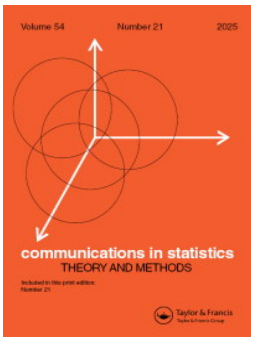

Taylor & Francis
Taylor & Francis Group

Communications in Statistics - Theory and Methods

ISSN: 0361-0926 (Print) 1532-415X (Online) Journal homepage: www.tandfonline.com/journals/lsta20

Analysis and implications of a negative parameter in Tikhonov regularisation

Joab R. Winkler & Marilena Mitrouli

To cite this article: Joab R. Winkler & Marilena Mitrouli (30 Sep 2025): Analysis and implications of a negative parameter in Tikhonov regularisation, Communications in Statistics - Theory and Methods, DOI: 10.1080/03610926.2025.2540866

To link to this article: https://doi.org/10.1080/03610926.2025.2540866

Published online: 30 Sep 2025.

COMMUNICATIONS IN STATISTICS—THEORY AND METHODS
https://doi.org/10.1080/03610926.2025.2540866

Taylor & Francis
Taylor & Francis Group

OPEN ACCESS

Analysis and implications of a negative parameter in Tikhonov regularisation

Joab R. Winkler[a] and Marilena Mitrouli[b]

[a]School of Computer Science, The University of Sheffield, Sheffield, UK; [b]Department of Mathematics, National and Kapodistrian University of Athens, Athens, Greece

ABSTRACT

The application of Tikhonov regularisation to the least squares (LS) problem arises frequently in machine learning, for example, in regression and the calculation of the excess risk (out-of-sample prediction error) from a given set of noisy observations. It requires the minimisation with respect to x of a function $f(x, \lambda)$, where λ is the regularisation parameter. If $\lambda \geq 0$, there exists an optimal value λ_{opt} of λ such that the vector $x(\lambda_{\text{opt}})$ that minimises $f(x, \lambda)$ is numerically stable and its error with respect to $x(0)$ is small. It has been claimed that λ_{opt} may be negative, and the aim of this article is the analysis of the consequences of this condition. It is shown theoretically that the condition $\lambda < 0$ yields a family of solutions $x(\lambda)$, each of whose members has a large error and is unstable. Furthermore, the L-curve, which is a method for the calculation of the value of λ_{opt}, yields a good result for $\lambda \geq 0$, and it also shows that $\lambda < 0$ yields unsatisfactory solutions. The L-curve implies, therefore, that $\lambda_{\text{opt}} \geq 0$, which is in accord with the theoretical analysis. Examples of LS problems that consider $\lambda < 0$ and $\lambda \geq 0$ are shown, and the unsatisfactory results for $\lambda < 0$ are evident.

ARTICLE HISTORY
13 February 2025
24 July 2025

KEYWORDS
Least squares; regularisation; negative regularisation parameter; condition estimation; regularisation error; L-curve

1. Introduction

It has been observed experimentally that overparameterised models in deep learning, that is, models for which $n \ll p$, where n is the number of data points and p is the number of predictors, that interpolate training data generalise well on new data, with little or no regularisation. This result is counter to classical theory, which suggests that models that overfit training data require significant regularisation for them to generalise on new data. There has therefore been research to understand this result and it is motivated by the many practical problems in which the condition $n \ll p$ arises, for example, variable selection in statistics, chemometrics and genomics.

Regularisation is used in statistics to reduce the complexity of a model in order that it has good generalisation properties, that is, it yields good results on new data. Variable selection is an example of this reduction in complexity because it allows the determination of the features that have the greatest effect on the response variable (Buccini et al. 2023; Koukoudakis et al. 2025; Winkler et al. 2022). Also, regularisation allows the computation of

CONTACT Joab Winkler j.r.winkler@sheffield.ac.uk School of Computer Science, The University of Sheffield, Regent Court, 211 Portobello, Sheffield S1 4DP, UK.

a stable solution of an ill conditioned set of linear algebraic equations, which arises in inverse problems.

Regularisation is applied to the least squares (LS) problem by the addition of a constraint on the norm of the solution x of the unconstrained LS problem, such that it is required to minimise the function $f(x, \lambda)$ with respect to x,

$$f(x, \lambda) = \|Ax - b\|^2 + \lambda \, \|x\|^2 , \qquad A \in \mathbb{R}^{n \times p}, \tag{1}$$

where $\|\cdot\| = \|\cdot\|_2$, rank $A = \min(n, p)$ and λ is the regularisation parameter. The vector $x(\lambda)$ that minimises $f(x, \lambda)$ is

$$x(\lambda) = \arg \min_{x \in \mathbb{R}^p} f(x, \lambda) = \left(A^T A + \lambda I_p\right)^{-1} A^T b, \tag{2}$$

where I_p is the identity matrix of order p. The stationary points of $f(x, \lambda)$ are defined by the value of λ:

- If $n \geq p$ and $\lambda > -\sigma_s^2 = -\sigma_p^2$, where $\sigma_i, i = 1, \ldots, s = \min(n, p)$, are the singular values of A, arranged in non-increasing order, then $x(\lambda)$ is the global minimum of $f(x, \lambda)$ because $A^T A + \lambda I_p$ is positive definite. If $n < p$, the expression (2) is written in a slightly different form because $A^T A$ is singular, and this modified form shows that $x(\lambda)$ is the global minimum of $f(x, \lambda)$ if $\lambda > -\sigma_s^2 = -\sigma_n^2, \lambda \neq 0$.
- If $-\sigma_1^2 \leq \lambda \leq -\sigma_s^2$, then $x^T (A^T A + \lambda I_p) x$ may be less than zero, equal to zero, or greater than zero, depending on the form of x, and thus $A^T A + \lambda I_p$ is indefinite. Also, $\|x(\lambda)\| \to \infty$ as $\lambda \to -\sigma_i^2, i = 1, \ldots, s$, because $A^T A + \lambda I_p$ is singular.
- If $\lambda < -\sigma_1^2$, then $x(\lambda)$ is the global maximum of $f(x, \lambda)$ because $A^T A + \lambda I_p$ is negative definite.

The second and third points are inconsistent with the minimisation function in (2).

A spiked covariance model is used in (Kobak et al. 2020) to analyse microarray data of $p = 3116$ genes in $n = 64$ rats. It is claimed the optimal value λ_{opt} of λ may be negative if p is much larger than n, where λ_{opt} is the value of λ such that $\left\|Ax(\lambda_{\text{opt}}) - b\right\|$ and $\left\|x(\lambda_{\text{opt}})\right\|$ are minimised approximately. The objective of this article is consideration of the properties of $x(\lambda)$ for $\lambda < 0$ and it is shown that $\lambda < 0$ yields a family of solutions $x(\lambda)$ that have undesirable properties, such that a solution $x(\lambda)$ for $\lambda < 0$ cannot be considered:

- The solutions $x(\lambda)$ for $\lambda < 0$ are unstable with respect to a perturbation in b, but $x(\lambda_{\text{opt}})$ is stable with respect to a perturbation in b, where $\lambda_{\text{opt}} \geq 0$ is the optimal value of λ.
- The error $\|x(\lambda) - x(0)\| / \|x(0)\|$ and residual $\|Ax(\lambda) - b\| / \|b\|$ for $\lambda < 0$ are much greater than their values for $\lambda \geq 0$. Furthermore, these error measures increase monotonically to their maximum value of one as $\lambda > 0$ increases. Their properties for $\lambda < 0$ are significantly different because they have many local minima and they are unbounded as $\lambda \to -\sigma_i^2, i = 1, \ldots, s$.
- The L-curve is a parametric plot of $\log_{10} \|Ax(\lambda) - b\|$ against $\log_{10} \|x(\lambda)\|$ that allows the value of λ_{opt} to be calculated (Hansen 1998, §4.6). If $\lambda > 0$ and the LS problem is ill conditioned, then λ_{opt} is the value of λ in the corner of the L, but the curve has a very different form if $\lambda < 0$, from which it is clear there does not exist an optimal value of λ that is negative.

The main claim of this article follows from these three points, specifically, $\lambda < 0$ cannot be considered because it yields solutions $x(\lambda)$ that are computationally unreliable. The optimal value of λ achieves a balance between the fidelity of the model and the suppression of the effects of noise in the data, and it is shown that this balance requires $\lambda \geq 0$. The conclusion of this article is achieved by error analysis and refined condition estimation of the LS problem, and examples of overdetermined and underdetermined LS problems that demonstrate the theory are shown.

Section 2 reviews work in which the consequences of a negative value of λ are analysed, and regularisation is considered in Section 3. The numerical stability of $x(\lambda)$ with respect to a perturbation in b, and the error and residual of $x(\lambda)$, are considered in Sections 4 and 5 respectively. These sections include an example of regression using exponential basis functions and the results show that $x(\lambda)$ is unstable and has a large error if $\lambda < 0$, which must be compared with the results for $x(\lambda_{\text{opt}})$, $\lambda_{\text{opt}} \geq 0$, which is stable and has a small error. The L-curve, which is a method for calculating the value of λ_{opt}, is discussed in Section 6 and it is shown that the results are consistent with the results in Sections 3, 4 and 5 because the form of the L-curve shows that λ cannot be negative, and that λ, and therefore λ_{opt}, must be greater than or equal to zero. Refined condition estimation and error analysis are used in Section 7 to analyse the simulation in (Kobak et al. 2020, § 2). The solution $x(0)$ of the LS problem for this data is stable, which differs from the solution of the LS problem for the example of regression in Sections 4, 5, and 6, which is unstable. The article is summarised in Section 8.

