The degrees of freedom of the Lasso for general design matrix

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

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) IMB, CNRS-Univ. Bordeaux 1
351 Cours de la Libération, F-33405 Talence, France
Charles.Dossal@math.u-bordeaux1.fr

(2) GREYC, CNRS-ENSICAEN-Univ. Caen
6 Bd du Maréchal Juin, 14050 Caen, France
Jalal.Fadili@greyc.ensicaen.fr
Maher.Kachour@greyc.ensicaen.fr

(3) Ceremade, CNRS-Univ. Paris-Dauphine
Place du Maréchal De Lattre De Tassigny, 75775 Paris 16, France
Gabriel.Peyre@ceremade.dauphine.fr

(4) LMNO, CNRS-Univ. Caen
Département de Mathématiques, UFR de Sciences, 14032 Caen, France
Chesneau.Christophe@math.unicaen.fr

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 cardinality 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 [30]. We also show, by providing a simple counterexample, that although the df theorem of [30] 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. as we illustrate in some numerical simulations.

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.

Let $\hat{x} = \delta(y)$ be an estimator of x^0, and $\hat{\mu} = A\hat{x}$ be the response or the predictor associated to $\hat{x}$. The concept of degrees of freedom plays a pivotal role in 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 [7], the degrees of freedom of the response $\hat{\mu}$ is defined by

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

Many model selection criteria involve $df(\hat{\mu})$, e.g. C_p (Mallows [13]), AIC (Akaike Information Criterion, [1]), BIC (Bayesian Information Citerion, [20]), GCV (Generalized Cross Validation, [2]) and SURE (Stein's unbiased risk estimation, see Section 2.2). Thus, the degrees of freedom is a quality of interest in model validation and selection and it intervenes also to finding the optimal hyperparameters of the estimator. Note that, the optimality here is in the sense of the prediction $\hat{\mu}$ and not the coefficients $\hat{x}$.

The well-known Stein's lemma [21], ensures that if $\hat{\mu}$ is continuous and almost differentiable then its divergence is an unbiased estimator of its degrees of freedom, i.e.

$$\widehat{df}(\hat{\mu}) = \mathrm{div}\, \hat{\mu} = \sum_{i=1}^{n} \frac{\partial \hat{\mu}_i}{\partial y_i}. \tag{3}$$

In order to estimate x^0, we consider the least absolute shrinkage and selection operator (Lasso) procedure, proposed originally by Tibshirani [24]. The Lasso estimate amounts to solving the following convex optimization problem

$$\mathrm{P}_1(y, \lambda) : \min_{x\in\mathbb{R}^p} \frac{1}{2}\|y - Ax\|_2^2 + \lambda\|x\|_1, \tag{4}$$

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 efficiently 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., [8, 9, 28, 29, 18, 15, 16, 6, 10, 25], to name just a few.

1.2 Contributions and related 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 (4). The main contribution of this paper is first to generalize the results of [30] 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 (4) has a unique solution, and in particular for the overdetermined case studied in [30]. Additionally, using the estimator at hand, we establish the reliability of the SURE for the Lasso.

While this paper was submitted, we became aware of the independent work of Tibshirani and Taylor [23], who studied the degrees on freedom 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 [30, Theorem 1] for the overdetermined case, and that of [23] and [26] for the general case.

1.3 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 discuss relation of our work with concurrent one in literature 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*

First, some notations and definitions are necessary. Let $x \in \mathbb{R}^p$, where x_i denotes the ith component of x. The support or the active set of x is defined by

$$I = \operatorname{supp}(x) = \{i : x_i \neq 0\},$$

and we denote its cardinality as $|\operatorname{supp}(x)| = |I|$. Moreover, we denote by x_I the reduced dimensional vector built upon the non-zero components of x. The active matrix A_I associated to a vector x is obtained by selecting the columns of A indexed by the support I of x. Let A_I^t be the transpose matrix of A_I. Suppose that A_I is full column rank, then the Moore-Penrose pseudo-inverse $(A_I^t A_I)^{-1} A_I^t$ of A_I is denoted A_I^+. $\operatorname{sign}(\cdot)$ represents the sign function: $\operatorname{sign}(a) = 1$ if $a > 0$; $\operatorname{sign}(a) = 0$ if $a = 0$; $\operatorname{sign}(a) = -1$ if $a < 0$. Let $\operatorname{sign}(x)$ be the sign vector of x, such that $\operatorname{sign}(x)_i = \operatorname{sign}(x_i)$.
Let now $I \subseteq \{1, 2, \cdots, p\}$, such that $A_I = (a_i)_{i \in I}$ is full column rank. We denote the cardinality 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

