Statistica Sinica **23** (2013), 809-828
doi:http://dx.doi.org/10.5705/ss.2011.281

THE DEGREES OF FREEDOM OF THE LASSO FOR GENERAL DESIGN MATRIX

C. Dossal[1] M. Kachour[2], M.J. Fadili[2], G. Peyré[3] and C. Chesneau[4]

[1]CNRS-Univ. Bordeaux 1, [2]CNRS-ENSICAEN-Univ. Caen,
[3]CNRS-Univ. Paris-Dauphine and [4]CNRS-Univ. Caen

Abstract: In this paper, we investigate and give a closed-form expression of the degrees of freedom (dof) of penalized ℓ_1 minimization (also known as the Lasso) for linear regression models. Namely, we show that for any given Lasso regularization parameter λ and any observed data y belonging to a set of full (Lebesgue) measure, the cardinality of the support of a particular solution of the Lasso problem is an unbiased estimator of the degrees of freedom. This is achieved without the need of uniqueness of the Lasso solution. Thus, our result holds true for both the underdetermined and the overdetermined case; the latter was originally studied in Zou, Hastie, and Tibshirani (2007). We also show, by providing a simple counterexample, that although the dof theorem of Zou, Hastie, and Tibshirani (2007) is correct, their proof contains a flaw since their divergence formula holds on a different set of a full measure than the one that they claim. An effective estimator of the number of degrees of freedom may have several applications including an objectively guided choice of the regularization parameter in the Lasso through the SURE framework. Our theoretical findings are illustrated through several numerical simulations.

Key words and phrases: Degrees of freedom, Lasso, model selection criteria, SURE.

1. Introduction

1.1. Problem statement

We consider the linear regression model

$$y = Ax^0 + \varepsilon, \qquad \mu = Ax^0, \tag{1.1}$$

where $y \in \mathbb{R}^n$ is the observed data, $A = (a_1, \cdots, a_p)$ is an $n \times p$ design matrix, $x^0 = \left(x^0_1, \cdots, x^0_p\right)^{\mathrm{T}}$ is the vector of unknown regression coefficients, and ε is a vector of i.i.d. centered Gaussian random variables with variance $\sigma^2 > 0$. In this paper, the number of observations n can be greater than p, the dimension of the regression vector to be estimated. When $n < p$, (1.1) is an underdetermined linear regression model, and when $n \geq p$ and all the columns of A are linearly independent, it is overdetermined.

Let $\widehat{x}(y)$ be an estimator of x^0, and $\widehat{\mu}(y) = A\widehat{x}(y)$ be the associated response or predictor. The concept of degrees of freedom plays a pivotal role in quantifying the complexity of a statistical modeling procedure. More precisely, since $y \sim \mathcal{N}(\mu = Ax^0, \sigma^2 \mathrm{Id}_{n\times n})$ ($\mathrm{Id}_{n\times n}$ is the identity on $\mathbb{R}^n$), according to Efron (1986), the degrees of freedom (dof) of the response $\widehat{\mu}(y)$ is defined by

$$df = \sum_{i=1}^{n} \frac{\mathrm{Cov}\,(\widehat{\mu}_i(y), y_i)}{\sigma^2}. \tag{1.2}$$

Many model selection criteria involve df, e.g. C_p (Mallows Mallows (1973)), AIC (Akaike Information Criterion, Akaike (1973)), BIC (Bayesian Information Citerion, Schwarz (1978)), GCV (Generalized Cross Validation, Craven and Wahba (1979)), and SURE (Stein's unbiased risk estimation Stein (1981), see Sect. 2.2). Thus, the dof is a quantity of interest in model validation and selection, and it can be used to get the optimal hyperparameters of the estimator. Note that the optimality here is intended in the sense of the prediction $\widehat{\mu}(y)$ and not the coefficients $\widehat{x}(y)$.

The well-known Stein's lemma Stein (1981) states that if $y \mapsto \widehat{\mu}(y)$ is weakly differentiable then its divergence is an unbiased estimator of its degrees of freedom,

$$\widehat{df}(y) = \mathrm{div}(\widehat{\mu}(y)) = \sum_{i=1}^{n} \frac{\partial \widehat{\mu}_i(y)}{\partial y_i}, \quad \text{and} \quad \mathbb{E}(\widehat{df}(y)) = df\ . \tag{1.3}$$

Here, in order to estimate x^0, we consider solutions to the Lasso problem, originally proposed in Tibshirani (1994). The Lasso amounts to solving the following convex optimization problem

$$\min_{x \in \mathbb{R}^p} \frac{1}{2}\|y - Ax\|_2^2 + \lambda\|x\|_1, \tag{$\mathrm{P}_1(y,\lambda)$}$$

where $\lambda > 0$ is called the Lasso regularization parameter and $\|\cdot\|_2$ (resp. $\|\cdot\|_1$) denotes the ℓ_2 (resp. ℓ_1) norm. An important feature of the Lasso is that it promotes sparse solutions. In the last years, there has been a huge amount of work in which efforts have focused on investigating the theoretical guarantees of the Lasso as a sparse recovery procedure from noisy measurements. See, e.g., Fan and Li (2001); Peng and Fan (2004); Zhao and Yu (2006); Zou (2006); Ravikumar et al. (2008); Nardi and Rinaldo (2008); Osborne, Presnell, and Turlach (2000a); Efron et al. (2004); Fuchs (2004); Tropp (2006), to name just a few.

1.2. Contributions and related work

Let $\widehat{\mu}_\lambda(y) = A\widehat{x}_\lambda(y)$ be the Lasso response vector, where $\widehat{x}_\lambda(y)$ is a solution of the Lasso problem ($\mathrm{P}_1(y,\lambda)$). Note that all minimizers of the Lasso share the same image under A, i.e. $\widehat{\mu}_\lambda(y)$ is uniquely defined; see Lemma 2 in Section 5 for details. Our main contribution is to provide an unbiased estimator of the degrees of freedom of the Lasso response for any design matrix. The estimator is valid everywhere except on a set of (Lebesgue) measure zero. We reach our goal without additional assumptions to ensure uniqueness of the Lasso solution. Thus, our result covers the challenging underdetermined case where the Lasso problem does not necessarily have a unique solution. It obviously holds when the Lasso problem $(\mathrm{P}_1(y,\lambda))$ has a unique solution and, in particular, in the overdetermined case originally studied in Zou, Hastie, and Tibshirani (2007). Using the estimator at hand, we also establish the reliability of the SURE as an unbiased estimator of the Lasso prediction risk.

While this paper was submitted, we became aware of the independent work of Tibshirani and Taylor Tibshirani and Taylor (2012), who studied the dof for general A both for the Lasso and the general (analysis) Lasso.

Section 3 is dedicated to a thorough comparison and discussion of connections and differences between our results and the one in Zou, Hastie, and Tibshirani (2007, Thm. 1) for the overdetermined case, and that of Kato (2009); Tibshirani and Taylor (2012); Vaiter et al. (2011) for the general case.

1.3. Overview of the paper

This paper is organized as follows. In Section 2, we state our main result. We provide an unbiased estimator of the dof of the Lasso, and we investigate the reliability of the SURE estimate of the Lasso prediction risk. Then, we discuss our work in relation to the literature in Section 3. Numerical illustrations are given in Section 4. The proofs of our results are postponed to Section 5. Conclusions are drawn and some perspectives of this work are provided in Section 6.

