The degrees of freedom of the Group Lasso for a General Design

Samuel Vaiter, Charles Deledalle, Gabriel Peyré, Jalal M. Fadili, Charles Dossal

The degrees of freedom of the Group Lasso for a General Design

Samuel Vaiter · Charles Deledalle · Gabriel Peyré · Jalal Fadili · Charles Dossal

Abstract In this paper, we are concerned with regression problems where covariates can be grouped in nonoverlapping blocks, and where only a few of them are assumed to be 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. With these results at hand, various model selection criteria, such as the Stein Unbiased Risk Estimator (SURE), are readily available which can provide an objectively guided choice of the optimal group Lasso fit.

Keywords Group Lasso · Degrees of freedom · Sparsity · Model selection criteria

Samuel Vaiter, Gabriel Peyré
CEREMADE, CNRS, Université Paris-Dauphine, Place du Maréchal De Lattre De Tassigny, 75775 Paris Cedex 16, France
E-mail: {samuel.vaiter,gabriel.peyre}@ceremade.dauphine.fr

Charles Deledalle, Charles Dossal
IMB, CNRS, Université Bordeaux 1, 351, Cours de la libération, 33405 Talence Cedex, France
E-mail: {charles.deledalle,charles.dossal}@math.u-bordeaux1.fr

Jalal Fadili
GREYC, CNRS-ENSICAEN-Université de Caen, 6, Bd du Maréchal Juin, 14050 Caen Cedex, France
E-mail: Jalal.Fadili@greyc.ensicaen.fr

1 Introduction

1.1 Group Lasso

Consider the linear regression problem

$$y = X\beta_0 + \varepsilon, \tag{1}$$

where $y \in \mathbb{R}^n$ is the response vector, $\beta_0 \in \mathbb{R}^p$ is the unknown vector of regression coefficients to be estimated, $X \in \mathbb{R}^{n\times p}$ 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 the number of observations n with respect to the number of predictors p. Recall that when $n < p$, (1) is an underdetermined linear regression model, whereas when $n \geqslant p$ and all the columns of X are linearly independent, it is overdetermined.

Regularization is now a central theme in many fields including statistics, machine learning and inverse problems. It allows to reduce the space of candidate solutions by imposing some prior structure on the object to be estimated. This regularization ranges from squared Euclidean or Hilbertian norms (Tikhonov and Arsenin 1997), to non-Hilbertian norms that have sparked considerable interest in the recent years. Of particular interest are sparsity-inducing regularizations such as the ℓ^1 norm which is an intensively active area of research, e.g. (Tibshirani 1996; Osborne et al 2000; Donoho 2006; Candès and Plan 2009; Bickel et al 2009); see (Bühlmann and van de Geer 2011) for a comprehensive review. When the covariates are assumed to be clustered in a few active groups/blocks, the group Lasso has been advocated since it promotes sparsity of the groups, i.e. it drives all the coefficients in one group to zero together hence leading to group selection, see (Bakin 1999; Yuan and Lin 2006; Bach 2008; Wei and Huang 2010) to cite a few.

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

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

where $\lambda > 0$ is the regularization parameter and $\|\cdot\|$ is the (Euclidean) ℓ^2-norm. By coercivity of the penalty norm, the set of minimizers of $(\mathcal{P}_\lambda(y))$ is a nonempty convex compact set. Note that the Lasso is a particular instance of $(\mathcal{P}_\lambda(y))$ that is recovered when each block b is of size 1.

1.2 Degrees of Freedom

We focus in this paper on sensitivity analysis of any solution to $(\mathcal{P}_\lambda(y))$ with respect to the observations y and the regularization parameter λ. This turns

out to be a central ingredient to compute an estimator of the degrees of freedom (DOF) of the group Lasso response. The DOF is usually used to quantify the complexity of a statistical modeling procedure (Efron 1986).

More precisely, let $\widehat{\mu}(y) = X\widehat{\beta}(y)$ be the response or the prediction associated to an estimator $\widehat{\beta}(y)$ of β_0, and let $\mu_0 = X\beta_0$. We recall that $\widehat{\mu}(y)$ is always uniquely defined (see Lemma 2), although $\widehat{\beta}(y)$ may not as is the case when X is a rank-deficient or underdetermined design matrix. Suppose that ε is an additive white Gaussian noise $\varepsilon \sim \mathcal{N}(0, \sigma^2 \mathrm{Id}_n)$. Following (Efron 1986), 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 $\widehat{\mu}(y)$ is a weakly differentiable function for which

$$\mathbb{E}_\varepsilon \left(\left| \frac{\partial}{\partial y_i} \widehat{\mu}_i(y) \right| \right) < \infty \ ,$$

then its divergence is an unbiased estimator of its DOF, i.e.

$$\widehat{df} = \mathrm{div}\widehat{\mu}(y) = \mathrm{tr}(\partial_y \widehat{\mu}(y)) \quad \text{and} \quad \mathbb{E}_\varepsilon(\widehat{df}) = df \ ,$$

where $\partial_y \widehat{\mu}(y)$ is the Jacobian of $\widehat{\mu}(y)$. It is well known that in Gaussian regression problems, an unbiased estimator of the DOF allows to get an unbiased of the prediction risk estimation $\mathbb{E}_\varepsilon \|\widehat{\mu}(y) - \mu_0\|^2$ through e.g. the Mallow's C_p (Mallows 1973), the AIC (Akaike 1973) or the SURE (Stein Unbiased Risk Estimate, Stein 1981). These quantities can serve as model selection criteria to assess the accuracy of a candidate model.