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

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

$$H_{I,j,S} = \{u \in \mathbb{R}^n : \langle P_{V_I^\perp}(a_j), u\rangle = \pm\lambda(1 - \langle a_j, (A_I^+)^t S\rangle)\}. \tag{5}$$

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

$$\Omega = \{(I,j,S) : a_j \notin V_I\} \tag{6}$$

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

$$G_\lambda = \mathbb{R}^n \setminus \bigcup_{(I,j,S)\in\Omega} H_{I,j,S}. \tag{7}$$

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,*

$$|I^*| = \min_{x_\lambda \in \mathcal{M}_{y,\lambda}} |\operatorname{supp}(x_\lambda)|. \tag{8}$$

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 $\hat{\mu}_\lambda(\cdot)$ satisfies

$$\hat{\mu}_\lambda(z) = \hat{\mu}_\lambda(y) + P_{V_{I^*}}(z-y). \tag{9}$$

As stated, this theorem assumes the existence of a solution whose active matrix A_{I^*} is full rank. Using a constructive proof, this can be shown to be true; see e.g. [4, Proof of Theorem 1] or [19, Theorem 3, Section B.1]. It is worth noting that this proof can be built on to construct a solution x^*_λ of $\mathrm{P}_1(y,\lambda)$ such that A_{I^*} is full rank from any solution x_λ whose active matrix has a nontrivial kernel.

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*

$$\mathrm{div}(\hat{\mu}_\lambda(y)) = |I^*|. \tag{10}$$

Therefore,

$$df(\hat{\mu}_\lambda) = \mathbb{E}(|I^*|). \tag{11}$$

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

In this work, we are interested in particular by the SURE as a model selection criteria. Indeed, suppose that $\hat{\mu}$ is an estimator of $\mu = Ax^0$, and the degrees of freedom has an unbiased estimator, denoted by $\widehat{df}(\hat{\mu})$, the SURE is defined as follows

$$\mathrm{SURE}(\hat{\mu}) = -n\sigma^2 + \|\hat{\mu} - y\|_2^2 + 2\sigma^2 \widehat{df}(\hat{\mu}). \tag{12}$$

It is well-known that the SURE is an unbiased estimate of the true risk, i.e.

$$\mathrm{Risk}(\hat{\mu}) = \mathbb{E}\left(\|\hat{\mu} - \mu\|_2^2\right) = \mathbb{E}\left(\mathrm{SURE}(\hat{\mu})\right).$$

In the previous section, we gave an unbiased estimator of the degree of freedom $\widehat{df}(\hat{\mu}_\lambda)$ of the Lasso response $\hat{\mu}_\lambda$. Consequently, the $\mathrm{SURE}(\hat{\mu}_\lambda)$ is an unbiased estimator of the true risk $\mathrm{Risk}(\hat{\mu}_\lambda)$. We now evaluate its reliability by computing the expected squared-error between SURE and SE, the true squared-error, that is

$$\mathrm{SE} = \|\hat{\mu}_\lambda - \mu\|_2^2. \tag{13}$$

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

$$\mathbb{E}\left(\left(\mathrm{SURE}(\hat{\mu}_\lambda) - \mathrm{SE}\right)^2\right) = -2\sigma^4 n + 4\sigma^2 \mathbb{E}\left(\|\hat{\mu}_\lambda - y\|_2^2\right) + 4\sigma^4 \mathbb{E}\left(|I^*|\right). \tag{14}$$

Moreover,

$$\mathbb{E}\left(\left(\frac{\mathrm{SURE}(\hat{\mu}_\lambda) - \mathrm{SE}}{n\sigma^2}\right)^2\right) = O\left(\frac{1}{n}\right). \tag{15}$$

3 Relation to prior work

Relation to [30]

The authors in [30] 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, $\mathrm{rank}(A) = p$. In fact, in this case the Lasso problem has a unique solution, denoted by $\hat{x}_\lambda$. Thus, before presenting the results of [30], 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 (4) are constant with respect to λ, for more details see [6], [16] and [17]. Hence, we call $\{\lambda_m\}$ the *transition points*.

