Appl. Numer. Math. 208 (2025) 297–309

Contents lists available at ScienceDirect

Applied Numerical Mathematics

journal homepage: www.elsevier.com/locate/apnum

Research Paper

Numerical properties of solutions of LASSO regression

Mayur V. Lakshmi [1], Joab R. Winkler [*]

The University of Sheffield, Department of Computer Science, Regent Court, 211 Portobello, Sheffield S1 4DP, United Kingdom

ARTICLE INFO

Keywords:
LASSO regression
Sparsity
Cross validation
L-curve

ABSTRACT

The determination of a concise model of a linear system when there are fewer samples m than predictors n requires the solution of the equation $Ax = b$, where $A \in \mathbb{R}^{m\times n}$ and rank $A = m$, such that the selected solution from the infinite number of solutions is sparse, that is, many of its components are zero. This leads to the minimisation with respect to x of $f(x,\lambda) = \|Ax - b\|_2^2 + \lambda\|x\|_1$, where λ is the regularisation parameter. This problem, which is called LASSO regression, yields a family of functions $x_{\text{lasso}}(\lambda)$ and it is necessary to determine the optimal value of λ, that is, the value of λ that balances the fidelity of the model, $\|Ax_{\text{lasso}}(\lambda) - b\| \approx 0$, and the satisfaction of the constraint that $x_{\text{lasso}}(\lambda)$ be sparse. The aim of this paper is an investigation of the numerical properties of $x_{\text{lasso}}(\lambda)$, and the main conclusion of this investigation is the incompatibility of sparsity and stability, that is, a sparse solution $x_{\text{lasso}}(\lambda)$ that preserves the fidelity of the model exists if the least squares (LS) solution $x_{\text{ls}} = A^{\dagger}b$ is unstable. Two methods, cross validation and the L-curve, for the computation of the optimal value of λ are compared and it is shown that the L-curve yields significantly better results. This difference between stable and unstable solutions x_{ls} of the LS problem manifests itself in the very different forms of the L-curve for these two solutions. The paper includes examples of stable and unstable solutions x_{ls} that demonstrate the theory.

1. Introduction

Many problems yield data in which there are fewer samples m than predictors n, and they require the computation of a solution z of the least squares (LS) problem[2]

$$z = \arg\min_{x\in\mathbb{R}^n} \|Ax - b\|_2^2, \qquad A \in \mathbb{R}^{m\times n}, \qquad m < n. \tag{1}$$

This is an underdetermined set of equations, and if rank $A = m$, the minimum norm solution is

$$x_{\text{ls}} = A^{\dagger}b = A^T(AA^T)^{-1}b. \tag{2}$$

[*] Corresponding author.
E-mail address: j.r.winkler@sheffield.ac.uk (J.R. Winkler).

[1] Mayur Lakshmi was supported by a grant from the EPSRC, reference EP/W011905/1.

[2] The 1- and 2-norms are denoted by the subscripts 1 and 2, respectively. If a subscript is omitted, then the quantity (the norm of a vector or matrix, and the condition number of a matrix) in both norms is considered.

https://doi.org/10.1016/j.apnum.2024.03.010
Received 21 August 2023; Received in revised form 15 January 2024; Accepted 11 March 2024

Available online 20 March 2024

The general solution of (1) is $z = x_{\rm ls} + v$ where v is an arbitrary vector that lies in the null space of A. This problem yields, therefore, an infinite number of solutions and it is necessary to select one solution from this infinite number of solutions. This selection is made using the criterion of sparsity, that is, the number of non-zero entries in $x_{\rm ls} + v$ is minimised because it is desired to construct a simple, concise and computationally efficient model. This criterion is motivated by the many examples, including the analysis of microarray data, image processing and face recognition, in which sparsity arises.

A sparse solution of (1) is achieved by constraining $\|x\|_1$, which leads to LASSO (Least Absolute Shrinkage and Selection Operator) regression [1],

$$x_{\rm lasso}(\lambda) = \arg \min_{x\in\mathbb{R}^n} f(x,\lambda) = \arg \min_{x\in\mathbb{R}^n} \left\{ \|Ax-b\|_2^2 + \lambda \|x\|_1 \right\}, \tag{3}$$

where $\lambda \geq 0$ is the regularisation parameter and

$$\|x\|_p = \left(\sum_{i=1}^n |x_i|^p \right)^{\frac{1}{p}}, \qquad p = 1, 2, \infty, \qquad x = \{x_i\}_{i=1}^n .$$

The solution $x_{\rm lasso}(\lambda)$ is a family of functions parameterised by λ, and two methods, coordinate descent [2,3] and LARS (Least Angle Regression) [4], are used to solve (3) for a given value of λ. The methods differ because LARS provides a piecewise linear solution of (3) but the solution from coordinate descent is defined on a grid of points.

It is often stated that $x_{\rm lasso}(0)$ is equal to $x_{\rm ls}$ for $m < n$, but this is incorrect because $x_{\rm lasso}(0)$ is equal to the set $\mathcal{S}$ of the infinite number of solutions z defined in (1), and $x_{\rm ls}$ is the member of $\mathcal{S}$ that has minimum 2-norm [5, p. 16]. Coordinate descent does not return $x_{\rm ls}$ when $\lambda = 0$ because it does not use the minimum norm criterion for the selection of a solution in $\mathcal{S}$.

The aim of this paper is an investigation of the numerical properties of $x_{\rm lasso}(\lambda)$, and the main result is that a necessary condition for the existence of an optimal sparse solution of (3), that is, a sparse solution that preserves the fidelity of the model, is that $x_{\rm ls}$ be unstable, such that a small relative error in b yields a much larger relative error in $x_{\rm ls}$. Furthermore, an optimal sparse solution does not exist if $x_{\rm ls}$ is stable because there does not exist a value of λ such that $\|Ax_{\rm lasso}(\lambda) - b\|_2$ and $\|x_{\rm lasso}(\lambda)\|_1$ assume, approximately, their minimum values. This result requires that a refined condition number of the minimum norm solution $x_{\rm ls}$ be developed, and this is addressed in Section 2. This refined condition number shows a difference between stable and unstable LS problems, but the condition number $\kappa(A)$ of A does not reveal this difference.

The incompatibility of sparsity and numerical stability has been established in feature selection and fMRI (functional magnetic resonance imaging), and its extension to LASSO regression is the main result of the work described in this paper. In particular, Xu et al. [6] show that an algorithm that is stable cannot identify redundant features and an algorithm that identifies redundant features is not stable. Also, Baldassarre et al. [7] consider fMRI for model selection in brain decoding and they show that sparse models can be unstable due to under sampling or slight changes in experimental conditions. The refined condition number developed in Section 2 quantifies the stability of the LS problem and it is required for the determination of the existence, or otherwise, of an optimal sparse solution of (3). It is also shown in Section 2 that the association between an unstable LS problem and sparsity can be motivated by comparison of the forms of LASSO regression and Tikhonov regularisation (ridge regression).

There does not exist, in general, a closed form expression for $x_{\rm lasso}(\lambda)$, but some of its properties are considered in Section 3. In particular, it is shown that an optimal sparse solution that has many zero entries may not exist. The determination of the optimal value λ^* of λ is critical because a value that is too small may yield a model that is too complex and can be made simpler by the elimination of more predictors, but a value that is too large may lead to a large error because the fidelity of the model is not preserved. The method of cross validation (CV) [5, §6.2] is frequently used to compute the value of λ^* but it requires the determination of a shallow minimum of a function, which is difficult. Another method, the L-curve [8, §4.6], is discussed in Section 4 and it is shown that the curve has the form of an L if an optimal sparse solution that has few zero entries exists, and the value of λ^* is the value of λ in the corner of the L. Section 5 contains examples of stable and unstable LS problems that demonstrate the theoretical analysis, and they show that the L-curve yields better results than CV. The paper is summarised in Section 6.

2. Condition estimation

The condition number $\kappa(A) = \|A\| \left\|A^\dagger\right\|$ of a rectangular matrix A, where $A^\dagger$ is the pseudo-inverse of A, is a measure of the stability of $x_{\rm ls} = A^\dagger b$, but it is shown in this section that $\kappa(A)$ may lead to an incorrect conclusion about the stability of $x_{\rm ls}$. A refined condition number that is a function of A and b is introduced and it is shown that, unlike $\kappa(A)$, it allows discrimination between stable and unstable solutions $x_{\rm ls}$.