1.3 Contributions

This paper establishes a general result (Theorem 1) on local parameterization of any solution to the group Lasso ($\mathcal{P}_\lambda(y)$) as a function of the observation vector y. This local behavior result does not need X to be full column rank. With such a result at hand, we derive an expression of the divergence of the group Lasso response. Using tools from semialgebraic geometry, we prove that this divergence formula is valid Lebesgue-almost everywhere (Theorem 2), and thus, this formula is a provably unbiased estimate of the DOF (Theorem 3). In turn, this allows us to deduce an unbiased estimate of the prediction risk of the group Lasso through the SURE.

1.4 Relation to prior works

In the special case of standard Lasso with a linearly independent design, (Zou et al 2007) show that the number of nonzero coefficients is an unbiased estimate for the degrees of freedom. This work is generalized in (Dossal et al 2012) to any arbitrary design matrix. The DOF of the analysis sparse regularization

(a.k.a. generalized Lasso in statistics) is studied in (Tibshirani and Taylor 2012; Vaiter et al 2012b).

A formula of an estimate of the DOF for the group Lasso when the design is orthogonal within each group is conjectured in (Yuan and Lin 2006). Its unbiasedness is proved but only for an orthogonal design. (Kato 2009) studies the DOF of a general shrinkage estimator where the regression coefficients are constrained to a closed convex set C. This work extended that of (Meyer and Woodroofe 2000) which treats the case where C is a convex polyhedral cone. When X is full column rank, (Kato 2009) derived a divergence formula under a smoothness condition on the boundary of C, from which he obtained an unbiased estimator of the degrees of freedom. When specializing to the constrained version of the group Lasso, the author provided an unbiased estimate of the corresponding DOF under the same group-wise orthogonality assumption on X as (Yuan and Lin 2006). An estimate of the DOF for the group Lasso is also given by (Solo and Ulfarsson 2010) using heuristic derivations that are valid only when X is full column rank, though its unbiasedness is not proved.

In (Vaiter et al 2012a), we derived an estimator of the DOF of the group Lasso and proved its unbiasedness when X is full column rank, but without the orthogonality assumption required in (Yuan and Lin 2006; Kato 2009). In this paper, we remove the full column rank assumption, which enables us to tackle the much more challenging rank-deficient or underdetermined case where $p > n$.

1.5 Notations

We start by some notations used in the rest of the paper. We extend the notion of support, commonly used in sparsity by defining the $\mathcal{B}$-support $\mathrm{supp}_{\mathcal{B}}(\beta)$ of $\beta \in \mathbb{R}^n$ as

$$\mathrm{supp}_{\mathcal{B}}(\beta) = \{b \in \mathcal{B} \setminus \|\beta_b\| \neq 0\}\,.$$

The size of $\mathrm{supp}_{\mathcal{B}}(\beta)$ is defined as $|\mathrm{supp}_{\mathcal{B}}(\beta)| = \sum_{b\in\mathcal{B}} |b|$. The set of all $\mathcal{B}$-supports is denoted $\mathcal{I}$. We denote by X_I, where I is a $\mathcal{B}$-support, the matrix formed by the columns X_i where i is an element of $b \in I$. To lighten the notation in our derivations, we introduce the following block-diagonal operators

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

where $\mathrm{Proj}_{\beta_b^\perp} = \mathrm{Id} - \beta_b\beta_b^{\mathrm{T}}$ is the orthogonal projector on $\beta_b^\perp$. For any matrix A, A^{T} denotes its transpose.

1.6 Paper organization

The paper is organized as follows. Sensitivity analysis of the group Lasso solutions to perturbations of the observations is given in Section 2. Then we turn to

the degrees of freedom and unbiased prediction risk estimation in Section 3. The proofs are deferred to Section 4 awaiting inspection by the interested reader.

2 Local Behavior of the Group Lasso

The first difficulty we need to overcome when X is not full column rank is that $\widehat{\beta}(y)$ is not uniquely defined. 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 $\{X_b\beta_b \setminus b \in I\}$ 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. $|b| = 1, \forall b \in I$.

Let us now turn to sensitivity of the minimizers $\widehat{\beta}(y)$ of $(\mathcal{P}_\lambda(y))$ to perturbations of y. Toward this end, we will exploit the fact that $\widehat{\beta}(y)$ obeys an implicit parameterization. But as optimal solutions turns out to be not everywhere differentiable, we will concentrate on a local analysis where y is allowed to vary in a neighborhood where non-differentiability will not occur. This is why we need to introduce the following transition space $\mathcal{H}$.

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

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

where we have denoted

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

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

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

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 $\mathcal{O}$, i.e.*

$$\forall \bar{y} \in \mathcal{O}, \quad \mathrm{supp}_{\mathcal{B}}(\widetilde{\beta}(\bar{y})) = I,$$

3. *the mapping $\widetilde{\beta}$ is $\mathcal{C}^1(\mathcal{O})$ and its Jacobian is such that $\forall \bar{y} \in \mathcal{O}$,*

$$\partial_{\bar{y}} \widetilde{\beta}_{I^c}(\bar{y}) = 0 \quad \textit{and} \quad \partial_{\bar{y}} \widetilde{\beta}_I(\bar{y}) = d(y, \lambda) \tag{2}$$

$$\textit{where} \quad d(y, \lambda) = \left(X_I^{\mathrm{T}} X_I + \lambda \delta_{\widehat{\beta}(y)} \circ P_{\widehat{\beta}(y)} \right)^{-1} X_I^{\mathrm{T}} \tag{3}$$

$$\textit{and} \quad I^c = \{ b \in \mathcal{B} \setminus b \notin I \}. \tag{4}$$

3 Degrees of freedom and Risk Estimation

As remarked earlier and stated formally in Lemma 2, all solutions of the Lasso share the same image under X, hence allowing us to denote the prediction $\widehat{\mu}(y)$ without ambiguity as a single-valued mapping. The next theorem provides a closed-form expression of the local variations of $\widehat{\mu}(y)$ with respect to the observation y. In turn, 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. Moreover, The mapping $y \mapsto \widehat{\mu}(y) = X\widehat{\beta}(y)$ is $\mathcal{C}^1(\mathbb{R}^n \setminus \mathcal{H})$ and,*

$$\mathrm{div}(\widehat{\mu}(y)) = \mathrm{tr}(X_I d(y, \lambda)) \tag{5}$$

where $\widehat{\beta}(y)$ is such that $(\mathbf{A}(\widehat{\beta}(y)))$ holds.