Now, let $\lambda \in (\lambda_{m+1}, \lambda_m)$. Thus, from Lemma 1 (see Section 5), we have the following implicit form of $\hat{x}_\lambda$,

$$(\hat{x}_\lambda)_{I_m} = A^+_{I_m} y - \lambda (A^t_{I_m} A_{I_m})^{-1} \mathrm{S}_m, \tag{16}$$

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 (16), [30] 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

$$\mathcal{K}_\lambda = \mathbb{R}^n \setminus \mathcal{N}_\lambda, \tag{17}$$

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, [30] 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 [30, Lemma 5].

We recall that the overdetermined case, considered in [30], 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 [30] 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 [30] ? 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 *df* formula [30, Theorem 1] is correct, unfortunately, their proof contains a flaw since their divergence formula [30, Lemma 5] 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 $\{e_1, e_2\}$ 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 $\lambda > 0$,

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

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 1 (see Section 5), 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 [30], 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 1 (see Section 5), 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.

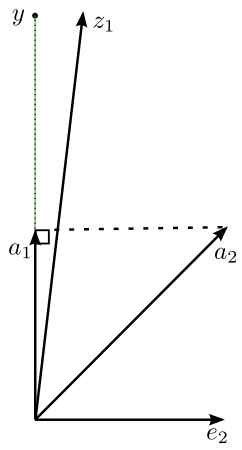


Figure 1: A counterexample for $n = p = 2$ of vectors in G_λ but not in $\mathcal{K}_\lambda$. See text for a detailed discussion.

Relation to [23] and [26]

In [23, Theorem 2], the authors proved that

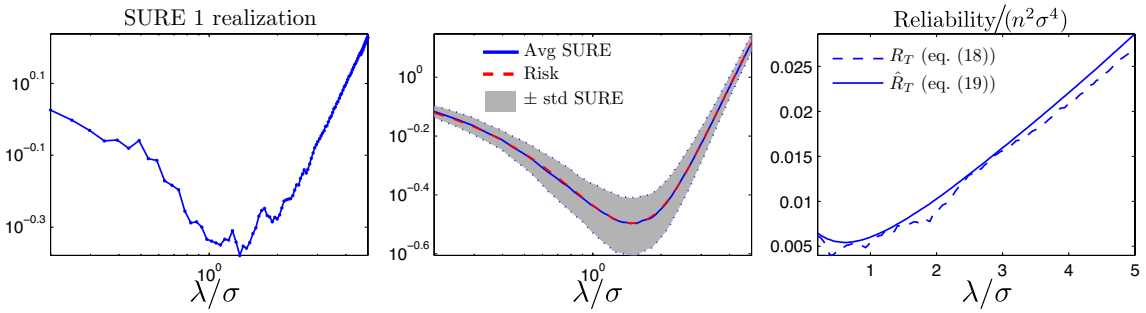
$$df(\hat{\mu}_\lambda) = \mathbb{E}(\text{rank}(A_I))$$

where $I = I(y)$ is the active set of any solution $x_\lambda(y)$ to $\mathrm{P}_1(y, \lambda)$. This coincides with Corollary 1 when A_I is full rank with $\text{rank}(A_I) = \text{rank}(A_{I^*})$. Note that in general, there exist vectors $y \in \mathbb{R}^n$ where the smallest cardinality among all active sets of Lasso solutions is different from the rank of the active matrix associated to the largest support. But these vectors are those excluded in G_λ. In the case of the generalized Lasso (a.k.a. analysis sparsity prior in the signal processing community), Vaiter et al. [26, Corollary 1] and Tibshirani and Taylor [23, Theorem 3] provide a formula of the df. This formula reduces to that of Corollary 1 when the analysis operator is the identity.

