The degrees of freedom of penalized l1 minimization

Charles H Dossal, Maher Kachour, Jalal M. Fadili, Gabriel Peyré, Christophe Chesneau

The degrees of freedom of penalized ℓ_1 minimization

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

(1) IMB, Université Bordeaux 1
351 cours de la Libération, F-33405 Talence Cedex, France
E-mail : Charles.Dossal@math.u-bordeaux1.fr

(2) GREYC, ENSICAEN
6 Bd du Maréchal Juin, 14050 Caen Cedex, France
{maher.kachour, jalal.fadili}@ensicaen.fr

(3) CNRS and Ceremade, Université Paris-Dauphine
Place du Maréchal De Lattre De Tassigny, 75775 Paris Cedex 16, France
E-mail : gabriel.peyre@ceremade.dauphine.fr

(4) LMNO, Université de Caen
Département de Mathématiques, UFR de Sciences, 14032 Caen, France
E-mail : christophe.chesneau@gmail.com

Abstract

In this paper, we investigate the degrees of freedom (df) of penalized ℓ_1 minimization (also known as the Lasso) for linear regression models. We give a closed-form expression of the degrees of freedom of the Lasso response. Namely, we show that for any given Lasso regularization parameter λ and any observed data y belongs to a set of full measure, the cardinal of the support of a particular solution of the Lasso problem is an unbiased estimator of the degrees of freedom of the Lasso response. This work is achieved without any assumption on the uniqueness of the Lasso solution. Thus, our result remains true for both the underdetermined and the overdetermined case studied originally in [27]. We also prove that a key result in [27] is not true by providing a simple counterexample. 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.

Keywords: Lasso, model selection criteria, degrees of freedom, SURE.

AMS classification code: Primary $62M10$, secondary $62M20$.

1 Introduction

1.1 Problem statement

We consider the following linear regression model

$$y = Ax^0 + \varepsilon, \qquad \mu = Ax^0, \tag{1}$$

where $y \in \mathbb{R}^n$ is the observed data or the response vector, $A = (a_1, \cdots, a_p)$ is an $n \times p$ deterministic design matrix, $x^0 = \left(x_1^0, \cdots, x_p^0\right)^t$ 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 observation number n can be greater or less than the number of the parameter to be estimated p. Recall that when $n < p$, (1) is called underdetermined linear regression model, which is probably the most famous example of statistical problems in high dimensional. On the other hand, when all the vectors of the design matrix A are linearly independent, which is only possible if $n \geq p$, (1) is called overdetermined linear regression model.

To estimate x^0, we consider the least absolute shrinkage and selection operator (Lasso) procedure, proposed originally by Tibshirani [22]. The Lasso estimate amounts to solving the following convex optimization problem

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

where $\lambda > 0$ is called the Lasso regularization parameter and $\|\cdot\|_2$ (resp. $\|\cdot\|_1$) denotes the ℓ_2 (resp. ℓ_1) norm. The convexity of this minimization problem ensures that the estimator can be computed even if $n < p$ and with very large p. An important feature of the Lasso is that, depending on the regularization parameter, some coefficients are exactly set to zero. In the last years, there has been a huge amount of work where 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 [7], Fan and Peng [8], Zhao and Bin [25], Zou [26], Ravikumar et al. [18], Nardi and Rinaldo [15], Osborne et al. [16], Efron et al. [5], Fuchs [10] and Tropp [23], to mention just a few.

Degrees of freedom *df* is a familiar phrase in statistics. More generally, degrees of freedom is often used to quantify the complexity of a statistical modeling procedure. However, there is no exact correspondence between the degrees of freedom *df* and the number of parameters in the model. Now, let us introduce a precise definition of the degrees of freedom of any fitting procedure and reveals its statistical importance. Let $\hat{x} = \delta(y)$ be an estimator of x^0, and let $\hat{\mu} = A\hat{x}$ be the response or the predictor associated to $\hat{x}$. Since $y \sim N(\mu = Ax^0, \sigma^2 I)$, and according to Efron [6], the degrees of freedom of the response $\hat{\mu}$ is defined by

$$df(\hat{\mu}) = \sum_{i=1}^{n} \frac{\mathrm{cov}(\hat{\mu}_i, y_i)}{\sigma^2}. \tag{3}$$

For example, when $\hat{\mu}$ is given by a linear function of y, i.e. $\hat{\mu} = \delta(y) = Sy$, with some matrix S being independent of y, the degrees of freedom equals to the trace of S, i.e. $df(\hat{\mu}) = \mathrm{tr}(S)$.

With *df* defined in (3), we can employ the covariance penalty method to construct a C_p-type statistic (Mallows [13]) as

$$C_p(\hat{\mu}) = -n\sigma^2 + \|\hat{\mu} - y\|_2^2 + 2\sigma^2 df(\hat{\mu}). \tag{4}$$

Note that C_p is an unbiased estimator of the true risk or the true prediction error

$$\mathrm{Risk}(\hat{\mu}) = \mathbb{E}\|\hat{\mu} - \mu\|_2^2 = \mathbb{E}\|A(\hat{x} - x^0)\|_2^2. \tag{5}$$

Moreover, Efron [4] showed that in some settings C_p offers substantially better accuracy than cross-validation and related nonparametric methods. Many others model selection criteria involve $df(\hat{\mu})$, e.g. AIC (Akaike Information Criterion, [1]), BIC (Bayesian Information Citerion, [19]), GCV (Generalized Cross Validation, [2]) and SURE (Stein's unbiased risk estimation, see below). Thus, the concept of degrees of freedom plays an important role in model validation and selection.

The degrees of freedom intervenes also to finding the optimal hyperparameters of the estimator, e.g the regularization parameter λ in the Lasso. Note that, the optimality here is in the sense of the prediction and not for the estimation. Finally, we find the degrees of freedom in the formula of the Fischer statistic used for the global test on the prediction.

In this work, we are interested in particular by the SURE as a model selection criteria. Indeed, suppose that the degrees of freedom has an unbiased estimator, denoted by $\hat{df}(\hat{\mu})$, the SURE is defined as follows

$$\text{SURE}(\hat{\mu}) = -n\sigma^2 + \|\hat{\mu} - y\|_2^2 + 2\sigma^2 \hat{df}(\hat{\mu}). \tag{6}$$

Hence, by replacing $df(\hat{\mu})$ in (4) by its its unbiased estimate $\hat{df}(\hat{\mu})$, the SURE or the modified C_p statistic is still unbiased as an estimate of the true risk (5). Lemma 1 below, called the Stein's lemma, ensures that if $\hat{\mu}$ is continuous and almost differentiable then its divergence is an unbiased estimator of its degrees of freedom.