Theorem 3 *Let $\lambda > 0$. Assume $y = X\beta_0 + \varepsilon$ where $\varepsilon \sim \mathcal{N}(0, \sigma^2 \mathrm{Id}_n)$. The set $\mathcal{H}$ has Lebesgue measure zero, and therefore (5) is an unbiased estimate of the DOF of the group Lasso. Moreover, an unbiased estimator of the prediction risk $\mathbb{E}_\varepsilon \|\widehat{\mu}(y) - \mu_0\|^2$ is given by the* SURE *formula*

$$\mathrm{SURE}(\widehat{\mu}(y)) = \|y - \widehat{\mu}(y)\|^2 - n\sigma^2 + 2\sigma^2 \,\mathrm{tr}(X_I d(y, \lambda)) \ . \tag{6}$$

Although not given here explicitly, Theorem 3 can be straightforwardly extended to unbiasedly of measures of the risk, including the *projection* risk, or the *estimation* risk (in the full rank case) through the Generalized Stein Unbiased Risk Estimator as proposed in (Vaiter et al 2012b).

An immediate corollary of Theorem 3 is obtained when X is orthogonal, and without loss of generality $X = \mathrm{Id}_n$, i.e. $\widehat{\mu}(y)$ is the block soft thresholding estimator. We then recover the expression found by (Yuan and Lin 2006).

Corollary 1 *If $X = \mathrm{Id}_n$, then*

$$\widehat{df} = |I| - \lambda \sum_{b \in I} \frac{|b| - 1}{\|y_b\|}$$

where $I = \bigcup \{b \in \mathcal{B} \setminus \|y_b\| > \lambda\}$. Moreover, the SURE *is given by*

$$\mathrm{SURE}(\widehat{\mu}(y)) = -n\sigma^2 + (2\sigma^2 + \lambda^2)|I| + \sum_{b \notin I} \|y_b\|^2 - 2\sigma^2\lambda \sum_{b \in I} \frac{|b| - 1}{\|y_b\|} \ .$$

We finally quantify the (relative) reliability of the SURE by computing the expected squared-error between $\mathrm{SURE}(\widehat{\mu}(y))$ and the true squared-error

$$\mathrm{SE}(\widehat{\mu}(y)) = \|\widehat{\mu}(y) - \mu_0\|^2 \ .$$

Proposition 1 *Under the assumptions of Theorem 3, the relative reliability obeys*

$$\mathbb{E}_w \left[\frac{(\mathrm{SURE}(\widehat{\mu}(y)) - \mathrm{SE}(\widehat{\mu}(y)))^2}{n^2\sigma^4} \right] \leqslant \frac{18 + 4\mathbb{E}_w\left(\|U_I\|^2\right)}{n} + \frac{8\|\mu_0\|^2}{n^2\sigma^2} \ .$$

$$\textit{where} \quad U_I = X_I^{\mathrm{T}} X_I \big(X_I^{\mathrm{T}} X_I + \lambda \delta_{\widehat{\beta}(y)} \circ P_{\widehat{\beta}(y)}\big)^{-1}.$$

In particular, it decays at the rate $O(1/n)$ if $\mathbb{E}_w\left(\|U_I\|^2\right) = O(1)$.

Note that when $X = \mathrm{Id}_n$, the proof of Corollary 1 yields that $\|U_I\| = 1$.

4 Proofs

This section details the proofs of our results. For a vector β whose $\mathcal{B}$-support is I, we introduce the following normalization operator

$$\mathcal{N}(\beta_I) = v_I \quad \text{where} \quad \forall b \in I, v_b = \frac{\beta_b}{\|\beta_b\|}.$$

4.1 Preparatory lemmata

By standard arguments of convex analysis and using the subdifferential of the group Lasso $\ell^1 - \ell^2$ penalty, the following lemma gives the first-order sufficient and necessary optimality condition of a minimizer of $(\mathcal{P}_\lambda(y))$; see e.g. Bach (2008).

Lemma 1 *A vector $\beta^\star \in \mathbb{R}^p$ is a solution of $(\mathcal{P}_\lambda(y))$ if, and only if the following holds*

1. *On the $\mathcal{B}$-support $I = \mathrm{supp}_{\mathcal{B}}(\beta^\star)$,*

$$X_I^{\mathrm{T}}(y - X_I\beta_I^\star) = \lambda \mathcal{N}(\beta_I^\star).$$

2. For all $b \in \mathcal{B}$ such that $b \notin I$, one has

$$\|X_b^{\mathrm{T}}(y - X_I\beta_I^\star)\| \leqslant \lambda.$$

We now show that all solutions of $(\mathcal{P}_\lambda(y))$ share the same image under the action of X, which in turn implies that the prediction/response vector $\widehat{\mu}$ is a single-valued mapping of y.