The simplest situation occurs when $m = n$, in which case $x = A^{-1}b$ and

$$\max_{\delta b, b\in\mathbb{R}^m} \frac{\Delta x}{\Delta b} = \|A\| \left\|A^{-1}\right\| = \kappa(A), \quad \Delta x = \frac{\|\delta x\|}{\|x\|}, \quad \Delta b = \frac{\|\delta b\|}{\|b\|}. \tag{4}$$

There are three points to note about $\kappa(A)$. It is assumed that A has full rank.

1. The expression (4) for $\kappa(A)$ assumes A is square and thus b lies in the column space $C(A)$ of A, and $\kappa(A)$ and $\Delta x/\Delta b$ are finite. If, however, $b \notin C(A)$, then $\kappa(A) = \|A\| \left\|A^\dagger\right\|$ is finite but $\Delta x/\Delta b$ may be infinite.

2. It is assumed in the expression (4) for $\kappa(A)$ that there are errors in b only and that A is exact. A more realistic scenario requires that errors in A and b be considered, but the restriction of errors to b allows linear perturbation analysis to be used.
3. The maximum in (4) is taken with respect to all vectors $\delta b, b \in \mathbb{R}^m$, but b is specified in a given problem. It is therefore necessary to consider the maximum value of the ratio ${}^{\Delta x_{\text{ls}}}/_{\Delta b}$, where $\Delta x_{\text{ls}} = {}^{\|\delta x_{\text{ls}}\|}/_{\|x_{\text{ls}}\|}$, with respect to all perturbations δb for the given vector b. This leads to the effective condition number of x_{ls}, and it is considered in Section 2.1.

2.1. The effective condition number

The condition number $\kappa(A)$ is a function of A only, but $x_{\text{ls}} = A^\dagger b$ is a function of A and b, and it is therefore necessary to develop a more general expression for the stability of x_{ls}. This revised measure, which is called the *effective condition number* $\eta(A,b)$, is defined in Definition 2.1, and an expression for it is derived in Theorem 2.1.

Definition 2.1 *(Effective condition number).* The effective condition number $\eta(A,b)$ of x_{ls} is

$$\eta(A,b) = \max_{\delta b \in \mathbb{R}^m} \frac{\Delta x_{\text{ls}}}{\Delta b}, \qquad \Delta x_{\text{ls}} = \frac{\|\delta x_{\text{ls}}\|}{\|x_{\text{ls}}\|},$$

where Δb is defined in (4).

Theorem 2.1. *The effective condition number $\eta(A,b)$ of x_{ls} is*

$$\eta(A,b) = \frac{\left\|A^\dagger\right\| \|b\|}{\|x_{\text{ls}}\|} = \frac{\left\|A^\dagger\right\| \|b\|}{\left\|A^\dagger b\right\|}, \tag{5}$$

where $A \in \mathbb{R}^{m\times n}$.

Proof. It follows from (2) that

$$\|\delta x_{\text{ls}}\| = \left\|A^\dagger \delta b\right\| \le \left\|A^\dagger\right\| \|\delta b\| = \left\|A^\dagger\right\| \|b\| \, \Delta b, \tag{6}$$

and division by $\|x_{\text{ls}}\| = \left\|A^\dagger b\right\|$ yields the expressions (5). Equality in (6) in the 2-norm holds when

$$\left\|A^\dagger \delta b\right\|_2 = \left\|A^\dagger\right\|_2 \|\delta b\|_2 = \left\|\Sigma^\dagger\right\|_2 \|\delta c\|_2, \qquad \delta c = U^T \delta b,$$

where the singular value decomposition (SVD) of A is $U\Sigma V^T$ and the singular values σ_i, $i = 1,\dots,p$, $p = \min(m,n)$, are arranged in non-increasing order, $\sigma_i \ge \sigma_{i+1}$, $i = 1,\dots,p-1$, [9, §2.4]. It follows that $\left\|\Sigma^\dagger\right\|_2 = {}^1/_{\sigma_p}$, and if

$$\delta c = \begin{bmatrix} 0 & \cdots & 0 & \delta c_p & 0 & \cdots & 0 \end{bmatrix}^T = \delta c_p e_p, \qquad \delta c \in \mathbb{R}^m,$$

where $e_p \in \mathbb{R}^m$ is the pth unit basis vector, then $\delta b = U\delta c = \delta c_p \left(U e_p\right)$. Thus,

$$\|\delta b\|_2 = \left|\delta c_p\right| \qquad \text{and} \qquad \left\|A^\dagger \delta b\right\|_2 = \left\|\Sigma^\dagger \delta c\right\|_2 = \frac{\left|\delta c_p\right|}{\sigma_p} = \left\|A^\dagger\right\|_2 \|\delta b\|_2,$$

and hence equality in (6) in the 2-norm is achieved when δb is aligned along the pth column of U.

Equality in (6) in the 1-norm is attained when δb is aligned along the column of $A^\dagger$ whose 1-norm is a maximum. In particular, if $A_i^\dagger$ is the ith column of $A^\dagger$, the index t is defined as

$$t = \arg \max_{i=1,\dots,m} \left\|A_i^\dagger\right\|_1, \qquad \left\|A^\dagger\right\|_1 = \left\|A_t^\dagger\right\|_1,$$

and $\delta b = \delta b_t e_t$, where $\delta b_t \in \mathbb{R}$ and $e_t \in \mathbb{R}^m$, then

$$\left\|A^\dagger \delta b\right\|_1 = |\delta b_t| \left\|A^\dagger e_t\right\|_1 = |\delta b_t| \left\|A_t^\dagger\right\|_1 = \left\|A^\dagger\right\|_1 \|\delta b\|_1 . \quad \square$$

The superiority of $\eta(A,b)$ with respect to $\kappa(A)$ follows because it is a function of A and b, and thus the conditions satisfied by A and b such that x_{ls} is stable, and x_{ls} is unstable, can be deduced. The stability and instability of x_{ls} are defined in Definitions 2.2 and 2.3, respectively.

Definition 2.2 *(Stability).* The vector x_{ls} is stable if a relative error Δb in b of order of magnitude ϵ causes a relative error Δx_{ls} in x_{ls} of the same order of magnitude,

$$\Delta b = \mathcal{O}(\epsilon) \qquad \text{and} \qquad \Delta x_{\text{ls}} = \mathcal{O}(\epsilon).$$

Definition 2.3 *(Instability).* The vector x_{ls} is unstable if a relative error Δb in b of order of magnitude ϵ causes a relative error Δx_{ls} in x_{ls} of much greater magnitude,

$$\Delta b = \mathcal{O}(\epsilon) \qquad \text{and} \qquad \Delta x_{\text{ls}} \gg \mathcal{O}(\epsilon).$$

It is noted that x_{ls} may be stable for a perturbation δb_1 in $b = b_1$, and unstable for a perturbation δb_2 in $b = b_2$.

This information on the stability or instability of x_{ls} cannot be obtained from $\kappa(A)$, which is an advantage of $\eta(A,b)$, but it follows from (5) that $\eta(A,b)$ is stable (unstable) if x_{ls} is stable (unstable) [10, §4]. In particular, it is shown in [11, §3] that a condition number of $\eta(A,b)$ in the 2-norm with respect to a perturbation in b is

$$\frac{\Delta\eta_2(A,b)}{\Delta b} \leq 1 + \eta_2(A,b), \tag{7}$$

to first order in δb, where, following the notation in (4) and (5),

$$\eta_2(A,b) = \frac{\left\|A^{\dagger}\right\|_2 \|b\|_2}{\left\|A^{\dagger}b\right\|_2}, \qquad \Delta\eta_2(A,b) = \frac{\left|\delta\eta_2(A,b)\right|}{\eta_2(A,b)}, \qquad \Delta b = \frac{\|\delta b\|_2}{\|b\|_2},$$

and $\delta\eta_2(A,b)$ is the perturbation in $\eta_2(A,b)$.

