mathcal{G}_k}(\lambda)\|}{\|\boldsymbol{\beta}^{\mathrm{ols}}_{\mathcal{G}_k}\|}(p_k - 1)\bigg), \tag{23}$$

where $\mathbb{1}(\cdot)$ is the indicator function and p_k is the number of elements in $\mathcal{G}_k$. For orthonormal design matrices, (23) 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 (23)] is generally comparable with that of fivefold cross-validation and is sometimes better."

This approximation of $\mathrm{df}(\lambda)$ 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

$$\hat{\boldsymbol{\beta}}^{\mathrm{ridge}}(\gamma) = (\mathbf{X}^{\intercal}\mathbf{X} + \gamma\mathbf{I})^{-1}\mathbf{X}^{\intercal}\mathbf{y}, \tag{24}$$

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

PROPOSITION 1. *Consider the coop-Lasso estimator* $\hat{\boldsymbol{\beta}}^{\mathrm{coop}}(\lambda)$ *defined by* (5). *Assuming that data are generated according to model* (1), *and that* $\mathbf{X}$ *is orthonormal*, *the following expression of* $\widetilde{\mathrm{df}}_{\mathrm{coop}}(\lambda)$ *is an unbiased estimate of* $\mathrm{df}(\lambda)$ *defined in* (22) *for the coop-Lasso fit*:

$$\begin{aligned}\widetilde{\mathrm{df}}_{\mathrm{coop}}(\lambda) = \sum_{k=1}^{K} &\mathbb{1}\big(\|(\hat{\boldsymbol{\beta}}^{\mathrm{coop}}_{\mathcal{G}_k}(\lambda))^{+}\| > 0\big)\bigg(1 + \frac{p_+^k - 1}{1+\gamma}\frac{\|(\hat{\boldsymbol{\beta}}^{\mathrm{coop}}_{\mathcal{G}_k}(\lambda))^{+}\|}{\|(\hat{\boldsymbol{\beta}}^{\mathrm{ridge}}_{\mathcal{G}_k}(\gamma))^{+}\|}\bigg) \\ &+ \mathbb{1}\big(\|(\hat{\boldsymbol{\beta}}^{\mathrm{coop}}_{\mathcal{G}_k}(\lambda))^{-}\| > 0\big)\bigg(1 + \frac{p_-^k - 1}{1+\gamma}\frac{\|(\hat{\boldsymbol{\beta}}^{\mathrm{coop}}_{\mathcal{G}_k}(\lambda))^{-}\|}{\|(\hat{\boldsymbol{\beta}}^{\mathrm{ridge}}_{\mathcal{G}_k}(\gamma))^{-}\|}\bigg),\end{aligned} \tag{25}$$

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

Proposition 1 raises a practical issue regarding the choice of a good reference $\hat{\boldsymbol{\beta}}^{\mathrm{ridge}}(\gamma)$. In our numerous simulations (most of which are not reported here), we did not observe a high sensitivity to γ, though high values degrade performances. When $\mathbf{X}$ is full rank we use $\gamma = 0$ (the OLS estimate) and, correspondingly, a vanishing γ (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 nonorthonormal settings, plugging expression (25) for the degrees of freedom df(λ) of the coop-Lasso in BIC (21) or AIC (20) provides sensible model selection criteria. As expected, BIC, which is more stringent than AIC, is better at retrieving the sparsity pattern of $\boldsymbol{\beta}^\star$, 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 $\boldsymbol{\beta}^\star \in \mathbb{R}^p$ 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 $j \in \{1, \dots, 9\}$, $(\boldsymbol{\beta}^\star_{\mathcal{G}_k})_j \propto \nu_k((h - |5 - j|)^+)^2$, where $\nu_k \in \{0, 1\}$ is a switch at the group level and $h \in \{3, 4, 5\}$ governs the wave width, that is, the within-group sparsity, with respectively $|\mathcal{S}_k| \in \{5, 7, 9\}$ nonzero coefficients in each group included in the support. The covariates are drawn from a multivariate normal distribution $X \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Psi})$ with, for all $(j, j') \in \{1, \dots, p\}^2$, covariances $\Psi_{jj'} = \rho^{|j-j'|}$, 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 $\boldsymbol{\beta}^\star$ 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 $\boldsymbol{\beta}^\star$ matches the group structure.

5.2. *Results.* Model selection is performed with BIC (21) for Lasso, group-Lasso and coop-Lasso. The estimation of the degrees of freedom for the Lasso is the number of nonzero entries in $\hat{\boldsymbol{\beta}}^{\text{lasso}}(\lambda)$ [Zou, Hastie and Tibshirani (2007)]. As there is no such analytical estimate of the degrees of freedom for the sparse group-Lasso, we tested two alternative model selection strategies: standard five-fold cross-validation (CV), selecting the model with minimum cross-validation error, and the so-called "1-SE rule" [Breiman et al. (1984)], which selects the most constrained model whose cross-validation error is within one standard error of the minimum.

First, we display in Figure 4 an example of the regularization paths obtained for each method for a small training set size ($n = p/2 = 45$) drawn from the model with three active groups having two zero coefficients each ($|\mathcal{S}_k| = 7$, $p_k = 9$) and a moderate positive correlation level ($\rho = 0.4$). As expected, the nonzero coef-