Lemma 1. *(**Stein's lemma** [20]). Let $y \sim N(\mu, \sigma^2 I)$ and $\hat{\mu}$ be an estimator of μ. Suppose that $\hat{\mu}_i : \mathbb{R}^n \to \mathbb{R}$ is absolutely continuous in i-th coordinate for $i = 1, \cdots, n$. If $\mathbb{E}|\dfrac{\partial \hat{\mu}_i}{\partial y_i}| < \infty$ for each i, then*

$$\sum_{i=1}^{n} \frac{\text{cov}(\hat{\mu}_i, y_i)}{\sigma^2} = \mathbb{E}\left(\text{div}\, \hat{\mu}\right), \tag{7}$$

where

$$\text{div}\, \hat{\mu} = \sum_{i=1}^{n} \frac{\partial \hat{\mu}_i}{\partial y_i}.$$

Therefore an unbiased estimator of the degrees of freedom is given by

$$\hat{df}(\hat{\mu}) = \text{div}\, \hat{\mu}. \tag{8}$$

1.2 Some characteristics of the Lasso solution

In this section, we recall some classical properties of the Lasso solution (see, e.g., Osborne et al. [16], Efron et al. [5], Fuchs [10] and Tropp [23]). First, some notations are necessary. Let $\tilde{x} \in \mathbb{R}^p$. $\tilde{x}_i$ denotes the ith component of $\tilde{x}$. The support or the active set of $\tilde{x}$ is defined by

$$I = \text{supp}(\tilde{x}) = \{i : \tilde{x}_i \neq 0\},$$

and we denote its cardinal as $|\,\text{supp}(\tilde{x})| = |I|$. Moreover, we denote by $\tilde{x}_I$ the reduced dimensional vector built upon the non-zero components of $\tilde{x}$. The active matrix A_I associated to a vector $\tilde{x}$ is obtained by selecting the colomns of A indexed by the support I of $\tilde{x}$. Let A_I^t be the transpose matrix of A_I. Suppose that A_I is full rank, then the pseudo-inverse $(A_I^t A_I)^{-1} A_I^t$ of A_I is denoted A_I^+. $\text{sign}(\cdot)$ represents the sign function: $\text{sign}(a) = 1$ if $a > 0$; $\text{sign}(a) = 0$ if $a = 0$; $\text{sign}(a) = -1$ if $a < 0$. Let $sign(\tilde{x})$ be the sign vector of $\tilde{x}$, such that $sign(\tilde{x})_i = \text{sign}(\tilde{x}_i)$.
Next, we recall the first order optimality conditions for the Lasso estimator, see [9] and [10].

Lemma 2. *A necessary and sufficient condition for $\tilde{x}$ to be a minimizer of the Lasso problem $P_1(y, \lambda)$ is that $\tilde{x}$ satisfies the two following conditions:*

1. $A_I^t(y - A\tilde{x}) = \lambda sign(\tilde{x}_I)$, *i.e.* $\langle a_k, y - A\tilde{x}\rangle = \lambda \,\text{sign}\, \tilde{x}_k$, $\forall\ k \in I$,
2. $|\langle a_j, y - A\tilde{x}\rangle| \leq \lambda$, $\forall\ j \in I^c$,

where I^c is the complement of I. Moreover, if A_I is full rank, then $\tilde{x}$ satisfies the following implicit relationship:

$$\tilde{x}_I = A_I^+ y - \lambda (A_I^t A_I)^{-1} \,\text{sign}(\tilde{x}_I). \tag{9}$$

Note that if the inequality in the condition 2 is strict, then $\tilde{x}$ is the unique minimizer of the Lasso problem $\text{P}_1(y, \lambda)$. Lemma 3 below shows that all the solutions of $\text{P}_1(y, \lambda)$ have the same image by A. In others words, the lasso response, denoted by $\hat{\mu}_\lambda(y)$, is unique, see [3].

Lemma 3. *If $\tilde{x}_1$ and $\tilde{x}_2$ are solutions of $P_1(y, \lambda)$, then*

$$A\tilde{x}_1 = A\tilde{x}_2 = \hat{\mu}_\lambda(y).$$

Uniqueness of the Lasso solution In the statistical framework of (1), and under the sparsity assumption on x^0 the vector of unknown regression parameters, we distinguish two cases. First, the overdetermined case, that is, when all the vectors of the design matrix A are linearly independent i.e. $\text{rank}(A) = p$. In this case the Lasso problem has a unique solution. On the other hand, the underdetermined case presented in the introduction of this paper. Thus, in this case the Lasso problem may have several solutions. If points $(\pm a_i)_{i\leqslant p}$ are in general position, that is if any affine subspace of $\mathbb{R}^n$ of dimension k contains less than $k+1$ points amongst $(\pm a_i)_{i\leqslant p}$, (exluding antipodal paires), then the solution of the LASSO is unique for all $y \in \mathbb{R}^n$ and for all $\lambda > 0$. If these points are in general position we will say that A satisfies condition (GP).

Moreover, this condition is satisfied by most matrices. Precisely, for any matrix A, the matrix $A+W$, where W is a random matrix whose columns are independent and follows a probability law with a density, satisfies (GP) with probability 1. For instance, it is the case when the entries of the design matrix A are identically and independently sampled from a standard normal distribution.

1.3 Contributions and relationship to prior work

Let $\hat{\mu}_\lambda(y) = A\hat{x}_\lambda(y)$ be the unique Lasso response vector, where $\hat{x}_\lambda(y)$ is a solution of the Lasso problem (2). The main contribution of this paper is first to generalize the results of [27] to the more challenging underdetermined case where the Lasso solution may not be unique. We provide an unbiased estimator of the degrees of freedom of the Lasso response valid everywhere except on a set of measure zero. Let's mention that we reach our goal without any additional assumption to ensure the uniqueness of the Lasso solution. Thus, our result is valid when the Lasso problem (2) has a unique solution, and in particular for the overdetermined case studied by Zou et al. [27]. Indeed, for the overdetermined case, authors [27] shows that for a fixed λ, and y outside a finite union of hyperplanes, the number of non-zero coefficients of the unique solution of the Lasso problem is an unbiased estimator of the degrees of the freedom of the response Lasso. In this work, we arrive at a similar expression of the degrees of freedom as in [27] for the overdetermined case, but with the notable distinction that it holds on a different set (of full measure) for the observed data y. Section 3 is dedicated to a thorough comparison and discussion of differences between our results, when specialized to the overdetermined case, and that in [27, Theorem 1]. On the other hand, using the estimator at hand, we establish the reliability of the SURE for the Lasso.

1.4 Overview of the paper

This paper is organized as follows. Section 2 is the core contribution of this work where we state our main results. There, we provide the unbiased estimator of the degrees of freedom of the Lasso, and we investigate the reliability of the SURE estimate of the Lasso response. Then, we compare our result with that of [27] for the overdetermined case, in Section 3. Numerical illustrations are given in Section 4. The proofs of our results are postponed to Section 5. A final discussion and perspectives of this work are provided in Section 6.

