Group Lasso with Overlap and Graph Lasso

Laurent Jacob LAURENT.JACOB@MINES-PARISTECH.FR
Mines ParisTech – CBIO, INSERM U900, Institut Curie, France

Guillaume Obozinski GOBO@STAT.BERKELEY.EDU
Department of Statistics, University of California, Berkeley, USA

Jean-Philippe Vert JEAN-PHILIPPE.VERT@MINES-PARISTECH.FR
Mines ParisTech – CBIO, INSERM U900, Institut Curie, France

Abstract

We propose a new penalty function which, when used as regularization for empirical risk minimization procedures, leads to sparse estimators. The support of the sparse vector is typically a union of potentially overlapping groups of covariates defined a priori, or a set of covariates which tend to be connected to each other when a graph of covariates is given. We study theoretical properties of the estimator, and illustrate its behavior on simulated and breast cancer gene expression data.

1. Introduction

Estimation of sparse linear models by the minimization of an empirical error penalized by a regularization term is a very popular and successful approach in statistics and machine learning. Controlling the trade-off between data fitting and regularization, one can obtain estimators with good statistical properties, even in very large dimension. Moreover, sparse classifiers lend themselves particularly well to interpretation, which is often of primary importance in many applications such as biology or social sciences. A popular example of such procedures is the penalization of a least-square criterion by the ℓ_1 norm of the estimator, known as *lasso* (Tibshirani, 1996) or *basis pursuit* (Chen et al., 1998). Interestingly, the lasso is able to recover the exact support of a sparse model from data generated by this model if the covariates are not too correlated (Zhao & Yu, 2006; Wainwright, 2006).

While the ℓ_1 norm penalty leads to sparse models, it does not contain any prior information about, *e.g.*, possible groups of covariates that one may wish to see selected jointly. Several authors have recently proposed new penalties to enforce the estimation of models with specific sparsity patterns. For example, when the covariates are partitioned into groups, the *group lasso* leads to the selection of groups of covariates (Yuan & Lin, 2006). The group lasso penalty for a model, also called ℓ_1/ℓ_2 penalty, is the sum (*i.e.*, ℓ_1 norm) of the ℓ_2 norms of the restrictions of the model to the different groups of covariates. It recovers the support of a model if the support is a union of groups and if covariates of different groups are not too correlated. It can be generalized to an infinite-dimensional setting (Bach, 2008). Other variants of the group lasso include joint selection of covariates for multi-task learning (Obozinski et al., 2009) and penalties to enforce hierarchical selection of covariates, *e.g.*, when one has a hierarchy over the covariates and wants to select covariates only if their ancestors in the hierarchy are also selected (Zhao et al., 2009; Bach, 2009).

In this paper we are interested in a more general situation. We assume that either (i) groups of covariates are given, potentially with overlap between the groups, and we wish to estimate a model whose support is a union of groups, or (ii) that a graph with covariates as vertices is given, and we wish to estimate a model whose support contains covariates which tend to be connected to each others on the graph. Although quite general, this framework is motivated in particular by applications in bioinformatics, when we have to solve classification or regression problems with few samples in high dimension, such as predicting the class of a tumour from gene expression measurements with microarrays, and simultaneously select a few genes to establish a predictive signature (Roth, 2002). Selecting a few genes that either belong to the same functional groups, where the groups are given a priori and may overlap, or tend to be connected to each other in a given biological network,

Appearing in *Proceedings of the 26^{th} International Conference on Machine Learning*, Montreal, Canada, 2009.

could then lead to increased interpretability of the signature and potential better performances (Rapaport et al., 2007).

To reach this goal, we propose and study a new penalty which generalizes the ℓ_1/ℓ_2 norm to overlapping groups for the first case, and propose to cast the problem of selecting connected covariates in a graph as the problem of selecting a union of overlapping groups, with adequate definition of groups, for the second case. We mention various properties of this penalty, and provide conditions for the consistency of support estimation in the regression setting. Finally, we report promising results on both simulated and real data.

2. Problem and notations

For any vector $w \in \mathbb{R}^p$, $\|w\|$ denotes the Euclidean norm of w, and $\text{supp}\,(w) \subset [1,p]$ denotes the support of w, *i.e.*, the set of covariates $i \in [1,p]$ such that $w_i \neq 0$. 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 $[1,p]$. Throughout the paper, $\mathcal{G} \subset \mathcal{P}([1,p])$ denotes a set of groups, usually fixed in advance for each application. We say that two groups overlap if they have at least one covariate in common. For any vector $w \in \mathbb{R}^p$, and any group $g \in \mathcal{G}$, we denote $w_g \in \mathbb{R}^p$ the vector whose entries are the same as w for the covariates in g, and are 0 for other other covariates. However, we use a different convention for elements of $\mathcal{V}_\mathcal{G} \subset \mathbb{R}^{p \times \mathcal{G}}$ the set of $|\mathcal{G}|$-tuples of vectors $\mathbf{v} = (v_g)_{g \in \mathcal{G}}$, where each v_g is this time a separate vector in $\mathbb{R}^p$, which satisfies $\text{supp}\,(v_g) \subset g$ for each $g \in \mathcal{G}$. For any differentiable function $f : \mathbb{R}^p \to \mathbb{R}$, we denote by $\nabla f(w) \in \mathbb{R}^p$ the gradient of f at $w \in \mathbb{R}^p$ and by $\nabla_g f(w) \in \mathbb{R}^g$ the partial gradient of f with respect to to the covariates in g.