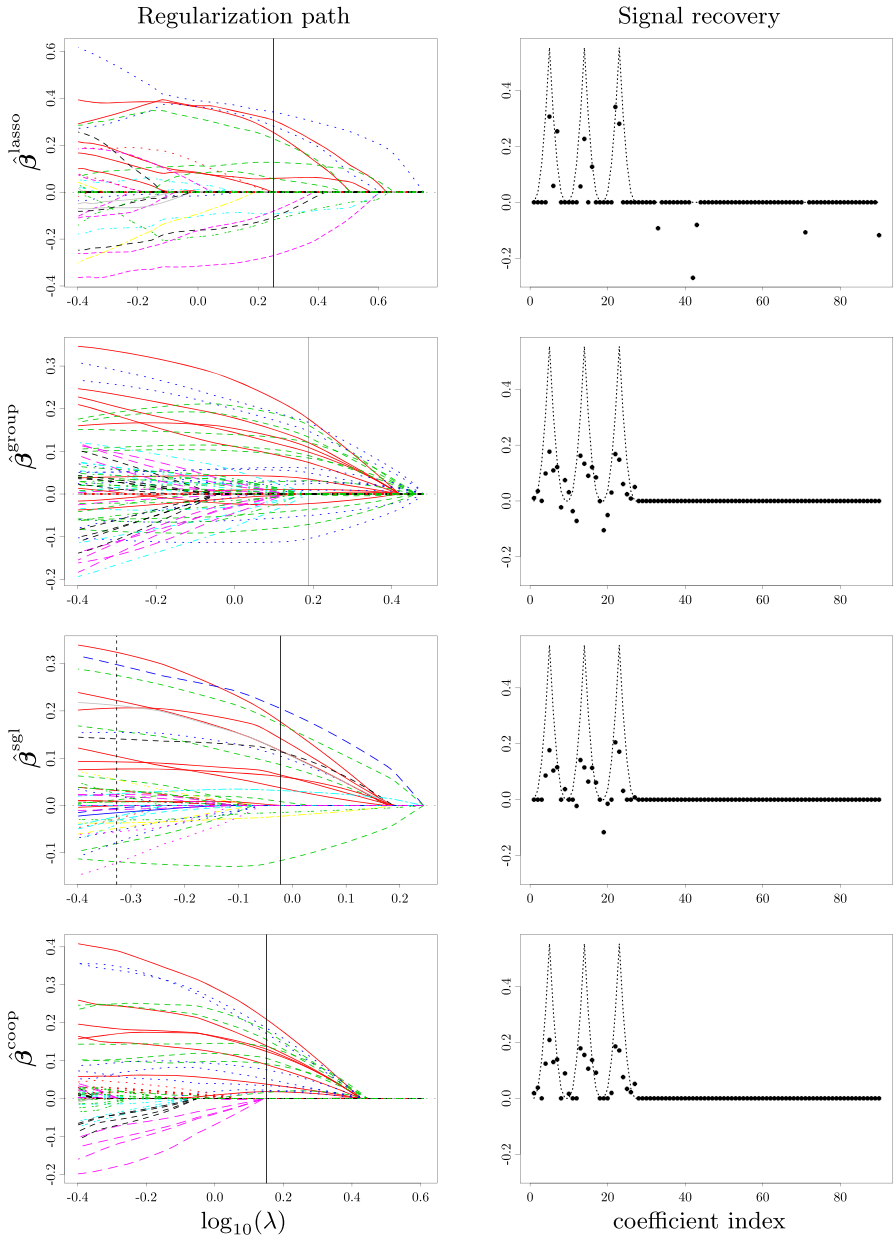

FIG. 4. *Lasso, group, sparse group and coop Lasso estimates for a training set of size* $n = 45$ *drawn from the generation process of Section* 5.1, *with* 3 *active waves out of* 10, $|\mathcal{S}_k|/p_k = 7/9$ *and* $\rho = 0.4$. *Left*: *regularization paths, where each line type/color represents a group of parameters and the plain vertical line marks the model selected by the* 1-*SE rule for sparse group-Lasso and BIC otherwise*; *right*: *true signal* (*dotted line*) *and estimated parameters for the selected model* (*filled circles*).

ficients appear one at a time along the Lasso regularization path and groupwise for the other methods, which detect the relevant groups early, with some coefficients kept to zero for the sparse group-Lasso and the coop-Lasso. The sparse group-Lasso is qualitatively intermediate between the group-Lasso and the coop-Lasso, setting many parameters to zero, but keeping a few negative coefficients in the solution. The coefficients of the model estimated by BIC or the 1-SE rule are displayed on the right of each path. The Lasso estimate includes some nonzero coefficients from irrelevant groups, but is otherwise quite conservative, excluding many nonzero parameters from its support. This conservative trend is also observed for the group methods, which exclude all irrelevant groups. The three group estimates mostly agree on truly important coefficients, and differ in the treatment of the spurious negative values that are frequent for group-Lasso, rarer for sparse group-Lasso and do not occur for coop-Lasso.

Table 1 provides a more objective evaluation of the compared methods, based on the root mean square error (RMSE) and the support recovery (more precisely, recovery of the sign of true parameters); prediction error (not shown) is tightly

TABLE 1

Average errors, with standard deviations, on 1000 *simulations from the setup described in Section* 5.1. *Each scenario differs in the number of observations n and the number of active variables per active group* $|\mathcal{S}_k|$ ($p_k = 9$). *Sparse-cv and sparse-*1*-se designate the sparse group-Lasso with* λ *selected by cross-validation and by the* 1*-SE rule, respectively*

		Lasso	**Group**	**Sparse-cv**	**Sparse-1-se**	**Coop**
Scenario				RMSE ($\times 10^3$)		
$\lvert\mathcal{S}_k\rvert = 5$	$n = 45$	87.1 (0.5)	95.0 (0.5)	82.5 (0.5)	88.1 (0.6)	84.2 (0.5)
	$n = 180$	43.7 (0.2)	49.1 (0.2)	41.7 (0.2)	44.9 (0.2)	43.5 (0.2)
	$n = 450$	28.8 (0.1)	33.4 (0.1)	27.2 (0.1)	30.9 (0.1)	29.4 (0.1)
$\lvert\mathcal{S}_k\rvert = 7$	$n = 45$	93.0 (0.5)	85.8 (0.5)	79.7 (0.4)	83.6 (0.5)	76.8 (0.5)
	$n = 180$	48.4 (0.2)	44.5 (0.2)	42.2 (0.2)	43.7 (0.2)	40.4 (0.2)
	$n = 450$	31.8 (0.1)	30.3 (0.1)	27.7 (0.1)	30.0 (0.1)	27.6 (0.1)
$\lvert\mathcal{S}_k\rvert = 9$	$n = 45$	99.2 (0.4)	82.0 (0.5)	81.0 (0.4)	83.2 (0.5)	73.7 (0.5)
	$n = 180$	52.5 (0.2)	41.9 (0.2)	43.3 (0.2)	43.8 (0.2)	39.0 (0.2)
	$n = 450$	34.1 (0.1)	28.7 (0.1)	28.8 (0.1)	30.6 (0.1)	27.1 (0.1)
Scenario				Mean sign error (%)		
$\lvert\mathcal{S}_k\rvert = 5$	$n = 45$	13.8 (0.1)	18.3 (0.2)	36.7 (0.4)	16.9 (0.3)	13.3 (0.2)
	$n = 180$	8.4 (0.1)	19.3 (0.2)	36.1 (0.4)	10.7 (0.2)	13.0 (0.2)
	$n = 450$	6.1 (0.1)	16.7 (0.2)	35.5 (0.4)	7.1 (0.2)	10.3 (0.2)
$\lvert\mathcal{S}_k\rvert = 7$	$n = 45$	18.9 (0.1)	12.9 (0.2)	34.6 (0.4)	16.8 (0.3)	10.1 (0.2)
	$n = 180$	11.9 (0.1)	12.7 (0.2)	34.5 (0.4)	10.5 (0.2)	9.8 (0.2)
	$n = 450$	8.8 (0.1)	10.4 (0.2)	34.9 (0.4)	7.1 (0.2)	7.7 (0.2)
$\lvert\mathcal{S}_k\rvert = 9$	$n = 45$	24.4 (0.1)	8.1 (0.2)	34.2 (0.4)	17.3 (0.3)	7.9 (0.2)
	$n = 180$	15.3 (0.1)	6.3 (0.2)	33.5 (0.4)	10.0 (0.2)	6.7 (0.2)
	$n = 450$	11.2 (0.1)	4.3 (0.1)	32.6 (0.4)	6.0 (0.2)	4.5 (0.1)

correlated with RMSE in our setup. Regarding the relative merits of the different methods, we did not observe a crucial role of the number of active groups and the covariate correlation level ρ. We report results for a true support comprising 3 groups out of 10 and $\rho = 0.4$, with various within-group sparsity and sample size scenarios.

All estimators perform about equally in RMSE, the sparse group-Lasso with CV having a slight advantage over the coop-Lasso when many zero coefficients belong to the active groups, and the coop-Lasso being marginally but significantly better elsewhere.

Regarding support recovery, model selection with CV leads to models overestimating the support of parameters. The 1-SE rule, which slightly harms RMSE, is greatly beneficial in this respect. BIC also performs very well, incurring a very small loss due to model selection compared to the oracle solution picking the model with best support recovery. The Lasso dominates all the groups methods when many zero coefficients belong to the active groups. Elsewhere, group methods (with appropriate model selection criteria) perform systematically significantly better for the small sample sizes. The coop-Lasso ranks first or a close second among group methods in all experimental conditions. It thus appears as the method of choice regarding inference issues when groups conform to the sign-coherence assumption.