2 Main results

2.1 An unbiased estimator of *df*

We first define some notation. Let $I \subseteq \{1, 2, \cdots, p\}$, such that $A_I = (a_i)_{i\in I}$ is full rank. We denote the cardinal of I by $|I|$, the range of A_I by V_I, the orthogonal projection onto V_I by P_{V_I}, and the orthogonal projection onto the orthogonal complement $V_I^\perp$ of V_I by $P_{V_I^\perp}$. We recall

$$V_I = \text{span}(a_i)_{i\in I}, \quad P_{V_I} = A_I A_I^+, \quad \text{and} \quad P_{V_I^\perp} = I_{n\times n} - P_{V_I}.$$

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

$$H_{I,j,S} = \{u \in \mathbb{R}^n : \langle P_{V_I^\perp}(a_j), u\rangle = \pm\lambda(1 - \langle a_j, (A_I^+)^t S\rangle)\}. \tag{10}$$

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

$$\Omega = \{(I, j, S) : a_j \notin V_I\} \tag{11}$$

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

$$G_\lambda = \mathbb{R}^n \setminus \bigcup_{(I,j,S)\in\Omega} H_{I,j,S}. \tag{12}$$

To cut a long story short, $\bigcup_{(I,j,S)\in\Omega} H_{I,j,S}$ is a set of (Lebesgue) measure zero (Hausdorff dimension $n-1$), and therefore G_λ is a set of full measure.

Now, we are now ready to introduce our main theorem.

Theorem 1. *Fix $\lambda > 0$. For any $y \in G_\lambda$, consider $\mathcal{M}_{y,\lambda}$ the set of solutions of $P_1(y, \lambda)$. Let $x^*_\lambda \in \mathcal{M}_{y,\lambda}$ with support I^* such that A_{I^*} is full rank. Then,*

$$|I^*| = \min_{x_\lambda \in \mathcal{M}_{y,\lambda}} |\operatorname{supp}(x_\lambda)|. \tag{13}$$

Furthermore, there exists $\varepsilon > 0$ such that for all $z \in \operatorname{Ball}(y, \varepsilon)$, the n-dimensional ball with center y and radius ε, we have

$$\hat{\mu}_\lambda(z) = \hat{\mu}_\lambda(y) + P_{V_{I^*}}(z - y). \tag{14}$$

Thus, a direct consequence of our main theorem is given by Corollary 1 below. The latter shows that if y belongs to G_λ, then the number of nonzero coefficients of the solution x^*_λ is an unbiased estimator of the degrees of freedom of the Lasso response.

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

$$\operatorname{div}(\hat{\mu}_\lambda(y)) = |I^*|. \tag{15}$$

Therefore,

$$df(\hat{\mu}_\lambda) = \mathbb{E}(|I^*|). \tag{16}$$

Obviously, in the particular case where the Lasso problem has a unique solution, our result remains true. Precisely, the cardinal of the support of this solution is an unbiased estimator of the degrees of freedom of the Lasso response.

2.2 Reliability of the SURE estimate of the Lasso response

From the estimator of the degree of freedom of the Lasso response $\hat{df}(\hat{\mu}_\lambda)$, it follows that the $\mathrm{SURE}(\hat{\mu}_\lambda)$ (see (6)) is an unbiased estimator of $\mathrm{Risk}(\hat{\mu}_\lambda)$ the true risk, defined by (5). We now evaluate its reliability by computing the expected squared-error between SURE and SE, the true squared-error, that is

$$\mathrm{SE} = \|\hat{\mu}_\lambda - \mu\|_2^2. \tag{17}$$

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

$$\mathbb{E}\left(\left(\mathrm{SURE}(\hat{\mu}_\lambda) - \mathrm{SE}\right)^2\right) = -2\sigma^4 n + 4\sigma^2 \mathbb{E}\left(\|\hat{\mu}_\lambda - y\|_2^2\right) + 4\sigma^4 \mathbb{E}\left(|I^*|\right). \tag{18}$$

Moreover,

$$\mathbb{E}\left(\left(\frac{\mathrm{SURE}(\hat{\mu}_\lambda) - \mathrm{SE}}{n\sigma^2}\right)^2\right) = O\left(\frac{1}{n}\right). \tag{19}$$

3 Comparison with prior work

The authors in [27] studied the degrees of freedom of the Lasso response but in the *overdetermined* case. Precisely, when all the vectors of the design matrix A are linearly independent, which is only possible if $n \geq p$. In other words, they consider that the design matrix A is full rank, that is, $\text{rank}(A) = p$. In fact, in this case the Lasso problem has a unique solution, denoted by $\hat{x}_\lambda$. Thus, before presenting the results of [27], it is necessary to point out a feature on the optimum $\hat{x}_\lambda$ when λ varies from 0 to $+\infty$:

- For $\lambda \geq \|A^t y\|_\infty$, the optimum is attained at $\hat{x}(\lambda) = 0$.
- The interval $]0, \|A^t y\|_\infty[$ can be divided into finite number of subintervals characterized by the fact that within each such subinterval, the support and the sign vector of the optimum of $\mathrm{P}_1(y, \lambda)$ are constant, with respect to λ. Explicitly, let $\{\lambda_m\}$ be the finite sequence of λ's values corresponding to a variation of the support and the sign of $\hat{x}(\lambda)$, defined by

  $$\|A^t y\|_\infty = \lambda_0 > \lambda_1 > \lambda_2 > \cdots > \lambda_K = 0.$$

  Thus, in the interior of the interval $(\lambda_{m+1}, \lambda_m)$, the support and the sign vector of the optimum of (2) are constant with respect to λ, for more details see [5], [16] and [17]. Hence, we call $\{\lambda_m\}$ the *transition points*.

Now, let $\lambda \in (\lambda_{m+1}, \lambda_m)$. Thus, from Lemma 2, we have the following implicit form of $\hat{x}_\lambda$,

$$(\hat{x}_\lambda)_{I_m} = A^+_{I_m} y - \lambda (A^t_{I_m} A_{I_m})^{-1} \mathrm{S}_m, \tag{20}$$

where I_m and S_m are respectively the constant support and the constant vector sign of $\hat{x}_\lambda$ with respect to λ. Hence, based on (20), [27] showed that for all $\lambda > 0$, there exists a set of measure zero $\mathcal{N}_\lambda$, which is a finite collection of hyperplanes in $\mathbb{R}^n$, and they defined

$$\mathcal{K}_\lambda = \mathbb{R}^n \setminus \mathcal{N}_\lambda, \tag{21}$$

so that $\forall\ y \in \mathcal{K}_\lambda$, λ is not any of the transition points, that is, $\lambda \notin \{\lambda_m\}$.
Then, for the overdetermined case, [27] stated that for all $y \in \mathcal{K}_\lambda$, the number of nonzero coefficients of the unique solution of $\mathrm{P}_1(y, \lambda)$ is an unbiased estimator of the degrees of freedom of the Lasso response. In fact, their main argument is that, by eliminating the vectors associated to transition points, the support and the sign of the lasso solution are locally constant with respect to y, see [27, Lemma 5].