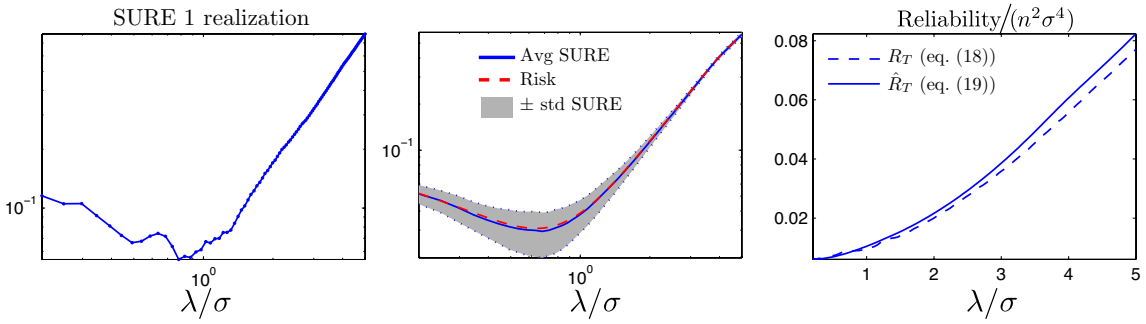
4 Numerical experiments

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. Here is the outline of these experiments.

For our first study, we consider two kinds of simulated design matrices A, a random Gaussian matrix with $n = 256$ and $p = 1024$ whose entries are $\sim_{\text{iid}} \mathcal{N}(0, 1/n)$, and a deterministic convolution design matrix A with $n = p = 256$ and a Gaussian blurring function. 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 $\epsilon \sim \mathcal{N}(0, \sigma^2)$. Then, for a given λ, we compute the Lasso response $\hat{\mu}_\lambda$ using the now popular iterative soft-thresholding algorithm [3], and we calculate the SURE and the SE. After $K = 100$



(a) Gaussian



(b) Convolution

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

independent replications, we compute the empirical mean and the standard deviation of $(\text{SURE}_k)_k$ (the sequence of the computed SURE values), the empirical mean of $(\text{SE}_k)_k$ (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 (14),

$$R_T = \frac{1}{K}\sum_{k=1}^{K}\left(\frac{\text{SURE}_k - \text{SE}_k}{n\sigma^2}\right)^2. \tag{18}$$

Moreover, based on the right-hand side of (14), we compute $\hat{R}_T$ as

$$\hat{R}_T = -\frac{2}{n} + \frac{4}{n^2\sigma^2}\left(\frac{1}{K}\sum_{k=1}^{K}\left(\|\left(\hat{\mu}_\lambda\right)_k - y_k\|_2^2\right)\right) + \frac{4}{n^2}\left(\frac{1}{K}\sum_{k=1}^{K}\left(|I^*|_k\right)\right), \tag{19}$$

where $(\hat{\mu}_\lambda)_k$, y_k and $|I^*|_k$ are respectively the response Lasso, the observed data, and the cardinality of the support of the Lasso solution at the kth replication. Finally, we repeat all these computations for various values of λ, for the two kinds of design matrices introduced above.

Figure 2 depicts the obtained results. For each kind of design matrix, we associate a panel, which contains three plots. Hence, for each case, from left to right, the first plot represents the SURE for one realization of the noise as a function of λ. In the second graph, we plot the computed true Risk curve and the empirical mean of the SURE as a function of the regularization parameter λ. Namely, the dashed (red) curve represents the calculated true Risk, the solid (blue) curve represents 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. This suggests that the SURE is consistent, and then our estimator of the degrees of freedom of the Lasso response is also consistent.