5.3. *Robustness*. The robustness to violations of the sign-coherence assumption is assessed by switching a proportion P_σ of signs in the vector $\boldsymbol{\beta}^\star$, otherwise generated as before. The sign of the corresponding covariates are switched accordingly, to ensure that only the coop-Lasso estimators are affected in the process.

Table 2 displays the coop-Lasso RMSE that degrades gradually with the amount of perturbation, becoming eventually worse than the Lasso, except for full groups. Regarding sign error, for small proportions of sign flip, the coop-Lasso stays at par with either Lasso or group-Lasso (see Table 1), but it eventually becomes significantly worse than both of them in most situations. Thus, if the sign-coherence

TABLE 2

Average errors, with standard deviations, on 1000 *simulations from the setup of Table* 1 *with* $n = 180$, *perturbed by switching a proportion* P_σ *of signs in* $\boldsymbol{\beta}^\star$

	RMSE ($\times 10^3$)			**Mean sign error (%)**		
P_σ	$\lvert\mathcal{S}_k\rvert = 5$	$\lvert\mathcal{S}_k\rvert = 7$	$\lvert\mathcal{S}_k\rvert = 9$	$\lvert\mathcal{S}_k\rvert = 5$	$\lvert\mathcal{S}_k\rvert = 7$	$\lvert\mathcal{S}_k\rvert = 9$
0.1	46.9 (0.2)	45.3 (0.2)	45.8 (0.2)	15.3 (0.2)	12.4 (0.2)	8.8 (0.2)
0.2	49.5 (0.3)	48.9 (0.2)	48.7 (0.2)	17.8 (0.2)	14.3 (0.2)	9.8 (0.2)
0.3	51.0 (0.3)	50.4 (0.3)	50.4 (0.2)	19.3 (0.2)	14.8 (0.2)	10.3 (0.2)
0.4	51.6 (0.2)	51.0 (0.2)	50.2 (0.2)	19.7 (0.2)	14.8 (0.2)	9.8 (0.2)
0.5	52.3 (0.3)	51.3 (0.2)	50.8 (0.2)	20.0 (0.2)	14.6 (0.2)	9.3 (0.2)

assumption is not firmly grounded, either group-Lasso or its sparse version seem to be better options: coop-Lasso only remains a second-best choice when there are less than 10% of sign mismatches within groups.

6. Illustrations on real data. This section illustrates the applicability of the coop-Lasso on two types of predictors, that is, categorical and continuous covariates. The first proposal may be widely applied to ordered categorical variables; the second one is specific to microarray data, but should apply more generally when groups of variables are produced by clustering.

In the first application, each group is formed by a set of variables coding an ordered categorical variable. Ordinal data are often processed either by omitting the order property, treating them as nominal, or by replacing each level with a prescribed value, treating them as quantitative. The latter procedure, combined with generalized linear regression, leads to monotone mapping from levels to responses. Section 6.1 describes how coop-Lasso can bias the estimate toward monotone mappings using a categorical treatment of ordinal variables.

In the second application of Section 6.2, the groups are formed by continuous variables that are redundant noisy measurements (probe signals) pertaining to a common higher-level unobserved variable (gene activity). Sign-coherence is expected here, since each measurement should be positively correlated with the activity of the common unobserved variable. A similar behavior should also be anticipated when groups of variables are formed by a clustering preprocessing step based on the Euclidean distance, such as k-means or average linkage hierarchical clustering [Eisen et al. (1998); Park, Hastie and Tibshirani (2007); Ma, Song and Huang (2007)].

6.1. *Monotonicity of responses to ordinal covariates.* Monotonicity is easily dealt with by transforming ordinal covariates into quantitative variables, but this approach is arbitrary and subject to many criticisms when there is no well-defined numerical difference between levels, which often lacks even for interval data when the lower or the upper interval is not bounded [Gertheiss and Tutz (2009)]. Hence, the categorical treatment is often preferred, even if it fails to fully grasp the order relation.

The Lasso, group-Lasso or fused-Lasso have been applied to the categorical treatment of ordinal features, with the aim to select variables or aggregate adjacent levels [see Gertheiss and Tutz (2010) and references within]. The coop-Lasso is used here to make a stronger usage of the order relationship, by biasing the mapping from levels to the response variable toward monotonic solutions. Note that our proposal does not impose monotonicity and neither does it prescribe an order (although several variations would be possible here). In these respects, we depart from the approaches imposing hard constraints on regression coefficients [Rufibach (2010)].

TABLE 3
Contrasts and codings for comparing the adjacent levels of a covariate with 4 *levels*

Level	**Contrasts**			**Codings**		
0	-1	0	0	$-3/4$	$-1/2$	$-1/4$
1	1	-1	0	$1/4$	$-1/2$	$-1/4$
2	0	1	-1	$1/4$	$1/2$	$-1/4$
3	0	0	1	$1/4$	$1/2$	$3/4$

6.1.1. *Methodology.* When not treated as numerical, ordinal variables are often coded by a set of variables that code differences between levels. Several types of codings have been developed in the ANOVA setting, with relatively little impact in the regression setting, where the so-called dummy codings are intensively used. Indeed, least squares fits are not sensible to coding choices provided there is a one-to-one mapping from one to the other, so that codings only matter regarding the direct interpretation of regression coefficients. However, codings evidently affect the solution in penalized regression, and we will use here specific codings to penalize targeted variations. In order to build a monotonicity-based penalty, we simply use contrasts that compare two adjacent levels. An example of these contrasts is displayed in Table 3, with the corresponding codings, known as backward difference codings, which are simply obtained by solving a linear system [Serlin and Levin (1985)]. Note that several codings are possible for the contrasts given in Table 3. They differ in the definition of a global reference level, whose effect is relegated to the intercept. As we do not penalize the intercept here, the particular choice has no outcome on the solution.

Irrespective of the coding, group penalties act as a selection tool for factors, that is, at variable level [Yuan and Lin (2006)]. On top of this, the sparse group penalty usually presents the ability to discard a level. With difference codings, some increments between adjacent levels may be set to zero, that is, levels may be fused [Gertheiss and Tutz (2010)]. With the coop-Lasso penalty, all increments are urged to be sign-coherent, thereby favoring monotonicity. As a side effect, level fusion may also be obtained.

6.1.2. *Experimental setup.* We illustrate the approach on the Statlog "German Credit" data set [available at the UCI machine learning repository, Frank and Asuncion (2010)], which gathers information about people classified as low or high credit risks. This binary response requires an appropriate model, such as logistic regression. The coop-Lasso fitting algorithm is easily adaptable to generalized linear models, following exactly the structure provided in Algorithm 1, where the appropriate likelihood function replaces the sum of square residuals in Step 1.

All quantitative variables are used for the analysis, but we focus here on the regression coefficients of four variables, encoded as integers or nominal in the Statlog project, which seem better interpreted as ordered nominal, namely: `history`, with 4 levels describing the ability to pay back credits in the past and now; `savings`, with 4 levels giving the balance of the saving account in currency intervals; `employment`, with 5 levels reporting the duration of the present employment in year intervals; and `job`, with 4 levels representing an employment qualification scale. Two other variables, related to the checking account status and property, were also encoded as nominal, but are not described here in full details since they do not show distinct qualitative behaviors between methods. We excluded from the ordinal variables categories merging two subcategories possibly corresponding to different ranks, such as "critical account/other credits existing (not at this bank)" in `history`, or "unknown/no savings account" in `savings`. For simplicity, we suppressed the corresponding examples, thus ending with a total of 330 observations, split into three equal-size learning, validation and test sets. We estimate the logistic regression coefficients on the learning set, perform model selection from deviance or misclassification error on the validation set, and finally keep the test set to estimate prediction performances.