2. Related work

The literature on regularisation for $\lambda > 0$ is extensive, but there has been much less consideration of the properties of $x(\lambda)$ for $\lambda < 0$. The first work in machine learning that considers $\lambda < 0$ is (Kobak et al. 2020), and more detailed consideration of this condition is in (LeJeune et al. 2024, §6.2; Patil, Du, and Tibshirani 2024, pp. 24–26; Tsigler and Bartlett 2023, §8; Wu and Xu 2020, §5). These papers consider linear regression,

$$a_i^T\theta = b_i + \epsilon_i, \qquad i = 1, \dots, n, \tag{3}$$

where the entries of each feature vector $a_i \in \mathbb{R}^p$ are independent and identically distributed random variables with zero mean, $\epsilon_i \in \mathbb{R}$ is a noise sample whose mean is zero, $b_i \in \mathbb{R}$ is the corresponding response variable and $\theta \in \mathbb{R}^p$ is the parameter vector. The determination of the properties of the excess risk of a new data sample $(\tilde{a}, \tilde{b})$, where $\tilde{a} \in \mathbb{R}^p$ and $\tilde{b} \in \mathbb{R}$, for overparameterised models is the focus of much research because, as noted in Section 1, these models interpolate training data and yield excellent results on new data with little or no regularisation, which is inconsistent with established knowledge.

The application of random projections to overparameterised systems is considered in (LeJeune et al. 2024, §6.2), and the condition that leads to $\lambda < 0$ is considered. The properties of λ_{opt} and the excess risk when the distributions of the training data and test data differ are considered in (Patil, Du, and Tibshirani 2024, Thm. 4) and it is shown that λ_{opt} may be negative in an overparameterised system if covariate shift (a change in the distribution of the training data and the new data) occurs. Also, the differences between positive and negative regularisation parameters for underdetermined and overdetermined LS problems are shown geometrically in (Patil, Du, and Tibshirani 2024, pp. 24-26). Sufficient conditions for $\lambda_{\text{opt}} < 0$ to hold are established in (Tsigler and Bartlett 2023, §8) from the upper bound of the excess risk for some negative values of λ. The effect of noise on models derived from linear regression is considered in (Ullah and Welsh 2024, p. 14 and §3.6) and it is claimed that a negative value of λ may reduce shrinkage, that is, the effects of sampling variation, due to noise.

The excess risk of $\theta(\lambda)$ for the generalised ridge regression,

$$\theta(\lambda) = \left(A^T A + \lambda \Sigma_w\right)^{-1} A^T b, \tag{4}$$

where Σ_w is a positive definite weighting matrix, is considered in (Wu and Xu 2020, §5). The conditions that define the sign of λ_{opt}, based on the covariance matrix of A and the prior on the covariance matrix of the true coefficients of the predictors, in the limit $p/n \to \gamma \in (1, \infty)$ are derived. Also, a negative value of λ has been used in linear models in climatology (Cannon 2009) and in the calculation of estimators for variable selection (Hua and Gunst 1983).

3. Regularisation

Regularisation is a method for the computation of an approximate and stable solution of an ill conditioned problem by the formation of a neighbouring well conditioned problem, such that the error in this approximate solution with respect to the exact solution of the ill conditioned problem is small. Regularisation is applied to ill conditioned problems that include the calculation of the joint angles of a robot when it loses one or more degrees of freedom (Berthet-Rayne et al. 2018), the computation of the solution of fractional diffusion equations (Djennadi, Shawagfeh, and Arqub 2021a, 2021b; Djennadi et al. 2021) and the restoration of a blurred image to its exact form (Hansen, Nagy, and O'Leary 2006). The well conditioned problem is formed by the addition of a penalty whose effect is the inclusion of a property in the approximate solution that the theoretically exact solution must satisfy.

The application of regularisation to the LS problem yields (1) and the vector $x(\lambda)$ that minimises $f(x, \lambda)$ is stated in (2), where the regularisation parameter λ controls the severity with which the constraint on $\|x(\lambda)\|$ is imposed. The value $\lambda = 0$ yields the LS problem, $\|x(\lambda)\| \to 0$ as $\lambda \to \infty$, and the L-curve, which is considered in Section 6, is a method for the computation of the optimal value of λ. The expression for $x(\lambda)$ in (2) is not valid for $\lambda = 0$ if $n < p$ because $A^T A$ is singular. It must therefore be written in a different form, and this is considered in Lemma 1.

Lemma 1. *If $A^T A + \lambda I_p$ and $AA^T + \lambda I_n$ are non-singular, then*

$$(A^T A + \lambda I_p)^{-1} A^T \equiv A^T (AA^T + \lambda I_n)^{-1}, \tag{5}$$

where $\lambda \neq 0$ if A is strictly rectangular.

It follows from (5) and the singular value decomposition (SVD) $U\Sigma V^T$ of A that (2) can be written as

$$x(\lambda) = \begin{cases} \left(A^T A + \lambda I_p\right)^{-1} A^T b = V\left(\Sigma^T \Sigma + \lambda I_p\right)^{-1} \Sigma^T U^T b, & n \geq p, \\ A^T \left(AA^T + \lambda I_n\right)^{-1} b = V\Sigma^T \left(\Sigma\Sigma^T + \lambda I_n\right)^{-1} U^T b, & n < p, \end{cases} \tag{6}$$

and the forms on the right hand side of this equation are used in the sequel.

It is stated above that $x(\lambda_{\text{opt}})$ is an acceptable approximation to $x(0)$ if the regularisation error is small, and $x(\lambda_{\text{opt}})$ must be stable with respect to a perturbation in b. This leads to the premise on which regularisation is based:

$$\begin{tikzcd}
\begin{array}{l}\text{Problem } \mathcal{P}\text{:} \\ x(0) = \arg\min_z \|Az-b\|^2\end{array} \arrow[r, "\text{neighbouring problem}"] \arrow[d, "\mathcal{P} \text{ is ill conditioned}"'] \arrow[dr, "\begin{array}{l}\text{computed} \\ \text{solution of} \\ \text{problem } \mathcal{P}\end{array}"] & \begin{array}{l}\text{Problem } \mathcal{P}^*\text{:} \\ x(\lambda) = \arg\min_z \left\{ \|Az-b\|^2 + \lambda \|z\|^2 \right\}\end{array} \arrow[d, "\mathcal{P}^* \text{ is well conditioned}"] \\
x(0) = A^\dagger b & \begin{array}{l} x(\lambda_{\text{opt}}) = (A^TA + \lambda_{\text{opt}} I)^{-1} A^T b, \;\; m \geq n \\ x(\lambda_{\text{opt}}) = A^T(AA^T + \lambda_{\text{opt}} I)^{-1} b, \;\; m < n \end{array}
\end{tikzcd}$$

Figure 1. The LS problem $\mathcal{P}$ whose solution $x(0)$ is unstable and the neighbouring problem $\mathcal{P}^*$ whose solution $x(\lambda_{\text{opt}})$ is stable and regularisation error is small.

There is a trade-off between the regularisation error and the stability of the regularised solution $x(\lambda_{\text{opt}})$: *The solution* $x(\lambda_{\text{opt}})$ *is accepted because (i) its error with respect to* $x(0)$ *is small, and (ii) it is stable, and much more stable than* $x(0)$.

This trade-off between the regularisation error and the stability of the regularised solution is shown in Figure 1. The figure shows the problem $\mathcal{P}$, which is the LS problem whose solution $x(0)$ is unstable, and the neighbouring problem $\mathcal{P}^*$ whose solution $x(\lambda_{\text{opt}})$ is stable and a very good approximation to $x(0)$. This trade-off is acceptable, that is, the regularisation error is small and $x(\lambda_{\text{opt}})$ is stable with respect to a perturbation in b, if the discrete Picard condition, which is a condition on the rate of decay of the singular values of A, is satisfied (Hansen 1998, p. 82). This condition requires that

$$\frac{|c_i|}{\sigma_i} \to 0 \qquad \text{as} \qquad i \to s = \min(n,p), \qquad c = \{c_i\}_{i=1}^{n} = U^T b, \tag{7}$$

and it is shown in Section 4 that it can be derived from a refined condition number of the LS problem. If (7) is satisfied, there exists an optimal value λ_{opt} of λ such that regularisation yields a stable solution $x(\lambda_{\text{opt}})$ whose error is small (Winkler and Mitrouli 2020). If, however, (7) is not satisfied, regularisation yields an unacceptable solution because the regularisation error is large for all values of $\lambda > 0$. The properties of $x(\lambda_{\text{opt}})$ with respect to condition estimation and overfitting for $\lambda \geq 0$ are considered in (Winkler 2024), but the properties of $x(\lambda)$ for $\lambda < 0$ are significantly different and they are considered in Section 3.1.

