

Group Lasso with Overlaps: the Latent Group Lasso approach

Guillaume Obozinski*

GUILLAUME.OBOZINSKI@ENS.FR

*Sierra team - INRIA
Ecole Normale Supérieure
(INRIA/ENS/CNRS UMR 8548)
Paris, France*

Laurent Jacob*

LAURENT@STAT.BERKELEY.EDU

*Department of Statistics
University of California
Berkeley CA 94720, USA*

Jean-Philippe Vert

JEAN-PHILIPPE.VERT@MINES.ORG

*Centre for Computational Biology
Mines ParisTech
Fontainebleau, F-77300, France
INSERM U900
Institut Curie
Paris, F-75005, France*

Abstract

We study a norm for structured sparsity which leads to sparse linear predictors whose supports are unions of predefined overlapping groups of variables. We call the obtained formulation *latent group Lasso*, since it is based on applying the usual group Lasso penalty on a set of latent variables. A detailed analysis of the norm and its properties is presented and we characterize conditions under which the set of groups associated with latent variables are correctly identified. We motivate and discuss the delicate choice of weights associated to each group, and illustrate this approach on simulated data and on the problem of breast cancer prognosis from gene expression data.

Keywords: group Lasso, sparsity, graph, support recovery, block regularization, feature selection

1. Introduction

Sparsity has triggered much research in statistics, machine learning and signal processing recently. Sparse models are attractive in many application domains because they lend themselves particularly well to interpretation and data compression. Moreover, from a statistical viewpoint, betting on sparsity is a way to reduce the complexity of inference tasks in large dimensions with limited amounts of observations. While sparse models have traditionally been estimated with greedy feature selection approaches, more recent formulations as optimization problems involving a non-differentiable convex penalty have proven very successful both theoretically and practically. The canonical example is the penalization of a least-square criterion by the ℓ_1 norm of the estimator, known as *Lasso* in statistics (Tibshirani, 1996) or *basis pursuit* in signal processing (Chen et al., 1998). Under appropriate

*. Equal contribution

assumptions, the Lasso can be shown to recover the exact support of a sparse model from data generated by this model if the covariates are not too correlated (Wainwright, 2009; Zhao and Yu, 2006). It is consistent even in high dimensions, with fast rates of convergence (Bickel et al., 2009; Lounici, 2008). We refer the reader to van de Geer (2010) for a detailed review.

While the ℓ_1 norm penalty leads to sparse models, it does not encode any prior information about the structure of the sets of covariates that one may wish to see selected jointly, such as predefined groups of covariates. An extension of the Lasso for the selection of variables in groups was proposed under the name group Lasso by Yuan and Lin (2006), who considered the case where the groups form a partition of the sets of variables. The group Lasso penalty, also called ℓ_1/ℓ_2 penalty, is defined as the sum (i.e., ℓ_1 norm) of the ℓ_2 norms of the restrictions of the parameter vector of the model to the different groups of covariates. The work of several authors shows that when the support can be encoded well by the groups defining the norm, support recovery and estimation are improved (Huang and Zhang, 2010; Kolar et al., 2011; Lounici et al., 2010, 2009; Negahban and Wainwright, 2011; Obozinski et al., 2010).

Subsequently, the notion of *structured sparsity* emerged as a natural generalization of the selection in groups, where the support of the model one wishes to recover is not anymore required to be just sparse but also to display certain structure. One of the first natural approaches to *structured sparsity* has been to consider extensions of the ℓ_1/ℓ_2 penalty to situations in which the set of groups considered overlap, so that the possible support pattern exhibit some structure (Bach, 2009; Zhao et al., 2009). Jenatton et al. (2011) formalized this approach and proposed an ℓ_1/ℓ_2 norm construction for families of allowed supports stable by *intersection*. Other approaches to structured sparsity are quite diverse: Bayesian or non-convex approaches that directly exploit the recursive structure of some sparsity patterns such as trees (Baraniuk et al., 2010; He and Carin, 2009), greedy approaches based on *block-coding* (Huang et al., 2009), relaxation of submodular penalties (Bach, 2010), generic variational formulations (Micchelli et al., 2011).

While Jenatton et al. (2011) proposed a norm inducing supports that arise as *intersections* of a sub-collection of groups defining the norm, we consider in this work norms which, albeit defined as well by a collection of overlapping groups, induce supports that are rather *unions* of a sub-collection of the groups encoding prior information. The main idea is that instead of directly applying the ℓ_1/ℓ_2 norm to a vector, we apply it to a set of latent variables each supported by one of the groups, which are combined linearly to form the estimated parameter vector. In the regression case, we therefore call our approach *latent group Lasso*.

The corresponding decomposition of a parameter vector into latent variables calls for the notion of *group-support*, which we introduce and which corresponds to the set of non-zero latent variables. In the context of a learning problem regularized by the norm we propose, we study the problem of *group-support recovery*, a notion stronger than the classical support recovery. Group-support recovery typically implies support recovery (although not always) if the support of a parameter vector is exactly a union of groups. We provide sufficient conditions for consistent group-support recovery.

In the definition of our norm, a weight is associated with each group. These weights play a much more important role in the case of overlapping groups than in the case of disjoint

groups, since in the former case they determine the set of recoverable supports and the complexity of the class of possible models. We discuss the delicate question of the choice of these weights.

While the norm we consider is quite general and has potentially many applications, we illustrate its potential on the particular problem of learning sparse predictive models for cancer prognosis from high-dimensional gene expression data. The problem of identifying a predictive molecular signature made of a small set of genes is often ill-posed and so noisy that exact variable selection may be elusive. We propose that, instead, selecting genes in groups that are involved in the same biological process or connected in a functional or interaction network could be performed more reliably, and potentially lead to better predictive models. We empirically explore this application, after extensive experiments on simulated data illustrating some of the properties of our norm.

To summarize, the main contributions of this paper, which rephrases and extends a preliminary version published in [Jacob et al. \(2009\)](#), are the following:

- We define the latent group Lasso penalty to infer sparse models with unions of predefined groups as supports, and analyze in details some of its mathematical properties.
- We introduce the notion of *group-support* and group-support recovery results. Using correspondence theory, we show under appropriate conditions, that, in a classical asymptotic setting, estimators for the linear regression regularized with $\Omega_{\mathcal{G}}^{\mathcal{L}}$ are consistent for the estimation of a sufficiently sparse *group-support*.
- We discuss in length the choice of weights associated to each group, which play a crucial role in the presence of overlapping groups of different sizes.
- We provide extended experimental results both on simulated data — addressing support-recovery, estimation error and role of weights — and on breast cancer data, using biological pathways and genes networks as prior information to construct latent group Lasso formulations.

The rest of the paper is structured as follows. We first introduce the latent group Lasso penalty and position it in the context of related work in Section 3. In Section 4 we show that it is a norm and provide several characterizations and variational formulations; we also show that regularizing with this norm is equivalent to covariate duplication (Section 4.6) and derive a corresponding multiple kernel learning formulation (Section 4.7). We briefly discuss algorithms in Section 4.8. In Section 5, we introduce the notion of *group-support* and consider in Section 6 a few toy examples to illustrate the concepts and properties discussed so far. We study group support-consistency in Section 7. The difficult question of the choice of the weighting scheme is discussed in Section 8. Section 9 presents the latent graph Lasso, a variant of the latent group Lasso when covariates are organized into a graph. Finally, in Section 10, we present several experiments: first, on artificial data to illustrate the gain in support recovery and estimation over the classical Lasso, as well as the influence of the choice of the weights; second, on the real problem of breast cancer prognosis from gene expression data.

2. Notations

In this section we introduce notations that will be used throughout the article. For any vector $\mathbf{w} \in \mathbb{R}^p$ and any $q \geq 1$, $\|\mathbf{w}\|_q = (\sum_{i=1}^p |\mathbf{w}_i|^q)^{1/q}$ denotes the ℓ_q norm of \mathbf{w} . We simply use the notation $\|\mathbf{w}\| = \|\mathbf{w}\|_2$ for the Euclidean norm. $\text{supp}(\mathbf{w}) \subset [1, p]$ denotes the support of \mathbf{w} , *i.e.*, the set of covariates $i \in [1, p]$ such that $\mathbf{w}_i \neq 0$. A *group* of covariates is a subset $g \subset [1, p]$. The set of all possible groups is therefore $\mathcal{P}([1, p])$, the power set of $[1, p]$. For any group g , $g^c = [1, p] \setminus g$ denotes the complement of g in $[1, p]$, *i.e.*, the covariates which are not in g . $\Pi_g : \mathbb{R}^p \rightarrow \mathbb{R}^p$ denotes the projection onto $\{\mathbf{w} : \mathbf{w}_i = 0 \text{ for } i \in g^c\}$, *i.e.*, $\Pi_g \mathbf{w}$ is the vector whose entries are the same as \mathbf{w} for the covariates in g , and are 0 for other other covariates. We will usually use the notation $\mathbf{w}_g \triangleq \Pi_g \mathbf{w}$. We say that two groups *overlap* if they have at least one covariate in common.

Throughout the article, $\mathcal{G} \subset \mathcal{P}([1, p])$ denotes a set of groups, usually fixed in advance for each application, and we denote $m \triangleq |\mathcal{G}|$ the number of groups in \mathcal{G} . We require that all covariates belong to at least one group, *i.e.*,

$$\bigcup_{g \in \mathcal{G}} g = [1, p] .$$

We note $\mathcal{V}_{\mathcal{G}} \subset \mathbb{R}^{p \times \mathcal{G}}$ the set of m -tuples of vectors $\bar{\mathbf{v}} = (\mathbf{v}^g)_{g \in \mathcal{G}}$, where each \mathbf{v}^g is a vector in \mathbb{R}^p , that satisfy $\text{supp}(\mathbf{v}^g) \subset g$ for each $g \in \mathcal{G}$.

For any differentiable function $f : \mathbb{R}^p \rightarrow \mathbb{R}$, we denote by $\nabla f(\mathbf{w}) \in \mathbb{R}^p$ the gradient of f at $\mathbf{w} \in \mathbb{R}^p$ and by $\nabla_g f(\mathbf{w}) \in \mathbb{R}^g$ the partial gradient of f with respect to the covariates in g .

In optimization problems throughout the paper we will use the convention that $\frac{0}{0} = 0$ so that the \mathbb{R} -valued function $(x, y) \mapsto \frac{x^2}{y}$ is well defined and jointly convex on $\mathbb{R} \times \mathbb{R}_+$.

3. Group Lasso with overlapping groups

Given a set of groups \mathcal{G} which form a partition of $[1, p]$, the group Lasso penalty (Yuan and Lin, 2006) is a norm over \mathbb{R}^p defined as :

$$\forall \mathbf{w} \in \mathbb{R}^p, \quad \|\mathbf{w}\|_{\ell_1/\ell_2} = \sum_{g \in \mathcal{G}} d_g \|\mathbf{w}_g\| , \quad (1)$$

where $(d_g)_{g \in \mathcal{G}}$ are positive weights. This is a norm whose balls have singularities when some \mathbf{w}_g are equal to zero. Minimizing a smooth convex loss functional $L : \mathbb{R}^p \rightarrow \mathbb{R}$ over such a ball, or equivalently solving the following optimization problem for some $\lambda > 0$:

$$\min_{\mathbf{w} \in \mathbb{R}^p} L(\mathbf{w}) + \lambda \sum_{g \in \mathcal{G}} d_g \|\mathbf{w}_g\| , \quad (2)$$

often leads to a solution that lies on a singularity, *i.e.*, to a vector \mathbf{w} such that $\mathbf{w}_g = \mathbf{0}$ for some of the groups g in \mathcal{G} . Equivalently, the solution is sparse at the group level, in the sense that coefficients within a group are usually zero or nonzero together. The hyperparameter $\lambda \geq 0$ in (2) is used to adjust the tradeoff between minimizing the loss and finding a solution which is sparse at the group level.

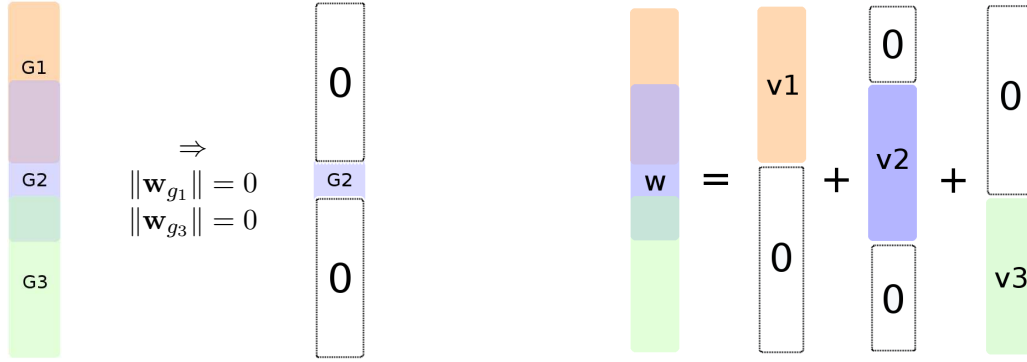


Figure 1: (a) *Left*: Effect of penalty (1) on the support: removing *any* group containing a variable removes the variable from the support. When variables in groups 1 and 3 are shrunk to zero, the support of the solution consists of the variables of the second group which are neither in the first, nor in the third. (b) *Right*: Latent decomposition of \mathbf{w} over $(\mathbf{v}^g)_{g \in \mathcal{G}}$: applying the ℓ_1/ℓ_2 penalty to the decomposition instead of applying it to the \mathbf{w}_g removes only the variables which do not belong to *any* selected group. The support of the solution if latent vectors \mathbf{v}_1 and \mathbf{v}_3 are shrunk to zero will be all variables in the second group.

When \mathcal{G} is not a partition anymore and some of its groups overlap, the penalty (1) is still a norm, because we assume that all covariates belong to at least one group. However, while the Lasso is sometimes loosely presented as *selecting* covariates and the group Lasso as *selecting* groups of covariates, the group Lasso estimator (2) does not necessarily select groups in that case. The reason is that the precise effect of non-differentiable penalties is to *set* covariates, or groups of covariates, to zero, and not to select them. When there is no overlap between groups, setting groups to zero leaves the other full groups to nonzero, which can give the impression that group Lasso is generally appropriate to select a small number of groups. When the groups overlap, however, setting one group to zero shrinks its covariates to zero even if they belong to other groups, in which case these other groups will not be entirely selected. This is illustrated in Figure 1(a) with three overlapping groups of covariates. If the penalty leads to an estimate in which the norm of the first and of the third group are zero, what remains nonzero is not the second group, but the covariates of the second group which are neither in the first nor in the third one. More formally, the overlapping case has been extensively studied by Jenatton et al. (2009), who showed that in the case where $L(\mathbf{w})$ is an empirical risk and under very general assumptions on the data, the support of a solution $\hat{\mathbf{w}}$ of (2) almost surely satisfies

$$\text{supp}(\hat{\mathbf{w}}) = \left(\bigcup_{g \in \mathcal{G}_0} g \right)^c$$

for some $\mathcal{G}_0 \subset \mathcal{G}$, *i.e.*, the support is almost surely the complement of a union of groups. Equivalently, the support is an intersection of the complements of some of groups considered.

In this work, we are interested in penalties which induce a different effect: we want the estimator to select entire groups of covariate, or more precisely we want the support of the solution $\hat{\mathbf{w}}$ to be a *union of groups*. For that purpose, we introduce a set of latent variables $\bar{\mathbf{v}} = (\mathbf{v}^g)_{g \in \mathcal{G}}$ such that $\mathbf{v}^g \in \mathbb{R}^p$ and $\text{supp}(\mathbf{v}^g) \subset g$ for each group $g \in \mathcal{G}$, and propose to solve the following problem instead of (2):

$$\min_{\mathbf{w} \in \mathbb{R}^p, \bar{\mathbf{v}} \in \mathcal{V}_{\mathcal{G}}} L(\mathbf{w}) + \lambda \sum_{g \in \mathcal{G}} d_g \|\mathbf{v}^g\| \quad \text{s.t.} \quad \mathbf{w} = \sum_{g \in \mathcal{G}} \mathbf{v}^g. \quad (3)$$

Problem (3) is always feasible since we assume that all covariates belong to at least one group. Intuitively, the vectors $\bar{\mathbf{v}} = (\mathbf{v}^g)_{g \in \mathcal{G}}$ in (3) represent a decomposition of \mathbf{w} as a sum of latent vectors whose supports are included in each group, as illustrated in Figure 1(b). Applying the ℓ_1/ℓ_2 penalty to these latent vectors favors solutions which shrink some \mathbf{v}^g to 0, while the non-shrunk components satisfy $\text{supp}(\mathbf{v}^g) = g$. On the other hand, since we enforce $\mathbf{w} = \sum_{g \in \mathcal{G}} \mathbf{v}^g$, a \mathbf{w}_i can be nonzero as long as i belongs to at least one non-shrunk group. More precisely, if we denote by $\mathcal{G}_1 \subset \mathcal{G}$ the set of groups g with $\hat{\mathbf{v}}^g \neq \mathbf{0}$ for the solution of (3), then we immediately get $\hat{\mathbf{w}} = \sum_{g \in \mathcal{G}_1} \hat{\mathbf{v}}^g$, and therefore we can expect:

$$\text{supp}(\hat{\mathbf{w}}) = \bigcup_{g \in \mathcal{G}_1} g.$$

In other words, this formulation leads to sparse solutions whose support is likely to be a union of groups.

Interestingly, problem (3) can be reformulated as the minimization of the cost function $L(\mathbf{w})$ penalized by a new regularizer which is a function of \mathbf{w} only. Indeed since the minimization over $\bar{\mathbf{v}}$ only involves the penalty term and the constraints, we can rewrite (3) as

$$\min_{\mathbf{w} \in \mathbb{R}^p} L(\mathbf{w}) + \lambda \Omega_{\bigcup}^{\mathcal{G}}(\mathbf{w}), \quad (4)$$

with

$$\Omega_{\bigcup}^{\mathcal{G}}(\mathbf{w}) \triangleq \min_{\bar{\mathbf{v}} \in \mathcal{V}_{\mathcal{G}}, \sum_{g \in \mathcal{G}} \mathbf{v}^g = \mathbf{w}} \sum_{g \in \mathcal{G}} d_g \|\mathbf{v}^g\|. \quad (5)$$

We call this penalty the *latent group Lasso* penalty, in reference to its formulation as a group Lasso over latent variables. When the groups do not overlap and form a partition, there exists a unique decomposition of $\mathbf{w} \in \mathbb{R}^p$ as $\mathbf{w} = \sum_{g \in \mathcal{G}} \mathbf{v}^g$ with $\text{supp}(\mathbf{v}^g) \subset g$, namely, $\mathbf{v}^g = \mathbf{w}_g$ for all $g \in \mathcal{G}$. In that case, both the group Lasso penalty (1) and the latent group Lasso penalty (5) are equal and boil down to the same standard group Lasso. When some groups overlap, however, the two penalties differ. For example, Figure 2 shows the unit ball for both norms in \mathbb{R}^3 with groups $\mathcal{G} = \{\{1, 2\}, \{2, 3\}\}$. The pillow shaped ball of $\|\cdot\|_{\ell_1/\ell_2}$ has four singularities corresponding to cases where either only \mathbf{w}_1 or only \mathbf{w}_3 is nonzero. By contrast, $\Omega_{\bigcup}^{\mathcal{G}}$ has two circular sets of singularities corresponding to cases where $(\mathbf{w}_1, \mathbf{w}_2)$ only or $(\mathbf{w}_2, \mathbf{w}_3)$ only is nonzero. For comparison, we also show the unit ball when we consider the partition $\mathcal{G} = \{\{1, 2\}, \{3\}\}$, in which case both norms coincide: singularities appear for $(\mathbf{w}_1, \mathbf{w}_2) = \mathbf{0}$ or $\mathbf{w}_3 = \mathbf{0}$.

To summarize, we enforce a prior we have on \mathbf{w} by introducing new variables in the optimization problem (3). The constraint we impose is that some groups should be shrunk

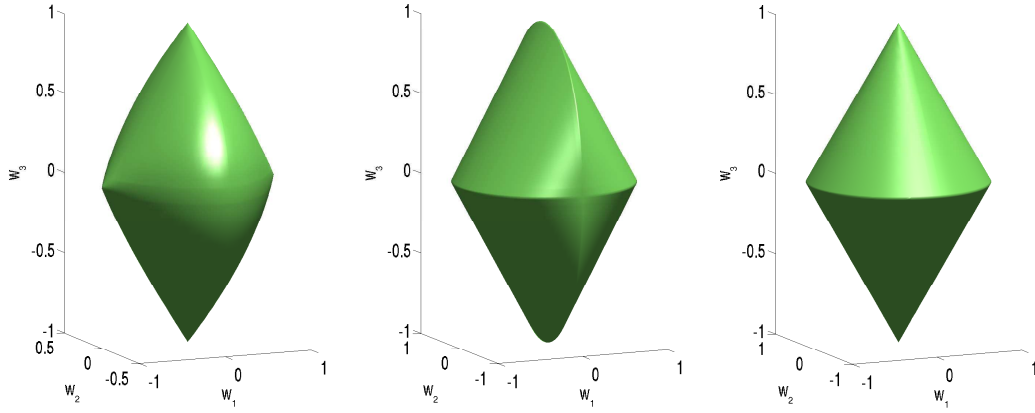


Figure 2: Unit balls for $\|\cdot\|_{\ell_1/\ell_2}$ (left), proposed by [Jenatton et al. \(2009\)](#), and $\Omega_{\mathcal{G}}^{\mathcal{G}}$ (middle), proposed in this paper, for the groups $\mathcal{G} = \{\{1, 2\}, \{2, 3\}\}$. \mathbf{w}_2 is represented as the vertical coordinate. We note that singularities exist in both cases, but occur at different positions: for $\|\cdot\|_{\ell_1/\ell_2}$ they correspond to situations where only \mathbf{w}_1 or only \mathbf{w}_2 is nonzero, *i.e.*, where all covariates of one group are shrunk to 0; for $\Omega_{\mathcal{G}}^{\mathcal{G}}$, they correspond to situations where only \mathbf{w}_1 or only \mathbf{w}_3 is equal to 0, *i.e.*, where all covariates of one group are nonzero. For comparison, we show on the right the unit ball of both norms for the partition $\mathcal{G} = \{\{1, 2\}, \{3\}\}$, where they both reduce to the classical group Lasso penalty.

to zero, and a covariate should have zero weight in \mathbf{w} if all the groups to which it belongs are set to zero. Equivalently, the support of \mathbf{w} should be a union of groups. This new problem can be re-written as a classical minimization of the empirical risk, penalized by a particular penalty $\Omega_{\mathcal{U}}^{\mathcal{G}}$ defined in (5). This penalty itself associates to each vector \mathbf{w} the solution of a particular constrained optimization problem. While this formulation may not be the most intuitive, it allows to reframe the problem in the classical context of penalized empirical risk minimization. In the remaining of this article, we investigate in more details the latent group Lasso penalty $\Omega_{\mathcal{U}}^{\mathcal{G}}$, both theoretically and empirically.

3.1 Related work

The idea of decomposing a parameter vector into some latent components and to regularize each of these components separately has appeared recently in the literature independently of this work. In particular Jalali et al. (2010) proposed to consider such a decomposition in the case of multi-task learning, where each task specific parameter vector is decomposed into a first ℓ_1 regularized vector and another vector, regularized with an ℓ_1/ℓ_∞ norm; so as to share its sparsity pattern with all other tasks. The norm considered in that work could be interpreted as a special case of the latent group Lasso, where the set of groups consists of all singletons and groups of coefficients associated with the same feature across task. The decomposition into latent variables is even more natural in the context of the work of Chen et al. (2011), Candès et al. (2009), or Agarwal et al. (2011) on robust PCA and matrix decomposition in which a matrix is decomposed in a low rank matrix regularized by the trace norm and a sparse or column-sparse matrix regularized by an ℓ_1 or group ℓ_1 -norm.

Another type of decompositions which is related to this norm is the idea of *cover* of the support. In particular it is interesting to consider the ℓ_0 counterpart to this norm, which could be written as

$$\Omega_0^{\mathcal{G}} = \min_{\tilde{\mathcal{G}}_1 \subset \mathcal{G}} \sum_{g \in \tilde{\mathcal{G}}_1} d_g \quad \text{s.t.} \quad \mathbf{w} = \sum_{g \in \tilde{\mathcal{G}}_1} \mathbf{v}^g, \text{ supp}(\mathbf{v}^g) \subset g.$$

$\Omega_0^{\mathcal{G}}$ can then be interpreted as the value of a min set-cover. This penalization has been considered in Huang et al. (2009) under the name *block coding*, since, indeed, when d_g is interpreted as a coding length, this penalization induces a code length on all sets, which can be interpreted in the MDL framework.

More generally, one could consider $\Omega_q^{\mathcal{G}}$ penalties, for all $q \geq 0$, by replacing the ℓ_2 norm used in the definition of the latent group Lasso penalty (5) by a ℓ_q norm. It should be noted then that, unlike the support, the definition of group-support we introduce in Section 5 changes if one considers the latent group Lasso with a different ℓ_q -norm, and even if the weights d_g change ¹.

Obozinski and F. (2011) considers the case of $\Omega_q^{\mathcal{G}}$, when the weights are given by a set function and shows that $\Omega_q^{\mathcal{G}}$ is then the tightest convex “ ℓ_q relaxation of the block-coding scheme of Huang et al. (2009). It also shows that when $\mathcal{G} = 2^V$ and the weights are an appropriate power of a submodular function then $\Omega_q^{\mathcal{G}}$ is the norm that naturally extends the norm considered by Bach (2010).

1. We discuss the choice of weights in detail in Section 8.

It should be noted that recent theoretical analyses of the norm studied in this paper have been proposed by [Percival \(2011\)](#) and [Maurer and Pontil \(2011\)](#). They adopt points of views or focus on questions that are complementary of this work; we discuss those in [section 7.3](#).

4. Some properties of the latent group Lasso penalty

In this section we study a few properties of the latent group Lasso $\Omega_{\mathcal{G}}^{\mathcal{G}}$, which will be in particular useful to prove consistency results in the next section. After showing that $\Omega_{\mathcal{G}}^{\mathcal{G}}$ is a valid norm, we compute its dual norm and provide two variational formulas. We then characterize its unit ball as the convex hull of basic disks, and compute its subdifferential. When used as a penalty for statistical inference, we further reinterpret it in the context of covariate duplication and multiple kernel learning. To lighten notations, in the rest of the paper we simply denote $\Omega_{\mathcal{G}}^{\mathcal{G}}$ by Ω .

4.1 Basic properties

We first analyze the decomposition induced by (5) of a vector $\mathbf{w} \in \mathbb{R}^p$ as $\sum_{g \in \mathcal{G}} \mathbf{v}^g$. We denote by $\mathbf{V}(\mathbf{w}) \subset \mathcal{V}_{\mathcal{G}}$ the set of m -tuples of vectors $\bar{\mathbf{v}} = (\mathbf{v}^g)_{g \in \mathcal{G}} \in \mathcal{V}_{\mathcal{G}}$ that are solutions to the optimization problem in (5), *i.e.*, which satisfy

$$\mathbf{w} = \sum_{g \in \mathcal{G}} \mathbf{v}^g \quad \text{and} \quad \Omega(\mathbf{w}) = \sum_{g \in \mathcal{G}} d_g \|\mathbf{v}^g\|.$$

We first have that:

Lemma 1 *For any $\mathbf{w} \in \mathbb{R}^p$, $\mathbf{V}(\mathbf{w})$ is non-empty, compact and convex.*

Proof The objective of problem (5) is a proper closed convex function with no *direction of recession*. Lemma 1 is then the consequence of classical results in convex analysis, such as Theorem 27.2 page 265 of [Rockafellar \(1997\)](#). \blacksquare

The following statement shows that, unsurprisingly, we can regard Ω as a classical norm-based penalty.

Lemma 2 $\mathbf{w} \mapsto \Omega(\mathbf{w})$ is a norm.