The condition numbers $\kappa(A)$ and $\eta(A,b)$ differ because $\kappa(A)$ is finite if A has full rank, but $\eta(A,b)$ is infinite if x_{ls} does not have a component that lies in the column space of A, even if A has full rank, as shown in Example 2.1.

Example 2.1. Consider the matrix A and vector b,

$$A = \begin{bmatrix} 2 & -1 \\ 2 & 0 \\ 1 & -1 \end{bmatrix} \qquad \text{and} \qquad b = \begin{bmatrix} -2 \\ 1 \\ 2 \end{bmatrix}.$$

The condition number of A is $\kappa_2(A) = 3.37$ but it follows from (5) that $\eta(A,b)$ is infinite because $x_{\text{ls}} = 0$ since b lies in the space that is orthogonal to the column space of A,[3]

$$b^T A = \begin{bmatrix} 0 & 0 \end{bmatrix}, \qquad x_{\text{ls}} = \left(A^T A\right)^{-1} A^T b = \begin{bmatrix} 0 \\ 0 \end{bmatrix},$$

and thus $\kappa(A)$ is an incorrect measure of the stability of x_{ls}. □

Example 2.2 considers the variation of $\eta_2(H,b)$, where H is the Hilbert matrix of order 11, as the components of b along the columns of U, where $U\Sigma V^T$ is the SVD of H, change.

Example 2.2. Consider the Hilbert matrix H of order 11. The vectors b_i,

$$b_i = U(:,i-1) + U(:,i) + U(:,i+1), \qquad i = 2,\ldots,10,$$

were formed and the values of $\eta_2(H,b_i)$ of the solutions $x_i = H^{-1}b_i$ of $Hx_i = b_i$ were computed. It follows that b_k is equal to the sum of the $(k-1)$th, kth and $(k+1)$th columns of U, or equivalently,

$$c_i = e_{i-1} + e_i + e_{i+1}, \qquad i = 2,\ldots,10, \qquad c = U^T b.$$

Fig. 1 shows that $\eta_2(H,b_i)$ decreases monotonically as i increases, that is, as the space spanned by b changes from the columns of U associated with the large singular values of H to the space spanned by the columns of U associated with the small singular values of H. The maximum value of $\eta_2(H,b_i)$ occurs when b is a linear combination of the 1st, 2nd and 3rd columns of U, and the minimum value occurs when b is a linear combination of the 9th, 10th and 11th columns of U. The condition number $\kappa_2(H)$ is independent of i and the error between it and $\eta_2(H,b_i)$ increases as i increases. □

Example 2.2 shows that if $m = n$, then $\eta_2(A,b) \approx \kappa_2(A)$ if the dominant components of $b = Uc$ are aligned along the first few columns of U, that is,

$$\left|c_1\right|, \left|c_2\right|, \ldots, \left|c_r\right| \gg \left|c_{r+1}\right|, \left|c_{r+2}\right|, \ldots, \left|c_m\right|, \qquad r \ll m.$$

Also, $\eta_2(A,b) \approx 1$ if the dominant components of b are aligned along the last few columns of U, that is,

$$\left|c_1\right|, \left|c_2\right|, \ldots, \left|c_s\right| \ll \left|c_{s+1}\right|, \left|c_{s+2}\right|, \ldots, \left|c_m\right|, \qquad s \gg 1.$$

These results follow from (5) and the SVD $U\Sigma V^T$ of A because, if $m = n$,

[3] A subscript to denote the norm is not assigned to $\eta(A,b)$ because it follows from (5) that $\eta(A,b) \to \infty$ in the $1, 2$ and ∞-norms.

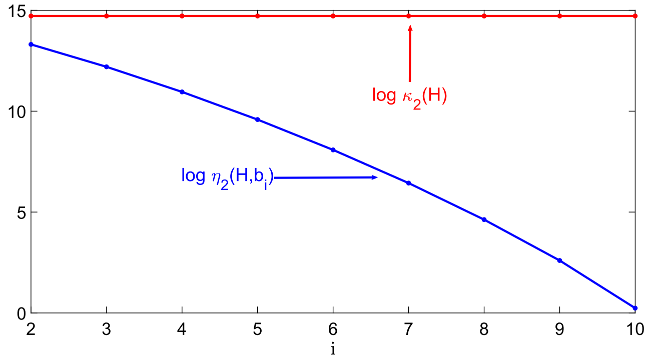


Fig. 1. The condition number $\log_{10} \kappa_2(H)$ and effective condition number $\log_{10} \eta_2(H, b_i)$ against i, for Example 2.2.

$$\eta_2(A,b) = \frac{\left\|A^\dagger\right\|_2 \|b\|_2}{\left\|A^\dagger b\right\|_2} = \left(\frac{1}{\sigma_m}\right) \frac{\|c\|_2}{\left(\sum_{i=1}^m \left(\frac{c_i}{\sigma_i}\right)^2\right)^{\frac{1}{2}}}, \qquad c = \left\{c_i\right\}_{i=1}^m = U^T b.$$

It follows that the limiting values of $\eta_2(A,b)$ are

$$\begin{array}{lllll} \eta_2(A,b) \approx 1 & \text{if} & \left|c_i\right| \to \infty & \text{as} & i \to m, \\ \eta_2(A,b) \approx \kappa_2(A) & \text{if} & \frac{|c_i|}{\sigma_i} \to 0 & \text{as} & i \to m, \end{array}$$

where the condition $\left|c_i\right| \to \infty$ as $i \to m$ implies that the dominant components of c occur for large values of i, and the condition $|c_i|/\sigma_i \to 0$ as $i \to m$ implies that the constants $\left|c_i\right|$ decay to zero faster than the singular values decay to zero. This condition is called the discrete Picard condition [12].

Equation (7) has implications for LASSO regression (3) and Tikhonov regularisation,

$$x_{\mathrm{Tikh}}(\mu) = \arg\min_{x\in\mathbb{R}^n} \left\{\|Ax-b\|_2^2 + \mu \|x\|_2^2\right\}, \qquad \mu \geq 0, \tag{8}$$

because it is shown in [11,13] that an optimal value μ^* of μ exists only if the discrete Picard condition is satisfied, that is, x_{ls} is unstable. It follows, however, from (7) that $\eta(A,b)$ cannot be determined accurately if x_{ls} is unstable, and this problem is addressed by the inclusion of prior information. This is demonstrated by the application of Tikhonov regularisation (8) to image deblurring because exact images satisfy the discrete Picard condition, which is the prior information. Similarly, it must be known *a priori* that the LS problem (1) admits a sparse solution, in which case x_{ls} and $\eta(A,b)$ are unstable.

The association between stability and sparsity in LASSO regression can be motivated by comparison of (3) with Tikhonov regularisation (8) because, as noted above, a necessary condition for a regularised solution $x_{\mathrm{Tikh}}(\mu^*)$ to exist is that x_{ls} is unstable. Furthermore, the application of Tikhonov regularisation to a stable LS problem yields a solution $x_{\mathrm{Tikh}}(\mu)$ whose error is large for all $\mu > 0$. Lasso regression and Tikhonov regularisation differ in the form of the constraint but similarities between them are, nonetheless, expected.