The application of regularisation requires that (7) be satisfied by the theoretically exact solution $x(0)$, but it is not satisfied in the presence of noise $\delta b = U\delta c$. The solution of the LS problem in the presence of noise is $x(0) + \delta x(0)$ and it is shown in (Winkler and Mitrouli 2020, Fig. 5) that if $x(0)$ satisfies (7), then

$$\sum_{i=1}^{s} \frac{|c_i + \delta c_i|}{\sigma_i} \approx \frac{|\delta c_s|}{\sigma_s} \qquad \text{and} \qquad x(0) + \delta x(0) \approx \left(\frac{|\delta c_s|}{\sigma_s} \right) v_s, \tag{8}$$

in the presence of noise, where v_s is the sth column of V and $s = \min(n, p)$. It follows that the solution of the LS problem in the presence of noise is dominated by the noise and the small singular values of A, and it is concluded that:

- The theoretically exact solution $x(0)$ is dominated by the large singular values of A if (7) is satisfied.
- If $x(0)$ satisfies (7), then the perturbed solution $x(0) + \delta x(0)$ is dominated by the small singular values of A.
- Regularisation removes the components of $x(0) + \delta x(0)$ that are associated with the small singular values of A, such that it is dominated by the large singular values of A, which yields a small regularisation error.

3.1. Properties of $x(\lambda)$ for $\lambda < 0$

It is stated in Section 1 that $x(\lambda)$ is the unique minimum of $f(x,\lambda)$ if $\lambda > 0$, but $f(x,\lambda)$ may not possess a minimum if $\lambda < 0$. This section considers the properties of $x(\lambda)$ as λ assumes different values, which allows the nature of the stationary point(s) of $f(x,\lambda)$ to be analysed.

Example 1. Consider the matrix A and the diagonal matrix Σ of its singular values,

$$A = \begin{bmatrix} -2 & 1 \\ 3 & -2 \\ -4 & -3 \end{bmatrix}, \quad A^TA = \begin{bmatrix} 29 & 4 \\ 4 & 14 \end{bmatrix}, \quad \Sigma = \begin{bmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} \sqrt{30} & 0 \\ 0 & \sqrt{13} \\ 0 & 0 \end{bmatrix}.$$

Figure 2 shows the surface $h(x,\lambda)$ for six values of λ,

$$h(x,\lambda) = x^T\left(A^TA + \lambda I_2\right)x = z^T\left(\Sigma^T\Sigma + \lambda I_2\right)z, \quad z = V^Tx, \quad x = \begin{bmatrix} x_1 & x_2 \end{bmatrix}^T. \tag{9}$$

1. If $\lambda = -\sigma_1^2 = -30$, then $A^TA + \lambda I_2$ is negative semidefinite and the maximum value of $h(x,\lambda)$ is zero. This value occurs at all points $r = \alpha[x_1\ x_2]^T = \alpha[4\ 1]^T$, where α is an arbitrary constant and $(A^TA - 30I_2)r = 0$.
2. If $\lambda = -\sigma_2^2 = -13$, then $A^TA + \lambda I_2$ is positive semidefinite and the minimum value of $h(x,\lambda)$ is zero. This value occurs at all points $r = \alpha[x_1\ x_2]^T = \alpha[1\ \ -4]^T$, where α is an arbitrary constant and $(A^TA - 13I_2)r = 0$.
3. The surface $h(x,\lambda)$ has a saddle point for $\lambda = -20$ and $\lambda = -25$. □

Example 1 shows that $f(x,\lambda)$ does not necessarily have a minimum if $\lambda < 0$ and thus this condition in (1) requires further consideration. The differences between the properties of $x(\lambda)$ for $\lambda \geq 0$ and $\lambda < 0$ show that the stability of $x(\lambda)$ and the regularisation error for $\lambda < 0$, that is, the fundamental properties of $x(\lambda)$ to be considered when regularisation is applied, must be addressed. These topics are considered in Sections 4 and 5, respectively.

3.2. The filters $f_i(\lambda)$

This section introduces filters $f_i(\lambda), i = 1,\dots,s = \min(n,p)$, in which $x(\lambda)$ can be expressed and it is shown they define the stability and error of $x(\lambda)$.

It follows from (6) and the SVD of A that $x(\lambda)$ can be written as

$$x(\lambda) = \sum_{i=1}^{s}\left(\frac{\sigma_i^2}{\sigma_i^2 + \lambda}\right)\left(\frac{c_i}{\sigma_i}\right)v_i = \sum_{i=1}^{s}\left(f_i(\lambda)\left(\frac{c_i}{\sigma_i}\right)\right)v_i, \tag{10}$$

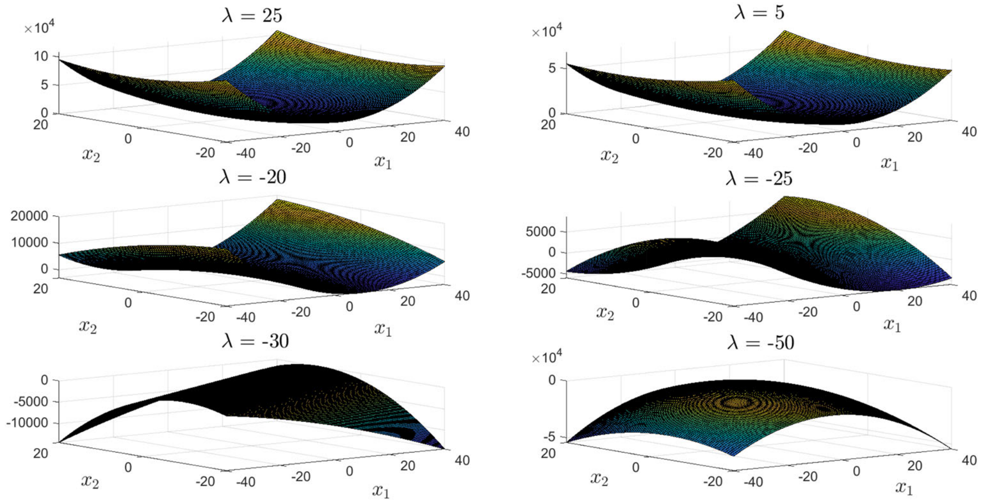

Figure 2. The surface (9) for $-40 \le x_1 \le 40$, $-20 \le x_2 \le 20$ and $\lambda = 25, 5, -20, -25, -30, -50$, for Example 1.

where v_i is the ith column of V, the right singular matrix of A from its SVD, and

$$f_i(\lambda) = \frac{\sigma_i^2}{\sigma_i^2 + \lambda}, \qquad i = 1, \ldots, s, \tag{11}$$

is a filter that removes the coefficient c_i/σ_i from $x(\lambda)$ if $\lambda > 0$, and more coefficients, and therefore more singular values, are removed as λ increases because $f_i(\lambda)$ is a monotonically decreasing function and $f_i(\lambda) > 0$. The properties of $f_i(\lambda)$ for $\lambda < 0$ are different because it is infinite at $\lambda = -\sigma_i^2$, it is positive for $\lambda > -\sigma_i^2$, and it is negative for $\lambda < -\sigma_i^2$. It follows that the condition $\lambda < 0$ may amplify some or all of the components of $x(\lambda)$, and it is shown in Section 4 it may lead to a decrease in the stability of $x(0)$, which is undesirable. The error and stability of $x(\lambda)$ require that bounds on the minimum and maximum eigenvalues of the product of symmetric positive definite matrices be considered, and they are established in Theorem 1.

Theorem 1. *Let P and Q be symmetric positive definite matrices and let $\mu(P)$, $\mu(Q)$ and $\mu(PQ)$ be the set of eigenvalues of P, Q and PQ, respectively. Bounds on the minimum and maximum eigenvalues of PQ are*

$$\mu_{\min}(PQ) \ge \mu_{\min}(P)\mu_{\min}(Q) \qquad \textit{and} \qquad \mu_{\max}(PQ) \le \mu_{\max}(P)\mu_{\max}(Q). \tag{12}$$

Proof. The bound on the maximum eigenvalue of PQ follows from $\|PQ\| \le \|P\|\,\|Q\|$ and $\mu_{\max}(P) = \|P\|$ for a symmetric positive definite matrix P.

The bound on the minimum eigenvalue of PQ follows from $\left\|(PQ)^{-1}\right\| \le \left\|P^{-1}\right\|\left\|Q^{-1}\right\|$ and the eigenvalues of a non-singular matrix X, which are equal to the reciprocals of the eigenvalues of X^{-1}. It therefore follows that

$$\frac{1}{\mu_{\min}(PQ)} \le \left(\frac{1}{\mu_{\ \min}(P)}\right)\left(\frac{1}{\mu_{\min}(Q)}\right), \tag{13}$$

which establishes the result. □

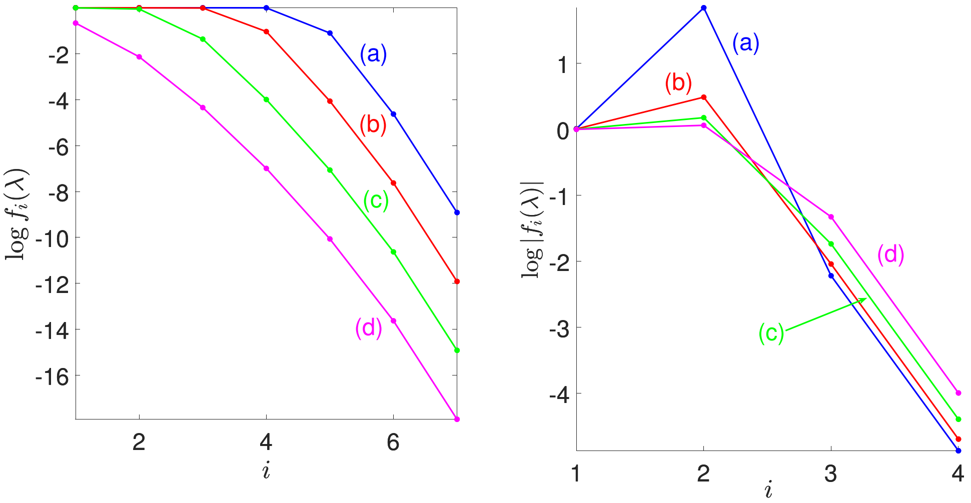

Figure 3. Left: The variation of $\log_{10} f_i(\lambda)$ with i for (a) $\lambda = 10^{-8}$, (b) $\lambda = 10^{-5}$, (c) $\lambda = 10^{-2}$, (d) $\lambda = 10$. Right: The variation of $\log_{10} |f_i(\lambda)|$ with i for (a) $\lambda = -0.075$, (b) $\lambda = -0.050$, (c) $\lambda = -0.025$, (d) $\lambda = -0.010$, for Example 2.