Lemma 2 *If β^0 and β^1 are two solutions of $(\mathcal{P}_\lambda(y))$, then $X\beta^0 = X\beta^1$.*

Proof Let β^0, β^1 be two solutions of $(\mathcal{P}_\lambda(y))$ such that $X\beta^0 \neq X\beta^1$. Take any convex combination $\beta^\rho = (1-\rho)\beta^0 + \rho\beta^1$, $\rho \in]0,1[$. Strict convexity of $u \mapsto \|y-u\|^2$ implies that the Jensen inequality is strict, i.e.

$$\frac{1}{2}\|y - X\beta^\rho\|^2 < \frac{1-\rho}{2}\|y - X\beta^0\|^2 + \frac{\rho}{2}\|y - X\beta^1\|^2 \ .$$

Denote the $\ell^1 - \ell^2$ norm $\|\beta\|_{\mathcal{B}} = \sum_{b \in B b} \|\beta_b\|$. Jensen's inequality applied to $\|\cdot\|_{\mathcal{B}}$ gives

$$\|\beta^\rho\|_{\mathcal{B}} \leqslant (1-\rho)\|\beta^0\|_{\mathcal{B}} + \rho\|\beta^1\|_{\mathcal{B}} \ .$$

Summing these two inequalities we arrive at $\frac{1}{2}\|y - X\beta^\rho\|^2 + \lambda\|\beta^\rho\|_{\mathcal{B}} < \frac{1}{2}\|y - X\beta^0\|^2 + \lambda\|\beta^0\|_{\mathcal{B}}$, a contradiction since β^0 is a minimizer of $(\mathcal{P}_\lambda(y))$.

4.2 Proof of Theorem 1

We first need the following lemma.

Lemma 3 *Let $\beta \in \mathbb{R}^p$ and $\lambda > 0$. Assume that $(\mathbf{A}(\beta))$ holds for I the $\mathcal{B}$-support of β. Then $X_I^{\mathrm{T}}X_I + \lambda\delta_\beta \circ P_\beta$ is invertible.*

Proof We prove that $X_I^{\mathrm{T}}X_I + \lambda\delta_\beta \circ P_\beta$ is actually symmetric definite positive. First observe that $X_I^{\mathrm{T}}X_I$ and $\delta_\beta \circ P_\beta$ are both symmetric semidefinite positive. Indeed, δ_β is diagonal (with strictly positive diagonal entries), and P_β is symmetric since it is a block-wise orthogonal projector, and we have

$$\langle x,\, \delta_\beta \circ P_\beta(x)\rangle = \sum_{b \in I} \frac{\|\mathrm{Proj}_{\beta_b^\perp}(x)\|^2}{\|\beta_b\|} \geqslant 0, \quad \forall x \in \mathbb{R}^{|I|} \ .$$

The inequality becomes an equality if and only if $x = \beta_I$, i.e. $\operatorname{Ker} \delta_\beta \circ P_\beta = \{\beta_I\}$.

It remains to show that $\operatorname{Ker} X_I^{\mathrm{T}}X_I \cap \operatorname{Ker} \delta_\beta \circ P_\beta = \{0\}$. Suppose that $\beta_I \in \operatorname{Ker} X_I^{\mathrm{T}}X_I$. This is equivalent to $\beta_I \in \operatorname{Ker} X_I$ since

$$\langle \beta_I,\, X_I^{\mathrm{T}}X_I\beta_I\rangle = \|X_I\beta_I\|^2 \ .$$

But this would mean that

$$X_I\beta_I = \sum_{b \in I} X_b\beta_b = 0$$

which is in contradiction with the linear independence assumption $(\mathbf{A}(\beta))$. □

Let $y \notin \mathcal{H}$. We define $I = \mathrm{supp}_{\mathcal{B}}(\widehat{\beta}(y))$ the $\mathcal{B}$-support of a solution $\widehat{\beta}(y)$ of $(\mathcal{P}_\lambda(y))$. We define the following mapping

$$\Gamma(\beta_I, y) = X_I^{\mathrm{T}}(X_I\beta_I - y) + \lambda\mathcal{N}(\beta_I).$$

Observe that the first statement of Lemma 1 is equivalent to $\Gamma(\widehat{\beta}_I(y), y) = 0$.

Any $\beta_I \in \mathbb{R}^{|I|}$ such that $\Gamma(\beta_I, y) = 0$ is solution of the problem

$$\min_{\beta_I \in \mathbb{R}^{|I|}} \frac{1}{2}\|y - X_I\beta_I\|^2 + \lambda \sum_{g \in I} \|\beta_g\| \ . \qquad (\mathcal{P}_\lambda(y)_I)$$

Our proof will be split in three steps. We first prove the first statement by showing that there exists a mapping $\bar{y} \mapsto \widetilde{\beta}(\bar{y})$ and an open neighborhood $\mathcal{O}$ of y such that every element $\bar{y}$ of $\mathcal{O}$ satisfies $\Gamma(\widetilde{\beta}_I(\bar{y}), \bar{y}) = 0$ and $\widetilde{\beta}_{I^c}(\bar{y}) = 0$. Then, we prove the second assertion that $\widetilde{\beta}(\bar{y})$ is a solution of $(\mathcal{P}_\lambda(\bar{y}))$ for $\bar{y} \in \mathcal{O}$. Finally, we obtain (2) from the implicit function theorem.

