The degrees of freedom of the group Lasso for a general design

Samuel Vaiter and Gabriel Peyré
CEREMADE, CNRS-U. Paris-Dauphine
Place du Maréchal De Lattre De Tassigny,
75775 Paris Cedex 16, France.

Jalal M. Fadili
GREYC, CNRS-ENSICAEN-U. Caen
6, Bd du Maréchal Juin
14050 Caen Cedex, France.

Charles Deledalle and Charles Dossal
IMB, CNRS-U. Bordeaux 1
351 cours de la Libration
33405 Talence cedex, France.

***Abstract*—In this paper, we are concerned with regression problems where covariates can be grouped in nonoverlapping blocks, from which a few are active. In such a situation, the group Lasso is an attractive method for variable selection since it promotes sparsity of the groups. We study the sensitivity of any group Lasso solution to the observations and provide its precise local parameterization. When the noise is Gaussian, this allows us to derive an unbiased estimator of the degrees of freedom of the group Lasso. This result holds true for any fixed design, no matter whether it is under- or overdetermined. Our results specialize to those of [1], [2] for blocks of size one, i.e. ℓ^1 norm. These results allow objective choice of the regularisation parameter through e.g. the SURE.**

I. Group Lasso and Degrees of Freedom

Consider the linear regression problem $Y = \mathrm{X}\beta_0 + \varepsilon$, where Y is the real n-dimensional response vector, $\beta_0 \in \mathbb{R}^p$ is the unknown vector of regression coefficients to be estimated, $\mathrm{X} \in \mathbb{R}^{n\times p}$ is the design matrix whose columns are the p covariate vectors, and ε is the error term. In this paper, we do not make any specific assumption on n with respect to p.

Let $\mathcal{B}$ be a disjoint union of the set of indices i.e. $\bigcup_{b\in\mathcal{B}} = \{1,\dots,p\}$ such that $b, b' \in \mathcal{B}, b \cap b' = \emptyset$. For $\beta \in \mathbb{R}^p$, for each $b \in \mathcal{B}$, $\beta_b = (\beta_i)_{i\in b}$ is a subvector of β whose entries are indexed by the block b, and $|b|$ is the cardinality of b. The group Lasso amounts to solving

$$\widehat{\beta}(y) \in \operatorname*{Argmin}_{\beta\in\mathbb{R}^p} \frac{1}{2}\|y - \mathrm{X}\beta\|^2 + \lambda \sum_{b\in\mathcal{B}} \|\beta_b\|, \qquad (\mathcal{P}_\lambda(y))$$

from an observation $y \in \mathbb{R}^n$ of the regression model, where $\lambda > 0$ is the regularization parameter and $\|\cdot\|$ is the ℓ^2-norm.

Let $y \mapsto \widehat{\mu}(y) = \mathrm{X}\widehat{\beta}(y)$ be the response or the prediction associated to $\widehat{\beta}(y)$, and let $\mu_0 = \mathrm{X}\beta_0$. We recall that $\widehat{\mu}(y)$ is always uniquely defined, although $\widehat{\beta}(y)$ may not as is the case when X is a rank-deficient or underdetermined design matrix. Suppose that $\varepsilon \sim \mathcal{N}(0, \sigma^2\mathrm{Id}_n)$. Following [3], the DOF is given by $df = \sum_{i=1}^n \frac{\mathrm{cov}(Y_i, \widehat{\mu}_i(Y))]}{\sigma^2}$. The well-known Stein's lemma asserts that, if $y \mapsto \widehat{\mu}(y)$ is a weakly differentiable function with an essentially bounded gradient, then an unbiased estimator of df is $\widehat{df}(Y) = \mathrm{div}\widehat{\mu}(Y) = \mathrm{tr}(\partial\widehat{\mu}(Y))$ and $\mathbb{E}_\varepsilon(\widehat{df}(Y)) = df$, where $\partial\widehat{\mu}(\cdot)$ is the Jacobian of $\widehat{\mu}(\cdot)$ w.r.t. to its argument.