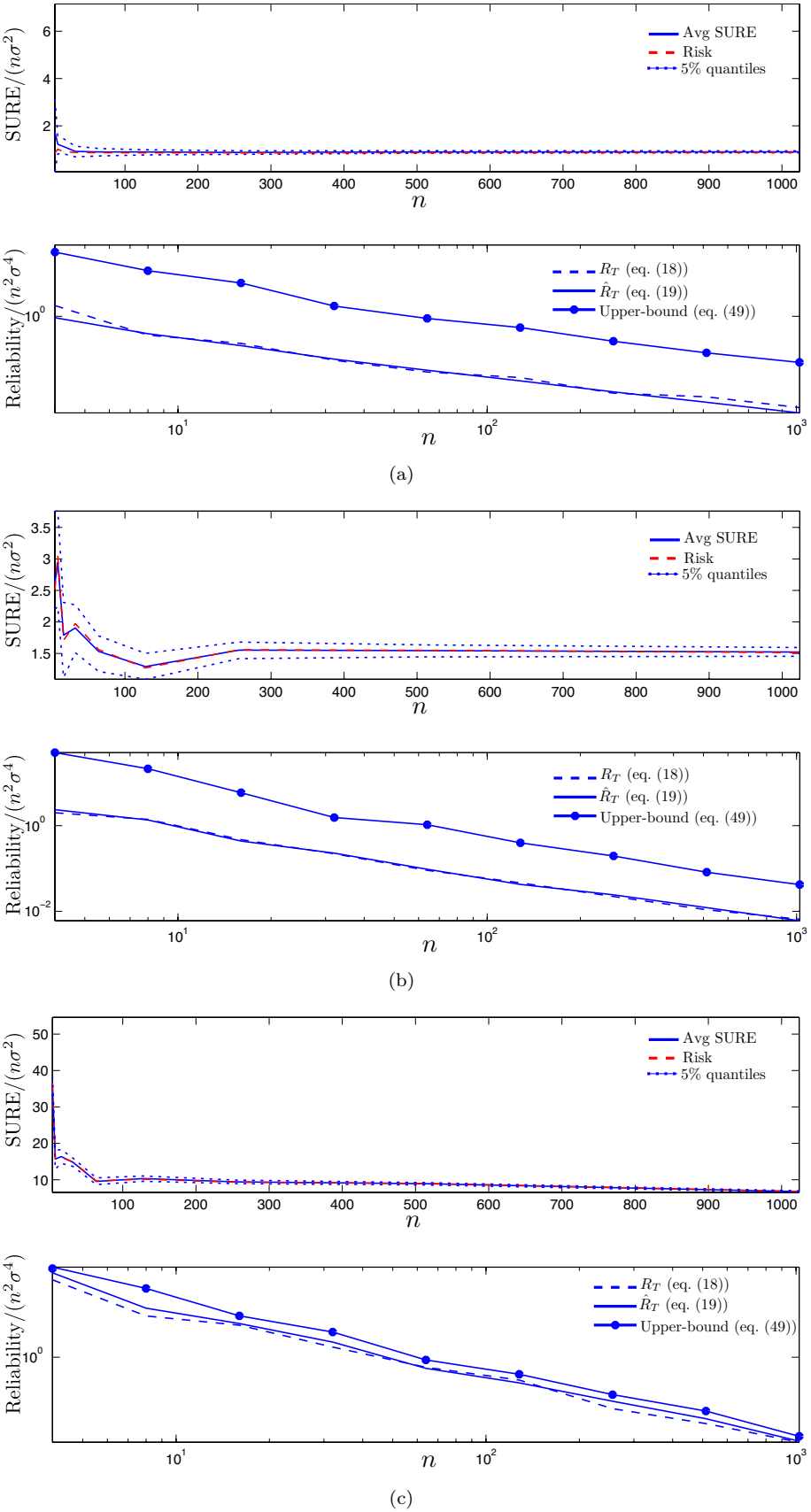


Figure 3: The SURE and its reliability as a function of the number of observations n. (a), (b) and (c) correspond respectively to $\lambda/\sigma = 0.1$ (small), 1 (medium) and 10 (large).

In the third graph, we plot the theoretical and empirical normalized reliability, defined respectively by (18) and (19), 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 the right hand term of (49). This confirms numerically that both sides (R_T and $\hat{R}_T$) of (14) 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.

$$\lambda_{\text{optimal}} = \underset{\lambda>0}{\operatorname{argmin}} \operatorname{SURE}(\hat{\mu}_\lambda). \tag{20}$$

This optimal value was found by a golden section search as $\operatorname{SURE}(\hat{\mu}_\lambda)$ turns out to be unimodal.

Now, for our second simulation study, we consider a partial Fourier design matrix, with $n < p = 4096$. For each value of three values of $\lambda/\sigma \in \{0.1, 1, 10\}$ (small, medium and large), we compute the Risk curve, the empirical mean of the SURE, as well as the values of the normalized reliability R_T and $\hat{R}_T$, as a function of $n \in \{4, \cdots, 1024\}$. The obtained results are shown in Figure 3. From each value of λ, the first plot (top panel) displays the empirical mean of the SURE (solid line) and its 5% quantiles (dotted) as well as the computed Risk (dashed). Unbiasedness is again clear whatever the value of λ. The second plot confirms that the SURE is an asymptotically reliable estimate of the risk with the rate established in Theorem 2. Moreover, as expected, the actual reliability gets closer to the upper-bound (49) as the number of samples n increases.

5 Proofs

First of all, we recall some classical properties of the Lasso solution (see, e.g., [16, 6, 10, 25]).