3. Group lasso with overlapping groups

When the groups in $\mathcal{G}$ do not overlap, the group lasso penalty (Yuan & Lin, 2006) is defined as:

$$\forall w \in \mathbb{R}^p\,, \quad \Omega^\mathcal{G}_{\text{group}}\,(w) = \sum_{g \in \mathcal{G}} \|w_g\|\,. \tag{1}$$

When the groups in $\mathcal{G}$ form a partition of the set of covariates, then $\Omega^\mathcal{G}_{\text{group}}\,(w)$ is a norm whose balls have singularities when some w_g are equal to zero. Minimizing a smooth convex risk functional over such a ball often leads to a solution that lies on a singularity, *i.e.*, to a vector w such that $w_g = 0$ for some of the g in $\mathcal{G}$.

When some of the groups in $\mathcal{G}$ overlap, the penalty (1) is still a norm (if all covariates are in at least one group) whose ball has singularities when some w_g are equal to zero. Indeed, for a vector w, if we denote by $\mathcal{G}_0 \subset \mathcal{G}$ the set of groups such that $w_g = 0$, then

$$\text{supp}\,(w) \subset \Big(\bigcup_{g \in \mathcal{G}_0} g\Big)^c\,.$$

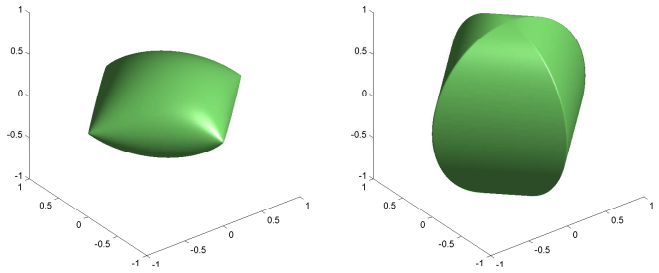

Figure 1. Balls for $\Omega^\mathcal{G}_{\text{group}}\,(\cdot)$ (left) and $\Omega^\mathcal{G}_{\text{overlap}}\,(\cdot)$ (right) for the groups $\mathcal{G} = \{\{1,2\},\{2,3\}\}$ where w_2 is represented as the vertical coordinate.

We see that this penalty induces the estimation of sparse vectors, whose support in typically the complement of a union of groups. Although this may be relevant for some applications, with appropriately designed families of groups — as considered by (Jenatton et al., 2009) — , we are interested in this paper in penalties which induce the opposite effect: that the support of w be a union of groups. For that purpose, we propose instead the following penalty:

$$\Omega^\mathcal{G}_{\text{overlap}}\,(w) = \inf_{\mathbf{v} \in \mathcal{V}_\mathcal{G}, \sum_{g \in \mathcal{G}} v_g = w} \sum_{g \in \mathcal{G}} \|v_g\|\,. \tag{2}$$

When the groups do not overlap and form a partition of $[1,p]$, there exists a unique decomposition of $w \in \mathbb{R}^p$ as $w = \sum_{g \in \mathcal{G}} v_g$ with $\text{supp}\,(v_g) \subset g$, namely, $v_g = w_g$ for all $g \in \mathcal{G}$. In that case, both penalties (1) and (2) are the same. If some groups overlap, then we show below that this penalty induces the selection of w that can be decomposed as $w = \sum_{g \in \mathcal{G}} v_g$ where some v_g are equal to 0. If we denote by $\mathcal{G}_1 \subset \mathcal{G}$ the set of groups g with $v_g \neq 0$, then we immediately get $w = \sum_{g \in \mathcal{G}_1} v_g$, and therefore:

$$\text{supp}\,(w) \subset \bigcup_{g \in \mathcal{G}_1} g\,.$$

In other words, the penalty (2) leads to sparse solutions whose support is typically a union of groups, matching the setting of applications that motivate this work. In the rest of this paper, we therefore investigate in more details $\Omega^\mathcal{G}_{\text{overlap}}\,(.)$, both theoretically and empirically.

Figure 1 shows the ball for both norms in $\mathbb{R}^3$ with groups $\mathcal{G} = \{\{1,2\},\{2,3\}\}$. The pillow shaped ball of $\Omega^\mathcal{G}_{\text{group}}\,(\cdot)$ has four singularities corresponding to cases where either only w_1 or only w_3 is non-zero. By contrast, $\Omega^\mathcal{G}_{\text{overlap}}\,(\cdot)$ has two circular sets of singularities corresponding to cases where (w_1, w_2) only or (w_2, w_3) only is non zero.