2. Main Results

2.1. An unbiased estimator of the dof

Some notations and definitions are necessary. For any vector x, x_i denotes its ith component. The support (or the active set) of x is $I = \mathrm{supp}(x) = \{i : x_i \neq 0\}$. and we denote its cardinality as $|\mathrm{supp}(x)| = |I|$. We denote by $x_I \in \mathbb{R}^{|I|}$ the vector built by restricting x to the entries indexed by I. The active matrix $A_I = (a_i)_{i\in I}$ associated to a vector x is obtained by selecting the columns of A indexed by the support I of x. Let $\cdot^{\mathrm{T}}$ denote the transpose. If A_I is full column rank, then we denote its Moore-Penrose pseudo-inverse by $A_I^+ = (A_I^{\mathrm{T}} A_I)^{-1} A_I^{\mathrm{T}}$. The sign function has $\mathrm{sign}(a) = 1$ if $a > 0$; $\mathrm{sign}(0) = 0$; $\mathrm{sign}(a) = -1$ if $a < 0$. For any $I \subseteq \{1, 2, \cdots, p\}$, let $V_I = \mathrm{span}(A_I)$, P_{V_I} the orthogonal projector onto V_I, and $P_{V_I^\perp}$ that onto the orthogonal complement $V_I^\perp$.

Let $S \in \{-1,1\}^{|I|}$ be a sign vector, and $j \in \{1,2,\cdots,p\}$. Fix $\lambda > 0$. We define the collection of hyperplanes

$$H_{I,j,S} = \{u \in \mathbb{R}^n : \langle P_{V_I^\perp}(a_j), u\rangle = \pm\lambda(1 - \langle a_j, (A_I^+)^{\mathrm{T}} S\rangle)\}. \tag{2.1}$$

Note that, if a_j does not belong to V_I, then $H_{I,j,S}$ becomes a finite union of two hyperplanes. Now, take the finite set of indices

$$\Omega = \{(I,j,S) : a_j \notin V_I\}\ , \tag{2.2}$$

and let G_λ be the subset of $\mathbb{R}^n$ that excludes the finite union of hyperplanes associated to Ω,

$$G_\lambda = \mathbb{R}^n \setminus \bigcup_{(I,j,S)\in\Omega} H_{I,j,S}. \tag{2.3}$$

Therefore, as $\bigcup_{(I,j,S)\in\Omega} H_{I,j,S}$ is a set of (Lebesgue) measure zero (Hausdorff dimension $n-1$), G_λ is a set of full measure.

We are now ready to state our main theorem.

Theorem 1. *Fix $\lambda > 0$. For any $y \in G_\lambda$, let $\mathcal{M}_{y,\lambda}$ be the set of solutions of $(\mathrm{P}_1(y,\lambda))$, and et $x_\lambda^*(y) \in \mathcal{M}_{y,\lambda}$ with support I^* such that A_{I^*} is full rank. Then*

$$|I^*| = \min_{\widehat{x}_\lambda(y)\in\mathcal{M}_{y,\lambda}} |\mathrm{supp}(\widehat{x}_\lambda(y))|. \tag{2.4}$$

Furthermore, there exists $\varepsilon > 0$ such that for all $z \in \mathrm{Ball}(y,\varepsilon)$, the n-dimensional ball with center y and radius ε, the Lasso response mapping $z \mapsto \widehat{\mu}_\lambda(z)$ satisfies

$$\widehat{\mu}_\lambda(z) = \widehat{\mu}_\lambda(y) + P_{V_{I^*}}(z-y). \tag{2.5}$$

This theorem assumes the existence of a solution whose active matrix A_{I^*} is full rank. This can be shown to be true; see e.g. Dossal (2007, Proof of Theorem 1) or Rosset, Zhu, and Hastie (2004, Thm. 3, Sec. B.1) It is worth noting that this proof is constructive, in that it yields a solution $x_\lambda^*(y)$ of $(\mathrm{P}_1(y,\lambda))$ such that A_{I^*} is full column rank from any solution $\widehat{x}_\lambda(y)$ whose active matrix has a nontrivial kernel. This will be exploited in Section 4 to derive an algorithm to get $x_\lambda^*(y)$, and hence I^*.

A direct consequence of Theorem 1 is that outside G_λ, the mapping $\widehat{\mu}_\lambda(y)$ is C^∞ and the sign and support are locally constant. Applying Stein's lemma one has that the number of nonzero coefficients of $x_\lambda^*(y)$ is an unbiased estimator of the dof of the Lasso.

Corollary 1. *Under the assumptions and with the same notations as in Theorem* 1*, we have the divergence formula*

$$\widehat{df}_\lambda(y) := \operatorname{div}(\widehat{\mu}_\lambda(y)) = |I^*|. \tag{2.6}$$

Therefore,

$$df = \mathrm{E}\,(\widehat{df}_\lambda(y)) = \mathrm{E}\,(|I^*|). \tag{2.7}$$

Obviously, in the particular case where the Lasso problem has a unique solution, our result holds true.

2.2. Reliability of the SUREestimate of the Lasso prediction risk

In this work, we focus on the SURE as a model selection criterion. For the Lasso,

$$\mathrm{SURE}(\widehat{\mu}_\lambda(y)) = -n\sigma^2 + \|\widehat{\mu}_\lambda(y) - y\|_2^2 + 2\sigma^2\widehat{df}_\lambda(y), \tag{2.8}$$

where $\widehat{df}(y)$ is an unbiased estimator of the dof as given in Corollary 1. It follows that $\mathrm{SURE}(\widehat{\mu}_\lambda(y))$ is an unbiased estimate of the prediction risk,

$$\mathrm{MSE}\,(\mu) = \mathrm{E}\,\left(\|\widehat{\mu}_\lambda(y) - \mu\|_2^2\right) = \mathrm{E}\,\left(\mathrm{SURE}(\widehat{\mu}_\lambda(y))\right).$$

We now evaluate its reliability by computing the expected squared-error between $\mathrm{SURE}(\widehat{\mu}_\lambda(y))$ and $\mathrm{SE}(\widehat{\mu}_\lambda(y))$, the true squared-error, that is

$$\mathrm{SE}(\widehat{\mu}_\lambda(y)) = \|\widehat{\mu}_\lambda(y) - \mu\|_2^2. \tag{2.9}$$

Theorem 2. *Under the assumptions of Theorem* 1*, we have*

$$\mathrm{E}\,\left((\mathrm{SURE}(\widehat{\mu}_\lambda(y)) - \mathrm{SE}(\widehat{\mu}_\lambda(y)))^2\right) = -2\sigma^4 n + 4\sigma^2\mathrm{E}\,\left(\|\widehat{\mu}_\lambda(y) - y\|_2^2\right) + 4\sigma^4\mathrm{E}\,(|I^*|)\,. \tag{2.10}$$

Moreover,

$$\mathrm{E}\,\left(\left(\frac{\mathrm{SURE}(\widehat{\mu}_\lambda(y)) - \mathrm{SE}(\widehat{\mu}_\lambda(y))}{n\sigma^2}\right)^2\right) = O\left(\frac{1}{n}\right). \tag{2.11}$$

3. Relation to Prior Work

3.1. Overdetermined case

Zou, Hastie, and Tibshirani (2007) studied the dof of the Lasso in the overdetermined case. Precisely, when $n \geq p$ and all the columns of the design matrix A are linearly independent, $\operatorname{rank}(A) = p$. In fact, in this case the Lasso problem has a unique minimizer $\widehat{x}_\lambda(y) = x^*_\lambda(y)$ (see Theorem 1).

Before discussing the result of Zou, Hastie, and Tibshirani (2007), we point out a popular feature of $\widehat{x}_\lambda(y)$ as λ varies in $]0, +\infty[$. For $\lambda \geq \|A^{\mathrm{T}}y\|_\infty$, the optimum is attained at $\widehat{x}_\lambda(y) = 0$. The interval $]0, \|A^{\mathrm{T}}y\|_\infty[$ is divided into a finite number of subintervals characterized by the fact that within each such subinterval, the support and the sign vector of $\widehat{x}_\lambda(y)$ are constant. Explicitly, let $(\lambda_m)_{0\leq m\leq K}$ be the finite sequence of λ's values corresponding to a variation of the support and the sign of $\widehat{x}_\lambda(y)$, defined by

$$\|A^{\mathrm{T}}y\|_\infty = \lambda_0 > \lambda_1 > \lambda_2 > \cdots > \lambda_K = 0.$$

Thus, in $]\lambda_{m+1}, \lambda_m[$, the support and the sign of $\widehat{x}_\lambda(y)$ are constant, see Efron et al. (2004); Osborne, Presnell, and Turlach (2000a,b). Hence, we call $(\lambda_m)_{0\leq m\leq K}$ the *transition points*.