1. The Jacobian of Γ with respect to the first variable reads on $\mathbb{R}^{I,*} \times \mathbb{R}^n$

   $$\partial_1\Gamma(\beta_I, y) = X_I^{\mathrm{T}}X_I + \lambda\delta_{\beta_I} \circ P_{\beta_I}.$$

   The mapping $\partial_1\Gamma$ is invertible according to Lemma 3. Hence, using the implicit function theorem, there exists a neighborhood $\widetilde{\mathcal{O}}$ of y such that we can define a mapping $\widetilde{\beta}_I : \widetilde{\mathcal{O}} \to \mathbb{R}^{|I|}$ which is $\mathcal{C}^1(\widetilde{\mathcal{O}})$, and satisfies for $\bar{y} \in \widetilde{\mathcal{O}}$

   $$\Gamma(\widetilde{\beta}_I(\bar{y}), \bar{y}) = 0 \quad \text{and} \quad \widetilde{\beta}_I(y) = \widehat{\beta}_I(y).$$

   We then extend $\widetilde{\beta}_I$ on I^c as $\widetilde{\beta}_{I^c}(\bar{y}) = 0$, which defines a continuous mapping $\widetilde{\beta} : \widetilde{\mathcal{O}} \to \mathbb{R}^p$.
2. From the second minimality condition of Lemma 1, we have

   $$\forall b \notin I, \quad \|X_b^{\mathrm{T}}(y - X_I\widehat{\beta}_I(y))\| \leqslant \lambda.$$

   We define the two following sets

   $$J_{\mathrm{sat}} = \left\{ b \notin I \setminus \|X_b^{\mathrm{T}}(y - X_I\widehat{\beta}_I(y))\| = \lambda \right\},$$
   $$J_{\mathrm{nosat}} = \left\{ b \notin I \setminus \|X_b^{\mathrm{T}}(y - X_I\widehat{\beta}_I(y))\| < \lambda \right\},$$

   which forms a disjoint union of $I^c = J_{\mathrm{sat}} \cup J_{\mathrm{nosat}}$.
   a) By continuity of $\bar{y} \mapsto \widetilde{\beta}_I(\bar{y})$ and since $\widetilde{\beta}_I(y) = \widehat{\beta}_I(y)$, we can find a neighborhood $\mathcal{O}$ of y included in $\widetilde{\mathcal{O}}$ such that

   $$\forall\, \bar{y} \in \mathcal{O}, \forall\, b \in J_{\mathrm{nosat}}, \quad \|X_b^{\mathrm{T}}(\bar{y} - X_I\widetilde{\beta}_I(\bar{y}))\| \leqslant \lambda.$$

b) Consider now a block $b \in J_{\text{sat}}$. Observe that the vector $(y, \widehat{\beta}_I(y), \mathcal{N}(\widehat{\beta}_I(y)))$ is an element of $\mathcal{A}_{I,b}$. In particular $y \in \pi(\mathcal{A}_{I,b})$. Since by assumption $y \notin \mathcal{H}$, one has $y \notin \mathrm{bd}(\pi(\mathcal{A}_{I,b}))$. Hence, there exists an open ball $\mathbb{B}(y,\varepsilon)$ for some $\varepsilon > 0$ such that $\mathbb{B}(y,\varepsilon) \subset \pi(\mathcal{A}_{I,b})$. Notice that every element of $\bar{y} \in \mathbb{B}(y,\varepsilon)$ is such that there exists $(\bar{\beta}_I, \bar{v}_I) \in \mathbb{R}^{I,*} \times \mathbb{R}^{I,*}$ with

$$\begin{aligned} \|X_b^{\mathrm{T}}(\bar{y} - X_I\bar{\beta}_I)\| &= \lambda \\ X_I^{\mathrm{T}}(X_I\bar{\beta}_I - \bar{y}) + \lambda\bar{v}_I &= 0 \\ \bar{v}_I &= \mathcal{N}(\bar{\beta}_I)\ . \end{aligned}$$

Using a similar argument as in the proof of Lemma 2, it is easy to see that all solutions of $(\mathcal{P}_\lambda(y)_I)$ share the same image under X_I. Thus the vector $(\bar{y}, \widetilde{\beta}_I(\bar{y}), \mathcal{N}(\widetilde{\beta}_I(\bar{y})))$ is an element $\mathcal{A}_{I,b}$, and we conclude that

$$\forall \bar{y} \in \mathbb{B}(y,\varepsilon), \quad f_{I,b}(\bar{y}) = \|X_b^{\mathrm{T}}(\bar{y} - X_I\widetilde{\beta}_I(\bar{y}))\| = \lambda.$$

Hence, $f_{I,b}$ is locally constant around y on an open ball $\bar{\mathcal{O}}$.
Moreover, by definition of the mapping $\widetilde{\beta}_I$, one has for all $\bar{y} \in \mathcal{O} \cap \bar{\mathcal{O}}$

$$X_I^{\mathrm{T}}(y - X_I\widetilde{\beta}_I(\bar{y})) = \lambda\mathcal{N}(\widetilde{\beta}_I(\bar{y})) \text{ and } \mathrm{supp}_{\mathcal{B}}(\widetilde{\beta}_I(\bar{y})) = I.$$

According to Lemma 1, the vector $\widetilde{\beta}(\bar{y})$ is a solution of $(\mathcal{P}_\lambda(\bar{y}))$.

3. By virtue of statement 1., we are in position to use the implicit function theorem, and we get the Jacobian of $\widetilde{\beta}_I$ as

$$\partial_{\bar{y}}\widetilde{\beta}_I(\bar{y}) = -\big(\partial_1 \mathit{\Gamma}(\widetilde{\beta}_I(y), y)\big)^{-1}\big(\partial_2 \mathit{\Gamma}(\widetilde{\beta}_I(y), y)\big)$$

where $\partial_2 \mathit{\Gamma}(\widetilde{\beta}_I(y), y) = X_I^{\mathrm{T}}$, which leads us to (2).