4. Some properties of $\Omega^\mathcal{G}_{\text{overlap}}\,(.)$

We first analyze the decomposition of a vector $w \in \mathbb{R}^p$ as $\sum_{g \in \mathcal{G}} v_g$ induced by (2). For that purpose, let $\mathbf{V}(w) \subset \mathcal{V}_\mathcal{G}$

be the set of $|\mathcal{G}|$-tuples of vectors $\mathbf{v} = (v_g)_{g\in\mathcal{G}}$ which reach the minimum in (2), *i.e.*, which satisfy

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

The optimization problem (2) defining $\Omega^{\mathcal{G}}_{\text{overlap}}(w)$ is a convex problem and its objective is coercive, so that the set of solutions $\mathbf{V}(w)$ is non-empty and convex. Moreover,

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

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

Using the conic dual of (2), we give another formulation of the norm $\Omega^{\mathcal{G}}_{\text{overlap}}(.)$ yelding some important properties.

Lemma 2. *1. It holds that:*

$$\Omega^{\mathcal{G}}_{overlap}(w) = \sup_{\alpha\in\mathbb{R}^p : \forall g\in\mathcal{G}, \|\alpha_g\|\leq 1} \alpha^\top w . \quad (3)$$

2. A vector $\alpha \in \mathbb{R}^p$ is a solution of (3) *if and only if there exists $\mathbf{v} = (v_g)_{g\in\mathcal{G}} \in \mathbf{V}(w)$ we have:*

$$\forall g \in \mathcal{G}, \ \text{if } v_g \neq 0, \ \alpha_g = \tfrac{v_g}{\|v_g\|} \ \text{else} \ \|\alpha_g\| \leq 1 \quad (4)$$

3. Conversely, a $\mathcal{G}$-tuple of vectors $\mathbf{v} = (v_g)_{g\in\mathcal{G}} \in \mathcal{V}_{\mathcal{G}}$ such that $w = \sum_g v_g$ is a solution to (2) *if and only if there exists a vector $\alpha \in \mathbb{R}^p$ such that* (4) *holds.*

Proof. Let us introduce slack variables $\mathbf{t} = (t_g)_{g\in\mathcal{G}} \in \mathbb{R}^{\mathcal{G}}$ and rewrite the optimization problem (2) as follows:

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

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

$$L = \sum_{g\in\mathcal{G}} t_g + \alpha^\top \Big(w - \sum_{g\in\mathcal{G}} v_g\Big) - \sum_{g\in\mathcal{G}} \left(\beta_g^\top v_g + \gamma_g t_g\right).$$

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

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

By strong duality (since, *e.g.*, Slater's condition is fulfilled), the optimal value $\Omega^{\mathcal{G}}_{\text{overlap}}(w)$ of the primal is equal to the maximum of the dual problem. Maximizing this dual function over $\gamma_g = 1$, $\|\beta_g\| \leq \gamma_g$ and $\alpha_g = -\beta_g$ is equivalent to maximizing $\alpha^\top w$ over the vectors $\alpha \in \mathbb{R}^p$ such that $\|\alpha_g\| \leq 1$ for all $g \in \mathcal{G}$, which proves (3). To prove the second point, we note that the variables $(\mathbf{t}, \mathbf{v}, \alpha, \beta, \gamma)$ 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 $g \in \mathcal{G}$:

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

Eliminating β and γ with the stationarity conditions, all conditions are fulfilled if and only if $w = \sum_{g\in\mathcal{G}} v_g$ and for all $g \in \mathcal{G}$, (i) either $v_g = 0$ and $\|\alpha_g\| \leq 1$, (ii) or $v_g \neq 0$ 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 $t_g = \|v_g\|$, $\beta_g = -\alpha_g$ and $\gamma_g = 1$. This proves points 2 and 3. □

Denote by $\mathcal{G}_1$ the group-support of w, i.e., the set of groups belonging to the support of at least one optimal decomposition of w: $\mathcal{G}_1 = \{g \in \mathcal{G} \mid \exists \mathbf{v} = (v_g)_g \in \mathbf{V}(w),\ v_g \neq 0\}$ and J_1 the corresponding set of variables $J_1 = \cup_{g\in\mathcal{G}_1} g$.

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

Proof. Consider any solution $\mathbf{v} = (v_g)_{g\in\mathcal{G}}$ of (2). Let α be any optimal solution of (3). Since $(\mathbf{v}, \alpha)$ form a primal/dual pair, they must satisfy the KKT conditions. In particular, for all g such that $v_g \neq 0$, 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. □

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

$$\|v_g\| \times \|v'_g\| = 0 \quad \textit{or} \quad \exists \gamma_g \geq 0 \ \textit{s.t.}\ v'_g = \gamma v_g . \quad (5)$$

Proof. If $v_g \neq 0$ and $v'_g \neq 0$, let α be solution of (3), by the previous lemma α_g is unique and $\alpha_g = \frac{v_g}{\|v_g\|} = \frac{v'_g}{\|v'_g\|}$. □

5. Using $\Omega^{\mathcal{G}}_{\text{overlap}}(.)$ as a penalty

We now consider a learning scenario where we use $\Omega^{\mathcal{G}}_{\text{overlap}}(w)$ as a regularization term to the minimization of an objective function $R(w)$, typically an empirical risk. We assume that $R(w)$ is convex and differentiable in w, and consider the optimization problem:

$$\min_{w\in\mathbb{R}^p} R(w) + \lambda \Omega^{\mathcal{G}}_{\text{overlap}}(w) , \quad (6)$$

where $\lambda > 0$ is a regularization parameter. We first derive optimality conditions for any solution of (6). For that purpose, let us denote $\mathcal{A}_{\mathcal{G}}(w)$ the set of vectors $\alpha \in \mathbb{R}^p$ solution of (3).