Lemma 1. *A necessary and sufficient condition for $\tilde{x}$ to be a minimizer of the Lasso problem $P_1(y, \lambda)$, defined by (4), is that $\tilde{x}$ satisfies the two following conditions:*

1. $A_I^t(y - A\tilde{x}) = \lambda sign(\tilde{x}_I)$, *i.e.* $\langle a_k, y - A\tilde{x}\rangle = \lambda \operatorname{sign}(\tilde{x}_k)$, $\forall\ k \in I = \{i : \tilde{x}_i \neq 0\}$,
2. $|\langle a_j, y - A\tilde{x}\rangle| \leq \lambda$, $\forall\ j \in I^c$,

where I^c is the complement of I, $A_I = (a_i)_{i\in I}$, A_I^t its transpose, and $\operatorname{sign}(\cdot)$ *represents the sign function. Moreover, if A_I is full rank, then $\tilde{x}$ satisfies the following implicit relationship:*

$$\tilde{x}_I = A_I^+ y - \lambda (A_I^t A_I)^{-1} \operatorname{sign}(\tilde{x}_I)\ . \tag{21}$$

Note that if the inequality in the condition 2 is strict, then $\tilde{x}$ is the unique minimizer of the Lasso problem $\mathrm{P}_1(y, \lambda)$. Lemma 2 below shows that all the solutions of $\mathrm{P}_1(y, \lambda)$ have the same image by A. In others words, the Lasso response, denoted by $\hat{\mu}_\lambda(y)$, is unique, see [4].

Lemma 2. *If $\tilde{x}_1$ and $\tilde{x}_2$ are solutions of $P_1(y, \lambda)$, defined by (4), then*

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

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

$$(J_{\hat{\mu}})_{i,j} \equiv \left(\frac{\partial \hat{\mu}}{\partial y}\right)_{i,j} = \frac{\partial \hat{\mu}_i}{\partial y_j}, \qquad i, j = 1, \cdots, n. \tag{22}$$

Then we can write

$$\operatorname{div}(\hat{\mu}) = \operatorname{tr}(J_{\hat{\mu}}) \equiv \operatorname{tr}\left(\frac{\partial \hat{\mu}}{\partial y}\right). \tag{23}$$

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 2. Here, we have

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

According to Lemma 1, 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 (21), we get the following implicit form of x^*_λ

$$(x^*_\lambda)_{I^*} = A^+_{I^*} y - \lambda (A^t_{I^*} A_{I^*})^{-1} S^*. \tag{24}$$

It follows that

$$\hat{\mu}_\lambda(y) = P_{V_{I^*}}(y) - \lambda d_{I^*,S^*}, \tag{25}$$

and

$$\hat{r}_\lambda(y) = y - \hat{\mu}_\lambda(y) = P_{V^\perp_{I^*}}(y) + \lambda d_{I^*,S^*}, \tag{26}$$

where $V_{I^*} = \operatorname{span}(a_i)_{i\in I^*}$, $P_{V_{I^*}} = A_{I^*}A^+_{I^*}$ is the orthogonal projection onto V_I, $P_{V^\perp_{I^*}} = \mathrm{Id}_{n\times n} - P_{V_{I^*}}$ 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

$$J = \{j : |\langle a_j, \hat{r}_\lambda(y)\rangle| = \lambda\}. \tag{27}$$

From lemma 1 we deduce that

$$I^* \subset J.$$

Since the orthogonal projection is a self-adjoint operator and from (26), for all $j \in J$, we have

$$|\langle P_{V^\perp_{I^*}}(a_j), y\rangle + \lambda \langle a_j, d_{I^*,S^*}\rangle| = \lambda. \tag{28}$$

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

$$a_j \in V_{I^*}, \text{ and then } |\langle a_j, d_{I^*,S^*}\rangle| = 1. \tag{29}$$

In fact, if $a_j \notin V_{I^*}$ then $(I^*, j, S^*) \in \Omega$ and from (28) 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 $V_J = \operatorname{span}(a_j)_{j\in J}$. Now, suppose that $\bar{x}_\lambda$ is an other solution of $\mathrm{P}_1(y,\lambda)$, such that its support $\bar{I}$ is different from 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

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

and then $|I^*|$ is equal to the minimum of the cardinalities of the supports of solutions of $\mathrm{P}_1(y, \lambda)$.