3. Properties of the solution of LASSO regression

A closed form expression for $x_{\mathrm{lasso}}(\lambda)$ does not exist, unless the columns of A are orthonormal, and it is therefore difficult to obtain theoretical results for LASSO regression. Some properties of the solution of LASSO regression can, however, be derived and they are considered in this section. The SVD of $A \in \mathbb{R}^{m\times n}$ is

$$U\left[\,\Sigma \quad 0_{m,n-m}\,\right] V^T = U\left[\,\Sigma \quad 0_{m,n-m}\,\right]\begin{bmatrix} V_1^T \\ V_2^T \end{bmatrix}, \tag{9}$$

where $\Sigma \in \mathbb{R}^{m\times m}$ is the diagonal matrix of the singular values σ_i, $i = 1,\dots,m$, of A, the columns of $V_1 \in \mathbb{R}^{n\times m}$ form an orthonormal basis for the row space of A and the columns of $V_2 \in \mathbb{R}^{n\times(n-m)}$ form an orthonormal basis for the null space of A. It follows from $x_{\mathrm{ls}} = A^\dagger b = V_1 \Sigma^{-1} U^T b$ that the general solution of (1) is z,

$$z = x_{\mathrm{ls}} + V_2 d_1, \qquad Az = A\left(x_{\mathrm{ls}} + V_2 d_1\right) = A x_{\mathrm{ls}} = b, \tag{10}$$

where $d_1 \in \mathbb{R}^{n-m}$ is arbitrary. If the solution of (3) is $x_{\mathrm{lasso}}(\lambda)$ and its residual is $\|r(\lambda)\|_2$ where $r(\lambda) = A x_{\mathrm{lasso}}(\lambda) - b$, then it follows from (10) that

$$A\left(x_{\mathrm{lasso}}(\lambda) - V_1 \Sigma^{-1} U^T b - V_2 d_1\right) = r(\lambda),$$

and thus

$$x_{\text{lasso}}(\lambda) - V_1\Sigma^{-1}U^T b - V_2 d_1 = A^\dagger r(\lambda) + V_2 d_2, \qquad A^\dagger = V_1\Sigma^{-1}U^T,$$

where $d_2 \in \mathbb{R}^n$ is arbitrary. It follows that

$$\begin{aligned} x_{\text{lasso}}(\lambda) &= V_1\Sigma^{-1}U^T\left(b + r(\lambda)\right) + V_2 d, \qquad d = d_1 + d_2, \\ &= \begin{bmatrix} V_1 & V_2 \end{bmatrix} \begin{bmatrix} \Sigma^{-1}U^T\left(b + r(\lambda)\right) \\ d \end{bmatrix} \\ &= V t(A, b, \lambda), \end{aligned} \tag{11}$$

where

$$t(A, b, \lambda) = \begin{bmatrix} \Sigma^{-1}U^T\left(b + r(\lambda)\right) \\ d \end{bmatrix},$$

and the premultiplication of both sides of (11) by V^T yields

$$V_1^T x_{\text{lasso}}(\lambda) = \Sigma^{-1}U^T\left(b + r(\lambda)\right) \qquad \text{and} \qquad V_2^T x_{\text{lasso}}(\lambda) = d.$$

Example 3.1 shows that if a sparse solution $x_{\text{lasso}}(\lambda)$ that has t zero components is desired, then the optimal value(s) of $\lambda > 0$ must satisfy t homogeneous equations and thus the number of equations that must be satisfied by λ increases as t increases. Also, if these t equations do not possess a common solution, then a sparse solution, t of whose components are zero, does not exist. Since $x_{\text{lasso}}(\lambda) \in \mathbb{R}^n$, the maximum number of forms of $x_{\text{lasso}}(\lambda)$ that have t zero components is $\binom{n}{t}$, and furthermore, each of these forms is, in general, defined by a different value of λ.

Example 3.1. Consider (11) as the number of zero entries $\mathcal{N}(\lambda)$ in $x_{\text{lasso}}(\lambda)$ increases. Let $v_i^T \in \mathbb{R}^n$, $i = 1, \ldots, n$, be the ith row of V. The values $\mathcal{N}(\lambda) = 1, 2, k, n$, are considered in order to show the trend in the results as $\mathcal{N}(\lambda)$ increases.

- $\mathcal{N}(\lambda) = 1$: It follows from (11) that if the solutions of q of the n equations

  $$v_i^T t(A, b, \lambda) = 0, \qquad i = 1, \ldots, n,$$

  are $\lambda_1, \lambda_2, \ldots, \lambda_q$, assuming the solution of each of these q equations is unique, then each solution $x_{\text{lasso}}(\lambda_i)$, $i = 1, \ldots, q$, of (3) has only one zero entry.
- $\mathcal{N}(\lambda) = 2$: The values of λ that satisfy the equations

  $$\begin{bmatrix} v_i^T \\ v_j^T \end{bmatrix} t(A, b, \lambda) = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \qquad 1 \le i < j \le n, \tag{12}$$

  for specified values of i and j return a solution $x_{\text{lasso}}(\lambda)$ that has two zero entries. There are $\binom{n}{2}$ pairs of equations (12) and the solutions $x_{\text{lasso}}(\lambda)$ of different pairs of equations, that is, different values of i and j, differ in the predictors whose values are set to zero.
- $\mathcal{N}(\lambda) = k$: The values of λ that satisfy the k equations

  $$\begin{bmatrix} v_{p_1}^T \\ v_{p_2}^T \\ \vdots \\ v_{p_k}^T \end{bmatrix} t(A, b, \lambda) = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \qquad 1 \le p_1 < p_2 < \cdots < p_k \le n, \tag{13}$$

  for specified values of $p_1, p_2, \ldots, p_k$, return a solution $x_{\text{lasso}}(\lambda)$ such that $\mathcal{N}(\lambda) = k$. There are $\binom{n}{k}$ sets of k equations (13) and different sets of k rows of V yield different values of λ. There may not, however, be a value of λ that satisfies any of these sets of equations, in which case there does not exist a solution $x_{\text{lasso}}(\lambda)$ that has k zero entries.
- $\mathcal{N}(\lambda) = n$: This condition implies $x_{\text{lasso}}(\lambda) = 0$ and thus the only solution of (11) is the trivial solution $t(A, b, \lambda) = 0$. It follows from (3) that the condition $\mathcal{N}(\lambda) = n$ is obtained when $\lambda \to \infty$, and thus from (11),

  $$\lim_{\lambda \to \infty} r(\lambda) = -b \qquad \text{and} \qquad \lim_{\lambda \to \infty} d = 0.$$

  The examples in Section 5 show that the solution $x_{\text{lasso}}(\lambda) = 0$ is also obtained for finite values of λ. □

Example 3.1 shows that the probability that $\mathcal{N}(\lambda)$ homogeneous equations from the n components of $x_{\text{lasso}}(\lambda)$ in (11) have an exact solution $\lambda = \lambda_0$ decreases as $\mathcal{N}(\lambda_0)$ increases, and thus the probability of a non-zero residual $\left\| r(\lambda_0) \right\|_2$ increases as $\mathcal{N}(\lambda_0)$ increases. This confirms that an optimal sparse solution has few zero entries ($\mathcal{N}(\lambda_0)$ is small) and it suggests that $\left\| r(\lambda_0) \right\|_2$ increases as λ_0 increases. The examples in Section 5 demonstrate these conclusions.

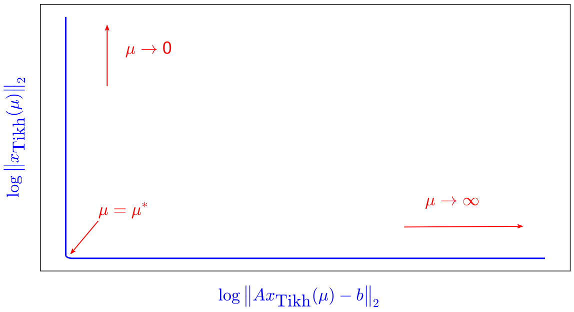


Fig. 2. The L-curve for Tikhonov regularisation.

4. The L-curve and the regularisation parameter

Example 3.1 shows that the value(s) of λ that yield sparse solutions satisfy a set of homogeneous equations (13). These solutions do not, however, necessarily guarantee the fidelity of the model and thus an efficient method for the computation of the optimal value of λ that yields a sparse solution that guarantees the fidelity of the model must be developed. This section considers two methods, cross validation (CV) [5, §2.3] and the L-curve [8, §4.6], for the computation of the optimal value λ^* of λ that yields a solution $x_{\text{lasso}}(\lambda^*)$ that satisfies these two criteria.

Cross validation requires that the data be divided at random into a training set and a test set, and the results from the training set are assessed by computing the response of the test set for a range of values of λ. This process of random division of the data into a training set and a test set, and assessing the results of the training set, is repeated, typically ten times, which yields ten estimates of the prediction error. The average error, for each value of λ, is calculated and it leads to 10-fold CV. The optimal value λ^* of λ is the value of λ for which the mean squared error (MSE), as a function of λ, assumes its minimum value [5, §2.3]. The examples in Section 5 show that this minimum may be shallow, which makes the computation of λ^* difficult.

