SOCIETE FRANÇAISE DE STATISTIQUE

Journal de la Société Française de Statistique

Vol. 152 No. 2 (2011)

Regularized learning in bioinformatics

Titre: Apprentissage régularisé en bioinformatique

Laurent Jacob[1]

Abstract: Regularization is an important theme in statistics and machine learning and provides a principled way to address problems which would otherwise be ill-posed. It can be thought of as a restriction of the set of functions in which an empirical risk minimization is performed. If the original empirical risk minimization problem is ill-posed in the sense that it admits several solutions or that the solution is very sensitive to small changes in the data, constraining the optimization to a small set of functions is known to sometimes yield better estimates of the true (population) risk minimizer. In particular, when one expects a good estimate to have a certain type of regularity, using this measure of regularity to build the constraint can decrease the variance of the estimator without adding too much bias. In a context of growing availability of biological data from high-throughput technologies like microarrays or next generation sequencing, being able to apply statistical learning methods to predict which treatment is best suited to a patient or how his disease is likely to evolve is of utmost importance. Since in practical situations few samples are available compared to the dimension of the data (typically tenth of thousand of measures), designing adequate regularity measures from biological prior information is important to make these problems amenable to statistical learning. Several such measures have been proposed in the recent years to address particular problems. In this work, we review some of these methods. We also present in more detail one of them, designed to enforce the support of a linear function to be a union of predefined overlapping groups of covariates, and discuss its performances on a breast cancer dataset.

Résumé : La régularisation est un thème important en statistiques et apprentissage automatique. Elle fournit un cadre général rigoureux pour résoudre des problèmes qui seraient autrement mal posés. On peut la présenter comme la restriction de l'ensemble des fonctions dans lequel on applique une minimisation du risque empirique. Lorsque le problème de minimisation du risque empirique est mal posé, dans le sens où il n'admet pas de solution unique ou que celle-ci est très sensible à de petits changements dans les données, contraindre l'optimisation dans un petit ensemble de fonctions améliore parfois l'estimation du minimum du vrai risque (en population). En particulier, si l'on s'attend à ce qu'un bon estimateur possède un certain type de régularité, utiliser cette mesure de régularité pour construire la contrainte peut permettre de diminuer la variance de l'estimateur sans pour autant trop augmenter son biais. La disponibilité grandissante des données biologiques issues de technologies dites à haut débit, telles que les puces à ADN ou le séquençage à haut débit rendent possible l'utilisation de méthodes d'apprentissage statistique pour prédire le traitement le plus adapté à un patient ou l'évolution la plus vraisemblable de sa maladie. Ces applications fondamentales sont limitées par le fait que peu d'échantillons sont généralement disponibles comparé à la dimension des données (typiquement des dizaines de milliers de mesures). La conception de mesures de régularité adaptées à ces problèmes est donc nécessaire. De nombreuses mesures, adaptées à des problèmes variés ont été récemment proposées. Nous proposons une revue de ces méthodes, et présentons plus en détail l'une d'entre elles, conçue pour contraindre le support de l'estimateur à une union de groupes de variables potentiellement chevauchants définis *a priori*. Nous présentons et discutons également ses performances sur un problème de prédiction impliquant des données de cancer du sein.

Keywords: Computational biology, Bioinformatics, Supervised learning, Regularization

Mots-clés : Bioinformatique, Apprentissage supervisé, Régularisation

AMS 2000 subject classifications: 62, 68, 92

[1] University of California, Berkeley, Department of Statistics.
E-mail: laurent@stat.berkeley.edu

ISSN: 2102-6238

1. Introduction

The recent development of high-throughput technologies in biology has lead to an explosion of available data. Along with the systematic organisation of experimental knowledge in large public databases, this opens the possibility of using statistical learning methods to address crucial problems. Of particular interest is the problem of predicting some properties of a biological sample based on its molecular features.

DNA microarrays [8], or more recently RNA-Seq [36, 59], are used to quantify the expression of several genes in a particular tissue. Each of the approximately $30,000$ genes in the human genome essentially codes for a protein, which is the functional element of the cell : variations of quantity of a particular protein in a cell has direct effects on its behavior, which suggests that the expression of genes can be a good feature on which to build functions predicting properties of a biological sample. More precisely, the actual genes are coded in DNA molecules, which stay in the cell nucleus and are (under normal circumstances) in constant quantity (*e.g.* 2 copies of each gene in the human genome) and are the same in every cell of the same organism. These genes are transcribed in RNA molecules which then leave the cell nucleus and are translated into proteins. DNA microarrays and RNA-Seq technologies aim at quantifying the RNA present in a cell for each gene in a particular sample, with the hope that these quantities relate to the behavior of the cell [2, 15]. Properties whose accurate prediction is of practical importance include tumor characterization, diagnostic, prognostic and drug response prediction [4, 6, 37, 54]. Better treatments can be designed if one can determine which type of cancer affects a patient, his risk of metastasis in the next 15 years or how much benefit he can expect from various therapies. In particular, one may want to take more risk with toxic treatments like chemotherapies for a patient with a poor prognosis, but spare it to a patient for whom surgery is likely to be enough.

In the particular case of cancer, dysfunctioning is often known to arise from changes in the genome, *i.e.*, at the DNA level. For example, some genes regulating the cell duplication may be deleted, or some genes positively involved in this same process can be duplicated, here again leading to a disappearing or change in relative quantities of certain proteins and to a functional deregulation of the cell activity. Therefore, a complementary approach is to measure the number of copies of each gene in a cell, using CGH arrays [10, 11, 39, 52] or DNA-Seq [33], and use these tenth of thousands of measures for each individual as descriptors on which to build functions predicting prognostic or response to a treatment [1, 51].

Another example of growing quantity of data is high-throughput screening, where the activity of a large number of molecules against a particular biological target is assessed *in vitro*. Statistical learning methods may be used on these data to help in the discovery of new drugs [32] by predicting which molecules among a very large number of candidates bind to the target. Once a deregulated protein is identified as causing a disease, a possible next step in the drug design process is indeed to find a small molecule binding the protein and inhibiting its activity [18].

In all these examples however, the number of available examples is typically very small compared to the description space dimensionality. Studies rarely involve more than a few hundreds of patients, a small number compared to the tenth of thousands of gene expression or copy number values being measured for each patient. In addition, the measuring process, including the technology itself, is prone to a lot of biological and technical noise. Classical statistics are not designed to handle this type of data, essentially because finding a function separating two sets of points in high dimension is an ill-posed problem, with no unique solution, even less a robust one. Fortunately, ill-posed problems, in particular those involving less data points than dimensions, are not new in statistics and statistical machine learning [17]. The answer often involves to bias the solution towards a simpler class of functions to make the problem well-posed and the estimate more robust. This approach is often known as *penalized* or *regularized* learning, and bears a direct relation to the definition of prior distributions in the context of Bayesian learning [5, 12]. While historically based on constraints on the Hilbert norm of the function and leading to smoother functions, this method lends itself well to non-Hilbertian measures of complexity. In particular, there has been a growing interest in the recent years for complexity measures of parametric functions based on ℓ_1 norms, which lead to "sparse" solutions [9, 49], *i.e.* to solutions with many zero weights.

If one has a good idea or prior on some regularity properties the true function should have, such as a particular type of smoothness or a particular sparsity pattern, then restricting oneself to the class of functions having this particular regularity should guide the learning process by making it work in a smaller set of functions which still contains good estimates of the true function, *i.e.*, by decreasing the variance without increasing too much the bias of the resulting estimator. Different biological questions, for which different priors were available, have led to the development of a variety of regularized statistical learning methods. Section 2 introduces the formal motivation of regularized learning in the context of structural risk minimization. Section 3 reviews some of these methods, and Section 4 gives a more detailed description of one of them along with its applications. Section 5 provides a discussion of the presented methods.

2. Structural risk minimization

Let $\mathscr{X}$ be the space of the observations (*e.g* $\mathbb{R}^p$ when observing samples in terms of p gene expressions), and $\mathscr{Y}$ the space of the output associated with each observation (*e.g.*, $\{-1,1\}$ for binary outputs such as good or bad prognostic).

2.1. Population and empirical risks

Supervised learning aims at finding the function f which minimizes

$$R(f) = \int_{\mathscr{X}\times\mathscr{Y}} L(y, f(x))d\mathbb{P}, \tag{1}$$

where R is the *risk* of f, *i.e.*, the mean cost of using it to predict y from x on the joint distribution for a particular cost or loss function L. L quantifies the error made when predicting output $f(x)$ if the true output is y. A typical classification loss is 1 when $f(x) = y$ and 0 otherwise, but convex surrogates such as the hinge and the logistic losses are often used in practical estimation [17]. A typical regression loss is the squared loss $(y_i - f(x_i))^2$.

This quantity defined in (1) will be referred to as the *population risk* hereafter. It cannot be computed in a real situation because the joint distribution $\mathbb{P}$ is unknown. In practice, this distribution is therefore replaced by the empirical distribution of the training pairs (x_i, y_i) yielding the following *empirical risk* :

$$R_n(f) = \frac{1}{n}\sum_{i=1}^{n} L(y_i, f(x_i)). \tag{2}$$

Choosing the f which minimizes R_n is a procedure known as *empirical risk minimization*. As stated in the introduction for a finite sample size n the function which minimizes R_n may not be unique and some of the minimizers may have a very high population risk R. For example in a regression problem where $\mathscr{Y} = \mathbb{R}$, any function f taking the correct values at the training points, *i.e.*, $f(x_i) = y_i, i = 1, \ldots, n$ and any value everywhere else will have a zero empirical risk for any reasonable regression loss function (*e.g.* the squared loss $(y_i - f(x_i))^2$).

2.2. Bias-variance tradeoff

The theory of statistical learning [56, 57] gives bounds on the distance between $R_n(f)$ and $R(f)$ as a function of the sample size n and the complexity of the class of functions $\mathscr{H}$ on which the empirical risk is minimized. The key to obtain a function having a low population risk and therefore better generalization abilities on (x, y) pairs which were not in the training set is to control the complexity of $\mathscr{H}$. Indeed when looking for the best classifier f in a given space of functions $\mathscr{H}$, the Bayes regret $R(f) - R^*$, where R^* is the population risk of the Bayes rule which is the optimal classifier knowing the true distribution, can classically be decomposed as :

$$R(f) - R^* = \left(R(f) - \inf_{g \in \mathscr{H}} R(g) \right) + \left(\inf_{g \in \mathscr{H}} R(g) - R^* \right). \tag{3}$$

The second term is called the *approximation error* and corresponds to the minimum excess risk which can be achieved by using a member of $\mathscr{H}$. It is a bias term, which does not depend on the data but only on the richness of $\mathscr{H}$, or at least on its ability to approximate the Bayes rule. It will be small if $\mathscr{H}$ contains a function whose population risk (not based on the data, therefore not observable) is not too high compared to the optimal classifier. The first term in turn is called the *estimation error*, and corresponds to the excess risk of f with respect to the best possible function in $\mathscr{H}$. Again, the reason why it is not trivial to find functions f for which this term is zero is that it is not possible to observe R. We only have access to R_n, and as explained above, a function with a low R_n may have an arbitrarily large R in the general case.