Now, let $\lambda \in]\lambda_{m+1}, \lambda_m[$. From Lemma 1 (see Section 5), we have the following implicit form of $\widehat{x}_\lambda(y)$,

$$(\widehat{x}_\lambda(y))_{I_m} = A^+_{I_m} y - \lambda (A^{\mathrm{T}}_{I_m} A_{I_m})^{-1} S^m, \tag{3.1}$$

where I_m and S^m are, respectively, the (constant) support and sign vector of $\widehat{x}_\lambda(y)$ for $\lambda \in]\lambda_{m+1}, \lambda_m[$. Hence, based on (3.1), Zou, Hastie, and Tibshirani (2007) showed that for all $\lambda > 0$, there exists a set of measure zero $\mathcal{N}_\lambda$, a finite collection of hyperplanes in $\mathbb{R}^n$ with

$$\mathcal{K}_\lambda = \mathbb{R}^n \setminus \mathcal{N}_\lambda \ . \tag{3.2}$$

λ is not any of the transition points $\forall\ y \in \mathcal{K}_\lambda$.

Then, for the overdetermined case, Zou, Hastie, and Tibshirani (2007) stated that the number of nonzero coefficients of the unique solution of $(\mathrm{P}_1(y,\lambda))$ is an unbiased estimator of the dof. The dof estimator formula is valid for all $y \in \mathcal{K}_\lambda$. In fact, their main argument is that, by eliminating the vectors associated to the transition points, the support and the sign of the Lasso solution are locally constant with respect to y, see Zou, Hastie, and Tibshirani (2007, Lemma 5).

We recall that the overdetermined case, considered in Zou, Hastie, and Tibshirani (2007), is a particular case of our result since the minimizer is unique. Thus, according to the Corollary 1, we find the same result as Zou, Hastie, and Tibshirani (2007) but valid on a different set $y \in G_\lambda = \mathbb{R}^n \setminus \bigcup_{(I,j,S)\in\Omega} H_{I,j,S}$. A natural question arises: can we compare our assumption to that of Zou, Hastie, and Tibshirani (2007) ? In other words, is there a link between $\mathcal{K}_\lambda$ and G_λ ?

The answer is that, depending on the matrix A, these two sets may be different. More importantly, it turns out that although the dof formula of Zou, Hastie, and Tibshirani (2007, Thm. 1) is correct, their proof contains a flaw since their divergence formula is not true on the set $\mathcal{K}_\lambda$. We prove this by providing a simple counterexample.

Example of vectors in G_λ but not in $\mathcal{K}_\lambda$

Let $\{e_1, e_2\}$ be an orthonormal basis of $\mathbb{R}^2$ and $a_1 = e_1$ and $a_2 = e_1 + e_2$, with A the matrix whose columns are a_1 and a_2.

Take $I = \{1\}$, $j = 2$, and $S = 1$. It turns out that $A_I^+ = a_1$ and $\langle (A_I^+)^{\mathrm{T}} S, a_j \rangle = 1$, which implies that for all $\lambda > 0$,

$$H_{I,j,S} = \{u \in \mathbb{R}^n : \langle P_{V_I^\perp}(a_j), u \rangle = 0\} = \operatorname{span}(a_1) \ .$$

Let $y = \alpha a_1$ with $\alpha > 0$, for any $\lambda > 0$, $y \in H_{I,j,S}$ (or, equivalently here, $y \notin G_\lambda$). Using Lemma 1 (see Section 5), one gets that for any $\lambda \in]0, \alpha[$, the solution of $(\mathrm{P}_1(y, \lambda))$ is $\widehat{x}_\lambda(y) = (\alpha - \lambda, 0)$ and that for any $\lambda \geq \alpha$, $\widehat{x}_\lambda(y) = (0, 0)$. Hence the only transition point is $\lambda_0 = \alpha$. It follows that for $\lambda < \alpha$, y belongs to $\mathcal{K}_\lambda$, but $y \notin G_\lambda$, see Figure 1.

We prove then that in any ball centered at y, there exists a vector z_1 such that the support of the solution of $(\mathrm{P}_1(z_1, \lambda))$ is different from the support of $(\mathrm{P}_1(y, \lambda))$. Choose $\lambda < \alpha$ and $\varepsilon \in]0, \alpha - \lambda[$ and take $z_1 = y + \varepsilon e_2$. From Lemma 1 (see Section 5), one deduces that the solution of $(\mathrm{P}_1(z_1, \lambda))$ is $\widehat{x}_\lambda(z_1) = (\alpha - \lambda - \varepsilon, \varepsilon)$ whose support is different from that of $\widehat{x}_\lambda(y) = (\alpha - \lambda, 0)$.

More generally, when there are sets $\{I, j, S\}$ such that $\langle (A_I^+)^{\mathrm{T}} S, a_j \rangle = 1$, a difference between the two sets G_λ and $\mathcal{K}_\lambda$ may arise. Clearly, G_λ is not only the set of transition points associated to λ.

Thus, in this specific situation, for any $\lambda > 0$ there may exist some vectors y that are not transition points, associated to λ, where the support of the solution of $(\mathrm{P}_1(y, \lambda))$ is not stable to infinitesimal perturbations of y. This situation may occur in under or overdetermined problems. In summary, even in the overdetermined case, excluding the set of transition points is not sufficient to guarantee stability of the support and sign of the Lasso solution.

Note that recently, in the overdetermined, the author in Zhang (2010) also proved that the cardinality of the support is an unbiased estimator of the dof of the Lasso (as a corollary of a more general result for sparsity penalties including some concave ones). However, there is not an explicit characterization of the set outside which the dof estimate formula is valid, nor reference to Zou, Hastie, and Tibshirani (2007).

3.2. General case

Kato (2009) studied the degrees of freedom of a generalization of the Lasso where the regression coefficients are constrained to a closed convex set. When the latter is a ℓ_1 ball and $p > n$, he proposes the cardinality of the support as an estimate of df, but under a restrictive assumption on A under which the Lasso problem has a unique solution.

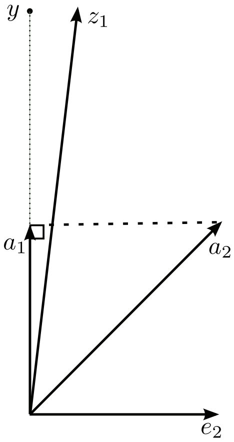

Figure 1. A counterexample for $n = p = 2$ of vectors in G_λ but not in $\mathcal{K}_\lambda$. See text for a detailed discussion.

Tibshirani and Taylor (2012, Thm. 2) proved that

$$df = \mathrm{E}\left(\mathrm{rank}(A_I)\right)$$

where $I = I(y)$ is the active set of any solution $\widehat{x}_\lambda(y)$ to $(\mathrm{P}_1(y, \lambda))$. This coincides with Corollary 1 when A_I is full rank with $\mathrm{rank}(A_I) = \mathrm{rank}(A_{I^*})$. Note that in general, there exist vectors $y \in \mathbb{R}^n$ where the smallest cardinality among all supports of Lasso solutions is different from the rank of the active matrix associated to the largest support. But these vectors are precisely those excluded in G_λ. In the case of the generalized Lasso (also known as analysis sparsity prior in the signal processing community), Vaiter et al. (2011, Corollary 1) and Tibshirani and Taylor (2012, Thm. 3) provide a formula of an unbiased estimator of df. This formula reduces to that of Corollary 1 when the analysis operator is the identity.