We recall that the overdetermined case, considered in [27], is a particular case for which the uniqueness of the solution of the Lasso problem is direct. Thus, according to the Corollary 1, we find the same result as [27] 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 [27] ? 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 unfortunately, the key Lemma 5 in [27] is not true on the set $\mathcal{K}_\lambda$. We prove this by providing a simple counterexample.

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

Let $\{e_1, e_2\}$ an orthonormal basis of $\mathbb{R}^2$ and let's define $a_1 = e_1$ and $a_2 = e_1 + e_2$ and A the matrix which first column is equal to a_1 and which second one is equal to a_2.

Let's define $I = \{1\}$, $j = 2$ and $S = 1$. It turns out that $A^+_I = a_1$ and $\langle (A^+_I)^t S, a_j \rangle = 1$ which implies that for all $\lambda > 0$,

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

Let $y = \alpha a_1$ with $\alpha > 0$, for any $\lambda > 0$, $y \in H_{I,j,S}$ that is $y \notin G_\lambda$. Using lemma 2, one gets that for any $\lambda \in]0, \alpha[$, the solution of $P_1(y, \lambda)$ is $x(\lambda) = (\alpha - \lambda, 0)$ and that for any $\lambda \geqslant \alpha$, $x(\lambda) = (0, 0)$.

Hence the only transition point is $\lambda_0 = \alpha$. It follows that for $\lambda < \alpha$, y belongs to $\mathcal{K}_\lambda$ defined in [27], but $y \notin G_\lambda$.

We prove then that in any ball centered at y, there exists a vector z_1 such that the support of the solution of $P_1(z_1, \lambda)$ is different from the support of $P_1(y, \lambda)$.
Let's choose $\lambda < \alpha$ and $\varepsilon \in]0, \alpha - \lambda[$ and let's define $z_1 = y + \varepsilon e_2$. From lemma 2, one deduces that the solution of $P_1(z_1, \lambda)$ is equal to $x^1(\lambda) = (\alpha - \lambda - \varepsilon, \varepsilon)$ whose support is different from $x(\lambda) = (\alpha - \lambda, 0)$.

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

According to the previous example, 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 $P_1(y, \lambda)$ is not stable to infinitesimal perturbations of y. This situation may occur for under- or over-determined problems. In summary, excluding the set of transition points is not sufficient to guarantee stability of the support of sign of the solution of the Lasso.

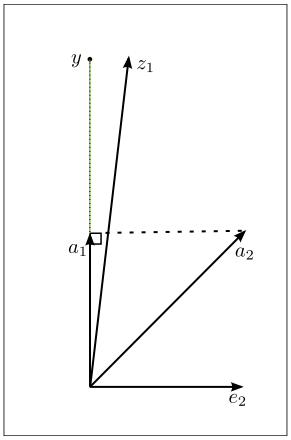


Figure 1: A counter-example for $n = p = 2$ of vectors in G_λ but not in $\mathcal{K}_\lambda$. See text for a detailed discussion.

4 Numerical experiments

In this section, we check the validity of our arguments by some numerical simulations. Indeed, in the previous section, we have introduced $\hat{df}(\hat{\mu}_\lambda)$ an unbiased estimator of the degrees of freedom of the Lasso response $\hat{\mu}_\lambda$. Note that this estimator is at the heart of the SURE, that reads

$$\mathrm{SURE}(\hat{\mu}_\lambda) = -n\sigma^2 + \|\hat{\mu}_\lambda - y\|_2^2 + 2\sigma^2 \hat{df}(\hat{\mu}_\lambda).$$

The SURE is an unbiased estimator of the true risk or the true predictor,

$$\mathbb{E}\{\mathrm{SURE}(\hat{\mu}_\lambda)\} = \mathrm{Risk}\{\hat{\mu}_\lambda\} = \mathbb{E}\,\mathrm{SE} = \mathbb{E}\{\|\hat{\mu}_\lambda - \mu\|_2\}, \text{ where } \mu = Ax^0.$$

Thus, to confirm our main theoretical results it is sufficient to verify numerically the above equality. Here is the outline of these experiments. For our first study, we consider three kinds of simulated design matrix A, Gaussian, partial Fourier and Hadamard, with $n = 1024$ and $p = 4096$, and a deterministic convolution design matrix A, with $n = p = 1024$. For each case, we simulate x^0 the actual parameter vector or the original signal, according to a mixed Gaussian-Bernoulli distribution, such that x^0 has 15 nonzero coefficients. For each design matrix A and vector x^0, we simulate n observations of the linear regression model (1), that is, $y = \mu + \epsilon$, with Ax^0 fixed and $\epsilon \sim N(0, \sigma^2)$. Then, for a given λ, we compute the Lasso response $\hat{\mu}_\lambda$ using the now popular iterative soft-thresholding algorithm, and we calculate the SURE and the SE. After $K = 100$ independent replications, we compute the empirical mean and the standard deviation of $(\mathrm{SURE}_k)_k$

(the sequence of the computed SURE values), the empirical mean of $(\mathrm{SE}_k)_k$ (the sequence of the obtained SE), which corresponds to the computed Risk, and we compute R_T the empirical normalized reliability on the left-hand side of (18),

$$R_T = \frac{1}{K}\sum_{k=1}^{K}\left(\frac{\mathrm{SURE}_k - \mathrm{SE}_k}{n\sigma^2}\right)^2. \tag{22}$$

Moreover, based on the right-hand side of (18), we compute $\hat{R}_T$ as

$$\hat{R}_T = -\frac{2}{n} + \frac{4}{n^2\sigma^2}\left(\frac{1}{K}\sum_{k=1}^{K}\left(\|\left(\hat{\mu}_\lambda\right)_k - y_k\|_2^2\right)\right) + \frac{4}{n^2}\left(\frac{1}{K}\sum_{k=1}^{K}\left(|I^*|_k\right)\right), \tag{23}$$