Interestingly, when choosing f by empirical risk minimization over $\mathscr{H}$, it can be shown that if f_n^* minimizes R_n, then :

$$R_n(f^*) - \inf_{f \in \mathscr{H}} R(f) \leq 2 \sup_{f \in \mathscr{H}} |R_n(f) - R(f)|,$$

so the estimation error can be thought of like a variance term, which grows with the size, or more precisely the complexity of $\mathscr{H}$.

The variance term in the Bayes regret decomposition (3) is larger for complex function spaces, while the bias term penalizes spaces which do not contain good approximations of the true function. In order to obtain a low population risk when minimizing the empirical risk, it is therefore necessary to minimize it over function spaces which are not too complex, but are rich enough to contain good approximations of the true function.

2.3. Structural risk minimization

This bias-variance trade-off is generally dealt with using the following *structural risk minimization* procedure [55] :

1. Define a structured family of function complexity classes,
2. Find the empirical risk minimizer on each complexity class,
3. Choose the minimizer giving the best generalization performances.

A practical approach to step 1-2 is to solve the sequence of problems :

$$\begin{cases} \min\limits_{f \in \mathscr{H}} R_n(f) \\ \Omega(f) \leq \mu, \end{cases} \tag{4}$$

where Ω is a measure of complexity of the function f and $\mu \in \mathbb{R}^+$. This constrained formulation highlights the fact that μ indeed parameterizes a “structured family of function complexity classes” as for each $\mu > 0$, the feasible f to solve (4) will include the ones that were feasible for smaller μ along with some new functions which are more complex in the sense of Ω. Each μ defines a new function space in which the empirical risk minimization can be performed, with the hope that one of them is simple enough to have a small variance term yet rich enough to have a small bias term. This procedure underlines the importance of designing good Ω functions : since it defines the sequence of function spaces in which empirical risk minimization is performed, an excellent complexity measure Ω may give function spaces which are simple yet contain good approximations of the true function. In particular, this may happen if the constraint enforces some *a priori* about what type of regularity the true function should have. Step 3 in turn is done by computing the empirical risk of the selected function on a hold out part of the training set which was not used in the two previous steps. More generally, one often resorts in practice to *cross validation* procedures [17].

In the remainder of this article, we will abuse the notation and denote $L(f)$ the empirical risk $R_n(f) = \sum_{i=1}^n L(x_i, f(x_i))$, because this matches the most used convention in machine learning. Furthermore, we will restrict our discussion to the case of linear functions $f(x) = \beta^\top x$ for some vector of parameters $\beta \in \mathbb{R}^p$. The empirical risk of such a function will be denoted by $L(\beta)$.

3. Review of some existing complexity measures Ω

As outlined in Section 2.3, effectiveness of structural risk minimization relies on the definition of an adequate structured family of function complexity classes for the problem at hand, often implicitely defined by a complexity measure Ω. Gradually decreasing the constraint of this complexity indeed defines a nested sequence of function spaces. This section presents several algorithms developed in this framework, with various convex complexity measures Ω enforcing in each case an appropriate type of regularity for the problem at hand.

3.1. Relation to classical methods

First note that for convex loss L and complexity measure Ω, (4) is equivalent to the following penalized (but unconstrained) optimization problem :

$$\min_{\beta \in \mathbb{R}^p} L(\beta) + \lambda \Omega(\beta), \tag{5}$$

sometimes referred to as Tikhonov formulation and where λ is a function of μ. This formulation subsumes several classical methods from the statistics and machine learning literatures, in particular :

– when $L(\beta) = \sum_{i=1}^{n}(y_i - \beta^\top x_i)^2$ and $\Omega(\beta) = \|\beta\|^2$, (5) is the ridge regression.
– when $L(\beta) = \sum_{i=1}^{n}\max(0, 1 - y_i\beta^\top x_i)$ and $\Omega(\beta) = \|\beta\|^2$, (5) is the support vector machine.
– when $L(\beta) = \sum_{i=1}^{n}(y_i - \beta^\top x_i)^2$ and $\Omega(\beta) = \|\beta\|_1$, (5) is the lasso.

The ℓ_2 and ℓ_1 norms used in these classical formulations enforce particular types of regularity for β. A function of small ℓ_2 norm is *smooth* in the sense that two close points x, x' have close evaluations by β. Indeed, by the Cauchy-Schwarz inequality,

$$|\beta^\top x - \beta^\top x'| \leq \|\beta\|.\|x - x'\|,$$

so if x and x' are close, $\|x - x'\|$ is small and since $\|\beta\|$ is constrained to be small $|\beta^\top x - \beta^\top x'|$ has to be small as well.

Because of its non-differentiability, the ℓ_1 norm leads to optima of (5) with a lot of coordinates of β exactly at zero. This is further discussed in Section 3.4. In both cases, if there exists a function of low population risk R which is smooth (resp. sparse), enforcing the penalty will not bias much the estimation, but will guide it by forcing the optimization to occur in a smaller function space (to which the correct function, or at least a good estimate, belongs).

3.2. *Smoothness on a network using the Laplacian norm*

An interesting example of design of such a complexity measure is the case of gene expression data classification when gene networks are available. Gene networks are formed by the accumulation of a particular empirical biological knowledge : which proteins activate or inhibit the transcription of some genes (regulation networks), which proteins are involved in a particular sequence of chemical reactions (metabolic networks) or which proteins interact with each other in an organism (protein-protein interaction networks). The growing availability of these types of network makes them relevant as a side information to build a penalty. The intuition is that genes which are close in the network can be expected to have similar weights in the classification function. Indeed, since two close genes are generally involved in the same or in related biological processes, if one of them has a strong weight in the classifier, the other one is likely to have a strong weight as well.

In terms of regularization, this should translate in a complexity measure Ω which is small for smooth functions on the network (in the sense that weights associated to close genes are similar), and large for functions which are less regular on the network. A typical approach is to use the following penalty in (5) :

$$\Omega_{\text{spectral}} = \sum_{k \sim l} (\beta_k - \beta_l)^2 , \tag{6}$$

where $k \sim l$ denotes that genes k and l are connected in the network. This penalty measures how different the weights of connected nodes are. It is straightforward to check that this penalty can also be written $\beta^\top \mathscr{L} \beta$ where $\mathscr{L} = D - A$ is the unormalized *graph Laplacian matrix* expressed in terms of the adjacency and degree matrices A and D. $\mathscr{L}$ being positive semidefinite, $\Omega_{\text{spectral}}(\beta)$ can be thought of as $\left(\Lambda^{\frac{1}{2}} U^\top \beta\right)^\top \left(\Lambda^{\frac{1}{2}} U^\top \beta\right)$ where $\mathscr{L} = U \Lambda U^\top$ is the spectral decomposition of $\mathscr{L}$ expressed in terms of an orthogonal matrix whose columns are the eigenvectors of $\mathscr{L}$ and Λ is a diagonal matrix whose diagonal elements are the eigenvalues. In other words, this amounts to projecting β on the eigenvectors of $\mathscr{L}$, multiplying each projection by the square root of the corresponding eigenvalue and taking the ℓ_2 norm of the resulting vector. Since for each eigenvector u_i and its corresponding eigenvalue λ_i, by definition $\Omega_{\text{spectral}}(u_i) = \lambda_i$, the eigenvectors of $\mathscr{L}$ can be thought of as a basis of vectors of increasing $\Omega_{\text{spectral}}(.)$, and a graph equivalent of the Fourier basis. Alternatively by a change of variables, it is possible to show that using (6) in (5) amounts to applying an ℓ_2 norm after projecting the data points x_i on $\mathscr{L}^{-\frac{1}{2}} = U\Lambda^{-\frac{1}{2}}U^{\top}$, that is, after mapping the data to the graph-Fourier space, shrinking each graph-Fourier coefficient by one over the square root of the corresponding eigen value, and mapping back to the original space, which can be thought of as a filtering of the data to shrink the high frequency parts of its signal. This observation is of practical interest, because it shows that minimization of an empirical risk under constraint on (6) can be directly implemented by projecting the data on $\mathscr{L}^{-\frac{1}{2}}$ and using any classical ℓ_2-penalized algorithm on the transformed data.

[41] used a similar approach to separate two sets of yeast samples (irradiated with a low radiation dose vs non-irradiated). The only difference is the shrinkage function they used for the graph-Fourier coefficients : instead of one over the square root of the eigenvalue, they tried an exponential function and a Heaviside function which set the coefficients corresponding to eigenvalues above a certain cutoff to 0. The resulting penalty doesn't have a simple interpretation in terms of differences of weights for connected coefficients like in (6), but is expected to yield similar results. In these experiments, the use of this particular notion of regularity didn't improve the performances, which were the same as the ones obtained by a regular SVM. This suggests that this prior didn't lead to complexity classes that were better than the regular smoothness assumption made when using the ℓ_2 norm. This may be because the assumption of regularity on the network is wrong in the first place, because the best classification function is poorly approximated by a linear function, regular or not, or because the available network is flawed, contains wrong annotations or misses important ones. An interesting result however, is that by construction, this method yields classifiers which are coherent with the network and therefore more interpretable while keeping the same accuracy.

Related works include the addition of a ℓ_1 norm penalty to (6) to enforce sparsity in addition to the network-smoothness [30] and alternative measures of measures of network-smoothness to the one based on the regular Laplacian matrix [45].

3.3. Multi-task learning

Another way of restricting the complexity of the function space in which empirical risk minimization is performed is to use data from related learning problems, or tasks, and to force the learned function to be similar in some sense to the ones learned for the related tasks. This family of approaches is known as multi-task learning. A typical way of relating the tasks is to force their ℓ_2 distance to be low, *i.e.*, to penalize

$$\Omega_{\text{mixed}}(\beta) = \sum_{t=1}^{T} ||\beta_t - \bar{\beta}||^2 + \lambda_2 \sum_{t=1}^{T} ||\beta_t||^2, \tag{7}$$

where β_t is the function used to classify the data from learning task t and $\bar{\beta} = \frac{1}{T}\sum_{t=1}^{T}\beta_t$ is the average of the T functions.

This approach was used in the context of vaccine and drug design [20, 23]. In both cases, the problem is to discriminate between elements that bind to a particular target and those who don't. For vaccine design, one wants to identify which fragments of a pathogen can bind to MHC-I molecules, as this is a necessary condition for them to be immunogenic. The MHC is a very polyallelic group of genes, and for the molecule resulting from some alleles, few binders are known. Reducing the complexity by using data from related problems is therefore crucial in this case. Similarly for drug discovery, one is interested in finding small molecules that bind to a particular protein, some of which have very few known binders. In both applications, using related learning tasks to restrict the complexity of the function space lead to better prediction performances on average over all the targets, and this improvement was particularly strong for targets with few known binders. Performances were further improved by generalizing this idea to use descriptors for both the ligands and the targets.