4. Numerical Experiments

4.1. Experiment description

In this section, we support the validity of our main theoretical findings with some numerical simulations, by checking the unbiasedness and the reliability of the SURE for the Lasso.

For our first study, we considered two kinds of design matrices A, a random Gaussian matrix with $n = 256$ and $p = 1,024$ whose entries are $\sim_{\mathrm{iid}} \mathcal{N}(0, 1/n)$, and a deterministic convolution design matrix A with $n = p = 256$ and a Gaussian blurring function. The original sparse vector x^0 was drawn randomly according to a mixed Gaussian-Bernoulli distribution, such that $|\mathrm{supp}(x^0) = 15|$. For each

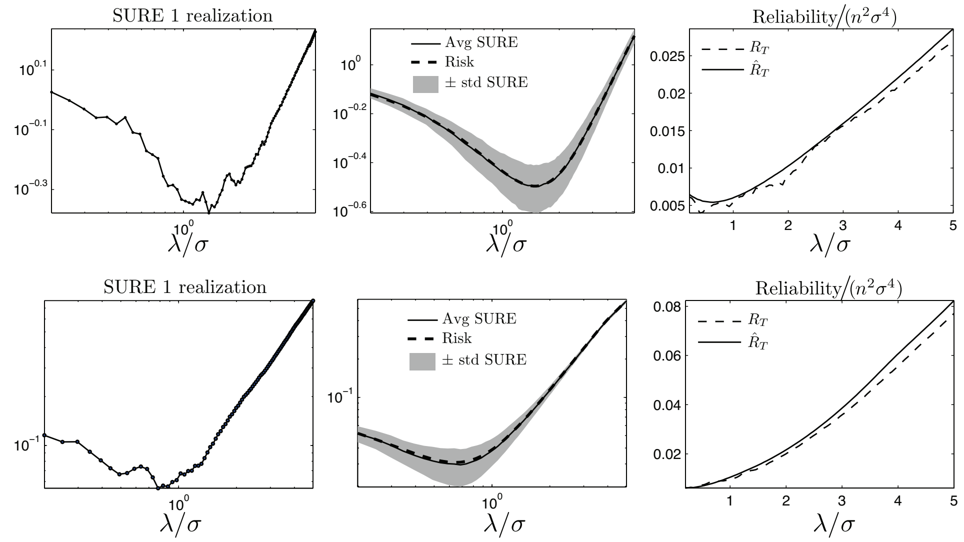

Figure 2. The SURE and its reliability as a function of λ for two types of design matrices: (a) Gaussian; (b) Convolution. For each design matrix, we associate three plots.

design matrix A and vector x^0, we generated $K = 100$ independent replications $y^k \in \mathbb{R}^n$ of the observation vector according to the linear regression model (1.1). Then, for each y^k and a given λ, we computed the Lasso response $\widehat{\mu}_\lambda(y^k)$ using the iterative soft-thresholding algorithm Daubechies, Defrise, and Mol (2004)[1], and we computed $\mathrm{SURE}(\widehat{\mu}_\lambda(y^k))$ and $\mathrm{SE}(\widehat{\mu}_\lambda(y^k))$. We then computed the empirical mean and the standard deviation of $\big(\mathrm{SURE}(\widehat{\mu}_\lambda(y^k))\big)_{1\le k\le K}$, the empirical mean of $\big(\mathrm{SE}(\widehat{\mu}_\lambda(y^k))\big)_{1\le k\le K}$, which corresponds to the computed prediction risk, and we computed R_T, the empirical normalized reliability on the left-hand side of (2.10),

$$R_T = \frac{1}{K}\sum_{k=1}^{K}\left(\frac{\mathrm{SURE}(\widehat{\mu}_\lambda(y^k)) - \mathrm{SE}(\widehat{\mu}_\lambda(y^k))}{n\sigma^2}\right)^2. \tag{4.1}$$

Moreover, based on the right-hand side of (2.10), we compute $\widehat{R}_T$ as

$$\widehat{R}_T = -\frac{2}{n} + \frac{4}{n^2\sigma^2}\left(\frac{1}{K}\sum_{k=1}^{K}\left(\|\widehat{\mu}_\lambda(y^k) - y^k\|_2^2\right)\right) + \frac{4}{n^2}\left(\frac{1}{K}\sum_{k=1}^{K}\left(|I^*|_k\right)\right) \tag{4.2}$$

where, at the kth replication, $|I^*|_k$ is the cardinality of the support of a Lasso solution whose active matrix is full column rank as stated in Theorem 1. Finally,

[1] Iterative soft-thresholding through block-coordinate relaxation was proposed in Sardy, Bruce, and Tseng (2000) for matrices A structured as the union of a finite number of orthonormal matrices.

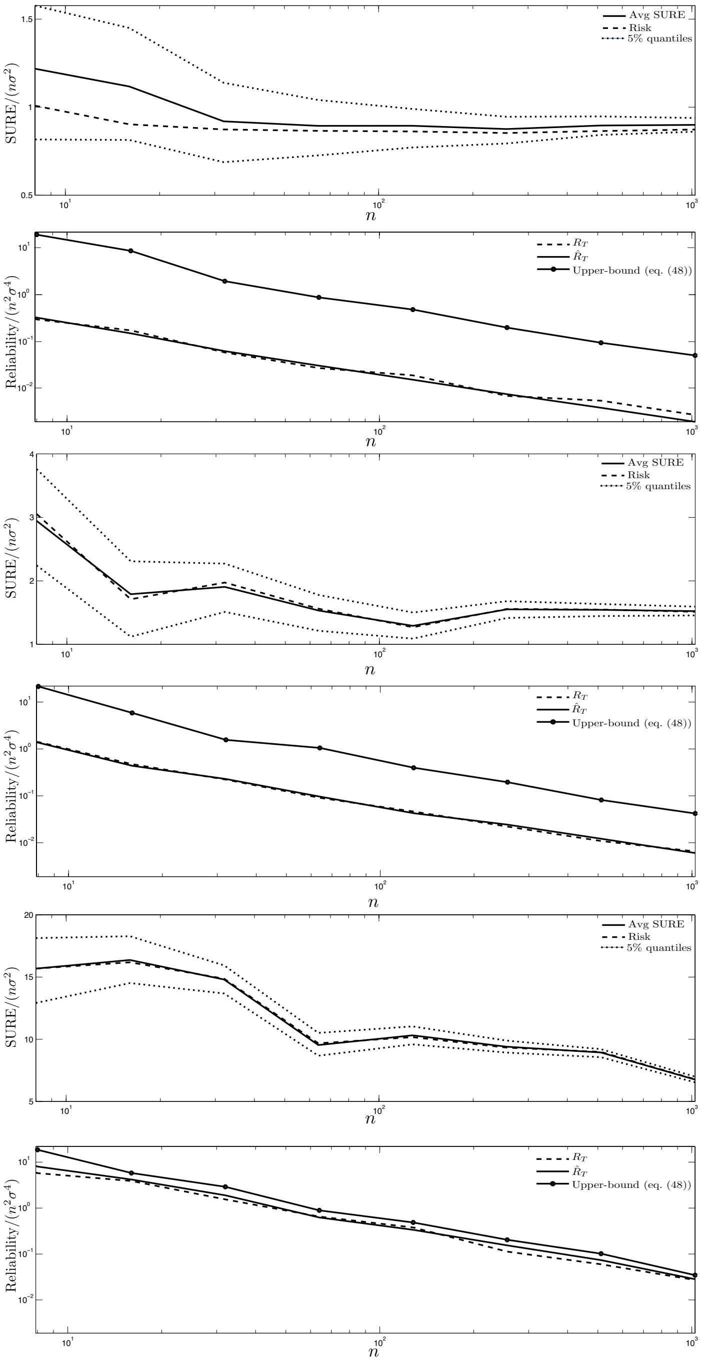

Figure 3. The SURE and its reliability as a function of the number of observations n.

we repeated all these computations for various values of λ, for the two kinds of design matrices considered above.