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 $\mathrm{Ball}(y, \varepsilon) \subsetneq G_\lambda$, that is, for all $z \in \mathrm{Ball}(y, \varepsilon)$, $z \in G_\lambda$. Now, let x^1_λ be the vector supported in I^* and defined by

$$(x^1_\lambda)_{I^*} = A^+_{I^*} z - \lambda (A^t_{I^*} A_{I^*})^{-1} S^* = (x^*_\lambda)_{I^*} + A^+_{I^*}(z - y). \tag{30}$$

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

$$\operatorname{sign}(x^1_\lambda)_{I^*} = \operatorname{sign}(x^*_\lambda)_{I^*} = S^*. \tag{31}$$

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

$$A^t_{I^*}(z - Ax^1_\lambda) = \lambda A^t_{I^*} d_{I^*, S^*} = \lambda S^* = \lambda \operatorname{sign}(x^1_\lambda)_{I^*}. \tag{32}$$

Moreover for all $j \in J \cap I^*$ from (29), we have that

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

and for all $j \notin J$

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

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

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

Therefore, we obtain

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

So, from Lemma 1, 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

$$\hat{\mu}_\lambda(z) = P_{V_{I^*}}(z) - \lambda d_{I^*, S^*}. \tag{33}$$

Therefore, from (25) and (33), we can deduce that for all $z \in \mathrm{Ball}(y, \varepsilon)$ 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

$$\|z - y\|_2 \leq \varepsilon \Rightarrow \hat{\mu}_\lambda(z) = \hat{\mu}_\lambda(y) + P_{V_{I^*}}(z - y). \tag{34}$$

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

$$\|\hat{\mu}_\lambda(z) - \hat{\mu}_\lambda(y)\|_2 = \|P_{V_{I^*}}(h)\|_2 = \|h\|_2 \leq \varepsilon. \tag{35}$$

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 [14] and [6].

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 (25) and (33). Therefore, $\hat{\mu}_\lambda(y)$ is a locally affine function of y, and then

$$J_{\hat{\mu}_\lambda(y)} = \frac{\partial \hat{\mu}_\lambda(y)}{\partial y} = P_{V_{I^*}} \tag{36}$$

Then the trace formula (23) implies that

$$\operatorname{div}\left(\hat{\mu}_\lambda(y)\right) = \operatorname{tr}\left(P_{V_{I^*}}\right) = |I^*|. \tag{37}$$

This holds almost everywhere since G_λ is of full measure, and (11) 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 Stein's lemma, we have

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

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

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

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

$$\mathbb{E}\left(\|y\|_2^2\right) = n\sigma^2 + \|\mu\|_2^2, \text{ and } \mathbb{V}\left(\|y\|_2^2\right) = 2\sigma^4\left(n + 2\frac{\|\mu\|_2^2}{\sigma^2}\right). \tag{38}$$

Now, we remark also that

$$Q_1(\hat{\mu}_\lambda) - \mathrm{SE} = 2\left(\sigma^2 \operatorname{div}(\hat{\mu}_\lambda) - \langle \varepsilon, \hat{\mu}_\lambda\rangle\right). \tag{39}$$

After an elementary calculation, we obtain

$$\mathbb{E}(\mathrm{SURE}(\hat{\mu}_\lambda) - \mathrm{SE})^2 = \mathbb{E}(Q_1(\hat{\mu}_\lambda) - \mathrm{SE})^2 + \mathbb{V}\left(\|y\|_2^2\right) + 4T, \tag{40}$$

where

$$T = \sigma^2 \mathbb{E}\left(\operatorname{div}(\hat{\mu}_\lambda)\|y\|_2^2\right) - \mathbb{E}\left(\langle \varepsilon, \hat{\mu}_\lambda\rangle \|y\|_2^2\right) = T_1 + T_2, \tag{41}$$

with

$$T_1 = 2\left(\sigma^2 \mathbb{E}\left(\operatorname{div}(\hat{\mu}_\lambda)\langle \varepsilon, \mu\rangle\right) - \mathbb{E}\left(\langle \varepsilon, \hat{\mu}_\lambda\rangle\langle \varepsilon, \mu\rangle\right)\right) \tag{42}$$

and

$$T_2 = \sigma^2 \mathbb{E}\left(\operatorname{div}(\hat{\mu}_\lambda)\|\varepsilon\|_2^2\right) - \mathbb{E}\left(\langle \varepsilon, \hat{\mu}_\lambda\rangle \|\varepsilon\|_2^2\right). \tag{43}$$