Alternative approaches include simultaneously learning which classifiers should be constrained to be close to each others and which ones should not [19] or enforcing other types of relation among the learning tasks such as their belonging to the same low-rank subspace [46, 47] or their sharing of a common sparsity pattern [38].

3.4. Sparsity-inducing penalties

Regularization by the ℓ_1 constraint was introduced independently in the statistics [49] and the signal processing [9] literatures in two very close formulations. In statistics, it was first proposed in a regularized regression problem, the *least absolute shrinkage and selection operator* (Lasso) :

$$\begin{cases} \min_{\beta} \sum_{i=1}^{n} (y_i - x_i\beta)^2 \\ \|\beta\|_1 \leq C, \end{cases} \tag{8}$$

whereas in signal processing it was introduced as the *basis pursuit* :

$$\begin{cases} \min_{\beta} \|\beta\|_1 \\ X\beta = y, \end{cases} \tag{9}$$

as a mean to recover exactly the signal β from a given overcomplete dictionary X (the *basis pursuit denoising* formulation is equivalent to the Lasso).

The most intuitive and visual way of understanding why constraining the ℓ_1 norm leads to sparse estimates is to see the ℓ_1 ball as the tightest convex relaxation of the ℓ_0 one. The ℓ_0 penalty simply counts the number of non-zero elements in a vector. It is not a proper norm because it is not positive homogeneous but it is the direct mathematical translation of the sparsity requirement. Figure 1 shows that the ℓ_1 ball is the tightest possible convex relaxation of the ℓ_0 constraint which intuitively make it a good candidate to enforce the sparsity prior while keeping the computational and analytical advantages of having an estimator defined as the solution of a convex optimization problem.

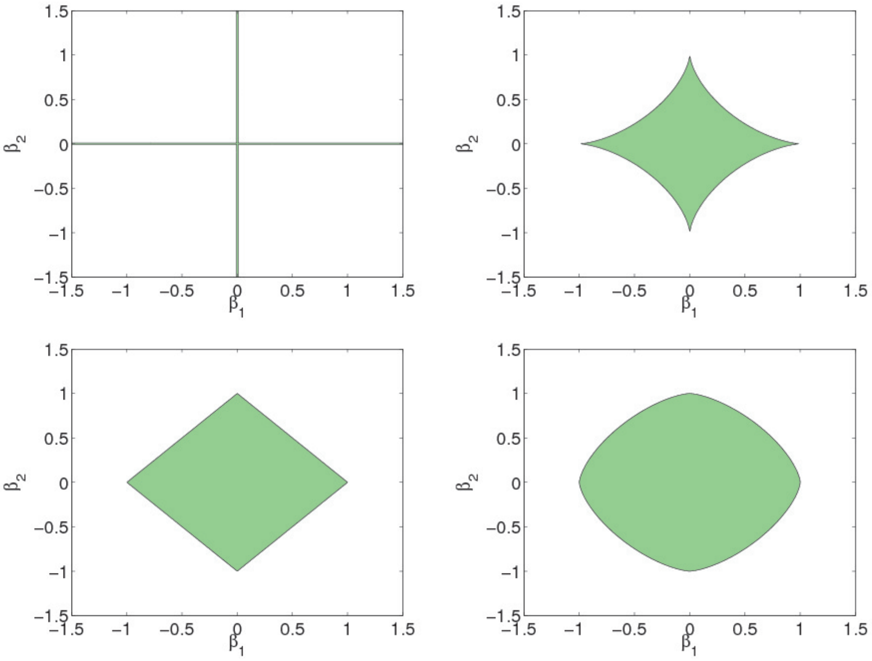

FIGURE 1. *Unit balls for the ℓ_p norms in two dimensions, for $p = 0, \frac{2}{3}, 1, \frac{3}{2}$.*

However, the fact that the ℓ_1 ball is a relaxation of the ℓ_0 does not suffice to explain that it still induces sparse estimates (for example, the ℓ_2 ball is also a relaxation of the ℓ_0 one and it does not yield sparse estimates). The two following arguments explain why penalizing by the ℓ_1 norm favors sparse solutions :

Geometric argument : Figure 2 illustrates in two dimensions the minimization of a smooth function, represented by the contour lines under ℓ_1 and ℓ_2 constraints represented by the green zones. It is visually clear that the shape of the ℓ_2 ball doesn't favor sparse solutions, as the ellipse doesn't hit it on one of the axes. In particular if the smooth function has a spherical shape and β_{min} is the minimizer of the smooth function (the center of the circles), the point inside the ℓ_2 ball of radius C which minimizes the function is the projection of β_{min} on the ball, $C.\frac{\beta_{min}}{\|\beta_{min}\|}$, which is colinear to β_{min} and has no reason to have any zero coordinate (unless β_{min} itself has zero coordinates, which has probability zero under any reasonable noise setting).

Under the ℓ_1 constraint on the other hand, the β minimizing the smooth function generally lies on one of the singularities of the ball as this is visually where the ellipse lines first hit the ℓ_1 ball, so the constrained solution typically has zero coordinates.

KKT argument : Consider the minimization of the following general problem :

$$\min_{\beta} L(\beta) + \lambda \|\beta\|_p, \tag{10}$$

for $1 \leq p \leq \infty$, where $\|\beta\|_p = \left(\sum_{j=1}^{d} |\beta_j|^p\right)^{\frac{1}{p}}$, denotes the ℓ_p norm of β. This problem is

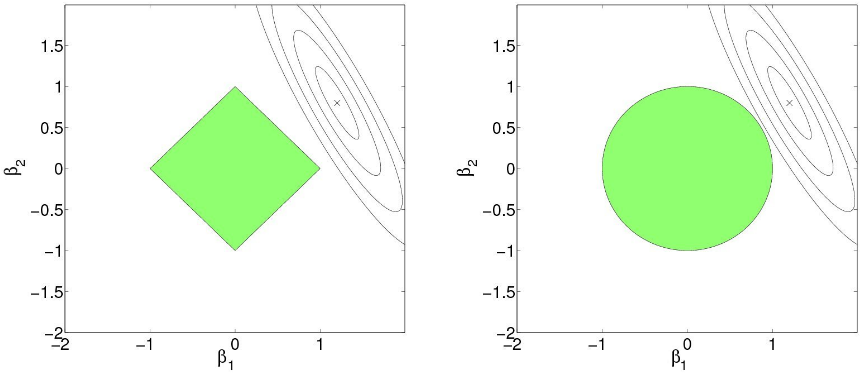

FIGURE 2. *Sparsity induction by the ℓ_1 norm.*

equivalent to $\min_\beta L(\beta)+\tilde{\lambda}\|\beta\|_p^p$ for some $\tilde{\lambda}$ and generalizes the ridge regression, the Lasso, and several other regularized problems.

For a convex loss function L, since $q \geq 1$, (10) is a convex optimization problem, whose solution is characterized by its KKT conditions [7]. In particular, at the optimum, β must satisfy the stationarity conditions. For $p>1$, $\|\beta\|_p$ is differentiable everywhere, and the stationarity conditions are :

$$\forall j, \frac{\partial L}{\partial \beta_j} = -\lambda \frac{\partial \|\beta\|_p}{\partial \beta_j} = -\lambda \text{sign}(\beta_j)\frac{|\beta_j|^{p-1}}{\|\beta\|_p^{p-1}}, \tag{11}$$

where $\text{sign}(\beta_j)$ is 1 for a positive β_j, -1 for a negative β_j and 0 for $\beta_j=0$. This condition simply means that at the optimum, the derivative of the loss function with respect to each parameter is cancelled out by the absolute value of the parameter (weighted by the strength λ of the constraint), which intuitively makes sense : the former tries to increase the parameter value in order to minimize the loss, whereas the latter penalizes large values of the parameter. As a consequence, for a β_j to be 0 at the optimum for any $\lambda>0$, the KKT conditions impose that the corresponding $\frac{\partial L}{\partial \beta_j}$ be 0 as well, which has probability 0 under any non-idealized setting.

For $p=1$ on the other hand, $|\beta_j|^p$ is non-differentiable at 0, so for $\beta_j=0$, the stationarity condition, in terms of the subdifferential ∂_{β_j} of the loss and the penalty functions becomes :

$$0 \in \partial_{\beta_j}\left(L(\beta)+\lambda\|\beta\|_p\right) \Leftrightarrow \left|\frac{\partial L}{\partial \beta_j}\right| \leq \lambda, \tag{12}$$

because the subdifferential of the absolute value is the $[-1,1]$ set. Therefore, any parameter with respect to which the gradient of the loss has an amplitude less than λ will be 0 at the optimum.

Note that in the previous argument, the zone where the parameter is left to 0 because it doesn't help enough the loss decrease is created by the non-differentiability of the penalty. This will be a useful fact to define more elaborate sparsity-inducing norms. Section 3.5 introduces an example where this strategy is used to constrain estimates to have some coordinates exactly equal. The entire Section 4 is devoted to the study of a penalty where this observation on non-differentiability regions is used to obtain estimates whose non-zero elements are unions of pre-defined groups of covariates. The Lasso penalty itself has been successfully used in several practical applications, including some biological ones [14, 43].

Finally as the ℓ_1 can serve two purposes (accurate estimation of y from x and estimation of the sparsity pattern of β), a natural question is whether the best parameter λ for risk minimization is also the best one for covariate selection. The answer is no [29], which means that selecting the regularization parameter by cross validation does not necessarily yield good performances in model selection.

3.5. *Piecewise-constant functions using the fused norm*

In the context of cancer data, it is also interesting to quantify the number of copies of each gene in the DNA of a particular tissue (typically a tumor). Indeed, amplification and deletion of particular genes are often at the center of tumorigenesis. These events do not occur specifically at the level of one gene, but rather of a DNA segment which can cover several genes, sometimes a whole chromosome. When using these counts in a classifier, a natural prior is therefore that two genes which are next to each other on the chromosome have exactly the same weight, since even if the quantitative values which are measured by the array are not exactly the same, they are likely to correspond to the same number of copies plus some noise. In other words, the linear classifier can be expected to be piecewise constant along the genome. This prior can be encoded using the following penalty :

$$\Omega_{\text{fused}}(\beta) = \sum_{i=1}^{p-1} |\beta_i - \beta_{i+1}|. \tag{13}$$

Just like the ℓ_1 norm leads to sparse estimates, (13) leads to piecewise constant classifiers, because it is simply an ℓ_1 norm applied to the differences of neighboring weights, and it therefore results in classifiers in which many of these differences are zero, *i.e.*, which are piecewise constant.

This penalty was first proposed in [28] and used in [16] for change-point detection. It was studied in combination with an ℓ_1 norm in [50] where it was applied to expression and mass spectrometry data. It was further refined in [42] where an adaptive formulation was proposed. [40] and [51] used this same combination of fused and ℓ_1 norm to classify array CGH (gene copy number) data. [40] tested the method on two cancer datasets and observed an improvement in the classification performances for one of them, no change for the other one.

