HAL
open science

The degrees of freedom of the Lasso for general design matrix

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

To cite this version:

Charles H Dossal, Maher Kachour, Jalal M. Fadili, Gabriel Peyré, Christophe Chesneau. The degrees of freedom of the Lasso for general design matrix. Statistica Sinica, 2013, 23 (2), pp.809-828. 10.5705/ss.2011.281 . hal-00638417v3

HAL Id: hal-00638417

https://hal.science/hal-00638417v3

Submitted on 28 May 2012

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

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 (dof) of penalized ℓ_1 minimization (also known as the Lasso) for linear regression models. We give a closed-form expression of the dof of the Lasso response. Namely, we show that for any given Lasso regularization parameter λ and any observed data y belonging to a set of full (Lebesgue) measure, the cardinality of the support of a particular solution of the Lasso problem is an unbiased estimator of the degrees of freedom. This is achieved without the need of uniqueness of the Lasso solution. Thus, our result holds true for both the underdetermined and the overdetermined case, where the latter was originally studied in [32]. We also show, by providing a simple counterexample, that although the dof theorem of [32] is correct, their proof contains a flaw since their divergence formula holds on a different set of a full measure than the one that they claim. An effective estimator of the number of degrees of freedom may have several applications including an objectively guided choice of the regularization parameter in the Lasso through the SURE framework. Our theoretical findings are illustrated through several numerical simulations.

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 $n \times p$ design matrix, $x^0 = \left(x_1^0, \cdots, x_p^0\right)^{\mathrm{T}}$ is the vector of unknown regression coefficients and ε is a vector of i.i.d. centered Gaussian random variables with variance $\sigma^2 > 0$. In this paper, the number of observations n can be greater than the ambient dimension p of the regression vector to be estimated. Recall that when $n < p$, (1) is an underdetermined linear regression model, whereas when $n \geq p$ and all the columns of A are linearly independent, it is overdetermined.

Let $\widehat{x}(y)$ be an estimator of x^0, and $\widehat{\mu}(y) = A\widehat{x}(y)$ be the associated response or predictor. The concept of degrees of freedom plays a pivotal role in quantifying the complexity of a statistical modeling procedure. More precisely, since $y \sim \mathcal{N}(\mu = Ax^0, \sigma^2 \mathrm{Id}_{n\times n})$ ($\mathrm{Id}_{n\times n}$ is the identity on $\mathbb{R}^n$), according to [8], the degrees of freedom (dof) of the response $\widehat{\mu}(y)$ is defined by

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

Many model selection criteria involve df, e.g. C_p (Mallows [14]), AIC (Akaike Information Criterion, [1]), BIC (Bayesian Information Citerion, [22]), GCV (Generalized Cross Validation, [3]) and SURE (Stein's unbiased risk estimation [23], see Section 2.2). Thus, the dof is a quantity of interest in model validation and selection and it can be used to get the optimal hyperparameters of the estimator. Note that the optimality here is intended in the sense of the prediction $\widehat{\mu}(y)$ and not the coefficients $\widehat{x}(y)$.

The well-known Stein's lemma [23] states that if $y \mapsto \widehat{\mu}(y)$ is weakly differentiable then its divergence is an unbiased estimator of its degrees of freedom, i.e.

$$\widehat{df}(y) = \mathrm{div}(\widehat{\mu}(y)) = \sum_{i=1}^{n} \frac{\partial \widehat{\mu}_i(y)}{\partial y_i}, \quad \text{and} \quad \mathbb{E}(\widehat{df}(y)) = df\ . \tag{3}$$

Here, in order to estimate x^0, we consider solutions to the Lasso problem, proposed originally in [26]. The Lasso amounts to solving the following convex optimization problem

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

where $\lambda > 0$ is called the Lasso regularization parameter and $\|\cdot\|_2$ (resp. $\|\cdot\|_1$) denotes the ℓ_2 (resp. ℓ_1) norm. An important feature of the Lasso is that it promotes sparse solutions. In the last years, there has been a huge amount of work where efforts have focused on investigating the theoretical guarantees of the Lasso as a sparse recovery procedure from noisy measurements. See, e.g., [9, 10, 30, 31, 19, 16, 17, 7, 11, 27], to name just a few.

1.2 Contributions and related work

Let $\widehat{\mu}_\lambda(y) = A\widehat{x}_\lambda(y)$ be the Lasso response vector, where $\widehat{x}_\lambda(y)$ is a solution of the Lasso problem ($\mathrm{P}_1(y,\lambda)$). Note that all minimizers of the Lasso share the same image under A, i.e. $\widehat{\mu}_\lambda(y)$ is

uniquely defined; see Lemma 2 in Section 5 for details. The main contribution of this paper is first to provide an unbiased estimator of the degrees of freedom of the Lasso response for any design matrix. The estimator is valid everywhere except on a set of (Lebesgue) measure zero. We reach our goal without any additional assumption to ensure uniqueness of the Lasso solution. Thus, our result covers the challenging underdetermined case where the Lasso problem does not necessarily have a unique solution. It obviously holds when the Lasso problem ($\mathrm{P}_1(y, \lambda)$) has a unique solution, and in particular in the overdetermined case studied in [32]. Using the estimator at hand, we also establish the reliability of the SURE as an unbiased estimator of the Lasso prediction risk.