Proof Positive homogeneity and positive definiteness hold trivially. We show the triangular inequality. Consider $\mathbf{w}, \mathbf{w}' \in \mathbb{R}^p$, and let $\bar{\mathbf{v}} \in \mathbf{V}(\mathbf{w})$ and $\bar{\mathbf{v}}' \in \mathbf{V}(\mathbf{w}')$ be respectively optimal decompositions of \mathbf{w} and \mathbf{w}' , so that $\Omega(\mathbf{w}) = \sum_g d_g \|\mathbf{v}^g\|$ and $\Omega(\mathbf{w}') = \sum_g d_g \|\mathbf{v}'^g\|$ with $\mathbf{w} = \sum_g \mathbf{v}^g$ and $\mathbf{w}' = \sum_g \mathbf{v}'^g$. Since $(\mathbf{v}^g + \mathbf{v}'^g)_{g \in \mathcal{G}}$ is a (a priori non-optimal) decomposition of $\mathbf{w} + \mathbf{w}'$, we clearly have :

$$\Omega(\mathbf{w} + \mathbf{w}') \leq \sum_{g \in \mathcal{G}} d_g \|\mathbf{v}^g + \mathbf{v}'^g\| \leq \sum_{g \in \mathcal{G}} d_g (\|\mathbf{v}^g\| + \|\mathbf{v}'^g\|) = \Omega(\mathbf{w}) + \Omega(\mathbf{w}').$$

\blacksquare

4.2 Dual norm and variational characterizations

Ω being a norm, by Lemma 1, we can consider its Fenchel dual norm Ω^* defined by:

$$\forall \alpha \in \mathbb{R}^p, \quad \Omega^*(\alpha) = \sup_{\mathbf{w} \in \mathbb{R}^p} \left\{ \mathbf{w}^\top \alpha \mid \Omega(\mathbf{w}) \leq 1 \right\}. \quad (6)$$

The following lemma shows that Ω^* has a simple closed form expression:

Lemma 3 (dual norm) *The Fenchel dual norm Ω^* of Ω satisfies:*

$$\forall \alpha \in \mathbb{R}^p, \quad \Omega^*(\alpha) = \max_{g \in \mathcal{G}} d_g^{-1} \|\alpha_g\|.$$

Proof We start from the definition of the dual norm (6) and compute:

$$\begin{aligned} \Omega^*(\alpha) &= \max_{\mathbf{w} \in \mathbb{R}^p} \mathbf{w}^\top \alpha && \text{s.t. } \Omega(\mathbf{w}) \leq 1 \\ &= \max_{\mathbf{w} \in \mathbb{R}^p, \bar{\mathbf{v}} \in \mathcal{V}_{\mathcal{G}}} \mathbf{w}^\top \alpha && \text{s.t. } \mathbf{w} = \sum_{g \in \mathcal{G}} \mathbf{v}^g, \sum_{g \in \mathcal{G}} d_g \|\mathbf{v}^g\| \leq 1 \\ &= \max_{\bar{\mathbf{v}} \in \mathcal{V}_{\mathcal{G}}} \sum_{g \in \mathcal{G}} \mathbf{v}^g{}^\top \alpha && \text{s.t. } \sum_{g \in \mathcal{G}} d_g \|\mathbf{v}^g\| \leq 1 \\ &= \max_{\bar{\mathbf{v}} \in \mathcal{V}_{\mathcal{G}}, \eta \in \mathbb{R}_+^m} \sum_{g \in \mathcal{G}} \mathbf{v}^g{}^\top \alpha && \text{s.t. } \sum_{g \in \mathcal{G}} \eta_g \leq 1, \forall g \in \mathcal{G}, d_g \|\mathbf{v}^g\| \leq \eta_g \\ &= \max_{\eta \in \mathbb{R}_+^m} \sum_{g \in \mathcal{G}} \eta_g d_g^{-1} \|\alpha_g\| && \text{s.t. } \sum_{g \in \mathcal{G}} \eta_g \leq 1 \\ &= \max_{g \in \mathcal{G}} d_g^{-1} \|\alpha_g\|. \end{aligned}$$

The second equality is due to the fact that :

$$\{\mathbf{w} \mid \Omega(\mathbf{w}) \leq 1\} = \left\{ \mathbf{w} \mid \exists \bar{\mathbf{v}} \in \mathcal{V}_{\mathcal{G}} \text{ s.t. } \mathbf{w} = \sum_g \mathbf{v}^g, \sum_g d_g \|\mathbf{v}^g\| \leq 1 \right\},$$

and the fifth results from the explicit solution $\mathbf{v}^g = \alpha_g \eta_g d_g^{-1} \|\alpha_g\|^{-1}$ of the maximization in $\bar{\mathbf{v}}$ in the fourth line. \blacksquare

Remark 4 *Remembering that the infimal convolution $f \star_{\text{inf}} g$ of two convex functions f and g is defined as $(f \star_{\text{inf}} g)(\mathbf{w}) = \inf_{\mathbf{v} \in \mathbb{R}^p} \{f(\mathbf{v}) + g(\mathbf{w} - \mathbf{v})\}$ (see [Rockafellar, 1997](#)), it could be noted that Ω is the infimal convolution of all functions ω_g for $g \in \mathcal{G}$ defined as $\omega_g : \mathbf{w} \mapsto \|\mathbf{w}_g\|_{\iota_g(\mathbf{w})}$ with $\iota_g(\mathbf{w}) = 0$ if $\text{supp}(\mathbf{w}) \subset g$ and $+\infty$ otherwise. One of the main properties motivating the notion of infimal convolution is the fact that it can be defined via $(f \star_{\text{inf}} g)^* = f^* + g^*$, where $*$ denotes Fenchel-Legendre conjugation. Several of the properties of Ω can be derived from this interpretation but we will however show them directly.*

The norm Ω was initially defined as the solution of an optimization problem in (5). From the characterization of Ω^* we can easily derive a second variational formulation:

Lemma 5 (second variational formulation) *For any $\mathbf{w} \in \mathbb{R}^p$, we have*

$$\Omega(\mathbf{w}) = \max_{\alpha \in \mathbb{R}^p} \alpha^\top \mathbf{w} \quad \text{s.t.} \quad \|\alpha_g\| \leq d_g \quad \text{for all } g \in \mathcal{G}. \quad (7)$$

Proof Since the bi-dual of a norm is the norm itself, we have the variational form

$$\Omega(\mathbf{w}) = \max_{\alpha \in \mathbb{R}^p} \alpha^\top \mathbf{w} \quad \text{s.t.} \quad \Omega^*(\alpha) \leq 1. \quad (8)$$

Plugging the characterization of Ω^* of Lemma 3 into this equation finishes the proof. \blacksquare

For any $\mathbf{w} \in \mathbb{R}^p$, we denote by $\mathcal{A}(\mathbf{w})$ the set of $\alpha \in \mathbb{R}^p$ in the dual unit sphere which solve the second variational formulation (7) of Ω , namely:

$$\mathcal{A}(\mathbf{w}) \triangleq \operatorname{argmax}_{\alpha \in \mathbb{R}^p, \Omega^*(\alpha) \leq 1} \alpha^\top \mathbf{w}. \quad (9)$$

With a few more efforts, we can also derive a third variational representation of the norm Ω , which will be useful in Section 7 in the proofs of consistency:

Lemma 6 (third variational formulation) *For any $\mathbf{w} \in \mathbb{R}^p$, we also have*

$$\Omega(\mathbf{w}) = \frac{1}{2} \min_{\lambda \in \mathbb{R}_+^m} \sum_{i=1}^p \frac{\mathbf{w}_i^2}{\sum_{g \ni i} \lambda_g} + \sum_{g \in \mathcal{G}} d_g^2 \lambda_g. \quad (10)$$

Proof For any $\mathbf{w} \in \mathbb{R}^p$, we can rewrite the solution of the constrained optimization problem of the second variational formulation (7) as the saddle point of the Lagrangian:

$$\Omega(\mathbf{w}) = \min_{\lambda \in \mathbb{R}_+^m} \max_{\alpha \in \mathbb{R}^p} \mathbf{w}^\top \alpha - \frac{1}{2} \sum_{g \in \mathcal{G}} \lambda_g (\|\alpha_g\|^2 - d_g^2).$$

Optimizing in α leads to α being solution of $\mathbf{w}_i = \alpha_i \sum_{g \ni i} \lambda_g$, which (distinguishing the cases $\mathbf{w}_i = 0$ and $\mathbf{w}_i \neq 0$) yields problem (10) when replacing α_i by its optimal value. \blacksquare

Let us denote by $\Lambda(\mathbf{w}) \subset \mathbb{R}_+^m$ the set of solutions to the third variational formulation (10). Note that there is not necessarily a unique solution to (10), because the Hessian of the objective function is not always positive definite (see lemma 48 in Appendix D for a characterization of cases in which positive definiteness can be guaranteed). For any $\mathbf{w} \in \mathbb{R}^p$, we now have three variational formulations for $\Omega(\mathbf{w})$, namely (5), (7) and (10), with respective solutions sets $\mathbf{V}(\mathbf{w})$, $\mathcal{A}(\mathbf{w})$ and $\Lambda(\mathbf{w})$. The following lemma shows that $\mathbf{V}(\mathbf{w})$ is in bijection with $\Lambda(\mathbf{w})$.

Lemma 7 *Let $\mathbf{w} \in \mathbb{R}^p$. The mapping*

$$\begin{aligned} \lambda &: \mathcal{V}_{\mathcal{G}} \rightarrow \mathbb{R}^m \\ \bar{\mathbf{v}} &\mapsto \lambda(\bar{\mathbf{v}}) = (d_g^{-1} \|\mathbf{v}^g\|)_{g \in \mathcal{G}} \end{aligned} \quad (11)$$

is a bijection from $\mathbf{V}(\mathbf{w})$ to $\Lambda(\mathbf{w})$. For any $\lambda \in \Lambda(\mathbf{w})$, the only vector $\bar{\mathbf{v}} \in \mathbf{V}(\mathbf{w})$ that satisfies $\lambda(\bar{\mathbf{v}}) = \lambda$ is given by $\mathbf{v}_g^g = \lambda_g \alpha_g$, where α is any vector of $\mathcal{A}(\mathbf{w})$.

Proof To express the penalty as a minimization problem, let us use the following basic equality valid for any $x \in \mathbb{R}_+$:

$$x = \frac{1}{2} \min_{\eta \geq 0} \left[\frac{x^2}{\eta} + \eta \right],$$

where the unique minimum in η is reached for $\eta = x$. From this we deduce that, for any $\mathbf{v} \in \mathbb{R}^p$ and $d > 0$:

$$d \|\mathbf{v}\| = \frac{1}{2} \min_{\eta \geq 0} d \left[\frac{\|\mathbf{v}\|^2}{\eta} + \eta \right] = \frac{1}{2} \min_{\lambda' \geq 0} \left[\frac{\|\mathbf{v}\|^2}{\lambda'} + d^2 \lambda' \right],$$

where the unique minimum in the last term is attained for $\lambda' = d^{-1} \|\mathbf{v}\|$. Using definition (5) we can therefore write $\Omega(\mathbf{w})$ as the optimum value of a jointly convex optimization problem in $\bar{\mathbf{v}} \in \mathcal{V}_{\mathcal{G}}$ and $\boldsymbol{\lambda}' = (\lambda'_g)_{g \in \mathcal{G}} \in \mathbb{R}_+^m$:

$$\Omega(\mathbf{w}) = \min_{\bar{\mathbf{v}} \in \mathcal{V}_{\mathcal{G}}, \sum_{g \in \mathcal{G}} \mathbf{v}^g = \mathbf{w}, \boldsymbol{\lambda}' \in \mathbb{R}_+^m} \frac{1}{2} \sum_{g \in \mathcal{G}} \left[\frac{\|\mathbf{v}^g\|^2}{\lambda'_g} + d_g^2 \lambda'_g \right], \quad (12)$$

where for any $\bar{\mathbf{v}}$, the minimum in $\boldsymbol{\lambda}'$ is uniquely attained for $\boldsymbol{\lambda}' = \boldsymbol{\lambda}(\bar{\mathbf{v}})$ defined in (11). By definition of $\mathbf{V}(\mathbf{w})$, the set of solutions of (12) is therefore exactly the set of pairs of the form $(\bar{\mathbf{v}}, \boldsymbol{\lambda}(\bar{\mathbf{v}}))$ for $\bar{\mathbf{v}} \in \mathbf{V}(\mathbf{w})$. Let us now isolate the minimization over $\bar{\mathbf{v}}$ in (12). To incorporate the constraint $\sum_{g \in \mathcal{G}} \mathbf{v}^g = \mathbf{w}$ we rewrite (12) with a Lagrangian:

$$\Omega(\mathbf{w}) = \min_{\boldsymbol{\lambda}' \in \mathbb{R}_+^m} \max_{\boldsymbol{\alpha}' \in \mathbb{R}^p} \min_{\bar{\mathbf{v}} \in \mathcal{V}_{\mathcal{G}}} \frac{1}{2} \sum_{g \in \mathcal{G}} \left[\frac{\|\mathbf{v}^g\|^2}{\lambda'_g} + d_g^2 \lambda'_g \right] + \boldsymbol{\alpha}'^\top (\mathbf{w} - \sum_{g \in \mathcal{G}} \mathbf{v}^g).$$

The inner minimization in $\bar{\mathbf{v}}$, for fixed $\boldsymbol{\lambda}'$ and $\boldsymbol{\alpha}'$, yields $\mathbf{v}_i^g = \lambda'_g \boldsymbol{\alpha}'_i$. The constraint $\mathbf{w} = \sum_{g \in \mathcal{G}} \mathbf{v}^g$ therefore implies that, after optimization in $\bar{\mathbf{v}}$ and $\boldsymbol{\alpha}'$, we have $\boldsymbol{\alpha}'_i = \frac{w_i}{\sum_{g \ni i} \lambda'_g}$, and as a consequence that $\mathbf{v}_i^g = \frac{\lambda'_g}{\sum_{h \ni i} \lambda'_h} \mathbf{w}_i$. A small computation now shows that, after optimization in $\bar{\mathbf{v}}$ and $\boldsymbol{\alpha}'$ for a fixed $\boldsymbol{\lambda}'$, we have:

$$\sum_{g \in \mathcal{G}} \frac{\|\mathbf{v}^g\|^2}{\lambda'_g} = \sum_{i=1}^p \sum_{g \ni i} \frac{(\mathbf{v}_i^g)^2}{\lambda'_g} = \sum_{i=1}^p \sum_{g \ni i} \frac{\lambda'_g \mathbf{w}_i^2}{\left(\sum_{h \ni i} \lambda'_h \right)^2} = \sum_{i=1}^p \frac{\mathbf{w}_i^2}{\sum_{h \ni i} \lambda'_h}.$$

Plugging this into (12), we see that after optimization in $\bar{\mathbf{v}}$, the optimization problem in $\boldsymbol{\lambda}'$ is exactly (10), which by definition admits $\boldsymbol{\Lambda}(\mathbf{w})$ as solutions, while we showed that (12) admits $\boldsymbol{\lambda}(\mathbf{V}(\mathbf{w}))$ as solutions. This shows that $\boldsymbol{\lambda}(\mathbf{V}(\mathbf{w})) = \boldsymbol{\Lambda}(\mathbf{w})$, and since for any $\boldsymbol{\lambda}' \in \boldsymbol{\Lambda}(\mathbf{w})$ there exists a unique $\bar{\mathbf{v}} \in \mathbf{V}(\mathbf{w})$ that satisfies $\boldsymbol{\lambda}(\bar{\mathbf{v}}) = \boldsymbol{\lambda}'$, namely, $\mathbf{v}_i^g = \frac{\lambda'_g}{\sum_{h \ni i} \lambda'_h} \mathbf{w}_i$, $\boldsymbol{\lambda}$ is indeed a bijection from $\mathbf{V}(\mathbf{w})$ to $\boldsymbol{\Lambda}(\mathbf{w})$. Finally, we noted in the proof of Lemma 6 that for any $\boldsymbol{\lambda} \in \boldsymbol{\Lambda}(\mathbf{w})$ and $\boldsymbol{\alpha} \in \mathcal{A}(\mathbf{w})$, $\mathbf{w}_i = \boldsymbol{\alpha}_i \sum_{h \ni i} \lambda_h$. This shows that the unique $\bar{\mathbf{v}} \in \mathbf{V}(\mathbf{w})$ associated to a $\boldsymbol{\lambda} \in \boldsymbol{\Lambda}(\mathbf{w})$ can equivalently be written $\mathbf{v}_i^g = \lambda_g \boldsymbol{\alpha}_g$, which concludes the proof of Lemma 7. \blacksquare

4.3 Characterization of the unit ball of Ω as a convex hull

Figure 2(b) suggests visually that the unit ball of Ω is just the convex hull of a horizontal disk and a vertical one. This impression is correct and formalized more generally in the following lemma.

Lemma 8 *For any group $g \in \mathcal{G}$, define the hyperdisks $\mathcal{D}_g = \{\mathbf{w} \in \mathbb{R}^p \mid \|\mathbf{w}_g\| \leq d_g^{-1}, \mathbf{w}_{g^c} = \mathbf{0}\}$. Then, the unit ball of Ω is the convex hull of the union of hyper-disks $\cup_{g \in \mathcal{G}} \mathcal{D}_g$.*

Proof Let $\mathbf{w} \in \text{ConvHull}(\cup_{g \in \mathcal{G}} \mathcal{D}_g)$, then there exist $\boldsymbol{\alpha}^g \in \mathcal{D}_g$ and $t_g \in \mathbb{R}_+$, for all $g \in \mathcal{G}$, such that $\sum_{g \in \mathcal{G}} t_g \leq 1$ and $\mathbf{w} = \sum_{g \in \mathcal{G}} t_g \boldsymbol{\alpha}^g$. Letting $\bar{\mathbf{v}} = (t_g \boldsymbol{\alpha}^g)_{g \in \mathcal{G}}$ as a suboptimal decomposition of \mathbf{w} , we easily get

$$\Omega(\mathbf{w}) \leq \sum_{g \in \mathcal{G}} d_g \|t_g \boldsymbol{\alpha}^g\| \leq \sum_{g \in \mathcal{G}} t_g \leq 1.$$

Conversely, if $\Omega(\mathbf{w}) \leq 1$, then there exists $\bar{\mathbf{v}} \in \mathcal{V}_{\mathcal{G}}$, such that $\sum_{g \in \mathcal{G}} d_g \|\mathbf{v}^g\| \leq 1$ and we obtain $\boldsymbol{\alpha}^g \in \mathcal{D}_g$ and t_g in the simplex by letting $t_g = d_g \|\mathbf{v}^g\|$ and

$$\boldsymbol{\alpha}^g = \begin{cases} \mathbf{0} & \text{if } t_g = 0, \\ \frac{\mathbf{v}^g}{d_g \|\mathbf{v}^g\|} & \text{else.} \end{cases}$$

■

It should be noted that this lemma shows that Ω is the gauge of the convex hull of the disks \mathcal{D}_g , in other words, Ω is, in the terminology introduced by [Chandrasekaran et al. \(2010\)](#), the unit ball of the *atomic norm* associated with the union of disks \mathcal{D}_g .

4.4 Subdifferential of Ω

The subdifferential of Ω at \mathbf{w} is, by definition:

$$\partial\Omega(\mathbf{w}) \triangleq \{\mathbf{s} \in \mathbb{R}^p \mid \forall \mathbf{h} \in \mathbb{R}^p, \Omega(\mathbf{w} + \mathbf{h}) - \Omega(\mathbf{w}) \geq \mathbf{s}^\top \mathbf{h}\}.$$

It is a standard result of convex optimization (resulting e.g. from characterization (b*) of the subdifferential in Theorem 23.5, p. 218, [Rockafellar, 1997](#)) that for all $\mathbf{w} \in \mathbb{R}^p$, $\partial\Omega(\mathbf{w}) = \mathcal{A}(\mathbf{w})$, where $\mathcal{A}(\mathbf{w})$ was defined in (9).

We can now show a simple relationship between the decomposition $(\mathbf{v}^g)_{g \in \mathcal{G}}$ of a vector \mathbf{w} induced by Ω , and the subdifferential of Ω .

Lemma 9 *For any $\boldsymbol{\alpha} \in \mathcal{A}(\mathbf{w}) = \partial\Omega(\mathbf{w})$ and any $\bar{\mathbf{v}} \in \mathbf{V}(\mathbf{w})$,*

$$\begin{cases} \text{either } \mathbf{v}^g \neq \mathbf{0} & \text{and } \boldsymbol{\alpha}_g = d_g \frac{\mathbf{v}^g}{\|\mathbf{v}^g\|}, \\ \text{or } \mathbf{v}^g = \mathbf{0} & \text{and } \|\boldsymbol{\alpha}_g\| \leq d_g. \end{cases}$$

Proof Let $\bar{\mathbf{v}} \in \mathbf{V}(\mathbf{w})$ and $\boldsymbol{\alpha} \in \mathcal{A}(\mathbf{w})$. Since $\Omega^*(\boldsymbol{\alpha}) \leq 1$, we have $\|\boldsymbol{\alpha}_g\| \leq d_g$ which implies $\boldsymbol{\alpha}^\top \mathbf{v}^g \leq d_g \|\mathbf{v}^g\|$. On the other hand, we also have $\boldsymbol{\alpha}^\top \mathbf{w} = \Omega(\mathbf{w})$ so that $0 = \Omega(\mathbf{w}) - \boldsymbol{\alpha}^\top \mathbf{w} = \sum_g (d_g \|\mathbf{v}^g\| - \boldsymbol{\alpha}_g^\top \mathbf{v}^g)$, which is a sum of non-negative terms. We conclude that, for all $g \in \mathcal{G}$, we have $\boldsymbol{\alpha}_g^\top \mathbf{v}^g = d_g \|\mathbf{v}^g\|$ which yields the result. ■

We can deduce a general property of all decompositions of given vector:

Corollary 10 *Let $\mathbf{w} \in \mathbb{R}^p$. For all $\bar{\mathbf{v}}, \bar{\mathbf{v}}' \in \mathbf{V}(\mathbf{w})$, and for all $g \in \mathcal{G}$ we have $\mathbf{v}^g = \mathbf{0}$ or $\mathbf{v}'^g = \mathbf{0}$ or there exists $\gamma \in \mathbb{R}$ such that $\mathbf{v}^g = \gamma \mathbf{v}'^g$.*

Proof By Lemma 9, if $\mathbf{v}^g \neq \mathbf{0}$ and $\mathbf{v}'^g \neq \mathbf{0}$, then $\alpha_g = d_g \frac{\mathbf{v}^g}{\|\mathbf{v}^g\|} = d_g \frac{\mathbf{v}'^g}{\|\mathbf{v}'^g\|}$ so that $\mathbf{v}^g = \frac{\|\mathbf{v}^g\|}{\|\mathbf{v}'^g\|} \mathbf{v}'^g$. \blacksquare

4.5 Ω as a regularizer

We consider in this section the situation where Ω is used as a regularizer for an empirical risk minimization problem. Specifically, let us consider a convex differentiable loss function $\ell : \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}$, such as the squared error $\ell(t, y) = (t - y)^2$ for regression problems or the logistic loss $\ell(t, y) = \log(1 + e^{-yt})$ for classification problems where $y = \pm 1$. Given a set of n training pairs $(\mathbf{x}^{(i)}, y^{(i)}) \in \mathbb{R}^p \times \mathbb{R}$, $i = 1, \dots, n$, we define the empirical risk $L(\mathbf{w}) = \frac{1}{n} \sum_{i=1}^n \ell(\mathbf{w}^\top \mathbf{x}^{(i)}, y^{(i)})$ and consider the regularized empirical risk minimization problem

$$\min_{\mathbf{w} \in \mathbb{R}^p} L(\mathbf{w}) + \lambda \Omega(\mathbf{w}). \quad (13)$$

Its solutions are characterized by optimality conditions from subgradient calculus:

Lemma 11 *A vector $\mathbf{w} \in \mathbb{R}^p$ is a solution of (13) if and only if one of the following equivalent conditions is satisfied*

$$(a) \quad -\nabla L(\mathbf{w})/\lambda \in \mathcal{A}(\mathbf{w})$$

$$(b) \quad \mathbf{w} \text{ can be decomposed as } \mathbf{w} = \sum_{g \in \mathcal{G}} \mathbf{v}^g \text{ for some } \bar{\mathbf{v}} \in \mathcal{V}_{\mathcal{G}} \text{ with for all } g \in \mathcal{G}:$$

$$\text{either } \mathbf{v}_g \neq \mathbf{0} \text{ and } \nabla_g L(\mathbf{w}) = -\lambda d_g \mathbf{v}^g / \|\mathbf{v}^g\| \quad \text{or} \quad \mathbf{v}^g = \mathbf{0} \text{ and } d_g^{-1} \|\nabla_g L(\mathbf{w})\| \leq \lambda.$$

Proof (a) is immediate from subgradient calculus and the fact that $\partial \Omega(\mathbf{w}) = \mathcal{A}(\mathbf{w})$ (see Section 4.4). (b) is immediate from Lemma 9. \blacksquare

4.6 Covariate duplication

In this section we show that empirical risk minimization penalized by Ω is equivalent to a regular group Lasso in a covariate space of higher dimension obtained by duplication of the covariates belonging to several groups. This has implications for practical implementation of Ω as a regularizer and for its generalization to non-linear classification.

More precisely, let us consider the duplication operator:

$$\begin{aligned} \mathbb{R}^p &\rightarrow \mathbb{R}^{\sum_{g \in \mathcal{G}} |g|} \\ \mathbf{x} &\mapsto \tilde{\mathbf{x}} = \bigoplus_{g \in \mathcal{G}} (\mathbf{x}_i)_{i \in g}. \end{aligned} \quad (14)$$

In other words, $\tilde{\mathbf{x}}$ is obtained by stacking the restrictions of \mathbf{x} to each group on top of each other, resulting in a $(\sum_{g \in \mathcal{G}} |g|)$ -dimensional vector. Note that any coordinate of \mathbf{x} that

occurs in several groups will be duplicated as many times in $\tilde{\mathbf{x}}$. Similarly, for a vector $\mathbf{v} \in \mathcal{V}_g$, let us denote by $\tilde{\mathbf{v}}$ the $(\sum_{g \in \mathcal{G}} |g|)$ -dimensional vector obtained by stacking the restrictions of the successive \mathbf{v}^g on their corresponding groups g on top of each other (resulting in no loss of information, since \mathbf{v}^g is null outside of g). This operation is illustrated in (18) below. Then for any $\mathbf{w} \in \mathbb{R}^p$ and $\mathbf{v} \in \mathcal{V}_g$ such that $\mathbf{w} = \sum_{g \in \mathcal{G}} \mathbf{v}^g$, we easily get, for any $\mathbf{x} \in \mathbb{R}^p$:

$$\mathbf{w}^\top \mathbf{x} = \sum_{g \in \mathcal{G}} \mathbf{v}^{g\top} \mathbf{x} = \tilde{\mathbf{v}}^\top \tilde{\mathbf{x}}. \quad (15)$$

Consider now a learning problem with training points $\mathbf{x}^1, \dots, \mathbf{x}^n \in \mathbb{R}^p$ where we minimize over $\mathbf{w} \in \mathbb{R}^p$ a penalized risk function that depends of \mathbf{w} only through inner products with the training points, *i.e.*, or the form

$$\min_{\mathbf{w} \in \mathbb{R}^p} \tilde{L}(\mathbf{X}\mathbf{w}) + \lambda \Omega(\mathbf{w}), \quad (16)$$

where \mathbf{X} is the $n \times p$ matrix of training points and $\mathbf{X}\mathbf{w}$ is therefore the vector of inner products of \mathbf{w} with the training points. Many problems, in particular those considered in Section 4.5, have this form. By definition of Ω we can rewrite (16) as

$$\min_{\mathbf{w} \in \mathbb{R}^p, \mathbf{v} \in \mathcal{V}_g, \sum_g \mathbf{v}^g = \mathbf{w}} \tilde{L}(\mathbf{X}\mathbf{w}) + \lambda \sum_{g \in \mathcal{G}} d_g \|\mathbf{v}^g\|,$$

which by (15) is equivalent to

$$\min_{\tilde{\mathbf{v}} \in \mathbb{R}^{\sum_{g \in \mathcal{G}} |g|}} \tilde{L}(\tilde{\mathbf{X}}\tilde{\mathbf{v}}) + \lambda \sum_{g \in \mathcal{G}} d_g \|\tilde{\mathbf{v}}_g\|, \quad (17)$$

where $\tilde{\mathbf{X}}$ is the $n \times (\sum_{g \in \mathcal{G}} |g|)$ matrix of duplicated training points, and $\tilde{\mathbf{v}}_g$ refers to the restriction of $\tilde{\mathbf{v}}$ to the coordinates of group g . In other words, we have eliminated \mathbf{w} from the optimization problem and reformulated it as a simple group Lasso problem without overlap between groups in an expanded space of size $\sum_{g \in \mathcal{G}} |g|$.

On the example of Figure 1, with 3 overlapping groups, this duplication trick can be rewritten as follows :

$$\mathbf{X}\mathbf{w} = \mathbf{X} \cdot \begin{array}{|c|} \hline \tilde{\mathbf{v}}^1 \\ \hline 0 \\ \hline \end{array} + \mathbf{X} \cdot \begin{array}{|c|} \hline 0 \\ \hline \tilde{\mathbf{v}}^2 \\ \hline 0 \\ \hline \end{array} + \mathbf{X} \cdot \begin{array}{|c|} \hline 0 \\ \hline 0 \\ \hline \tilde{\mathbf{v}}^3 \\ \hline \end{array} = (\mathbf{X}_{g_1}, \mathbf{X}_{g_2}, \mathbf{X}_{g_3}) \cdot \begin{array}{|c|} \hline \tilde{\mathbf{v}}_1 \\ \hline \tilde{\mathbf{v}}_2 \\ \hline \tilde{\mathbf{v}}_3 \\ \hline \end{array} \triangleq \tilde{\mathbf{X}}\tilde{\mathbf{v}}. \quad (18)$$

This formulation as a classical group Lasso problem in an expanded space has several implications, detailed in the next two sections. On the one hand, it allows to extend the penalty to non-linear functions by considering infinite-dimensional duplicated spaces endowed with

positive definite kernels (Section 4.7). On the other hand, it leads to straightforward implementations by borrowing classical group Lasso implementations after feature duplications (Section 4.8). Note, however, that the theoretical results we will show in Section 7, on the consistency of the estimator proposed, are not mere consequences of existing results for the classical group Lasso, because, in the case we consider, not only is the design matrix $\tilde{\mathbf{X}}$ rank deficient, but so are all of its restriction to sets of variables corresponding to any union of overlapping groups.

4.7 Multiple Kernel Learning formulations

Given the reformulation in a duplicated variable space presented above, we provide in this section a multiple kernel learning (MKL) interpretation to the regularization by our norm and show that it extends naturally the case with disjoint groups.