Theorem 1 is extended in the sequel to the product of an arbitrary number of matrices, which will include the diagonal matrices $F(\lambda)$ and $I_s - F(\lambda)$,

$$\begin{aligned} F(\lambda) &= \operatorname{diag}\left\{f_i(\lambda)\right\}_{i=1}^{s} &&= \operatorname{diag}\left\{\frac{\sigma_i^2}{\sigma_i^2+\lambda}\right\}_{i=1}^{s}, \\ I_s - F(\lambda) &= \operatorname{diag}\left\{1 - f_i(\lambda)\right\}_{i=1}^{s} &&= \operatorname{diag}\left\{\frac{\lambda}{\sigma_i^2+\lambda}\right\}_{i=1}^{s}. \end{aligned} \tag{14}$$

The matrix $F(\lambda)$ is positive definite for $\lambda > -\sigma_s^2$ and it is negative definite for $\lambda < -\sigma_1^2$, and $I_s - F(\lambda)$ is positive definite for $\lambda > 0$ or $\lambda < -\sigma_1^2$, and it is negative definite for $-\sigma_s^2 < \lambda < 0$. The product of symmetric positive definite matrices is used in Section 7 to analyse the results of the computations in (Kobak et al. 2020, §2.3). Example 2 considers the properties of the filters $f_i(\lambda)$ for $\lambda \geq 0$ and $\lambda < 0$ for the Hilbert matrix of order seven and it is shown there are significant differences between these two situations.

Example 2. Figure 3 shows the variation of the filters $f_i(\lambda)$ with i for positive and negative values of λ for the Hilbert matrix of order seven. Figure 3 (left) shows that the filters are monotonically decreasing functions if $\lambda > 0$, but Figure 3 (right) shows that this monotonicity is not preserved for negative values of λ. The filters $f_i(\lambda)$ may be larger than one for $\lambda < 0$, in which case they amplify the components of $x(\lambda)$ and thus they do not perform a filtering operation. The filters assume positive and negative values if $\lambda < 0$, which marks a difference between $\lambda < 0$ and $\lambda \geq 0$ because the filters are strictly positive for $\lambda \geq 0$. Figure 4 shows the variation of $f_i(\lambda)$ with positive and negative values of λ, and it assumes very large negative values and very large positive values when $\lambda \approx -\sigma_i^2$, but $0 < f_i(\lambda) \leq 1$ if $\lambda \geq 0$. □

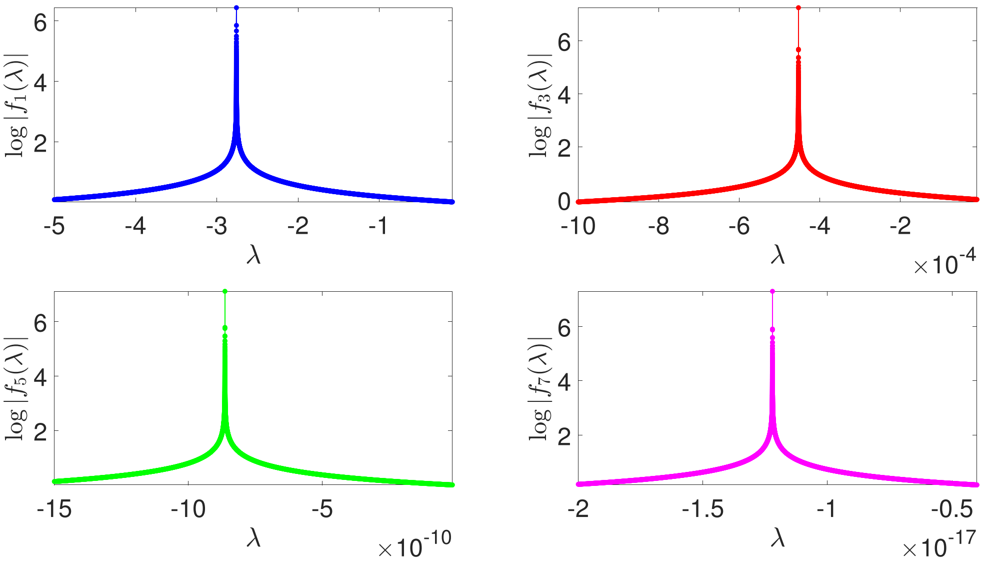

Figure 4. The variation of $\log_{10}|f_i(\lambda)|$ with $\lambda < 0$ for $i = 1, 3, 5, 7$, for Example 2.

4. Condition estimation

Regularisation imposes stability on the solution of the LS problem, and this section considers the numerical condition of this problem. It is shown that its properties for $\lambda < 0$ are significantly different from its properties for $\lambda \geq 0$.

It follows from (6) that

$$\delta x(\lambda) = \begin{cases} \left(A^T A + \lambda I_p\right)^{-1} A^T \delta b, & n \geq p, \\ A^T\left(AA^T + \lambda I_n\right)^{-1} \delta b, & n < p, \end{cases} \tag{15}$$

and thus the effective condition number $\eta(A, b, \lambda)$, which is the maximum value of the ratio of the relative error in $x(\lambda)$ to the relative error in b with respect to a perturbation δb in b, is (Winkler and Mitrouli 2020, §5.2)

$$\eta(A, b, \lambda) = \max_{\delta b \in \mathbb{R}^n} \frac{\Delta x(\lambda)}{\Delta b} = \begin{cases} \frac{\left\|(A^T A+\lambda I_p)^{-1} A^T\right\| \|b\|}{\left\|(A^T A+\lambda I_p)^{-1} A^T b\right\|} = \frac{\left\|(\Sigma^T \Sigma+\lambda I_p)^{-1} \Sigma^T\right\| \|c\|}{\left\|(\Sigma^T \Sigma+\lambda I_p)^{-1} \Sigma^T c\right\|}, & n \geq p, \\ \frac{\left\|A^T (AA^T+\lambda I_n)^{-1}\right\| \|b\|}{\left\|A^T (AA^T+\lambda I_n)^{-1} b\right\|} = \frac{\left\|\Sigma^T (\Sigma\Sigma^T+\lambda I_n)^{-1}\right\| \|c\|}{\left\|\Sigma^T (\Sigma\Sigma^T+\lambda I_n)^{-1} c\right\|}, & n < p, \end{cases} \tag{16}$$

where $\Delta x(\lambda)$ and Δb are the relative errors in $x(\lambda)$ and b,

$$\Delta x(\lambda) = \frac{\|\delta x(\lambda)\|}{\|x(\lambda)\|} \qquad \text{and} \qquad \Delta b = \frac{\|\delta b\|}{\|b\|}, \tag{17}$$

the terms on the right follow from the SVD of A and $c = U^T b$. The expressions (16) for $\eta(A, b, \lambda)$ can be combined,

$$\eta(A, b, \lambda) = \frac{\max_{i=1,\dots,s}\left\{\frac{\sigma_i}{|\sigma_i^2+\lambda|}\right\}\|c\|}{\left\|\operatorname{diag}\left\{\frac{\sigma_i}{\sigma_i^2+\lambda}\right\}_{i=1}^{s}\{c_i\}_{i=1}^{s}\right\|} = \frac{\max_{i=1,\dots,s}\left\{\frac{\sigma_i}{|\sigma_i^2+\lambda|}\right\}\|c\|}{\left\|\operatorname{diag}\left\{f_i(\lambda)\right\}_{i=1}^{s}\left\{\frac{c_i}{\sigma_i}\right\}_{i=1}^{s}\right\|}, \tag{18}$$

where the filters $f_i(\lambda)$ are defined in (11) and the denominator shows that numerical problems may occur if $\lambda < 0$. These problems do not occur if $\lambda \geq 0$ and the distinction between these two regimes is therefore in accord with Example 2.

The condition of the LS problem ($\lambda = 0$) must be considered because it determines if regularisation is required. This issue is addressed in Theorem 2.

Theorem 2. *The effective condition number $\eta(A, b, 0)$ of the LS problem satisfies*

$$\eta(A,b,0) \leq \frac{\kappa(A)}{\cos\theta}, \qquad \cos\theta = \frac{\sqrt{\sum_{i=1}^{s} c_i^2}}{\|c\|}, \qquad s = \min(n,p), \tag{19}$$

where θ is the angle between b and its component that lies in the column space of A.