The L-curve is used for the computation of the optimal value of μ in Tikhonov regularisation (8) and it can be extended to LASSO regression for the computation of the optimal value of λ. Consider initially the L-curve in Tikhonov regularisation, which is a plot of $\log_{10} \|x_{\text{Tikh}}(\mu)\|_2$ against $\log_{10} \|Ax_{\text{Tikh}}(\mu) - b\|_2$ as a function of μ. If the discrete Picard condition [12],

$$\frac{|c_i|}{\sigma_i} \to 0 \qquad \text{as} \qquad i \to m, \qquad c = \{c_i\}_{i=1}^{m} = U^T b, \tag{14}$$

where U is defined in (9), is satisfied and the noise is white, the curve assumes the form of an L, as shown in Fig. 2.[4] As μ increases from zero, $\|x_{\text{Tikh}}(\mu)\|_2$ decreases and $\|Ax_{\text{Tikh}}(\mu) - b\|_2$ is approximately constant, until $\mu = \mu^*$, which is the value of μ in the corner of the L. As μ increases from μ^*, $\|x_{\text{Tikh}}(\mu)\|_2$ is approximately constant and $\|Ax_{\text{Tikh}}(\mu) - b\|_2$ increases. The value μ^* is the optimal value of μ because $\|Ax_{\text{Tikh}}(\mu) - b\|_2$ and $\|x_{\text{Tikh}}(\mu)\|_2$ attain, approximately, their minimum values for $\mu = \mu^*$, and thus this value of μ balances the fidelity of the model and the constraint on $\|x\|_2$.

The discrete Picard condition (14) implies that the constants $|c_i|$ decay to zero faster than the singular values σ_i decay to zero, in which case the effective condition number $\eta_2(A, b)$ of $x_{\text{ls}} = A^{\dagger} b$ is approximately equal to its maximum value, $\kappa_2(A)$, if b lies in the column space of A and the decay (14) is sufficiently rapid [10, §4]. Equation (14) must be satisfied in order that a parametric plot of $\log_{10} \|x_{\text{Tikh}}(\mu)\|_2$ against $\log_{10} \|Ax_{\text{Tikh}}(\mu) - b\|_2$ assume the form of an L [8, §4.6]. If, however, (14) is not satisfied, the curve assumes a different form.

The extension of the L-curve from Tikhonov regularisation to LASSO regression yields a plot of $\log_{10} \|x_{\text{lasso}}(\lambda)\|_1$ against $2\log_{10} \|Ax_{\text{lasso}}(\lambda) - b\|_2$ as a function of λ. The curve assumes the form of an L if x_{ls} is unstable and $x_{\text{lasso}}(\lambda^*)$ has few zero entries, in which case the optimal value λ^* of λ is the value of λ in the corner of the L. The minimum norm solution x_{ls} is used in Tikhonov regularisation but as stated in Section 1, coordinate descent does not select this solution from the infinite number of solutions z in (1). Also, the solutions x_{ls} and $x_{\text{Tikh}}(\mu)$ are orthogonal to the null space of A, but this orthogonality property does not extend to $x_{\text{lasso}}(\lambda)$ because the solution from coordinate descent includes a component that lies in the null space of A. This difference between the solutions of LASSO regression and Tikhonov regularisation may account for differences in the L-curve for these problems.

5. Examples

This section contains three examples that demonstrate the theory in the preceding sections. All computations were performed using MATLAB R2022a, which uses coordinate descent to compute $x_{\text{lasso}}(\lambda)$ for given values of λ. The default values for the number of iterations and threshold for convergence of the algorithm, 10^5 and 10^{-4} respectively, in the function `lasso` were used.

[4] White noise is a stationary time series whose autocorrelation is zero. Thus the correlation coefficient of all pairs of values of white noise taken at different times is zero.

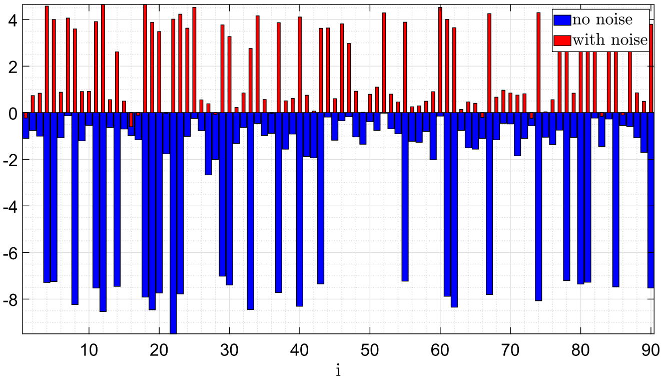


Fig. 3. The components $\log_{10}|x_{\text{ls},i}|$ of $x_{\text{ls}} = A^{\dagger}b$, with and without noise, for Example 5.1.

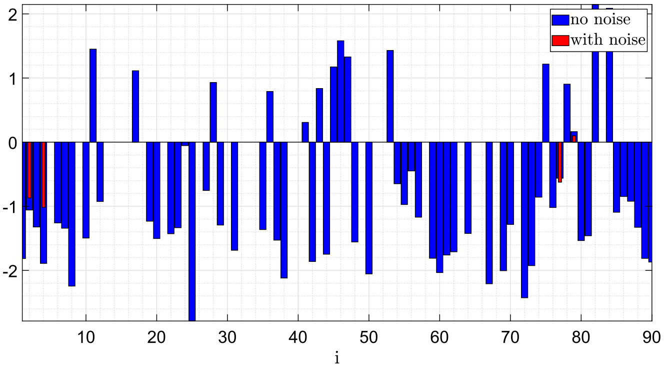


Fig. 4. The components $\log_{10}|x_{\text{lasso},i}(\lambda^*)|$ of $x_{\text{lasso}}(\lambda^*)$, where λ^* is the optimal value of λ computed by 10-fold CV, with and without noise, for Example 5.1.

Example 5.1. Consider a random matrix $A \in \mathbb{R}^{50\times 90}$ and a vector $b \in \mathbb{R}^{50}$ to which noise δb is added such that the signal-to-noise ratio (SNR) is $\|b\|_2/\|\delta b\|_2 = 30$. The condition number and effective condition number are $\kappa_2(A) = 6.87 \times 10^7$ and $\eta_2(A,b) = 7.81 \times 10^6$ respectively.

Fig. 3 shows, on a semi-logarithmic plot, the LS solution $x_{\text{ls}} = A^{\dagger}b$, with and without noise, and the large difference between the two solutions follows from the large value of $\eta_2(A,b)$. The matrix A and vector b were chosen such that 40 components of x_{ls} are about 10,000 times larger than its other 50 components and thus the constraint on $\|x\|_1$ is expected to have the greatest effect on these large components, such that $x_{\text{lasso}}(\lambda^*)$ has 40 zero components.

The solution of (3) using 10-fold CV, with and without noise, is shown in Fig. 4. The solution in the absence of noise has 25 zero components, and the effect of the noise δb is significant because there is a large difference between the solutions with and without noise. It follows that the results using 10-fold CV are computationally unreliable.