Lemma 4. *A vector $w \in \mathbb{R}^p$ is a solution of* (6) *if and only if $-\nabla R(w)/\lambda \in \mathcal{A}_{\mathcal{G}}(w)$.*

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 $w \in \mathbb{R}^p$ is a solution of* (6) *if and only if it can be decomposed as $w = \sum_{g\in\mathcal{G}} v_g$ where, for any $g \in \mathcal{G}$, $v_g \in \mathbb{R}^p$, supp $(v_g) = g$, and if $v_g = 0$ then $\|\nabla_g R(w)\| \leq \lambda$, and $\nabla_g R(w) = -\lambda v_g/\|v_g\|$ otherwise.*

6. Consistency

Before we present a consistency result on $\Omega^{\mathcal{G}}_{\text{overlap}}(.)$, we will need the following lemma.

Lemma 5. *Assume that for all w' in a small neighborhood U of w, w' admits a unique decomposition $(v'_g)_{g\in\mathcal{G}}$ of minimal norm supported by the same set of groups $\mathcal{G}_1$ as w. Writing $\eta_g = \|v_g\|$, there exists a neighborhood U_0 of w_{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 $\phi : w_{J_1} \mapsto (\alpha_{J_1}(w), \eta_{\mathcal{G}_1}(w))$ from U_0 to U'_0.*

Proof. The dual problem (3) is equivalent to the saddle-point problem $\min_\alpha \max_\eta L'(\alpha, \eta, w)$ s.t. $\eta_g \in \mathbb{R}_+$ with lagrangian $L'(\alpha, \eta, w) = -\alpha^\top w + \sum_{g\in\mathcal{G}} \frac{\eta_g}{2}(\|\alpha_g\|^2 - 1)$ and KKT conditions:

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

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

$$\begin{cases} \forall i \in J_1, -w_i + \left(\sum_{g\ni i} \eta_g\right) \alpha_i = 0 \\ \forall g \in \mathcal{G}_1, \|\alpha_g\|^2 - 1 = 0 \end{cases} \tag{7}$$

In other words consider the function

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

then (7) is equivalent to $F(w_{J_1}, \alpha_{J_1}, \eta_{\mathcal{G}_1}) = 0$. We use the implicit function theorem for non-differentiable function of (Kumagai, 1980). The theorem states that for a continuous function $F : \mathbb{R}^{|J_1|} \times \mathbb{R}^{|J_1|\times|\mathcal{G}_1|} \to \mathbb{R}^{|J_1|\times|\mathcal{G}_1|}$ such that $F(w_0, (\alpha_0, \eta_0)) = 0$, if there exist open neighborhoods $U \subset R^{|J_1|}$ and $U' \subset R^{|J_1|\times|\mathcal{G}_1|}$ of w_0 and (α_0, η_0) respectively, such that, for all $w \in U$, $F(w, \cdot) : U' \to R^{|J_1|\times|\mathcal{G}_1|}$ is locally one-to-one then there exist open neighborhoods $U_0 \subset R^{|J_1|}$ and $U'_0 \subset R^{|J_1|\times|\mathcal{G}_1|}$ of w_0 and (α_0, η_0), such that, for all $w \in U_0$, the equation $F(w, (\alpha, \eta)) = 0$ has a unique solution $(\alpha, \eta) = \phi(w) \in U'_0$, 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 w fixed, $F(w, \cdot)$ is bijective, because of the assumption of the existence of a unique decomposition in a neighborhood of w. Applying the theorem of (Kumagai, 1980) then yields the desired result.

□

We are now ready to prove the consistency of $\Omega^{\mathcal{G}}_{\text{overlap}}(.)$. Consider the linear regression model $Y = X\bar{w} + \epsilon$, where $X \in \mathbb{R}^{n\times p}$ is a design matrix, $Y \in \mathbb{R}^p$ is the response vector and $\epsilon \in \mathbb{R}^p$ is a vector of i.i.d. random variables with mean 0 and finite variance. We denote the true regression function by $\bar{w}$. We assume that

1. (H1) $\Sigma := \frac{1}{n} X^\top X \succ 0$
2. (H2) There exists a neighborhood of $\bar{w}$ in which (2) has a unique solution.