Proof. It follows from (16) that

$$\eta(A,b,0) = \frac{1}{\sigma_s}\frac{\|b\|}{\left\|A^{\dagger}b\right\|} = \frac{\|c\|}{\sigma_s\sqrt{\sum_{i=1}^{s}\left(\frac{c_i}{\sigma_i}\right)^2}} \geq 1, \tag{20}$$

which is a refined measure of the stability of $x(0)$ because it is a function of A and b. The relationship between $\eta(A, b, 0)$ and the condition number $\kappa(A)$ of A requires that the conditions $n \leq p$ and $n > p$ be considered separately because $b \in \mathcal{C}(A)$ if $n \leq p$, but $b \notin \mathcal{C}(A)$ if $n > p$, where $\mathcal{C}(A)$ is the column space of A.

Consider the situation $n \leq p$, and thus $s = n$ and $\cos\theta = 1$, and

$$\max_{b\in\mathbb{R}^n} \eta(A,b,0) = \max_{c\in\mathbb{R}^n} \eta(A,b,0) = \max_{\delta b, b\in\mathbb{R}^n} \left.\frac{\Delta x(\lambda)}{\Delta b}\right|_{\lambda=0} = \kappa(A), \qquad n \leq p, \tag{21}$$

where $\kappa(A) = \sigma_1/\sigma_n$. Equality of $\eta(A, b, 0)$ and $\kappa(A)$ occurs when $c = e_1$ where e_i is the ith unit basis vector, and thus $b = Uc = Ue_1$, that is, b is equal to the first column of U. More generally, it follows from (20) that

$$\max_{b\in\mathbb{R}^n} \eta(A,b,0) \approx \kappa(A), \qquad n \leq p, \tag{22}$$

when the discrete Picard condition (7) is satisfied.

Consider now the situation $n > p$, and thus $s = p$, for which $b \notin \mathcal{C}(A)$. It follows from (20) that

$$\eta(A,b,0) = \frac{1}{\sigma_p}\left(\frac{\|c\|}{\sqrt{\sum_{i=1}^{p}\left(\frac{c_i}{\sigma_i}\right)^2}}\right) \leq \kappa(A)\left(\frac{\|c\|}{\sqrt{\sum_{i=1}^{p} c_i^2}}\right), \qquad \kappa(A) = \frac{\sigma_1}{\sigma_p}, \tag{23}$$

and thus the result (19) is established.

If the SVD of A is written as

$$A = U\Sigma V^T = \begin{bmatrix} U_1 & U_2 \end{bmatrix}\begin{bmatrix} \Sigma_1 \\ 0 \end{bmatrix} V^T, \qquad U_1^T U_1 = I_p, \tag{24}$$

where $U_1 \in \mathbb{R}^{n\times p}$, $U_2 \in \mathbb{R}^{n\times(n-p)}$ and $\Sigma_1 \in \mathbb{R}^{p\times p}$, and if c is partitioned as

$$c = \{c_i\}_{i=1}^{n} = \begin{bmatrix} \tilde{c}_1 \\ \tilde{c}_2 \end{bmatrix} = \begin{bmatrix} U_1^T \\ U_2^T \end{bmatrix} b, \qquad \tilde{c}_1 \in \mathbb{R}^p, \qquad \tilde{c}_2 \in \mathbb{R}^{n-p}, \tag{25}$$

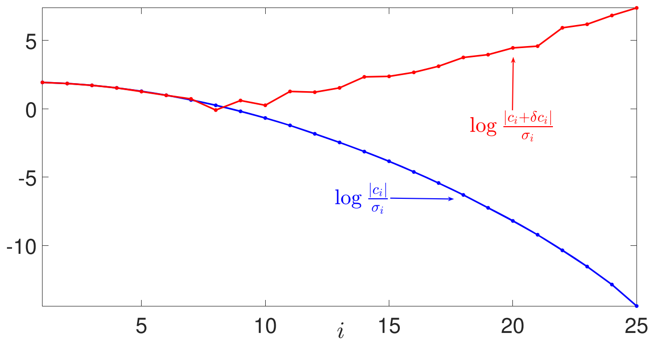

Figure 5. The ratio $\log_{10}|c_i|/\sigma_i$ for the exact data, and the ratio $\log_{10}|c_i+\delta c_i|/\sigma_i$ for the perturbed data, for Example 3.

then it follows from (19) that

$$\cos^2\theta = \frac{\|\tilde{c}_1\|^2}{\|c\|^2} = \frac{\left\|U_1^T b\right\|^2}{\|b\|^2} = \frac{\left\|b^T U_1\left(U_1^T U_1\right) U_1^T b\right\|}{\|b\|^2} = \frac{\left\|U_1 U_1^T b\right\|^2}{\|b\|^2}, \tag{26}$$

where $U_1 U_1^T b$ is the component of b that lies in the column space of A. □

The dependence of $\eta(A,b,0)$ on b suggests that $\eta(A,b,0)$, rather than $\kappa(A)$, which is a function of A only, should be used to compute the stability of $x(0)$. This is incorrect because it is shown in (Winkler 2024, Theorem 4) that to first order in δb,

$$\frac{\Delta\eta(A,b,0)}{\Delta b} \leq 1+\eta(A,b,0), \qquad \Delta\eta(A,b,0) = \frac{|\delta\eta(A,b,0)|}{\eta(A,b,0)}, \tag{27}$$

and thus $\eta(A,b,0)$ is unstable, and therefore $x(0)$ is also unstable, with respect to a perturbation in b if $\eta(A,b,0) \gg 1$. This is a disadvantage of this condition number, but it is instructive to consider it because the discrete Picard condition (7) follows from the denominator of the expression for $\eta(A,b,0)$ in (20).

Example 3. Consider the approximation of a function $f(x)$ that is defined at $m=76$ points by $n=25$ exponential basis functions,

$$f(x_i) = \sum_{j=1}^{n} a_j \exp\left(\frac{-(x_i-\mu_j)^2}{2\sigma^2}\right), \qquad i=1,\ldots,m, \tag{28}$$

where $\sigma=1.35$ and the 76 points are uniformly distributed in the interval $I=[0,\ldots,15]$. The centres μ_j of the basis functions are uniformly distributed in I, $A\in\mathbb{R}^{76\times 25}$ and $b\in\mathbb{R}^{76}$, where the entries of b are the function values $f(x_i)$. The condition number and effective condition number are $\kappa(A)=1.53\times 10^8$ and $\eta(A,b,0)=1.41\times 10^8$, respectively, and thus the coefficients a_j are unstable with respect to a perturbation in the function values $f(x_i)$.

Noise δb was added to b such that $\|b\|/\|\delta b\|=42.6$, and Figure 5 shows the ratio (7) for the exact and perturbed data. The effect of noise is significant, which follows from the large value of $\eta(A,b,0)$, and thus the LS problem must be regularised. Figure 6 shows the variation of $\kappa(A)$, $\eta(A,b,\lambda)$ and $\eta(A,b+\delta b,\lambda)$ with $\lambda\geq 0$. It is seen that $\eta(A,b,\lambda)$ is a decreasing function of λ for $\lambda\leq 1$, and it is approximately constant at its minimum value of one for

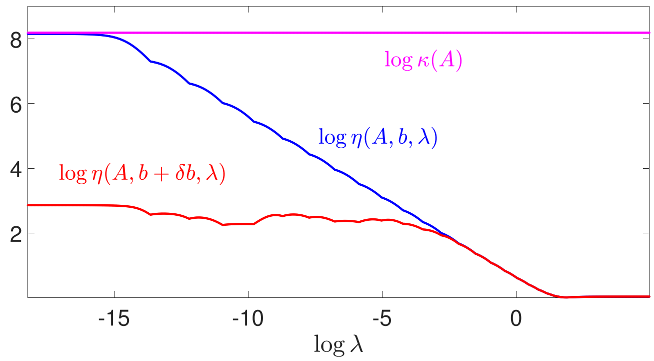

Figure 6. The variation of the condition number $\log_{10}\kappa(A)$, and effective condition numbers $\log_{10}\eta(A,b,\lambda)$ and $\log_{10}\eta(A,b+\delta b,\lambda)$, with λ, for Example 3.

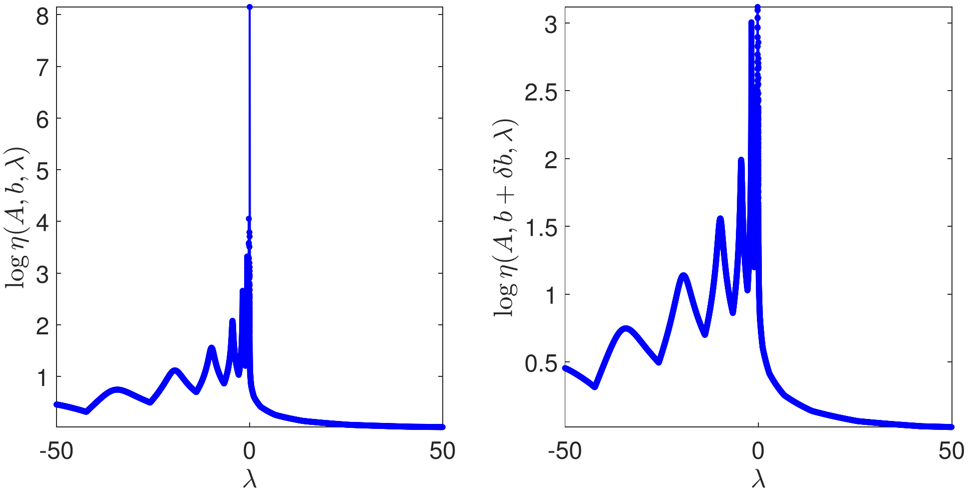

Figure 7. The variation of, left, $\log_{10}\eta(A,b,\lambda)$, and right, $\log_{10}\eta(A,b+\delta b,\lambda)$, with λ, for Example 3.