While this paper was submitted, we became aware of the independent work of Tibshirani and Taylor [25], who studied the dof for general A both for the Lasso and the general (analysis) Lasso.

Section 3 is dedicated to a thorough comparison and discussion of connections and differences between our results and the one in [32, Theorem 1] for the overdetermined case, and that of [12, 25, 28] 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 dof of the Lasso, and we investigate the reliability of the SURE estimate of the Lasso prediction risk. Then, we discuss relation of our work with concurrent one in the 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 the dof

First, some notations and definitions are necessary. For any vector x, x_i denotes its ith component. The support or the active set of x is defined by

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

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

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

$$H_{I,j,S} = \{u \in \mathbb{R}^n : \langle P_{V_I^\perp}(a_j), u\rangle = \pm\lambda(1 - \langle a_j, (A_I^+)^{\mathrm{T}} S\rangle)\}. \tag{4}$$

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

$$\Omega = \{(I, j, S) : a_j \notin V_I\} \tag{5}$$

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

$$G_\lambda = \mathbb{R}^n \setminus \bigcup_{(I,j,S)\in\Omega} H_{I,j,S}. \tag{6}$$

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.

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 $(\mathrm{P}_1(y,\lambda))$. Let $x^*_\lambda(y) \in \mathcal{M}_{y,\lambda}$ with support I^* such that A_{I^*} is full rank. Then,*

$$|I^*| = \min_{\widehat{x}_\lambda(y)\in\mathcal{M}_{y,\lambda}} |\operatorname{supp}(\widehat{x}_\lambda(y))|. \tag{7}$$

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

$$\widehat{\mu}_\lambda(z) = \widehat{\mu}_\lambda(y) + P_{V_{I^*}}(z-y). \tag{8}$$

As stated, this theorem assumes the existence of a solution whose active matrix A_{I^*} is full rank. This can be shown to be true; see e.g. [5, Proof of Theorem 1] or [20, Theorem 3, Section B.1][1]. It is worth noting that this proof is constructive, in that it yields a solution $x^*_\lambda(y)$ of $(\mathrm{P}_1(y,\lambda))$ such that A_{I^*} is full column rank from any solution $\widehat{x}_\lambda(y)$ whose active matrix has a nontrivial kernel. This will be exploited in Section 4 to derive an algorithm to get $x^*_\lambda(y)$, and hence I^*.

A direct consequence of our main theorem is that outside G_λ, the mapping $\widehat{\mu}_\lambda(y)$ is C^∞ and the sign and support are locally constant. Applying Stein's lemma yields Corollary 1 below. The latter states that the number of nonzero coefficients of $x^*_\lambda(y)$ is an unbiased estimator of the dof of the Lasso.

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

$$\widehat{df}_\lambda(y) := \operatorname{div}(\widehat{\mu}_\lambda(y)) = |I^*|. \tag{9}$$

Therefore,

$$df = \mathbb{E}(\widehat{df}_\lambda(y)) = \mathbb{E}(|I^*|). \tag{10}$$

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

2.2 Reliability of the SURE estimate of the Lasso prediction risk

In this work, we focus on the SURE as a model selection criterion. The SURE applied to the Lasso reads

$$\mathrm{SURE}(\widehat{\mu}_\lambda(y)) = -n\sigma^2 + \|\widehat{\mu}_\lambda(y) - y\|_2^2 + 2\sigma^2\widehat{df}_\lambda(y), \tag{11}$$

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

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

[1] This proof is alluded to in the note at the top of [21, Page 363].

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

$$\mathrm{SE}(\widehat{\mu}_\lambda(y)) = \|\widehat{\mu}_\lambda(y) - \mu\|_2^2. \tag{12}$$

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

$$\mathbb{E}\left((\mathrm{SURE}(\widehat{\mu}_\lambda(y)) - \mathrm{SE}(\widehat{\mu}_\lambda(y)))^2\right) = -2\sigma^4 n + 4\sigma^2 \mathbb{E}\left(\|\widehat{\mu}_\lambda(y) - y\|_2^2\right) + 4\sigma^4 \mathbb{E}\left(|I^*|\right). \tag{13}$$

Moreover,

$$\mathbb{E}\left(\left(\frac{\mathrm{SURE}(\widehat{\mu}_\lambda(y)) - \mathrm{SE}(\widehat{\mu}_\lambda(y))}{n\sigma^2}\right)^2\right) = O\left(\frac{1}{n}\right). \tag{14}$$

3 Relation to prior work

Overdetermined case [32]

The authors in [32] studied the dof of the Lasso in the overdetermined case. Precisely, when $n \geq p$ and all the columns of the design matrix A are linearly independent, i.e. $\mathrm{rank}(A) = p$. In fact, in this case the Lasso problem has a unique minimizer $\widehat{x}_\lambda(y) = x^*_\lambda(y)$ (see Theorem 1).