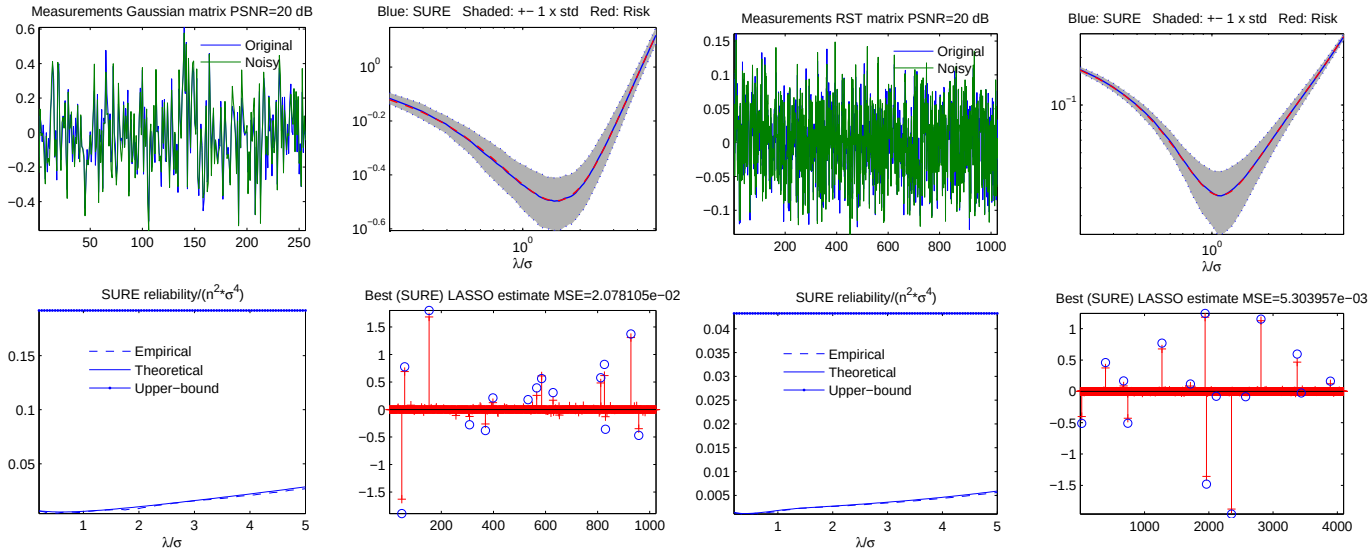
where $\left(\hat{\mu}_\lambda\right)_k$, y_k and $|I^*|_k$ are respectively the response lasso, the observed data, and the cardinal of the support of the lasso solution at the kth replication. Finally, we repeat all this computations for various values of λ, for the four kind of design matrices introduced above.

Figure 2 below shows all obtained results for the four cases. For each kind of design matrix, we associate a panel, which contains four plots. Hence, for each case, from left to right and top to bottom, the first plot represents the observed data without and with noise, that is, the fixed μ and an observation of y. In the second graph, we plot the calculated true Risk curve and the empirical mean of the SURE as a function of the regularization parameter λ. Namely, the red curve represents the calculated true Risk, the blue curve represent the empirical mean of the SURE, and the shaded area represent the empirical mean of the sure $\pm$ the empirical standard deviation of the SURE. The latter shows that the SURE is an unbiased estimator of the true Risk with a controlled variance SURE. This suggests that the SURE is consistent, and then our estimator of the degrees of freedom of the Lasso response is also consistent. In the third graph, we plot the theoretical and empirical normalized reliability, defined respectively by (22) and (23), as a function of the regularization parameter λ. More precisely, the solid and dashed blue curves represent respectively R_T and $\hat{R}_T$, and the horizontal blue line is the upper-bound of the normalized reliability given by right hand term of (52). This confirms numerically that both sides (R_T and $\hat{R}_T$) of (18) indeed coincide.

As discussed in the introduction, 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, we compute the optimal λ that minimizes the SURE (see the second plot), i.e.

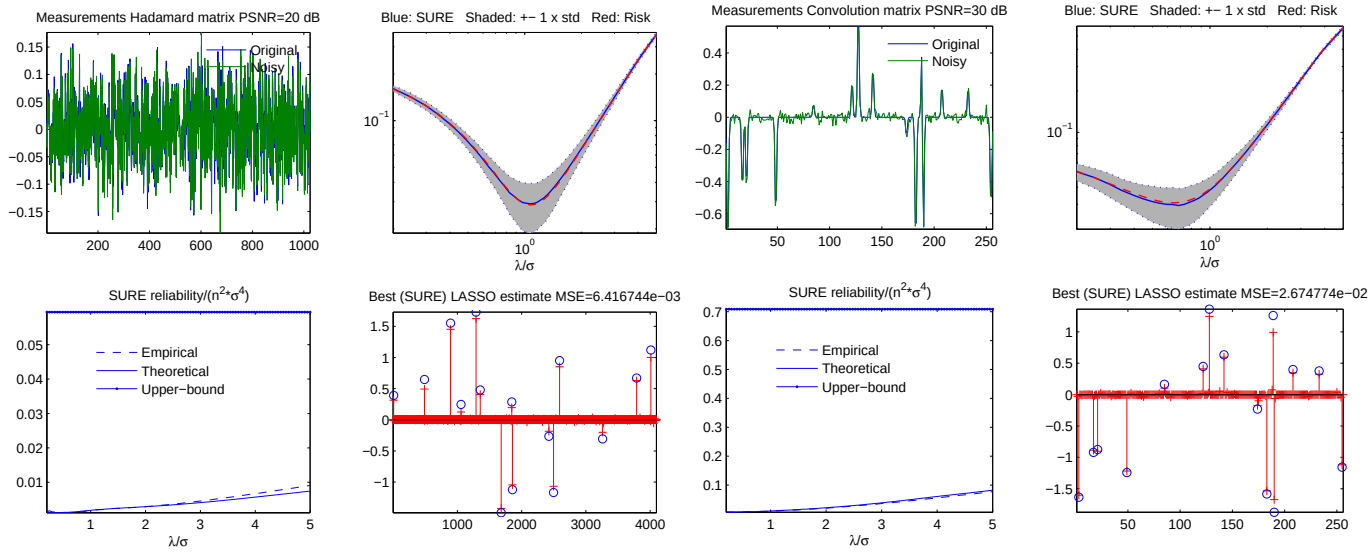
$$\lambda_{\text{optimal}} = \min_{\lambda} \mathrm{SURE}(\hat{\mu}_\lambda). \tag{24}$$

In the fourth graph, we compare the original signal x^0, represented by the blue circles, and the Lasso solution associated to λ_{optimal}, denoted by $\hat{x}_{\lambda_{\text{optimal}}}$ plotted with red crosses. We remark that some coefficients of $\hat{x}_{\lambda_{\text{optimal}}}$ are nonzero outside the support of x^0. This is not a real surprise, since the optimality is in the sense of the prediction variable estimation rather than the regression coefficients.



(a) Gaussian

(b) Random Fourier



(c) Random Hadamard

(d) Convolution

Figure 2: The SURE and its reliability as a function of λ for four types of design matrices. (a) Gaussian; (b) Random Fourier; (c) Random Hadamard; (d) Convolution. For each kind of design matrix, we associate four plots.