6.1.3. *Results*. The performances of the three group methods are identical, either evaluated in terms of deviance, classification error rate or weighted misclassification (unbalanced misclassification losses are provided with the data set). The regression coefficients differ, however, as shown in Figure 5 displaying the regularization paths for all methods. Recall that we only represent the ordinal covariates `history`, `savings`, `employement` and `job`. Each coefficient represents the increment between two adjacent levels, with positive and negative values resulting in an increase and decrease, respectively. Monotonicity with respect to all levels is reached if all the values corresponding to a factor are nonnegative or nonpositive. We also provide an alternative view of the coop-Lasso path, with the overall effects corresponding to levels, obtained by summing up the increments.

Most factors are not obviously amenable to quantitative coding since there is no natural distance between levels, but we, however, underline that using the usual quantitative transformation with equidistant values followed by linear regression would correspond here to identical increments between levels. Obviously, all displayed solutions radically contradict this linear trend hypothesis.

Our three solutions differ regarding monotonicity, which is almost never observed along the group-Lasso regularization path. The sparse group-Lasso paths have long sign-coherent sections, where group-Lasso infers slight wiggles. These sections extend further with the coop-Lasso. However, as the coop penalty goes to zero, sign-coherence is no longer preserved, and all methods eventually reach the same solution.

The sparse group and the coop-Lasso set some increments to zero, leading to the fusion of adjacent levels that should be welcomed regarding interpretation. The

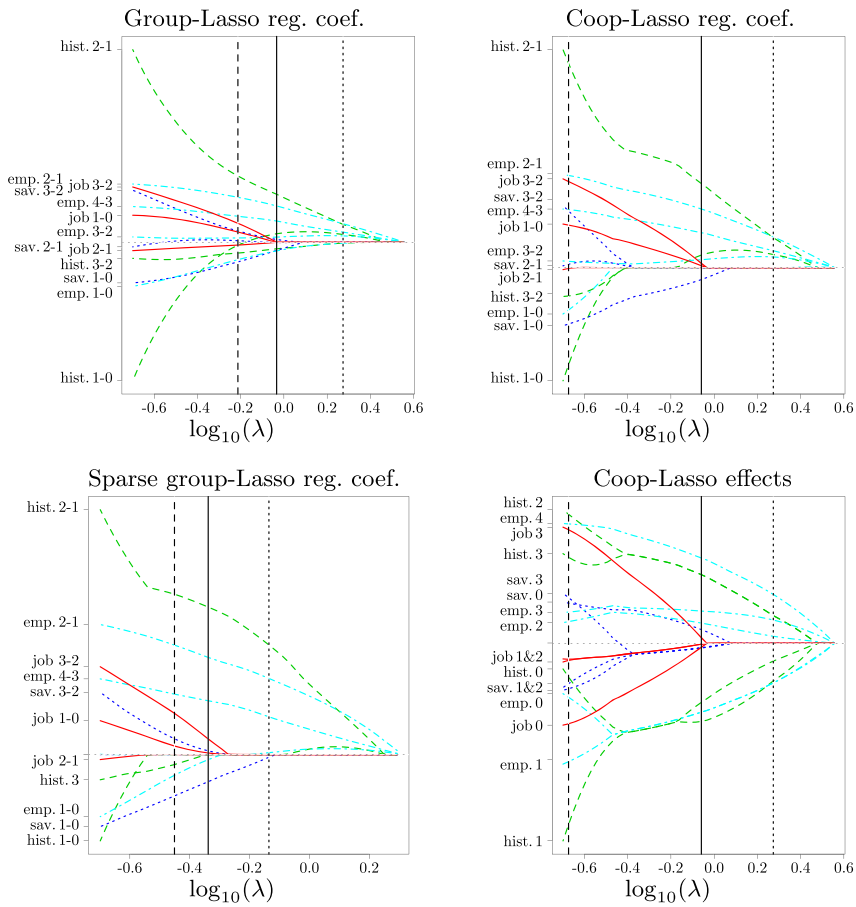

FIG. 5. *Regularization paths for four ordinal covariates* (*history*, *savings*, *job and employment*) *for the group*, *coop*, *and sparse group-Lasso on the contrast coefficients obtained from backward difference coding* (*top left*, *top right and bottom left*, *respectively*). *The transcription of contrasts to levels is also displayed for coop-Lasso* (*bottom right*). *The vertical lines mark the model selected by cross-validation on the validation set*, *for different criteria*: *deviance* (*plain*), *misclassification rate* (*dashed*), *and weighted misclassification error* (*dotted*).

solutions tend to agree on these fusions on long sections of the paths, with some additional fusions of the sparse group-Lasso when slight monotonic solutions are provided by the coop-Lasso (see `employment`, levels 2 and 3, and `savings` levels 1 and 2). These fusions are perceived more directly on the coop-Lasso path of effects, displayed in the bottom right of Figure 5, where the effect of each level is displayed directly.

6.2. *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 data set 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[2] examining the human genome, each probe being related to a unique gene. Following Jeanmougin, Guedj and Ambroise (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.

6.2.1. *Methodology.* The usual processing of microarray data relies 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 gathering all the measurements associated to gene entities at an early stage of 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 nonadaptive 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 increasing 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.

6.2.2. *Experimental setup.* We select the first $d = 200$ most differentiated probes, as identified by the analysis of Jeanmougin, Guedj and Ambroise (2011),

[2]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.

TABLE 4

CV *scores for misclassification error and binomial deviance on the basal tumor data. The minimizer of* CV *for misclassification and deviance are respectively denoted by* λ^{err} *and* λ^{dev}; *the number of selected groups and features respectively refers to genes and probes*

		Probes	**Lasso**	**Group**	**Sparse**	**Coop**
Model selection rule			CV score $\times 100$ (standard error)			
Classification	λ^{err}	10.3 (5.8)	6.9 (4.9)	3.4 (3.5)	3.4 (3.5)	3.4 (3.5)
Deviance	λ^{dev}	76.5 (37.6)	67.2 (32.3)	13.7 (8.1)	20.5 (10.0)	13.8 (7.9)
Model selection rule			# selected groups (features)			
Classification	λ^{err}	17 (17)	16 (17)	11 (15)	14 (21)	9 (11)
Deviance	λ^{dev}	19 (19)	17 (18)	13 (21)	16 (26)	14 (18)

on the 22,269 probes for the $n = 29$ patients with basal tumor. These 200 probes correspond to 172 genes, themselves associated to $p = 381$ probes on the microarray as a whole, with 1 to 13 probes per gene. We clearly enter the high-dimensional setup with $p > 13 \times n$.

All signals are normalized to have a unitary within-class variance. We compare then the Lasso on the $d = 200$ most differentiated probes, with the Lasso and group, sparse group and coop Lasso on the $p = 381$ probes. All fits are produced 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: we evaluate the CV error for each method with the same block partition using either the binomial deviance or the unweighted classification error.

6.2.3. *Results*. The 5-folds CV scores, either based on deviance or misclassification losses, are reported for each estimation method in Table 4, which also displays the number of selected groups and features for the models selected by minimizing the CV score.

Expanding the set of probes from d to p slightly improves the performances of the Lasso, and considerable further progresses are brought by all group methods, which misclassify about 1 patient among the 29 and quarter deviance scores.[3] As expected, less genes are selected by group methods; the difference is more important for the minimizers of the misclassification score, and, among those, for the

[3]A note of caution regarding performances: scores comparisons are fair here, in the sense that the CV scores are optimized with respect to a single parameter λ, whose role is analog for all. Additional simulations (not reported here) show that, for all group methods, the CV error is stable with respect to the random choice of folds and that the CV curves are smooth around their minima. However, the minimizers of CV are biased estimates of out-of-sample scores, and the representativeness of their observed difference can be questioned.