4.2. Construction of full rank active matrix

As stated in the discussion just after Theorem 1, in situations where the Lasso problem has non-unique solutions and the minimization algorithm returns a solution whose active matrix is rank deficient, one can construct an alternative optimal solution whose active matrix is full column rank, and then get the estimator of the degrees of freedom.

More precisely, let $\widehat{x}_\lambda(y)$ be a solution of the Lasso problem with support I such that its active matrix A_I has a non-trivial kernel. The construction is as follows:

1. Take $h \in \ker A_I$ such that $\operatorname{supp} h \subset I$.
2. For $t \in \mathbb{R}$, $A\widehat{x}_\lambda(y) = A\left(\widehat{x}_\lambda(y) + th\right)$ and the mapping $t \mapsto \|\widehat{x}_\lambda(y) + th\|_1$ is locally affine in a neighborhood of 0, i.e. for $|t| < \min_{j \in I} |(\widehat{x}_\lambda(y))_j| / \|h\|_\infty$. $\widehat{x}_\lambda(y)$ being a minimizer of $(\mathrm{P}_1(y, \lambda))$, this mapping is constant in a neighborhood of 0. We have then constructed a whole collection of solutions to $(\mathrm{P}_1(y, \lambda))$ having the same image and the same ℓ_1 norm, which lives on a segment.
3. Move along h with the largest step $t_0 > 0$ until an entry of $\widehat{x}^1_\lambda(y) = \widehat{x}_\lambda(y) + t_0 h$ vanishes yielding $\operatorname{supp}(\widehat{x}^1_\lambda(y) + t_0 h) \subsetneq I$.
4. Repeat this process to get a vector $x^*_\lambda(y)$ with a full column rank active matrix A_{I^*}.

Note that this construction bears similarities with the one in Rosset, Zhu, and Hastie (2004).

4.3. Results discussion

Figure 2 depicts the obtained results. For each design matrix, we associate a panel containing three plots. From left to right, the first plot gives the SURE for one realization of the noise as a function of λ. In the second graph, as a function of the regularization parameter λ, the dashed curve is the calculated prediction risk, the solid curve is the empirical mean of the SURE, and the shaded area is the empirical mean of the SURE ± the empirical standard deviation of the SURE. The latter confirms that the SURE is an unbiased estimator of the prediction risk with a controlled variance. This suggests that the SURE is consistent, and then so is our estimator of the degrees of freedom. In the third graph, the solid and dashed blue curves are respectively R_T and $\widehat{R}_T$ as a function of the regularization parameter λ. This confirms numerically that both sides R_T and $\widehat{R}_T$ indeed coincide as predicted by (2.10).

As discussed above, one of the motivations of having an unbiased estimator of the degrees of freedom of the Lasso is to provide a data-driven objective way for selecting the optimal Lasso regularization parameter λ. For this, one can compute the optimal λ by minimizing the SURE

$$\lambda_{\text{optimal}} = \underset{\lambda > 0}{\operatorname{argmin}} \ \text{SURE}(\widehat{\mu}_\lambda(y)). \tag{4.3}$$

In practice, this optimal value can be found either by a exhaustive search over a fine grid or, alternatively, by any dicothomic search algorithm (e.g. golden section) if $\lambda \mapsto \text{SURE}(\widehat{\mu}_\lambda(y))$ is unimodal.

For our second simulation study, we considered a partial Fourier design matrix with $n < p$ and a constant underdeterminacy factor $p/n = 4$. x^0 was again simulated according to a mixed Gaussian-Bernoulli distribution with $\lceil 0.1p \rceil$ nonzero entries. For each of three values of $\lambda/\sigma \in \{0.1, 1, 10\}$ (small, medium and large), we computed the prediction risk curve, the empirical mean of the SURE, as well as the values of the normalized reliabilities R_T and $\widehat{R}_T$, as a function of $n \in \{8, \ldots, 1,024\}$. The results are shown in Figure 3. For each value of λ, the first plot (top panel) displays the normalized empirical mean of the SURE (solid line) and its 5% quantiles (dotted) as well as the computed normalized prediction risk (dashed). Unbiasedness is again clear whatever the value of λ. The trend on the prediction risk (and average SURE) is in agreement with rates known for the Lasso, see e.g., Bickel et al. (2009). The second plot confirms that the SURE is an asymptotically reliable estimate of the prediction risk with the rate established in Theorem 2. Moreover, as expected, the actual reliability gets closer to the upper-bound (5.29) as the number of samples n increases.

5. Proofs

First of all, we recall some classical properties of any solution of the Lasso (see, e.g., Osborne, Presnell, and Turlach (2000a); Efron et al. (2004); Fuchs (2004); Tropp (2006)). To lighten the notation, we drop the dependency of the minimizers of $(\text{P}_1(y,\lambda))$ on either λ or y.

Lemma 1. *$\widehat{x}$ is a (global) minimizer of the Lasso problem $(\text{P}_1(y,\lambda))$ if, and only if*

1. *$A_I^{\mathrm{T}}(y - A\widehat{x}) = \lambda \text{sign}(\widehat{x}_I)$, where $I = \{i : \widehat{x}_i \neq 0\}$, and*
2. *$|\langle a_j, y - A\widehat{x}\rangle| \leq \lambda$, $\forall\ j \in I^c$,*

where $I^c = \{1, \ldots, p\} \setminus I$. Moreover, if A_I is full column rank, then $\widehat{x}$ satisfies the implicit relationship

$$\widehat{x}_I = A_I^{+} y - \lambda (A_I^{\mathrm{T}} A_I)^{-1} \text{sign}(\widehat{x}_I) \ . \tag{5.1}$$

Note that if the inequality in condition 2 above is strict, then $\widehat{x}$ is the unique minimizer of the Lasso problem $(\mathrm{P}_1(y,\lambda))$ Fuchs (2004).

Lemma 2 shows the Lasso response $\widehat{\mu}_\lambda(y)$ is well-defined as a single-valued mapping of y, see Dossal (2007).

Lemma 2. *If $\widehat{x}^1$ and $\widehat{x}^2$ are two solutions of $(\mathrm{P}_1(y,\lambda))$, then $A\widehat{x}^1 = A\widehat{x}^2 = \widehat{\mu}_\lambda(y)$.*

Before delving into the technical details, we recall a trace formula for the divergence. Let $J_{\widehat{\mu}(y)}$ be the Jacobian matrix of a mapping $y \mapsto \widehat{\mu}(y)$

$$\left(J_{\widehat{\mu}(y)}\right)_{i,j} := \frac{\partial \widehat{\mu}(y)_i}{\partial y_j}, \qquad i,j = 1,\cdots,n. \tag{5.2}$$

Then we can write

$$\mathrm{div}\,(\widehat{\mu}(y)) = \mathrm{tr}\left(J_{\widehat{\mu}(y)}\right). \tag{5.3}$$

Proof of Theorem 1. Let $x^*_\lambda(y)$ be a solution of the Lasso problem $(\mathrm{P}_1(y,\lambda))$, and I^* its support such that A_{I^*} is full column rank. Let $(x^*_\lambda(y))_{I^*}$ be the restriction of $x^*_\lambda(y)$ to its support and $S^* = \mathrm{sign}\,((x^*_\lambda(y))_{I^*})$. From Lemma 2 we have

$$\widehat{\mu}_\lambda(y) = Ax^*_\lambda(y) = A_{I^*}(x^*_\lambda(y))_{I^*}.$$

According to Lemma 1, we know that

$$\begin{aligned} &A^{\mathrm{T}}_{I^*}(y - \widehat{\mu}_\lambda(y)) = \lambda S^*; \\ &|\langle a_k, y - \widehat{\mu}_\lambda(y)\rangle| \le \lambda, \forall\ k \in (I^*)^c. \end{aligned}$$