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

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

and

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

It follows that

$$T = -2\sigma^2\big(\mathbb{E}\left(\langle \hat{\mu}_\lambda, \mu\rangle\right) + \sigma^2 \mathbb{E}\left(\operatorname{div}(\hat{\mu}_\lambda)\right)\big). \tag{44}$$

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

$$\mathbb{E}(Q_1(\hat{\mu}_\lambda) - \mathrm{SE})^2 = 4\sigma^2\bigg(\mathbb{E}\left(\|\hat{\mu}_\lambda\|_2^2\right) + \sigma^2 \mathbb{E}\left(\operatorname{tr}\left(\left(J_{\hat{\mu}_\lambda}\right)^2\right)\right)\bigg), \tag{45}$$

where $J_{\hat{\mu}_\lambda} = \left(\dfrac{\partial\left(\hat{\mu}_\lambda\right)_i}{\partial y_j}\right)_{1\le i,j\le n}$ 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 $\left(J_{\hat{\mu}_\lambda}\right)^2 = J_{\hat{\mu}_\lambda}$, and $\operatorname{tr}\left(J_{\hat{\mu}_\lambda}\right) = \operatorname{div}(\hat{\mu}_\lambda) = |I^*|$. Therefore, we get

$$\mathbb{E}(Q_1(\hat{\mu}_\lambda) - \mathrm{SE})^2 = 4\sigma^2\left(\mathbb{E}\left(\|\hat{\mu}_\lambda\|_2^2\right) + \sigma^2 \mathbb{E}\left(|I^*|\right)\right). \tag{46}$$

Furthermore, observe that

$$\mathbb{E}\left(\mathrm{SURE}(\hat{\mu}_\lambda)\right) = -n\sigma^2 + \mathbb{E}\left(\|\hat{\mu}_\lambda - y\|_2^2\right) + 2\sigma^2 \mathbb{E}\left(|I^*|\right). \tag{47}$$

Therefore, by combining (38), (40), (44) and (46), we obtain

$$\begin{aligned}
\mathbb{E}(\mathrm{SURE}(\hat{\mu}_\lambda) - \mathrm{SE})^2 &= 2n\sigma^4 + 4\sigma^2 \mathbb{E}\left(\mathrm{SE}\right) - 4\sigma^4 \mathbb{E}\left(|I^*|\right) \\
&= 2n\sigma^4 + 4\sigma^2 \mathbb{E}\left(\mathrm{SURE}(\hat{\mu}_\lambda)\right) - 4\sigma^4 \mathbb{E}\left(|I^*|\right) \\
\text{(by using (47))} &= -2n\sigma^4 + 4\sigma^2 \mathbb{E}\left(\|\hat{\mu}_\lambda - y\|_2^2\right) + 4\sigma^4 \mathbb{E}\left(|I^*|\right).
\end{aligned}$$

On the other hand, since x_λ^* is a minimizer 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

$$\mathbb{E}\left(\|\hat{\mu}_\lambda - y\|_2^2\right) \le \mathbb{E}\left(\|y\|_2^2\right) = n\sigma^2 + \|\mu\|_2^2. \tag{48}$$

Then, since $|I^*| = O(n)$ and from (48), we have

$$\mathbb{E}\left(\left(\frac{\mathrm{SURE}(\hat{\mu}_\lambda) - \mathrm{SE}}{n\sigma^2}\right)^2\right) \le \frac{6}{n} + \frac{4\|\mu\|_2^2}{n^2\sigma^2}. \tag{49}$$

Finally, since $\|\mu\|_2 < +\infty$, 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 [30] in the overdetermined case. We also highlight the fact 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

[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] Daubechies, I., Defrise, M., and Mol, C. D. 2004. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint, Communications on Pure and Applied Mathematics 57, 1413-1541.

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

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

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

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

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

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

[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 shape restricted 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] Rosset, S., Zhu, J., Hastie, T. (2004). Boosting as a Regularized Path to a Maximum Margin Classifier. J. Mach. Learn. Res. 5 941?973.

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

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

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

[23] Tibshirani, R. and Taylor, J. (2012). Degrees of Freedom in Lasso Problems. Technical report, arXiv:1111.0653.

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

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

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

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

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

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

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