$\lambda > 1$, which shows that regularisation imposes stability on $x(\lambda)$. The unstable property of $\eta(A,b,\lambda)$ for $\lambda \approx 0$ is evident because $\eta(A,b,0)$ and $\eta(A,b+\delta b,0)$ differ by five orders of magnitude, which confirms (27).

Figure 7 shows the variation of $\eta(A,b,\lambda)$ and $\eta(A,b+\delta b,\lambda)$ with positive and negative values of λ. The figure shows that $\eta(A,b,\lambda)$ and $\eta(A,b+\delta b,\lambda)$ have two important differences for $\lambda < 0$ and $\lambda \geq 0$:

- $\eta(A,b,\lambda)$ and $\eta(A,b+\delta b,\lambda)$ decay rapidly as $\lambda \geq 0$ increases, and their values for $\lambda > 0$ are smaller than their values for $\lambda < 0$.
- An increase in the value of $\lambda < 0$ to $\lambda = 0$ may cause a significant increase in the values of $\eta(A,b,\lambda)$ and $\eta(A,b+\delta b,\lambda)$, which is unacceptable because an increase in the value of λ should yield a decrease in their values. □

Example 3 shows that negative values of λ yield large values of $\eta(A, b, \lambda)$, and the values of $\eta(A, b, \lambda)$ for $\lambda < 0$ are larger than their values for $\lambda \geq 0$. It follows that negative values of λ cannot regularise the solution of the LS problem.

It was shown in Section 3 that regularisation requires a trade-off between the stability and regularisation error of $x(\lambda)$. The stability of $x(\lambda)$ is considered in this section, and the regularisation error is considered in the next section.

5. Errors in the regularised solution

The vector $x(\lambda)$ in (2) that minimises $f(x, \lambda)$, $\lambda > 0$, in (1) is not equal to the solution $x(0)$ of the LS problem, and there is therefore an error between $x(\lambda)$ and $x(0)$ for $\lambda > 0$. This error is called the regularisation error (residual), and the optimal value λ_{opt} of λ yields a solution $x(\lambda_{\text{opt}})$ (i) whose error with respect to $x(0)$ is small, and (ii) that is much more stable than $x(0)$ with respect to a perturbation in b. The value $\lambda = \lambda_{\text{opt}}$ achieves an optimal trade-off between the regularisation error, which is small, and a large increase in the stability of $x(\lambda)$, and the L-curve, which is considered in Section 6, is a method for the calculation of the value of λ_{opt}. This section considers the variation of the regularisation error and relative error of $x(\lambda)$ with λ. These errors are, respectively,

$$e_{\text{res}}(\lambda) = \frac{\|Ax(\lambda) - b\|}{\|b\|} \quad \text{and} \quad e_{\text{rel}}(\lambda) = \frac{\|x(\lambda) - x(0)\|}{\|x(0)\|}, \tag{29}$$

and expressions for them that are derived from (10) are stated in Theorem 3.

Theorem 3. *The absolute residual is*

$$\begin{aligned} \|Ax(\lambda) - b\| &= \left\| \begin{bmatrix} \operatorname{diag}\left\{\frac{\lambda}{\sigma_i^2+\lambda}\right\}_{i=1}^{s} & \\ & I_{n-s} \end{bmatrix} \begin{bmatrix} \tilde{c}_1 \\ \tilde{c}_2 \end{bmatrix} \right\| \\ &= \left(\sum_{i=1}^{s} \left(\frac{\lambda}{\sigma_i^2 + \lambda} \right)^2 \tilde{c}_{1,i}^2 + \sum_{i=s+1}^{n} \tilde{c}_{2,i}^2 \right)^{\frac{1}{2}}, \end{aligned} \tag{30}$$

where the filters $f_i(\lambda)$ are defined in (11),

$$c = \{c_i\}_{i=1}^{n} = \begin{bmatrix} \tilde{c}_1 \\ \tilde{c}_2 \end{bmatrix} = \begin{bmatrix} U_1^T \\ U_2^T \end{bmatrix} b, \qquad \tilde{c}_1 \in \mathbb{R}^s, \qquad \tilde{c}_2 \in \mathbb{R}^{n-s}, \tag{31}$$

and the relative error is

$$e_{\text{rel}}(\lambda) = \left(\frac{\sum_{i=1}^{s} \left(\frac{\tilde{c}_{1,i}}{\sigma_i}\right)^2 \left(\frac{\lambda}{\sigma_i^2+\lambda}\right)^2}{\sum_{i=1}^{s} \left(\frac{\tilde{c}_{1,i}}{\sigma_i}\right)^2} \right)^{\frac{1}{2}} = \left(\frac{\sum_{i=1}^{s} \left(\frac{\tilde{c}_{1,i}}{\sigma_i}\right)^2 (1 - f_i(\lambda))^2}{\sum_{i=1}^{s} \left(\frac{\tilde{c}_{1,i}}{\sigma_i}\right)^2} \right)^{\frac{1}{2}}. \tag{32}$$

□

It follows from (29) and (30) that

$$e_{\text{res}}^2(0) = \frac{\|Ax(0) - b\|^2}{\|b\|^2} = \frac{\|\tilde{c}_2\|^2}{\|c\|^2} = 1 - \frac{\|\tilde{c}_1\|^2}{\|c\|^2} = 1 - \cos^2\theta, \tag{33}$$

where $\cos\theta$ is defined in (26).

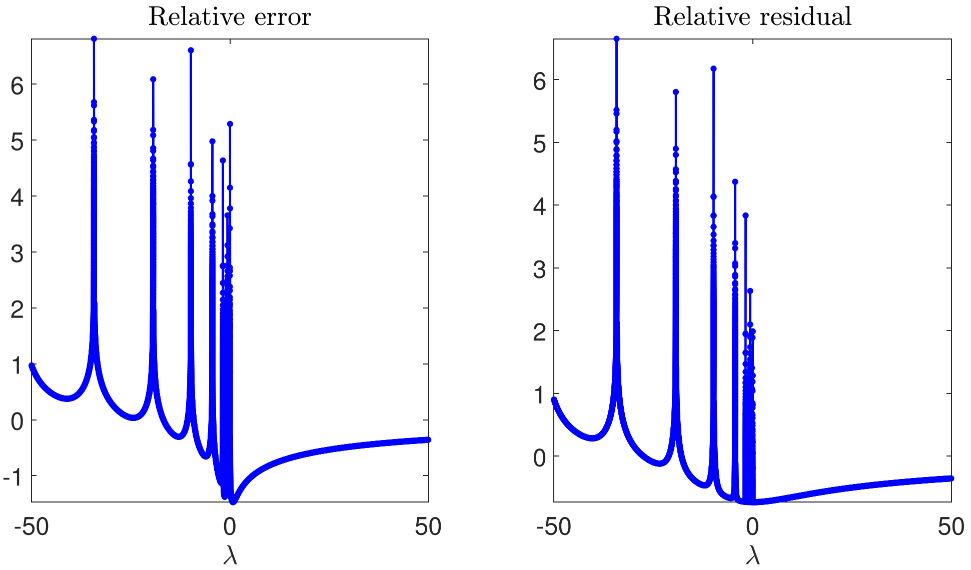

Figure 8. The variation of, left, $\log_{10} e_{\text{rel}}(\lambda)$, and right, $\log_{10} e_{\text{res}}(\lambda)$, with λ, for Example 4.

Example 4. The variation of $\log_{10} e_{\text{res}}(\lambda)$ and $\log_{10} e_{\text{rel}}(\lambda)$ with λ for the problem in Example 3 is shown in Figure 8. The curves are very similar and both of them have local maxima of large magnitude for $\lambda < 0$, they increase monotonically for $\lambda \geq 0$, and the errors for $\lambda < 0$ are much larger than the errors for $\lambda \geq 0$.