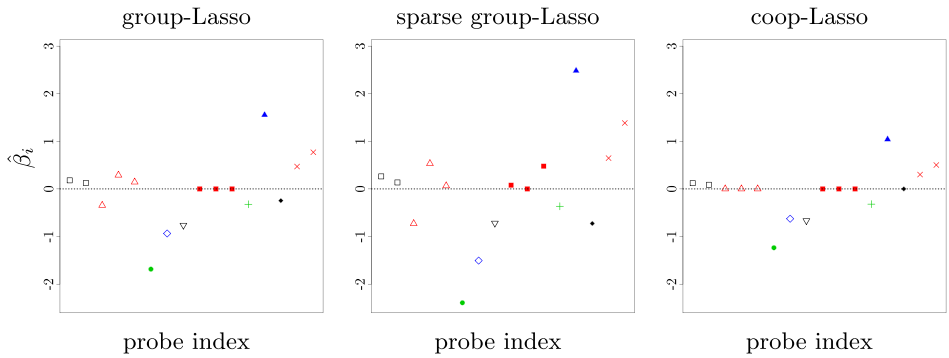

FIG. 6. *Logistic regression coefficients attached to each probe for group, sparse group and coop Lasso. Each marker (color and symbol) designates the gene associated to the probe*: RNPS1 (□), MSH6 (△), PRPS2 (●), H1FX (◇), MFGE8 (▽), SULF1 (■), RNF115 (+), RNF38 (▲), THNSL2 (◆) *and* EDEM3 (×).

group-Lasso and coop-Lasso that comply more stringently to the group structure. These observations indicate that the group structure defined by genes provides truly useful guidelines for inference.

The sparsity numbers differ among the group methods, coop-Lasso selecting as many genes as group-Lasso and fewer probes, and sparse group-Lasso retaining slightly more genes and probes. A more detailed picture is provided in Figure 6, which shows the regression coefficients for the three group estimators adjusted on the whole data set with their respective λ^{err} values. Among the three methods, a total of 15 groups (i.e., genes) are selected. For readability, we only represent the 10 leading groups of regression coefficients (according to their average norm). We first oberve that the magnitude of coefficients differs for each method, the coop-Lasso having the smallest one. In fact, there is a wide range of λ^{err} values for which the miclassification score is minimal for the coop-Lasso, enabling to choose a highly penalized solution without affecting accuracy. The magnitude apart, the group methods have qualitatively the same behaviors for all unitary groups but one, with THNSL2 (◆) being set to zero by the coop-Lasso. The same patterns are observed for two other groups, RNPS1 (□) and EDEM3 (×), whose regression coefficients are consistently estimated to be sign-coherent. Then, SULF1 (■), though being estimated sign-coherent by the sparse group-Lasso, is excluded from the support of the group and coop Lasso. Finally, MSH6 (△), estimated as sign incoherent with the two groups methods, is excluded from the support for the coop-Lasso.

Overall, the probe enrichment scheme we propose here leads to considerable improvements in prediction performance. This better statistical explanation is obtained without impairing interpretability, since sign-coherence is actually often satisfied by all methods and strictly enforced by the coop-Lasso. As often in this type of study, several methods provided similar prediction performances, but the explanation provided by the coop-Lasso is simpler, both from a statistical and from

a biological viewpoint. Note that the coefficient paths (not shown) diverge early between the group and coop methods, so that the above-mentioned discrepancies are not simply due to model selection issues. As a final remark, we observed qualitatively similar behaviors when the initial number of probes d ranged from 10 to 2000. For $d \leq 1000$, the group methods always performed best, with approximately identical classification errors, the group-Lasso and coop-Lasso slightly dominating the sparse group-Lasso in terms of deviance. With larger initial sets of probes, the enrichment procedure becomes less efficient, and all methods provide similar decaying results. The chosen setup displayed here, with $d = 200$, leads to the smallest classification error for all methods, and was chosen for being representative of the most interesting regime.

7. Discussion. The *coop-Lasso* is a variant of the *group-Lasso* that was originally proposed in the context of multi-task learning, for inferring related networks with Gaussian Graphical Models [Chiquet, Grandvalet and Ambroise (2011)]. Here we develop its analysis in the linear regression setup and demonstrate its value for prediction and inference with generalized linear models. Along with 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 and sparse group-Lasso [Friedman, Hastie and Tibshirani (2010)] by the assumption that the group structure is *sign-coherent*, namely, that groups gather either nonpositive, nonnegative or null parameters, enabling the recovery of various within-group sign patterns (positive, negative, null, nonpositive, nonnegative, nonnull). This flexibility greatly reduces the incentive to drive within-group sparsity with an additional parameter that later leads to an unwieldy model selection step. However, the relevance of the sign-coherence assumption should be firmly established since it plays an essential role in the performance of coop-Lasso compared to the sparse group-Lasso.

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 the 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.

We illustrate the merits of the coop-Lasso applied to the analysis to ordinal and continuous predictors. With an apposite coding, such as forward or backward difference coding, the sign-coherence assumption is transcribed in a monotonicity

assumption, which does not require to stipulate the usual and controversial mapping from levels to quantitative variables. 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.

APPENDIX: PROOFS

A.1. Proof of Lemma 1. Let us use $\mathcal{T}_k$ as a shorthand for $\mathcal{S}_k(\boldsymbol{\beta})$, Chiquet, Grandvalet and Ambroise (2011) show that the subdifferential $\boldsymbol{\theta}$ obey the following conditions:

$$\max(\|\boldsymbol{\theta}^+_{\mathcal{G}_k}\|, \|\boldsymbol{\theta}^-_{\mathcal{G}_k}\|) \le w_k \qquad \text{if } \boldsymbol{\beta}_{\mathcal{G}_k} = \mathbf{0}, \tag{26a}$$

$$\boldsymbol{\theta}_{\mathcal{T}_k} = \frac{w_k \boldsymbol{\beta}_{\mathcal{T}_k}}{\|\boldsymbol{\beta}_{\mathcal{T}_k}\|}, \qquad \|\boldsymbol{\theta}^-_{\mathcal{T}_k^c}\| \le w_k, \qquad \|\boldsymbol{\theta}^+_{\mathcal{T}_k^c}\| = 0 \qquad \text{if } \|\boldsymbol{\beta}^+_{\mathcal{G}_k}\| > 0, \|\boldsymbol{\beta}^-_{\mathcal{G}_k}\| = 0, \tag{26b}$$

$$\boldsymbol{\theta}_{\mathcal{T}_k} = \frac{w_k \boldsymbol{\beta}_{\mathcal{T}_k}}{\|\boldsymbol{\beta}_{\mathcal{T}_k}\|}, \qquad \|\boldsymbol{\theta}^+_{\mathcal{T}_k^c}\| \le w_k, \qquad \|\boldsymbol{\theta}^-_{\mathcal{T}_k^c}\| = 0 \qquad \text{if } \|\boldsymbol{\beta}^-_{\mathcal{G}_k}\| > 0, \|\boldsymbol{\beta}^+_{\mathcal{G}_k}\| = 0, \tag{26c}$$

$$\forall j \in \mathcal{G}_k \qquad \boldsymbol{\theta}_j = w_k \boldsymbol{\beta}_j \|\operatorname{sign}(\beta_j)\boldsymbol{\beta}\|^{-1} \qquad \text{if } \|\boldsymbol{\beta}^-_{\mathcal{G}_k}\| > 0, \|\boldsymbol{\beta}^+_{\mathcal{G}_k}\| > 0. \tag{26d}$$

We thus simply have to prove the equivalence of conditions (8) and (26) for all $\boldsymbol{\beta}_{\mathcal{G}_k}$ values.