Now, for our second simulation study, we fix λ and we consider a partial Fourier design matrix, with $n < p = 4096$. Then, we compute the calculated true Risk curve, the empirical mean of the SURE, the values of the normalized reliability R_T and $\hat{R}_T$, as a function of n. The obtained results are shown in Figure 4. From top to bottom, the first plot displays the empirical mean and standard deviation of the SURE and the true Risk. Unbiasedness is again clear. The second plot confirms again that the SURE is an asymptotically reliable estimate of the risk with the rate established in Theorem 2.

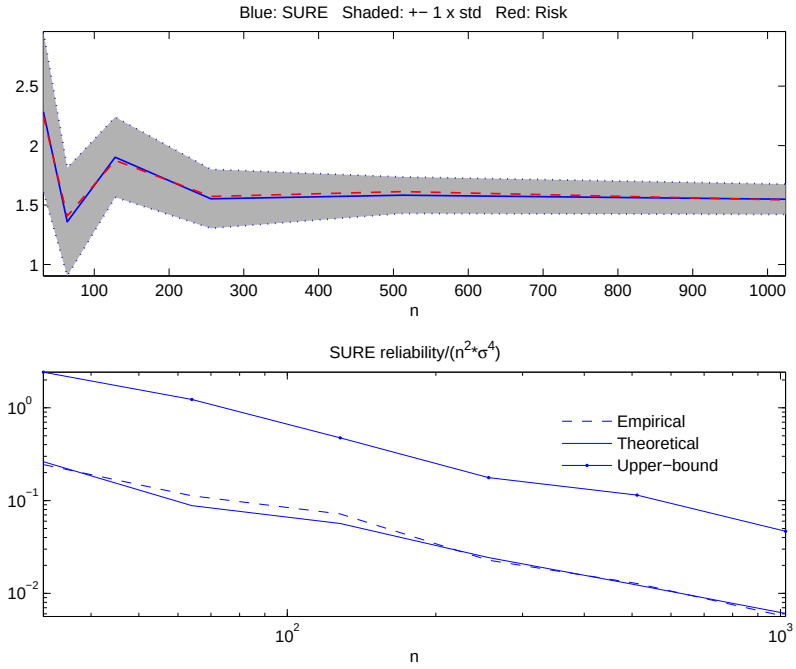


Figure 3: The SURE and its reliability as a function of the number of observations n.

5 Proofs

Before delving into the technical details, let us introduce the following matrix representation of the divergence. Let $\hat{\mu}$ be a function of y and $J_{\hat{\mu}} \equiv \frac{\partial \hat{\mu}}{\partial y}$ be the Jacobian matrix of $\hat{\mu}$, defined as follows

$$(J_{\hat{\mu}})_{i,j} \equiv \left(\frac{\partial \hat{\mu}}{\partial y}\right)_{i,j} = \frac{\partial \hat{\mu}_i}{\partial y_j}, \qquad i,j = 1, \cdots, n. \tag{25}$$

Then we can write

$$\operatorname{div}(\hat{\mu}) = \operatorname{tr}(J_{\hat{\mu}}) \equiv \operatorname{tr}\left(\frac{\partial \hat{\mu}}{\partial y}\right). \tag{26}$$

The above trace expression will be used in our proofs.

Proof of Theorem 1. Recall that x^*_λ is a solution of the Lasso problem $\mathrm{P}_1(y, \lambda)$ and I^* its support. Let $(x^*_\lambda)_{I^*}$ be the restricted vector of x^*_λ into its support, $S^* = \operatorname{sign}((x^*_\lambda)_{I^*})$ and $\hat{\mu}_\lambda(y)$ be the unique Lasso response of $\mathrm{P}_1(y, \lambda)$, see Lemma 3. Here, we have

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

According to Lemma 2, we know that

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

Furthermore, from (9), we get the following implicit form of x^*_λ

$$(x^*_\lambda)_{I^*} = A^+_{I^*} y - \lambda (A^t_{I^*} A_{I^*})^{-1} S^*. \tag{27}$$

It follows that

$$\hat{\mu}_\lambda(y) = P_{V_{I^*}}(y) - \lambda d_{I^*,S^*}, \tag{28}$$

and

$$\hat{r}_\lambda(y) = y - \hat{\mu}_\lambda(y) = P_{V^\perp_{I^*}}(y) + \lambda d_{I^*,S^*}, \tag{29}$$

where $V_{I^*} = \text{span}(a_i)_{i\in I^*}$, $P_{V_{I^*}} = A_{I^*} A^+_{I^*}$ is the orthogonal projection onto V_I, $P_{V^\perp_{I^*}} = I_{n\times n} - P_{V_{I^*}}$ is the orthogonal projection onto the orthogonal complement $V^\perp_{I^*}$ of V_{I^*}, and $d_{I^*,S^*} = (A^+_{I^*})^t S^*$. We define the following set of indices

$$J = \{j : |\langle a_j, \hat{r}_\lambda(y)\rangle| = \lambda\}. \tag{30}$$

From lemma 2 we deduce that

$$I^* \subset J.$$

Since the orthogonal projection is a self-adjoint operator and from (29), for all $j \in J$, we have

$$|\langle P_{V^\perp_{I^*}}(a_j), y\rangle + \lambda \langle a_j, d_{I^*,S^*}\rangle| = \lambda. \tag{31}$$

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

$$a_j \in V_{I^*}, \text{ and then } |\langle a_j, d_{I^*,S^*}\rangle| = 1. \tag{32}$$

In fact, if $a_j \notin V_{I^*}$ then $(I^*, j, S^*) \in \Omega$ and from (31) we have that $y \in H_{I^*,j,S^*}$, which is a contradiction with $y \in G_\lambda$.
Therefore, the finite set of vectors $(a_i)_{i\in I^*}$ forms a basis of $V_J = \text{span}(a_j)_{j\in J}$. Now, suppose that $\bar{x}_\lambda$ is an other solution of $\mathrm{P}_1(y, \lambda)$, such that its support $\bar{I}$ is different than I^*. If $A_{\bar{I}}$ is full rank, then by using the same above arguments we can deduce that $(a_i)_{i\in\bar{I}}$ forms also a basis of V_J. Therefore, we have

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

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

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

So, for any solution $\hat{x}$ of the Lasso problem, we have

$$|\,\text{supp}(\hat{x})| \geq |I^*|,$$

and then $|I^*|$ equals to the minimum of the cardinal's support of solutions of the Lasso problem.