To introduce it, we return first to the concept of MKL (Bach et al., 2004; Lanckriet et al., 2004) which we can present as follows. If one considers a learning problem of the form

$$H = \min_{\mathbf{w} \in \mathbb{R}^p} \tilde{L}(\mathbf{X}\mathbf{w}) + \frac{\lambda}{2} \|\mathbf{w}\|^2, \quad (19)$$

then by the representer theorem the optimal value of the objective H only depends on the input data \mathbf{X} through the Gram matrix $\mathbf{K} = \mathbf{X}\mathbf{X}^\top$, which therefore can be replaced by any positive definite (p.d.) kernel between the datapoints. Moreover H can be shown to be a convex function of \mathbf{K} (Lanckriet et al., 2004). Given a collection of p.d. kernels $\mathbf{K}_1, \dots, \mathbf{K}_k$, any convex combination $\mathbf{K} = \sum_{i=1}^k \eta_i \mathbf{K}_i$ with $\eta_i \geq 0$ and $\sum_i \eta_i = 1$ is itself a p.d. kernel. The multiple kernel learning problem consists in finding the best such combination in the sense of minimizing H :

$$\min_{\boldsymbol{\eta} \in \mathbb{R}_+^k} H(\sum_i \eta_i \mathbf{K}_i) \quad \text{s.t.} \quad \sum_i \eta_i = 1. \quad (20)$$

The kernels considered in the linear combination above are typically reproducing kernels associated with different reproducing kernel Hilbert spaces (RKHS).

Bach et al. (2004) showed that problems regularized by a squared ℓ_1/ℓ_2 -norm and multiple kernel learning were intrinsically related. More precisely he shows that, if \mathcal{G} forms a partition of $\{1, \dots, p\}$, letting problems (P) and (P') be defined through

$$(P) \min_{\mathbf{w} \in \mathbb{R}^p} \tilde{L}(\mathbf{X}\mathbf{w}) + \frac{\lambda}{2} (\sum_{g \in \mathcal{G}} d_g \|\mathbf{w}_g\|)^2 \quad \text{and} \quad (P') \min_{\boldsymbol{\eta} \in \mathbb{R}_+^m} H(\sum_{g \in \mathcal{G}} \eta_g \mathbf{K}_g) \quad \text{s.t.} \quad \sum_{g \in \mathcal{G}} d_g^2 \eta_g = 1,$$

with $\mathbf{K}_g = \mathbf{X}_g \mathbf{X}_g^\top$, then (P) and (P') are equivalent in the sense that the optimal values of both objectives are equal with a bijection between the optimal solutions. Note that such an equivalence does not hold if the groups $g \in \mathcal{G}$ overlap.

Now turning to the norm we introduced, using the same derivation as the one leading from problem (16) to problem (17), we can show that minimizing $\tilde{L}(\mathbf{X}\mathbf{w}) + \frac{\lambda}{2} \Omega(\mathbf{w})^2$ w.r.t. \mathbf{w} is equivalent to minimizing $\tilde{L}(\tilde{\mathbf{X}}\tilde{\mathbf{v}}) + \frac{\lambda}{2} (\sum_g \|\mathbf{v}^g\|)^2$ and setting $\mathbf{w} = \sum_{g \in \mathcal{G}} \mathbf{v}^g$. At this point, the result of Bach et al. (2004) applied to the latter formulation in the space of duplicates shows that it is equivalent to the multiple kernel learning problem

$$\min_{\boldsymbol{\eta} \in \mathbb{R}_+^m} H(\sum_{g \in \mathcal{G}} \eta_g \mathbf{K}_g) \quad \text{s.t.} \quad \sum_{g \in \mathcal{G}} d_g^2 \eta_g = 1, \quad \text{with} \quad \mathbf{K}_g = \mathbf{X}_g \mathbf{X}_g^\top. \quad (21)$$

This shows that minimizing $\tilde{L}(\mathbf{X}\mathbf{w}) + \frac{\lambda}{2}\Omega(\mathbf{w})^2$ is equivalent to the MKL problem above. Compared with the original result of [Bach et al. \(2004\)](#), it should be noted now that, because of the duplication mechanism implicit in our norm, the original sets $g \in \mathcal{G}$ are no longer required to be disjoint. In fact this derivation shows that, in some sense, the norm we introduced is the one that corresponds to the most natural extension of multiple kernel learning to the case of overlapping groups.

Conversely, it should be noted that, while one of the application of multiple kernel learning is *data fusion* and thus allows to combine kernels corresponding to functions of intrinsically different input variables, MKL can also be used to select and combine elements from different function spaces defined on the same input. In general these function spaces are not orthogonal and are typically not even disjoint. In that case the MKL formulation corresponds implicitly to using the norm presented in this paper.

Finally, another MKL formulation corresponding to the norm is possible. If we denote $\mathbf{K}_i = \mathbf{X}_i\mathbf{X}_i^\top$ the rank one kernel corresponding to the i th feature, then we can write $\mathbf{K}_g = \sum_{i \in g} \mathbf{K}_i$. If $\mathbf{B} \in \mathbb{R}^{p \times m}$ is the binary matrix defined by $\mathbf{B}_{ig} = \mathbf{1}_{\{i \in g\}}$, and $Z = \{\mathbf{B}\boldsymbol{\eta} \mid \boldsymbol{\eta} \in \mathbb{R}_+^m, \sum_{g \in \mathcal{G}} \eta_g = 1\}$ is the image of the canonical simplex of \mathbb{R}^m by the linear transformation associated with \mathbf{B} , then with $\boldsymbol{\zeta} \in Z$ obtained through $\zeta_i = \sum_{g \ni i} \eta_g$, the MKL problem above can be reformulated as

$$\min_{\boldsymbol{\zeta} \in Z} H\left(\sum_{i=1}^p \zeta_i \mathbf{K}_i\right). \quad (22)$$

This last formulation can be viewed as the structured MKL formulation associated with the norm Ω (see [Bach et al., 2011](#), sec. 1.5.4). It is clearly more interesting computationally when $m \gg p$. It is however restricted to a particular form of kernel \mathbf{K}_g for each group, which has to be a sum of feature kernels \mathbf{K}_i . In particular, it doesn't allow for interactions among features in the group.

In the two formulations above, it is obviously possible to replace the linear kernel used for the derivation by a non-linear kernel. In the case of (21) the combinatorial structure of the problem is a priori lost in the sense that the different kernels are no longer linear combinations of a set of "primary" kernels, while this is still the case for (22).

Using non-linear kernels like RBF, or kernels on discrete structures such as sequence- or graph-kernels may prove useful in cases where the relationship between the covariates in the groups and the output is expected to be non-linear. For example if g is a group of genes and the coexpression patterns of genes within the group are associated with the output, the group will be deemed important by a non linear kernel while a linear one may miss it. More generally, it allows for structured non-linear feature selection.

4.8 Algorithms

There are several possible algorithmic approaches to solve the optimization problem (13), depending on the structure of the groups in \mathcal{G} . The approach we chose in this paper is based on the reformulation by *covariate duplication* of section 4.6, and applies an algorithm for the group Lasso in the space of duplicates. To be specific, for the experiments presented in section 10, we implemented the block-coordinate descent algorithm of [Meier et al. \(2008\)](#) combined with the working set strategy proposed by [Roth and Fischer \(2008\)](#). Note that

the covariate duplication of the input matrix X needs not to be done explicitly in computer memory, since only fast access to the corresponding entries in X is required. Only the vector \tilde{v} which is optimized has to be stored in the duplicated space $\mathbb{R}^{\sum_{g \in \mathcal{G}} |g|}$ and is potentially of large dimension (although sparse) if \mathcal{G} has many groups.

Alternatively, efficient algorithms which do not require working in the space of duplicated covariates are possible. Such an algorithm was proposed by Mosci et al. (2010) who suggested to use a proximal algorithm, and to compute the proximal operator of the norm Ω via an approximate projection on the unit ball of the dual norm in the input space. To avoid duplication, it would also be possible to use an approach similar to that of (Rakotomamonjy et al., 2008). Finally, one could also consider algorithms from the multiple kernel learning literature.

5. Group-support

A natural question associated with the norm Ω is what sparsity pattern are elicited when the norm is used as a regularizer. This question is natural in the context of support recovery. If the groups are disjoint, one could equivalently ask which patterns of selected group are possible, since answering the latter or the former questions are equivalent. This suggest a view in which the support is expressed in terms of groups. We formalize this idea through the concept of group-support of a vector \mathbf{w} , which, put informally, is the set of groups that are non-zero in a decomposition of \mathbf{w} . We will see that this notion is useful to characterize induced decompositions and recovery properties of the norm.

5.1 Definitions

More formally, we naturally call *group-support* of a decomposition $\bar{\mathbf{v}} \in \mathcal{V}_{\mathcal{G}}$, the set of groups g such that $\mathbf{v}^g \neq \mathbf{0}$. We extend this definition to a vector as follows:

Definition 12 (Strong group-support) *The strong group-support $\check{\mathcal{G}}_1(\mathbf{w})$ of a vector $\mathbf{w} \in \mathbb{R}^p$ is the union of the group-supports of all its optimal decompositions, namely:*

$$\check{\mathcal{G}}_1(\mathbf{w}) \triangleq \{g \in \mathcal{G} \mid \exists \bar{\mathbf{v}} \in \mathbf{V}(\mathbf{w}) \text{ s.t. } \mathbf{v}^g \neq \mathbf{0}\}.$$

If \mathbf{w} has a unique decomposition $\bar{\mathbf{v}}(\mathbf{w})$, then $\check{\mathcal{G}}_1(\mathbf{w}) = \{g \in \mathcal{G} \mid \mathbf{v}^g(\mathbf{w}) \neq \mathbf{0}\}$ is the group-support of its decomposition. We also define a notion of *weak group-support* in terms of uniqueness of the optimal dual variables.

Definition 13 (Weak group-support) *The weak group-support of a vector $\mathbf{w} \in \mathbb{R}^p$ is*

$$\mathcal{G}_1(\mathbf{w}) \triangleq \{g \in \mathcal{G} \mid \exists \alpha_g \in \mathbb{R}^p \text{ s.t. } \Pi_g \mathcal{A}(\mathbf{w}) = \{\alpha_g\} \text{ and } \|\alpha_g\| = d_g\}.$$

It follows immediately from Lemma 9 that $\check{\mathcal{G}}_1(\mathbf{w}) \subset \mathcal{G}_1(\mathbf{w})$. When $\check{\mathcal{G}}_1(\mathbf{w}) = \mathcal{G}_1(\mathbf{w})$, we refer to $\check{\mathcal{G}}_1(\mathbf{w})$ as the group-support of \mathbf{w} ; otherwise we say that the group-support is ambiguous.

The definitions of *strong group-support* and *weak group-support* are motivated by the fact that in the variational formulation (8), the *strong group-support* is the set of groups for which the constraints $\|\alpha_g\| \leq 1$ are *strongly active* whereas the *weak group-support* is the set of *weakly* or *strongly active* such constraints (Nocedal and Wright, 2006, p.342). We illustrate these two notions on a few examples in Section 6.

5.2 Supports induced by the group-support

For any $\mathbf{w} \in \mathbb{R}^p$, we denote by $J_1(\mathbf{w})$ (resp. $\check{J}_1(\mathbf{w})$) the set of variables in groups of the weak group-support (resp. strong group-support):

$$J_1(\mathbf{w}) \triangleq \bigcup_{g \in \mathcal{G}_1(\mathbf{w})} g \quad \text{and} \quad \check{J}_1(\mathbf{w}) \triangleq \bigcup_{g \in \check{\mathcal{G}}_1(\mathbf{w})} g.$$

Since $\check{\mathcal{G}}_1(\mathbf{w}) \subset \mathcal{G}_1(\mathbf{w})$, it immediately follows that $\check{J}_1(\mathbf{w}) \subset J_1(\mathbf{w})^2$. The following two lemmas show that, on $J_1(\mathbf{w})$, any dual variables $\alpha \in \mathcal{A}(\mathbf{w})$ are uniquely determined.

Lemma 14 *If $J_1(\mathbf{w}) \setminus \check{J}_1(\mathbf{w}) \neq \emptyset$, then for any $\alpha \in \mathcal{A}(\mathbf{w})$, $\alpha_{J_1(\mathbf{w}) \setminus \check{J}_1(\mathbf{w})} = \mathbf{0}$.*

Proof Note that $\mathbf{w}_{J_1(\mathbf{w}) \setminus \check{J}_1(\mathbf{w})} = \mathbf{0}$ since $\mathbf{v}^g = \mathbf{0}$ for $g \in \mathcal{G}_1(\mathbf{w}) \setminus \check{\mathcal{G}}_1(\mathbf{w})$. Let $g \in \mathcal{G}_1(\mathbf{w}) \setminus \check{\mathcal{G}}_1(\mathbf{w})$. If $g \setminus \check{J}_1(\mathbf{w}) \neq \emptyset$, and if $\Pi_{g \setminus \check{J}_1(\mathbf{w})} \mathcal{A}(\mathbf{w}) \neq \{\mathbf{0}\}$ then, let $i \in g \setminus \check{J}_1(\mathbf{w})$ such that there exists $\alpha \in \mathcal{A}(\mathbf{w})$ with $\alpha_i \neq 0$. Setting $\alpha_i = 0$ leads to another vector that solves the second variational formulation (7) and such that $\|\alpha_g\| < d_g$ which contradicts the hypothesis that $g \in \mathcal{G}_1(\mathbf{w})$. \blacksquare

Lemma 15 *For any $\mathbf{w} \in \mathbb{R}^p$, $\Pi_{J_1(\mathbf{w})} \mathcal{A}(\mathbf{w})$ is a singleton, i.e., there exists $\alpha_{J_1(\mathbf{w})}(\mathbf{w}) \in \mathbb{R}^{|J_1(\mathbf{w})|}$ such that, for all $\alpha' \in \mathcal{A}(\mathbf{w})$, $\alpha'_{J_1(\mathbf{w})} = \alpha_{J_1(\mathbf{w})}(\mathbf{w})$.*

Proof By definition of $\check{J}_1(\mathbf{w})$, for all $i \in \check{J}_1(\mathbf{w})$ there exists at least one $\mathbf{v} \in \mathbf{V}(\mathbf{w})$ and one group $g \ni i$, such that $(\mathbf{v}^g)_i \neq 0$. Now as a consequence of Lemma 9, for any two solutions $\alpha, \alpha' \in \mathcal{A}(\mathbf{w})$, we have that $\alpha_g = \alpha'_g = d_g \frac{\mathbf{v}^g}{\|\mathbf{v}^g\|}$, so in particular $\alpha_i = \alpha'_i$. For $i \in J_1(\mathbf{w}) \setminus \check{J}_1(\mathbf{w})$, Lemma 14 shows that $\alpha_i = 0$. \blacksquare

6. Illustrative examples

In this section, we consider a few examples that illustrate some of the properties of Ω , namely situations where weak and strong group support differ, or where there is an entire set of optimal decompositions. We will abuse notations and write \mathbf{v}_g for \mathbf{v}_g^g when writing explicit decompositions. We will denote by Sign the correspondence (or set-valued function) defined by $\text{Sign}(x) = 1$ if $x > 0$, $\text{Sign}(x) = -1$ if $x < 0$ and $\text{Sign}(0) = [-1, 1]$.

6.1 Two overlapping groups

We first consider the case $p = 3$ and $\mathcal{G} = \{\{1, 2\}, \{2, 3\}\}$.

Lemma 16 *We have $\Omega(\mathbf{w}) = \|(w_2, |w_1| + |w_3|)^\top\|$. If $(w_1, w_3) \neq \mathbf{0}$, the optimal decomposition is unique with*

$$\mathbf{v}_{\{12\}} = \left(w_1, \frac{|w_1|}{|w_1| + |w_3|} w_2 \right)^\top \quad \text{and} \quad \mathbf{v}_{\{23\}} = \left(\frac{|w_3|}{|w_1| + |w_3|} w_2, w_3 \right)^\top, \quad (23)$$

2. It is possible to have $\check{J}_1(\mathbf{w}) \neq J_1(\mathbf{w})$ consider $\mathcal{G} = \{\{1, 2\}, \{1, 3\}, \{2, 3, 4\}\}$ and $\mathbf{w} = \frac{1}{\sqrt{2}}(1, \mu, 1 - \mu, 0)$ for any $\mu \in (0, 1)$. We then have $\check{\mathcal{G}}_1 = \{\{1, 2\}, \{1, 3\}\}$ and $\mathcal{G}_1 = \mathcal{G}$ so that $\check{J}_1 = \{1, 2, 3\} \neq J_1 = \{1, 2, 3, 4\}$.

$$\mathcal{A}(\mathbf{w}) = \{ ((|w_1|+|w_3|) \gamma_1, w_2, (|w_1|+|w_3|) \gamma_3) / \Omega(\mathbf{w}) \mid \gamma_i \in \text{Sign}(w_i), i \in \{1, 3\} \},$$

$J_1 = \check{J}_1 = \text{supp}(\mathbf{w})$ and $\mathcal{G}_1 = \check{\mathcal{G}}_1$ includes $\{w_1, w_2\}$ if $w_1 \neq 0$ and $\{w_2, w_3\}$ if $w_3 \neq 0$. If $(w_1, w_3) = \mathbf{0}$, then $\mathbf{v}_{\{12\}} = (0, \gamma w_2)^\top$ and $\mathbf{v}_{\{23\}} = ((1 - \gamma) w_2, 0)^\top$ is an optimal decomposition for any $\gamma \in [0, 1]$, $\mathcal{A}(\mathbf{w}) = \{(0, \text{sign}(w_2), 0)\}$, $J_1 = \check{J}_1 = \{1, 2, 3\}$ and $\mathcal{G}_1 = \check{\mathcal{G}}_1 = \mathcal{G}$.

We prove this lemma in section C.1.1 (as a special case of the “cycle of length three” which we consider next). Here, the case where the decomposition is not unique seems to be a relatively pathological case where the true support is included in the intersection of two groups. However, note that the weak group-support and strong-group support coincide, even in the latter case.

6.2 Cycle of length 3

We now turn to the case $p = 3$ and $\mathcal{G} = \{\{1, 2\}, \{2, 3\}, \{1, 3\}\}$. Note that if at least one of the groups is not part of the weak-group support, we fall back on the case of two overlapping groups. We therefore have the following lemma:

Lemma 17 Define $\mathcal{W}_{bal} \triangleq \{\mathbf{w} \in \mathbb{R}^3 \mid |w_i| \leq \|\mathbf{w}_{\{i\}^c}\|_1, i \in \{1, 2, 3\}\}$. We have

$$\Omega_{\cup}^{\mathcal{G}}(\mathbf{w}) = \begin{cases} \frac{1}{\sqrt{2}} \|\mathbf{w}\|_1 & \text{if } \mathbf{w} \in \mathcal{W}_{bal} \\ \min_{i \in \{1, 2, 3\}} \|(w_i, \|\mathbf{w}_{\{i\}^c}\|_1)\| & \text{else.} \end{cases}$$

If $|\text{supp}(\mathbf{w})| \neq 1$ the optimal decomposition is unique. If in addition, $\mathbf{w} \in \mathcal{W}_{bal}$ we have for $(i, j, k) \in \{(1, 2, 3), (2, 3, 1), (3, 1, 2)\}$:

$$\mathbf{v}_{\{ij\}} = \frac{1}{2} (|w_i| + |w_j| - |w_k|) \begin{pmatrix} \text{sign}(w_i) \\ \text{sign}(w_j) \end{pmatrix} \quad \text{and} \quad \mathcal{A}(\mathbf{w}) = \left\{ \frac{1}{\sqrt{2}} (\text{sign}(w_1), \text{sign}(w_2), \text{sign}(w_3)) \right\}.$$

Moreover, we have $J_1 = \check{J}_1 = \{1, 2, 3\}$, $\mathcal{G}_1 = \mathcal{G}$ and for $\mathbf{w} \in \mathcal{W}_{bal}$, $\mathcal{G}_1 = \check{\mathcal{G}}_1 = \mathcal{G}$.

We prove this lemma in appendix C.1, and illustrate it on Figure 3 with the unit ball of the obtained norm. In this case it is interesting to note that the group-support (weak or strong) is not necessarily a *minimal cover*, where we say that a set of groups provides a minimal cover if it is impossible to remove a group while still covering the support. For instance, for \mathbf{w} in the interior of \mathcal{W}_{bal} , the group-support contains all three groups, while the support is covered by any two groups. This is clearly a consequence of the convexity of the formulation. The cycle of length 3 is also interesting because, for any \mathbf{w} on the boundary of \mathcal{W}_{bal} , the weak and strong group-support do not coincide, as illustrated on Figure 3 (right). Indeed if for example $|w_3| = |w_1| + |w_2|$, then $\mathbf{v}_{\{1,2\}} = (0, 0)^\top$, $\mathbf{v}_{\{1,3\}} = |w_1|(\text{sign}(w_1), \text{sign}(w_3))^\top$ and $\mathbf{v}_{\{2,3\}} = |w_2|(\text{sign}(w_2), \text{sign}(w_3))^\top$ so that by lemma 9 the dual variable satisfies $\|\alpha_{\{1,2\}}\| = 1$, which means that $\{1, 2\}$ is in the weak but not in the strong group-support.

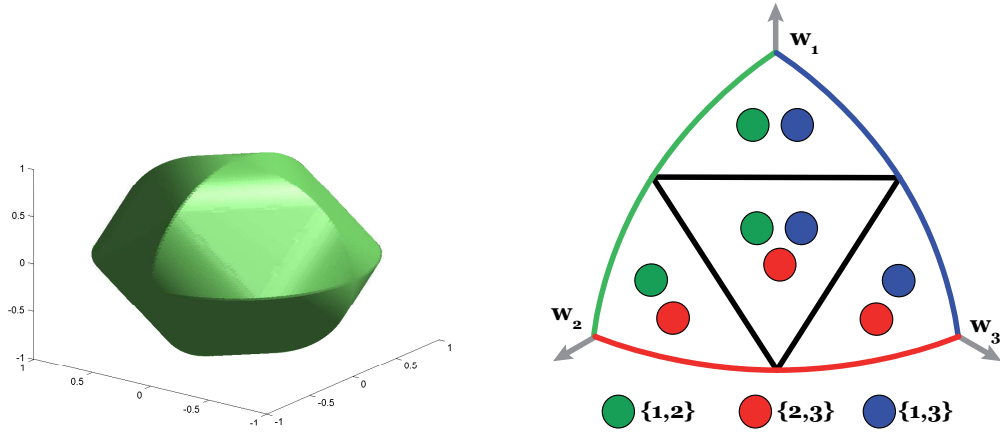


Figure 3: (Left) The unit ball of Ω for the groups $\{1, 2\}, \{1, 3\}, \{2, 3\}$ in \mathbb{R}^3 . (Right) a diagram that represents the restriction of the unit ball to the positive orthant. The black lines separate the surface in four regions. The triangular central region is \mathcal{W}_{bal} . On the interior of each region and on the colored outer boundaries, the group-support is constant, non-ambiguous (i.e., the weak and strong group-supports coincide) and represented by color bullets or the color of the edge, with one color associated to each group. On the boundary of \mathcal{W}_{bal} , the black lines indicate the group-support is ambiguous, the weak group-support containing all three groups, and the strong group-support being equal to that of the outer adjacent region for each black segment.

6.3 Cycle of length 4

We consider the case $p = 4$ and show the following result in appendix C.2.

Lemma 18 *For $\mathcal{G} = \{\{1, 2\}, \{1, 3\}, \{2, 4\}, \{3, 4\}\}$. Ω has the closed form*

$$\Omega(\mathbf{w}) = \|(|w_1| + |w_4|, |w_2| + |w_3|)\|_2.$$

However, if $|\text{supp}(\mathbf{w})| = 4$, the optimal decomposition is never unique.

This suggests that for a general \mathcal{G} , unique solutions are the exception rather than the rule. This motivates a posteriori definitions of group-support that are meaningful in the case where the decomposition is not unique. We consider a necessary and sufficient condition for uniqueness in lemma 48.

7. Model selection consistency

In this section we consider the estimator $\hat{\mathbf{w}}$ obtained as a solution of the learning problem (13) in the context of a well-specified model. Specifically, we consider the linear regression model:

$$\mathbf{y} = \mathbf{X}\mathbf{w}^* + \boldsymbol{\varepsilon}, \quad (24)$$

where $\mathbf{X} \in \mathbb{R}^{n \times p}$ is a design matrix, $\mathbf{y} \in \mathbb{R}^p$ is the response vector and $\boldsymbol{\varepsilon} \in \mathbb{R}^p$ is a vector of i.i.d. random variables with mean 0 and finite variance. We denote by \mathbf{w}^* the true regression function, and by $\hat{\mathbf{w}}$ the one we estimate as the solution of the following optimization problem, which is a particular case of (13):

$$\min_{\mathbf{w} \in \mathbb{R}^p} \frac{1}{2n} \|\mathbf{y} - \mathbf{X}\mathbf{w}\|^2 + \lambda_n \Omega(\mathbf{w}). \quad (25)$$

Several types of consistency results are of interest when using a sparsity-inducing norm as a regularizer. One typically distinguishes *classical consistency* where $\|\hat{\mathbf{w}} - \mathbf{w}^*\|_p$ converges in probability to zero, *prediction consistency* where $|L(\hat{\mathbf{w}}) - L(\mathbf{w}^*)|$ converges to zero in probability, and *model selection consistency* or *support recovery* where the support of $\hat{\mathbf{w}}$ coincides with the support of \mathbf{w}^* with high probability. We are interested in the discussion of the last type of result, support recovery, for solutions of (25).

As compared with the Lasso and the group Lasso in the case of disjoint supports, the discussion of support recovery is complicated by several factors here. First, supports that can be recovered are not exactly the ones that can be expressed as unions of groups in \mathcal{G} : as the reader might expect, the appropriate notion of support is $J_1(\mathbf{w}^*)$ (or $\check{J}_1(\mathbf{w}^*)$), the one induced by the concept of group-support introduced in section 5. Second, by contrast with the situation of the group Lasso with disjoint groups, the identification of the support $J_1(\mathbf{w}^*)$ (or $\check{J}_1(\mathbf{w}^*)$) is not equivalent to the identification of the group-support $\mathcal{G}_1(\mathbf{w}^*)$ (or $\check{\mathcal{G}}_1(\mathbf{w}^*)$), the latter being now a harder problem. As a consequence one should distinguish support recovery from group-support recovery, and, depending on the context, the appropriate notion to consider for model selection consistency might be one or the other. Third, the group-support is characterized by properties of the set $\mathbf{V}(\hat{\mathbf{w}})$ whose convergence is less trivial to study than that of a vector. For these reasons, we consider only in this paper the classical asymptotic regime in which the model generating the data is of fixed

finite dimension p while $n \rightarrow \infty$. However we focus on the harder problem of *group-support recovery*, which will then imply *support recovery* results.

The proof of consistency we present below follows a classical proof scheme (Bach, 2008a). However the originality of our work reside in that we characterize the group-support consistency here, which requires in particular to study the convergence of the set-valued map $\mathbf{V}(\hat{\mathbf{w}})$. We therefore start in the next section by introducing appropriate notions of continuity for set-valued functions.

7.1 Correspondence theory to the rescue

We appeal to the theory of *correspondences* developed by Claude Berge at the end of the 1950's (Berge, 1959). In particular, we follow closely its presentation by Border (1985).

Definition 19 (correspondence) *A correspondence ϕ from a set X to a set Y , denoted $\phi : X \rightrightarrows Y$, is a set-valued mapping which to each element $x \in X$ associates a set $\phi(x) \subset Y$.*

When X and Y are metric spaces, the usual notion of continuity of a function is replaced for correspondences by the following notions:

Definition 20 (hemicontinuity and continuity) *Given two metric spaces (X, d) and (Y, ρ) , a correspondence $\phi : X \rightrightarrows Y$ is said to be upper hemicontinuous or u.h.c. (resp. lower hemicontinuous or l.h.c.) if for any point $x \in X$ and any open set $U \subset Y$ such that $\phi(x) \subset U$ (resp. $\phi(x) \cap U \neq \emptyset$) there exists a neighborhood V of x such that, for all $x' \in V$, $\phi(x') \subset U$ (resp. $\phi(x') \cap U \neq \emptyset$). A correspondence is said to be continuous if it is both upper and lower hemicontinuous.*

Note that a singleton valued correspondence ϕ can be identified with the function f taking this unique value, and that f is continuous if and only if ϕ is lower or upper hemicontinuous, both notions being equivalent in that case. The following results, which we prove in appendix A, are key to study the consistency of our method in the next section.

Lemma 21 $\mathbf{w} \mapsto \mathcal{A}(\mathbf{w})$ is an upper hemicontinuous correspondence.

Lemma 22 If $\text{supp}(\mathbf{w}) = J_1$, then, on the domain $\mathcal{D} = \{\mathbf{u} \in \mathbb{R}^p \mid \text{supp}(\mathbf{u}) = J_1\}$, $\mathbf{u} \mapsto \mathbf{V}(\mathbf{w} + \mathbf{u})$ is a continuous correspondence at $\mathbf{u} = \mathbf{0}$.

7.2 Group-support recovery

In this section, we state and prove our main consistency results for group-support and support recovery in the least-square linear regression framework (24). We consider two main hypotheses:

$$(H1) \quad \Sigma \triangleq \frac{1}{n} \mathbf{X}^\top \mathbf{X} \succ 0, \quad (H2) \quad \text{supp}(\mathbf{w}^*) = J_1(\mathbf{w}^*).$$

We denote $\mathcal{G}_2(\mathbf{w}^*) \triangleq \mathcal{G} \setminus \mathcal{G}_1(\mathbf{w}^*)$ and $J_2(\mathbf{w}^*) \triangleq [1, p] \setminus J_1(\mathbf{w}^*)$. For convenience, for any group of covariates g we note \mathbf{X}_g the $n \times |g|$ design matrix restricted to the covariates in g , and for any two groups g, g' we note $\Sigma_{gg'} = \frac{1}{n} \mathbf{X}_g^\top \mathbf{X}_{g'}$.