Before discussing the result of [32], let's point out a popular feature of $\widehat{x}_\lambda(y)$ as λ varies in $]0, +\infty[$:

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

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

  Thus, in $]\lambda_{m+1}, \lambda_m[$, the support and the sign of $\widehat{x}_\lambda(y)$ are constant, see [7, 17, 18]. Hence, we call $(\lambda_m)_{0 \leq m \leq K}$ 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 $\widehat{x}_\lambda(y)$,

$$(\widehat{x}_\lambda(y))_{I_m} = A^+_{I_m} y - \lambda (A^{\mathrm{T}}_{I_m} A_{I_m})^{-1} S m, \tag{15}$$

where I_m and S^m are respectively the (constant) support and sign vector of $\widehat{x}_\lambda(y)$ for $\lambda \in]\lambda_{m+1}, \lambda_m[$. Hence, based on (15), [32] 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

$$\mathcal{K}_\lambda = \mathbb{R}^n \setminus \mathcal{N}_\lambda, \tag{16}$$

so that $\forall\, y \in \mathcal{K}_\lambda$, λ is not any of the transition points.
Then, for the overdetermined case, [32] 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 dof. In fact, their main argument is that, by eliminating the vectors associated to the transition points, the support and the sign of the Lasso solution are locally constant with respect to y, see [32, Lemma 5].

We recall that the overdetermined case, considered in [32], is a particular case of our result since the minimizer is unique. Thus, according to the Corollary 1, we find the same result as [32]

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 [32] ? In other words, is there a link between $\mathcal{K}_\lambda$ and G_λ ?

The answer is that, depending on the matrix A, these two sets may be different. More importantly, it turns out that although the dof formula [32, Theorem 1] is correct, unfortunately, their proof contains a flaw since their divergence formula [32, 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 $\{e_1, e_2\}$ be 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 whose columns are a_1 and a_2.

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

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

Let $y = \alpha a_1$ with $\alpha > 0$, for any $\lambda > 0$, $y \in H_{I,j,S}$ (or equivalently here $y \notin G_\lambda$). Using Lemma 1 (see Section 5), one gets that for any $\lambda \in]0, \alpha[$, the solution of $(\mathrm{P}_1(y, \lambda))$ is $\widehat{x}_\lambda(y) = (\alpha - \lambda, 0)$ and that for any $\lambda \geq \alpha$, $\widehat{x}_\lambda(y) = (0, 0)$. Hence the only transition point is $\lambda_0 = \alpha$. It follows that for $\lambda < \alpha$, y belongs to $\mathcal{K}_\lambda$ defined in [32], 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 $(\mathrm{P}_1(z_1, \lambda))$ is different from the support of $(\mathrm{P}_1(y, \lambda))$.
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 $(\mathrm{P}_1(z_1, \lambda))$ is equal to $\widehat{x}_\lambda(z_1) = (\alpha - \lambda - \varepsilon, \varepsilon)$ whose support is different from that of $\widehat{x}_\lambda(y) = (\alpha - \lambda, 0)$.

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

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 $(\mathrm{P}_1(y, \lambda))$ is not stable to infinitesimal perturbations of y. This situation may occur for under or overdetermined problems. In summary, even in the overdetermined case, excluding the set of transition points is not sufficient to guarantee stability of the support and sign of the Lasso solution.

General case [12, 25, 28]

In [12], the author studies the degrees of freedom of a generalization of the Lasso where the regression coefficients are constrained to a closed convex set. When the latter is a ℓ_1 ball and $p > n$, he proposes the cardinality of the support as an estimate of *df* but under a restrictive assumption on A under which the Lasso problem has a unique solution.

In [25, Theorem 2], the authors proved that

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

where $I = I(y)$ is the active set of any solution $\widehat{x}_\lambda(y)$ to $(\mathrm{P}_1(y, \lambda))$. This coincides with Corollary 1 when A_I is full rank with $\mathrm{rank}(A_I) = \mathrm{rank}(A_{I^*})$. Note that in general, there exist vectors $y \in \mathbb{R}^n$ where the smallest cardinality among all supports of Lasso solutions is different from the rank of the active matrix associated to the largest support. But these vectors are precisely those excluded in G_λ. In the case of the generalized Lasso (a.k.a. analysis sparsity prior in the signal processing

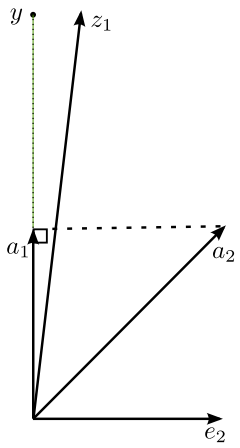


Figure 1: A counterexample for $n = p = 2$ of vectors in G_λ but not in $\mathcal{K}_\lambda$. See text for a detailed discussion.

community), Vaiter et al. [28, Corollary 1] and Tibshirani and Taylor [25, Theorem 3] provide a formula of an unbiased estimator of *df*. This formula reduces to that of Corollary 1 when the analysis operator is the identity.

4 Numerical experiments

Experiments description In this section, we support the validity of our main theoretical findings with some numerical simulations, by checking the unbiasedness and the reliability of the SURE for the Lasso. Here is the outline of these experiments.