If $\mathcal{G}_1$ is the set of group supporting the unique solution of (2), we denote $\mathcal{G}_2 \triangleq \mathcal{G}\backslash\mathcal{G}_1$ and $J_2 \triangleq [1,p]\backslash J_1$. 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 $\Omega^{\mathcal{G}}_{\text{overlap}}(w)$ leads to an estimator with the correct support. Consider the two conditions:

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

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

Lemma 6. *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* (6) *to estimate consistently the group-support of $\bar{w}$.*

Proof. We follow the line of proof of (Bach, 2008) but consider a fixed design for simplicity of notations. Let us first consider the subproblem of estimating a vector only on the support of $\bar{w}$ by using only the groups in J_1 in the penalty, *i.e.*, consider $w_1 \in \mathbb{R}^{J_1}$ a solution of

$\min_{w_{J_1} \in \mathbb{R}^{J_1}} \frac{1}{2n} \|Y - X_{J_1} w_{J_1}\|^2 + \lambda_n \Omega^{\mathcal{G}_1}_{\text{overlap}}(w_{J_1})$. By standard arguments, we can prove that w_1 converges in Euclidean norm to $\bar{w}$ restricted to J_1 as n tends to infinity (Fu & Knight, 2000). In the rest of the proof we show how to construct a vector $w \in \mathbb{R}^p$ from w_1 which under condition (C2) is with high probability a solution to (6). By adding null components to w_1, we obtain a vector $w \in \mathbb{R}^p$ whose support is also J_1, and $u = w - \bar{w}$ therefore satisfies $\text{supp}(u) \subset J_1$. A direct computation of the gradient of the risk $R(w) = \|Y - Xw\|^2$ gives $\nabla R(w) = \Sigma u - W$, where $W = \frac{1}{n} X\epsilon$. From this we deduce that $u = \Sigma^{-1}_{J_1 J_1}(\nabla_{J_1} R(w) + W_{J_1})$, and since $\nabla_{J_1} R(w) = -\lambda_n \alpha_{J_1}(w)$ we have :

$$\nabla_{J_2} R(w) = \Sigma_{J_2 J_1} \Sigma^{-1}_{J_1 J_1} (W_{J_1} - \lambda_n \alpha_{J_1}(w)) - W_{J_2} .$$

To show that w is a feasible solution to (6) it is enough to show that $\forall g \in \mathcal{G}_2$, $\|\nabla_g R(w)\| \leq \lambda_n$. Moreover, since the noise has bounded variance, $\Sigma_{J_2 J_1} \Sigma^{-1}_{J_1 J_1} W_{J_1} - W_{J_2} = X^\top_{J_2} \left[\frac{1}{n} X_{J_1} \Sigma^{-1}_{J_1 J_1} X^\top_{J_1} - I\right] \epsilon$ is $\sqrt{n}$-consistent and

$$\tfrac{1}{\lambda_n} \|\nabla_g R(w)\| \leq \|\Sigma_{g J_1} \Sigma^{-1}_{J_1 J_1} \alpha_{J_1}(w)\| + \mathcal{O}_p(\lambda_n^{-1} n^{-1/2}).$$

By Lemma 5, we have that α_{J_1} is a continuous function of w in a neighborhood of $\bar{w}$ so that $w_{J_1} \xrightarrow{\mathbb{P}} \bar{w}_{J_1}$ implies $\alpha_{J_1}(w) \xrightarrow{\mathbb{P}} \alpha_{J_1}(\bar{w})$. Since we chose λ_n such that $\lambda_n^{-1} n^{-1/2} \to 0$, we have

$$\tfrac{1}{\lambda_n} \|\nabla_g R(w)\| \leq \|\Sigma_{g J_1} \Sigma^{-1}_{J_1 J_1} \alpha_{J_1}(\bar{w})\| + o_p(1).$$

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

$$\tfrac{1}{\lambda_n} \|\nabla_g R(w)\| \geq \|\Sigma_{g J_1} \Sigma^{-1}_{J_1 J_1} \alpha_{J_1}(\bar{w})\| - o_p(1).$$

□

7. 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 consider the prior $\Omega^{\mathcal{G}}_{\text{overlap}}(.)$ where $\mathcal{G}$ is a set that generates by union the connected components. 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 $\Omega_{\text{graph}}(w) = \min_{v \in \mathcal{V}_E} \sum_{e \in E} \|v_e\|$ s.t. $\sum_{e \in E} v_e = w$, $\text{supp}(v_e) = e$.

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 chose 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.

8. Implementation

A simple way to implement empirical risk minimization using $\Omega^{\mathcal{G}}_{\text{overlap}}(.)$ as the regularizer is to explicitly duplicate the variables in the design matrix, *i.e.*, to replace $X \in \mathbb{R}^{n \times p}$ by $\tilde{X} \in \mathbb{R}^{n \times \sum |g|}$ defined by the concatenation of copies of the design matrix restricted each to a certain group g, i.e., $\tilde{X} = [X_{g_1}, X_{g_2}, ..., X_{g_{|\mathcal{G}|}}]$, where $\mathcal{G} = \{g_1, \ldots, g_{\mathcal{G}}\}$. To see this, denote $\tilde{v}_g = (v_{gi})_{i \in g}$ and $\tilde{\mathbf{v}} = (\tilde{v}^\top_{g_1}, \ldots, \tilde{v}^\top_{g_{|\mathcal{G}|}})^\top$, and consider that, for an empirical risk of the form $R(w) = \tilde{R}(Xw)$, we can eliminate w from (6) to get $R(w) = \tilde{R}(X(\sum_g v_g)) = \tilde{R}(\tilde{X}\tilde{\mathbf{v}})$ and thus for the full objective : $\tilde{R}(\tilde{X}\tilde{\mathbf{v}}) + \lambda \sum_g \|\tilde{v}_g\|$. That way the vector $\tilde{\mathbf{v}} \in \mathbb{R}^{\sum |g|}$ can be directly estimated from $\tilde{X}$ with a classical group lasso for non-overlapping groups. We implemented the approach of (Meier et al., 2008) to estimate the group lasso in the expanded space. Note that (Roth & Fischer, 2008) provides a faster algorithm for the group Lasso. When there are many groups with important overlap however, an alternative implementation without explicit data duplication, *e.g.*, with a variational formulation similar to the one of (Rakotomamonjy et al., 2008) might be more scalable.