Consider the following two conditions, where we denote $J_1(\mathbf{w}^*)$ simply by J_1 for sake of clarity:

$$\forall g \in \mathcal{G}_2(\mathbf{w}^*), \quad \left\| \Sigma_{gJ_1} \Sigma_{J_1J_1}^{-1} \alpha_{J_1}(\mathbf{w}^*) \right\| \leq d_g, \quad (\text{C1})$$

$$\forall g \in \mathcal{G}_2(\mathbf{w}^*), \quad \left\| \Sigma_{gJ_1} \Sigma_{J_1J_1}^{-1} \alpha_{J_1}(\mathbf{w}^*) \right\| < d_g. \quad (\text{C2})$$

Theorem 23 *Under assumption (H1), for $\lambda_n \rightarrow 0$ and $\lambda_n n^{1/2} \rightarrow \infty$, conditions (C1) and (C2) are respectively necessary and sufficient for the strong group-support of the solution of (13), $\check{\mathcal{G}}_1(\hat{\mathbf{w}})$ to satisfy with probability tending to 1 as $n \rightarrow +\infty$:*

$$\check{\mathcal{G}}_1(\hat{\mathbf{w}}) \subset \mathcal{G}_1(\mathbf{w}^*).$$

Proof We follow the line of proof of Bach (2008a) but consider a fixed design for simplicity of notations. Let us first consider the subproblem of estimating a vector only on the support of \mathbf{w}^* by using only the groups in $\mathcal{G}_1(\mathbf{w}^*)$ in the penalty, *i.e.*, consider $\mathbf{w}_1 \in \mathbb{R}^{J_1}$ a solution of

$$\min_{\mathbf{w}_{J_1} \in \mathbb{R}^{J_1}} \frac{1}{2n} \|\mathbf{y} - \mathbf{X}_{J_1} \mathbf{w}_{J_1}\|^2 + \lambda_n \Omega_{\cup}^{\mathcal{G}_1(\mathbf{w}^*)}(\mathbf{w}_{J_1}).$$

By standard arguments, we can prove that \mathbf{w}_1 converges in Euclidean norm to \mathbf{w}^* restricted to J_1 as n tends to infinity (Knight and Fu, 2000). In the rest of the proof we show how to construct a vector $\mathbf{w} \in \mathbb{R}^p$ from \mathbf{w}_1 which under condition (C2) is with high probability a solution to (25). By adding null components to \mathbf{w}_1 , we obtain a vector $\mathbf{w} \in \mathbb{R}^p$ whose support is also J_1 , and $\mathbf{u} = \mathbf{w} - \mathbf{w}^*$ therefore satisfies $\text{supp}(\mathbf{u}) \subset J_1$. A direct computation of the gradient of the loss $L(\mathbf{w}) = \frac{1}{2n} \|\mathbf{y} - \mathbf{X}\mathbf{w}\|^2$ gives $\nabla L(\mathbf{w}) = \Sigma \mathbf{u} - \mathbf{q}$, where $\mathbf{q} = \frac{1}{n} \mathbf{X}^\top \boldsymbol{\varepsilon}$. From this we deduce that $\mathbf{u}_{J_1} = \Sigma_{J_1J_1}^{-1} (\nabla_{J_1} L(\mathbf{w}) + \mathbf{q}_{J_1})$, and since, by Lemma 11, $-\nabla_{J_1} L(\mathbf{w}) \in \lambda_n \Pi_{J_1} \mathcal{A}(\mathbf{w})$, there exists $\alpha_{J_1} \in \Pi_{J_1} \mathcal{A}(\mathbf{w})$ such that we have

$$\nabla_{J_2} L(\mathbf{w}) = \Sigma_{J_2J_1} \mathbf{u} - \mathbf{q}_{J_2} = \Sigma_{J_2J_1} \Sigma_{J_1J_1}^{-1} (-\lambda_n \alpha_{J_1} + \mathbf{q}_{J_1}) - \mathbf{q}_{J_2}.$$

To show that \mathbf{w} is a feasible solution to (25) it is enough to show that $\forall g \in \mathcal{G}_2(\mathbf{w}^*), \|\nabla_g L(\mathbf{w})\| \leq \lambda_n d_g$. But since the noise has bounded variance,

$$\Sigma_{J_2J_1} \Sigma_{J_1J_1}^{-1} \mathbf{q}_{J_1} - \mathbf{q}_{J_2} = \frac{1}{n} \mathbf{X}_{J_2}^\top \left[\frac{1}{n} \mathbf{X}_{J_1} \Sigma_{J_1J_1}^{-1} \mathbf{X}_{J_1}^\top - I \right] \boldsymbol{\varepsilon}$$

is \sqrt{n} -consistent, and by the union bound we get $\mathcal{P}(\forall g \in \mathcal{G}_2(\mathbf{w}^*), \|\nabla_g L(\mathbf{w})\| \leq \lambda_n d_g) \geq 1 - \sum_{g \in \mathcal{G}_2(\mathbf{w}^*)} \mathcal{P}(\|\nabla_g L(\mathbf{w})\| > \lambda_n d_g)$. We therefore deduce that, for any $g \in \mathcal{G}_2(\mathbf{w}^*)$,

$$\frac{1}{\lambda_n} \|\nabla_g L(\mathbf{w})\| \leq \left\| \Sigma_{gJ_1} \Sigma_{J_1J_1}^{-1} \alpha_{J_1} \right\| + \mathcal{O}_p(\lambda_n^{-1} n^{-1/2}).$$

By Lemma 21, we have that $\Pi_{J_1} \mathcal{A}(\mathbf{w})$ is an upper hemicontinuous correspondence so that $\mathbf{w}_{J_1} \xrightarrow{\mathbb{P}} \mathbf{w}_{J_1}^*$ implies that

$$\max_{\alpha' \in \mathcal{A}(\mathbf{w})} \|\alpha'_{J_1} - \alpha_{J_1}(\mathbf{w}^*)\| \xrightarrow{\mathbb{P}} 0.$$

Since we chose λ such that $\lambda_n^{-1}n^{-1/2} \rightarrow 0$, we have

$$\frac{1}{\lambda_n} \|\nabla_g L(\mathbf{w})\| \leq \left\| \boldsymbol{\Sigma}_{gJ_1} \boldsymbol{\Sigma}_{J_1 J_1}^{-1} \boldsymbol{\alpha}_{J_1}(\mathbf{w}^*) \right\| + o_p(1).$$

This shows that, under (C2), \mathbf{w} is a feasible solution to (25) whose group-support is contained in $\mathcal{G}_1(\mathbf{w}^*)$, i.e., we have shown $\check{\mathcal{G}}_1(\hat{\mathbf{w}}) \subset \mathcal{G}_1(\mathbf{w}^*)$.

For the necessary condition, by contradiction, consider a solution supported on J_1 . Then, reusing the previous argument we have

$$\frac{1}{\lambda_n} \|\nabla_g L(\mathbf{w})\| \geq \left\| \boldsymbol{\Sigma}_{gJ_1} \boldsymbol{\Sigma}_{J_1 J_1}^{-1} \boldsymbol{\alpha}_{J_1}(\mathbf{w}^*) \right\| - o_p(1),$$

which shows that for the optimality conditions of Lemma 11(b) to hold, condition (C1) is necessary. \blacksquare

The previous theorem shows some partial consistency result in the sense that it guarantees that no group outside of the group-support will be selected. Since $\hat{\mathbf{w}}$ also converges with high probability in Euclidean norm to \mathbf{w}^* , this implies for the support that with high probability

$$\text{supp}(\mathbf{w}^*) \subset \text{supp}(\hat{\mathbf{w}}) \subset J_1(\mathbf{w}^*).$$

However, the theorem does not guarantee that all groups in $\check{\mathcal{G}}_1(\mathbf{w}^*)$ will be selected. This is not a shortcoming of the theorem: we provide an example in Appendix B which shows that it is possible that $\check{\mathcal{G}}_1(\hat{\mathbf{w}}) \subsetneq \check{\mathcal{G}}_1(\mathbf{w}^*)$ with probability 1. Nonetheless, we also show in the same appendix that with high probability there exists $\bar{\mathbf{v}}^* \in \mathbf{V}(\mathbf{w}^*)$ whose group-support is included in $\check{\mathcal{G}}_1(\hat{\mathbf{w}})$.

Theorem 24 *With assumptions (H1,H2) and for $\lambda_n \rightarrow 0$ and $\lambda_n n^{1/2} \rightarrow \infty$, condition (C1) is sufficient for the strong group-support of the solution of (25), $\check{\mathcal{G}}_1(\hat{\mathbf{w}})$, to satisfy with high probability:*

$$\check{\mathcal{G}}_1(\mathbf{w}^*) \subset \check{\mathcal{G}}_1(\hat{\mathbf{w}}) \subset \mathcal{G}_1(\mathbf{w}^*).$$

Proof The previous theorem shows that (C1) implies, with high probability, $\check{\mathcal{G}}_1(\hat{\mathbf{w}}) \subset \mathcal{G}_1(\mathbf{w}^*)$. However, by Lemma 22, we have that hypothesis (H2) guarantees that $\mathbf{w} \mapsto \mathbf{V}(\mathbf{w})$ is continuous at \mathbf{w}^* for \mathbf{w} with $\text{supp}(\mathbf{w}) \subset J_1(\mathbf{w}^*)$. Combined with the fact that $\hat{\mathbf{w}}$ converges in probability with \mathbf{w}^* , this implies that $\forall \epsilon > 0, \exists n_0, \forall n > n_0$, with probability larger than $1 - \epsilon$, $\forall \bar{\mathbf{v}}^* \in \mathbf{V}(\mathbf{w}^*)$, there exists $\bar{\mathbf{v}} \in \mathbf{V}(\hat{\mathbf{w}})$ such that $\|\bar{\mathbf{v}} - \bar{\mathbf{v}}^*\| < \epsilon$. For each $g \in \check{\mathcal{G}}_1(\mathbf{w}^*)$, for $\bar{\mathbf{v}}^* \in \mathbf{V}(\mathbf{w}^*)$ such that $\mathbf{v}^{*g} \neq 0$, there thus exists $\epsilon > 0$ such that the previous convergence results implies that $g \in \check{\mathcal{G}}_1(\hat{\mathbf{w}})$ with high probability. Finally, since $|\check{\mathcal{G}}_1(\mathbf{w}^*)|$ is finite, for n large enough, the union bound ensures that, with high probability, $\check{\mathcal{G}}_1(\mathbf{w}^*) \subset \check{\mathcal{G}}_1(\hat{\mathbf{w}})$. \blacksquare

The previous theorem shows the best result possible for the situation where $\check{\mathcal{G}}_1(\mathbf{w}^*) \neq \mathcal{G}_1(\mathbf{w}^*)$, as, in the example of the cycle of length 3 of section 6.2, the case of $\mathbf{w}^* = (2, 1, 1)$. If $\check{\mathcal{G}}_1(\mathbf{w}^*) = \mathcal{G}_1(\mathbf{w}^*)$, then we have the obvious corollary:

Corollary 25 *With assumptions (H1,H2), and assuming $\check{\mathcal{G}}_1(\mathbf{w}^*) = \mathcal{G}_1(\mathbf{w}^*)$, for $\lambda_n \rightarrow 0$ and $\lambda_n n^{1/2} \rightarrow \infty$, conditions (C1) and (C2) are respectively necessary and sufficient for the solution of (13) to estimate consistently the correct group-support $\mathcal{G}_1(\mathbf{w}^*)$.*

Remarks: For the Lasso and the usual group Lasso with disjoint groups, the most favorable case w.r.t. to condition (C2) is the case where the empirical covariance of the design is the identity (the same analysis can be done in the random design case), *i.e.*, the case where there is no correlations between groups. In that case, we have $\Sigma_{J_2 J_1} \Sigma_{J_1 J_1}^{-1} = 0$ and the mutual incoherence condition is 0. However, in the case of overlap, for $g \in \mathcal{G}$ such that $g \cap J_1 \neq \emptyset$, then $\Sigma_{g J_1} \Sigma_{J_1 J_1}^{-1} \neq 0$ and we have $\|\Sigma_{g J_1} \Sigma_{J_1 J_1}^{-1} \alpha_{J_1}\| = \|\alpha_{g \cap J_1}\|$. First, this gives yet another motivation to consider the weak-group support, since those groups in the weak-group support are exactly the ones for which $\|\alpha_{g \cap J_1}\| = 1$ (see Lemma 14). Second this show that if $g_1 \in \check{\mathcal{G}}_1(\mathbf{w}^*)$ and $g_2 \notin \mathcal{G}_1(\mathbf{w}^*)$ have a large overlap then $\|\alpha_{g_1 \cap g_2}\|$ can be fairly close to 1 even for a design with identity covariance. This means that it might be very difficult in practice to identify g_2 correctly as being outside of the support unless large amounts of data are available.

7.3 Related theoretical results

Two papers proposed recently some theoretical results on the estimator via regularization by Ω in the high-dimensional setting. Percival (2011) shows two types of results. First, he proposes a generalization of the restricted eigenvalue condition of Bickel et al. (2009) and generalize their proof to obtain fast-rate type of concentration results for the prediction error and convergence in ℓ_2 -norm. The bounds obtained scales as $\sqrt{B} \log(M)$, where M is the total number of groups and B is the largest group size. Then he considers an adaptive version of the regularization (in the sense of the adaptive Lasso) and shows for the resulting estimator a central limit theorem under high-dimensional scaling, under the conditions that the support is exactly a union of groups and that the decomposition of any point in a neighborhood of the optimum is unique. These results do not focus on support or group-support recovery. Also, it was one of our concerns to relax the assumption that the decomposition was unique or that the support was exactly a union of groups.

Maurer and Pontil (2011) give a bound on the Rademacher complexity of linear functions whose parameter vector lies in the unit ball of the norm $\Omega_{\mathcal{U}}^{\mathcal{G}}$, hence bounding the generalization error of such function. They consider as well extensions of this norm where each of the latent variables in the latent group Lasso are penalized by the norm of their image by some operator.

Our paper and these two papers have thus considered complementary aspects of estimation and recovery in statistical and compressed sensing based on $\Omega_{\mathcal{U}}^{\mathcal{G}}$ settings which should all contribute to understanding the high-dimensional learning setting.

8. Choice of the weights

The choice of the weights d_g associated to each group has been discussed in the literature on the classical group Lasso, when groups do not overlap. The main motivation for the introduction of these weights is to take into account the discrepancies of size existing between different groups. Yuan and Lin (2006) used $d_g = \sqrt{|g|}$, which yields solutions similar to the ANOVA test under a certain design. Bach et al. (2004) in the context of *multiple kernel learning* used $d_g \propto \sqrt{\text{tr} K_g}$, where $\{K_g\}_{g \in \mathcal{G}}$ are positive definite kernels, with $K_g = \mathbf{X}_g \mathbf{X}_g^{\top}$ in our context; for normalized features such as $\mathbf{X} \mathbf{X}^{\top} = I$, this yields $d_g = \sqrt{|g|}$ as well.

In the context of our latent group Lasso with overlapping groups, the choice of the weights is significantly more important than in the case of disjoint groups, and, arguably, than in the case of other formulations considering overlapping groups: indeed, the notions of group-support $\mathcal{G}_1(\mathbf{w})$ and $\check{\mathcal{G}}_1(\mathbf{w})$ and of support $J_1(\mathbf{w})$ and $\check{J}_1(\mathbf{w})$ associated to a vector \mathbf{w} through the norm $\Omega(\mathbf{w})$ themselves change according to the choice of the weights.

In this section we propose two types of arguments to study the effect of and guide the choice of weights:

- On the one hand we consider a vector \mathbf{w} and ask, independently of a learning problem, which groups participate in its group support: there is no point in introducing a group in \mathcal{G} if the weights are such that it can never be included in the group support. We show in Section 8.1 that, for all groups to be useful, weights should increase with the size of the groups, but not too quickly; in Section 8.2 we attempt to characterize when large groups are preferred over unions of smaller ones.
- On the other hand, we consider in Section 8.3 a simple regression scenario, and discuss the impact of the weights on the probability to correctly identify relevant groups, and simultaneously control the rate of false positives.

8.1 Redundant groups

Informally, we are concerned in this section with the fact that, if a group g contains a group h and d_g/d_h is too small, h will never enter the group support, and, conversely, if g is covered by a certain number of groups and d_g is too large, then g will never enter the group-support.

Formally, we say that a group $g \in \mathcal{G}$ is *redundant* for a certain set of weights $(d_g)_{g \in \mathcal{G}}$ if it can be removed without changing the value of the norm Ω for any \mathbf{w} ; this is equivalent to asking that the dual norm Ω^* is unchanged.

We first show that if there exists another group $g' \in \mathcal{G}$ such that $g \subset g'$, g is redundant unless we require that $d_g < d'_{g'}$:

Lemma 26 *If $g, g' \in \mathcal{G}$ satisfy $g \subset g'$ and $d_g \geq d_{g'}$, then for any \mathbf{w} , $(g \in \mathcal{G}_1(\mathbf{w})) \Rightarrow (g' \in \mathcal{G}_1(\mathbf{w}))$.*

Proof If $d_g \geq d_{g'}$, and if $g \in \mathcal{G}_1(\mathbf{w})$ then $1 = \frac{\|\alpha_g(\mathbf{w})\|}{d_g} \leq \frac{\|\alpha_{g'}(\mathbf{w})\|}{d_{g'}}$, which implies $g' \in \mathcal{G}_1(\mathbf{w})$. ■

It would be very natural to try and require that the weights are chosen so that, if $g = \text{supp}(\mathbf{w})$, its group-support is exactly g . Unfortunately, this is in general not possible: we show a negative result, which arises as a consequence of the previous lemma.

Lemma 27 *For some group sets \mathcal{G} , it is impossible to choose the weights d_g independently of \mathbf{w} so that $J_1(\mathbf{w}) = \text{supp}(\mathbf{w})$ (or $\check{J}_1(\mathbf{w}) = \text{supp}(\mathbf{w})$) if the latter is a union of groups.*

Proof Consider the groups $A = \{1, 2, 3\}$, $B = \{3, 4\}$, $C = \{2, 3, 4\}$:

- To have that $\check{J}_1(\mathbf{w}) = \text{supp}(\mathbf{w})$ for all \mathbf{w} Lemma 26 imposes that $d_B < d_C$ so that B is not redundant; this is necessary to have $\check{J}_1(\mathbf{w}) = \text{supp}(\mathbf{w}) = B$ for $\mathbf{w} = (0, 0, w, w)$.

- Then consider $\mathbf{w} = (0, w, \epsilon, \epsilon)$. $\check{J}_1(\mathbf{w}) = \text{supp}(\mathbf{w})$ requires that $\check{\mathcal{G}}_1(\mathbf{w}) = \{C\}$. But then $\mathbf{v}^C = \mathbf{w}$ so that $\boldsymbol{\alpha} = d_C \mathbf{w} / \|\mathbf{w}\|$. In particular $\|\boldsymbol{\alpha}_A\|^2 = d_C^2 (w^2 + \epsilon^2) / (w^2 + 2\epsilon^2)$ and $\|\boldsymbol{\alpha}_B\|^2 = d_C^2 2\epsilon^2 / (w^2 + 2\epsilon^2)$. For the inequality $\|\boldsymbol{\alpha}_A\| \leq d_A$ to hold for all $\epsilon > 0$, we need $d_A \geq d_C$.
- Finally consider $\mathbf{w} = (\epsilon, \epsilon, w, 0)$. Following the same line as for the previous case, $\check{J}_1(\mathbf{w}) = \text{supp}(\mathbf{w})$ requires that $\check{\mathcal{G}}_1(\mathbf{w}) = \{A\}$, which implies that $\mathbf{v}^A = \mathbf{w}$ so that $\boldsymbol{\alpha} = d_A \mathbf{w} / \|\mathbf{w}\|$. In particular $\|\boldsymbol{\alpha}_B\|^2 = d_A^2 w^2 / (w^2 + 2\epsilon^2)$ and $\|\boldsymbol{\alpha}_C\|^2 = d_A^2 (w^2 + \epsilon^2) / (w^2 + 2\epsilon^2)$. For the inequalities, $\|\boldsymbol{\alpha}_B\| \leq d_B$ and $\|\boldsymbol{\alpha}_C\| \leq d_C$ to hold for all $\epsilon > 0$, we need to have $d_A \leq d_B$.

These three inequalities are clearly incompatible and $\check{J}_1(\mathbf{w}) \subset J_1(\mathbf{w})$ which proves the result. \blacksquare

We now characterize more technically redundancy. The intuition behind the next lemma is the following geometric interpretation of the dual norm: the definition of Ω^* implies that its unit ball is the intersection of cylinders of the form $\{\boldsymbol{\alpha} \mid \|\boldsymbol{\alpha}_g\| \leq d_g\}$. This means that a group g is redundant if its associated cylinder contains the unit ball of the norm induced by the remaining groups. This can be formally stated as follows:

Lemma 28 *A group $g \in \mathcal{G}$ is not redundant if and only if there exists $\boldsymbol{\alpha} \in \mathbb{R}^p$ such that $\|\boldsymbol{\alpha}_g\| > d_g$ and $\forall h \in \mathcal{G} \setminus \{g\}, \|\boldsymbol{\alpha}_h\| \leq d_h$.*

Proof Define the unit balls: $\mathcal{U} = \{\boldsymbol{\alpha} \in \mathbb{R}^p \mid \forall h \in \mathcal{G}, \|\boldsymbol{\alpha}_h\| \leq d_h\}$ and $\mathcal{U}_g = \{\boldsymbol{\alpha} \in \mathbb{R}^p \mid \forall h \in \mathcal{G} \setminus \{g\}, \|\boldsymbol{\alpha}_h\| \leq d_h\}$. We have that g is redundant for Ω if and only if it is redundant for Ω^* , and the latter is true if and only if $\mathcal{U} = \mathcal{U}_g$. Since $\mathcal{U} \subset \mathcal{U}_g$, g is not redundant if and only if there exists $\boldsymbol{\alpha} \in \mathcal{U}_g \setminus \mathcal{U}$. \blacksquare

Corollary 29 *Let $g \in \mathcal{G}$ and $\mathcal{H} \subset \mathcal{G}$ such that g is covered by groups in \mathcal{H} , i.e., $g \subset \cup_{h \in \mathcal{H}} h$. Then g is redundant if $d_g^2 > \sum_{h \in \mathcal{H}} d_h^2$.*

Proof The fact that g is covered by groups in \mathcal{H} implies that, for any $\boldsymbol{\alpha} \in \mathbb{R}^p$, $\|\boldsymbol{\alpha}_g\|^2 \leq \sum_{h \in \mathcal{H}} \|\boldsymbol{\alpha}_h\|^2$. If g is part of the group-support, then necessarily $d_g^2 = \|\boldsymbol{\alpha}_g\|^2 \leq \sum_{h \in \mathcal{H}} \|\boldsymbol{\alpha}_h\|^2 \leq \sum_{h \in \mathcal{H}} d_h^2$. \blacksquare

In particular, if all singletons are part of \mathcal{G} with $d_{\{i\}} = 1$, $i \in [1, p]$, this imposes $d_g \leq \sqrt{|g|}$.

In the case where the weights depend only on the cardinality of the g , i.e., $d_g = d_k$ for $|g| = k$, we consider the following condition:

$$\forall k > 1, \quad d_{k-1} < d_k < \sqrt{\frac{k}{k-1}} d_{k-1}. \quad (\text{C})$$

Lemma 30 *Condition (C) is sufficient to guarantee that no group is redundant.*

Proof Assume that $(d_i)_{1 \leq i \leq m}$ satisfy condition (C), and let $g \in \mathcal{G}$ a group of cardinality k . Consider the vector $\alpha = \frac{d_k}{\sqrt{k}} \mathbf{1}_g$ with $\mathbf{1}_g \in \mathbb{R}^p$ the vector with entry i equal to 1 for $i \in g$ and 0 else. Since $|g| = k$ we have $\|\alpha_g\| = d_k$. Note that (C) implies $\frac{d_k}{\sqrt{k}} < \frac{d_{k-1}}{\sqrt{k-1}}$, which more generally implies by induction $\frac{d_k}{\sqrt{k}} < \frac{d_j}{\sqrt{j}}$ for any $j < k$. Now, for any group $g' \in \mathcal{G}$ of cardinality $j < k$, we have $\|\alpha_{g'}\| \leq \frac{d_k}{\sqrt{k}} \sqrt{j} < d_j$. Similarly, if $|g'| = j > k$ then $\|\alpha_{g'}\| \leq \|\alpha_g\| = d_k < d_j$, and if $g' \neq g$ but $|g'| = |g|$, then $\|\alpha_{g'}\| < \|\alpha_g\| = d_k = d_{g'}$. Since $\|\alpha_g\| = d_g$ and $\|\alpha_{g'}\| < d_{g'}$ for $g' \neq g$, it is possible to choose $\epsilon > 0$ sufficiently small such that the vector $\alpha' = \alpha + \epsilon \mathbf{1}_g$ satisfies $\|\alpha'_g\| > d_g$ and $\|\alpha'_{g'}\| < d_{g'}$ for any $g' \neq g$. Lemma 28 then shows that g is not redundant. ■

We would like insist that condition (C) is sufficient to guarantee non-redundancy but might be unnecessary for many restricted families of groups, for example as soon as each group contains an element which belongs to no other group. However, without any condition on the set of groups, the previous condition is the weakest possible if the weights depend only on the group sizes, since it becomes necessary in the following special case:

Lemma 31 *Assume that group g with cardinality $|g| = k$ contains all k groups of size $k-1$, then (C) is necessary for g to be non-redundant.*

Proof If $g \in \mathcal{G}$ is not redundant, by Lemma 28 we can find $\alpha \in \mathbb{R}^p$ such that $\|\alpha_g\| > d_g$ and $\|\alpha_h\| \leq d_h$ for $h \in \mathcal{G} \setminus \{g\}$. In particular, for all $i \in g$, $\|\alpha_{g \setminus \{i\}}\|^2 \leq d_{k-1}^2$ so that $(k-1)d_k^2 < (k-1)\|\alpha_g\|^2 = \sum_{i \in g} \|\alpha_{g \setminus \{i\}}\|^2 \leq k d_{k-1}^2$ which shows the result. ■

Condition (C) allows scalings of the weights which go from quasi uniform weights, in which case the larger groups dominate the smaller groups in the sense that they are preferably selected, to weights that scale like \sqrt{k} , in which case the smaller group dominate (and in particular if the singletons are included the norm approaches the ℓ_1 -norm). Condition (C) suggests to consider weights of the form $d_k = k^\gamma$, $\gamma \in (0, \frac{1}{2})$. We illustrate on Figure 4 the trade-offs obtained with the groups $\mathcal{G} = \{\{1\}, \{2\}, \{3\}, \{1, 2, 3\}\}$ and different γ . The first ball for $\gamma = 0$ is the ball we would have without considering the singletons since only the largest group is active. At the other extreme for $\gamma = \frac{1}{2}$ the ball is the one we would have without the $\{1, 2, 3\}$ group since only the singletons are active. In intermediate regimes, all the groups are active in some region. More specifically, the second ball for $\gamma = \frac{1}{4}$ corresponds to a limit case that we present in Section 8.3, while the third one for $\gamma = \frac{\log(2)}{2 \log(3)}$ illustrate another problem that we now introduce : the possibility that a group *dominates* other groups. Intuitively for $\gamma \geq \frac{\log(2)}{2 \log(3)}$, i.e., if the sphere gets any smaller than on the third ball, it becomes impossible to select a support of exactly two covariates even though (i) such a support would be a union of groups and (ii) no group is redundant. We detail this notion in the next section.

8.2 Dominating group

Let us first formalize the notion of group domination.

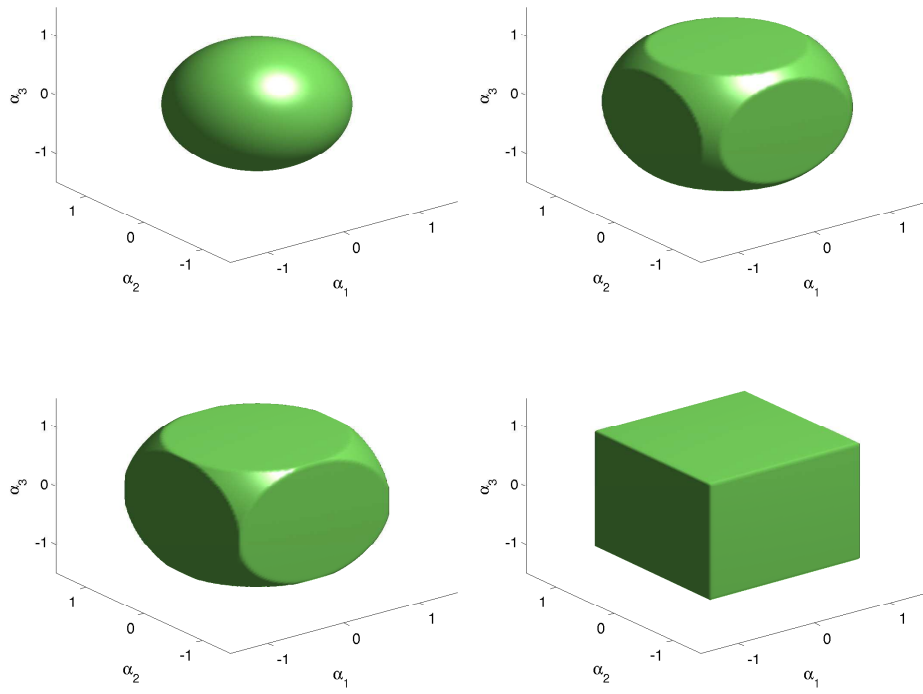


Figure 4: Balls for Ω^* for the groups $\mathcal{G} = \{\{1\}, \{2\}, \{3\}, \{1, 2, 3\}\}$ with $\gamma = 0, \frac{1}{4}, \frac{\log(2)}{2\log(3)}, \frac{1}{2}$ from top to bottom, left to right.