For our first study, we consider two kinds of 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. The original sparse vector x^0 was drawn randomly according to a mixed Gaussian-Bernoulli distribution, such that x^0 is 15-sparse (i.e. $|\operatorname{supp}(x^0) = 15|$). For each design matrix A and vector x^0, we generate $K = 100$ independent replications $y^k \in \mathbb{R}^n$ of the observation vector according to the linear regression model (1). Then, for each y^k and a given λ, we compute the Lasso response $\widehat{\mu}_\lambda(y^k)$ using the now popular iterative soft-thresholding algorithm [4][2], and we compute $\mathrm{SURE}(\widehat{\mu}_\lambda(y^k))$ and $\mathrm{SE}(\widehat{\mu}_\lambda(y^k))$. We then compute the empirical mean and the standard deviation of $\big(\mathrm{SURE}(\widehat{\mu}_\lambda(y^k))\big)_{1\le k\le K}$, the empirical mean of $\big(\mathrm{SE}(\widehat{\mu}_\lambda(y^k))\big)_{1\le k\le K}$, which corresponds to the computed prediction risk, and we compute R_T the empirical normalized reliability on the left-hand side of (13),

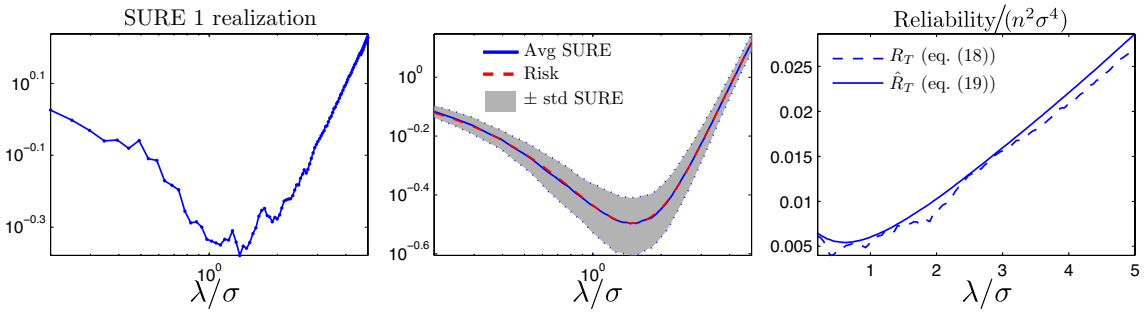
$$R_T = \frac{1}{K}\sum_{k=1}^{K}\left(\frac{\mathrm{SURE}(\widehat{\mu}_\lambda(y^k)) - \mathrm{SE}(\widehat{\mu}_\lambda(y^k))}{n\sigma^2}\right)^2. \tag{17}$$

Moreover, based on the right-hand side of (13), we compute $\widehat{R}_T$ as

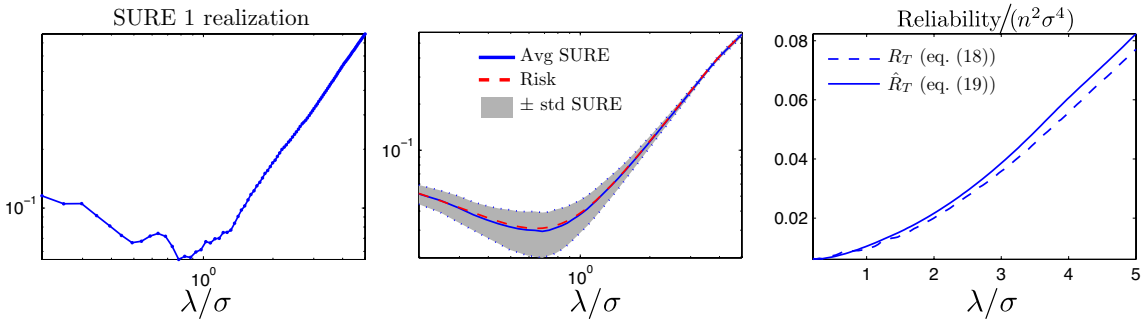
$$\widehat{R}_T = -\frac{2}{n} + \frac{4}{n^2\sigma^2}\left(\frac{1}{K}\sum_{k=1}^{K}\left(\|\widehat{\mu}_\lambda(y^k) - y^k\|_2^2\right)\right) + \frac{4}{n^2}\left(\frac{1}{K}\sum_{k=1}^{K}\left(|I^*|_k\right)\right), \tag{18}$$

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

[2]Iterative soft-thresholding through block-coordinate relaxation was proposed in [21] for matrices A structured as the union of a finite number of orthonormal matrices.



(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 three plots.

various values of λ, for the two kinds of design matrices considered above.