Fig. 5 shows the variation of the MSE with respect to λ, with and without noise. The minimum of each curve is very shallow and thus the optimal values of λ are badly defined. Fig. 6 shows the L-curves, with and without noise, and the optimal values of λ are their values in the corners of the curves, which are well defined. The optimal value of λ in the presence of noise is its value in the corner, on the line segment of maximum gradient on the curve, as shown in the graph. The normalised residuals at $\lambda = \lambda^*$ are

$$\begin{array}{lll} \text{No noise:} & \lambda^* = 8.68 \times 10^{-8}, & \dfrac{\|Ax_{\text{lasso}}(\lambda^*)-b\|_2}{\|b\|_2} = 10^{-3}, \\ \text{With noise:} & \lambda^* = 6.05 \times 10^{-7}, & \dfrac{\|Ax_{\text{lasso}}(\lambda^*)-b\|_2}{\|b\|_2} = 1.79 \times 10^{-2}. \end{array}$$

Fig. 7 shows the variation of the number of zero entries $\mathcal{N}(\lambda)$ in $x_{\text{lasso}}(\lambda)$ with λ, with and without noise, and the optimal values λ^* of λ computed by the L-curve and 10-fold CV are marked on the graphs. The value of λ^* computed by the L-curve is numerically stable because $\mathcal{N}(\lambda^*) = 40$, which is the number of components of x_{ls} that are significantly larger than its other 50 components, with and without noise. The values of $\mathcal{N}(\lambda^*)$ computed by 10-fold CV are $\mathcal{N}(\lambda^*) = 25$ in the absence of noise and $\mathcal{N}(\lambda^*) = 85$ in the presence of noise, which shows that the method is numerically unstable, and it explains the large difference in the graphs in Fig. 4. Fig. 8 shows the solution $x_{\text{lasso}}(\lambda^*)$, where λ^* is the optimal value of λ calculated by the L-curve (Fig. 6). The noise does not change the zero or non-zero property of the components of $x_{\text{lasso}}(\lambda^*)$, but it does change the values of its non-zero components. □

Example 5.2. Let $A \in \mathbb{R}^{50\times 90}$ be a random matrix and let noise δb be added to $b \in \mathbb{R}^{50}$ such that $\text{SNR} = 30$. Sixty components of x_{ls} are 10,000 times larger than its other 30 components and thus the constraint on $\|x\|_1$ is expected to have the greatest effect on these

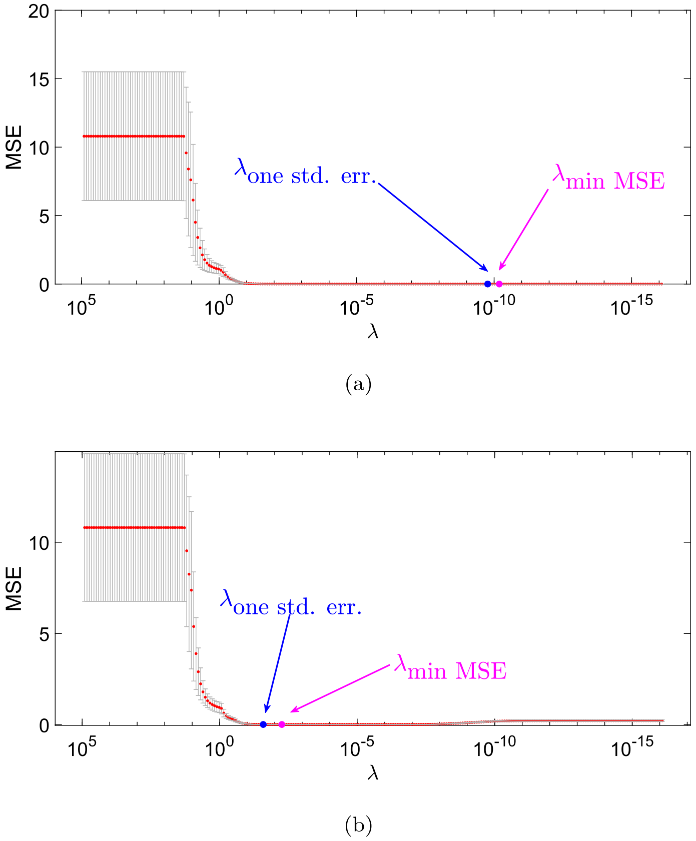


Fig. 5. The variation of the MSE with λ using 10-fold CV (a) without noise, and (b) with noise, for Example 5.1. The figures include error bars for each value of λ and the largest value of λ that is within one standard error of the minimum MSE.

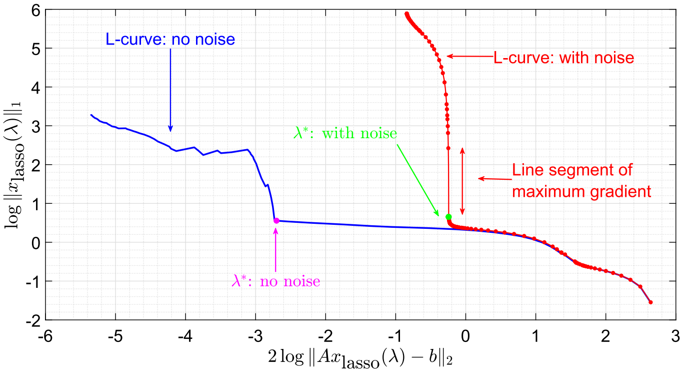
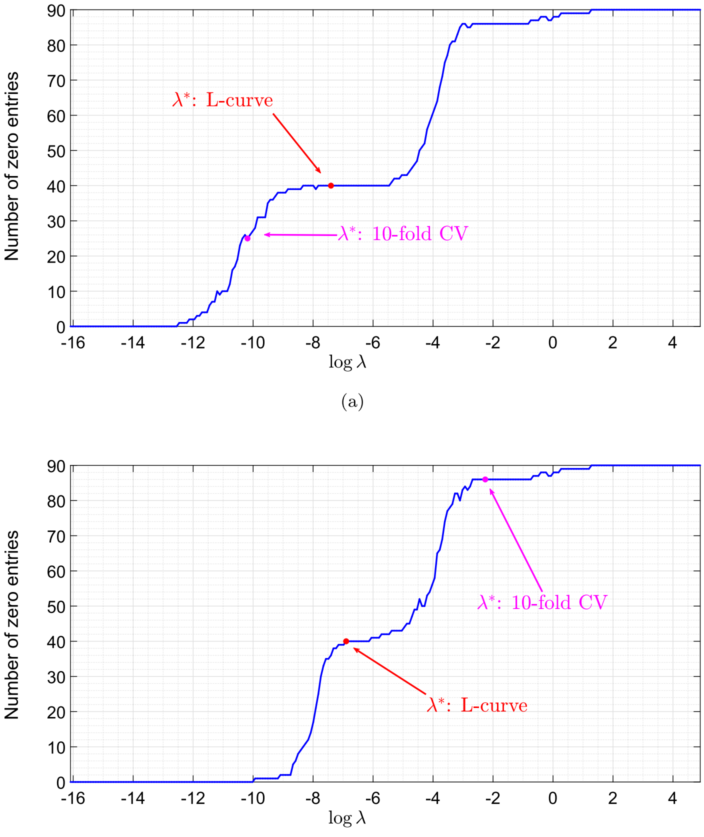


Fig. 6. The L-curves, with and without noise, for Example 5.1.

large components, such that $x_{\text{lasso}}(\lambda^*)$ has 60 zero components. The condition number and effective condition number are $\kappa_2(A) = 83$ and $\eta_2(A, b) = 24.3$ respectively, and thus x_{ls} is stable.

Fig. 9 shows the L-curves, with and without noise, and it is seen that $\|x_{\text{lasso}}(\lambda)\|_1$ is approximately constant and $\|Ax_{\text{lasso}}(\lambda) - b\|_2$ increases as λ increases from $\lambda = 0$ to $\lambda \approx 10^{-12}$. As λ increases further, $\|x_{\text{lasso}}(\lambda)\|_1$ decreases, and $\|Ax_{\text{lasso}}(\lambda) - b\|_2$ increases and it is then approximately constant at 10^{-4}. The figure shows there does not exist a value of λ such that $\|x_{\text{lasso}}(\lambda)\|_1$ and $\|Ax_{\text{lasso}}(\lambda) - b\|_2$ attain, approximately, their minimum values, and Fig. 10 shows the variation of the number of zero entries $\mathcal{N}(\lambda)$ in $x_{\text{lasso}}(\lambda)$ as a function of λ. It is seen that a sparse solution exists for many values of λ, but the figure does not reveal information on the fidelity of the model for each value of λ. This information is readily obtained from Fig. 9, which shows that a sparse solution with a small residual exists for many values of λ. Figs. 9 and 10 show, however, that an optimal value of λ and an optimal sparse solution do not exist. □



(a)

(b)

Fig. 7. The number of zero entries $\mathcal{N}(\lambda)$ in $x_{\text{lasso}}(\lambda)$ as a function of λ (a) without noise, and (b) with noise, for Example 5.1.