4.3 Proof of Theorem 2

We define the set

$$\mathit{\Delta}_I = \left\{ \beta_I \in \mathbb{R}^{|I|} \setminus \forall \mu \in \mathbb{R}^{\sharp I}, \sum_{i=1}^{\sharp I} \mu_i X_{b_i}\beta_{b_i} = 0 \Rightarrow \mu = 0 \right\}\ , \tag{7}$$

where $\sharp I$ is the number of blocks in I, and $b_i \in I$ is the i-th block in I. It is easy to see that $\beta^\star \in \mathit{\Delta}_I$ for I the $\mathcal{B}$-support of $\beta^\star$ if and only if $(\mathbf{A}(\beta^\star))$.

The following lemma proves that there exists a solution $\beta^\star$ of $(\mathcal{P}_\lambda(y))$ such that $(\mathbf{A}(\beta^\star))$ holds. A similar result with a different proof can be found in (Liu and Zhang 2009).

Lemma 4 *There exists a solution $\beta^\star$ of $(\mathcal{P}_\lambda(y))$ such that $\beta^\star \in \mathit{\Delta}_I$ where $I = \mathrm{supp}_{\mathcal{B}}(\beta^\star)$.*

Proof Let β^0 be a solution of $(\mathcal{P}_\lambda(y))$ and $I = \mathrm{supp}_{\mathcal{B}}(\beta^0)$ such that $\beta^0_I \notin \Delta_I$. There exists $\mu \in \mathbb{R}^{\sharp I}$ such that

$$\sum_{i=1}^{\sharp I} \mu_i X_{b_i} \beta^0_{b_i} = 0. \tag{8}$$

Consider now the family $t \mapsto \beta^t$ defined for every $t \in \mathbb{R}$

$$\forall b_i \in I, \quad \beta^t_{b_i} = (1 + t\mu_i)\beta^0_{b_i} \quad \text{and} \quad \beta^t_{I^c} = 0 \ . \tag{9}$$

Consider $t_0 = \min\{|t| \in \mathbb{R} \setminus \exists b_i \in I \text{ such that } 1 + t\mu_i = 0\}$. Without loss of generality, we assume that $t_0 > 0$. Remark that for all $t \in [0, t_0)$, β^t is a solution of $(\mathcal{P}_\lambda(y))$. Indeed, I is the $\mathcal{B}$-support of β^t and

$$X_I \beta^t_I = X_I \beta^0_I + t \underbrace{\sum_{i=1}^{\sharp I} \mu_i X_{b_i} \beta^0_{b_i}}_{=0 \text{ using } (8)} = X_I \beta^0_I. \tag{10}$$

Hence,

$$X_I^{\mathrm{T}}(y - X_I \beta^t_I) = X_I^{\mathrm{T}}(y - X_I \beta^0_I) = \lambda \mathcal{N}(\beta_I) = \lambda \mathcal{N}(\beta^t_I),$$

and

$$\|X_b^{\mathrm{T}}(y - X_I \beta^t_I)\| = \|X_b^{\mathrm{T}}(y - X_I \beta^0_I)\| \leqslant \lambda, \quad \forall b \in I^c \ .$$

Since the image of all solutions of $(\mathcal{P}_\lambda(y))$ are equal under X, one has

$$X\beta^t = X\beta^0 \quad \text{and} \quad \|\beta^t\|_{\mathcal{B}} = \|\beta^0\|_{\mathcal{B}}.$$

where $\|\cdot\|_{\mathcal{B}}$ is the $\ell^1 - \ell^2$ norm. Consider now the vector β^{t_0}. By continuity of $\beta \mapsto X\beta$ and $\beta \mapsto \|\beta\|_{\mathcal{B}}$, one has

$$X\beta^{t_0} = X\beta^0 \quad \text{and} \quad \|\beta^{t_0}\|_{\mathcal{B}} = \|\beta^0\|_{\mathcal{B}}.$$

Hence, β^{t_0} has a $\mathcal{B}$-support I_{t_0} strictly included in I (in the sense that for all $b \in I_{t_0}$ one has $b \in I$) and is a solution of $(\mathcal{P}_\lambda(y))$. Iterating this argument with $\beta^0 = \beta^{t_0}$ shows that there exists a solution $\beta^\star$ such that $\beta^\star \in \Delta_{\mathrm{supp}_{\mathcal{B}}(\beta^\star)}$. This concludes the proof of the lemma. □

According to Theorem 1, $y \mapsto \widehat{\beta}(y)$ is $\mathcal{C}^1(\mathbb{R}^n \setminus \mathcal{H})$. This property is preserved under the linear mapping X which shows that $\widehat{\mu}$ is also $\mathcal{C}^1(\mathbb{R}^n \setminus \mathcal{H})$. Thus, taking the trace of the Jacobian $X_I d(y, \lambda)$ gives the divergence formula (5) for any solution $\widehat{\beta}(y)$ such that $(\mathbf{A}(\widehat{\beta}(y)))$ holds.

4.4 Proof of Theorem 3

The next lemma shows that the transition space has zero measure.

Lemma 5 *Let $\lambda > 0$. The transition space $\mathcal{H}$ is of zero measure with respect to the Lebesgue measure of $\mathbb{R}^n$.*

Proof We obtain this result by proving that all $\mathcal{H}_{I,b}$ are of zero measure for all I and $b \notin I$, and that the union is over a finite set.