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

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

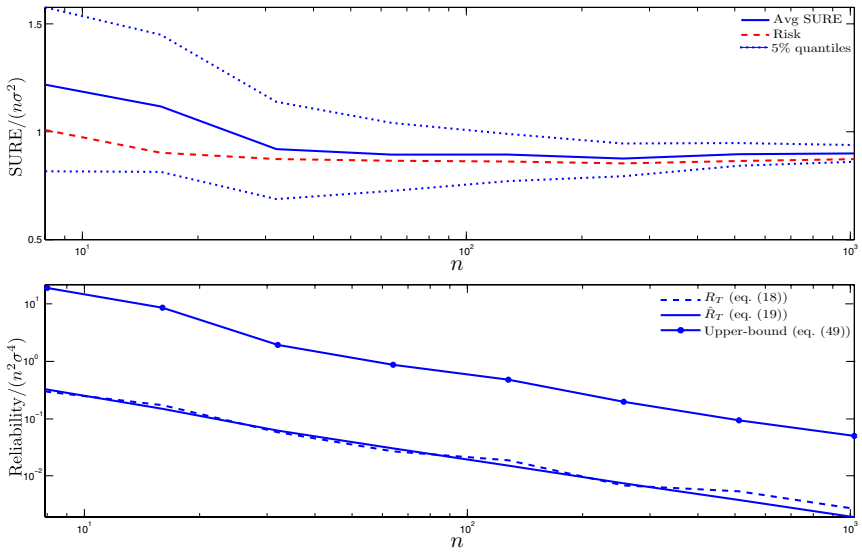
1. Take $h \in \ker A_I$ such that $\operatorname{supp} h \subset I$.

2. For $t \in \mathbb{R}$, $A\widehat{x}_\lambda(y) = A\left(\widehat{x}_\lambda(y) + th\right)$ and the mapping $t \mapsto \|\widehat{x}_\lambda(y) + th\|_1$ is locally affine in a neighborhood of 0, i.e. for $|t| < \min_{j\in I} |(\widehat{x}_\lambda(y))_j|/\|h\|_\infty$. $\widehat{x}_\lambda(y)$ being a minimizer of $(\mathrm{P}_1(y,\lambda))$, this mapping is constant in a neighborhood of 0. We have then constructed a whole collection of solutions to $(\mathrm{P}_1(y,\lambda))$ having the same image and the same ℓ_1 norm, which lives on a segment.

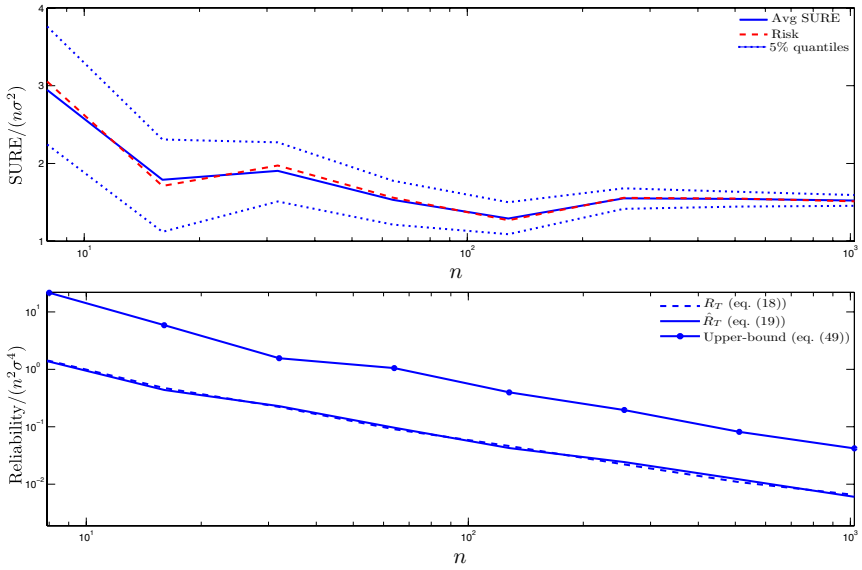
3. Move along h with the largest step $t_0 > 0$ until an entry of $\widehat{x}^1_\lambda(y) = \widehat{x}_\lambda(y) + t_0 h$ vanishes, i.e. $\operatorname{supp}(\widehat{x}^1_\lambda(y) + t_0 h) \subsetneq I$.

4. Repeat this process until getting a vector $x^*_\lambda(y)$ with a full column rank active matrix A_{I^*}.

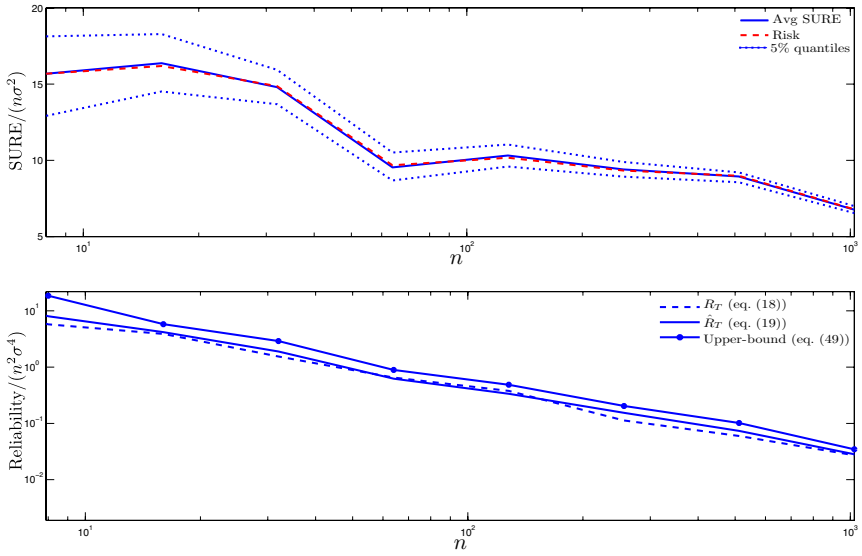
Note that this construction bears similarities with the one in [20].