4. Structured sparsity penalties for outcome prediction

We now focus on a particular penalty developed in [22] to deal with gene expression data. The motivation is that among all the genes whose expression is measured, only those involved in a small number of biological functions are worth using to predict the outcome (*e.g.*, metastasis). This is a slightly more general hypothesis than the usual sparsity, in which one adds a prior regarding which genes should be selected simultaneously in the model. We present this structured prior, some of its properties and its performances on real data.

4.1. Overlapping group lasso

When dealing with gene expression data, it is of course possible to directly use the ℓ_1 penalty to simultaneously learn a linear classifier and select a few genes to establish a predictive signature [14, 43]. In terms of feature selection however, it is known [62] that the Lasso fails to recover the correct sparsity pattern when the covariates are too correlated, which is likely to be the case with gene expression data. Indeed, the expression of each gene can be positively or negatively regulated by the expression of other genes, which systematically correlates their expressions. More generally, most biological mechanisms involve several interacting genes and therefore act as latent variables whose activation activates a whole group of genes, making their expressions strongly correlated. This suggests that instead of imposing sparsity at the gene level, one may try to constrain the number of biological functions involved in the classifier. This may also avoid the selection of genes spuriously associated with the output because of the noise, as such a spurious association is much less likely for a whole set of genes. While some algorithms like the elastic net [63] lean on the empirical covariance to jointly select correlated covariates, the method we present in this section relies on pre-defined groups of genes known to all be involved in a particular process, and available from public databases. None of the two paradigms really dominates the other : algorithms relying on the empirical covariance may suffer in the small sample regime when the estimation of the covariance becomes inaccurate, and algorithms relying on pre-defined structures on the covariates can perform arbitrarily bad in case of severe misannotation, *i.e.*, if the grouping they define is wrong or not relevant for the problem at hand.

A generalization of the Lasso penalty coined *group-lasso* was proposed by [60] to deal with group-sparsity, *i.e.*, the case where one expects some pre-defined groups of covariates to be either all selected in the model or all have zero weights :

$$\Omega_{\text{group}} = \sum_{g \in \mathscr{G}} \|\beta_g\|, \tag{14}$$

where $\mathscr{G}$ is the set of pre-defined groups forming a partition of the variables and $\beta_g \in \mathbb{R}^p$ are the vectors whose entries are the same as β for the covariates in g, and are 0 for other other covariates. (14) is the ℓ_1 norm of the ℓ_2 norms of the β_g, often referred to as a mixed norm. Just like the ℓ_1 norm leads to estimates with many weights at zero because it relaxes the ℓ_0 count of non-zero variables, (14) leads to estimates with many groups of variables at zero because it relaxes the count of groups of covariates whose corresponding restriction β_g has non-zero ℓ_2 norm. Generalizations of the geometric and KKT-based arguments developed to show that the ℓ_1 yielded sparse estimates can be straightforwardly derived for this ℓ_1/ℓ_2 norm. In particular, the ball corresponding to this norm (shown on Figure 3 for a special case in 3 dimensions) has singularities at points where some groups have exactly zero norm.

When dealing with gene sets corresponding to biological functions, the problem is slightly different, because these sets often overlap and therefore do not form a partition of the genes. When the groups in $\mathscr{G}$ overlap, (14) is still a norm (if all covariates are in at least one group), but the complexity that it measures, and the effect its penalization has on the estimate do not correspond to the prior that few biological functions should be involved in the model. Indeed, (14) by construction leads to estimates in which groups of weights are zero. By complementarity, this also means that the non-zero pattern is a union of groups when the groups form a partition, but in

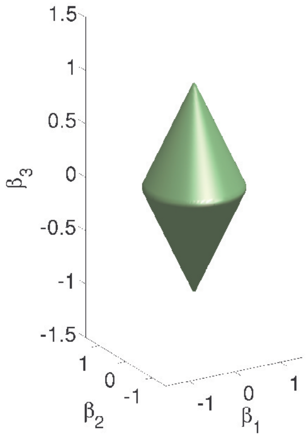

FIGURE 3. *Unit ball of the group-lasso in* 3 *dimensions with* $\mathscr{G} = \{\{1,2\},\{3\}\}$.

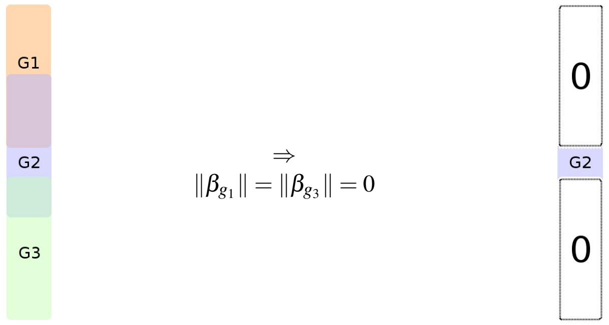

FIGURE 4. *Effect of penalty* (14) *on the support. Removing* any *group containing a variable removes the variable from the support.*

the general case it leads to non-zero patterns which are *complements* to unions of groups. This is illustrated on a simple example with three groups in Figure 4. When groups 1 and 3 are set to zero, the remaining weights are not the ones in group 2, but the ones that are neither in 1 nor in 3.

This type of effect can be relevant for some applications as studied *e.g.* in [24, 25]. In the case of gene expression data, this may not be the definition of regularity one wants to enforce. If a gene is an important marker of the property to be predicted, *all the gene sets* which include this gene will have to be non-zero for the gene to be involved in the classifier. In order to enforce the notion of regularity that we need, we should design a measure which is small when several gene sets are discarded but without necessarily enforcing that all genes in a discarded gene sets have zero weight. Rather, we want genes to have zero weight if *all* the gene sets to which they belong have been discarded.

In terms of penalized empirical risk minimization, this objective may be expressed under the following form :

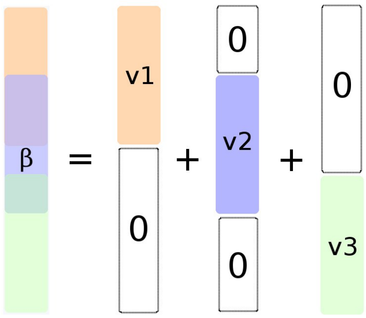

FIGURE 5. *Latent decomposition of β over $(v_g)_{v\in\mathscr{G}}$. Applying the ℓ_1/ℓ_2 penalty to the decomposition instead of applying it to the β_g removes only the variables which do not belong to* any *selected group.*

$$\begin{cases} \min_{\beta,v} L(\beta)+\lambda \sum_{g\in\mathscr{G}} \|v_g\| \\ \beta = \sum_{g\in\mathscr{G}} v_g \\ \operatorname{supp}(v_g) \subseteq g, \end{cases} \tag{15}$$

where each group is assigned a latent vector v_g on which the mixed norm is applied, while the classifier used in the loss function is formed by adding all the latent variables. The penalty term induces group-sparsity, but any covariate can remain in β as long as at least one v_g which includes the covariate has non-zero norm, as illustrated on Figure 5.

Sparsity patterns resulting from the minimization of (15) are by construction unions of groups. Indeed, if we denote by $\mathscr{G}_1 \subset \mathscr{G}$ the set of groups g with $v_g \neq 0$, then we immediately get $\beta = \sum_{g\in\mathscr{G}_1} v_g$, and therefore:

$$\operatorname{supp}(\beta) \subset \bigcup_{g\in\mathscr{G}_1} g\,.$$

For analysis purpose, it is good to notice that problem (15) can be reformulated under the more classical form (5). Indeed, the only part of problem (15) which depends on v is the penalty term and the constraints. It is therefore possible to move the minimization over v and the related constraints to the last term, and see the problem as a minimization over β of a new functional which includes another minimization problem (over v) :

$$\begin{cases} \min_{\beta,v} L(\beta)+\lambda \sum_{g\in\mathscr{G}} \|v_g\| \\ \beta = \sum_{g\in\mathscr{G}} v_g \\ \operatorname{supp}(v_g) \subseteq g, \end{cases} \quad = \quad \min_{\beta} L(\beta)+\lambda\Omega^{\mathscr{G}}_{\text{overlap}}(\beta)\,, \tag{16}$$

with

$$\Omega^{\mathscr{G}}_{\text{overlap}}(\beta) = \min_{\mathbf{v}\in\mathscr{V}_{\mathscr{G}},\sum_{g\in\mathscr{G}} v_g=\beta} \sum_{g\in\mathscr{G}} \|v_g\|\,. \tag{17}$$

where $\mathscr{V}_{\mathscr{G}} \subset \mathbb{R}^{p\times\mathscr{G}}$ is the set of $|\mathscr{G}|$-tuples of vectors $\mathbf{v} = (v_g)_{g\in\mathscr{G}}$, where each v_g is a vector in $\mathbb{R}^p$, that satisfy $\operatorname{supp}(v_g) \subset g$ for each $g\in\mathscr{G}$.

To summarize, it is possible to enforce the prior we have on the linear classifier by introducing new variables in the optimization problem (such that a gene has zero weight in the classifier if all the gene sets to which it belongs have been set to zero by the penalty). This new problem can be re-written as a classical minimization of the empirical risk, penalized by a particular penalty $\Omega^{\mathcal{G}}_{\text{overlap}}(.)$. This penalty itself associates to each function β 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 particular, the next sections show that $\Omega^{\mathcal{G}}_{\text{overlap}}(.)$ is a norm, derive some of its properties and show under which conditions its penalization leads to model-consistent estimation.

4.2. Properties of the resulting penalty

We now focus on the properties of $\Omega^{\mathcal{G}}_{\text{overlap}}(.)$ as defined in (17). The section follows the lines of [22]. We will prove that $\Omega^{\mathcal{G}}_{\text{overlap}}(.)$ is a norm. We will then derive a dual version of the norm, along with some of its properties, in particular the fact the optimal weights associated to the group-support are uniquely defined. Finally, one can notice that the set of solutions of the dual formulation corresponds to the subgradient of the norm. These last properties will be of interest in the next section to prove some consistency results. Proofs are given in appendix.

Let us first introduce some notation. $\mathbf{V}(\beta) \subset \mathcal{V}_{\mathcal{G}}$ denotes the set of $|\mathcal{G}|$-tuples of vectors $\mathbf{v} = (v_g)_{g\in\mathcal{G}}$ which reach the minimum in (17), *i.e.*, which satisfy

$$\beta = \sum_{g\in\mathcal{G}} v_g \quad \text{and} \quad \Omega^{\mathcal{G}}_{\text{overlap}}(\beta) = \sum_{g\in\mathcal{G}} \|v_g\|\,.$$

The objective of (17) is a convex and coercive function, so that the set of solutions $\mathbf{V}(\beta)$ is non-empty and convex. Moreover,

Lemma 1. *$\beta \mapsto \Omega^{\mathcal{G}}_{overlap}(\beta)$ is a norm.*

Again, in the general case this norm is expressed in a variational form (as the solution of an optimization problem). To the best of our knowledge, no closed form is available in the general case (although a closed form can be derived in some simple particular cases)

We now give a dual formulation of the norm $\Omega^{\mathcal{G}}_{\text{overlap}}(.)$ yielding some important properties.