Furthermore, from (5.1), we get the implicit form of $x^*_\lambda(y)$

$$(x^*_\lambda(y))_{I^*} = A^+_{I^*}y - \lambda(A^{\mathrm{T}}_{I^*}A_{I^*})^{-1}S^*. \tag{5.4}$$

It follows that

$$\widehat{\mu}_\lambda(y) = P_{V_{I^*}}(y) - \lambda d_{I^*,S^*}, \tag{5.5}$$

$$\widehat{r}_\lambda(y) = y - \widehat{\mu}_\lambda(y) = P_{V^\perp_{I^*}}(y) + \lambda d_{I^*,S^*}\ , \tag{5.6}$$

where $d_{I^*,S^*} = (A^+_{I^*})^{\mathrm{T}}S^*$. Let

$$J = \{j : |\langle a_j, \widehat{r}_\lambda(y)\rangle| = \lambda\}. \tag{5.7}$$

From Lemma 1 we deduce that $I^* \subset J$. Since the orthogonal projection is a self-adjoint operator and from (5.6), for all $j \in J$, we have

$$|\langle P_{V^\perp_{I^*}}(a_j), y\rangle + \lambda\langle a_j, d_{I^*,S^*}\rangle| = \lambda. \tag{5.8}$$

As $y \in G_\lambda$, we deduce that if $j \in J \cap (I^*)^c$ then we necessarily have

$$a_j \in V_{I^*}, \text{ and therefore } |\langle a_j, d_{I^*,S^*}\rangle| = 1. \tag{5.9}$$

In fact, if $a_j \notin V_{I^*}$ then $(I^*, j, S^*) \in \Omega$ and, from (5.8), we have that $y \in H_{I^*,j,S^*}$, which is a contradiction with $y \in G_\lambda$.

The collection of vectors $(a_i)_{i \in I^*}$ forms a basis of $V_J = \text{span}(a_j)_{j \in J}$. Now, suppose that $\widehat{x}_\lambda(y)$ is another solution of $(\mathrm{P}_1(y, \lambda))$ such that its support I is different from I^*. If A_I is full column rank, then by using the same arguments as above we can deduce that $(a_i)_{i \in I}$ also forms a basis of V_J. Thus

$$|I| = |I^*| = \dim(V_J).$$

On the other hand, if A_I is not full rank, then there exists a subset $I_0 \subsetneq I$ such that A_{I_0} is full rank (see the discussion following Theorem 1) and $(a_i)_{i \in I_0}$ also forms a basis of V_J, which implies that

$$|I| > |I_0| = \dim(V_J) = |I^*|.$$

We conclude that for any solution $\widehat{x}_\lambda(y)$ of $(\mathrm{P}_1(y, \lambda))$, we have

$$|\text{supp}(\widehat{x}_\lambda(y))| \geq |I^*|,$$

and then $|I^*|$ is the minimum of the cardinalities of the supports of solutions of $(\mathrm{P}_1(y, \lambda))$. This proves the first part of the theorem.

For the second statement, note that G_λ is an open set and all components of $(x^*_\lambda(y))_{I^*}$ are nonzero, so we can choose a small enough ε such that $\text{Ball}(y, \varepsilon) \subsetneq G_\lambda$. Now, let $x^1_\lambda(z)$ be the vector supported in I^*

$$(x^1_\lambda(z))_{I^*} = A^+_{I^*} z - \lambda (A^{\mathrm{T}}_{I^*} A_{I^*})^{-1} S^* = (x^*_\lambda(y))_{I^*} + A^+_{I^*}(z - y). \tag{5.10}$$

If ε is small enough, then for all $z \in \text{Ball}(y, \varepsilon)$ we have

$$\text{sign}(x^1_\lambda(z))_{I^*} = \text{sign}(x^*_\lambda(y))_{I^*} = S^*. \tag{5.11}$$

For the rest, we invoke Lemma 1 to show that, for ε small enough, $x^1_\lambda(z)$ is actually a solution of $(\mathrm{P}_1(z, \lambda))$. First we notice that $z - Ax^1_\lambda(z) = P_{V_I^\perp}(z) + \lambda d_{I^*,S^*}$. It follows that

$$A^{\mathrm{T}}_{I^*}(z - Ax^1_\lambda(z)) = \lambda A^{\mathrm{T}}_{I^*} d_{I^*,S^*} = \lambda S^* = \lambda \text{sign}(x^1_\lambda(z))_{I^*}. \tag{5.12}$$

Moreover for all $j \in J \cap I^*$, from (5.9), we have that

$$\begin{aligned} |\langle a_j, z - Ax^1_\lambda(z)\rangle| &= |\langle a_j, P_{V_{I^*}^\perp}(z) + \lambda d_{I^*,S^*}\rangle| \\ &= |\langle P_{V_{I^*}^\perp}(a_j), z\rangle + \lambda \langle a_j, d_{I^*,S^*}\rangle| \\ &= \lambda |\langle a_j, d_{I^*,S^*}\rangle| = \lambda. \end{aligned}$$

and for all $j \notin J$

$$|\langle a_j, z - Ax^1_\lambda(z)\rangle| \leq |\langle a_j, y - Ax^*_\lambda(y)\rangle| + |\langle P_{V^\perp_{I^*}}(a_j), z - y\rangle|.$$

Since for all $j \notin J$, $|\langle a_j, y - Ax^*_\lambda\rangle| < \lambda$, there exists ε such that for all $z \in \mathrm{Ball}(y, \varepsilon)$ and $\forall\ j \notin J$, we have

$$|\langle a_j, z - Ax^1_\lambda(z)\rangle| < \lambda.$$

Therefore, we obtain

$$|\langle a_j, z - Ax^1_\lambda(z)\rangle| \leq \lambda, \forall\ j \in (I^*)^c.$$

By Lemma 1, $x^1_\lambda(z)$ is a solution of $(\mathrm{P}_1(z, \lambda))$ and the unique Lasso response associated to $(\mathrm{P}_1(z, \lambda))$, denoted by $\widehat{\mu}_\lambda(z)$, is

$$\widehat{\mu}_\lambda(z) = P_{V_{I^*}}(z) - \lambda d_{I^*, S^*}. \tag{5.13}$$

Therefore, from (5.5) and (5.13), we can deduce that for all $z \in \mathrm{Ball}(y, \varepsilon)$ we have

$$\widehat{\mu}_\lambda(z) = \widehat{\mu}_\lambda(y) + P_{V_{I^*}}(z - y).$$

Proof of Corollary 1. We showed that there exists ε sufficiently small such that

$$\|z - y\|_2 \leq \varepsilon \Rightarrow \widehat{\mu}_\lambda(z) = \widehat{\mu}_\lambda(y) + P_{V_{I^*}}(z - y). \tag{5.14}$$

Let $h \in V_{I^*}$ such that $\|h\|_2 \leq \varepsilon$ and $z = y + h$. Thus we have that $\|z - y\|_2 \leq \varepsilon$, and then

$$\|\widehat{\mu}_\lambda(z) - \widehat{\mu}_\lambda(y)\|_2 = \|P_{V_{I^*}}(h)\|_2 = \|h\|_2 \leq \varepsilon. \tag{5.15}$$

Therefore, the Lasso response $\widehat{\mu}_\lambda(y)$ is uniformly Lipschitz on G_λ. Moreover, $\widehat{\mu}_\lambda(y)$ is a continuous function of y, and thus $\widehat{\mu}_\lambda(y)$ is uniformly Lipschitz on $\mathbb{R}^n$. Hence, $\widehat{\mu}_\lambda(y)$ is almost differentiable; see Meyer and Woodroofe (2000) and Efron et al. (2004).