(a) $\lambda/\sigma = 0.1$



(b) $\lambda/\sigma = 1$



(c) $\lambda/\sigma = 10$

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

Results discussion Figure 2 depicts the obtained results. For each design matrix, we associate a panel, each containing 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 prediction risk curve and the empirical mean of the SURE as a function of the regularization parameter λ. Namely, the dashed curve represents the calculated prediction risk, the solid 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 prediction risk with a controlled variance. This suggests that the SURE is consistent, and then so is our estimator of the degrees of freedom. In the third graph, we plot the theoretical and empirical normalized reliability, defined respectively by (17) and (18), as a function of the regularization parameter λ. More precisely, the solid and dashed blue curves represent respectively R_T and $\widehat{R}_T$. This confirms numerically that both sides (R_T and $\widehat{R}_T$) of (13) 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, one can compute the optimal λ that minimizes the SURE, i.e.

$$\lambda_{\text{optimal}} = \underset{\lambda>0}{\operatorname{argmin}} \ \text{SURE}(\widehat{\mu}_\lambda(y)). \tag{19}$$

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

Now, for our second simulation study, we consider a partial Fourier design matrix, with $n < p$ and a constant underdeterminacy factor $p/n = 4$. x^0 was again simulated according to a mixed Gaussian-Bernoulli distribution with $\lceil 0.1p \rceil$ non-zero entries. For each of three values of $\lambda/\sigma \in \{0.1, 1, 10\}$ (small, medium and large), we compute the prediction risk curve, the empirical mean of the SURE, as well as the values of the normalized reliability R_T and $\widehat{R}_T$, as a function of $n \in \{8, \cdots, 1024\}$. The obtained results are shown in Figure 3. For each value of λ, the first plot (top panel) displays the normalized empirical mean of the SURE (solid line) and its 5% quantiles (dotted) as well as the computed normalized prediction risk (dashed). Unbiasedness is again clear whatever the value of λ. The trend on the prediction risk (and average SURE) is in agreement with rates known for the Lasso, see e.g. [2]. The second plot confirms that the SURE is an asymptotically reliable estimate of the prediction risk with the rate established in Theorem 2. Moreover, as expected, the actual reliability gets closer to the upper-bound (48) as the number of samples n increases.

5 Proofs

First of all, we recall some classical properties of any solution of the Lasso (see, e.g., [17, 7, 11, 27]). To lighten the notation in the two following lemmas, we will drop the dependency of the minimizers of $(\mathrm{P}_1(y, \lambda))$ on either λ or y.

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

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

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

$$\widehat{x}_I = A_I^+ y - \lambda (A_I^{\mathrm{T}} A_I)^{-1} \operatorname{sign}(\widehat{x}_I) \ . \tag{20}$$

Note that if the inequality in condition 2 above is strict, then $\widehat{x}$ is the unique minimizer of the Lasso problem $(\mathrm{P}_1(y, \lambda))$ [11].

Lemma 2 below shows that all solutions of $(\mathrm{P}_1(y, \lambda))$ have the same image by A. In other words, the Lasso response $\widehat{\mu}_\lambda(y)$, is unique, see [5].

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

$$A\widehat{x}^1 = A\widehat{x}^2 = \widehat{\mu}_\lambda(y).$$

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

$$\left(J_{\widehat{\mu}(y)}\right)_{i,j} := \frac{\partial \widehat{\mu}(y)_i}{\partial y_j}, \qquad i, j = 1, \cdots, n. \tag{21}$$

Then we can write

$$\operatorname{div}\left(\widehat{\mu}(y)\right) = \operatorname{tr}\left(J_{\widehat{\mu}(y)}\right). \tag{22}$$

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

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

According to Lemma 1, we know that

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

Furthermore, from (20), we get the following implicit form of $x_\lambda^*(y)$

$$(x_\lambda^*(y))_{I^*} = A_{I^*}^+ y - \lambda (A_{I^*}^{\mathrm{T}} A_{I^*})^{-1} S^*. \tag{23}$$

It follows that

$$\widehat{\mu}_\lambda(y) = P_{V_{I^*}}(y) - \lambda d_{I^*, S^*}, \tag{24}$$

and

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

where $d_{I^*, S^*} = (A_{I^*}^+)^{\mathrm{T}} S^*$. We define the following set of indices

$$J = \{j : |\langle a_j, \widehat{r}_\lambda(y) \rangle| = \lambda\}. \tag{26}$$

From Lemma 1 we deduce that

$$I^* \subset J.$$

Since the orthogonal projection is a self-adjoint operator and from (25), for all $j \in J$, we have

$$|\langle P_{V_{I^*}^\perp}(a_j), y \rangle + \lambda \langle a_j, d_{I^*, S^*} \rangle| = \lambda. \tag{27}$$