In the sequel, we define the $\mathcal{B}$-support $\mathrm{supp}_\mathcal{B}(\beta)$ of $\beta \in \mathbb{R}^p$ as $\mathrm{supp}_\mathcal{B}(\beta) = \{b \in \mathcal{B} \setminus \|\beta_b\| \neq 0\}$. The size of $\mathrm{supp}_\mathcal{B}(\beta)$ is $|\mathrm{supp}_\mathcal{B}(\beta)| = \sum_{b\in\mathcal{B}} |b|$. The set of all $\mathcal{B}$-supports is denoted $\mathcal{I}$, and X_I, where I is a $\mathcal{B}$-support, is the matrix formed by the columns X_i where i is an element of $b \in I$. We also introduce the following block-diagonal operators

$$\delta_\beta : v \in \mathbb{R}^{|I|} \mapsto (v_b/\|\beta_b\|)_{b\in I} \in \mathbb{R}^{|I|}$$
$$\text{and} \quad P_\beta : v \in \mathbb{R}^{|I|} \mapsto (\mathrm{Proj}_{\beta_b^\perp}(v_b))_{b\in I} \in \mathbb{R}^{|I|},$$

where $\mathrm{Proj}_{\beta_b^\perp} = \mathrm{Id} - \beta_b\beta_b^\mathrm{T}$ is the orthogonal projector on $\beta_b^\perp$.

II. Main Contributions

The first difficulty we need to overcome when X is not full column rank is that $y \mapsto \widehat{\beta}(y)$ is set-valued. Toward this goal, we are led to impose the following assumption on X with respect to the block structure.

Assumption $(\mathbf{A}(\beta))$: Given a vector $\beta \in \mathbb{R}^p$ of $\mathcal{B}$-support I, we assume that the finite subset of vectors $\{\mathrm{X}_b\beta_b \setminus b \in I\}$ is linearly independent.

It is important to notice that $(\mathbf{A}(\beta))$ is weaker than imposing that X_I is full column rank, which is standard when analyzing the Lasso. The two assumptions coincide for the Lasso, i.e. $|b| = 1, \forall b \in I$.

Definition 1: Let $\lambda > 0$. The *transition space* $\mathcal{H}$ is defined as

$$\mathcal{H} = \bigcup_{I\subset\mathcal{I}} \bigcup_{b\notin I} \mathcal{H}_{I,b}, \quad \text{where} \quad \mathcal{H}_{I,b} = \mathrm{bd}(\pi(\mathcal{A}_{I,b})),$$

where we have denoted

$$\pi : \mathbb{R}^n \times \mathbb{R}^{I,*} \times \mathbb{R}^{I,*} \to \mathbb{R}^n \quad \text{where} \quad \mathbb{R}^{I,*} = \prod_{b\in I} (\mathbb{R}^{|b|} \setminus \{0\})$$

the canonical projection on $\mathbb{R}^n$ (with respect to the first component), $\mathrm{bd}\, C$ is the boundary of the set C, and

$$\mathcal{A}_{I,b} = \Big\{ (y, \beta_I, v_I) \in \mathbb{R}^n \times \mathbb{R}^{I,*} \times \mathbb{R}^{I,*} \setminus \|\mathrm{X}_b^\mathrm{T}(y - \mathrm{X}_I\beta_I)\| = \lambda,$$
$$\mathrm{X}_I^\mathrm{T}(\mathrm{X}_I\beta_I - y) + \lambda v_I = 0, \forall g \in I, v_g = \frac{\beta_g}{\|\beta_g\|} \Big\}.$$

We are now equipped to state our main sensitivity analysis result.