9. Experiments

9.1. Synthetic data: given overlapping groups

To assess the performance of our method when overlapping groups are given as a priori, we simulated data with $p = 82$ variables, covered by 10 groups of 10 variables with 2 variables of overlap between two successive groups: $\{1, \ldots, 10\}, \{9, \ldots, 18\}, \ldots, \{73, \ldots, 82\}$. We chose the support of w to be the union of groups 4 and 5 and sampled both the support weights 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, $\Omega^{\mathcal{G}}_{\text{overlap}}(.)$ can recover the right support, whereas by construction $\Omega^{\mathcal{G}}_{\text{group}}(\cdot)$ using the same groups would be unable to recover it.

The model is learned from n data points (x_i, y_i), with $y_i = w^\top x_i + \varepsilon$, $\varepsilon \sim \mathcal{N}(0, \sigma^2)$, $\sigma = |\mathbb{E}(Xw + b)|$. Using an ℓ_2 loss $R(w) = \|Y - Xw - b\|^2$, we learn models from 50 such training sets. On Figure 2, for each variable (on the vertical axis), we plot its frequency of selection in levels of gray as a function of the regularization parameter λ, both for the lasso penalty and $\Omega^{\mathcal{G}}_{\text{overlap}}(.)$.

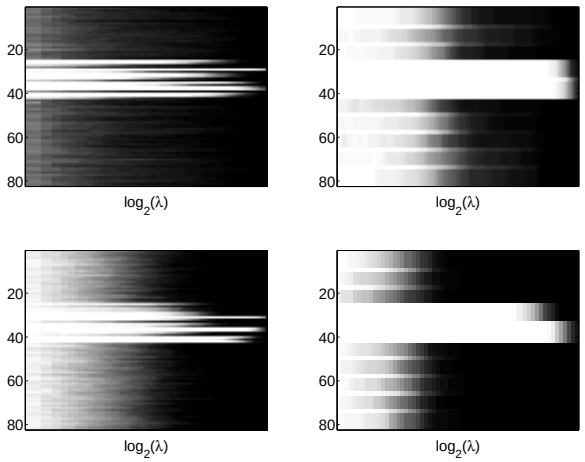

Figure 2. Frequency of selection of each variable with the lasso (left) and $\Omega^{\mathcal{G}}_{\text{overlap}}(.)$ (right) for $n = 50$ (top) and 100 (bottom).

For any choice of λ the lasso frequently misses some variables from the support, while $\Omega^{\mathcal{G}}_{\text{overlap}}(.)$ never misses any variable from the support for a large part of the regularization path. Besides, we observed that over the replicates, the lasso never selected the exact correct pattern for $n < 100$. For $n = 100$, the right pattern was selected with low frequency on a small part of the regularization path. $\Omega^{\mathcal{G}}_{\text{overlap}}(.)$ on the other hand selected it up to 92% of the times for $n = 50$ and more than 99% on more than one third of the path for $n = 100$. We tried the same experiment for various n and as long as n was too small for the lasso to recover the right support, the group regularization always helped.

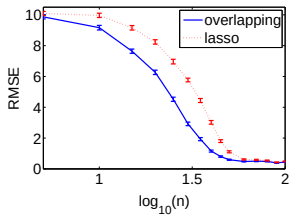

Figure 3. Root mean squared error of overlapped group lasso and lasso as a function of the number of training points.

Figure 3 shows the root mean squared error of both methods for various n. For both methods, the full regularization path is computed and tested on three replicates of n training and 100 testing points. The best average parameter is selected and used to train and test a model on a fourth replicate. On a large range of n, $\Omega^{\mathcal{G}}_{\text{overlap}}(.)$, not only helps to recover the right pattern, improves the regression performance. A possible explanation is that if several variables from the support are correlated in the design matrix X, the lasso selects one and is less robust than $\Omega^{\mathcal{G}}_{\text{overlap}}(.)$ which uses all the variables. Note that when enough training points become available (last point on Figure 3), Figure 2 shows that the selected model is generally better but still not correct whereas $\Omega^{\mathcal{G}}_{\text{overlap}}(.)$ selects the right model, even if it does not give much lower error anymore.

9.2. Synthetic data: given linear graph structure

We now consider that the prior given on the variables is a graph structure and that we are interested by solutions which are connected components on this graph. As a first simple illustration, we consider a chain. We use $w \in \mathbb{R}^p$, $p = 100$, $\text{supp}(w) = [20, 40]$. The nodes of the graph are the variables w_i, the edges are all the pairs $(w_i, w_{i+1}), i = 1, \ldots, n$. The model's weights, offset and the 50 training examples (x, y) are drawn using the same protocol as in the previous experiment. We take for the groups all the sub-chains of length k. We present the results for various choices of k and compare to the lasso ($k = 1$).