For $\boldsymbol{\beta}_{\mathcal{G}_k} = \mathbf{0}$, (8) reads

$$\|\boldsymbol{\theta}^+_{\mathcal{G}_k}\| \le w_k \quad \text{and} \quad \|\boldsymbol{\theta}^-_{\mathcal{G}_k}\| \le w_k, \tag{27}$$

which is equivalent to (26a).

For $\boldsymbol{\beta}_{\mathcal{G}_k} \neq \mathbf{0}$, the equalities for $\boldsymbol{\theta}_{\mathcal{T}_k}$ in (26b)–(26d) are equivalent to (8a), thus setting the equivalence between (8) and (26) for all nonzero coefficients. For $\boldsymbol{\beta}_{\mathcal{T}_k^c}$, let us consider the case (26b), where all nonzero parameters within group k are positive. The first equation of (26b) implies that $\|\boldsymbol{\theta}^+_{\mathcal{T}_k}\| = w_k$ and $\|\boldsymbol{\theta}^-_{\mathcal{T}_k}\| = 0$. Hence, $\|\boldsymbol{\theta}^-_{\mathcal{T}_k^c}\| \le w_k$ and $\|\boldsymbol{\theta}^+_{\mathcal{T}_k^c}\| = 0$ imply (27), so that (26b) implies (8). The contraposition is also easy to check. From (8a), when all coefficients are positive, we have that $\|\boldsymbol{\theta}^-_{\mathcal{T}_k}\| = 0$ and $\|\boldsymbol{\theta}^+_{\mathcal{T}_k}\| = w_k$. Then, this implies that (8b) reads

$$\|\boldsymbol{\theta}^-_{\mathcal{T}_k^c}\| \le w_k \quad \text{and} \quad \|\boldsymbol{\theta}^+_{\mathcal{T}_k^c}\| = 0,$$

which defines $\boldsymbol{\theta}_{\mathcal{T}_k^c}$ in (26b). The proof is similar for (26c) where all nonzero parameters within group k are positive.

A.2. Proof of Proposition 1. We assume here that $\mathbf{X}^\intercal\mathbf{X} = \mathbf{I}_p$. We introduce the ridge estimator in the computation of the trace in equation (22), through the chain rule, yielding an unbiased estimate of df:

$$\begin{aligned}\widetilde{\mathrm{df}}_{\mathrm{coop}}(\lambda) &= \mathrm{tr}\left(\frac{\partial \hat{\mathbf{y}}(\lambda)}{\partial \mathbf{y}}\right) = \mathrm{tr}\left(\frac{\partial \mathbf{X}^\intercal \hat{\boldsymbol{\beta}}^{\mathrm{coop}}(\lambda)}{\partial \hat{\boldsymbol{\beta}}^{\mathrm{ridge}}(\gamma)} \frac{\partial \hat{\boldsymbol{\beta}}^{\mathrm{ridge}}(\gamma)}{\partial \mathbf{y}}\right) \\ &= \frac{1}{1+\gamma} \sum_{k=1}^{K} \sum_{j \in \mathcal{G}_k} \frac{\partial \hat{\beta}_j^{\mathrm{coop}}(\lambda)}{\partial \hat{\beta}_j^{\mathrm{ridge}}(\gamma)},\end{aligned}$$

where the last equation derives from the definition (24) of the ridge estimator with regularization parameter γ. Then, the expression of the coop-Lasso as a function of the ridge regression estimate is simply obtained from equation (10), using that, in the orthonormal case, we have $\hat{\boldsymbol{\beta}}^{\mathrm{ols}} = (1+\gamma)\hat{\boldsymbol{\beta}}^{\mathrm{ridge}}(\gamma)$. Dropping the reference to λ and γ that is obvious from the context, we have, $\forall k \in \{1, \dots, K\}$ and $\forall j \in \mathcal{G}_k$,

$$\hat{\beta}_j^{\mathrm{coop}} = \left(1 - \frac{\lambda w_k}{(1+\gamma)\|\boldsymbol{\varphi}_j(\hat{\boldsymbol{\beta}}_{\mathcal{G}_k}^{\mathrm{ridge}})\|}\right)^+ (1+\gamma)\hat{\beta}_j^{\mathrm{ridge}}. \tag{28}$$

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

$$\begin{aligned}\frac{1}{1+\gamma} \frac{\partial \hat{\beta}_j^{\mathrm{coop}}}{\partial \hat{\beta}_j^{\mathrm{ridge}}} &= \mathbb{1}(\|\hat{\beta}_j^{\mathrm{coop}}\| > 0) \\ &\quad \times \left(1 - \frac{\lambda w_k}{(1+\gamma)}\left(\frac{1}{\|\boldsymbol{\varphi}_j(\hat{\boldsymbol{\beta}}_{\mathcal{G}_k}^{\mathrm{ridge}})\|} - \frac{(\hat{\beta}_j^{\mathrm{ridge}})^2}{\|\boldsymbol{\varphi}_j(\hat{\boldsymbol{\beta}}_{\mathcal{G}_k}^{\mathrm{ridge}})\|^3}\right)\right).\end{aligned}$$

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

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

From (28), we have, $\forall k \in \{1, \dots, K\}$ and $\forall j \in \mathcal{G}_k$,

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

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 (18). 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:

(1) restrict the estimation problem to the true support;
(2) complete this estimate by 0 outside the true support;
(3) 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 irrepresentable conditions (A4)–(A5) 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 $\lambda_n \to 0$. Lemma 3 provides the basis for the inequalities (16) and (17) that express our irrepresentable conditions.

LEMMA 2. *Assuming* (A1)–(A3), *let* $\tilde{\boldsymbol{\beta}}^n_{\mathcal{S}}$ *be the unique minimizer of the regression problem restricted to the true support* $\mathcal{S}$:

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

where $\|\cdot\|_n = \|\cdot\|/n$ *denotes the empirical norm.*

If $\lambda_n \to 0$, *then* $\tilde{\boldsymbol{\beta}}^n_{\mathcal{S}} \xrightarrow{P} \boldsymbol{\beta}^\star_{\mathcal{S}}$.

PROOF. This lemma stems from standard results of M-estimation [van der Vaart (1998)]. Let $\boldsymbol{\varepsilon} = \mathbf{y} - \mathbf{X}\boldsymbol{\beta}^\star$, and write $\boldsymbol{\Psi}^n = \mathbf{X}^\intercal\mathbf{X}/n$. If $\lambda_n \to 0$, then under

(A1)–(A2), for any $\mathbf{v} \in \mathbb{R}^{|\mathcal{S}|}$

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

tends in probability to

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

It follows from the strict convexity of Z_n that $\arg\min Z_n(\mathbf{v}) \xrightarrow{P} \arg\min Z(\mathbf{v}) = \boldsymbol{\beta}_{\mathcal{S}}^\star$ [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 μ the limit S is bounded away from* 1:

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

Then,

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

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

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

Convergence in probability of S_n to S concludes the proof:

$$\mathbb{P}\big(\mu(S_n - S) \leq \delta\big) \to 1 \qquad \text{therefore } \mathbb{P}\big(\mu(S_n) \leq 1\big) \to 1. \qquad \square$$

Let us consider the full vector $\tilde{\boldsymbol{\beta}}^n$ with coefficients $\tilde{\boldsymbol{\beta}}_{\mathcal{S}}^n$ defined as in Lemma 2 and other coefficients null, $\tilde{\boldsymbol{\beta}}_{\mathcal{S}^c}^n = \mathbf{0}$. We now proceed to the last step of the proof of Theorem 2, by proving that $\tilde{\boldsymbol{\beta}}^n$ satisfies the coop-Lasso optimality conditions with probability tending to 1 under the additional conditions (A4)–(A5). The final conclusion then results from the uniqueness of the coop-Lasso estimator.