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

$$a_j \in V_{I^*}, \text{ and therefore } |\langle a_j, d_{I^*,S^*}\rangle| = 1. \tag{28}$$

In fact, if $a_j \notin V_{I^*}$ then $(I^*, j, S^*) \in \Omega$ and from (27) we have that $y \in H_{I^*,j,S^*}$, which is a contradiction with $y \in G_\lambda$.
Therefore, the collection of vectors $(a_i)_{i\in I^*}$ forms a basis of $V_J = \text{span}(a_j)_{j\in J}$. Now, suppose that $\widehat{x}_\lambda(y)$ is another solution of $(\mathrm{P}_1(y,\lambda))$, such that its support I is different from I^*. If A_I is full column rank, then by using the same above arguments we can deduce that $(a_i)_{i\in I}$ forms also a basis of V_J. Therefore, we have

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

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

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

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

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

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

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

$$(x^1_\lambda(z))_{I^*} = A^+_{I^*} z - \lambda (A^{\mathrm{T}}_{I^*} A_{I^*})^{-1} S^* = (x^*_\lambda(y))_{I^*} + A^+_{I^*}(z-y). \tag{29}$$

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

$$\text{sign}(x^1_\lambda(z))_{I^*} = \text{sign}(x^*_\lambda(y))_{I^*} = S^*. \tag{30}$$

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

$$A^{\mathrm{T}}_{I^*}(z - Ax^1_\lambda(z)) = \lambda A^{\mathrm{T}}_{I^*} d_{I^*,S^*} = \lambda S^* = \lambda \,\text{sign}\,(x^1_\lambda(z))_{I^*}. \tag{31}$$

Moreover for all $j \in J \cap I^*$, from (28), we have that

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

and for all $j \notin J$

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

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

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

Therefore, we obtain

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

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

$$\widehat{\mu}_\lambda(z) = P_{V_{I^*}}(z) - \lambda d_{I^*,S^*}. \tag{32}$$

Therefore, from (24) and (32), we can deduce that for all $z \in \mathrm{Ball}(y,\varepsilon)$ we have

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

□

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

$$\|z-y\|_2 \leq \varepsilon \Rightarrow \widehat{\mu}_\lambda(z) = \widehat{\mu}_\lambda(y) + P_{V_{I^*}}(z-y). \tag{33}$$

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

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

Therefore, the Lasso response $\widehat{\mu}_\lambda(y)$ is uniformly Lipschitz on G_λ. Moreover, $\widehat{\mu}_\lambda(y)$ is a continuous function of y, and thus $\widehat{\mu}_\lambda(y)$ is uniformly Lipschitz on $\mathbb{R}^n$. Hence, $\widehat{\mu}_\lambda(y)$ is almost differentiable; see [15] and [7].

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 as $x^*_\lambda(y)$, and thus $\widehat{\mu}_\lambda(z)$ belongs to the vector space V_{I^*}, whose dimension equals to $|I^*|$, see (24) and (32). Therefore, $\widehat{\mu}_\lambda(y)$ is a locally affine function of y, and then

$$J_{\widehat{\mu}_\lambda(y)} = P_{V_{I^*}}\ . \tag{35}$$

Then the trace formula (22) implies that

$$\mathrm{div}\left(\widehat{\mu}_\lambda(y)\right) = \mathrm{tr}\left(P_{V_{I^*}}\right) = |I^*|. \tag{36}$$

This holds almost everywhere since G_λ is of full measure, and (10) is obtained by invoking Stein's lemma. □

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

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

From Stein's lemma, we have

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

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

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

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

$$\mathbb{E}\left(\|y\|_2^2\right) = n\sigma^2 + \|\mu\|_2^2, \text{ and } \mathbb{V}\left(\|y\|_2^2\right) = 2\sigma^4\left(n + 2\frac{\|\mu\|_2^2}{\sigma^2}\right). \tag{37}$$

Now, we remark also that

$$Q_1(\widehat{\mu}_\lambda(y)) - \mathrm{SE}(\widehat{\mu}_\lambda(y)) = 2\left(\sigma^2 \operatorname{div}(\widehat{\mu}_\lambda(y)) - \langle \varepsilon, \widehat{\mu}_\lambda(y)\rangle\right). \tag{38}$$

After an elementary calculation, we obtain

$$\mathbb{E}(\mathrm{SURE}(\widehat{\mu}_\lambda(y)) - \mathrm{SE}(\widehat{\mu}_\lambda(y)))^2 = \mathbb{E}(Q_1(\widehat{\mu}_\lambda(y)) - \mathrm{SE}(\widehat{\mu}_\lambda(y)))^2 + \mathbb{V}\left(\|y\|_2^2\right) + 4T, \tag{39}$$

where

$$T = \sigma^2\mathbb{E}\left(\operatorname{div}(\widehat{\mu}_\lambda(y))\|y\|_2^2\right) - \mathbb{E}\left(\langle \varepsilon, \widehat{\mu}_\lambda(y)\rangle\|y\|_2^2\right) = T_1 + T_2, \tag{40}$$

with

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

and

$$T_2 = \sigma^2\mathbb{E}\left(\operatorname{div}(\widehat{\mu}_\lambda(y))\|\varepsilon\|_2^2\right) - \mathbb{E}\left(\langle \varepsilon, \widehat{\mu}_\lambda(y)\rangle\|\varepsilon\|_2^2\right). \tag{42}$$