The differences in the residual between $\lambda < 0$ and $\lambda \geq 0$ follow from (14) and (30),

$$\|Ax(\lambda) - b\|^2 = c^T \begin{bmatrix} (I_s - F(\lambda))^2 & \\ & I_{n-s} \end{bmatrix} c = \tilde{c}_1^T (I_s - F(\lambda))^2 \tilde{c}_1 + \tilde{c}_2^T \tilde{c}_2, \tag{34}$$

where $\tilde{c}_1$ and $\tilde{c}_2$ are defined in (31), and thus

$$\frac{d\left(\|Ax(\lambda) - b\|^2\right)}{d\lambda} = -2\tilde{c}_1^T (I_s - F(\lambda)) \frac{dF(\lambda)}{d\lambda} \tilde{c}_1, \tag{35}$$

where, from (14),

$$\frac{dF(\lambda)}{d\lambda} = \frac{d}{d\lambda}\left(\text{diag } \{f_i(\lambda)\}_{i=1}^s\right) = -\text{diag}\left\{\frac{\sigma_i^2}{\left(\sigma_i^2 + \lambda\right)^2}\right\}_{i=1}^s, \tag{36}$$

and hence

$$\begin{aligned} \frac{dF(\lambda)}{d\lambda} &= -\text{diag}\left\{\left(\frac{1}{\lambda}\right)\left(\frac{\lambda}{\sigma_i^2 + \lambda}\right)\left(\frac{\sigma_i^2}{\sigma_i^2 + \lambda}\right)\right\}_{i=1}^s \\ &= -\left(\frac{1}{\lambda}\right)(I_s - F(\lambda))\, F(\lambda). \end{aligned} \tag{37}$$

It follows that

$$\begin{aligned} \frac{d\left(\|Ax(\lambda) - b\|^2\right)}{d\lambda} &= 2\left(\frac{d\left(\|Ax(\lambda) - b\|\right)}{d\lambda}\right)\|Ax(\lambda) - b\| \\ &= \left(\frac{2}{\lambda}\right)\tilde{c}_1^T (I_s - F(\lambda))^2 F(\lambda)\tilde{c}_1, \end{aligned} \tag{38}$$

and thus

$$\begin{aligned}\frac{d\left(\|Ax(\lambda)-b\|\right)}{d\lambda} &= \left(\frac{1}{\lambda}\right)\left(\frac{\tilde{c}_1^T\left(I_s-F(\lambda)\right)^2F(\lambda)\tilde{c}_1}{\|Ax(\lambda)-b\|}\right)\\ &= \frac{\sum_{i=1}^{s}\left(\frac{\lambda\sigma_i^2}{(\sigma_i^2+\lambda)^3}\right)\tilde{c}_{1,i}^2}{\left(\sum_{i=1}^{s}\left(\frac{\lambda}{\sigma_i^2+\lambda}\right)^2\tilde{c}_{1,i}^2+\sum_{i=s+1}^{n}\tilde{c}_{2,i}^2\right)^{\frac{1}{2}}}, \qquad (39)\end{aligned}$$

and since the entries of $F(\lambda)$ and $I_s - F(\lambda)$ are strictly positive for $\lambda > 0$, it follows that the residual $e_{\text{res}}(\lambda)$ increases monotonically to one as $\lambda \to \infty$. If, however, $\lambda < 0$, then $e_{\text{res}}(\lambda)$ is not a monotonic function because the signs of the entries of $F(\lambda)$ and $I_s - F(\lambda)$ are functions of λ, and therefore $e_{\text{res}}(\lambda)$ has local minima and maxima, and (39) shows that its derivative is unbounded as $\lambda \to -\sigma_i^2, i = 1, \ldots, s$.

The analysis of the relative error (32) is very similar, and thus the results in Figure 8 can be explained by the properties of $F(\lambda)$ and $I_s - F(\lambda)$. □

Example 4 shows that $\lambda < 0$ yields large values of the residual and relative error, and that they are much larger than their values for $\lambda \geq 0$. These results are consistent with the results in Section 4, which show that $\eta(A, b, \lambda < 0) > \eta(A, b, \lambda \geq 0)$.

Example 5 considers the relationship between the effective condition number and the regularisation error.

Example 5. Consider the Hilbert matrix H of order 12 and the four forms b_1, b_2, b_3 and b_4 of b shown in Figure 9,

$$b_1 = U(e_1 + e_2), \quad b_2 = U(e_4 + e_5), \quad b_3 = U(e_7 + e_8), \quad b_4 = U(e_{10} + e_{11}), \qquad (40)$$

where $Hx = b$, e_i is the ith unit basis vector and U is the left singular matrix from the SVD of H.

The progression $b_1 \to b_2 \to b_3 \to b_4$ is associated with a shift in the non-zero components of b from the first few columns of U, that is, the large singular values of H, to the last few columns of U, that is, the small singular values of H, and it causes a change in the effective condition number. In particular, Figure 10 shows the variation of the effective condition numbers $\eta(H, b_1, \lambda)$, $\eta(H, b_2, \lambda)$, $\eta(H, b_3, \lambda)$ and $\eta(H, b_4, \lambda)$ with λ, and the condition number $\kappa(H)$. Each effective condition number decreases to its minimum value, which is approximately equal to one, as λ increases, it then increases, and this increase is significant for $b = b_4$, but it is very small for $b = b_1$. It is seen that

$$\kappa(H) > \eta(H, b_1, 0) \gg \eta(H, b_2, 0) \gg \eta(H, b_3, 0) \gg \eta(H, b_4, 0), \qquad (41)$$

which confirms that the progression $b_1 \to b_2 \to b_3 \to b_4$ is associated with an increase in the stability of x with respect to a perturbation in b.

Figure 11 shows the variation of the regularisation error with λ for $b = b_1$, $b = b_2$, $b = b_3$ and $b = b_4$. The smallest and largest errors occur for $b = b_1$ and $b = b_4$, respectively, which is consistent with the graphs of the effective condition numbers in Figure 10. Figure 12 is a parametric plot of the regularisation error against the effective condition number as a function of λ, for $b = b_1$, $b = b_2$, $b = b_3$ and $b = b_4$. An increase in the value of λ from $\lambda = \lambda_{\min} = 10^{-33}$ yields a significant decrease in the value of $\eta(H, b_1, \lambda)$, but the value of the error is

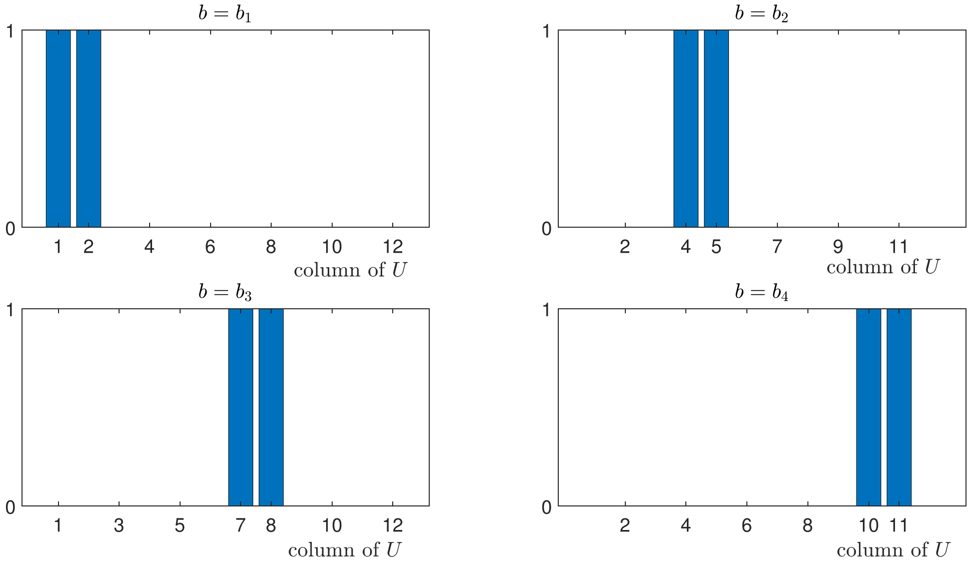

Figure 9. The vectors b_1, b_2, b_3 and b_4 for Example 5.

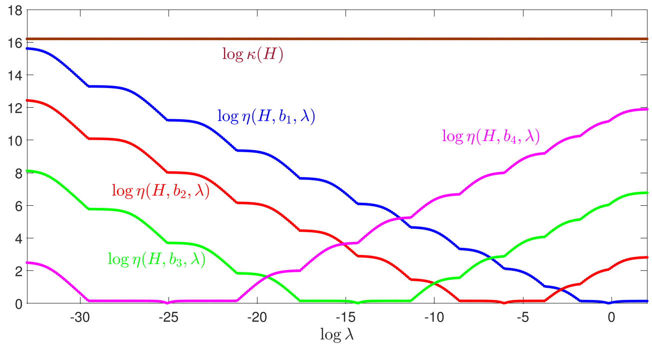

Figure 10. The variation of the effective condition numbers $\eta(H, b_1, \lambda)$, $\eta(H, b_2, \lambda)$, $\eta(H, b_3, \lambda)$ and $\eta(H, b_4, \lambda)$ with λ, and the condition number $\kappa(H)$, for Example 5.

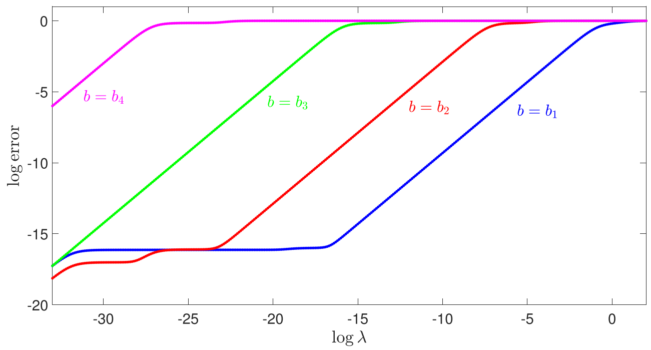

Figure 11. The variation of the regularisation error with λ, for $b = b_1$, $b = b_2$, $b = b_3$ and $b = b_4$, for Example 5.