Definition 32 Let $g \in \mathcal{G}$ and $\mathcal{H} \subset \mathcal{G}$ a set of subgroups satisfying $\forall h \in \mathcal{H}, h \subset g$. We say that g dominates \mathcal{H} if \mathcal{H} could be the weak group-support for some \mathbf{w} if g was removed from \mathcal{G} , but is the weak group support of no \mathbf{w} in the presence of g .

We can characterize the presence of domination in terms of weights as follows:

Lemma 33 A group g dominates a set of subgroups \mathcal{H} if and only if, on the one hand, \mathcal{H} is a possible group-support when g is removed from \mathcal{G} , and, on the other,

$$d_g < P(g, \mathcal{H}) \triangleq \min \{ \|\alpha_g\| \mid \alpha \in \mathbb{R}^p \text{ and } \|\alpha_h\| = d_h, \forall h \in \mathcal{H} \}.$$

Proof First note that the set of constraints $\|\alpha_h\| = d_h, \forall h \in \mathcal{H}$ is feasible since \mathcal{H} is assumed to be a possible group support without g . Then note that the condition is equivalent to saying that the ball $\{\alpha_g \in \mathbb{R}^{|g|} \mid \|\alpha_g\| \leq d_g\}$ does not intersect the previous feasible set, which characterizes the set of possible dual variables for which the weak group-support is \mathcal{H} . \blacksquare

As discussed previously, one natural property to require would be that if \mathbf{w} is exactly supported by a group g , its group-support should be g . As argued in Lemma 27, we can not have this property in general. We can however show that if the support of \mathbf{w} is a single group in \mathcal{G} , then this group is always in the group support of \mathbf{w} .

The following result shows that, under some conditions on the weights, we can ensure that a group g does not dominate any set of subgroups that do not cover it entirely.

Lemma 34 Let a group $g \in \mathcal{G}$ and a set of subgroups $\mathcal{H} \subset \mathcal{G}$ such that $\forall h \in \mathcal{H}, h \subset g$ and $\cup_{h \in \mathcal{H}} h \subsetneq g$. Assuming that \mathcal{H} could be in the group support of some \mathbf{w} if g was removed from \mathcal{G} , then g does not dominate \mathcal{H} if, for some constant $d_1 > 0$, weights satisfy $d_h \leq \sqrt{|h|} d_1$ for all $h \in \mathcal{H}$ and $d_g \geq \sqrt{|g| - 1} d_1$.

Proof By Lemma 33, g does not dominate \mathcal{H} if and only if $d_g \geq P(g, \mathcal{H})$. To prove this, let us rewrite $P(g, \mathcal{H})$ as the solution of the following optimization problem:

$$\min_{\mathbf{x} \in \mathbb{R}_+^p} \mathbf{x}^\top \mathbf{1}_g \quad \text{s.t.} \quad \forall h \in \mathcal{H}, \mathbf{x}^\top \mathbf{1}_h = d_h^2.$$

By strong duality of linear programs $P(g, \mathcal{H})$ is also the solution of the dual problem:

$$\max_{\mathbf{u} \in \mathbb{R}^{|\mathcal{H}|}} \sum_{h \in \mathcal{H}} u_h d_h^2 \quad \text{s.t.} \quad \forall i \in [1, p], \sum_{h \in \mathcal{H}} u_h 1_{\{i \in h\}} \leq 1_{\{i \in g\}}.$$

But if $\bar{h} \triangleq \cup_{h \in \mathcal{H}} h$, under the conditions on the weights in Lemma 34, we can upper bound the optimal value as follows:

$$\sum_{h \in \mathcal{H}} u_h d_h^2 \leq d_1^2 \sum_{h \in \mathcal{H}} u_h |h| = d_1^2 \sum_{i \in g} \sum_{h \in \mathcal{H}} u_h 1_{\{i \in h\}} \leq d_1^2 |g \cap \bar{h}| \leq d_1^2 (|g| - 1),$$

where the second inequality results from the constraints of the dual program and the fact that for $i \in g \setminus \bar{h}$, the corresponding terms in the sum are equal to 0. This shows that if $d_g^2 \geq (|g| - 1) d_1^2$, then $d_g \geq P(g, \mathcal{H})$. \blacksquare

Note that Lemma 33 is tight in the following case:

Lemma 35 *For any group $g \in \mathcal{G}$, if \mathcal{H} is a set of $|g| - 1$ singletons of g , each with weight d_1 , that could be in a group support if g was removed, then g dominates \mathcal{H} if and only if $d_g < d_1 \sqrt{|g| - 1}$.*

Proof This is a direct consequence of Lemma 33, where the value of $P(g, \mathcal{H})$ is trivially equal to $d_1 \sqrt{|g| - 1}$. \blacksquare

What the two previous lemmata indicate is that, if there are large gaps in size between a group of size k and many much smaller subgroups contained in it, it is necessary to choose a value for the weight which is possibly unreasonably large, to allow all combinations of subgroups to be selected (even *non-covering* ones). Lemma 35 is illustrated on Figure 4, with the the group $\mathcal{G} = \{\{1\}, \{2\}, \{3\}, \{1, 2, 3\}\}$. Giving singletons the weight $d_1 = 1$, the critical weight for $g = \{1, 2, 3\}$ to dominate or not pairs of singletons is $d_g = \sqrt{|g| - 1} = \sqrt{2}$. We represent it equivalently as $d_g = |g|^\gamma$ with $\gamma = \frac{\log(2)}{2 \log(3)}$ on Figure 4. This corresponds to the critical value, below which it is not possible to select two singletons only. The trade-off we are facing here is not surprising when the weights are thought to correspond to *code lengths*. Indeed, in light of the interpretation of the norm Ω as a relaxation of a *block coding* penalization, it is clear that allowing groups with quite large weights (i.e., code lengths) increases the expressiveness of the code at the expense of compressibility and reduces the strength of the prior on support, since large weight allows for a greater diversity of supports. Put more simply, there is a trade-off between how coarsely the supports are encoded and how informative the prior on the supports is. The trade-off can also be interpreted as a bias-variance trade-off, where biasing the estimate of the support with a coarser set of patterns reduces the variance in its estimation.

It should be noted that, as an important consequence of domination, the set of possible sparsity patterns (although consisting of unions of sets of \mathcal{G}) is in general *not* stable by union.

8.3 Importance of weights for support consistency, FDR and FWER control

In this section we consider the following regression setting:

$$\min_{\mathbf{w} \in \mathbb{R}^p} \frac{1}{2} \|\mathbf{w} - \mathbf{w}^* + \epsilon\|^2 + \lambda \Omega(\mathbf{w}), \quad (26)$$

where the design matrix is taken to be the identity and the noise to be Gaussian, bearing in mind that the analysis we propose here could be extended easily to the case of a design satisfying properties such as RIP with noise that could be taken more generally subgaussian. The mapping to the solution of this optimization problem is often called the soft-thresholding operator, shrinkage operator or proximal operator associated with the norm Ω . We denote this mapping $\mathbf{w} \mapsto \text{ST}(\mathbf{w})$. In terms of support recovery and group-support consistency, a reasonable minimal requirement is that for sufficiently large values of the coefficients and for small levels of noise, assuming that the distribution of the noise is absolutely continuous with respect to the Lebesgue measure, the solution to problem (26) should retrieve the correct support, provided the latter can be expressed as a union of groups.

We first show that redundant groups may never be selected by (26).

Lemma 36 Take $\mathcal{G} = \{g, g'\}$ with $g \subsetneq g'$ and $d_g \geq d_{g'}$. Then for any \mathbf{w} , $g \notin \mathcal{G}_1(\hat{\mathbf{w}})$ a.s. where $\hat{\mathbf{w}} = ST(\mathbf{w})$.

Proof We first note that the optimality condition for (26) is

$$\hat{\mathbf{w}} - \mathbf{w}^* + \epsilon = -\lambda \alpha, \quad (27)$$

where $\alpha \in \mathcal{A}(\hat{\mathbf{w}})$. We then reason by contradiction and assume $g \in \mathcal{G}_1(\hat{\mathbf{w}})$ so that $\|\alpha_g(\hat{\mathbf{w}})\| = d_g$. Then, because $g \subsetneq g'$, $\|\alpha_{g'}(\hat{\mathbf{w}})\|^2 = \|\alpha_g(\hat{\mathbf{w}})\|^2 + \|\alpha_{g' \setminus g}(\hat{\mathbf{w}})\|^2 \leq d_{g'}^2$, which implies $\alpha_{g' \setminus g} = 0 = \mathbf{w}_{g' \setminus g} + \epsilon_{g' \setminus g} - \hat{\mathbf{w}}_{g' \setminus g}$. But $\mathbf{w}_{g' \setminus g} + \epsilon_{g' \setminus g} \neq \mathbf{0}$ a.s., this implies $\hat{\mathbf{w}}_{g' \setminus g} \neq \mathbf{0}$, and therefore that $\mathbf{v}^{g'} \neq \mathbf{0}$. But $\mathbf{v}^{g'}$ restricted to $g' \setminus g$ should then both be equal to $\mathbf{0}$ by optimality condition, and be equal to $\hat{\mathbf{w}}_{g' \setminus g}$, which is a contradiction. \blacksquare

Lemma 36 should be compared to Lemma 26. While the later one shows that g can not be selected without g' , Lemma 36 shows that in the regression setting it may simply not be selected a.s. This shows in particular that $d_g \geq d_{g'}$ can pose a problem of support consistency because it implies that, if the only way to write the support as a union of elements of \mathcal{G} is $\text{supp}(\mathbf{w}) = g$, the support is a.s. never correctly estimated by solving problem (26).

We now discuss in more details the influence of the weights on the probability to select false positives (Section 8.3.1) and to have false negatives (Section 8.3.2)

8.3.1 FALSE POSITIVES

Let us consider a group $g \in \mathcal{G}$ of size $|g| = k$ which is outside of the support (i.e. $\mathbf{w}_g^* = \mathbf{0}$), and such that not other group intersecting it is selected. From the optimality condition (27) we see that $\hat{\mathbf{w}}_g = \mathbf{0}$ if and only if $\|\epsilon_g\|^2 \leq \lambda^2 d_k^2$.

If we assume that $\lambda = \sigma$, then setting

$$d_k = \sqrt{k + c\sqrt{k}} \quad (28)$$

is an interesting choice because this is, at second order, the smallest possible rate that ensures that each group has a vanishingly small probability of being selected by chance. Indeed, on the one hand, $\|\epsilon_g\|^2 \sim \sigma^2 \chi_k^2$ so the usual Chernoff bound yields:

$$\mathbb{P}(\|\epsilon_g\|^2 \geq tk\sigma^2) \leq e^{-\frac{k}{2}(t - \log(t) - 1)},$$

and it is easy to verify that for $t = 1 + \frac{c'}{\sqrt{k}}$, with c sufficiently large, the above probability can be made arbitrarily small uniformly in k . This implies that if d_k is fixed according to (28), then with c large enough we can make the probability that g is selected as small as possible. On the other hand, choosing d_k smaller, i.e., $d_k^2 - k = o(\sqrt{k})$, would fail to guarantee $\mathbb{P}(\|\epsilon\|^2 \leq \sigma^2 d_k^2) > 1 - \eta$ for k large because the central limit theorem implies that $\frac{X-k}{\sqrt{2k}} \xrightarrow{d} \mathcal{N}(0, 1)$. In summary, (28) is the smallest rate which ensures that we can control the probability of selecting a wrong group uniformly in k . Finally, note that for $d_k = \sqrt{k + c\sqrt{k}}$, condition (C) is satisfied; furthermore, we have $ck^{\frac{1}{4}} \leq d_k \leq (1+c)k^{\frac{1}{2}}$. In particular if we consider the case of $c \rightarrow \infty$, we retrieve a scaling of the form $d_k = k^{\frac{1}{4}}$.

Note that if we want to control the expected number of incorrectly selected variables instead of the number of incorrectly selected groups, then, using the same reasoning, but

based on a bound on the expected number of false positive of the form $\sum_{g \in G} |g| \mathbb{P}(X_g \geq t|g|)$ we would show similarly that an appropriate choice for d_k is $d_k = \sqrt{k + c\sqrt{k \log k}}$. Obtaining a control of the type FWER instead of FDR is possible by choosing $c \propto \sqrt{\log m}$. The reader probably noticed that the analysis in this section is ignoring the overlaps between groups, and for groups that have a quite significant overlap with a group of the support, the probability of being incorrectly selected is much larger. This issue can however be addressed by choosing c sufficiently large. Besides this point, the weights derived nonetheless satisfy constraints from the previous sections in which issues arising from overlaps were considered.

8.3.2 FALSE NEGATIVES

These choices for d_k allow to control for false positives, but it is interesting as well to ask which groups containing true non-zero elements will be selected, and which ones could be false negatives. For simplicity we assume that $\mathbf{w}_i^* \in \{0, 1\}$ and that the noise is Gaussian as previously. If the fraction of non-zero elements in \mathbf{w}_i^* is p and one assumes a null model H_0 under which group g is unrelated to the nonzero pattern of \mathbf{w}^* then it is reasonable to model the number of non-zero elements in g as a binomial random variable $\mathcal{B}in(k, p)$ with $k = |g|$. Using again the KKT conditions, if none of the groups intersecting g is selected, we will have $\mathbf{v}_g = \mathbf{0}$ if and only if $\|\mathbf{w}_g^* + \epsilon_g\|^2 \leq \lambda^2 d_k^2$.

Since $\|\mathbf{w}_g^*\|^2 \sim \mathcal{B}in(k, p)$ and $\|\epsilon_g\|^2 \sim \sigma^2 \chi_k^2$, we have $\mathbb{E}[\|\mathbf{w}_g^* + \epsilon_g\|^2] = kp + k\sigma^2$ and

$$\text{Var}(\|\mathbf{w}_g^* + \epsilon_g\|^2) = \text{Var}(\|\mathbf{w}_g^*\|^2) + \mathbb{E}[(\epsilon_g^\top \mathbf{w}_g^*)^2] + \text{Var}(\|\epsilon_g\|^2) = kp(1-p) + 4kp\sigma^2 + 2k\sigma^4.$$

If $\lambda^2 = p + \sigma^2$ and if d_k is chosen of the previous form $d_k = \sqrt{k + c\sqrt{k}}$, then, for an appropriate choice of c , namely $c = c' \frac{\sqrt{p(1-p) + 4p\sigma^2 + 2\sigma^4}}{p + \sigma^2}$, classical Chernoff bounds together with an analysis similar to that of the previous section shows that we have $\|\mathbf{w}_g^* + \epsilon_g\|^2 > \lambda^2 d_k^2$ with probability decreasing exponentially in c' . Therefore in this model, groups selected can be interpreted as groups that are “enriched” in non-zero coefficients, where we call a group enriched if the number of non-zero coefficients in that group is significantly larger than for a random group of the same size. To put things differently the false negatives correspond to groups that do not have a significant number of non-zero elements.

This property is certainly a feature that can be desirable, especially in the applications in genomics that we have in mind where it is common to test for biological processes (or other groups of genes) that are enriched in “active genes”.

Note that if a group g has elements in common with another selected group g' , the elements that are in g' are explained in part by g' and are therefore “discounted” for group g , in the sense that we only need

$$\|\mathbf{w}_g^* - \sum_{g' \cap g \neq \emptyset} \mathbf{v}_g^{g'} + \epsilon_g\|^2 \leq \lambda^2 d_k^2.$$

A group is therefore selected if it contains enough non zero components that it itself explains.

It should be stressed that the previous analysis depends on the assumption that the components of \mathbf{w}^* are of the same order of magnitude and fails if the distribution of the entries of \mathbf{w}^* has a long tail.

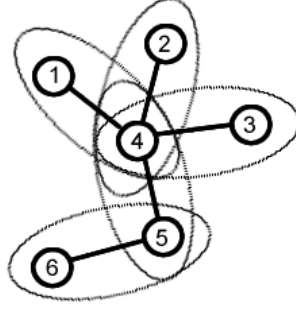


Figure 5: Graph-Lasso : if the penalty leads to the selection of connected sets of covariates like the edges, the resulting pattern should be more connected on the graph.

Finally, the analysis presented in these last two sections is heuristic in nature. It is by no means aimed at proving that a specific weighting scheme can be chosen universally for all possible collections of groups \mathcal{G} , but rather solely motivated by the need for an initial set of criteria to guide this choice. It is likely that finer analyses, namely under high-dimensional scaling and dedicated to specific collections of groups are required to make more definite recommendations for the choice of the weights. It should be noted that a different view on the weights can be adopted by considering them as defined through a set function; this is the point of view adopted in [Obozinski and F. \(2011\)](#) which relates the behavior of Ω to the set-function.

9. Graph Lasso

We now consider the situation where we have a simple undirected graph (I, E) , where the set of vertices $I = [1, k]$ is the set of covariates and $E \subset I \times I$ is a set of edges that connect covariates. We suppose that we wish to estimate a sparse model such that selected covariates tend to be connected to each other, *i.e.*, form a limited number of connected components on the graph. An obvious approach is to use the norm $\Omega_{\mathcal{G}}^{\mathcal{G}}$ where \mathcal{G} is a set that generates connected components by union. For example, we may consider for \mathcal{G} the set of edges, cliques, or small linear subgraphs. As an example, considering all edges, *i.e.*, $\mathcal{G} = E$ leads to :

$$\Omega_{\text{graph}}(\mathbf{w}) = \min_{v \in \mathcal{V}_E} \sum_{e \in E} d_e \|v_e\| \quad \text{s.t.} \quad \sum_{e \in E} v_e = w, \text{supp}(v_e) = e.$$

Alternatively, we will consider in the experiments the set of all linear subgraphs of length $k \geq 1$. Although we have no formal statement on how to choose k , it intuitively controls the size of the groups of connected variables which are selected, and should therefore be typically chosen to be slightly smaller than the size of the minimal connected component expected in the support of the model.

10. Experiments

To assess the performance of our method when either overlapping groups or a graph are provided as a priori information, and subsequently, to assess the influence of the weights d_g , we considered several synthetic examples of regression model in which the structure of the model generating the data matches the prior on supports induced by the norm.

10.1 Synthetic data: given overlapping groups

In this experiment, we simulated data with $p = 82$ variables, covered by 10 groups of 10 variables with 2 variables of overlap between two successive groups:

$$\mathcal{G} = \{\{1, \dots, 10\}, \{9, \dots, 18\}, \dots, \{73, \dots, 82\}\}.$$

We chose the support of \mathbf{w} to be the union of groups 4 and 5 and sampled both the coefficients on the support and the offset from i.i.d. Gaussian variables. Note that in this setting, the support can be expressed as a union of groups, but not as the complement of a union. Therefore, our latent group Lasso penalty $\Omega_{\mathcal{G}}^{\mathcal{G}}$ could recover the right support.

The model is learned from n data points (\mathbf{x}_i, y_i) , with $y_i = \mathbf{w}^\top \mathbf{x}_i + \varepsilon$, $\varepsilon \sim \mathcal{N}(0, \sigma^2)$, $\sigma = |\mathbb{E}(\mathbf{X}\mathbf{w} + b)|$. Using an ℓ_2 loss $L(\mathbf{w}) = \|\mathbf{y} - \mathbf{X}\mathbf{w} - b\|^2$, we learn models from 100 such training sets.

We report the empirical frequencies of the selection of each variable on Figure 6. For any choice of λ , the Lasso frequently misses some variables from the support, while $\Omega_{\mathcal{G}}^{\mathcal{G}}$ does not miss any variable from the support on a large part of the regularization path. Besides, we observe that over the replicates, the Lasso never selects the exact correct pattern for $n < 100$. For $n = 100$, the right pattern is selected with low frequency on a small part of the regularization path. $\Omega_{\mathcal{G}}^{\mathcal{G}}$ on the other hand selects it up to 92% of the times for $n = 50$ and more than 99% on more than one third of the path for $n = 100$.

Figure 7 shows the root mean squared error for both methods and several values of n . For both methods, the full regularization path is computed and tested on three replicates of n training and 100 testing points. We selected the best parameter in average and used it to train and test a model on a fourth replicate. For a large range of n , $\Omega_{\mathcal{G}}^{\mathcal{G}}$ not only helps to recover the right pattern, but also decreases the MSE compared to the classical Lasso.

10.2 Synthetic data: given linear graph structure

We now consider the case where the prior given on the variables is a graph structure and where we are interested by solutions which are highly connected components on this graph. As a first simple illustration, we consider a chain in which variables with successive indices are connected. We use $\mathbf{w} \in \mathbb{R}^p$, $p = 100$, $\text{supp}(\mathbf{w}) = [20, 40]$. The nodes of the graph correspond to the parameters w_i and the edges to the pairs (w_i, w_{i+1}) , $i = 1, \dots, n$. The parameters of the model and the 50 training examples (\mathbf{x}_i, y_i) are drawn using the same protocol as in the previous experiment. We use for the groups all the sub-chains of length k . Results are reported for various choices of k and compared to the Lasso ($k = 1$).

Figure 8 shows the frequency of each variable selection over 20 replications. Here again, using a group prior improves pattern recovery, with better results as k increases. However, for larger groups, two consecutive groups are very correlated, which makes it more difficult to identify the exact boundaries of the support.

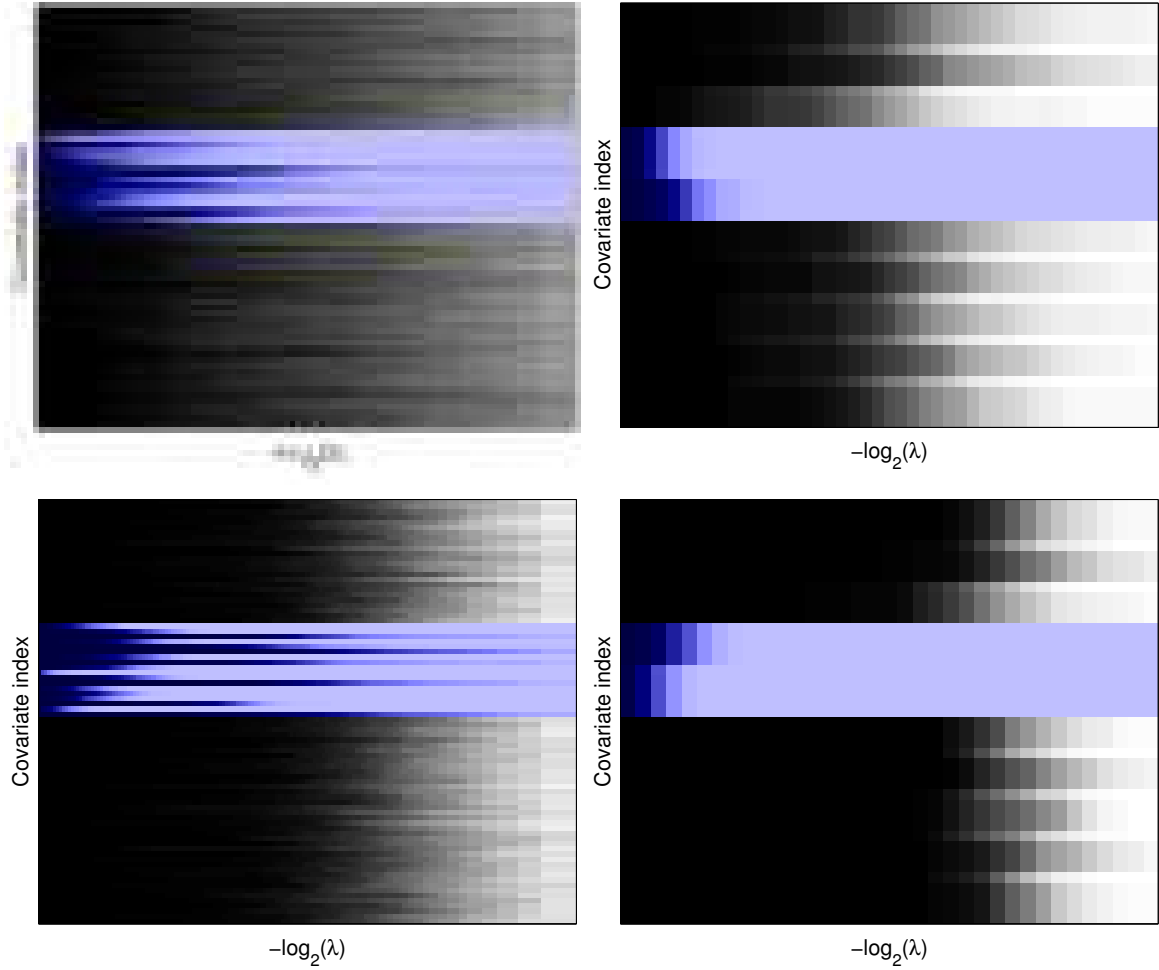


Figure 6: Frequency of selection of each variable with the Lasso (left) and $\Omega_{\mathcal{G}}$ (right) for $n = 50$ (top) and 100 (bottom). For each variable index (on the y-axis), its frequency of selection is represented in levels of gray as a function of the regularization parameter λ (on the x-axis), both for the Lasso penalty and $\Omega_{\mathcal{G}}$. The transparent blue band superimposed indicates the set of covariates that belong to the support.

10.3 Synthetic data: effect of the weights

As discussed in Section 8, the choice of a set of weights $\{d_g\}_{g \in \mathcal{G}}$ influences the variable selection behavior of the learning algorithm penalized by Ω . At one extreme, if the weights are uniform, only groups that are included in no other can be selected. At the other extreme, for weights growing as the square root of the group size, the group-support selected will be composed (almost surely) of the smallest groups possible covering the support.

To illustrate the effect of the weighting scheme on covariate selection, we run three experiments with respectively $p = 100, 200, 300$ covariates and $n = 100, 50, 30$ training

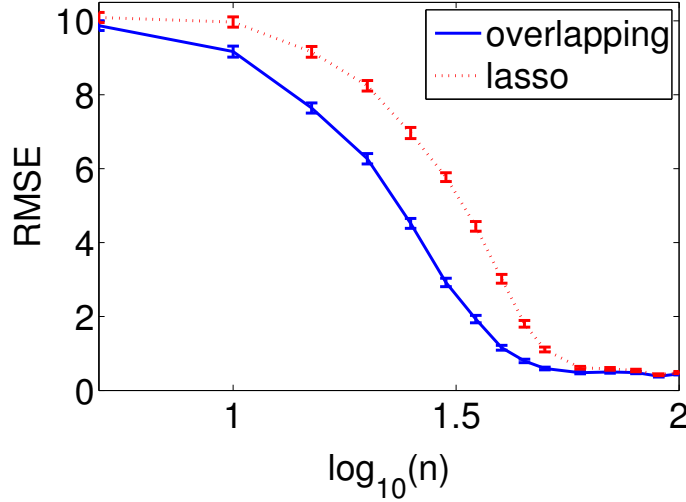


Figure 7: Root mean squared error of overlapped group lasso and Lasso as a function of the number of training points.

points. In each setting, the groups are all the sets of size from 1 to 20 formed by sequences of consecutive covariates, much like in 10.2 but with more groups. Note that this creates a lot of nested groups. The support is formed by covariates with indices from 5 to 24 and from 90 to 92, *i.e.*, 23 covariates. The noise level σ^2 is 0.1. For each of the three settings, we compare 6 weighting schemes over 50 replications. The first 4 schemes follow (28) and assign $d_s = \sqrt{s + c\sqrt{s}}$ to each group of size s , with $c = 0, 1, 4, 6$. We also try $d_s = \sqrt[4]{s}$ (the limit when c grows) and $d_s = 1$. Note that $d_s = 1$ and $c = 0$ ($d_s = \sqrt{s}$) correspond to the two extreme regimes in condition (C).

We evaluate the performance of the regularization in two different ways. First, we select by cross-validation the value of λ that yields the smallest MSE and return the corresponding value. Second, we return the best possible recovery error attainable on the entire regularization path. We consider these two criteria since it is known that the regularization regime corresponding to optimal support recovery and best MSE are not the same (Bach, 2008b; Leng et al., 2004).

Ideally, for support recovery, we would have to either use a theoretical value for λ or to use the OLS-hybrid two-step procedure (Efron et al., 2004) in which the models obtained in sequence along the regularization path are refitted with OLS and tested on a held out set to select the best model. This would obviously lead to a much heavier experimental setting, which is why we simply return the best performance along the path.