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

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

and

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

It follows that

$$T = -2\sigma^2\big(\mathbb{E}\left(\langle \widehat{\mu}_\lambda, \mu\rangle\right) + \sigma^2\mathbb{E}\left(\operatorname{div}(\widehat{\mu}_\lambda(y))\right)\big). \tag{43}$$

Moreover, from [13, Property 1], we know that

$$\mathbb{E}(Q_1(\widehat{\mu}_\lambda(y)) - \mathrm{SE}(\widehat{\mu}_\lambda(y)))^2 = 4\sigma^2\left(\mathbb{E}\left(\|\widehat{\mu}_\lambda(y)\|_2^2\right) + \sigma^2\mathbb{E}\left(\operatorname{tr}\left(\left(J_{\widehat{\mu}_\lambda(y)}\right)^2\right)\right)\right), \tag{44}$$

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

$$\mathbb{E}(Q_1(\widehat{\mu}_\lambda(y)) - \mathrm{SE}(\widehat{\mu}_\lambda(y)))^2 = 4\sigma^2\left(\mathbb{E}\left(\|\widehat{\mu}_\lambda(y)\|_2^2\right) + \sigma^2\mathbb{E}\left(|I^*|\right)\right). \tag{45}$$

Furthermore, observe that

$$\mathbb{E}\left(\mathrm{SURE}(\widehat{\mu}_\lambda(y))\right) = -n\sigma^2 + \mathbb{E}\left(\|\widehat{\mu}_\lambda(y) - y\|_2^2\right) + 2\sigma^2\mathbb{E}\left(|I^*|\right). \tag{46}$$

Therefore, by combining (37), (39), (43) and (45), we obtain

$$\begin{aligned}
\mathbb{E}(\mathrm{SURE}(\widehat{\mu}_\lambda(y)) - \mathrm{SE}(\widehat{\mu}_\lambda(y)))^2 &= 2n\sigma^4 + 4\sigma^2\mathbb{E}\left(\mathrm{SE}(\widehat{\mu}_\lambda(y))\right) - 4\sigma^4\mathbb{E}\left(|I^*|\right) \\
&= 2n\sigma^4 + 4\sigma^2\mathbb{E}\left(\mathrm{SURE}(\widehat{\mu}_\lambda(y))\right) - 4\sigma^4\mathbb{E}\left(|I^*|\right) \\
\text{(by using (46))} &= -2n\sigma^4 + 4\sigma^2\mathbb{E}\left(\|\widehat{\mu}_\lambda(y) - y\|_2^2\right) + 4\sigma^4\mathbb{E}\left(|I^*|\right).
\end{aligned}$$

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

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

Therefore, we have

$$\mathbb{E}\left(\|\widehat{\mu}_\lambda(y) - y\|_2^2\right) \le \mathbb{E}\left(\|y\|_2^2\right) = n\sigma^2 + \|\mu\|_2^2. \tag{47}$$

Then, since $|I^*| = O(n)$ and from (47), we have

$$\mathbb{E}\left(\left(\frac{\mathrm{SURE}(\widehat{\mu}_\lambda(y)) - \mathrm{SE}(\widehat{\mu}_\lambda(y))}{n\sigma^2}\right)^2\right) \leq \frac{6}{n} + \frac{4\|\mu\|_2^2}{n^2\sigma^2}. \tag{48}$$

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

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

□

6 Discussion

In this paper we proved that the number of nonzero coefficients of a particular solution of the Lasso problem is an unbiased estimate of the degrees of freedom of the Lasso response for linear regression models. This result covers both the over and underdetermined cases. This was achieved through a divergence formula, valid almost everywhere except on a set of measure zero. We gave a precise characterization of this set, and the latter turns out to be larger than the set of all the vectors associated to the transition points considered in [32] in the overdetermined case. We also highlight the fact that even in the overdetermined case, the set of transition points is not sufficient for the divergence formula to hold.

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

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

References

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

[2] Bickel, P. J., Ritov, Y., and Tsybakov, A., (2009). Simultaneous analysis of Lasso and Dantzig selector. Annals of Statistics. 37 1705?1732.

[3] Craven, P. and Wahba, G. (1979). Smoothing Noisy Data with Spline Functions: estimating the correct degree of smoothing by the method of generalized cross validation. Numerische Mathematik 31, 377-403.

[4] 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.

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

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

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

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

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

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

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

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

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

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

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

[16] 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.

[17] 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.

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

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

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

[21] Sardy, S., Bruce, A., and Tseng, P. (2000). Block coordinate relaxation methods for nonparametric wavelet denoising. J. of Comp. Graph. Stat. 9(2) 361?379.

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

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

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

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

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

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

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

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

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

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

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