We recall from (Coste 2002) that any semialgebraic set $S \subseteq \mathbb{R}^n$ can be decomposed in a disjoint union of q semialgebraic subsets C_i each diffeomorphic to $(0,1)^{d_i}$. The dimension of S is thus

$$d = \max_{i \in \{1,\dots,q\}} d_i \leqslant n.$$

The set $\mathcal{A}_{I,b}$ is an algebraic, hence a semialgebraic, set. By the fundamental Tarski-Seidenberg principle, the canonical projection $\pi(\mathcal{A}_{I,b})$ is also semialgebraic. The boundary $\operatorname{bd}(\pi(\mathcal{A}_{I,b}))$ is also semialgebraic with a strictly smaller dimension than $\pi(\mathcal{A}_{I,b})$

$$\dim \mathcal{H}_{I,b} = \dim \operatorname{bd}(\pi(\mathcal{A}_{I,b})) < \dim \pi(\mathcal{A}_{I,b}) \leqslant n$$

whence we deduce that $\mathcal{H}$ is of zero measure with respect to the Lebesgue measure on $\mathbb{R}^n$. □

As $\widehat{\mu}$ is uniformly Lipschitz over $\mathbb{R}^n$, using similar arguments as in (Meyer and Woodroofe 2000), we get that $\widehat{\mu}$ is weakly differentiable with an essentially bounded gradient. Moreover, the divergence formula (5) holds valid almost everywhere, except on the set $\mathcal{H}$ which is of Lebesgue measure zero. We conclude by invoking Stein's lemma (Stein 1981) to establish unbiasedness of the estimator $\widehat{df}$ of the DOF.

Plugging the DOF expression into that of the SURE (Stein 1981, Theorem 1), we get (6).

4.5 Proof of Corollary 1

When $X = \mathrm{Id}_n$, we have $X_I^{\mathrm{T}} X_I = \mathrm{Id}_I$, which in turn implies that $\mathrm{Id}_I + \lambda \delta_{\widehat{\beta}_b(y)} \circ P_{\widehat{\beta}_b(y)}$ is block-diagonal. Thus, specializing the divergence formula of Theorem 2 to $X = \mathrm{Id}_n$ yields

$$\begin{aligned}
\widehat{df} &= \operatorname{tr}\left(\left(X_I^{\mathrm{T}} X_I + \lambda \delta_{\widehat{\beta}(y)} \circ P_{\widehat{\beta}(y)}\right)^{-1}\right) \\
&= \sum_{b \in I} \operatorname{tr}\left(\left(\mathrm{Id}_b + \tfrac{\lambda}{\|\widehat{\beta}_b(y)\|}\left(\mathrm{Id}_b - \tfrac{\widehat{\beta}_b(y)\widehat{\beta}_b(y)^{\mathrm{T}}}{\|\widehat{\beta}_b(y)\|^2}\right)\right)^{-1}\right) \\
&= \sum_{b \in I} \left(1 + \frac{|b| - 1}{1 + \frac{\lambda}{\|\widehat{\beta}_b(y)\|}}\right)
\end{aligned}$$

where the last equality follows from the fact that $\mathrm{Proj}_{\widehat{\beta}(y)_b^\perp} = \mathrm{Id}_b - \frac{\widehat{\beta}_b(y)\widehat{\beta}_b(y)^{\mathrm{T}}}{\|\widehat{\beta}_b(y)\|^2}$ is the orthogonal projector on a subspace of dimension $|b|-1$.

Furthermore, for $X = \mathrm{Id}_n$, $\widehat{\beta}_b(y)$ has a closed-form given by block soft thresholding

$$\widehat{\beta}_b(y) = \begin{cases} 0 & \text{if } \|y_b\| \leqslant \lambda \\ (1 - \frac{\lambda}{\|y_b\|})y_b & \text{otherwise} \end{cases} . \tag{11}$$

It then follows that

$$\frac{1}{1+\frac{\lambda}{\|\widehat{\beta}_b(y)\|}} = \frac{1}{1+\frac{\lambda}{\|y_b\|-\lambda}} = 1 - \frac{\lambda}{\|y_b\|} .$$

Piecing everything together, we obtain

$$\widehat{df} = \sum_{b\in I}\left(1 + (|b|-1)\left(1 - \frac{\lambda}{\|y_b\|}\right)\right) = \sum_{b\in I}|b| - \lambda\sum_{b\in I}\frac{|b|-1}{\|y_b\|} .$$

As $|I| = \sum_{b\in I}|b|$, we get the desired result. Note that this result can be obtained directly by differentiating (11).

4.6 Proof of Proposition 1

Let's introduce the shorhand notation for the reliability

$$R = \mathbb{E}_w\left[\left(\mathrm{SURE}(\widehat{\mu}(y)) - \mathrm{SE}(\widehat{\mu}(y))\right)^2\right] .$$

Applying (Vaiter et al 2012b, Theorem 4), we get

$$R = 2n\sigma^4 - 4\sigma^4\,\mathrm{tr}\left(X_I B(y,\lambda) X_I^{\mathrm{T}}\left(2\mathrm{Id}_n - X_I B(y,\lambda) X_I^{\mathrm{T}}\right)\right) + 4\sigma^2\mathbb{E}_w\left(\|\widehat{\mu}(y) - \mu_0\|^2\right)$$

where $B(y,\lambda) = \left(X_I^{\mathrm{T}}X_I + \lambda\delta_{\widehat{\beta}(y)} \circ P_{\widehat{\beta}(y)}\right)^{-1}$ is positive definite by Lemma 3.