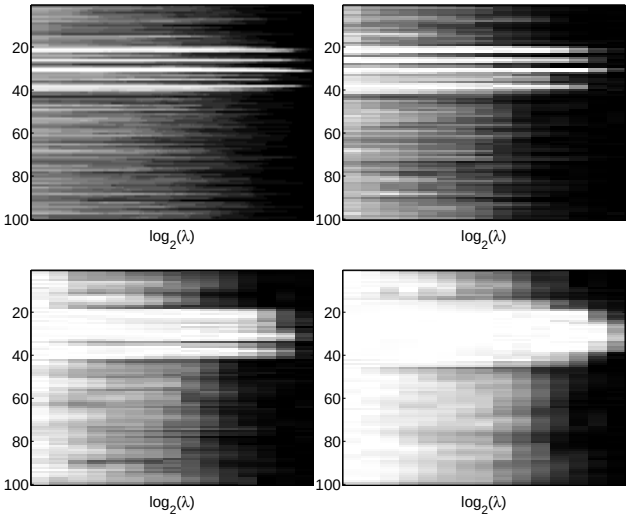

Figure 4. Variable selection frequency with $\Omega^{\mathcal{G}}_{\text{overlap}}(.)$ using the chains of length k (left) as groups, for $k = 1, 2, 4, 8$.

Figure 4 shows the frequency of each variable selection over 20 replications. Here again, using a group prior helps the pattern recovery. We also observe as expected that the choice of k plays a role in the improvement.

9.3. Synthetic data: given non-linear graph structure

Here we consider the same setting as in the linear case, except that instead of a chain we are given a grid structure on the variables. Each node is connected to the 4 nodes above, below, left and right. The support is a 20-variable region in the center of the grid, x-axis 4 to 7, y-axis 4 to 8. As groups, we use all the 4-cycles, which is a natural prior given the graph topology and the expected pattern.

Figure 5 shows the variable selection frequency of each

variable for both methods at a fixed λ (chosen in both cases to give the best behavior). $\Omega^{\mathcal{G}}_{\text{overlap}}(.)$ seems to generally give better selection performances than lasso.

Besides, we observed that on each run, variables incorrectly selected where always unions of groups whereas the lasso selected disconnected variables on the graph. We made the same observation for the linear graph case. This is an expected property of our method, and implies that even if variables which are not in the model are selected, they enter the model as large connected components, whereas the false positive of the lasso are more randomly distributed on the graph, often as isolated variables. This is an interesting property for real applications because it may then be easier to discard manually a few large connected components of false positives, than many isolated variables (assuming of course that the right variables are selected as well).

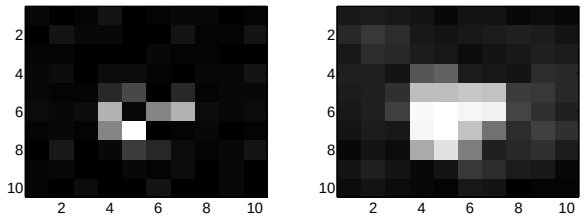

Figure 5. Grid view of the variable selection frequencies with the graph setting. Left: lasso, right: $\Omega^{\mathcal{G}}_{\text{overlap}}(.)$ using 4-cycles as groups. $n = 30$ training points, λ is arbitrarily fixed.

9.4. 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. For example, one may want to analyse microarrays in terms of biologically meaningful gene sets. In most such analysis, genes discriminating the classes (*e.g.* tumors leading to metastasis versus non-metastasis) are selected in a first step, then enrichment analysis is performed by looking for gene sets in which selected genes are overrepresented (Subramanian et al., 2005). Several organizations of the 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.

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 3510 genes which are in at least one pathway. Since the dataset is very unbalanced, we balance it by using 3 replicates of each metastasis patient (keeping all duplicates in the same fold during cross-validation).

We estimate by 3-fold cross validation the accuracy of a logistic regression with ℓ_1 and $\Omega^{\mathcal{G}}_{\text{overlap}}(.)$ penalties, using the pathways as groups. As a pre-processing, we keep the 300 genes most correlated with the output (on each training set). λ is selected by cross validation on each training set.

Table 1. Classification error, number and proportion of pathways selected by the ℓ_1 and $\Omega^{\mathcal{G}}_{\text{overlap}}(.)$ on the 3 folds.

METHOD	ℓ_1	$\Omega^{\mathcal{G}}_{\text{OVERLAP}}(.)$
ERROR	0.38 ± 0.04	0.36 ± 0.03
♯ PATH.	$148, 58, 183$	$6, 5, 78$
PROP. PATH.	$0.32, 0.14, 0.41$	$0.01, 0.01, 0.17$

Table 1 shows the results of both methods. Using $\Omega^{\mathcal{G}}_{\text{overlap}}(.)$ instead of the ℓ_1 penalty leads to a slight improvement in the prediction performances, and much sparser solutions at the pathway level, which makes the selected model easier to interpret.

9.5. Breast cancer data: graph analysis

Another important application in 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 connected components on a graph carrying biological information 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 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 7, taking all the edges on the network as the groups, 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 preprocessing as for the pathway analysis.