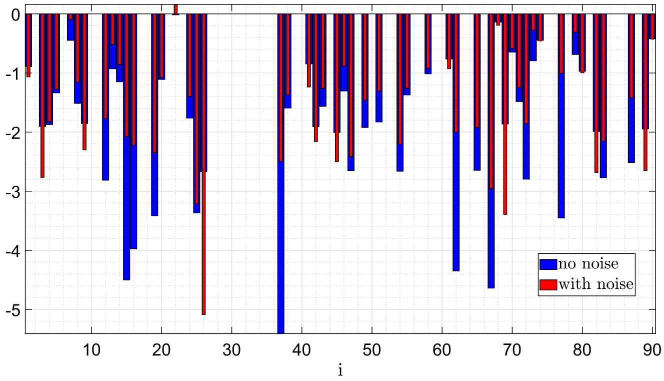


Fig. 8. The components $\log_{10}\left|x_{\text{lasso},i}(\lambda^*)\right|$ of $x_{\text{lasso}}(\lambda^*)$, where λ^* is the optimal value of λ computed by the L-curve, with and without noise, for Example 5.1.

Example 5.3. Let $A \in \mathbb{R}^{50\times 90}$ be a random matrix and let noise δb be added to $b \in \mathbb{R}^{50}$ such that SNR $= 30$. The condition number and effective condition number are $\kappa_2(A) = 7.50 \times 10^7$ and $\eta_2(A,b) = 4.33$ respectively, and thus x_{ls} is stable. Although $\kappa_2(A)$ is very large in Example 5.1 and this example, the values of $\eta_2(A,b)$ show that Example 5.1 and this example define, respectively, an unstable LS problem and a stable LS problem.

Fig. 11 shows the components of x_{ls}, with and without noise, and they are similar because $\eta_2(A,b) = 4.33$. Fig. 12 shows the variation of the MSE with λ, with and without noise, and it is seen that neither graph possesses a unique minimum because the MSE is constant for $\lambda \geq 100$ in both graphs and thus the optimal values of λ satisfy $\lambda^* \geq 100$. It follows from Fig. 14, which shows the variation of the number of zero entries $\mathcal{N}(\lambda)$ in $x_{\text{lasso}}(\lambda)$ with λ, that $x_{\text{lasso}}(\lambda^*) = 0$, with and without noise.

Fig. 13 shows the L-curves, with and without noise, and it is seen that the curves are similar, that they are significantly different from the L-curves in Fig. 6, and that they are similar to the L-curves in Fig. 9. It is noted that the L-curves in Fig. 6 are for an unstable

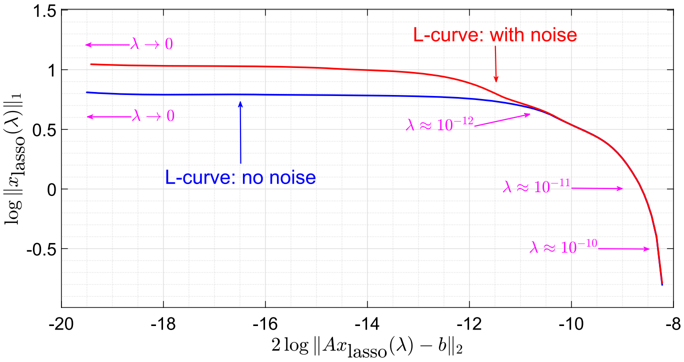


Fig. 9. The L-curves, with and without noise, for Example 5.2.

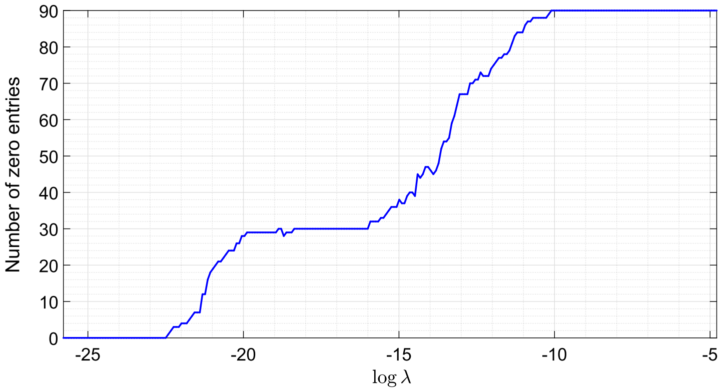


Fig. 10. The number of zero entries $\mathcal{N}(\lambda)$ in $x_{\text{lasso}}(\lambda)$ as a function of λ, for Example 5.2.

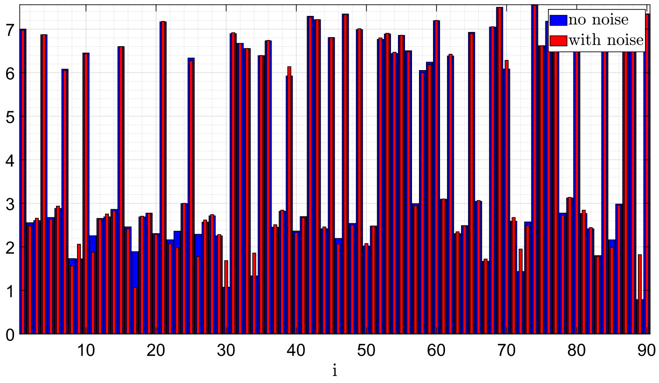


Fig. 11. The components $\log_{10} \left| x_{\text{ls,i}} \right|$ of $x_{\text{ls}} = A^{\dagger} b$, with and without noise, for Example 5.3.

problem, and the L-curves in Figs. 9 and 13 are for stable problems. Fig. 13 shows that there does not exist a value of λ for which $\left\| x_{\text{lasso}}(\lambda) \right\|_1$ and $\left\| Ax_{\text{lasso}}(\lambda) - b \right\|_2$ attain, approximately, their minimum values, and thus an optimal sparse solution does not exist. It is noted that the residuals $\left\| Ax_{\text{lasso}}(\lambda) - b \right\|_2$ are many orders of magnitude larger in Fig. 13 than in Fig. 9.

Fig. 14 shows the variation of the number of zero entries $\mathcal{N}(\lambda)$ in $x_{\text{lasso}}(\lambda)$ with λ, with and without noise. The results for these four scenarios are almost identical and thus the graph in Fig. 14 is representative of the results for all the scenarios. The figure shows that a sparse solution exists for $-6 < \log_{10} \lambda < 1.5$, but Fig. 13 shows that there does not exist a value of λ such that $\left\| Ax_{\text{lasso}}(\lambda) - b \right\|_2$ and $\left\| x_{\text{lasso}}(\lambda) \right\|_1$ assume, approximately, their minimum values, and thus an optimal sparse solution does not exist. □

Some components of x_{ls} in the examples are significantly larger than its other components, and they therefore allowed the ability of the L-curve to identify the number of zero entries in $x_{\text{lasso}}(\lambda^*)$ to be considered. Many other examples, which are not included in the paper, showed that the results are unchanged when an unstable LS problem, without specification of the magnitudes of the components of x_{ls}, is considered. In particular, (i) a sparse solution exists if x_{ls} is unstable and $x_{\text{lasso}}(\lambda^*)$ has few zero components,