Lemma 2. 1. *It holds that:*

$$\Omega^{\mathcal{G}}_{overlap}(\beta) = \sup_{\alpha\in\mathbb{R}^p:\forall g\in\mathcal{G}, \|\alpha_g\|\le 1} \alpha^\top \beta\,. \tag{18}$$

2. *A vector $\alpha \in \mathbb{R}^p$ is a solution of* (18) *if and only if there exists $\mathbf{v} = (v_g)_{g\in\mathcal{G}} \in \mathbf{V}(\beta)$ such that:*

$$\forall g \in \mathcal{G}\,,\ \text{if } v_g \neq 0,\ \alpha_g = \frac{v_g}{\|v_g\|} \text{ else } \|\alpha_g\| \le 1 \tag{19}$$

3. *Conversely, a $\mathcal{G}$-tuple of vectors $\mathbf{v} = (v_g)_{g\in\mathcal{G}} \in \mathcal{V}_{\mathcal{G}}$ such that $\beta = \sum_g v_g$ is a solution to* (17) *if and only if there exists a vector $\alpha \in \mathbb{R}^p$ such that* (19) *holds.*

Recall that neither the primal nor the dual formulation admit a unique minimizer or maximizer, *i.e.*, several feasible $\mathbf{v}$ (resp. α) can reach the minimum of (17) (resp. the maximum of (18)). The next result however shows that the part of the optimal α associated to the group-support of β is uniquely defined. Denote by $\mathscr{G}_1$ the group-support of β, *i.e.*, the set of groups belonging to the support of at least one optimal decomposition of β: $\mathscr{G}_1 = \{g \in \mathscr{G} \,|\, \exists \mathbf{v} = (v_g)_g \in \mathbf{V}(\beta),\, v_g \neq 0\}$ and J_1 the corresponding set of variables $J_1 = \cup_{g \in \mathscr{G}_1} g$.

Lemma 3. *Let α be an optimum in the formulation* (18) *of the $\Omega^{\mathscr{G}}_{overlap}(\cdot)$ norm, then α_{J_1} is uniquely defined.*

The following corrolary transposes this result to the primal formulation :

Corollary 1. *For any $\mathbf{v}, \mathbf{v}' \in \mathbf{V}(\beta)$ and for any $g \in \mathscr{G}$,*

$$\|v_g\| \times \|v'_g\| = 0 \quad or \quad \exists \gamma_g \geq 0 \ s.t. \ v'_g = \gamma v_g \,. \tag{20}$$

So far, we have only considered properties of $\Omega^{\mathscr{G}}_{\text{overlap}}(.)$ itself, not in the context of any statistical learning task. The next result expresses the optimality conditions of a regularized learning problem involving $\Omega^{\mathscr{G}}_{\text{overlap}}(.)$ in terms of the optimum of (18). Consider a learning scenario where we use $\Omega^{\mathscr{G}}_{\text{overlap}}(\beta)$ as the regularization term in (5), with a convex and differentiable objective function L :

$$\min_{\beta \in \mathbb{R}^p} L(\beta) + \lambda \Omega^{\mathscr{G}}_{\text{overlap}}(\beta) \,, \tag{21}$$

where $\lambda > 0$ is a regularization parameter. We first derive optimality conditions for any solution of (21). For that purpose, let us denote $\mathscr{A}_{\mathscr{G}}(\beta)$ the set of vectors $\alpha \in \mathbb{R}^p$ solution of (18).

Lemma 4. *A vector $\beta \in \mathbb{R}^p$ is a solution of* (21) *if and only if $-\nabla L(\beta)/\lambda \in \mathscr{A}_{\mathscr{G}}(\beta)$.*

Proof. The proof follows from the same Lagrangian based derivation as for Lemma 2, adding only the loss term. □

Remark 1. *By point* 2 *of Lemma 2, an equivalent formulation is the following: a vector $\beta \in \mathbb{R}^p$ is a solution of* (21) *if and only if it can be decomposed as $\beta = \sum_{g \in \mathscr{G}} v_g$ where, for any $g \in \mathscr{G}$, $v_g \in \mathbb{R}^p$, supp$(v_g) = g$, and if $v_g = 0$ then $\|\nabla_g L(\beta)\| \leq \lambda$, and $\nabla_g L(\beta) = -\lambda v_g / \|v_g\|$ otherwise.*

4.3. Model consistency of $\Omega^{\mathscr{G}}_{overlap}(.)$ minimization

An important problem when studying sparsity-inducing penalties is their *model consistency* : under which settings do they recover the correct sparsity pattern when the number of data points grows? [61] and [62] showed that in the case of a finite number of parameters and for an adequate choice of the regularization parameter, the Lasso is consistent under some irrepresentable conditions, stating that when the variables of the model are not too correlated with the variables which are not in the model. [3] gave model consistency conditions for the group-lasso, yielding similar irrepresentable conditions as for the Lasso. In this section, we give conditions for the consistency of $\Omega^{\mathscr{G}}_{\text{overlap}}(.)$ minimization.

Consider the linear regression model $Y = X\bar{\beta} + \varepsilon$, where $X \in \mathbb{R}^{n \times p}$ is a design matrix, $Y \in \mathbb{R}^p$ is the response vector and $\varepsilon \in \mathbb{R}^p$ is a vector of i.i.d. random variables with mean 0 and finite variance. We denote the true regression function by $\bar{\beta}$. We assume that

1. (H1) $\Sigma := \frac{1}{n}X^\top X$ is positive definite.
2. (H2) There exists a neighborhood of $\bar{\beta}$ in which (17) has a unique solution.

If $\mathscr{G}_1$ is the set of group supporting the unique solution of (17), we denote $\mathscr{G}_2 \overset{\Delta}{=} \mathscr{G}\backslash\mathscr{G}_1$ and $J_2 \overset{\Delta}{=} [1,p]\backslash J_1$. For convenience, for any group of covariates g we note X_g the $n \times |g|$ design matrix restricted to the predictors in g, and for any two groups g, g' we note $\Sigma_{gg'} = X_g^\top X_{g'}$. We can then provide a condition under which minimizing the least-square error penalized by $\Omega^{\mathscr{G}}_{\text{overlap}}(\beta)$ leads to an estimator with the correct support. Consider the two conditions:

$$\forall g \in \mathscr{G}_2,\ \|\Sigma_{gJ_1}\Sigma^{-1}_{J_1J_1}\alpha_{J_1}(\bar{\beta})\| \leq 1 \tag{C1}$$

$$\forall g \in \mathscr{G}_2,\ \|\Sigma_{gJ_1}\Sigma^{-1}_{J_1J_1}\alpha_{J_1}(\bar{\beta})\| < 1 \tag{C2}$$

Lemma 5. *With assumptions (H1-2), for $\lambda_n \to 0$ and $\lambda_n n^{1/2} \to \infty$, conditions* (C1) *and* (C2) *are respectively necessary and sufficient for the solution of* (21) *to estimate consistently the group-support of $\bar{\beta}$.*

These conditions are very similar to the ones obtained in [3] for the usual group lasso. The $\frac{\beta_g}{\|\beta_g\|}$ factor of the usual group lasso is replaced here by $\alpha_{J_1}(\bar{\beta})$. These expressions are the subgradients of the corresponding norms, which makes this result a direct generalization of the group-lasso one. In particular, in the case where the groups in $\mathscr{G}$ form a partition, (17) boils down to the ℓ_1/ℓ_2 norm, and $\alpha_g(\bar{\beta}) = \frac{\beta_g}{\|\beta_g\|}$. Another difference arises from the Σ_{gJ_1} factor. The equivalent factor in the case of non-overlapping groups only depends on the correlation among the unique covariates. In the context of an orthogonal design for example, it only has zero elements by definition. In the case of overlapping groups, the term also depends on the level of intersection between g and J_1. If a group is almost included in J_1, the term can therefore become very close to 1 even in the case of an orthogonal design. Note that this is not really a weakness of the algorithm, but more an illustration of the fact that identifying the correct group support in the case of overlapping group can be a much more difficult task than in the case of a partition.

4.4. *Algorithm*

We now present a practical way to optimize (16). We consider loss functions L which only depend on β through dot products with the data points X_i, *i.e.* $L(\beta) = \tilde{L}(X\beta)$, which is the case of many loss functions of interest.

In this case, a simple way to optimize (16) is to explicitly duplicate the variables in the design matrix. Indeed, eliminating β in (16) by plugging the first constraint $\beta = \sum_g v_g$ into $\tilde{L}(X\beta)$, we can write :

$$\begin{cases} \min\limits_{\beta,v} \tilde{L}(X\beta) + \lambda \sum\limits_g \|v_g\| \\ \beta = \sum_g v_g \\ \operatorname{supp}(v_g) \subseteq g. \end{cases} = \min_{\tilde{v}} \tilde{L}(\tilde{X}\tilde{v}) + \lambda \sum_g \|\tilde{v}_g\|, \tag{22}$$

where $\tilde{X} \in \mathbb{R}^{n\times\Sigma|g|}$ is defined by the concatenation of copies of the design matrix restricted each to a certain group g, i.e., $\tilde{X} = [X_{g_1}, X_{g_2}, ..., X_{g_{|\mathscr{G}|}}]$, with $\mathscr{G} = \{g_1, \dots, g_{|\mathscr{G}|}\}$, and where we denote $\tilde{v}_g = (v_{gi})_{i\in g}$ and $\tilde{\mathbf{v}} = (\tilde{v}_{g_1}^\top, \dots, \tilde{v}_{g_{|\mathscr{G}|}}^\top)^\top$.

On our simple example with 3 overlapping groups used in Figures 4 and 5, this gives :

$$X\beta = X.\begin{pmatrix}\tilde{v}_1\\ 0\end{pmatrix} + X.\begin{pmatrix}0\\ \tilde{v}_2\\ 0\end{pmatrix} + X.\begin{pmatrix}0\\ \tilde{v}_3\end{pmatrix} = (X_{g_1}, X_{g_2}, X_{g_3}).\begin{pmatrix}\tilde{v}_1\\ \tilde{v}_2\\ \tilde{v}_3\end{pmatrix} \triangleq \tilde{X}\tilde{v}.$$

That way the vector $\tilde{\mathbf{v}} \in \mathbb{R}^{\Sigma|g|}$ can be directly estimated from $\tilde{X}$ with a classical group lasso for non-overlapping groups : the right hand side of (22) is the same as (5) with penalty (14), *i.e.*, a classical group-lasso, performed on new variables.

For the experiments presented in the next section, we implemented the approach of [34] to estimate the group lasso in the expanded space, combined with the active set strategy of [44]. This active set strategy allows experiments to run quite fast even when a large number of groups/covariates are present, as long as only the beginning of the regularization path (for large values of λ) is needed. This is the case in our experiments, as the perfomances degrade quickly when including too many covariates, and the 5 5-fold double cross-validations of Section 4.5 run in a couple of hours on a laptop.