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

$$(x^1_\lambda)_{I^*} = A^+_{I^*} z - \lambda (A^t_{I^*} A_{I^*})^{-1} S^* = (x^*_\lambda)_{I^*} + A^+_{I^*}(z - y). \tag{33}$$

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

$$\text{sign}(x^1_\lambda)_{I^*} = \text{sign}(x^*_\lambda)_{I^*} = S^*. \tag{34}$$

Here, we use Lemma 2 to prove that, for ε small enough, x^1_λ is a solution of $\mathrm{P}_1(z, \lambda)$. First we notice that $z - Ax^1_\lambda = P_{V^\perp_I}(z) + \lambda d_{I^*,S^*}$. It follows that

$$A^t_{I^*}(z - Ax^1_\lambda) = \lambda A^t_{I^*} d_{I^*,S^*} = \lambda S^* = \lambda\, \text{sign}\,(x^1_\lambda)_{I^*}. \tag{35}$$

Moreover for all $j \in J \cap I^*$ from (32), we have that

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

and for all $j \notin J$

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

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

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

Therefore, we obtain

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

So, from Lemma 2, we have that x^1_λ is a solution of $\mathrm{P}_1(z, \lambda)$, and the unique Lasso response associated to $\mathrm{P}_1(z, \lambda)$, denoted by $\hat{\mu}_\lambda(z)$, is defined by

$$\hat{\mu}_\lambda(z) = P_{V_{I^*}}(z) - \lambda d_{I^*,S^*}. \tag{36}$$

Therefore, from (28) and (36), we can deduce that for all $z \in \text{Ball}(y, \varepsilon)$ we have

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

□

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

$$\|z - y\|_2 \leq \varepsilon \Rightarrow \hat{\mu}_\lambda(z) = \hat{\mu}_\lambda(y) + P_{V_{I^*}}(z - y). \tag{37}$$

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

$$\|\hat{\mu}_\lambda(z) - \hat{\mu}_\lambda(y)\|_2 = \|P_{V_{I^*}}(h)\|_2 = \|h\|_2 \leq \varepsilon. \tag{38}$$

Therefore, the Lasso response $\hat{\mu}_\lambda(y)$ is uniformly Lipschitz on G_λ. Moreover, $\hat{\mu}_\lambda(y)$ is a continuous function of y, and thus $\hat{\mu}_\lambda(y)$ is uniformly Lipschitz on $\mathbb{R}^n$. Hence, $\hat{\mu}_\lambda(y)$ is almost differentiable; see Meyer and Woodroofe [14] and Efron et al. [5].

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 $\mathrm{P}_1(z, \lambda)$, which has the same support and the same sign of x^*_λ, and thus $\hat{\mu}_\lambda(z)$ belongs to the vector space V_{I^*}, whose dimension equals to $|I^*|$, see (28) and (36). Therefore, $\hat{\mu}_\lambda(y)$ is a locally affine function of y, and then

$$J_{\hat{\mu}_\lambda(y)} = \frac{\partial \hat{\mu}_\lambda(y)}{\partial y} = P_{V_{I^*}} \tag{39}$$

Then the trace formula (26) implies that

$$\operatorname{div}(\hat{\mu}_\lambda(y)) = \operatorname{tr}(P_{V_{I^*}}) = |I^*|. \tag{40}$$

This holds almost everywhere since G_λ is of full measure, and (16) is obtained by invoking Stein's lemma. □

Proof of Theorem 2. First, consider the following random variable

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

From Lemma 1, we have

$$\mathbb{E}\langle \varepsilon, \hat{\mu}_\lambda\rangle = \sigma^2 \mathbb{E} \operatorname{div}(\hat{\mu}_\lambda).$$

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

$$\mathbb{E}\,\mathrm{SURE}(\hat{\mu}_\lambda) = \mathbb{E}Q_1(\hat{\mu}_\lambda) = \mathbb{E}\{\mathrm{SE}\} = \mathrm{Risk}(\hat{\mu}_\lambda).$$

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

$$\mathbb{E}\left(\|y\|_2^2\right) = n\sigma^2 + \|\mu\|_2^2, \text{ and } \mathbb{V}\{\|y\|_2^2\} = 2\sigma^4\left(n + 2\frac{\|\mu\|_2^2}{\sigma^2}\right). \tag{41}$$

Now, we remark also that

$$Q_1(\hat{\mu}_\lambda) - \mathrm{SE} = 2\left(\sigma^2\,\mathrm{div}(\hat{\mu}_\lambda) - \langle\varepsilon, \hat{\mu}_\lambda\rangle\right). \tag{42}$$

After an elementary calculation, we obtain

$$\mathbb{E}(\mathrm{SURE}(\hat{\mu}_\lambda) - \mathrm{SE})^2 = \mathbb{E}(Q_1(\hat{\mu}_\lambda) - \mathrm{SE})^2 + \mathbb{V}\{\|y\|_2^2\} + 4T, \tag{43}$$

where

$$T = \sigma^2\mathbb{E}\,\mathrm{div}(\hat{\mu}_\lambda)\|y\|_2^2 - \mathbb{E}\langle\varepsilon, \hat{\mu}_\lambda\rangle\|y\|_2^2 = T_1 + T_2, \tag{44}$$

with

$$T_1 = 2\left(\sigma^2\mathbb{E}\,\mathrm{div}(\hat{\mu}_\lambda)\langle\varepsilon, \mu\rangle - \mathbb{E}\langle\varepsilon, \hat{\mu}_\lambda\rangle\langle\varepsilon, \mu\rangle\right) \tag{45}$$

and

$$T_2 = \sigma^2\mathbb{E}\,\mathrm{div}(\hat{\mu}_\lambda)\|\varepsilon\|_2^2 - \mathbb{E}\langle\varepsilon, \hat{\mu}_\lambda\rangle\|\varepsilon\|_2^2. \tag{46}$$

Hence, by using the fact that a gaussian probability density $f(\varepsilon_i)$ satisfies $\varepsilon_i f(\varepsilon_i) = -\sigma^2 f^{'}(\varepsilon_i)$ and integrations by parts, we find that

$$T_1 = -2\sigma^2\mathbb{E}\langle\hat{\mu}_\lambda, \mu\rangle$$

and

$$T_2 = -2\sigma^4\mathbb{E}\left[\mathrm{div}(\hat{\mu}_\lambda)\right].$$

It follows that

$$T = -2\sigma^2\left(\mathbb{E}\langle\hat{\mu}_\lambda, \mu\rangle + \sigma^2\mathbb{E}\,\mathrm{div}(\hat{\mu}_\lambda)\right). \tag{47}$$