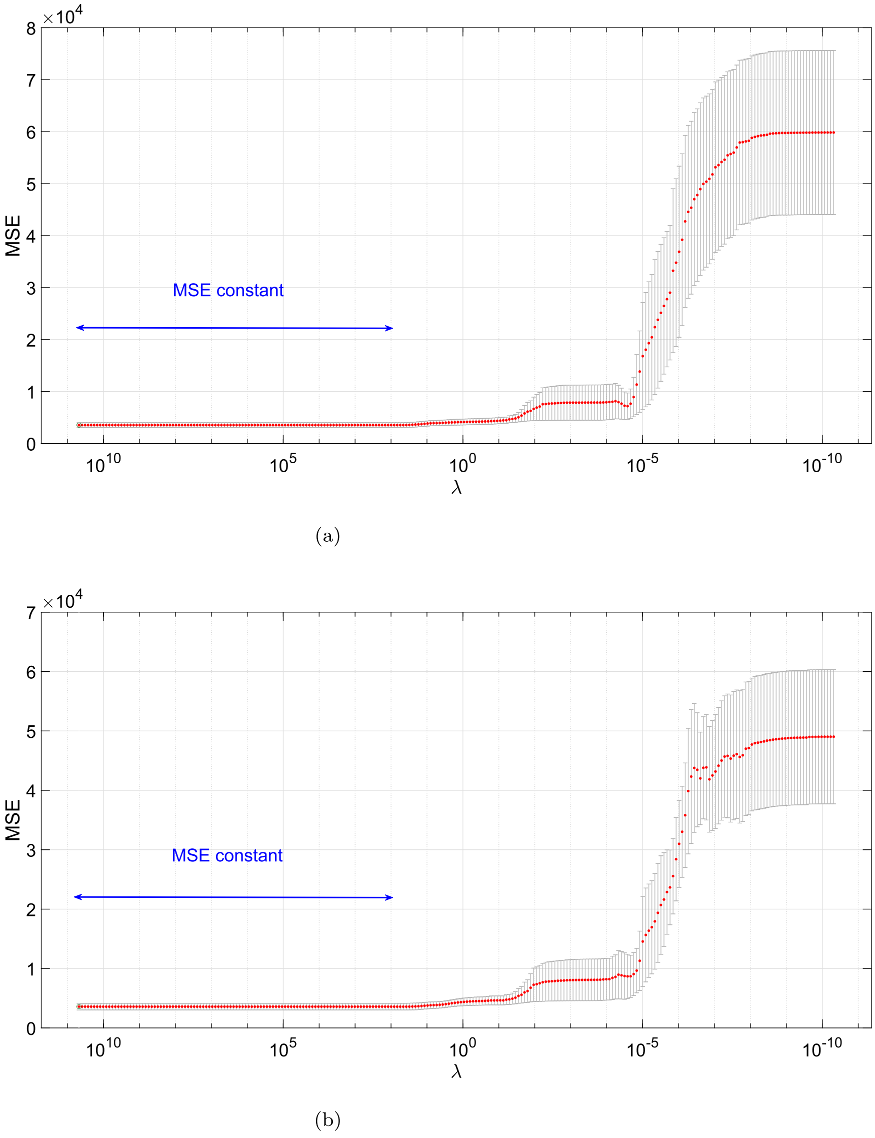


Fig. 12. The variation of the MSE with λ using 10-fold CV (a) without noise, and (b) with noise, for Example 5.3. The figures include error bars for each value of λ and the largest value of λ that is within one standard error of the minimum MSE.

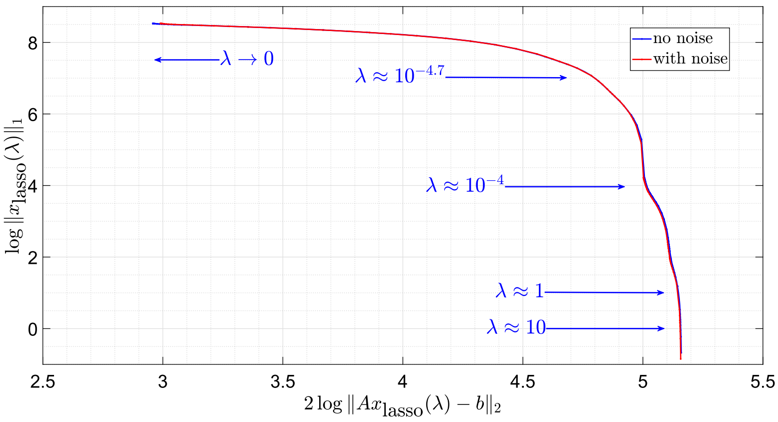


Fig. 13. The L-curves, with and without noise, for Example 5.3.

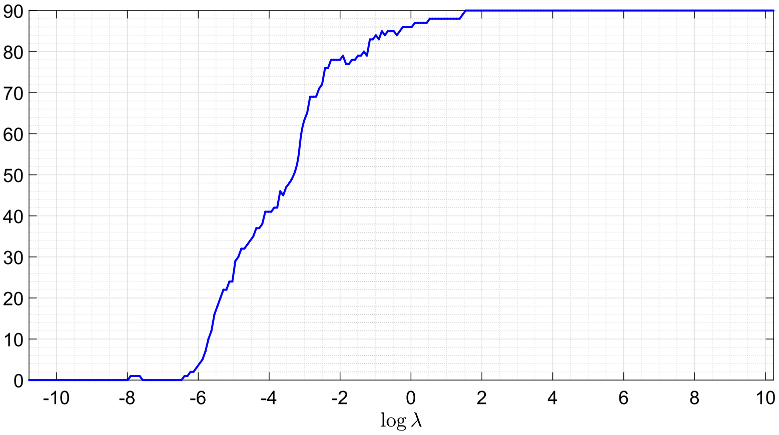


Fig. 14. The number of zero entries $\mathcal{N}(\lambda)$ in $x_{\text{lasso}}(\lambda)$ as a function of λ, from 10-fold CV and the L-curve, with and without noise, for Example 5.3.

in which case the L-curve displays a sharp corner and the number of zero components in $x_{\text{lasso}}(\lambda^*)$ does not change in the presence of noise, (ii) the variation of the MSE with λ displays a shallow minimum and it is therefore difficult to calculate the value of λ^*, (iii) the L-curve returned much better results than 10-fold CV, and (iv) a sparse model can be computed for a wide range of values of λ, but the existence of a sparse model does not guarantee that it is an accurate model, and thus an optimal sparse solution may not exist.

6. Summary

This paper has considered some properties of the solution of LASSO regression for an underdetermined equation and it was shown that they are dependent on the stability, or otherwise, of the solution of the LS problem. In particular, an optimal sparse solution with few zero entries exists if x_{ls} is unstable, and an optimal sparse solution does not exist if x_{ls} is stable. It was also shown that the L-curve yields much better results than CV and it is stable in the presence of noise. The number of equations that must be satisfied by λ increases as the number of zero components in $x_{\text{lasso}}(\lambda)$ increases, and thus an optimal sparse solution that has many zero components may not exist.

References

[1] R. Tibshirani, Regression shrinkage and selection via the lasso, J. R. Stat. Soc. B 58 (1) (1996) 267–288.
[2] J. Friedman, T. Hastie, H. Höfling, R. Tibshirani, Pathwise coordinate optimization, Ann. Appl. Stat. 1 (2) (2007) 302–322.
[3] T.T. Wu, K. Lange, Coordinate descent algorithms for lasso penalized regression, Ann. Appl. Stat. 2 (1) (2008) 224–244.
[4] B. Efron, T. Hastie, I. Johnstone, R. Tibshirani, Least angle regression, Ann. Stat. 32 (2) (2004) 407–499.
[5] T. Hastie, R. Tibshirani, M. Wainwright, Statistical Learning with Sparsity: The Lasso and Generalizations, Chapman and Hall, New York, USA, 2015.
[6] H. Xu, C. Caramanis, S. Mannor, Sparse algorithms are not stable: a no-free-lunch theorem, IEEE Trans. Pattern Anal. Mach. Intell. 34 (1) (2012) 187–193.
[7] L. Baldassarre, M. Pontil, J. Mourão-Miranda, Sparsity is better with stability: combining accuracy and stability for model selection in brain decoding, https://doi.org/10.3389/fnins.2017.00062, 2017.
[8] P.C. Hansen, Rank-Deficient and Discrete Ill-Posed Problems: Numerical Aspects of Linear Inversion, SIAM, Philadelphia, USA, 1998.
[9] G.H. Golub, C.F. Van Loan, Matrix Computations, John Hopkins University Press, Baltimore, USA, 2013.
[10] J.R. Winkler, M. Mitrouli, C. Koukouvinos, The application of regularisation to variable selection in statistical modelling, J. Comput. Appl. Math. 404 (2022) 113884.
[11] J.R. Winkler, Regularisation, overfitting and condition estimation in regression, 2024, submitted for publication.
[12] P.C. Hansen, The discrete Picard condition for discrete ill–posed problems, BIT Numer. Math. 30 (1990) 658–672.
[13] J.R. Winkler, M. Mitrouli, Condition estimation for regression and feature estimation, J. Comput. Appl. Math. 373 (2020) 112212.