The code is freely available at
`http://cbio.ensmp.fr/~ljacob/documents/overlasso-package.tgz`.

4.5. Result

To illustrate the effect of using the gene set information through $\Omega^{\mathscr{G}}_{\text{overlap}}(.)$ penalization, we compare the performances of a classifier learned using $\Omega^{\mathscr{G}}_{\text{overlap}}(.)$ to the performances of a classifier learned using ℓ_1 penalization. The latter is a natural alternative to $\Omega^{\mathscr{G}}_{\text{overlap}}(.)$ which was used as a comparison in [22] and corresponds to the simpler prior that few genes are involved in the true prediction function (as opposed to few gene sets).

Like in [22], we use the gene expression dataset compiled by [53] and the canonical gene sets or *pathways* from MSigDB [48] containing 639 groups of genes, 637 of which involve genes from our study. The gene expression dataset 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 3510 genes which are in at least one pathway.

[22] obtained interesting results from a gene set selection point of view, namely they managed to build classifiers as accurate as the one learned with the ℓ_1 penalty, but involving much fewer pathways. This was a positive result from an interpretability perspective, but somehow disappointing because it did not lead to a substantial increase in performances, suggesting that the

TABLE 1. *Balanced classification error for the ℓ_1 and $\Omega^{\mathcal{G}}_{overlap}(.)$ on average over 5 folds, for 5 different folding choices.*

METHOD	$\Omega^{\mathcal{G}}_{\text{OVERLAP}}(.)$	ℓ_1
ERROR FOLDING 1	0.29 ± 0.05	0.36 ± 0.04
ERROR FOLDING 2	0.30 ± 0.08	0.42 ± 0.04
ERROR FOLDING 3	0.34 ± 0.14	0.37 ± 0.10
ERROR FOLDING 4	0.31 ± 0.11	0.37 ± 0.08
ERROR FOLDING 5	0.35 ± 0.05	0.37 ± 0.05

enforced prior was not a good one. In the experiment we present here, we didn't take into account the gene sets of more than 50 genes in the penalty. Intuitively, including very large gene sets in the regularization defeats the purpose of parsimony-based regularization : adding a large group to the model is cheap (in terms of ℓ_0 cost, it would be the same cost as adding a small group) and gives more degrees of freedom to fit the model and decrease the empirical risk. In practice, one can observe that very large groups of genes always enter the model at the beginning of the regularization path.

We learn a linear classifier by solving problem (5) for a balanced logistic loss function and either an ℓ_1 or $\Omega^{\mathcal{G}}_{\text{overlap}}(.)$ penalty. The other difference with the experiments of [22] is that we estimate error on 5 fold cross validation and since we observed a high variance of the performance across the choice of the 5 folds, we present the results for 5 folding choices. As in [22], λ is selected by internal cross validation on each training set, *i.e.*, for each train/test pair, we do another cross-validation *within* the training set along a grid of λ parameters, pick the parameter yielding the best internal cross-validation performances, train a new model using this parameter and the whole training set and evaluate its error on the test set. The chosen criterion is the balanced error, *i.e.*, the mean of sensitivity and specificity.

The results are presented in Table 1, and show a uniform improvement of the classification performances of the $\Omega^{\mathcal{G}}_{\text{overlap}}(.)$ regularized method against the ℓ_1 one. This suggests that the prior of gene set parsimony enforced by $\Omega^{\mathcal{G}}_{\text{overlap}}(.)$ is relevant and can lead to performances improvements under the condition that the gene set family is cleaned by removing very large pathways.

5. Discussion

We have presented an introduction to regularized empirical risk minimization : why it is expected to improve performances in practical cases of learning in high dimension, and several of its applications in computational biology along with existing penalties designed for these cases.

We have also presented more in details one particular penalty designed to restrict the function space to those whose support is formed by a union of pre-defined overlapping groups of covariates, along with its application to outcome prediction from gene expression data. Interestingly, this particular problem lead to the design of a new non-Hilbertian norm rather than the use or combination of existing ones.

The main motivation behind the design of these penalties is generally to reduce the complexity of the function space in which empirical risk minimization is performed, thereby reducing the variance without increasing too much the bias if the true function is expected to be well approximated by a function lying in the simpler function space, *i.e.*, having low penalty. To obtain this effect, it is necessary that the prior which is used to design the penalty is accurate enough. A first condition for this is of course that the intuition behind the prior is correct *e.g.*, there exists a good approximating function which is smooth on the gene network, or piecewise constant along the genome. But it is also necessary that the side information used to build the prior, like the position of the genes along the genome, the wiring of the gene network or the grouping of genes in pathways is accurate enough, and relevant in terms of the prior. This may explain why the prior enforced in (6) didn't lead to improvements in classification accuracy : the annotation of gene networks is known to be incomplete and inaccurate, and a good linear classifier may be smooth on some gene networks (say, an hypothetical complete and accurate regulation network) but not on metabolic networks. Similarly, the results presented in Section 4.5 suggest that using the raw, unfiltered set of gene pathways from MSigDB doesn't lead to a good penalty (17), while the same penalty yields substantial improvements when using only gene sets of reasonable size. Finally in some cases, working with linear functions may be a bad choice : if no linear function is a good approximation of the function of lowest population risk, *i.e.* if the bias incurred by restricting oneself to linear function is large enough, playing with the variance term by designing penalties may not change much the performances.

A second motivation can be the effect on the properties on the estimated function. In some cases, this effect can actually be the main goal, for example when the objective is to detect breakpoints in several copy number profiles using a group fused lasso penalty with a regression loss [58] : in this case, the goal is not really to minimize the population risk, *i.e.* to find good approximation of the copy number profiles in terms of ℓ_2 error, but to detects regions of the genome which are amplified or deleted. The used of the fused norm is showed to lead to higher accuracy in breakpoint detection on synthetic data. More generally, one may be interested in both the learned function and its characteristics induced by the penalty. Typically, enforcing some regularity constraints makes the function more interpretable, *e.g.* in the context of a known network or in terms of genes or biological functions involved (not necessarily as a cause though). The Laplacian norm used by [41] for example allowed to obtain a linear classifier whose weights were coherent with known biology of the studied phenomenon, such as upregulation of the oxidative phosphorylation pathway as well as DNA and RNA replication and repair. Clinicians are generally more comfortable with interpretable gene signatures than with black box classification functions, and functions involving few covariates (gene expressions or protein levels) are easier to use in practical conditions, *e.g.* using dedicated chips measuring the expression level of a few genes [35]. Moreover, obtaining sparse classifiers can also be a way to detect important genes or gene sets as potential drug targets. This is classically done by statistical hypothesis testing, namely two-sample test at the gene level, and can be extended at the pathway level [21, 31]. Sparse learning approaches are expected to yield less redundant results in the sense that only one among several correlated genes or gene sets should be selected, which can be a good or a bad thing depending if the dropped entities actually correspond to non-causal elements, or if on the contrary important causal elements are dropped because correlated non-causal ones are selected.

A last remark is that controlling the bias-variance tradeoff on very noisy data with few samples in high dimension may require more extreme biases than just working in particular balls of the space of linear functions, and some alternative approaches consider much simpler classification functions like ratios of two gene expressions, yielding good and very stable performances [13].

Acknowledgements

The author thanks the reviewer for his helpful comments and the Genentech postdoctoral innovation fellowship and the Stand Up To Cancer Program for their generous fundings.

References

[1] C. Aliferis, D. Hardin, and P. Massion. Machine Learning Models For Lung Cancer Classification Using Array Comparative Genomic Hybridization. In *Proceedings of the 2002 American Medical Informatics Association (AMIA) Annual Symposium*, pages 7–11, 2002.

[2] A. A. Alizadeh, M. B. Eisen, R. E. Davis, C. Ma, I. S. Lossos, A. Rosenwald, J. C. Boldrick, H. Sabet, T. Tran, X. Yu, J. I. Powell, L. Yang, G. E. Marti, T. Moore, J. Hudson, L. Lu, D. B. Lewis, R. Tibshirani, G. Sherlock, W. C. Chan, T. C. Greiner, D. D. Weisenburger, J. O. Armitage, R. Warnke, R. Levy, W. Wilson, M. R. Grever, J. C. Byrd, D. Botstein, P. O. Brown, and L. M. Staudt. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, 403(6769):503–511, Feb 2000.

[3] F. Bach. Consistency of the group lasso and multiple kernel learning. *J. Mach. Learn. Res.*, 9:1179–1225, 2008.

[4] A. Ben-Dor, L. Bruhn, N. Friedman, I. Nachman, M. Schummer, and Z. Yakhini. Tissue classification with gene expression profiles. *J. Comput. Biol.*, 7(3-4):559–583, 2000.

[5] J. Berger. *Statistical Decision Theory and Bayesian Analysis*. Springer-Verlag, 1985.

[6] A. Bhattacharjee, W. G. Richards, J. Staunton, C. Li, S. Monti, P. Vasa, C. Ladd, J. Beheshti, R. Bueno, M. Gillette, M. Loda, G. Weber, E. J. Mark, E. S. Lander, W. Wong, B. E. Johnson, T. R. Golub, D. A. Sugarbaker, and M. Meyerson. Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. *Proc. Natl. Acad. Sci. USA*, 98(24):13790–13795, Nov 2001.

[7] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, New York, NY, USA, 2004.

[8] P. Brown and D. Botstein. Exploring the new world of the genome with DNA microarrays. *Nat. Genet.*, 21:33–37, 2000.

[9] S. S. Chen, D. L. Donoho, and M. Saunders. Atomic decomposition by basis pursuit. *SIAM J. Sci. Comput.*, 20(1):33–61, 1998.

[10] S. F. Chin, A. E. Teschendorff, J. C. Marioni, Y. Wang, N. L. Barbosa-Morais, N. P. Thorne, J. L. Costa, S. E. Pinder, M. A. van de Wiel, A. R. Green, I. O. Ellis, P. L. Porter, S. Tavaré, J. D. Brenton, B. Ylstra, and C. Caldas. High-resolution aCGH and expression profiling identifies a novel genomic subtype of ER negative breast cancer. *Genome Biol.*, 8(10):R215, 2007.

[11] S.-F. Chin, Y. Wang, N. P. Thorne, A. E. Teschendorff, S. E. Pinder, M. Vias, A. Naderi, I. Roberts, N. L. Barbosa-Morais, M. J. Garcia, N. G. Iyer, T. Kranjac, J. F. R. Robertson, S. Aparicio, S. Tavare, I. Ellis, J. D. Brenton, and C. Caldas. Using array-comparative genomic hybridization to define molecular portraits of primary breast cancers. *Oncogene*, 26(13):1959–1970, Sept. 2006.

[12] M. H. De Groot. *Optimal statistical decisions / Morris H. De Groot*. McGraw-Hill, New York :,, 1970.

[13] D. Geman, C. d'Avignon, D. Q. Naiman, and R. L. Winslow. Classifying gene expression profiles from pairwise mrna comparisons. *Stat Appl Genet Mol Biol*, 3:Article19, 2004.