The results are shown in Table 1, 2 and 3. In each case, the best average MSE across the 50 runs and along the regularization path is given along with the corresponding point on the regularization path (λ^*), average number of selected variables in the corresponding model (Model size*), pattern recovery error of the selected model (Rec err*) and lowest

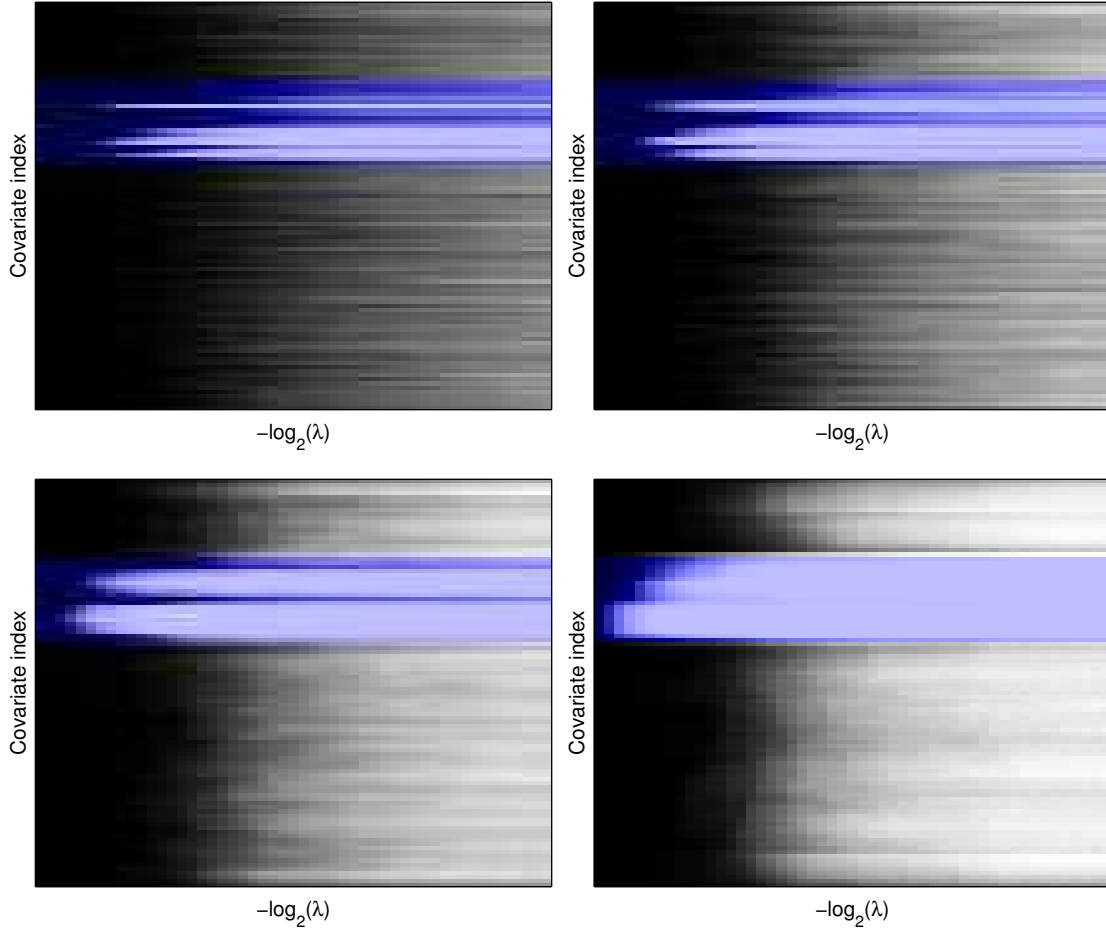


Figure 8: Variable selection frequency with $\Omega_{\mathcal{U}}^{\mathcal{G}}$ using the chains of length k (left) as groups, for $k = 1$ (Lasso), 2, 4, 8. For each variable index (on the y-axis), its frequency of selection is represented in levels of gray as a function of the regularization parameter λ (on the x-axis), both for the Lasso penalty and $\Omega_{\mathcal{U}}^{\mathcal{G}}$. The transparent blue band superimposed indicates the covariates that belong to the support.

pattern recovery error along the regularization path (Rec err min). The pattern recovery error is the average of the proportion of covariates that were in the support and were not selected, and the proportion of covariates that were not in the support and were selected. The standard deviation is given for each measured quantity as well. The regularization path was approximated by a grid of 51 values of λ between 2^{-7} and 2^3 . For Table 2, a longer grid of 76 values starting at 2^{-12} was used to make sure that the end of the regularization path was reached.

The last column of Table 1 illustrates the effect of the weighting scheme on pattern recovery.

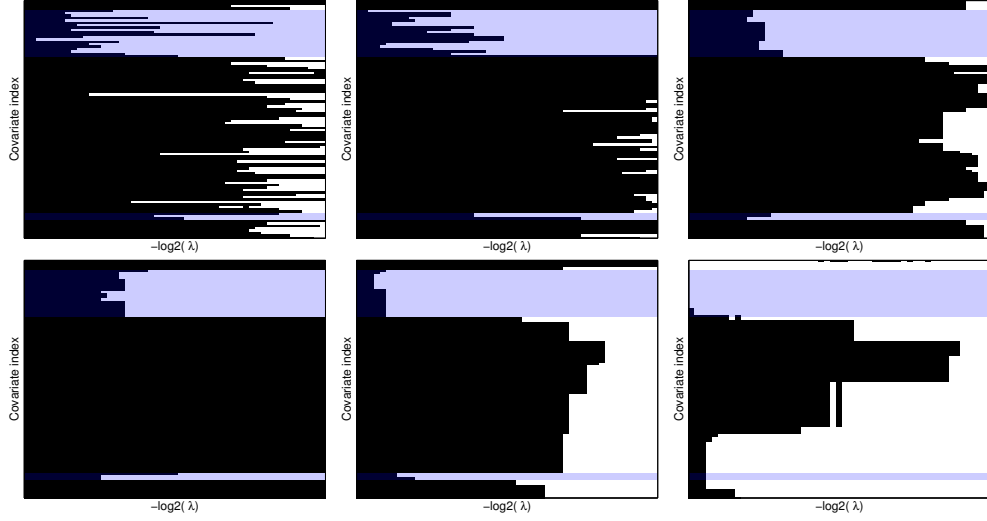


Figure 9: Variable selection for one of the 50 runs with $\Omega_{\mathcal{U}}^{\mathcal{G}}$ using the chains up to length 20 as groups and weights of the form $d_k = \sqrt{k + c\sqrt{k}}$, $c = 0, 1, 4, 6, \infty$ and uniform weights (from left to right, top to bottom). A transparent blue band is superimposed to indicate the covariates that belong to the support.

The results of Table 1 correspond to $n = 100$, $p = 100$ so that if $s = 23$ is the size of the support, we have $n/(2s \log(p)) \approx 0.47$ which means that the sample size is slightly too small for the Lasso to recover the support exactly. Note that as expected from the theory, the fifth column shows that the model selected based on the MSE is not optimal in term of variable selection. The fourth column shows that more uniform weights encourage the selection of more variables, which is expected given that they favor the selection of larger groups. Lastly, the values of the MSE suggest that in this regime of sparsity, dimension and number of training points, the performances in pattern recovery have little influence on the MSE, because there are enough training points to deal with the noise created by the selection of spurious covariates. Here again however, the two extreme regimes lead to higher MSE.

Figure 9 illustrates the influence of the weights on the selection behavior. As expected from theory, uniform weights ($d_k = 1$) only allow selection of the largest groups *i.e.*, chains of size 20 while at the other extreme, for $d_k = \sqrt{k}$, only the small groups (singletons) are active. In intermediate regimes, all groups are active and allow to recover the correct support at some point on the regularization path, except $c = 1$ which on this particular run doesn't yield perfect recovery. More adequate choices of c lead to correct recovery on a larger portion of the regularization path.

Table 2 corresponds to a harder regime, with fewer training points and in higher dimension. As in the first regime, the fourth and last columns shows that the weighting scheme has a significant influence on the variable selection behavior, with more uniform schemes leading to more variables selected, and a better pattern recovery being achieved

Table 1: Effect of c on the MSE, the λ giving the best average MSE, the pattern recovery error at the optimal MSE, and the best pattern recovery error possible. 100 training points, 100 dimensions, 50 replications.

c	MSE	λ^*	MODEL SIZE*	REC ERR*	REC ERR MIN
0	0.06709 ± 0.1814	0.02368	37.08 ± 12.8	0.1068 ± 0.07444	0.07148 ± 0.03768
1	0.02891 ± 0.09583	0.01031	41.8 ± 18.4	0.1245 ± 0.12	0.02951 ± 0.02057
4	0.04513 ± 0.07202	0.0136	49.72 ± 27.21	0.1759 ± 0.1759	0.01468 ± 0.01599
6	0.03877 ± 0.1116	0.01031	45.78 ± 26.63	0.1506 ± 0.1741	0.01804 ± 0.01579
$d_s = \sqrt[4]{s}$	0.04318 ± 0.08945	0.0359	51.72 ± 27.11	0.1878 ± 0.1757	0.02461 ± 0.02585
$d_s = 1$	0.09263 ± 0.2278	0.04737	81.22 ± 17.16	0.3764 ± 0.1129	0.09788 ± 0.03598

Table 2: Effect of c on the MSE, the λ giving the best average MSE, the pattern recovery error at the optimal MSE, and the best pattern recovery error possible. 50 training points, 200 dimensions, 50 replications.

c	MSE	λ^*	MODEL SIZE*	REC ERR*	REC ERR MIN
0	8.264 ± 5.187	0.04123	47.54 ± 7.149	0.2706 ± 0.06144	0.2661 ± 0.06096
1	6.317 ± 4.809	0.0002441	61.3 ± 3.824	0.1957 ± 0.07468	0.1823 ± 0.08499
4	2.428 ± 2.401	0.0002441	101.4 ± 13.74	0.2301 ± 0.04765	0.08716 ± 0.05194
6	2.2 ± 2.404	0.0002441	111.9 ± 17.29	0.2572 ± 0.05094	0.06944 ± 0.03839
$D(s) = \sqrt[4]{s}$	1.66 ± 1.593	0.0007401	141.2 ± 15.52	0.3366 ± 0.04511	0.0823 ± 0.05281
$D(s) = 1$	3.707 ± 2.836	0.0002441	155.4 ± 14.44	0.3757 ± 0.0409	0.08228 ± 0.02283

for an intermediate scheme ($c = 6$). The reason for the optimal c to be higher than in the previous regime may be that in higher dimension with less training points, it is not possible anymore to recover the fine structure of the true pattern and a better alternative is to select a less precise but more stable selection of larger groups. In terms of MSE, the minimum is reached for $d_s = \sqrt[4]{s}$, and for all the other weightings the optimum λ is the last one in the grid, for which a large fraction of the covariates have entered the model.

In the last regime (30 training points, 300 dimensions), Table 3 shows that the best pattern recovery is performed with uniform weights, which suggests that at this level of noise, using the fine structure of the groups is more harmful than helpful, and that the best choice is to only use the largest groups. The same reasoning applies to the MSE.

10.4 Breast cancer data: pathway analysis

An important motivation for our method is the possibility to perform gene selection from microarray data using priors which are overlapping groups. Genes are known to modify each other's expression through various regulation mechanisms. More generally, some genes are known to be involved in the same biological function, so the presence of a particular gene

Table 3: Effect of c on the MSE, the λ giving the best average MSE, the pattern recovery error at the optimal MSE, and the best pattern recovery error possible. 30 training points, 300 dimensions, 50 replications.

C	MSE	λ^*	MODEL SIZE*	REC ERR*	REC ERR MIN
0	18.78 ± 7.021	1.32	15.74 ± 3.451	0.4059 ± 0.07167	0.396 ± 0.07169
1	17.21 ± 6.763	0.5743	23.22 ± 3.501	0.3841 ± 0.06413	0.3693 ± 0.07547
4	17.21 ± 8.195	0.125	51.5 ± 10.74	0.2281 ± 0.1294	0.2181 ± 0.1285
6	14.74 ± 7.398	0.125	66.86 ± 17.36	0.2037 ± 0.1122	0.1996 ± 0.1198
$D(s) = \sqrt[4]{s}$	11.81 ± 5.307	0.007812	119.8 ± 23.15	0.2259 ± 0.08258	0.1546 ± 0.1197
$D(s) = 1$	11.82 ± 5.31	0.007812	159.2 ± 24.22	0.268 ± 0.0401	0.1284 ± 0.05387

in a predictive models can be indicative of the presence of related genes. In other words, when we select one gene in our predictive model, we can expect that genes which are known to either regulate or to be regulated by this gene, or more generally to be involved in the same biological function should also be selected. Since an increasing amount of information on gene interaction is being gathered from empirical biological knowledge and organized in databases (Subramanian et al., 2005), our hope is to use this information to :

Improve prediction accuracy : Functions involving a small number of pre-defined gene sets, form a smaller hypothesis sets in which we can hope to better estimate. Since genes present in the same biological function are likely to be either all involved in the studied phenomenon (disease outcome, subtype, response to a treatment) or all not involved, we can expect to find a function predicting the phenomenon correctly in this class.

Build accurate sparse prediction functions : Building sparse estimators has practical implications in this context because it is technically easier to measure the expression level of a small number of genes in a patient than a whole transcriptome. Selecting a small number of gene sets is a more robust procedure than selecting a small number of genes, because it is easy to spuriously select a gene from a noisy training set while the evidences add up for a set of genes. In addition, selecting a few genes that belong to the same functional groups could lead to increased interpretability of the signature.

To reach this goal we use our $\Omega_{\cup}^{\mathcal{G}}$ penalty with an (overlapping) predefined gene sets as groups. Several groupings of genes into gene sets are available in various databases. We use the canonical pathways from MSigDB (Subramanian et al., 2005) containing 639 groups of genes, 637 of which involve genes from our study. Among these, we restricted ourselves to the 589 groups that contained less than 50 genes. Indeed we observed empirically that keeping very large pathways in the penalty lead to poor regularization, which makes sense because the presence of very large groups allows the penalty to select a very large number of covariates at a low cost, partially breaking the purpose of regularization. As discussed in Section 8, it is possible to penalize large groups more heavily, but weighting cannot correct extreme size discrepancies such as combinations of groups of size two and groups of size

100. In addition, we are interested in identifying a small number of well defined biological functions that predict the outcome. Selecting a large pathway which contains one third of the genes would not be very informative.

We use the breast cancer dataset compiled by [van de Vijver et al. \(2002\)](#), which consists of gene expression data for 8,141 genes in 295 breast cancer tumors (78 metastatic and 217 non-metastatic). We restrict the analysis to the 2465 genes which are in at least one pathway. Since the dataset is very unbalanced, we use a balanced logistic loss, weighting each positive example by the proportion of negative examples and each negative example by the proportion of positive examples.

We estimate by 5-fold cross validation the balanced accuracy (average of specificity and sensitivity) of the balanced logistic regression with ℓ_1 and $\Omega_{\mathcal{G}}^{\mathcal{G}}$ penalties, using the pathways as groups. As a pre-processing, we keep the 500 genes most correlated with the output (on each training set). This type of prefiltering is common practice with microarray data, and all the results are quite robust to changes in the number of genes kept. λ is selected by internal cross validation on each training set.

In our experiments on this very noisy dataset, we noticed that results changed a lot with the choice of the split, often more than between methods. In order to make sure that observed differences were actually caused by algorithms and not by particular choices of the 5 foldings, we repeated each experiment on 5 choices of the 5 foldings, and show the result for each of these choices separately.

Table 4 gives the balanced accuracies using $\Omega_{\mathcal{G}}^{\mathcal{G}}$ with and without weights, and using ℓ_1 . We observe a consistent improvement in the performances when using $\Omega_{\mathcal{G}}^{\mathcal{G}}$ against ℓ_1 (between 2% and 12% depending on the fold). The weighted version of $\Omega_{\mathcal{G}}^{\mathcal{G}}$ using $c = 4$ also leads to consistent improvement over ℓ_1 but is outperformed by the unweighted version of the penalty. Table 5 shows that the unweighted version of the penalty tends to select groups that are larger than average, since the average size of the initial set of pathways (after the preprocessing step that keeps only 500 genes) is 5 genes with a standard deviation of slightly above 5. The weighted penalty allows to correct this bias: it leads to the selection of groups of average size 5 but typically selects a much larger number of groups.

Table 6 shows the average number of genes involved in the model learned by each of the methods. As expected, Ω selects more genes, since it enforces sparsity at the gene set level but doesn't enforce sparsity at the gene level. Note however that the number of involved genes remains reasonable. As expected given the numbers of Table 5 the number of genes selected in the model learned by the weighted version of $\Omega_{\mathcal{G}}^{\mathcal{G}}$ is even larger.

Finally, we should mention, as a caveat, that the regularization coefficient was chosen here to minimize the classification error, i.e., in a regime which typically overestimates the support. A more tedious two-stage approach allowing to remove the bias of the estimator, would probably lead to smaller supports, as suggested by the comparison of Rec Err and Rec Err Min in Tables 1,2 and 3.

10.5 Breast cancer data: graph analysis

Another important application of microarray data analysis is the search for potential drug targets. In order to identify genes which are related to a disease, one would like to find groups of genes forming densely connected components on a graph carrying biological information

Table 4: Balanced classification error for the ℓ_1 and $\Omega_{\cup}^{\mathcal{G}}$ (with and without weights) on average over 5 folds, for 5 different folding choices.

METHOD	$\Omega_{\cup}^{\mathcal{G}}$	WEIGHTED $\Omega_{\cup}^{\mathcal{G}}$	ℓ_1
ERROR FOLDING 1	0.29 ± 0.05	0.35 ± 0.05	0.36 ± 0.04
ERROR FOLDING 2	0.30 ± 0.08	0.39 ± 0.05	0.42 ± 0.04
ERROR FOLDING 3	0.34 ± 0.14	0.34 ± 0.1	0.37 ± 0.10
ERROR FOLDING 4	0.31 ± 0.11	0.33 ± 0.07	0.37 ± 0.08
ERROR FOLDING 5	0.35 ± 0.05	0.35 ± 0.05	0.37 ± 0.05

Table 5: Number (and size) of involved pathways in the $\Omega_{\cup}^{\mathcal{G}}$ (with and without weights) signatures on average over 5 folds, for 5 different folding choices.

METHOD	$\Omega_{\cup}^{\mathcal{G}}$	WEIGHTED $\Omega_{\cup}^{\mathcal{G}}$
FOLDING 1	$6 \pm 1.225(16.73 \pm 2.378)$	$45.8 \pm 21.11(5.35 \pm 0.6635)$
FOLDING 2	$12.6 \pm 7.765(13.86 \pm 3.589)$	$48.8 \pm 23.13(5.092 \pm 0.4939)$
FOLDING 3	$7.6 \pm 3.209(14.86 \pm 2.584)$	$43.8 \pm 12.13(5.147 \pm 0.7176)$
FOLDING 4	$8.6 \pm 7.266(16.7 \pm 4.477)$	$30.6 \pm 17.3(5.045 \pm 0.7267)$
FOLDING 5	$8 \pm 1(14.82 \pm 1.191)$	$48.4 \pm 10.62(5.347 \pm 0.2867)$

such as regulation, involvement in the same chain of metabolic reactions, or protein-protein interaction. Similarly to what is done in pathway analysis, [Chuang et al. \(2007\)](#) built a network by compiling several biological networks and performed such a graph analysis by identifying discriminant subnetworks in one step and using these subnetworks to learn a classifier in a separate step. We use this network and the approach described in section 9, treating all the edges on the network as groups of size two, on the breast cancer dataset. Here again, we restrict the data to the 7910 genes which are present in the network, and use the same correlation-based pre-processing as for the pathway analysis to reduce the set to 500 genes.

Table 7 shows the prediction accuracy of the balanced logistic regression with ℓ_1 and $\Omega_{\cup}^{\mathcal{G}}$. Both methods yield almost exactly the same performance in average, suggesting that this particular network is not a particularly informative prior for this learning problem.

Nonetheless, while ℓ_1 mostly selects isolated variables on the graph, $\Omega_{\cup}^{\mathcal{G}}$ tends to select variables which are clustered into larger connected components. Table 8 shows, for each of the 5 foldings, the size of the largest connected component of the network restricted to the selected genes (the average and standard deviations are computed over the 5 folds of each folding). The average size of the largest connected component in the network after preprocessing (*i.e.*, keeping only 500 genes in each training set) is 68. One might suspect that the increase of connectivity is merely caused by the fact that overall the $\Omega_{\cup}^{\mathcal{G}}$ selects more genes. While it is clear that selecting more genes makes it more likely to select

Table 6: Number of involved genes in the ℓ_1 and $\Omega_{\mathcal{U}}^{\mathcal{G}}$ (with and without weights) signatures on average over 5 folds, for 5 different folding choices.

METHOD	$\Omega_{\mathcal{U}}^{\mathcal{G}}$	WEIGHTED $\Omega_{\mathcal{U}}^{\mathcal{G}}$	ℓ_1
FOLDING 1	98 ± 18	159.4 ± 60.1	41.2 ± 20.6
FOLDING 2	86.4 ± 18	143.4 ± 32	59.4 ± 22.5
FOLDING 3	125 ± 37.7	156.4 ± 36.7	59.4 ± 21.4
FOLDING 4	91.6 ± 25	115.2 ± 57.9	45.6 ± 28.4
FOLDING 5	98 ± 36	178.4 ± 33.9	56 ± 97

Table 7: Balanced classification error of the ℓ_1 and $\Omega_{\mathcal{U}}^{\mathcal{G}}$ (using the edges as the groups) on the 5 folds.

METHOD	$\Omega_{\mathcal{U}}^{\mathcal{G}}$	ℓ_1
FOLDING 1	0.3625 ± 0.04538	0.3367 ± 0.03788
FOLDING 2	0.4142 ± 0.05885	0.4042 ± 0.06035
FOLDING 3	0.3681 ± 0.04773	0.3782 ± 0.07497
FOLDING 4	0.3749 ± 0.06476	0.3834 ± 0.06449
FOLDING 5	0.3317 ± 0.04318	0.3443 ± 0.04414

larger connected components, the last two columns of Table 8 suggest that the increased connectivity is not simply caused by the selection of a larger number of genes. For example in folding 5, $\Omega_{\mathcal{U}}^{\mathcal{G}}$ selects many more genes than ℓ_1 but leads to the most modest increase in connectivity, while in folding 4 the number of selected genes is practically the same, although the $\Omega_{\mathcal{U}}^{\mathcal{G}}$ estimate is still much more connected than that of ℓ_1 .

This gain of connectivity without loss of prediction accuracy could potentially make the interpretation of the classifier and the search for new drug targets easier in practice.

11. Conclusion

We have presented the latent group Lasso, a generalization of the group lasso penalty which leads to sparse models with sparsity patterns that are unions of pre-defined groups of covariates, or, given a graph of covariates, groups of connected covariates in the graph. We studied various properties of the penalty function, and gave both sufficient and necessary conditions for *group-support recovery*, i.e., the correct recovery of the same union of groups as in the decomposition induced by the penalty on the true optimal parameter vector. We have highlighted the importance of setting weights correctly, and obtained promising empirical results on both simulated and real data.

In future work it would be interesting to characterize further for which collections of groups the latent group Lasso penalty and the estimators obtained by regularizing with it are computable efficiently; which form of structures can be encoded via such collections; and

Table 8: Average size of the largest connected components and average number of genes selected by the ℓ_1 and $\Omega_{\mathcal{G}}^{\mathcal{G}}$ (using the edges as the groups) on the 5 folding.

METHOD	$\Omega_{\mathcal{G}}^{\mathcal{G}}$ LARGEST CC	ℓ_1 LARGEST CC	$\Omega_{\mathcal{G}}^{\mathcal{G}}$ # GENES	ℓ_1 # GENES
FOLDING 1	10.2 ± 5.586	1.8 ± 0.4472	75.4 ± 47.54	37.2 ± 17.68
FOLDING 2	6.2 ± 3.633	2 ± 0	58.4 ± 30.81	50 ± 9.301
FOLDING 3	8.6 ± 4.278	2 ± 0.7071	53.2 ± 8.012	43.2 ± 5.357
FOLDING 4	8 ± 6.205	2.2 ± 0.4472	48.6 ± 30.25	45.6 ± 20.63
FOLDING 5	6 ± 3.082	1.8 ± 0.4472	69 ± 31.2	37.2 ± 12.3

what are the appropriate choice of weights in those cases, which will have to be determined based on specific analyses of the consistency of these estimators under high-dimensional scaling. Finally, more systematic comparisons with other group Lasso formulations, such as that proposed by Jenatton et al. (2009), would be important.

Acknowledgments

LJ gratefully acknowledges the support of the Stand Up to Cancer Program. JPV was supported by ANR grants ANR-07-BLAN-0311-03 and ANR-09-BLAN-0051-04. GO acknowledges funding from the European Research Council grant SIERRA: Project 239993. The authors would like to thank Rodophe Jenatton, Julien Mairal and Francis Bach for useful discussions.

References

- A. Agarwal, S. Negahban, and M.J. Wainwright. Noisy matrix decomposition via convex relaxation: Optimal rates in high dimensions. Technical Report 1102.4807, arXiv, 2011. URL <http://arxiv.org/abs/1102.4807>.
- F. Bach. Consistency of the group lasso and multiple kernel learning. *J. Mach. Learn. Res.*, 9:1179–1225, 2008a. URL <http://jmlr.csail.mit.edu/papers/v9/bach08b.html>.
- F. Bach. Exploring large feature spaces with hierarchical multiple kernel learning. In *Adv. Neural. Inform. Process Syst.*, volume 21, pages 105–112, 2009.
- F. Bach. Structured sparsity-inducing norms through submodular functions. Technical Report 1008.4220, arXiv, 2010. URL <http://arxiv.org/abs/1008.4220>.
- F. Bach, R. Jenatton, J. Mairal, and G. Obozinski. Optimization with sparsity-inducing penalties. Technical Report hal-00613125, HAL, 2011. URL <http://hal.archives-ouvertes.fr/hal-00613125/fr/>.
- F. R. Bach. Bolasso: model consistent lasso estimation through the bootstrap. In *ICML '08: Proceedings of the 25th international conference on Machine learning*, pages 33–40,

- New York, NY, USA, 2008b. ACM. ISBN 978-1-60558-205-4. doi: <http://doi.acm.org/10.1145/1390156.1390161>.
- F. R. Bach, G. R. G. Lanckriet, and M. I. Jordan. Multiple kernel learning, conic duality, and the SMO algorithm. In *Proceedings of the Twenty-First International Conference on Machine Learning*, page 6, New York, NY, USA, 2004. ACM. doi: <http://doi.acm.org/10.1145/1015330.1015424>.
- R.G. Baraniuk, V. Cevher, M.F. Duarte, and C. Hegde. Model-based compressive sensing. *Information Theory, IEEE Transactions on*, 56(4):1982–2001, 2010.
- C. Berge. *Espaces topologiques et fonctions multivoques*. Dunod, Paris, 1959.
- P. J. Bickel, Y. Ritov, and A. Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *Ann. Stat.*, 37(4):1705–1732, 2009.
- K. C. Border. *Fixed point theorems with applications to economics and game theory*. Cambridge University Press, Cambridge, UK, 1985.
- E.J. Candes, X. Li, Y. Ma, and J. Wright. Robust principal component analysis? *Journal of the ACM*, 58(1):1–37, 2009.
- V. Chandrasekaran, B. Recht, P.A. Parrilo, and A.S. Willsky. The convex geometry of linear inverse problems. Technical Report 1012.0621, arXiv, 2010. URL <http://arxiv.org/abs/1012.0621>.
- S. S. Chen, D. L. Donoho, and M. Saunders. Atomic decomposition by basis pursuit. *SIAM J. Sci. Comput.*, 20(1):33–61, 1998. doi: 10.1137/S1064827596304010. URL <http://dx.doi.org/10.1137/S1064827596304010>.
- Y. Chen, H. Xu, C. Caramanis, and S. Sanghavi. Robust matrix completion with corrupted columns. Technical Report 1102.2254, arXiv, 2011. URL <http://arxiv.org/abs/1102.2254>.
- H.-Y. Chuang, E. Lee, Y.-T. Liu, D. Lee, and T. Ideker. Network-based classification of breast cancer metastasis. *Mol. Syst. Biol.*, 3:140, 2007. doi: 10.1038/msb4100180. URL <http://dx.doi.org/10.1038/msb4100180>.
- B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *Ann. Stat.*, 32(2):407–499, 2004.
- L. He and L. Carin. Exploiting structure in wavelet-based Bayesian compressive sensing. *IEEE Transactions on Signal Processing*, 57:3488–3497, 2009.
- J.B. Hiriart-Urruty and C. Lemaréchal. *Convex Analysis and Minimization Algorithms I: Fundamentals*. Springer-Verlag, 1994.
- J. Huang and T. Zhang. The benefit of group sparsity. *Ann. Stat.*, 38(4):1978–2004, 2010. doi: 10.1214/09-AOS778. URL <http://dx.doi.org/10.1214/09-AOS778>.

- J. Huang, T. Zhang, and D. Metaxas. Learning with structured sparsity. Technical Report 0903.3002, arXiv, 2009. URL <http://arxiv.org/abs/0903.3002>.
- L. Jacob, G. Obozinski, and J.-P. Vert. Group lasso with overlap and graph lasso. In *ICML '09: Proceedings of the 26th Annual International Conference on Machine Learning*, pages 433–440, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-516-1. doi: <http://doi.acm.org/10.1145/1553374.1553431>.
- Ali Jalali, Pradeep Ravikumar, Sujay Sanghavi, and Chao Ruan. A dirty model for multi-task learning. In J. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R.S. Zemel, and A. Culotta, editors, *Adv. Neural. Inform. Process Syst.*, pages 964–972. Kaufmann publishers, 2010.
- R. Jenatton, J.-Y. Audibert, and F. Bach. Structured variable selection with sparsity-inducing norms. Technical Report 0904.3523, arXiv, 2009. URL <http://fr.arxiv.org/abs/0904.3523>.
- R. Jenatton, J. Mairal, G. Obozinski, and F. Bach. Proximal methods for hierarchical sparse coding. *J. Mach. Learn. Res.*, 12(Jul):2297–2334, 2011. URL <http://jmlr.csail.mit.edu/papers/v12/jenatton11a.html>.
- K. Knight and W. Fu. Asymptotics for lasso-type estimators. *Ann. Stat.*, 28(5):1356–1378, 2000. doi: [doi:10.1214/aos/1015957397](https://doi.org/10.1214/aos/1015957397). URL <http://dx.doi.org/10.1214/aos/1015957397>.
- M. Kolar, J. Lafferty, and L. Wasserman. Union support recovery in multi-task learning. *J. Mach. Learn. Res.*, 12:2415–2435, 2011.
- G.R.G. Lanckriet, N. Cristianini, P. Bartlett, L. El Ghaoui, and M.I. Jordan. Learning the kernel matrix with semidefinite programming. *J. Mach. Learn. Res.*, 5:27–72, 2004. URL <http://www.jmlr.org/papers/v5/lanckriet04a.html>.
- C. Leng, Y. Lin, and G. Wahba. A note on the Lasso and related procedures in model selection. *Statistica Sinica*, 16(4):1273–1284, 2004.
- K. Lounici. Sup-norm convergence rate and sign concentration property of lasso and dantzig estimators. *Electron. J. Statist.*, 2:90–102, 2008. doi: [10.1214/08-EJS177](https://doi.org/10.1214/08-EJS177). URL <http://dx.doi.org/10.1214/08-EJS177>.
- K. Lounici, M. Pontil, A.B. Tsybakov, and S. Van De Geer. Oracle inequalities and optimal inference under group sparsity. Technical Report 1007.1771, arXiv, 2010. URL <http://arxiv.org/abs/1007.1771>. To appear in the Annals of Statistics.
- Karim Lounici, Massimiliano Pontil, Alexandre B. Tsybakov, and Sara van de Geer. Taking advantage of sparsity in multi-task learning. In *Proceedings of COLT*, 2009.
- A. Maurer and M. Pontil. Structured sparsity and generalization. Technical Report 1108.3476, arXiv, 2011. URL <http://arxiv.org/abs/1108.3476>.