First, consider optimality conditions with respect to $\boldsymbol{\beta}_{\mathcal{S}}$. As a result of Lemma 2, the probability that $\tilde{\boldsymbol{\beta}}_j^n \neq 0$ for every $j \in \mathcal{S}$ tends to 1. Thereby, $\tilde{\boldsymbol{\beta}}_{\mathcal{S}}^n$ satisfies (9a) on the restriction of $\mathbf{X}$ to covariates in $\mathcal{S}$ with probability tending to 1. As $\tilde{\boldsymbol{\beta}}_{\mathcal{S}^c}^n = \mathbf{0}$, then $\mathbf{X}\tilde{\boldsymbol{\beta}}^n = \mathbf{X}_{\cdot\mathcal{S}}\tilde{\boldsymbol{\beta}}_{\mathcal{S}}^n$ and for every $j \in \mathcal{S}$, $\|\boldsymbol{\varphi}_j(\tilde{\boldsymbol{\beta}}_{\mathcal{S}_k})^n\| = \|\boldsymbol{\varphi}_j(\tilde{\boldsymbol{\beta}}_{\mathcal{G}_k}^n)\|$, therefore, $\tilde{\boldsymbol{\beta}}_{\mathcal{S}}^n$ satisfies (9a) in the original problem with probability tending to 1.

Second, $\tilde{\boldsymbol{\beta}}_{\mathcal{S}^c}^n$ should also verify the optimality conditions (9) 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

$$\mathbb{P}\big(\max(\|((\mathbf{X}_{.\mathcal{S}_k^c})^\intercal(\mathbf{X}\tilde{\boldsymbol{\beta}}^n-\mathbf{y}))^+\|_n, \|((\mathbf{X}_{.\mathcal{S}_k^c})^\intercal(\mathbf{X}\tilde{\boldsymbol{\beta}}^n-\mathbf{y}))^-\|_n) \le \lambda_n w_k\big) \to 1; \tag{29}$$

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

$$\mathbb{P}\big(\{\nu_k(\mathbf{X}_{.\mathcal{S}_k^c})^\intercal(\mathbf{X}\tilde{\boldsymbol{\beta}}^n-\mathbf{y}) \succeq \mathbf{0}\} \cap \{\|(\mathbf{X}_{.\mathcal{S}_k^c})^\intercal(\mathbf{X}\tilde{\boldsymbol{\beta}}^n-\mathbf{y})\|_n \le \lambda_n w_k\}\big) \to 1. \tag{30}$$

To prove (29) and (30), we study the asymptotics of $(\mathbf{X}_{.\mathcal{S}_k^c})^\intercal(\mathbf{X}\tilde{\boldsymbol{\beta}}^n-\mathbf{y})/n$ for any group such that $\mathcal{S}_k^c$ 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

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

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

$$\begin{aligned}\frac{1}{n}(\mathbf{X}_{.\mathcal{S}_k^c})^\intercal(\mathbf{X}\tilde{\boldsymbol{\beta}}^n-\mathbf{y}) &= \frac{1}{n}(\mathbf{X}_{.\mathcal{S}_k^c})^\intercal\mathbf{X}(\tilde{\boldsymbol{\beta}}^n-\boldsymbol{\beta}^\star) - \frac{1}{n}(\mathbf{X}_{.\mathcal{S}_k^c})^\intercal\boldsymbol{\varepsilon}\\ &= \frac{1}{n}(\mathbf{X}_{.\mathcal{S}_k^c})^\intercal\mathbf{X}_{.\mathcal{S}}(\tilde{\boldsymbol{\beta}}^n_{\mathcal{S}}-\boldsymbol{\beta}^\star_{\mathcal{S}}) + O_P(n^{-1/2})\\ &= \boldsymbol{\Psi}_{\mathcal{S}_k^c\mathcal{S}}(\tilde{\boldsymbol{\beta}}^n_{\mathcal{S}}-\boldsymbol{\beta}^\star_{\mathcal{S}}) + O_P(n^{-1/2}),\end{aligned} \tag{31}$$

while the combination of (31) and optimality conditions (9a) on $\tilde{\boldsymbol{\beta}}^n_{\mathcal{S}}$ leads to

$$\boldsymbol{\Psi}_{\mathcal{S}\mathcal{S}}(\tilde{\boldsymbol{\beta}}^n_{\mathcal{S}}-\boldsymbol{\beta}^\star_{\mathcal{S}}) = -\lambda_n\mathbf{D}(\tilde{\boldsymbol{\beta}}^n_{\mathcal{S}})\tilde{\boldsymbol{\beta}}^n_{\mathcal{S}} + O_P(n^{-1/2}), \tag{32}$$

where $\mathbf{D}(\cdot)$ is the weighting matrix (15). Put (31) and (32) together to finally obtain

$$\frac{1}{n}(\mathbf{X}_{.\mathcal{S}_k^c})^\intercal(\mathbf{X}\tilde{\boldsymbol{\beta}}^n-\mathbf{y}) = -\lambda_n\boldsymbol{\Psi}_{\mathcal{S}_k^c\mathcal{S}}\boldsymbol{\Psi}_{\mathcal{S}\mathcal{S}}^{-1}\mathbf{D}(\tilde{\boldsymbol{\beta}}^n_{\mathcal{S}})\tilde{\boldsymbol{\beta}}^n_{\mathcal{S}} + O_P(n^{-1/2}). \tag{33}$$

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

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

Limits (29) and (30) are expressed:

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

$$\mathbb{P}\big(\max(\|R_{k,n}^+\|, \|R_{k,n}^-\|) \le 1\big) \to 1;$$

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

$$\mathbb{P}\big(\{\nu_k R_{k,n} \succeq \mathbf{0}\} \cap \{\|(\nu_k R_{k,n})^+\| \le 1\}\big) \to 1.$$

Remark that, as a continuous function of $\tilde{\boldsymbol{\beta}}^n_{\mathcal{S}}$, $\mathbf{D}(\tilde{\boldsymbol{\beta}}^n_{\mathcal{S}})\tilde{\boldsymbol{\beta}}^n_{\mathcal{S}}$ converges in probability to $\mathbf{D}(\boldsymbol{\beta}^\star_{\mathcal{S}})\boldsymbol{\beta}^\star_{\mathcal{S}}$. Therefore, with a decrease rate for λ_n chosen such that $n^{1/2}\lambda_n \to \infty$, equation (33) implies

$$R_{k,n} \xrightarrow{P} R_k. \tag{34}$$

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

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

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

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

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

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

$$\begin{aligned}
&\mathbb{P}\big(\{\|(\nu_k R_{k,n})^+\| \le 1\} \cap \{\nu_k R_{k,n} \succeq \mathbf{0}\}\big)\\
&\quad = 1 - \mathbb{P}\big(\{\|(\nu_k R_{k,n})^+\| > 1\} \cup \{\nu_k R_{k,n} \prec \mathbf{0}\}\big)\\
&\quad \ge 1 - \mathbb{P}\big(\|(\nu_k R_{k,n})^+\| > 1\big) - \mathbb{P}(\nu_k R_{k,n} \prec \mathbf{0})\\
&\quad \ge 1 - \mathbb{P}\big(\max(\|R_{k,n}^+\|, \|R_{k,n}^-\|) > 1\big) - \mathbb{P}(\nu_k R_{k,n} \prec \mathbf{0}).
\end{aligned}$$

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, $P(A_{k,n}) \to 1$. This implies that

$$\mathbb{P}\Big(\bigcup_{k:\,\mathcal{S}_k^c \ne \varnothing} A_{k,n}^c\Big) \le \sum_{k:\,\mathcal{S}_k^c \ne \varnothing} \mathbb{P}(A_{k,n}^c) \to 0,$$

which in turn concludes the proof:

$$\mathbb{P}\Big(\bigcap_{k:\,\mathcal{S}_k^c \ne \varnothing} A_{k,n}\Big) \to 1.$$

Acknowledgments. We would like to thank Marine 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.