Moreover, from [[12], Property 1 page 25], we know that

$$\mathbb{E}(Q_1(\hat{\mu}_\lambda) - \mathrm{SE})^2 = 4\sigma^2\left(\mathbb{E}\left[\|\hat{\mu}_\lambda\|_2^2\right] + \sigma^2\mathbb{E}\left[\mathrm{tr}\{(J_{\hat{\mu}_\lambda})^2\}\right]\right), \tag{48}$$

where $J_{\hat{\mu}_\lambda} = \left(\dfrac{\partial\,(\hat{\mu}_\lambda)_i}{\partial y_j}\right)_{1\le i,j\le n}$ is the Jacobian matrix of $\hat{\mu}_\lambda$. Thus, since $J_{\hat{\mu}_\lambda} = P_{V_{I^*}}$ which is a self-adjoint projection, we have $(J_{\hat{\mu}_\lambda})^2 = J_{\hat{\mu}_\lambda}$, and $\mathrm{tr}\,(J_{\hat{\mu}_\lambda}) = \mathrm{div}(\hat{\mu}_\lambda) = |I^*|$. Therefore, we get

$$\mathbb{E}(Q_1(\hat{\mu}_\lambda) - \mathrm{SE})^2 = 4\sigma^2\left(\mathbb{E}\left(\|\hat{\mu}_\lambda\|_2^2\right) + \sigma^2\mathbb{E}\left(|I^*|\right)\right). \tag{49}$$

Furthermore, observe that

$$\mathbb{E}\,\mathrm{SURE}(\hat{\mu}_\lambda) = -n\sigma^2 + \mathbb{E}\left(\|\hat{\mu}_\lambda - y\|_2^2\right) + 2\sigma^2\mathbb{E}\left(|I^*|\right). \tag{50}$$

Therefore, by combining (41), (43), (47) and (49), we obtain

$$\begin{aligned}
\mathbb{E}(\mathrm{SURE}(\hat{\mu}_\lambda) - \mathrm{SE})^2 &= 2n\sigma^4 + 4\sigma^2\mathbb{E}\,\mathrm{SE} - 4\sigma^4\mathbb{E}\left(|I^*|\right) \\
&= 2n\sigma^4 + 4\sigma^2\mathbb{E}\left(\mathrm{SURE}(\hat{\mu}_\lambda)\right) - 4\sigma^4\mathbb{E}\left(|I^*|\right) \\
\text{(by using (50))} &= -2n\sigma^4 + 4\sigma^2\mathbb{E}\left(\|\hat{\mu}_\lambda - y\|_2^2\right) + 4\sigma^4\mathbb{E}\left(|I^*|\right).
\end{aligned}$$

On the other hand, since x_λ^* is an optimum of the Lasso problem $\mathrm{P}_1(y, \lambda)$, we observe that

$$\frac{1}{2}\|\hat{\mu}_\lambda - y\|_2^2 \le \frac{1}{2}\|\hat{\mu}_\lambda - y\|_2^2 + \lambda\|x_\lambda^*\|_1 \le \frac{1}{2}\|A.0 - y\|_2^2 + \lambda\|0\|_1 = \frac{1}{2}\|y\|_2^2.$$

Therefore, we have

$$\mathbb{E}\left(\|\hat{\mu}_\lambda - y\|_2^2\right) \leq \mathbb{E}\left(\|y\|_2^2\right) = n\sigma^2 + \|\mu\|_2^2. \tag{51}$$

Then, since $|I^*| = o(n)$ and from (51), we have

$$\mathbb{E}\left(\left(\frac{\mathrm{SURE}(\hat{\mu}_\lambda) - \mathrm{SE}}{n\sigma^2}\right)^2\right) \leq \frac{6}{n} + \frac{4\|\mu\|_2^2}{n^2\sigma^2}. \tag{52}$$

Finally, since $\|\mu\|_2^2 < +\infty$, we can deduce that

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

□

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 case. 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 the latter turns out to be larger than the set of all the vectors associated to the transition points considered in [27] in the overdetermined case. We also highlight the fact the set of transition points is not sufficient for the divergence formula to hold, hence providing a counterexample to some of the key results in [27].

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 extend it to other penalties such as those promoting structured sparsity, e.g. the group Lasso.

References

[1] Akaike, H. (1973). Information theory and an extension of the maximum liklihood principle. Second International Symposium on Information Theory 267-281.

[2] Craven, p. and Wahba, G. Smoothing Noisy Data with Spline Functions: estimating the correct degree of smoothing by the method of generalized cross validation. Numerische Mathematik 31, 377-403.

[3] Dossal, C (2007). A necessary and sufficient condition for exact recovery by l1 minimization. Technical report, HAL-00164738:1.

[4] Efron, B. (2004). The estimation of prediction error: Covariance penalties and cross-validation (with discussion). J. Amer. Statist. Assoc. 99 619-642.

[5] Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004). Least angle regression (with discussion). Ann. Statist. 32 407-499.

[6] Efron, B. (1981). How biased is the apparent error rate of a prediction rule. J. Amer. Statist. Assoc. vol. 81 pp. 461-470.

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

[8] Fan, J. and Peng, H. (2004). Nonconcave penalized likelihood with a diverging number of parameters. The Annals of Statistics, 32(3), 928-961.

[9] Fletcher, R. Practical Methods of Optimization. John Wiley and Sons, Inc., 2nd edition, 1987.

[10] Fuchs, J. J. (2004). On sparse representations in arbitrary redundant bases. IEEE Trans. Inform. Theory, vol. 50, no. 6, pp. 1341-1344.

[11] Kato, K. (2009). On the degrees of freedom in shrinkage estimation. Journal of Multivariate Analysis 100(7), 1338-1352.

[12] Luisier, F. (2009). The SURE-LET approach to image denoising. Ph.D. dissertation, EPFL, Lausanne. Available: http://library.epfl.ch/theses/?nr=4566.

[13] Mallows, C. (1973). Some comments on C_p. Technometrics 15, 661-675.

[14] Meyer, M. and Woodroofe, M. (2000). On the degrees of freedom in shaperestricted regression. Ann. Statist. 28 1083-1104

[15] Nardi, Y. and Rinaldo, A (2008). On the asymptotic properties of the group Lasso estimator for linear models. Electronic Journal of Statistics, 2: 605-633.

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

[17] Osborne, M. R., Presnell, B. and Turlach, B. (2000b). On the LASSO and its dual. J. Comput. Graph. Statist. 9 319-337.

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

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

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

[21] Tibshirani, R. and Taylor, J. (2011). The Solution Path of the Generalized Lasso. Annals of Statistics. In Press.

[22] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. J. Roy. Statist. Soc. Ser. B 58(1) 267-288.

[23] Tropp J. A. (2006). Just relax: convex programming methods for identifying sparse signals in noise, IEEE Trans. Info. Theory 52 (3), 1030-1051.

[24] Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. J. Roy. Statist. Soc. Ser. B 68 49-67.

[25] Zhao, P. and Bin, Y. (2006). On model selection consistency of Lasso. Journal of Machine Learning Research, 7, 2541-2563.

[26] Zou, H. (2006). The adaptive Lasso and its oracle properties. Journal of the American Statistical Association, 101(476), 1418-1429

[27] Zou, H., Hastie, T. and Tibshirani, R. (2007). On the "degrees of freedom" of the Lasso. Ann. Statist. Vol. 35, No. 5. 2173-2192.