- L. Meier, S. van de Geer, and P. Bühlmann. The group lasso for logistic regression. *J. R. Stat. Soc. Ser. B*, 70(1):53–71, 2008. doi: 10.1111/j.1467-9868.2007.00627.x. URL <http://dx.doi.org/10.1111/j.1467-9868.2007.00627.x>.
- C.A. Micchelli, J.M. Morales, and M. Pontil. Regularizers for structured sparsity. Technical Report 1010.0556, arXiv, 2011. URL <http://arxiv.org/abs/1010.0556>.
- S. Mosci, S. Villa, A. Verri, and L. Rosasco. A primal-dual algorithm for group sparse regularization with overlapping groups. In J. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R.S. Zemel, and A. Culotta, editors, *Adv. Neural. Inform. Process Syst.*, pages 2604–2612. Kaufmann publishers, 2010.
- S.N. Negahban and M.J. Wainwright. Simultaneous support recovery in high dimensions: Benefits and perils of block ℓ_1/ℓ_∞ -regularization. *Information Theory, IEEE Transactions on*, 57(6):3841–3863, 2011.
- J. Nocedal and S. Wright. *Numerical optimization*. Springer, 2006.
- G. Obozinski and Bach F. Convex relaxation of combinatorial penalties. Technical report, 2011. In preparation.
- G. Obozinski, B. Taskar, and M.I. Jordan. Joint covariate selection and joint subspace selection for multiple classification problems. *Statistics and Computing*, 20(2):231–252, 2010.
- D. Percival. Theoretical properties of the overlapping groups lasso. Technical Report 1103.4614, arXiv, 2011. URL <http://arxiv.org/abs/1103.4614>.
- A. Rakotomamonjy, F. Bach, S. Canu, and Y. Grandvalet. SimpleMKL. *J. Mach. Learn. Res.*, 9:2491–2521, 2008.
- R.T. Rockafellar. *Convex Analysis*. Princeton Univ. Press, 1997.
- V. Roth and B. Fischer. The group-lasso for generalized linear models: uniqueness of solutions and efficient algorithms. In *ICML '08: Proceedings of the 25th international conference on Machine learning*, pages 848–855, 2008.
- A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, and J. P. Mesirov. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. USA*, 102(43):15545–15550, Oct 2005. doi: 10.1073/pnas.0506580102. URL <http://dx.doi.org/10.1073/pnas.0506580102>.
- R. Tibshirani. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B*, 58(1):267–288, 1996.
- S. van de Geer. ℓ_1 -regularization in high-dimensional statistical models. In *Proceedings of the International Congress of Mathematicians*, volume 4, pages 2251–2369, 2010.

- M. J. van de Vijver, Y. D. He, L. J. van't Veer, H. Dai, A. A. M. Hart, D. W. Voskuil, G. J. Schreiber, J. L. Peterse, C. Roberts, M. J. Marton, M. Parrish, D. Atsma, A. Witteveen, A. Glas, L. Delahaye, T. van der Velde, H. Bartelink, S. Rodenhuis, E. T. Rutgers, S. H. Friend, and R. Bernards. A gene-expression signature as a predictor of survival in breast cancer. *N. Engl. J. Med.*, 347(25):1999–2009, Dec 2002. doi: 10.1056/NEJMoa021967. URL <http://dx.doi.org/10.1056/NEJMoa021967>.
- M. J. Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1 -constrained quadratic programming (lasso). *IEEE T. Inform. Theory.*, 55(5):2183–2202, 2009. doi: 10.1109/TIT.2009.2016018. URL <http://dx.doi.org/10.1109/TIT.2009.2016018>.
- M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. Ser. B*, 68(1):49–67, 2006.
- P. Zhao and B. Yu. On model selection consistency of lasso. *J. Mach. Learn. Res.*, 7:2541, 2006. URL <http://jmlr.csail.mit.edu/papers/v7/zhao06a.html>.
- P. Zhao, G. Rocha, and B. Yu. Grouped and hierarchical model selection through composite absolute penalties. *Ann. Stat.*, 37(6A):3468–3497, 2009.

Appendix A. Proofs of Lemmata 21 and 22

Lemmata 21 and 22 are about the continuity of the correspondences $\mathbf{w} \mapsto \mathcal{A}(\mathbf{w})$ and $\mathbf{w} \mapsto \mathbf{V}(\mathbf{w})$. In order to prove them, we start by reviewing general results in correspondence theory (Section A.1), notably Berge’s maximum theorem which is the main ingredient to prove the two lemmas. We prove Lemma 21 directly in Section A.2. We then prove several continuity properties of auxiliary correspondences in Section A.3 and A.4 in order to finally prove Lemma 22 in Section A.5.

A.1 Elements of correspondence theory

We start with a couple of useful technical lemmas from correspondence theory.

Lemma 37 *If f is a continuous function at p and ϕ is a correspondence u.h.c. (resp. l.h.c.) at $f(p)$, then $\phi \circ f$ is a correspondence u.h.c. (resp. l.h.c.) at p . If $\phi : P \rightarrow X$ is a correspondence u.h.c. (resp. l.h.c.) at p and f is a continuous function on X then $f \circ \phi$ is a correspondence u.h.c. (resp. l.h.c.) at p .*

Proof The proofs are straightforward from the definitions. ■

Lemma 38 *An elementwise product of u.h.c. (resp. l.h.c.) correspondences is itself u.h.c. (resp. l.h.c.).*

Proof It is easy to check that a cartesian product of l.h.c. (resp. u.h.c.) correspondences has itself the same property. Moreover, the product is a continuous application, so the

result is proved by Lemma 37. ■

We now state without proof the celebrated maximum theorem (Berge, 1959).

Theorem 39 (Berge maximum theorem) *Let $\phi : P \rightrightarrows X$ be a compact-valued correspondence. Let $f : X \times P \rightarrow \mathbb{R}$ be a continuous real valued function. Define the “argmax” correspondence $\mu : P \rightrightarrows X$ by $\mu(p) = \{x \in \phi(p) \mid f(x, p) = \max_{x' \in \phi(p)} f(x', p)\}$. If ϕ is continuous at p , then μ is non-empty, compact-valued and u.h.c. at p .*

A.2 Proof of Lemma 21

Lemma 21 is a simple consequence of Theorem 39. Indeed, remember that, by definition, $\mathcal{A}(\mathbf{w}) = \operatorname{argmax}_{\boldsymbol{\alpha}} \boldsymbol{\alpha}^\top \mathbf{w}$ s.t. $\Omega^*(\boldsymbol{\alpha}) \leq 1$. Since $(\boldsymbol{\alpha}, \mathbf{w}) \mapsto \boldsymbol{\alpha}^\top \mathbf{w}$ is continuous and since the correspondence $\mathbf{w} \mapsto \{\boldsymbol{\alpha} \in \mathbb{R}^p \mid \Omega^*(\boldsymbol{\alpha}) \leq 1\}$ is compact-valued and continuous (it is constant), Theorem 39 applies and shows that the correspondence $\mathbf{w} \mapsto \mathcal{A}(\mathbf{w})$ is u.h.c. (For more general results on the continuity of the subdifferential viewed as a multi-function see Hiriart-Urruty and Lemaréchal (1994, chap. VI.6.2 p. 282)). ■

A.3 Continuity properties of $\mathbf{V}(\mathbf{w})$, $\boldsymbol{\Lambda}(\mathbf{w})$ and $\mathbf{Z}(\mathbf{w})$

The fact that $\mathbf{w} \mapsto \mathbf{V}(\mathbf{w})$ is u.h.c. is also a direct consequence of Berge’s maximum theorem. We show this in the following two lemmata.

Lemma 40 *The correspondence ϕ defined by*

$$\phi(\mathbf{w}) = \{\bar{\mathbf{v}} \in \mathcal{V}^g \mid \mathbf{w} = \sum \mathbf{v}^g, \operatorname{sign}(\mathbf{v}_i^g) = \operatorname{sign}(\mathbf{w}_i), 1 \leq i \leq p\} \quad (29)$$

is a continuous correspondence.

Proof We have $\phi(\mathbf{w}) = \prod_{i=1}^p \phi_i(\mathbf{w}_i)$ with

$$\phi_i(\mathbf{w}_i) = \{(\mathbf{v}_i^g)_{g \in \mathcal{G}} \in \mathbb{R}^m \mid \mathbf{w}_i = \sum_{g \in \mathcal{G}} \mathbf{v}_i^g, \forall i \in g, \operatorname{sign}(\mathbf{v}_i^g) = \operatorname{sign}(\mathbf{w}_i), \text{ and } \mathbf{v}_i^g = 0, i \notin g\}.$$

It is easy to verify that a Cartesian product of compact-valued continuous correspondences is also continuous, so that we only need to show that ϕ_i is compact-valued and continuous. We therefore focus on $\phi_i(\mathbf{w}_i) \subset \mathbb{R}^m$. First note that ϕ_i is compact valued because the sign constraints in the definition of ϕ_i imply that for all $\mathbf{v}_i = (\mathbf{v}_i^g)_{g \in \mathcal{G}} \in \phi_i(\mathbf{w}_i)$ we have $\|\mathbf{v}_i\|_1 \leq \|\mathbf{w}_i\|_1$. We first show that ϕ_i is u.h.c.. Let U be an open set containing $\phi_i(\mathbf{w}_i)$. For two sets $A, B \subset \mathbb{R}^m$, we define $d_\infty(A, B) \triangleq \inf_{a \in A, b \in B} \|a - b\|_\infty$. Let $u_0 \in U^c$, $d_0 \triangleq d_\infty(\{u_0\}, \phi_i(\mathbf{w}_i))$ and define $K \triangleq \{u \in \mathbb{R}^m \mid d_\infty(\{u\}, \phi_i(\mathbf{w}_i)) \leq d_0\}$. By construction $K \cap U^c \neq \emptyset$, and we have $d_\infty(U^c, \phi_i(\mathbf{w}_i)) = d_\infty(U^c \cap K, \phi_i(\mathbf{w}_i))$. Moreover, it is classical to show that the compactness of $\phi_i(\mathbf{w}_i)$ implies that K is compact as well. Since $U^c \cap K$ and $\phi_i(\mathbf{w}_i)$ are compact sets the infimum in the definition of d_∞ is attained, which means that there are $\mathbf{u}^* \in U^c \cap K$ and $\mathbf{v}^* \in \phi_i(\mathbf{w}_i)$ such that $d_\infty(U^c \cap K, \phi_i(\mathbf{w}_i)) = \|\mathbf{u}^* - \mathbf{v}^*\|_\infty$. But we must have $\|\mathbf{u}^* - \mathbf{v}^*\|_\infty > 0$ otherwise $\mathbf{u}^* = \mathbf{v}^* \in U^c \cap \phi_i(\mathbf{w}_i)$ which would contradict the

hypothesis that $\phi_i(\mathbf{w}_i) \subset U$. If $\varepsilon \triangleq \|\mathbf{u}^* - \mathbf{v}^*\|_\infty/2$, we just showed that for all $\delta \in \mathbb{R}^m$ such that $\|\delta\|_\infty \leq \varepsilon$, $\phi_i(\mathbf{w}_i) + \delta \subset U$.

If $\mathbf{w}_i = 0$, then any decomposition of $\mathbf{w}_i \pm \varepsilon$, say $\check{\mathbf{v}}_i$ is such that $\|\check{\mathbf{v}}_i\|_\infty \leq \varepsilon$, and $\phi(\mathbf{w}_i \pm \varepsilon) \subset U$. If $\mathbf{w}_i \neq 0$, w.l.o.g. assume that $\mathbf{w}_i > 0$; consider a decomposition $\check{\mathbf{v}}_i \in \mathbb{R}^m$ of $\mathbf{w}_i + \varepsilon'$ with $|\varepsilon'| \leq \min(\varepsilon, |\mathbf{w}_i|/2)$; if $\varepsilon' < 0$ then $\mathbf{v}_i \triangleq \check{\mathbf{v}}_i + \varepsilon' \mathbf{e}_1$ is a decomposition of \mathbf{w}_i and $\|\mathbf{v}_i - \check{\mathbf{v}}_i\|_\infty \leq \varepsilon'$; if $\varepsilon' > 0$ then it is easy to show that the projection \mathbf{v}_i of $\check{\mathbf{v}}_i$ on the simplex $\phi(\mathbf{w}_i)$ satisfies $\|\mathbf{v}_i - \check{\mathbf{v}}_i\|_\infty < \varepsilon'$. In all cases $\phi(\mathbf{w}_i + \varepsilon') \subset U$ for some $\varepsilon > 0$, which shows that ϕ is u.h.c..

We can show similarly that ϕ is l.h.c. : if $\mathbf{v}_i \in U \cap \phi(\mathbf{w}_i)$, then for some $\varepsilon > 0$, U contains a closed ℓ_∞ ball of radius ε centered at \mathbf{v}_i , which contains a decomposition of $\mathbf{w}_i \pm \varepsilon$ so that $U \cap \phi(\mathbf{w}_i \pm \varepsilon) \neq \emptyset$. ■

Lemma 41 *The correspondence $\mathbf{w} \mapsto \mathbf{V}(\mathbf{w})$ is compact-valued and u.h.c.*

Proof Define $f(\bar{\mathbf{v}}, \mathbf{w}) = \sum_{g \in \mathcal{G}} \|\mathbf{v}^g\|$ and ϕ as in (29).

We have that $\mathbf{V}(\mathbf{w}) = \text{Argmin}_{\bar{\mathbf{v}} \in \phi(\mathbf{w})} f(\bar{\mathbf{v}}, \mathbf{w})$ since it can be shown easily that any optimal decomposition satisfies $\text{sign}(\mathbf{v}_i^g) = \text{sign}(\mathbf{w}_i)$.

Since the previous lemma shows that ϕ is a compact-valued continuous correspondence, theorem 39 applies and proves the result. ■

Remember that $\Lambda(\mathbf{w}) \subset \mathbb{R}^m$ is the set of solutions to (10). For a vector $\boldsymbol{\lambda} \in \mathbb{R}^m$ we consider the vector $\boldsymbol{\zeta}(\boldsymbol{\lambda}) \in \mathbb{R}^p$ defined by $\zeta_i(\boldsymbol{\lambda}) = \sum_{g \ni i} \lambda_g$, and denote $\mathbf{Z}(\mathbf{w}) = \{\boldsymbol{\zeta}(\boldsymbol{\lambda}) \in \mathbb{R}^p, \boldsymbol{\lambda} \in \Lambda(\mathbf{w})\}$.

Lemma 42 *$\Lambda(\mathbf{w})$ and $\mathbf{Z}(\mathbf{w})$ are u.h.c. correspondences.*

Proof Since \mathbf{V} is u.h.c., by lemma 37, the continuity of $(\mathbf{v}^g)_{g \in \mathcal{G}} \mapsto (\|\mathbf{v}^g\|)_{g \in \mathcal{G}}$ shows that $\Lambda(\mathbf{w})$ is u.h.c. and the continuity of $\boldsymbol{\lambda} \mapsto (\sum_{g \ni i} \lambda_g)_{1 \leq i \leq p}$ shows that $\mathbf{Z}_i(\mathbf{w})$ is u.h.c.. ■

Lemma 43 *For all i such that $\mathbf{w}_i \neq 0$, $\mathbf{Z}_i(\mathbf{w})$ is a singleton, and if we denote this unique value by $\zeta_i(\mathbf{w})$ then the function $\mathbf{w}' \mapsto \zeta_i(\mathbf{w}')$ is uniquely defined in a neighborhood of \mathbf{w} and it is continuous at \mathbf{w} .*

Proof Uniqueness of $\zeta_i(\mathbf{w})$ at \mathbf{w} such that $\mathbf{w}_i \neq 0$ is granted by the fact that if $\mathbf{w}_i \neq 0$, then $\alpha_i \neq 0$, α_i is unique (cf lemma 9) and the proof of lemma 6 shows that $\zeta_i = \frac{w_i}{\alpha_i}$. Thus, $\zeta_i(\mathbf{w})$ is unique, but so is $\zeta_i(\mathbf{w}')$ for \mathbf{w}' in a small neighborhood of \mathbf{w} since $\mathbf{w}'_i \neq 0$.

Moreover we have $\zeta_i(\mathbf{w}) = \sum_{g \in \mathcal{G}} \lambda_g$ for any $\boldsymbol{\lambda} \in \Lambda(\mathbf{w})$. Finally the upper hemicontinuity of $\mathbf{w} \mapsto \mathbf{Z}_i(\mathbf{w})$ shown in the previous lemma implies the continuity of ζ_i . ■

Lemma 44 Let $\mathcal{S} = \{\mathbf{u} \in \mathbb{R}^p \mid \text{supp}(\mathbf{u}) \subset J_1\}$. Consider \mathbf{w} such that $\forall i \in J_1$ and for all \mathbf{u} in a neighborhood of $\mathbf{0}$ in \mathcal{S} , $Z_i(\mathbf{w} + \mathbf{u})$ is a singleton, then if $\Pi_{\mathcal{G}_1}$ denotes the projection on $\{\boldsymbol{\lambda} \in \mathbb{R}^m \mid \boldsymbol{\lambda}_{\mathcal{G}_1^c} = \mathbf{0}\}$ we have that

$$\begin{aligned} \Lambda|_{J_1}^{\mathcal{G}_1} : \mathcal{S} &\rightarrow \mathbb{R}^{|\mathcal{G}_1|} \\ \mathbf{w}' &\mapsto \Pi_{\mathcal{G}_1} \Lambda(\mathbf{w}') \end{aligned}$$

is a lower hemicontinuous correspondence at \mathbf{w} .

Proof Let $\mathbf{B} \in \mathbb{R}^{p \times m}$ the adjacency matrix associated to \mathcal{G} , defined by $\mathbf{B}_{ig} = 1$ if $i \in g$ and 0 else. To simplify notations we denote $\tilde{\mathbf{B}} = \mathbf{B}_{J_1 \mathcal{G}_1}$ the submatrix obtained by keeping rows in J_1 and columns in \mathcal{G}_1 , $\tilde{\boldsymbol{\zeta}} = \boldsymbol{\zeta}_{J_1}(\mathbf{w}')$ and $\tilde{\boldsymbol{\Lambda}} = \Pi_{\mathcal{G}_1} \Lambda(\mathbf{w}')$. Given $\tilde{\boldsymbol{\zeta}}$, then $\tilde{\boldsymbol{\Lambda}} = \{\tilde{\boldsymbol{\lambda}} \in \mathbb{R}_+^{|\mathcal{G}_1|} \mid \tilde{\boldsymbol{\zeta}} = \tilde{\mathbf{B}} \tilde{\boldsymbol{\lambda}}\}$ which means that if $\tilde{\mathbf{B}}^+$ denotes the Moore-Penrose pseudo-inverse of $\tilde{\mathbf{B}}$ then $\tilde{\boldsymbol{\Lambda}} = (\tilde{\mathbf{B}}^+ \tilde{\boldsymbol{\zeta}} + \text{Ker}(\tilde{\mathbf{B}})) \cap \mathbb{R}_+^{|\mathcal{G}_1|}$.

We now show that this correspondence is l.h.c.. The uniqueness of $\tilde{\boldsymbol{\zeta}}$ implies its continuity, since by lemma 42, $Z_i(\mathbf{w})$ is u.h.c.. Denoting by \mathbf{H} a matrix whose columns form a basis of $\text{Ker}(\tilde{\mathbf{B}})$, \mathbf{h}^g and \mathbf{b}^g the g^{th} row of \mathbf{H} and $\tilde{\mathbf{B}}^+$ respectively, then an element of $\tilde{\boldsymbol{\Lambda}}$ is of the form $(\mathbf{b}^g \tilde{\boldsymbol{\zeta}} + \mathbf{h}^g \mathbf{q})_{g \in \mathcal{G}_1}$ for some \mathbf{q} . Given an element $\tilde{\mathbf{B}}^+ \tilde{\boldsymbol{\zeta}} + \mathbf{H} \mathbf{q} \in U \cap \mathbb{R}_+^{|\mathcal{G}_1|}$, we show that there exists an element $\lambda(\mathbf{w} + \mathbf{u}, \mathbf{q}') \triangleq \tilde{\mathbf{B}}^+ \tilde{\boldsymbol{\zeta}}(\mathbf{w} + \mathbf{u}) + \mathbf{H} \mathbf{q}' \in U \cap \mathbb{R}_+^{|\mathcal{G}_1|}$ for \mathbf{u} in neighborhood of $\mathbf{0}$ in \mathcal{S} . Without loss of generality we can take U a cartesian product of open sets $U = \bigotimes_{g \in \mathcal{G}_1} U_g$.

Let $\mathcal{Q} = \{\mathbf{q}' \mid \tilde{\mathbf{B}}^+ \tilde{\boldsymbol{\zeta}}(\mathbf{w}) + \mathbf{H} \mathbf{q}' \in \mathbb{R}_+^{|\mathcal{G}_1|}\}$. For all $g \in \check{\mathcal{G}}_1$, there exists $\mathbf{q}^{(g)} \in \mathcal{Q}$ such that $\mathbf{b}^g \tilde{\boldsymbol{\zeta}} + \mathbf{h}^g \mathbf{q}^{(g)} > 0$. Set $\mathbf{q}' = (1 - \epsilon) \mathbf{q} + \frac{\epsilon}{|\check{\mathcal{G}}_1|} \sum_{g \in \check{\mathcal{G}}_1} \mathbf{q}^{(g)}$. For ϵ sufficiently small, $\lambda_g(\mathbf{w}, \mathbf{q}') \in U_g \cap \mathbb{R}_+^*$, for all $g \in \check{\mathcal{G}}_1$ so that for \mathbf{u} sufficiently small $\lambda_g(\mathbf{w} + \mathbf{u}, \mathbf{q}') \in U_g \cap \mathbb{R}_+^*$ as well. For all $g \notin \check{\mathcal{G}}_1$, $\Lambda_g(\mathbf{w}) = \{0\}$ and since Λ is u.h.c., for any $\eta > 0$, for \mathbf{u} sufficiently small we have $\Lambda_g(\mathbf{w} + \mathbf{u}) \subset [0, \eta)$, $g \notin \check{\mathcal{G}}_1$. Choosing η such that $\forall g \notin \check{\mathcal{G}}_1$, $[0, \eta) \subset U_g$ shows the result. \blacksquare

A.4 Continuity properties of \mathcal{G}_1 and $\check{\mathcal{G}}_1$

Lemma 45 There exists a neighborhood U of $\mathbf{0}$ in \mathbb{R}^p such that for all $\mathbf{u} \in U$ with $\text{supp}(\mathbf{u}) \subset J_1(\mathbf{w})$, $\mathcal{G}_1(\mathbf{w} + \mathbf{u}) \subset \mathcal{G}_1(\mathbf{w})$.

Proof By definition of $\mathcal{G}_1(\mathbf{w} + \mathbf{u})$, if $g \in \mathcal{G}_1(\mathbf{w} + \mathbf{u})$, then $\boldsymbol{\alpha}_g(\mathbf{w} + \mathbf{u})$ is unique by lemma 15, since $g \subset J_1(\mathbf{w} + \mathbf{u})$. For any $g \in \mathcal{G}_1(\mathbf{w} + \mathbf{u})$, $g \cap J_1(\mathbf{w}) \neq \emptyset$; indeed if $g \cap J_1(\mathbf{w}) = \emptyset$, then $\mathbf{w}_g = \mathbf{u}_g = \mathbf{0}$. If $g \subset J_1(\mathbf{w})$, $\boldsymbol{\alpha}_g(\mathbf{w})$ is unique and since $\boldsymbol{\alpha}_g(\mathbf{w} + \mathbf{u})$ is unique, the upper hemicontinuity of \mathcal{A} implies that $\boldsymbol{\alpha}_g$ is continuous at \mathbf{w} so that $(\|\boldsymbol{\alpha}_g(\mathbf{w} + \mathbf{u})\| = 1 \Rightarrow \|\boldsymbol{\alpha}_g(\mathbf{w})\| = 1)$. If $g \setminus J_1(\mathbf{w}) \neq \emptyset$, then it has to be the case that $\boldsymbol{\alpha}_{g \setminus J_1(\mathbf{w})}(\mathbf{w} + \mathbf{u}) = \mathbf{0}$, because it is indeed a possible value for $\boldsymbol{\alpha}_{g \setminus J_1(\mathbf{w})}(\mathbf{w} + \mathbf{u})$ (given that $\mathbf{w}_{g \setminus J_1(\mathbf{w})} = \mathbf{u}_{g \setminus J_1(\mathbf{w})} = \mathbf{0}$) and because $\boldsymbol{\alpha}_g(\mathbf{w} + \mathbf{u})$ is unique. This implies that $\|(\boldsymbol{\alpha}_{g \cap J_1(\mathbf{w})}(\mathbf{w} + \mathbf{u}))\| = 1$ and since $\boldsymbol{\alpha}_{g \cap J_1(\mathbf{w})}(\mathbf{w})$ is unique, upper hemicontinuity of \mathcal{A} implies that $\mathbf{w}' \mapsto \boldsymbol{\alpha}_{g \cap J_1(\mathbf{w})}(\mathbf{w}')$ is continuous at \mathbf{w} so that we have by continuity

$\|\alpha_g(\mathbf{w})\| \geq \|\alpha_{g \cap J_1}(\mathbf{w})\| = 1$ which proves that $\|\alpha_g(\mathbf{w})\| = 1$; but this is a contradiction because this would imply $g \in \mathcal{G}_1$ and therefore $g \subset J_1$. \blacksquare

Lemma 46 *Let $\mathcal{D}_{J_1} = \{\mathbf{u} \in \mathbb{R}^p \mid \|\mathbf{u}\| \leq 1, \mathbf{u}_{J_1^c} = \mathbf{0}\}$; then*

$$\mathcal{G}_1(\mathbf{w}) = \bigcap_{\epsilon > 0} \bigcup_{\mathbf{u} \in \mathcal{D}_{J_1}} \check{\mathcal{G}}_1(\mathbf{w} + \epsilon \mathbf{u}).$$

Proof One inclusion is already shown by the previous Lemma 45. For the other inclusion, let $\bar{\mathbf{v}}$ be an optimal decomposition of \mathbf{w} and α the unique element of $\mathcal{A}(\mathbf{w})$ such that $\alpha_{J_1^c} = \mathbf{0}$. Let $\lambda_g = \|\mathbf{v}_g\|$. The case of $g \in \check{\mathcal{G}}_1(\mathbf{w})$ is straightforward, and we concentrate therefore on $g \in \mathcal{G}_1(\mathbf{w}) \setminus \check{\mathcal{G}}_1(\mathbf{w})$. By lemma 9, we have $\mathbf{w} = \sum_{g \in \check{\mathcal{G}}_1} \lambda_g \alpha_g$. Consider $\mathbf{w}_{(g_0, \epsilon)} = \mathbf{w} + \epsilon \alpha_{g_0}$ for some $g_0 \in \mathcal{G}_1(\mathbf{w}) \setminus \check{\mathcal{G}}_1(\mathbf{w})$. By construction, $\alpha \in \mathcal{D}_{J_1}$ and for all $\beta \in \mathbb{R}^p$ such that $\Omega^*(\beta) \leq 1$ we have

$$\mathbf{w}_{(g_0, \epsilon)}^\top \beta = \sum_{g \in \mathcal{G}} \lambda_g \alpha_g^\top \beta_g + \epsilon \alpha_{g_0}^\top \beta_{g_0} \leq \sum_{g \in \mathcal{G}} \lambda_g + \epsilon = \mathbf{w}_{(g_0, \epsilon)}^\top \alpha$$

which shows that $\bar{\mathbf{v}}'$ defined by $\mathbf{v}'_{g_0} = \epsilon \alpha_{g_0}$ and $\mathbf{v}'_g = \mathbf{v}_g$, $g \neq g_0$ is an optimal decomposition of $\mathbf{w}_{(g_0, \epsilon)}$ with group-support $\check{\mathcal{G}}_1(\mathbf{w}) \cup g_0$. Since this is true for any ϵ and any $g_0 \in \mathcal{G}_1(\mathbf{w}) \setminus \check{\mathcal{G}}_1(\mathbf{w})$, this proves the statement. \blacksquare

A.5 Proof of Lemma 22

We know from Lemma 41 that $\mathbf{w} \mapsto \mathbf{V}(\mathbf{w})$ is a compact-valued u.h.c. correspondence. If $\text{supp}(\mathbf{w}) = J_1$ then lemma 43 implies that for all $i \in J_1$, $\zeta_i(\mathbf{w} + \mathbf{u})$ is unique for all \mathbf{u} in a neighborhood of $\mathbf{0}$. From lemma 44, this implies that $\mathbf{u} \mapsto \Pi_{\mathcal{G}_1} \Lambda(\mathbf{w} + \mathbf{u})$ is l.h.c at $\mathbf{u} = \mathbf{0}$. This extends to $\mathbf{u} \mapsto \Lambda(\mathbf{w} + \mathbf{u})$ since we know from Lemma 45 that there exists a neighborhood of zero such that, for all \mathbf{u} in that neighborhood, $\Pi_{\mathcal{G}_1^c} \Lambda(\mathbf{w} + \mathbf{u}) = \mathbf{0}$. Given that $\mathbf{V}(\mathbf{w} + \mathbf{u}) = \alpha(\mathbf{w} + \mathbf{u}) \Lambda(\mathbf{w} + \mathbf{u})$, since $\alpha(\mathbf{w})$ is l.h.c. from Lemma 21 and since a product of l.h.c. correspondences is l.h.c. (cf. Lemma 38), we have shown that $\mathbf{u} \mapsto \mathbf{V}(\mathbf{w} + \mathbf{u})$ is also l.h.c. at $\mathbf{u} = \mathbf{0}$. \blacksquare

Appendix B. Partial group-support recovery

Theorem 23, which only assumes hypothesis (H1), does not give a lower bound (in the sense of inclusion) for $\check{\mathcal{G}}_1(\mathbf{w})$, suggesting that hypothesis (H2) is necessary to guarantee group-support recovery. In this section, we first consider an example in which $\check{\mathcal{G}}_1(\mathbf{w})$ is strictly included in $\check{\mathcal{G}}_1(\mathbf{w}^*)$.