[14] D. Ghosh and A. M. Chinnaiyan. Classification and Selection of Biomarkers in Genomic Data Using LASSO. *J Biomed Biotechnol*, 2005(2):147–54, 2005.

[15] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander. Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. *Science*, 286:531–537, 1999.

[16] Z. Harchaoui and C. Levy-Leduc. Catching change-points with lasso. In J. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems 20*, pages 617–624. MIT Press, Cambridge, MA, 2008.

[17] T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning: data mining, inference, and prediction*. Springer, 2001.

[18] A. L. Hopkins and C. R. Groom. The druggable genome. *Nat. Rev. Drug Discov.*, 1(9):727–730, Sep 2002.

[19] L. Jacob, F. Bach, and J.-P. Vert. Clustered multi-task learning: A convex formulation. In *Advances in Neural Information Processing Systems 21*, pages 745–752. MIT Press, 2009.

[20] L. Jacob, B. Hoffmann, B. Stoven, and J.-P. Vert. Virtual screening of GPCRs: an *in silico* chemogenomics approach. *BMC Bioinformatics*, 9:363, 2008.

[21] L. Jacob, P. Neuvial, and S. Dudoit. Gains in power from structured two-sample tests of means on graphs. Technical Report arXiv:q-bio/1009.5173v1, arXiv, 2010.

[22] 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.

[23] L. Jacob and J.-P. Vert. Efficient peptide-MHC-I binding prediction for alleles with few known binders. *Bioinformatics*, 24(3):358–366, Feb 2008.

[24] R. Jenatton, J.-Y. Audibert, and F. Bach. Structured Variable Selection with Sparsity-Inducing Norms. Research report, WILLOW - INRIA, 2009.

[25] R. Jenatton, G. Obozinski, and F. Bach. Structured sparse principal component analysis. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2010.

[26] K. Knight and W. Fu. Asymptotics for lasso-type estimators. *Ann. Stat.*, 28(5):1356–1378, 2000.

[27] S. Kumagai. An implicit function theorem: Comment. *Journal of Optimization Theory and Applications*, 31:285–288, Jun 1980.

[28] S. R. Land and J. H. Friedman. Variable fusion: A new adaptive signal regression method. Technical Report 656, Department of Statistics, Carnegie Mellon University Pittsburgh, 1997.

[29] 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.

[30] C. Li and H. Li. Network-constrained regularization and variable selection for analysis of genomic data. *Bioinformatics*, 24(9):1175–1182, May 2008.

[31] Y. Lu, P.-Y. Liu, P. Xiao, and H.-W. Deng. Hotelling's t2 multivariate profiling for detecting differential expression in microarrays. *Bioinformatics*, 21(14):3105–3113, Jul 2005.

[32] C. Manly, S. Louise-May, and J. Hammer. The impact of informatics and computational chemistry on synthesis and screening. *Drug Discov. Today*, 6(21):1101–1110, Nov 2001.

[33] E. R. Mardis. Next-generation dna sequencing methods. *Annu. Rev. Genomics Hum. Genet.*, 9:387–402, 2008.

[34] 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.

[35] S. Mook, L. J. V. Veer, E. J. T. Rutgers, M. J. Piccart-Gebhart, and F. Cardoso. Individualization of therapy using mammaprint: from development to the mindact trial. *Cancer Genomics Proteomics*, 4(3):147–155, 2007.

[36] A. Mortazavi, B. A. Williams, K. McCue, L. Schaeffer, and B. Wold. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods*, 5(7):621–628, Jul 2008.

[37] S. Mukherjee, P. Tamayo, J. P. Mesirov, D. Slonim, A. Verri, and T. Poggio. Support vector machine classification of microarray data. Technical Report 182, C.B.L.C., 1998. A.I. Memo 1677.

[38] G. Obozinski, M. J. Wainwright, and M. I. Jordan. Union support recovery in high-dimensional multivariate regression. Technical Report 0808.0711v1, arXiv, August 2008.

[39] D. Pinkel, R. Segraves, D. Sudar, S. Clark, I. Poole, D. Kowbel, C. Collins, W. L. Kuo, C. Chen, Y. Zhai, S. H. Dairkee, B. M. Ljung, J. W. Gray, and D. G. Albertson. High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nat. Genet.*, 20(2):207–211, Oct 1998.

[40] F. Rapaport, E. Barillot, and J.-P. Vert. Classification of arrayCGH data using fused SVM. *Bioinformatics*, 24(13):i375–i382, Jul 2008.

[41] F. Rapaport, A. Zynoviev, M. Dutreix, E. Barillot, and J.-P. Vert. Classification of microarray data using gene networks. *BMC Bioinformatics*, 8:35, 2007.

[42] A. Rinaldo. Properties and refinements of the fused lasso. *Annals of Statistics*, 37(5B):2922–2952, 2009.

[43] V. Roth. The generalized lasso: a wrapper approach to gene selection for microarray data. In *Proc. CADE-14, 252–255*, 2002.

[44] 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.

[45] T. Sandler, J. Blitzer, P. Talukdar, and F. Pereira. Regularized learning with networks of features. In *Neural Information Processing Systems*, Cambridge, MA, 2009. MIT Press.

[46] N. Srebro, J. D. M. Rennie, and T. S. Jaakkola. Maximum-margin matrix factorization. In L. K. Saul, Y. Weiss, and L. Bottou, editors, *Adv. Neural. Inform. Process Syst. 17*, pages 1329–1336, Cambridge, MA, 2005. MIT Press.

[47] N. Srebro and A. Shraibman. Rank, trace-norm and max-norm. In *COLT*, pages 545–560, 2005.

[48] 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.

[49] R. Tibshirani. Regression shrinkage and selection via the lasso. *J. Royal. Statist. Soc. B.*, 58(1):267–288, 1996.

[50] R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, and K. Knight. Sparsity and smoothness via the fused lasso. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 67(1):91–108, 2005.

[51] R. Tibshirani and P. Wang. Spatial smoothing and hot spot detection for cgh data using the fused lasso. *Biostatistics (Oxford, England)*, 9(1):18–29, January 2008.

[52] E. van Beers and P. Nederlof. Array-CGH and breast cancer. *Breast Cancer Research*, 8(3):210, 2006.

[53] 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.

[54] L. J. van 't Veer, H. Dai, M. J. van de Vijver, Y. D. He, A. A. M. Hart, M. Mao, H. L. Peterse, K. van der Kooy, M. J. Marton, A. T. Witteveen, G. J. Schreiber, R. M. Kerkhoven, C. Roberts, P. S. Linsley, R. Bernards, and S. H. Friend. Gene expression profiling predicts clinical outcome of breast cancers. *Nature*, 415(6871):530–536, Jan 2002.

[55] V. Vapnik and A. Y. Chervonenkis. Teoriya raspoznavaniya obrazov: Statisticheskie problemy obucheniya. (Russian) [Theory of Pattern Recognition: Statistical Problems of Learning]. Moscow: Nauka, 1974.

[56] V. N. Vapnik. *The nature of statistical learning theory*. Springer-Verlag New York, Inc., New York, NY, USA, 1995.

[57] V. N. Vapnik. *Statistical Learning Theory*. Wiley, New-York, 1998.

[58] J.-P. Vert and K. Bleakley. Fast detection of multiple change-points shared by many signals using group lars. In J. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems 23*, pages 2343–2351. MIT Press, 2010.

[59] Z. Wang, M. Gerstein, and M. Snyder. Rna-seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.*, 10(1):57–63, Jan 2009.

[60] 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.

[61] M. Yuan and Y. Lin. On the non-negative garrotte estimator. *Journal Of The Royal Statistical Society Series B*, 69(2):143–161, 2007.

[62] P. Zhao and B. Yu. On model selection consistency of lasso. *J. Mach. Learn. Res.*, 7:2541, 2006.

[63] H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society B*, 67:301–320, 2005.

Appendix A: Proof of Lemma 1

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

$$\Omega^{\mathcal{G}}_{\text{overlap}}(\beta+\beta') \leq \sum_{g\in\mathcal{G}} \|v_g + v'_g\| \leq \sum_g (\|v_g\| + \|v'_g\|) = \Omega^{\mathcal{G}}_{\text{overlap}}(\beta) + \Omega^{\mathcal{G}}_{\text{overlap}}(\beta') \,.$$

□

Appendix B: Proof of Lemma 2

Let us introduce slack variables $\mathbf{t} = (t_g)_{g\in\mathscr{G}} \in \mathbb{R}^{\mathscr{G}}$ and rewrite the optimization problem (17) as follows:

$$\min_{\mathbf{t}\in\mathbb{R}^{\mathscr{G}},\mathbf{v}\in\mathscr{V}_{\mathscr{G}}} \sum_{g\in\mathscr{G}} t_g \text{ s.t. } \sum_{g\in\mathscr{G}} v_g = w \text{ and } \forall g \in \mathscr{G}, \|v_g\| \le t_g\,.$$

We can form a Lagrangian [7] for this problem with the dual variables $\alpha \in \mathbb{R}^p$ for the constraint $\sum_{g\in\mathscr{G}} v_g = w$, and $(\delta,\gamma) \in \mathscr{V}_{\mathscr{G}} \times \mathbb{R}^{\mathscr{G}}$ with $\|\delta_g\| \le \gamma_g$ for the conic constraints $\|v_g\| \le t_g$, and get:

$$L = \sum_{g\in\mathscr{G}} t_g + \alpha^\top\Big(w - \sum_{g\in\mathscr{G}} v_g\Big) - \sum_{g\in\mathscr{G}}\Big(\delta_g^\top v_g + \gamma_g t_g\Big)\,.$$

The minimum of L with respect to the primal variables $\mathbf{t}$ and $\mathbf{v}$ is non trivial only if $\gamma_g = 1$ and $\alpha_g = -\delta_g$ for any $g \in \mathscr{G}$. Therefore, we get the dual function:

$$\min_{\mathbf{t},\mathbf{v}} L = \begin{cases} \alpha^\top w & \text{if } \gamma_g = 1 \text{ and } \alpha_g = -\delta_g \text{ for all } g \in \mathscr{G}\,, \\ -\infty & \text{otherwise.} \end{cases}$$

By strong duality (since, *e.g.*, Slater's condition is fulfilled), the optimal value $\Omega^{\mathscr{G}}_{\text{overlap}}(w)$ of the primal is equal to the maximum of the dual problem. Maximizing this dual function over $\gamma_g = 1$, $\|\delta_g\| \le \gamma_g$ and $\alpha_g = -\delta_g$ is equivalent to maximizing $\alpha^\top w$ over the vectors $\alpha \in \mathbb{R}^p$ such that $\|\alpha_g\| \le 1$ for all $g \in \mathscr{G}$, which proves (18). To prove the second point, we note that the variables $(\mathbf{t},\mathbf{v},\alpha,\delta,\gamma)$ are primal/dual optimal for this convex optimization problem if and only if the Karush-Kuhn-Tucker (KKT) conditions are satisfied, *i.e.*, if and only if, for all $g \in \mathscr{G}$:

$$\begin{cases} \operatorname{supp}(v_g) = g, \|v_g\| \le t_g \quad \text{and} \quad w = \sum_{g\in\mathscr{G}} v_g \\ \operatorname{supp}(\delta_g) = g, \|\delta_g\| \le \gamma_g \\ \alpha_g = -\delta_g \text{ and } \gamma_g = 1 \\ \delta_g^\top v_g + \gamma_g t_g = 0 \end{cases}$$