Let's bound the last term. By Jensen's inequality and the fact that $\widehat{\beta}(y)$ is a (global) minimizer of $(\mathcal{P}_\lambda(y))$, we have

$$\begin{aligned}
\mathbb{E}_w\left(\|\widehat{\mu}(y) - \mu_0\|^2\right) &\leqslant 2(\mathbb{E}_w\left(\|y - \widehat{\mu}(y)\|^2\right) + \mathbb{E}_w\left(\|y - \mu_0\|^2\right) \\
&\leqslant 4\mathbb{E}_w\left(\tfrac{1}{2}\|y - \widehat{\mu}(y)\|^2 + \lambda\sum_{b\in\mathcal{B}}\|\widehat{\beta}_b(y)\|\right) + 2n\sigma^2 \\
&\leqslant 4\mathbb{E}_w\left(\tfrac{1}{2}\|y\|^2\right) + 2n\sigma^2 = 2\|\mu_0\|^2 + 4n\sigma^2 .
\end{aligned}$$

Let's turn to the second term. We have

$$\begin{aligned}
\mathrm{tr}\left(X_I B(y,\lambda) X_I^{\mathrm{T}} X_I B(y,\lambda) X_I^{\mathrm{T}}\right) &= \mathrm{tr}\left(X_I^{\mathrm{T}} X_I B(y,\lambda) X_I^{\mathrm{T}} X_I B(y,\lambda)\right) \\
&= \|X_I^{\mathrm{T}} X_I B(y,\lambda)\|_F^2 \leqslant n\|X_I^{\mathrm{T}} X_I B(y,\lambda)\|^2 .
\end{aligned}$$

In addition, $X_I B(y,\lambda) X_I^{\mathrm{T}}$ is semidefinite positive and therefore

$$\begin{aligned}
R &\leqslant 2n\sigma^4 + 4\sigma^4\mathbb{E}_w\left(\mathrm{tr}\left(X_I B(y,\lambda) X_I^{\mathrm{T}} X_I B(y,\lambda) X_I^{\mathrm{T}}\right)\right) + 4\sigma^2\mathbb{E}_w\left(\|\widehat{\mu}(y) - \mu_0\|^2\right) \\
&\leqslant 2n\sigma^4 + 4n\sigma^4\mathbb{E}_w\left(\|X_I^{\mathrm{T}} X_I B(y,\lambda)\|^2\right) + 16n\sigma^4 + 8\sigma^2\|\mu_0\|^2 ,
\end{aligned}$$

whence we get the desired bound after dividing both sides by $n^2\sigma^4$.

References

Akaike H (1973) Information theory and an extension of the maximum likelihood principle. In: Second international symposium on information theory, Springer Verlag, vol 1, pp 267–281

Bach F (2008) Consistency of the group lasso and multiple kernel learning. Journal of Machine Learning Research 9:1179–1225

Bakin S (1999) Adaptive regression and model selection in data mining problems. Thesis (Ph.D.)–Australian National University, 1999

Bickel PJ, Ritov Y, Tsybakov A (2009) Simultaneous analysis of lasso and Dantzig selector. Annals of Statistics 37:1705–1732

Bühlmann P, van de Geer S (2011) Statistics for High-Dimensional Data: Methods, Theory and Applications. Springer

Candès E, Plan Y (2009) Near-ideal model selection by ℓ_1 minimization. Annals of Statistics 37(5A):2145–2177

Coste M (2002) An introduction to semialgebraic geometry. Tech. rep., Institut de Recherche Mathematiques de Rennes

Donoho D (2006) For most large underdetermined systems of linear equations the minimal ℓ^1-norm solution is also the sparsest solution. Communications on pure and applied mathematics 59(6):797–829

Dossal C, Kachour M, Fadili J, Peyré G, Chesneau C (2012) The degrees of freedom of penalized ℓ_1 minimization. to appear in Statistica Sinica URL http://hal.archives-ouvertes.fr/hal-00638417

Efron B (1986) How biased is the apparent error rate of a prediction rule? Journal of the American Statistical Association 81(394):461–470

Kato K (2009) On the degrees of freedom in shrinkage estimation. Journal of Multivariate Analysis 100(7):1338–1352

Liu H, Zhang J (2009) Estimation consistency of the group lasso and its applications. Journal of Machine Learning Research 5:376–383

Mallows CL (1973) Some comments on cp. Technometrics 15(4):661–675

Meyer M, Woodroofe M (2000) On the degrees of freedom in shape-restricted regression. Annals of Statistics 28(4):1083–1104

Osborne M, Presnell B, Turlach B (2000) A new approach to variable selection in least squares problems. IMA journal of numerical analysis 20(3):389

Solo V, Ulfarsson M (2010) Threshold selection for group sparsity. In: Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on, IEEE, pp 3754–3757

Stein C (1981) Estimation of the mean of a multivariate normal distribution. The Annals of Statistics 9(6):1135–1151

Tibshirani R (1996) Regression shrinkage and selection via the Lasso. Journal of the Royal Statistical Society Series B Methodological 58(1):267–288

Tibshirani RJ, Taylor J (2012) Degrees of freedom in Lasso problems. Tech. rep., arXiv:1111.0653

Tikhonov AN, Arsenin VY (1997) Solutions of Ill-posed Problems. V. H. Winston and Sons

Vaiter S, Deledalle C, Peyré G, Fadili J, Dossal C (2012a) Degrees of freedom of the group Lasso. In: ICML'12 Workshops, pp 89–92

Vaiter S, Deledalle C, Peyré G, Fadili J, Dossal C (2012b) Local behavior of sparse analysis regularization: Applications to risk estimation. to appear in Applied and Computational Harmonic Analysis URL http://hal.archives-ouvertes.fr/hal-00687751/

Wei F, Huang J (2010) Consistent group selection in high-dimensional linear regression. Bernoulli 16(4):1369–1384

Yuan M, Lin Y (2006) Model selection and estimation in regression with grouped variables. J of The Roy Stat Soc B 68(1):49–67

Zou H, Hastie T, Tibshirani R (2007) On the "degrees of freedom" of the Lasso. The Annals of Statistics 35(5):2173–2192