Example with partial recovery. Take $\mathcal{G} = \{\{0, 1, 2\}, \{0, 1, 3\}, \{0, 2, 3\}\}$ for $\mathbf{w} = (w_0, w_1, w_2, w_3) \in \mathbb{R}^4$. It is easy to check that $\lambda_{\{0, 1, 2\}} = \gamma(|w_1| + |w_2| - |w_3|)_+$, $\lambda_{\{0, 1, 3\}} = \gamma(|w_1| + |w_3| - |w_2|)_+$ and $\lambda_{\{0, 2, 3\}} = \gamma(|w_2| + |w_3| - |w_1|)_+$ with γ determined by the equation

$\sum_{i=0}^2 \frac{w_i^2}{\zeta_i^2} = 1$. In particular if we consider $\mathbf{w}^* = (1, 0, 0, 0)$, then taking the identity as the design matrix and assuming independent Gaussian noise, we have $y = (1 + \epsilon_0, \epsilon_1, \epsilon_2, \epsilon_3)$ with ϵ_i i.i.d. $\mathcal{N}(0, \sigma^2)$. Thus solving the first order approximation of the KKT in the neighborhood of \mathbf{w}^* we get $\mathbf{w} = ((1 + \epsilon_0 - \lambda)_+, \epsilon_1, \epsilon_2, \epsilon_3)$. We have $\check{\mathcal{G}}_1(\mathbf{w}^*) = \mathcal{G}_1(\mathbf{w}^*) = \mathcal{G}$ but for any value of σ^2 , with probability μ, μ, μ and $1 - 3\mu$, $\check{\mathcal{G}}_1(\mathbf{w})$ takes respectively the values $\mathcal{G} \setminus \{0, 1, 2\}$, $\mathcal{G} \setminus \{0, 1, 3\}$, $\mathcal{G} \setminus \{0, 2, 3\}$ and \mathcal{G} , with $\mu \approx 0.216$.

However, the following lemma shows that the group-support recovered contains at least the group-support of one of the decomposition of the true support.

Lemma 47 *If \mathbf{w}_n is a sequence converging to \mathbf{w} , then denoting $\text{gsupp}(\bar{\mathbf{v}})$ the group support of a decomposition $\bar{\mathbf{v}}$, we have*

$$\exists n_0, \forall n \geq n_0, \forall \bar{\mathbf{v}}_n \in \mathbf{V}(\mathbf{w}_n), \exists \bar{\mathbf{v}} \in \mathbf{V}(\mathbf{w}), \text{gsupp}(\bar{\mathbf{v}}) \subset \text{gsupp}(\bar{\mathbf{v}}_n).$$

Proof Reason by contradiction and assume that

$$\forall n_0, \exists n \geq n_0, \exists \bar{\mathbf{v}}_n \in \mathbf{V}(\mathbf{w}_n), \forall \bar{\mathbf{v}} \in \mathbf{V}(\mathbf{w}), \text{gsupp}(\bar{\mathbf{v}}) \not\subset \text{gsupp}(\bar{\mathbf{v}}_n).$$

We can therefore extract a subsequence $(\mathbf{w}_{\varphi(n)})_n$ with this property and the corresponding subsequence $(\bar{\mathbf{v}}_{\varphi(n)})_n$ illustrating it. There exists at least one $\mathcal{G}_0 \in 2^{\mathcal{G}}$ such that there are infinitely many elements $\bar{\mathbf{v}}_{\varphi(n)}$ in the subsequence which satisfies $\text{gsupp}(\bar{\mathbf{v}}_{\varphi(n)}) = \mathcal{G}_0$. We consider the subsequence $(\bar{\mathbf{v}}_{\varphi'(n)})_n$ composed of those elements. From the sequence $(\bar{\mathbf{v}}_{\varphi'(n)})_n$, since we can assume without loss of generality it lives in the compact set $\{\bar{\mathbf{v}} \mid \forall g \in \mathcal{G}, \|\mathbf{v}^g\| \leq 2\|\mathbf{w}\|\}$, we can extract a converging subsequence $(\bar{\mathbf{v}}_{\varphi''(n)})_n$. Since $(\mathbf{w}_{\varphi''(n)})_n$ converges to \mathbf{w} and by upper hemicontinuity of $\mathbf{V}(\cdot)$ the subsequence $(\bar{\mathbf{v}}_{\varphi''(n)})_n$ converges to an optimal decomposition $\bar{\mathbf{v}}_\infty$ of \mathbf{w} . This implies that $\text{gsupp}(\bar{\mathbf{v}}_\infty) \subset \mathcal{G}_0 = \text{gsupp}(\bar{\mathbf{v}}_{\varphi''(n)})$ which is a contradiction. \blacksquare

The simpler example with $\mathcal{G} = \{\{1, 2\}, \{2, 3\}\}$ and $\mathbf{w}^* = (0, 1, 0)$ could be expected to be problematic since $(0, 1, \epsilon)$ and $(\epsilon, 1, 0)$ have respectively group-support $\{\{2, 3\}\}$ and $\{\{1, 2\}\}$. However, this case is consistent since it can be shown that \mathbf{w}_1 and \mathbf{w}_3 are almost surely non-zero, which implies that both groups are part of the group-support.

Appendix C. Derivations for the illustrative examples

C.1 Graph Lasso for the cycle of length 3

We consider the overlap norm in \mathbb{R}^3 with groups $\mathcal{G} = \{\{1, 2\}, \{1, 3\}, \{2, 3\}\}$. If $\boldsymbol{\alpha}$ denotes a dual variable. The dual norm takes the form:

$$\Omega^*(\boldsymbol{\alpha}) \triangleq \max (\|(\alpha_1, \alpha_2)\|, \|(\alpha_1, \alpha_3)\|, \|(\alpha_2, \alpha_3)\|)$$

By Fenchel duality, $\Omega(\mathbf{w}) = \max_{\boldsymbol{\alpha} \in \mathbb{R}^3} \boldsymbol{\alpha}^\top \mathbf{w}$ s.t. $\max_{g \in \mathcal{G}} \|\boldsymbol{\alpha}_g\|^2 \leq 1$. Consider the Lagrangian

$$\begin{aligned} L^*(\boldsymbol{\alpha}, \lambda, \mathbf{w}) &= -(\alpha_1 w_1 + \alpha_2 w_2 + \alpha_3 w_3) \\ &\quad + \frac{1}{2} [(\lambda_{12} + \lambda_{13}) \alpha_1^2 + (\lambda_{12} + \lambda_{23}) \alpha_2^2 + (\lambda_{13} + \lambda_{23}) \alpha_3^2 - (\lambda_{12} + \lambda_{13} + \lambda_{23})] \end{aligned}$$

and consider the optimization problem $\min_{\alpha \in \mathbb{R}^p} L^*(\alpha, \lambda, \mathbf{w})$ s.t. $\lambda_g \geq 0, g \in \mathcal{G}$.

A singular point of the Lagrangian satisfies

$$w_1 = (\lambda_{12} + \lambda_{13}) \alpha_1, \quad w_2 = (\lambda_{12} + \lambda_{23}) \alpha_2, \quad w_3 = (\lambda_{13} + \lambda_{23}) \alpha_3. \quad (30)$$

C.1.1 AT MOST TWO GROUPS ARE ACTIVE

Assume that $\lambda_{13} = 0$. Note that this case reduces to the case of $\mathcal{G} = \{\{1, 2\}, \{2, 3\}\}$, which is of interest on its own. Eq. 30 simplifies and the singular points of the Lagrangian solve

$$w_1 = (\lambda_{12}) \alpha_1, \quad w_2 = (\lambda_{12} + \lambda_{23}) \alpha_2, \quad w_3 = (\lambda_{23}) \alpha_3. \quad (31)$$

We assume first that $\lambda_{12} > 0, \lambda_{23} > 0, |w_1| > 0, |w_3| > 0$. Since, by complementary slackness, $\|\alpha_{12}\| = 1$ and $\|\alpha_{23}\| = 1$, using (30), we have

$$\frac{w_1^2}{\lambda_{12}^2} + \frac{w_2^2}{(\lambda_{12} + \lambda_{23})^2} = 1 \quad \text{and} \quad \frac{w_2^2}{(\lambda_{12} + \lambda_{23})^2} + \frac{w_3^2}{\lambda_{23}^2} = 1. \quad (32)$$

So that $\frac{w_1^2}{\lambda_{12}^2} = \frac{w_2^2}{\lambda_{23}^2}$ or equivalently $\lambda_{23} = \frac{|w_3|}{|w_1|} \lambda_{12}$ and by substitution in (32) we get respectively:

$$\lambda_{12} = \frac{|w_1|}{|w_1| + |w_3|} \|(w_2, |w_1| + |w_3|)\| \quad \text{and} \quad \lambda_{23} = \frac{|w_3|}{|w_1| + |w_3|} \|(w_2, |w_1| + |w_3|)\|.$$

Substituting these expressions for λ_{12} and λ_{23} in the singular point equations (31), we get:

$$\alpha_1 = \text{sign}(w_1) \frac{|w_1| + |w_3|}{\|(w_2, |w_1| + |w_3|)\|} \quad \text{and} \quad \alpha_2 = \frac{w_2}{\|(w_2, |w_1| + |w_3|)\|}. \quad (33)$$

α_3 has a similar expression as α_1 , where the roles of w_3 and w_1 are exchanged. Finally, the decomposition is:

$$v_{12} = \left(w_1, \frac{|w_1|}{|w_1| + |w_3|} w_2 \right)^\top \quad \text{and} \quad v_{23} = \left(\frac{|w_3|}{|w_1| + |w_3|} w_2, w_3 \right)^\top, \quad (34)$$

and the norm then takes the closed form $\Omega(w) = \|(w_2, |w_1| + |w_3|)\|$. Remains to consider the cases where $w_1 = 0$, or $w_3 = 0$, which we do not develop here.

C.1.2 ALL GROUPS ARE ACTIVE

We first consider the case $\lambda_{12} > 0, \lambda_{13} > 0, \lambda_{23} > 0$. By complementary slackness we have $\|\alpha_g\| = 1, g \in \mathcal{G}$. Introducing $\zeta_1 = \lambda_{12} + \lambda_{13}, \zeta_2 = \lambda_{12} + \lambda_{23}$ and $\zeta_3 = \lambda_{13} + \lambda_{23}$, (30) rewrites as

$$\frac{w_1^2}{\zeta_1^2} + \frac{w_2^2}{\zeta_2^2} = 1, \quad \frac{w_2^2}{\zeta_2^2} + \frac{w_3^2}{\zeta_3^2} = 1, \quad \frac{w_1^2}{\zeta_1^2} + \frac{w_3^2}{\zeta_3^2} = 1.$$

which taking pairwise differences yields:

$$\frac{1}{\gamma} \triangleq \frac{w_1^2}{\zeta_1^2} = \frac{w_2^2}{\zeta_2^2} = \frac{w_3^2}{\zeta_3^2} \quad (35)$$

Or in other words:

$$\begin{pmatrix} |w_1| \\ |w_2| \\ |w_3| \end{pmatrix} = \frac{1}{\gamma} \begin{pmatrix} \zeta_1 \\ \zeta_2 \\ \zeta_3 \end{pmatrix} = \frac{1}{\gamma} \begin{pmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \end{pmatrix} \begin{pmatrix} \lambda_{12} \\ \lambda_{13} \\ \lambda_{23} \end{pmatrix}$$

which yields

$$\lambda_{12} = \gamma(|w_1| + |w_2| - |w_3|), \quad \lambda_{13} = \gamma(|w_1| + |w_3| - |w_2|), \quad \lambda_{23} = \gamma(|w_2| + |w_3| - |w_1|).$$

But since we have assumed $\lambda_g > 0$, the solution found is only valid if no coordinate dominates in the sense that $\mathbf{w} \in \mathcal{W}_{\text{bal}}$ with

$$\mathcal{W}_{\text{bal}} \triangleq \{ \mathbf{w} \in \mathbb{R}^3 \mid |w_1| \leq |w_2| + |w_3|, |w_2| \leq |w_1| + |w_3|, |w_3| \leq |w_1| + |w_2| \}$$

By re-substituting (35) in (30), we can solve for γ and find that

$$\boxed{\boldsymbol{\alpha} = \frac{1}{\sqrt{2}} \text{sign}(\mathbf{w}) \quad \text{and thus} \quad \Omega(\mathbf{w}) = \frac{1}{\sqrt{2}} \|\mathbf{w}\|_1}$$

The unit ball of the norm therefore has some flat faces. Finally, since $(\mathbf{v}_g)_g$ is an optimal decomposition of w we have $\mathbf{v}_g = \lambda_g \boldsymbol{\alpha}_g$, the decomposition is unique and can be written

$$\mathbf{v}_{\{12\}} = \frac{1}{2} \begin{pmatrix} w_1 + (|w_2| - |w_3|) \text{sign}(w_1) \\ w_2 + (|w_1| - |w_3|) \text{sign}(w_2) \end{pmatrix}, \quad \mathbf{v}_{\{13\}} = \frac{1}{2} \begin{pmatrix} w_1 + (|w_3| - |w_2|) \text{sign}(w_1) \\ w_3 + (|w_1| - |w_2|) \text{sign}(w_3) \end{pmatrix},$$

$$\text{and} \quad \mathbf{v}_{\{23\}} = \frac{1}{2} \begin{pmatrix} w_2 + (|w_3| - |w_1|) \text{sign}(w_2) \\ w_3 + (|w_2| - |w_1|) \text{sign}(w_3) \end{pmatrix}.$$

If $\mathbf{w} \notin \mathcal{W}_{\text{bal}}$, then one of $\lambda_{12}, \lambda_{13}$ or λ_{23} equals 0, and this reduces to the situation where only two groups are active which we considered in section C.1.1 above.

C.1.3 CLOSED FORM EXPRESSION FOR THE NORM

Finally, summarizing the analysis, we obtain the closed form expression:

$$\Omega_{\cup}^{\mathcal{G}}(w) = \begin{cases} \frac{1}{\sqrt{2}} \|\mathbf{w}\|_1 & \text{if } \mathbf{w} \in \mathcal{W}_{\text{bal}} \\ \min \begin{cases} \|(w_1, |w_2| + |w_3|)\|, \\ \|(w_2, |w_1| + |w_3|)\|, \\ \|(w_3, |w_1| + |w_2|)\| \end{cases} & \text{else.} \end{cases}$$

C.2 Graph Lasso for the cycle of length 4

We consider here the case where the groups are $\mathcal{G} = \{\{1, 2\}, \{1, 3\}, \{2, 4\}, \{3, 4\}\}$. This case is interesting because we will show that non-sparse \mathbf{w} on the cycle always admit several optimal decompositions. The dual norm takes the form:

$$\Omega^*(\boldsymbol{\alpha}) \triangleq \max (\|(\alpha_1, \alpha_2)\|, \|(\alpha_1, \alpha_3)\|, \|(\alpha_2, \alpha_4)\|, \|(\alpha_3, \alpha_4)\|)$$

We use again Fenchel duality, write $\Omega(\mathbf{w}) = \max_{\boldsymbol{\alpha} \in \mathbb{R}^4} \boldsymbol{\alpha}^\top \mathbf{w}$ s.t. $\Omega^*(\boldsymbol{\alpha})^2 \leq 1$ and we construct the Lagrangian:

$$L^*(\boldsymbol{\alpha}, \lambda, \mathbf{w}) = -(\alpha_1 w_1 + \alpha_2 w_2 + \alpha_3 w_3 + \alpha_4 w_4) + \frac{1}{2} [\zeta_1 \alpha_1^2 + \zeta_2 \alpha_2^2 + \zeta_3 \alpha_3^2 + \zeta_4 \alpha_4^2 - (\lambda_{12} + \lambda_{23} + \lambda_{24} + \lambda_{34})]$$

with $\zeta_1 = \lambda_{12} + \lambda_{23}$, $\zeta_2 = \lambda_{12} + \lambda_{24}$, $\zeta_3 = \lambda_{13} + \lambda_{34}$ and $\zeta_4 = \lambda_{24} + \lambda_{34}$. A singular point of the Lagrangian satisfies $w_i = \zeta_i \alpha_i$, $1 \leq i \leq 4$.

C.3 All groups are active

We first consider the case $\lambda_{12}, \lambda_{13}, \lambda_{24}, \lambda_{34} > 0$. By complementary slackness

$$\|\alpha_g\| = 1, \quad g \in \mathcal{G} \quad (\text{CS})$$

which, using (30), rewrites as

$$\frac{w_1^2}{\zeta_1^2} + \frac{w_2^2}{\zeta_2^2} = 1, \quad \frac{w_1^2}{\zeta_1^2} + \frac{w_3^2}{\zeta_3^2} = 1, \quad \frac{w_2^2}{\zeta_2^2} + \frac{w_4^2}{\zeta_4^2} = 1 \quad \text{and} \quad \frac{w_3^2}{\zeta_3^2} + \frac{w_4^2}{\zeta_4^2} = 1. \quad (36)$$

Taking differences between pairs of equations above that share a common variable w_i we get

$$\begin{cases} |w_1|(\lambda_{24} + \lambda_{34}) = |w_4|(\lambda_{12} + \lambda_{13}) \\ |w_2|(\lambda_{13} + \lambda_{34}) = |w_3|(\lambda_{12} + \lambda_{24}) \end{cases}$$

Thus, isolating λ_{12} in both equations and eliminating it yields

$$\frac{|w_1|}{|w_4|}(\lambda_{24} + \lambda_{34}) - \lambda_{13} = \frac{|w_2|}{|w_3|}(\lambda_{13} + \lambda_{34}) - \lambda_{24}$$

Now isolating λ_{13} we get

$$\lambda_{13} = \left(1 + \frac{|w_2|}{|w_3|}\right)^{-1} \left(\frac{|w_1|}{|w_4|}(\lambda_{24} + \lambda_{34}) + \lambda_{24} - \frac{|w_2|}{|w_3|}\lambda_{34}\right)$$

Adding λ_{34} on both sides yields

$$\lambda_{13} + \lambda_{34} = \frac{\left(1 + \frac{|w_1|}{|w_4|}\right)\lambda_{24} + \left(1 + \frac{|w_2|}{|w_3|}\right)\lambda_{34}}{1 + \frac{|w_2|}{|w_3|}}$$

Inserting this expression into the only equation of (36) which doesn't contain λ_{12} we get

$$\frac{w_3^2 \left(1 + \frac{|w_2|}{|w_3|}\right)^2}{\left(1 + \frac{|w_1|}{|w_4|}\right)^2 (\lambda_{24} + \lambda_{34})^2} + \frac{w_4^2}{(\lambda_{24} + \lambda_{34})^2} = 1$$

which reduces to

$$\zeta_4 \triangleq \lambda_{24} + \lambda_{34} = \frac{|w_4|}{|w_1| + |w_4|} \left[(|w_2| + |w_3|)^2 + (|w_1| + |w_4|)^2 \right]^{\frac{1}{2}} \quad (37)$$

By symmetry, we get similar expressions for $\lambda_{12} + \lambda_{13}$, $\lambda_{12} + \lambda_{24}$, and $\lambda_{13} + \lambda_{34}$. Since $\Omega_{\mathcal{U}}^{\mathcal{G}}(w) = \lambda_{12} + \lambda_{13} + \lambda_{24} + \lambda_{34}$, we get immediately that

$$\Omega_{\mathcal{U}}^{\mathcal{G}}(w) = \left[(|w_2| + |w_3|)^2 + (|w_1| + |w_4|)^2 \right]^{\frac{1}{2}} = \|(|w_1| + |w_4|, |w_2| + |w_3|)\|$$

The above derivation gave us values for $\zeta_1, \zeta_2, \zeta_3, \zeta_4$. We discuss now the existence and the uniqueness of the $(\lambda_g)_g$. Given the vectors $\zeta \in \mathbb{R}^4$ and $\lambda \in \mathbb{R}^4$ we have $\zeta = B\lambda$ where B is the incidence matrix of the groups, with $B_{ig} = \mathbf{1}_{\{i \in g\}}$. To be precise we have

$$\begin{pmatrix} \zeta_1 \\ \zeta_2 \\ \zeta_3 \\ \zeta_4 \end{pmatrix} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \end{bmatrix} \begin{pmatrix} \lambda_{12} \\ \lambda_{13} \\ \lambda_{24} \\ \lambda_{34} \end{pmatrix}$$

Clearly, in this case, B is not invertible, and the kernel of B is the span of $(-1, 1, 1, -1)^T$. Since the matrix is symmetric, $\text{Ker}(B) = \text{Im}(B)^T$, and since $\zeta_1 + \zeta_4 = \Omega(\mathbf{w}) = \zeta_2 + \zeta_3$, we have $\zeta_1 - \zeta_2 + \zeta_3 - \zeta_4 = 0$. The vector λ exists provided the pre-image of ζ_i has a non-empty intersection with the positive orthant. Moreover, if all λ are positive then the solution is not unique. The Moore-Penrose pseudo-inverse of B is

$$B^+ = \frac{1}{8} \begin{bmatrix} 3 & 3 & -1 & -1 \\ 3 & -1 & 3 & -1 \\ -1 & 3 & -1 & 3 \\ -1 & -1 & 3 & 3 \end{bmatrix}.$$

Since $\zeta_1 + \zeta_4 = \zeta_2 + \zeta_3 = \omega \triangleq \Omega(\mathbf{w})$, the set of solutions is given by

$$\begin{pmatrix} \lambda_{12} \\ \lambda_{13} \\ \lambda_{24} \\ \lambda_{34} \end{pmatrix} = B^+ \cdot \begin{pmatrix} \zeta_1 \\ \zeta_2 \\ \omega - \zeta_2 \\ \omega - \zeta_1 \end{pmatrix} + \frac{\delta}{2} \begin{pmatrix} -1 \\ 1 \\ 1 \\ -1 \end{pmatrix} = \frac{1}{2} \begin{pmatrix} \zeta_1 + \zeta_2 - \delta \\ \zeta_1 - \zeta_2 + \delta \\ \zeta_2 - \zeta_1 + \delta \\ 2\omega - \zeta_1 - \zeta_2 - \delta \end{pmatrix}$$

for values of δ such that $\lambda_g \geq 0$. The latter constraint implies that we necessarily have

$$|\zeta_2 - \zeta_1| \leq \delta \leq \min(\zeta_1 + \zeta_2, 2\omega - \zeta_1 - \zeta_2)$$

W.l.o.g., we assume that $\zeta_1 \leq \zeta_2 \leq \omega - \zeta_2 \leq \omega - \zeta_1$. In that case the set of solutions in λ is parametrized by $\nu \in [0, 1]$ with

$$\lambda_{12} = \nu \zeta_1, \quad \lambda_{13} = (1 - \nu) \zeta_1, \quad \lambda_{24} = \zeta_2 - \nu \zeta_1, \quad \lambda_{34} = \omega - \zeta_2 - (1 - \nu) \zeta_1.$$

In particular, we see that setting $\nu = 0$ or $\nu = 1$ respectively removes $\{1, 2\}$ and $\{1, 3\}$ from the group-support of $\bar{\mathbf{v}}$.

The case considered here is an example of the situation where the decomposition is not unique, which is characterised by lemma 48 in the next section.

Appendix D. Uniqueness of the decomposition

In this section we give necessary and sufficient conditions for the support to be unique. As in lemma 44, we consider \mathbf{B} the incidence matrix of the groups defined by $\mathbf{B}_{ig} = \mathbf{1}_{\{i \in g\}}$. As before we denote $\check{\mathcal{G}}_1$ the strong group-support, $\check{J}_1 = \cup_{g \in \check{\mathcal{G}}_1} g$ and $J_0 = \text{supp}(\mathbf{w})$. Denote by $\mathbf{B}_{J_0 \check{\mathcal{G}}_1}$ the submatrix of \mathbf{B} whose rows are indexed by elements of the support of \mathbf{w} and whose columns are indexed by elements of $\check{\mathcal{G}}_1$.

Lemma 48 *The decomposition is unique if and only if $\mathbf{B}_{J_0 \check{\mathcal{G}}_1}$ has full row rank.*

Proof By lemma 7, the uniqueness of the decomposition is equivalent to the uniqueness of the solution λ to problem (10), which we can rewrite

$$\min_{\lambda \in \mathbb{R}_+^m} \frac{1}{2} \sum_{i \in J_0} \frac{w_i^2}{\sum_{g \ni i} \lambda_g} + \frac{1}{2} \sum_{g \in \check{\mathcal{G}}_1} \lambda_g. \quad (38)$$

Notice that only the terms indexed by $i \in J_0$ and $g \in \check{\mathcal{G}}_1$ contribute. Since the objective is a proper closed convex function with no direction of recession, this optimization problem admits at least one solution (the proof is the same as for 1). Since the gradient of the previous objective depends on λ_g only through $\zeta_i = \sum_{g \ni i} \lambda_g$, $i \in J_0$, then any other vector $\lambda_{\check{\mathcal{G}}_1}$ such that $\zeta_{J_0} = \mathbf{B}_{J_0 \check{\mathcal{G}}_1} \lambda_{\check{\mathcal{G}}_1}$ is also solution. It is therefore clear that it is sufficient that the kernel of $\mathbf{B}_{J_0 \check{\mathcal{G}}_1}$ is not trivial, i.e., $\mathbf{B}_{J_0 \check{\mathcal{G}}_1}$ is row rank deficient, to have multiple solutions. Indeed let $\mathbf{H} \in \mathbb{R}^{|\check{\mathcal{G}}_1| \times K}$ be a basis of the kernel of $\mathbf{B}_{J_0 \check{\mathcal{G}}_1}$ and consider that, by definition of $\check{\mathcal{G}}_1$, for all $g \in \check{\mathcal{G}}_1$, $\lambda_g > 0$. As a consequence, there must exist a neighborhood \mathcal{U} of 0 in \mathbb{R}^K such that for all $\mathbf{q} \in \mathcal{U}$, $\lambda_{\check{\mathcal{G}}_1} + \mathbf{H}\mathbf{q}$ has positive components. Since $\zeta_{J_0} = \mathbf{B}_{J_0 \check{\mathcal{G}}_1}(\lambda_{\check{\mathcal{G}}_1} + \mathbf{H}\mathbf{q})$, we have that $\lambda_{\check{\mathcal{G}}_1} + \mathbf{H}\mathbf{q}$ is another solution of the KKT conditions.

We now prove that $\mathbf{B}_{J_0 \check{\mathcal{G}}_1}$ being of full row rank is sufficient to ensure the uniqueness of the decomposition. Indeed, we show next that when $\mathbf{B}_{J_0 \check{\mathcal{G}}_1}$ is of full row rank, the hessian of the objective, restricted to the non-zero λ_g of (38) is positive definite, so that the objective is strictly convex and the optimum is therefore unique. The hessian is $\mathbf{Q} = (\mathbf{Q}_{gg'})_{g, g' \in \check{\mathcal{G}}_1}$ with

$$\mathbf{Q}_{gg'} = \sum_{i \in g \cap g'} \frac{w_i^2}{\left(\sum_{\tilde{g} \ni i} \lambda_{\tilde{g}}\right)^3} = \mathbf{B}_{J_0 \check{\mathcal{G}}_1}^\top \mathbf{D} \mathbf{B}_{J_0 \check{\mathcal{G}}_1} \quad \text{and} \quad \mathbf{D} = \text{diag} \left(w_i^2 \left(\sum_{\tilde{g} \ni i} \lambda_{\tilde{g}} \right)^{-3} \right)_{i \in J_0}.$$

Since \mathbf{D} is a diagonal matrix with non-zero coefficients, \mathbf{H} is p.s.d. iff $\mathbf{B}_{J_0 \check{\mathcal{G}}_1}$ is full row rank which concludes the proof. \blacksquare