Eliminating δ and γ with the stationarity conditions, all conditions are fulfilled if and only if $w = \sum_{g\in\mathscr{G}} v_g$ and for all $g \in \mathscr{G}$, (i) either $v_g = 0$ and $\|\alpha_g\| \le 1$, (ii) or $v_g \ne 0$ and $\alpha_g = v_g/\|v_g\|$. If a pair $(\alpha,\mathbf{v})$ fulfills these conditions, then we obtain a primal/dual solution by taking $t_g = \|v_g\|$, $\delta_g = -\alpha_g$ and $\gamma_g = 1$. This proves points 2 and 3. □

Appendix C: Proof of Lemma 3

Consider any solution $\mathbf{v} = (v_g)_{g\in\mathscr{G}}$ of (17). Let α be any optimal solution of (18). Since $(\mathbf{v},\alpha)$ form a primal/dual pair, they must satisfy the KKT conditions. In particular, for all g such that $v_g \ne 0$, α_g is defined uniquely by $\alpha_g = \frac{v_g}{\|v_g\|}$. Since this is true for all solutions $\mathbf{v} \in \mathbf{V}(w)$, α_{J_1} is uniquely defined. □

Appendix D: Proof of Corollary 1

If $v_g \ne 0$ and $v'_g \ne 0$, let α be solution of (18), by the previous lemma α_g is unique and $\alpha_g = \frac{v_g}{\|v_g\|} = \frac{v'_g}{\|v'_g\|}$. □

Appendix E: Proof of Lemma 5 (consistency)

In order to prove the consistency result on $\Omega^{\mathcal{G}}_{\text{overlap}}(.)$, we will need the following lemma.

Lemma 6. *Assume that for all β' in a small neighborhood U of β, β' admits a unique decomposition $(v'_g)_{g\in\mathcal{G}}$ of minimal norm supported by the same set of groups $\mathcal{G}_1$ as β. Writing $\eta_g = \|v_g\|$, there exists a neighborhood U_0 of β_{J_1} in $\mathbb{R}^{|J_1|}$ and a neighborhood U'_0 of $(\alpha_{J_1}, \eta_{\mathcal{G}_1})$ in $\mathbb{R}^{|J_1|\times|\mathcal{G}_1|}$ such that there exists a unique continuous function*

$$\phi : \beta_{J_1} \mapsto (\alpha_{J_1}(\beta), \eta_{\mathcal{G}_1}(\beta))$$

from U_0 to U'_0.

Proof. The dual problem (18) is equivalent to the saddle-point problem

$$\min_{\alpha}\max_{\eta} L'(\alpha,\eta,\beta) \text{ s.t. } \eta_g \in \mathbb{R}_+,$$

with Lagrangian

$$L'(\alpha,\eta,\beta) = -\alpha^\top\beta + \sum_{g\in\mathcal{G}} \frac{\eta_g}{2}(\|\alpha_g\|^2 - 1)$$

and KKT conditions:

$$\begin{cases} \forall g \in \mathcal{G}, \|\alpha_g\|^2 \leq 1, & \text{(primal feas.)} \\ \forall g \in \mathcal{G}, \eta_g \geq 0, & \text{(dual feas.)} \\ \forall i \in [1,p], -\beta_i + \left(\sum_{g\ni i}\eta_g\right)\alpha_i = 0, & \text{(stationarity)} \\ \forall g \in \mathcal{G}, \eta_g(\|\alpha_g\|^2 - 1) = 0, & \text{(comp.slack.)} \end{cases}$$

By stationarity, $(v_g)_{g\in\mathcal{G}}$ defined by $v_g = \eta_g\alpha_g$ is a decomposition of β; it is optimal because it satisfies property 3 of lemma 2; finally we have $\eta_g = \|v_g\|$ consistently with our definition of $\eta_g(\beta)$. For any β with the same set of supporting groups $\mathcal{G}_1$, we have $\|\alpha_g(\beta)\| = 1$ for all $g \in \mathcal{G}_1$ and $\eta_g = 0$ for all $g \in \mathcal{G}\backslash\mathcal{G}_1$. For all β_{J_1} with group-support no smaller than $\mathcal{G}_1$, the corresponding pair $(\alpha_{J_1}(\beta), \eta_{\mathcal{G}_1}(\beta))$ is therefore a solution of the set of non-linear equations:

$$\begin{cases} \forall i \in J_1, -\beta_i + \left(\sum_{g\ni i}\eta_g\right)\alpha_i = 0 \\ \forall g \in \mathcal{G}_1, \|\alpha_g\|^2 - 1 = 0 \end{cases} \tag{23}$$

In other words consider the function

$$\begin{array}{rcl} F : \mathbb{R}^{|J_1|\times|J_1|\times|\mathcal{G}_1|} & \rightarrow & \mathbb{R}^{|J_1|\times|\mathcal{G}_1|} \\ (\beta_{J_1}, \alpha_{J_1}, \eta_{\mathcal{G}_1}) & \mapsto & \begin{pmatrix} \left(-\beta_i + \left[\sum_{g\ni i}\eta_g\right]\alpha_i\right)_{i\in J_1} \\ (\|\alpha_g\|^2 - 1)_{g\in\mathcal{G}_1} \end{pmatrix}, \end{array}$$

then (23) is equivalent to $F(\beta_{J_1}, \alpha_{J_1}, \eta_{\mathcal{G}_1}) = 0$. We use the implicit function theorem for non-differentiable function of [27]. The theorem states that for a continuous function

$$F : \mathbb{R}^{|J_1|} \times \mathbb{R}^{|J_1|\times|\mathcal{G}_1|} \rightarrow \mathbb{R}^{|J_1|\times|\mathcal{G}_1|},$$

such that $F(\beta_0,(\alpha_0,\eta_0)) = 0$, if there exist open neighborhoods $U \subset R^{|J_1|}$ and $U' \subset R^{|J_1|\times|\mathscr{G}_1|}$ of β_0 and (α_0,η_0) respectively, such that, for all $\beta \in U$, $F(\beta,\cdot) : U' \to R^{|J_1|\times|\mathscr{G}_1|}$ is locally one-to-one then there exist open neighborhoods $U_0 \subset R^{|J_1|}$ and $U_0' \subset R^{|J_1|\times|\mathscr{G}_1|}$ of β_0 and (α_0,η_0), such that, for all $\beta \in U_0$, the equation $F(\beta,(\alpha,\eta)) = 0$ has a unique solution $(\alpha,\eta) = \phi(\beta) \in U_0'$, where ϕ is a continuous function from U_0 into U_0'. By continuity of the addition, the product and the Euclidean norm, the above defined F is continuous. For each β fixed, $F(\beta,\cdot)$ is bijective, because of the assumption of the existence of a unique decomposition in a neighborhood of β. Applying the theorem of [27] then yields the desired result. □

We follow the line of proof of [3] but consider a fixed design for simplicity of notations. Let us first consider the subproblem of estimating a vector only on the support of $\bar{\beta}$ by using only the groups in J_1 in the penalty, *i.e.*, consider $\beta_1 \in \mathbb{R}^{J_1}$ a solution of $\min_{\beta_{J_1}\in\mathbb{R}^{J_1}} \frac{1}{2n}\|Y - X_{J_1}\beta_{J_1}\|^2 + \lambda_n\Omega^{\mathscr{G}_1}_{\text{overlap}}(\beta_{J_1})$. By standard arguments, we can prove that β_1 converges in Euclidean norm to $\bar{\beta}$ restricted to J_1 as n tends to infinity [26]. In the rest of the proof we show how to construct a vector $\beta \in \mathbb{R}^p$ from β_1 which under condition (C2) is with high probability a solution to (21). By adding null components to β_1, we obtain a vector $\beta \in \mathbb{R}^p$ whose support is also J_1, and $u = \beta - \bar{\beta}$ therefore satisfies $\operatorname{supp}(u) \subset J_1$. A direct computation of the gradient of the loss $L(\beta) = \|Y - X\beta\|^2$ gives $\nabla L(\beta) = \Sigma u - B$, where $B = \frac{1}{n}X\varepsilon$. From this we deduce that $u = \Sigma^{-1}_{J_1J_1}(\nabla_{J_1}L(\beta) + B_{J_1})$, and since $\nabla_{J_1}L(\beta) = -\lambda_n\alpha_{J_1}(\beta)$ we have :

$$\nabla_{J_2}L(\beta) = \Sigma_{J_2J_1}\Sigma^{-1}_{J_1J_1}(B_{J_1} - \lambda_n\alpha_{J_1}(\beta)) - B_{J_2}\,.$$

To show that β is a feasible solution to (21) it is enough to show that $\forall g \in \mathscr{G}_2$, $\|\nabla_g L(\beta)\| \leq \lambda_n$. Moreover, since the noise has bounded variance,

$$\Sigma_{J_2J_1}\Sigma^{-1}_{J_1J_1}B_{J_1} - B_{J_2} = X_{J_2}^\top\left[\frac{1}{n}X_{J_1}\Sigma^{-1}_{J_1J_1}X_{J_1}^\top - I\right]\varepsilon$$

is $\sqrt{n}$-consistent and

$$\frac{1}{\lambda_n}\|\nabla_g L(\beta)\| \leq \|\Sigma_{gJ_1}\Sigma^{-1}_{J_1J_1}\alpha_{J_1}(\beta)\| + \mathscr{O}_p(\lambda_n^{-1}n^{-1/2}).$$

By Lemma 6, we have that α_{J_1} is a continuous function of β in a neighborhood of $\bar{\beta}$ so that $\beta_{J_1} \xrightarrow{\mathbb{P}} \bar{\beta}_{J_1}$ implies $\alpha_{J_1}(\beta) \xrightarrow{\mathbb{P}} \alpha_{J_1}(\bar{\beta})$. Since we chose λ_n such that $\lambda_n^{-1}n^{-1/2} \to 0$, we have

$$\frac{1}{\lambda_n}\|\nabla_g L(\beta)\| \leq \|\Sigma_{gJ_1}\Sigma^{-1}_{J_1J_1}\alpha_{J_1}(\bar{\beta})\| + o_p(1).$$

Hence the result for the sufficient condition. Symmetrically, for the necessary condition we have

$$\frac{1}{\lambda_n}\|\nabla_g L(\beta)\| \geq \|\Sigma_{gJ_1}\Sigma^{-1}_{J_1J_1}\alpha_{J_1}(\bar{\beta})\| - o_p(1).$$

□