Theorem 1: Let $\lambda > 0$. Let $y \notin \mathcal{H}$, and $\widehat{\beta}(y)$ a solution of $(\mathcal{P}_\lambda(y))$. Let $I = \mathrm{supp}_\mathcal{B}(\widehat{\beta}(y))$ be the $\mathcal{B}$-support of $\widehat{\beta}(y)$ such that $(\mathbf{A}(\widehat{\beta}(y)))$ holds. Then, there exists an open neighborhood of y $\mathcal{O} \subset \mathbb{R}^n$, and a mapping $\widetilde{\beta} : \mathcal{O} \to \mathbb{R}^p$ such that

1) For all $\bar{y} \in \mathcal{O}$, $\widetilde{\beta}(\bar{y})$ is a solution of $(\mathcal{P}_\lambda(\bar{y}))$, and $\widetilde{\beta}(y) = \widehat{\beta}(y)$.
2) the $\mathcal{B}$-support of $\widetilde{\beta}(\bar{y})$ is constant on $\mathcal{O}$.
3) the mapping $\widetilde{\beta}$ is $\mathcal{C}^1(\mathcal{O})$ and its Jacobian is such that $\forall \bar{y} \in \mathcal{O}$,

$$\partial\widetilde{\beta}_{I^c}(\bar{y}) = 0 \quad \text{and} \quad \partial\widetilde{\beta}_I(\bar{y}) = d(y,\lambda) \qquad (1)$$
$$\text{where} \quad d(y,\lambda) = \left(\mathrm{X}_I^\mathrm{T}\mathrm{X}_I + \lambda\delta_{\widehat{\beta}(y)} \circ P_{\widehat{\beta}(y)}\right)^{-1}\mathrm{X}_I^\mathrm{T} \qquad (2)$$
$$\text{and} \quad I^c = \{b \in \mathcal{B} \setminus b \notin I\}. \qquad (3)$$

The next theorem provides a closed-form expression of the local variations of $y \mapsto \widehat{\mu}(y)$. In turn, when $\varepsilon \sim \mathcal{N}(0, \sigma^2\mathrm{Id})$, this will yield an unbiased estimator of the degrees of freedom and of the prediction risk of the group Lasso.

Theorem 2: Let $\lambda > 0$. For all $y \notin \mathcal{H}$, there exists a solution $\widehat{\beta}(y)$ of $(\mathcal{P}_\lambda(y))$ with $\mathcal{B}$-support $I = \mathrm{supp}_\mathcal{B}(\widehat{\beta}(y))$ such that $(\mathbf{A}(\widehat{\beta}(y)))$ is fulfilled. The mapping $y \mapsto \widehat{\mu}(y) = \mathrm{X}\widehat{\beta}(y)$ is $\mathcal{C}^1(\mathbb{R}^n \setminus \mathcal{H})$ and,

$$\mathrm{div}(\widehat{\mu}(y)) = \mathrm{tr}(\mathrm{X}_I d(y,\lambda)) \qquad (4)$$

where $\widehat{\beta}(y)$ is such that $(\mathbf{A}(\widehat{\beta}(y)))$ holds. Moreover, The set $\mathcal{H}$ has Lebesgue measure zero. If $Y = \mathrm{X}\beta_0 + \varepsilon$ where $\varepsilon \sim \mathcal{N}(0, \sigma^2\mathrm{Id}_n)$, then $\mathrm{tr}(\mathrm{X}_I d(Y,\lambda))$ is an unbiased estimate of the DOF of the group Lasso.

References

[1] H. Zou, T. Hastie, and R. Tibshirani, "On the "degrees of freedom" of the Lasso," *The Annals of Statistics*, vol. 35, no. 5, pp. 2173–2192, 2007.

[2] C. Dossal, M. Kachour, J. Fadili, G. Peyré, and C. Chesneau, "The degrees of freedom of penalized ℓ_1 minimization," *to appear in Statistica Sinica*, 2012. [Online]. Available: http://hal.archives-ouvertes.fr/hal-00638417

[3] B. Efron, "How biased is the apparent error rate of a prediction rule?" *Journal of the American Statistical Association*, vol. 81, no. 394, pp. 461–470, 1986.