On the other hand, we proved that there exists a neighborhood of y such that for all z in this neighborhood, there exists a solution of the Lasso problem $(\mathrm{P}_1(z, \lambda))$, that has the same support and the same sign as $x^*_\lambda(y)$, and thus $\widehat{\mu}_\lambda(z)$ belongs to the vector space V_{I^*} whose dimension is $|I^*|$, see (5.5) and (5.13). Therefore, $\widehat{\mu}_\lambda(y)$ is a locally affine function of y and

$$J_{\widehat{\mu}_\lambda(y)} = P_{V_{I^*}}\ . \tag{5.16}$$

Then the trace formula (5.3) implies that

$$\mathrm{div}\left(\widehat{\mu}_\lambda(y)\right) = \mathrm{tr}\left(P_{V_{I^*}}\right) = |I^*|. \tag{5.17}$$

This holds almost everywhere since G_λ is of full measure, and (2.7) is obtained by invoking Stein's Lemma.

Proof of Theorem 2. First, consider the random variable

$$Q_1(\widehat{\mu}_\lambda(y)) = \|\widehat{\mu}_\lambda(y)\|_2^2 + \|\mu\|_2^2 - 2\langle y, \widehat{\mu}_\lambda(y)\rangle + 2\sigma^2 \mathrm{div}(\widehat{\mu}_\lambda(y)).$$

From Stein's Lemma, we have

$$\mathrm{E}\,\langle \varepsilon, \widehat{\mu}_\lambda(y)\rangle = \sigma^2 \mathrm{E}\,\left(\mathrm{div}(\widehat{\mu}_\lambda(y))\right).$$

Thus, we can deduce that $Q_1(\widehat{\mu}_\lambda(y))$ and $\mathrm{SURE}(\widehat{\mu}_\lambda(y))$ are unbiased estimator of the prediction risk, i.e.

$$\mathrm{E}\,\left(\mathrm{SURE}(\widehat{\mu}_\lambda(y))\right) = \mathrm{E}\,\left(Q_1(\widehat{\mu}_\lambda(y))\right) = \mathrm{E}\,\left(\mathrm{SE}(\widehat{\mu}_\lambda(y))\right) = \mathrm{MSE}\,(\mu).$$

Moreover, note that $\mathrm{SURE}(\widehat{\mu}_\lambda(y)) - Q_1(\widehat{\mu}_\lambda(y)) = \|y\|_2^2 - \mathrm{E}\,\left(\|y\|_2^2\right)$, where

$$\mathrm{E}\,\left(\|y\|_2^2\right) = n\sigma^2 + \|\mu\|_2^2, \text{ and } \mathbb{V}\left(\|y\|_2^2\right) = 2\sigma^4\left(n + 2\frac{\|\mu\|_2^2}{\sigma^2}\right). \tag{5.18}$$

Now, we remark also that

$$Q_1(\widehat{\mu}_\lambda(y)) - \mathrm{SE}(\widehat{\mu}_\lambda(y)) = 2\left(\sigma^2 \mathrm{div}(\widehat{\mu}_\lambda(y)) - \langle \varepsilon, \widehat{\mu}_\lambda(y)\rangle\right). \tag{5.19}$$

After an elementary calculation, we obtain

$$\mathrm{E}\left(\mathrm{SURE}(\widehat{\mu}_\lambda(y)) - \mathrm{SE}(\widehat{\mu}_\lambda(y))\right)^2 = \mathrm{E}\left(Q_1(\widehat{\mu}_\lambda(y)) - \mathrm{SE}(\widehat{\mu}_\lambda(y))\right)^2 + \mathbb{V}\left(\|y\|_2^2\right) + 4T, \tag{5.20}$$

where

$$T = \sigma^2 \mathrm{E}\,\left(\mathrm{div}(\widehat{\mu}_\lambda(y))\|y\|_2^2\right) - \mathrm{E}\,\left(\langle \varepsilon, \widehat{\mu}_\lambda(y)\rangle \|y\|_2^2\right) = T_1 + T_2, \tag{5.21}$$

with

$$T_1 = 2\left(\sigma^2 \mathrm{E}\,\left(\mathrm{div}(\widehat{\mu}_\lambda(y))\langle \varepsilon, \mu\rangle\right) - \mathrm{E}\,\left(\langle \varepsilon, \widehat{\mu}_\lambda(y)\rangle\langle \varepsilon, \mu\rangle\right)\right) \tag{5.22}$$

$$T_2 = \sigma^2 \mathrm{E}\,\left(\mathrm{div}(\widehat{\mu}_\lambda(y))\|\varepsilon\|_2^2\right) - \mathrm{E}\,\left(\langle \varepsilon, \widehat{\mu}_\lambda(y)\rangle\|\varepsilon\|_2^2\right). \tag{5.23}$$

Hence, by using the fact that a Gaussian probability density $\varphi(\varepsilon_i)$ satisfies $\varepsilon_i\varphi(\varepsilon_i) = -\sigma^2\varphi'(\varepsilon_i)$ and integration by parts, we find that

$$T_1 = -2\sigma^2 \mathrm{E}\,\left(\langle \widehat{\mu}_\lambda, \mu\rangle\right)$$

and

$$T_2 = -2\sigma^4 \mathrm{E}\,\left(\mathrm{div}(\widehat{\mu}_\lambda(y))\right).$$

It follows that

$$T = -2\sigma^2\big(\mathrm{E}\,(\langle\widehat{\mu}_\lambda,\mu\rangle) + \sigma^2\mathrm{E}\,(\mathrm{div}(\widehat{\mu}_\lambda(y)))\,\big). \tag{5.24}$$

Moreover, from Luisier (2009, Property 1), we know that

$$\mathrm{E}\,(Q_1(\widehat{\mu}_\lambda(y)) - \mathrm{SE}(\widehat{\mu}_\lambda(y)))^2 = 4\sigma^2\bigg(\mathrm{E}\,\left(\|\widehat{\mu}_\lambda(y)\|_2^2\right) + \sigma^2\mathrm{E}\,\left(\mathrm{tr}\left(\left(J_{\widehat{\mu}_\lambda(y)}\right)^2\right)\right)\bigg)\,. \tag{5.25}$$

Since $J_{\widehat{\mu}_\lambda(y)} = P_{V_{I^*}}$ which is an orthogonal projector (hence self-adjoint and idempotent), we have $\mathrm{tr}\left(\left(J_{\widehat{\mu}_\lambda(y)}\right)^2\right) = \mathrm{div}(\widehat{\mu}_\lambda(y)) = |I^*|$. Therefore, we get

$$\mathrm{E}\,(Q_1(\widehat{\mu}_\lambda(y)) - \mathrm{SE}(\widehat{\mu}_\lambda(y)))^2 = 4\sigma^2\left(\mathrm{E}\,\left(\|\widehat{\mu}_\lambda(y)\|_2^2\right) + \sigma^2\mathrm{E}\,(|I^*|)\right). \tag{5.26}$$

Furthermore, observe that

$$\mathrm{E}\,(\mathrm{SURE}(\widehat{\mu}_\lambda(y))) = -n\sigma^2 + \mathrm{E}\,\left(\|\widehat{\mu}_\lambda(y) - y\|_2^2\right) + 2\sigma^2\mathrm{E}\,(|I^*|)\,. \tag{5.27}$$

By combining (5.18), (5.20), (5.24), and (5.26), we obtain

$$\begin{aligned}\mathrm{E}\,(\mathrm{SURE}(\widehat{\mu}_\lambda(y)) - \mathrm{SE}(\widehat{\mu}_\lambda(y)))^2 &= 2n\sigma^4 + 4\sigma^2\mathrm{E}\,(\mathrm{SE}(\widehat{\mu}_\lambda(y))) - 4\sigma^4\mathrm{E}\,(|I^*|)\\ &= 2n\sigma^4 + 4\sigma^2\mathrm{E}\,(\mathrm{SURE}(\widehat{\mu}_\lambda(y))) - 4\sigma^4\mathrm{E}\,(|I^*|)\\ (\text{by using (5.27)}) &= -2n\sigma^4 + 4\sigma^2\mathrm{E}\,\left(\|\widehat{\mu}_\lambda(y) - y\|_2^2\right) + 4\sigma^4\mathrm{E}\,(|I^*|)\,.\end{aligned}$$