REFERENCES

BACH, F. R. (2008). Consistency of the group lasso and multiple kernel learning. *J. Mach. Learn. Res.* **9** 1179–1225. MR2417268

BAKIN, S. (1999). Adaptive regression and model selection in data mining problems. Ph.D. thesis, Australian National Univ., Canberra.

BECK, A. and TEBOULLE, M. (2009). A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Imaging Sci.* **2** 183–202. MR2486527

BREIMAN, L. (1995). Better subset regression using the nonnegative garrote. *Technometrics* **37** 373–384. MR1365720

BREIMAN, L. (1996). Heuristics of instability and stabilization in model selection. *Ann. Statist.* **24** 2350–2383. MR1425957

BREIMAN, L., FRIEDMAN, J. H., OLSHEN, R. and STONE, C. J. (1984). *Classification and Regression Trees*. Wadsworth, Belmont, CA.

CHIQUET, J., GRANDVALET, Y. and AMBROISE, C. (2011). Inferring multiple graphical structures. *Statistic and Computing* **21** 537–553.

EFRON, B. (2004). The estimation of prediction error: Covariance penalties and cross-validation. *J. Amer. Statist. Assoc.* **99** 619–642. MR2090899

EISEN, M. B., SPELLMAN, P. T., BROWN, P. O. and BOTSTEIN, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA* **95** 14863–14868.

FOYGEL, R. and DRTON, M. (2010). Exact block-wise optimization in group lasso for linear regression. Technical report. Available at arXiv:1010.3320.

FRANK, A. and ASUNCION, A. (2010). UCI machine learning repository.

FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2010). A note on the group Lasso and a sparse group Lasso. Technical report. Available at arXiv:1001.0736.

GERTHEISS, J. and TUTZ, G. (2009). Penalized regression with ordinal predictors. *International Statistical Review* **77** 345–365.

GERTHEISS, J. and TUTZ, G. (2010). Sparse modeling of categorial explanatory variables. *Ann. Appl. Stat.* **4** 2150–2180. MR2829951

GRANDVALET, Y. and CANU, S. (1999). Outcomes of the equivalence of adaptive ridge with least absolute shrinkage. In *Advances in Neural Information Processing Systems* 11 (*NIPS* 1998) 445–451.

HESS, K. R., ANDERSON, K., SYMMANS, W. F., VALERO, V., IBRAHIM, N., MEJIA, J. A., BOOSER, D., THERIAULT, R. L., BUZDAR, U., DEMPSEY, P. J., ROUZIER, R., SNEIGE, N., ROSS, J. S., VIDAURRE, T., GÓMEZ, H. L., HORTOBAGYI, G. N. and PUSTZAI, L. (2006). Pharmacogenomic predictor of sensitivity to preoperative chemotherapy with Paclitaxel and Fluorouracil, Doxorubicin, and Cyclophosphamide in breast cancer. *Journal of Clinical Oncology* **24** 4236–4244.

HESTERBERG, T., CHOI, N. H., MEIER, L. and FRALEY, C. (2008). Least angle and l_1 penalized regression: A review. *Stat. Surv.* **2** 61–93. MR2520981

HUANG, J. and ZHANG, T. (2010). The benefit of group sparsity. *Ann. Statist.* **38** 1978–2004. MR2676881

JEANMOUGIN, M., GUEDJ, M. and AMBROISE, C. (2011). Defining a robust biological prior from pathway analysis to drive network inference. *J. SFdS* **152** 97–110. MR2821224

KNIGHT, K. and FU, W. (2000). Asymptotics for lasso-type estimators. *Ann. Statist.* **28** 1356–1378. MR1805787

MA, S., SONG, X. and HUANG, J. (2007). Supervised group Lasso with applications to microarray data analysis. *BMC Bioinformatics* **8** 60.

MEIER, L., VAN DE GEER, S. and BÜHLMANN, P. (2008). The group Lasso for logistic regression. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **70** 53–71. MR2412631

NARDI, Y. and RINALDO, A. (2008). On the asymptotic properties of the group lasso estimator for linear models. *Electron. J. Stat.* **2** 605–633. MR2426104

OSBORNE, M. R., PRESNELL, B. and TURLACH, B. A. (2000). On the LASSO and its dual. *J. Comput. Graph. Statist.* **9** 319–337. MR1822089

PARK, M. Y., HASTIE, T. and TIBSHIRANI, R. (2007). Averaged gene expressions for regression. *Biostatistics* **8** 212–227.

ROTH, V. and FISCHER, B. (2008). The group-Lasso for generalized linear models: Uniqueness of solutions and efficient algorithms. In *ICML*'08: *Proceedings of the* 25*th International Conference on Machine Learning* 848–855.

RUFIBACH, K. (2010). An active set algorithm to estimate parameters in generalized linear models with ordered predictors. *Comput. Statist. Data Anal.* **54** 1442–1456. MR2600846

SERLIN, R. C. and LEVIN, J. R. (1985). Teaching how to derive directly interpretable coding schemes for multiple regression analysis. *Journal of Educational Statistics* **10** 223–238.

STEIN, C. M. (1981). Estimation of the mean of a multivariate normal distribution. *Ann. Statist.* **9** 1135–1151. MR0630098

TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267–288. MR1379242

VAN DER VAART, A. W. (1998). *Asymptotic Statistics. Cambridge Series in Statistical and Probabilistic Mathematics* **3**. Cambridge Univ. Press, Cambridge. MR1652247

YUAN, M. and LIN, Y. (2006). Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **68** 49–67. MR2212574

YUAN, M. and LIN, Y. (2007). On the non-negative garrote estimator. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **69** 143–161. MR2325269

ZHAO, P., ROCHA, G. and YU, B. (2009). The composite absolute penalties family for grouped and hierarchical variable selection. *Ann. Statist.* **37** 3468–3497. MR2549566

ZHOU, H., SEHL, M. E., SINSHEIMER, J. S. and LANGE, K. (2010). Association screening of common and rare genetic variants by penalized regression. *Bioinformatics* **26** 2375–2382.

ZOU, H. and HASTIE, T. (2005). Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **67** 301–320. MR2137327

ZOU, H., HASTIE, T. and TIBSHIRANI, R. (2007). On the "degrees of freedom" of the lasso. *Ann. Statist.* **35** 2173–2192. MR2363967

LABORATOIRE STATISTIQUE ET GÉNOME
23, BOULEVARD DE FRANCE
91037 ÉVRY
FRANCE
E-MAIL: julien.chiquet@genopole.cnrs.fr
yves.grandvalet@utc.fr
camille.charbonnier@genopole.cnrs.fr
URL: http://stat.genopole.cnrs.fr