Table 2 shows the results of the logistic regression with ℓ_1 and $\Omega^{\mathcal{G}}_{\text{overlap}}(.)$. Here again, both methods give similar performances, with a slight advantage for $\Omega^{\mathcal{G}}_{\text{overlap}}(.)$. On the other hand, while the ℓ_1 mostly selects disconnected variables on the graph, $\Omega^{\mathcal{G}}_{\text{overlap}}(.)$ tends to select variables which are grouped into larger connected components on the graph. This would make the interpretation and the search for new drug targets easier.

10. Discussion

We have presented a generalization of the group lasso penalty, which leads to sparse models with sparsity patterns that are unions of pre-defined groups of covariates,

Table 2. Classification error and average size of the connected components selected by the ℓ_1 and $\Omega^{\mathcal{G}}_{\text{overlap}}(.)$ on the 3 folds.

METHOD	ℓ_1	$\Omega^{\mathcal{G}}_{\text{OVERLAP}}(.)$
ERROR	0.39 ± 0.04	0.36 ± 0.01
AV. SIZE C.C.	$1.1, 1, 1.0$	$1.3, 1.4, 1.2$

or, given a graph of covariates, groups of connected covariates in the graph. We obtained promising results on both simulated and real data.

From a theoretical point of view, we gave both sufficient and necessary conditions for the correct recovery of the same union of groups as in the decomposition induced by $\Omega^{\mathcal{G}}_{\text{overlap}}(\cdot)$ on the true optimal parameter vector. It still remains to characterize when the latter decomposition has the smallest number of groups. The situation where several decompositions exist should be analyzed. Also, the construction of an adaptive version of the Group Lasso with overlap that could possibly generalize the scheme proposed by (Bach, 2008) would be of interest.

From a practical point of view, although algorithms for the standard group Lasso can be used to implement $\Omega^{\mathcal{G}}_{\text{overlap}}(\cdot)$, more dedicated and scalable algorithms could be designed for cases with large overlaps.

Future work should compare more systematically $\Omega^{\mathcal{G}}_{\text{overlap}}(\cdot)$ and $\Omega^{\mathcal{G}}_{\text{group}}(\cdot)$ empirically and theoretically.

Acknowledgments

This work was supported by ANR grant ANR-07-BLAN-0311 and the France-Berkeley fund. The authors thank Bin Yu and Michael Jordan for useful discussions.

References

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

Bach, F. (2009). Exploring large feature spaces with hierarchical multiple kernel learning. *Adv. Neural. Inform. Process Syst.*.

Chen, S. S., Donoho, D. L., & Saunders, M. (1998). Atomic decomposition by basis pursuit. *SIAM J. Sci. Comput.*, *20*, 33–61.

Chuang, H.-Y., Lee, E., Liu, Y.-T., Lee, D., & Ideker, T. (2007). Network-based classification of breast cancer metastasis. *Mol. Syst. Biol.*, *3*, 140.

Fu, W., & Knight, K. (2000). Asymptotics for Lasso-type estimators. *Ann. Stat.*, *28*, 1356–1378.

Jenatton, R., Audibert, J.-Y., & Bach, F. (2009). *Structured Variable Selection with Sparsity-Inducing Norms*. INRIA - Ecole Normale Supérieure de Paris.

Kumagai, S. (1980). An implicit function theorem: Comment. *Journal of Optimization Theory and Applications*, *31*, 285–288.

Meier, L., van de Geer, S., & Bühlmann, P. (2008). The group lasso for logistic regression. *Journal Of The Royal Statistical Society Series B*, *70*, 53–71.

Obozinski, G., Taskar, B., & Jordan, M. (2009). Joint covariate selection and joint subspace selection for multiple classification problems. *Statistics and Computing*. To appear.

Rakotomamonjy, A., Bach, F., Canu, S., & Grandvalet, Y. (2008). SimpleMKL. *J. Mach. Learn. Res.*, *9*, 2491–2521.

Rapaport, F., Zynoviev, A., Dutreix, M., Barillot, E., & Vert, J.-P. (2007). Classification of microarray data using gene networks. *BMC Bioinformatics*, *8*, 35.

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

Roth, V., & Fischer, B. (2008). The group-lasso for generalized linear models: uniqueness of solutions and efficient algorithms. *ICML* (pp. 848–855).

Subramanian, A., et al., (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. USA*, *102*, 15545–15550.

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

Van de Vijver, M. J., et al., (2002). A gene-expression signature as a predictor of survival in breast cancer. *N. Engl. J. Med.*, *347*, 1999–2009.

Wainwright, M. J. (2006). *Sharp thresholds for high-dimensional and noisy recovery of sparsity* (Technical Report 709). UC Berkeley, Department of Statistics.

Yuan, M., & Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. Ser. B*, *68*, 49–67.

Zhao, P., Rocha, G., & Yu, B. (2009). Grouped and hierarchical model selection through composite absolute penalties. *Ann. Stat.* To appear.

Zhao, P., & Yu, B. (2006). On model selection consistency of lasso. *J. Mach. Learn. Res.*, *7*, 2541.