On the other hand, since $x^*_\lambda(y)$ is a minimizer of the Lasso problem $(\mathrm{P}_1(y,\lambda))$, we observe that

$$\frac{1}{2}\|\widehat{\mu}_\lambda(y) - y\|_2^2 \leq \frac{1}{2}\|\widehat{\mu}_\lambda(y) - y\|_2^2 + \lambda\|x^*_\lambda(y)\|_1 \leq \frac{1}{2}\|A.0 - y\|_2^2 + \lambda\|0\|_1 = \frac{1}{2}\|y\|_2^2.$$

Therefore,

$$\mathrm{E}\,\left(\|\widehat{\mu}_\lambda(y) - y\|_2^2\right) \leq \mathrm{E}\,\left(\|y\|_2^2\right) = n\sigma^2 + \|\mu\|_2^2. \tag{5.28}$$

Then, since $|I^*| = O(n)$ and from (5.28), we have

$$\mathrm{E}\,\left(\left(\frac{\mathrm{SURE}(\widehat{\mu}_\lambda(y)) - \mathrm{SE}(\widehat{\mu}_\lambda(y))}{n\sigma^2}\right)^2\right) \leq \frac{6}{n} + \frac{4\|\mu\|_2^2}{n^2\sigma^2}. \tag{5.29}$$

Finally, since $\|\mu\|_2 < +\infty$, we can deduce that

$$\mathrm{E}\,\left(\left(\frac{\mathrm{SURE}(\widehat{\mu}_\lambda(y)) - \mathrm{SE}(\widehat{\mu}_\lambda(y))}{n\sigma^2}\right)^2\right) = O\left(\frac{1}{n}\right).$$

6. Discussion

In this paper we proved that the number of nonzero coefficients of a particular solution of the Lasso problem is an unbiased estimate of the degrees of freedom of the Lasso response for linear regression models. This result covers both the over and underdetermined cases. This was achieved through a divergence formula, valid almost everywhere except on a set of measure zero. We gave a precise characterization of this set, and it turns out to be larger than the set of all the vectors associated to the transition points considered in Zou, Hastie, and Tibshirani (2007) in the overdetermined case. We also highlight the fact that, even in the overdetermined case, the set of transition points is not sufficient for the divergence formula to hold.

We think that some techniques developed in this article can be applied to derive the degrees of freedom of other nonlinear estimating procedures. Typically, a natural extension of this work is to consider other penalties such as those promoting structured sparsity, e.g. the group Lasso.

Acknowledgement

This work was partly funded by the ANR grant NatImages, ANR-08-EMER-009.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory*, 267-281.

Bickel, J., Ritov, Y. and Tsybakov, B. (2009). Simultaneous analysis of lasso and dantzig selector. *Ann. Statist.* **37**, 1705-1732.

Craven, P. and Wahba, G. (1979). Smoothing noisy data with spline functions: Estimating the correct degree of smoothing by the method of generalized cross-validation. *Numer. Math.* **31**, 377-403.

Daubechies, I., Defrise, M. and Mol, C. D. (2004). An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Comm. Pure Appl. Math.* **57**, 1413-1457.

Dossal, C. (2007). A necessary and sufficient condition for exact recovery by ℓ_1 minimization. Technical Report HAL-00164738:1.

Efron, B. (1986). How Biased is the Apparent Error Rate of a Prediction Rule? *J. Amer. Statist. Assoc.* **81**, 461-470.

Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004). Least angle regression. *Ann. Statist.* **32**, 407-499.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96**, 1348-1360.

Fuchs, J.-J. (2004). On sparse representations in arbitrary redundant bases. *IEEE Trans. Information Theory* **50**, 1341-1344.

Kato, K. (2009). On the degrees of freedom in shrinkage estimation. *J. Multivariate Anal.* **100**, 1338-1352.

Luisier, F. (2009). The SURE-LET approach to image denoising. Ph.D. thesis, EPFL, Lausanne.

Mallows, C. L. (1973). Some comments on C_p. *Technometrics* **15**, 661-675.

Meyer, M. and Woodroofe, M. (2000). On the degrees of freedom in shape restricted regression. *Ann. Statist.* **28**, 1083-1104.

Nardi, Y. and Rinaldo, A. (2008). On the asymptotic properties of the group lasso estimator in least squares problems. *Electronic J. Statist.* **2**, 605-633.

Osborne, M. R., Presnell, B. and Turlach, B. A. (2000a). A new approach to variable selection in least squares problems. *IMA J. Numerical Anal.* **20**, 389-403.

Osborne, M. R., Presnell, B. and Turlach, B. A. (2000b). On the lasso and its dual. *J. Comput. Graph. Statist.* **9**, 319-337.

Peng, H. and Fan, J. (2004). Nonconcave penalized likelihood with a diverging number of parameters. *Ann. Statist.* **32**, 928-961.

Ravikumar, P., Liu, H., Lafferty, J. and Wasserman, L. (2008). Spam: Sparse additive models. In *Advances in Neural Information Processing Systems* (*NIPS*), volume 22.

Rosset, S., Zhu, J. and Hastie, T. (2004). Boosting as a regularized path to a maximum margin classifier. *J. Mach. Learn. Res.* **5**, 941-973.

Sardy, S., Bruce, A. and Tseng, P. (2000). Block coordinate relaxation methods for nonparametric wavelet denoising. *J. Comput. Graph. Statist.* **9**, 361-379.

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6**, 461-464.

Stein, C. M. (1981). Estimation of the mean of a multivariate normal distribution. *Ann. Statist.* **9**, 1135-1151.

Tibshirani, R. (1994). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58**, 267-288.

Tibshirani, R. J. and Taylor, J. (2012). Degrees of freedom in lasso problems. Technical Report arXiv:1111.0653v4.

Tropp, T. (2006). Just relax: Convex programming methods for subset selection and sparse approximation. *IEEE Trans. Inform. Theory* **52**, 1030-1051.

Vaiter, S., Peyré, G., Dossal, C. and Fadili, M. J. (2011). Robust sparse analysis regularization. Technical Report arXiv:1109.6222.

Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist.* **38**, 894-942.

Zhao, P. and Yu, B. (2006). On model selection consistency of lasso. *J. Mach. Learn. Res.* **7**, 2541-2563.

Zou, H. (2006). The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc.* **101**, 1418-1429.

Zou, H., Hastie, T. and Tibshirani, R. (2007). On the "degrees of freedom" of the lasso. *Ann. Statist.* **35**, 2173-2192.

IMB, CNRS-Univ. Bordeaux 1, 351 Cours de la Libération, F-33405 Talence, France.

E-mail: Charles.Dossal@math.u-bordeaux1.fr

GREYC, CNRS-ENSICAEN-Univ. Caen, 6 Bd du Maréchal Juin, 14050 Caen, France.

E-mail: Jalal.Fadili@greyc.ensicaen.fr

GREYC, CNRS-ENSICAEN-Univ. Caen, 6 Bd du Maréchal Juin, 14050 Caen, France.

E-mail: Maher.Kachour@greyc.ensicaen.fr

Ceremade, CNRS-Univ. Paris-Dauphine, Place du Maréchal De Lattre De Tassigny, 75775 Paris 16, France.

E-mail: Gabriel.Peyre@ceremade.dauphine.fr

LMNO, CNRS-Univ. Caen, Département de Mathématiques, UFR de Sciences, 14032 Caen, France.

E-mail: Chesneau.Christophe@math.unicaen.fr

(Received November 2011; accepted August 2012)