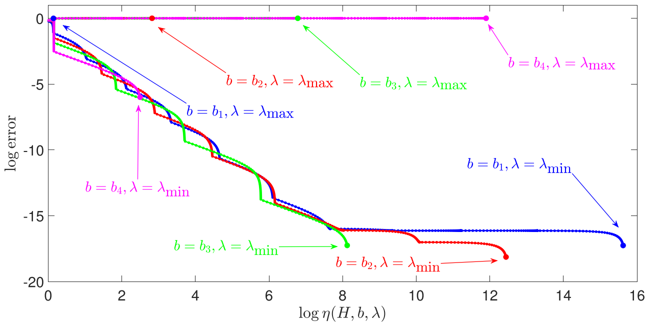

Figure 12. The variation of the regularisation error and effective condition number as functions of λ, for $b = b_1, b = b_2, b = b_3$ and $b = b_4$, for Example 5. The end points, $\lambda = \lambda_{\min} = 10^{-33}$ and $\lambda = \lambda_{\max} = 100$, are marked on each graph.

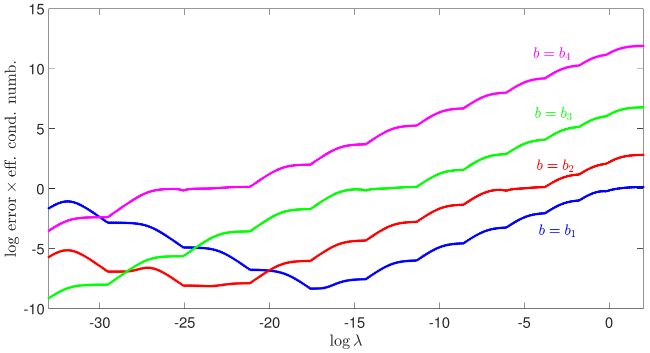

Figure 13. The variation of the product of the regularisation error and effective condition number as a function of λ, for $b = b_1, b = b_2, b = b_3$ and $b = b_4$, for Example 5.

approximately constant at 10^{-16}. The error increases when $\lambda > \lambda^*$, where $\eta(H, b_1, \lambda^*) \approx 10^8$, and similar properties are observed for $b = b_2$, but to a lesser extent. The graphs for $b = b_3$ and $b = b_4$ are significantly different because the error is not constant as λ increases from $\lambda = \lambda_{\min}$. There is, however, a range of values of λ for each vector $b = b_2$, $b = b_3$ and $b = b_4$ for which the error is approximately equal to one.

Figure 13 shows the variation of the product of the regularisation error and effective condition number with λ, for $b = b_1$, $b = b_2$, $b = b_3$ and $b = b_4$. The curve for $b = b_1$ has a well defined minimum, which confirms the trade-off between the regularisation error and effective condition number at the optimal regularisation parameter. This minimum is poorly defined for $b = b_2$, and the curves for $b = b_3$ and $b = b_4$ do not possess a minimum, and thus regularisation is effective in imposing stability on $x_1 = H^{-1}b_1$, but it is less effective in imposing stability on $x_2 = H^{-1}b_2$. Regularisation cannot, however, be applied to $x_3 = H^{-1}b_3$ and $x_4 = H^{-1}b_4$. □

6. The L-curve

It was shown in Sections 3, 4 and 5 that an acceptable solution $x(\lambda)$ requires $\lambda \geq 0$, and this section considers the L-curve, which is a method for calculating the value of λ_{opt} (Hansen

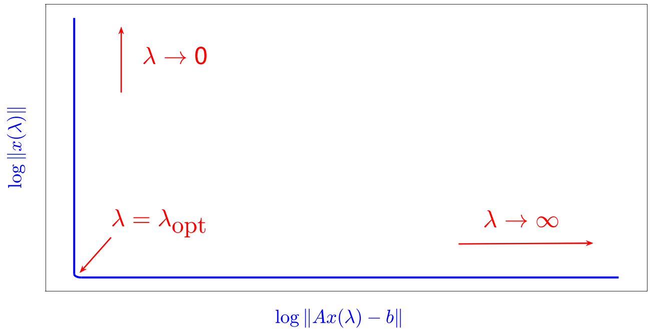

Figure 14. The L-curve.

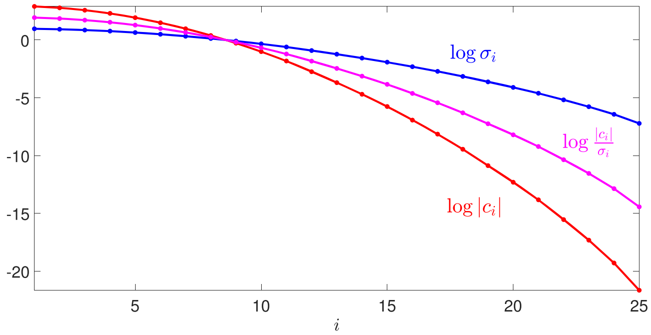

Figure 15. The variation of $\log_{10} |c_i|$, $\log_{10} \sigma_i$ and $\log_{10} |c_i|/\sigma_i$ with i, for Example 6.

1998, §4.6). The L-curve is a parametric plot of $\log_{10} \|Ax(\lambda) - b\|$ (horizontal axis) against $\log_{10} \|x(\lambda)\|$ (vertical axis), and if the discrete Picard condition (7) is satisfied, the curve has the form of an L, as shown in Figure 14. As λ increases from zero, $\|x(\lambda)\|$ decreases and $\|Ax(\lambda) - b\|$ is approximately constant, until $\lambda = \lambda_{\rm opt}$, which is the value of λ in the corner of the L. As λ increases from $\lambda_{\rm opt}$, $\|x(\lambda)\|$ is approximately constant and $\|Ax(\lambda) - b\|$ increases. The value $\lambda = \lambda_{\rm opt}$ is the optimal value of λ because $\|Ax(\lambda) - b\|$ and $\|x(\lambda)\|$ attain, approximately, their minimum values for this value of λ, and thus $\lambda_{\rm opt}$ balances the fidelity of the model and the satisfaction of the constraint on $\|x\|$.

Example 6. Consider the problem in Example 3, for which Figure 15 shows the variation of $\log_{10} |c_i|$, $\log_{10} \sigma_i$ and $\log_{10} |c_i|/\sigma_i$ with i. The discrete Picard condition (7) is satisfied because the constants $|c_i|$ decay to zero faster than the singular values σ_i decay to zero. Figure 5 shows that the noise δb has a significant effect because the exact solution is dominated by the large singular values of A, but the perturbed solution is dominated by its small singular values.

The L-curve for $\lambda \geq 0$ is shown in Figure 16 and the optimal value $\lambda_{\rm opt} = 0.47014$ of λ minimises, approximately, the error $\|Ax(\lambda) - b\|$ in the regression model and the magnitude $\|x(\lambda)\|$ of the solution. Figure 17 shows the L-curve for $-50 \leq \lambda < 0$ and the difference between this L-curve and the L-curve in Figure 16 is clear. Figure 17 shows there does not

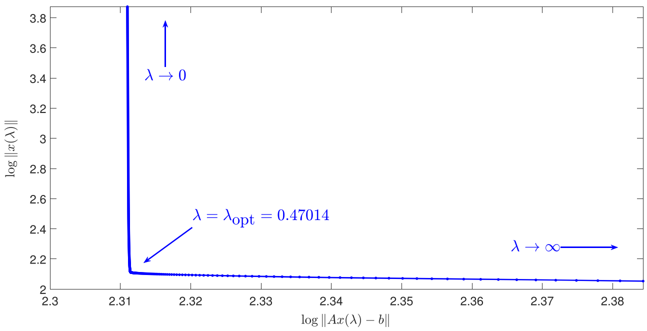

Figure 16. The L-curve for $\lambda \geq 0$, for Example 6.

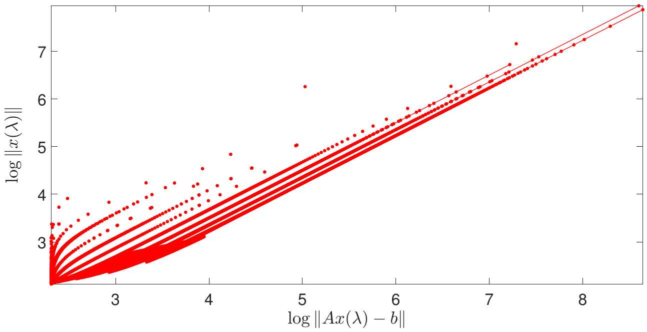

Figure 17. The L-curve for $-50 \leq \lambda < 0$, for Example 6.

exist an optimal value of λ that is negative, and in particular, it follows from (10) that

$$\|x(\lambda)\| = \left\| \operatorname{diag}\left\{\frac{\sigma_i^2}{\sigma_i^2 + \lambda}\right\}_{i=1}^{s} \left\{\frac{\tilde{c}_{1,i}}{\sigma_i}\right\}_{i=1}^{s} \right\| = \left\| \operatorname{diag}\left\{f_i(\lambda)\right\}_{i=1}^{s} \left\{\frac{\tilde{c}_{1,i}}{\sigma_i}\right\}_{i=1}^{s} \right\|, \tag{42}$$

where $\tilde{c}_1 = \left\{\tilde{c}_{1,i}\right\}_{i=1}^{s}$ is defined in (31), and hence

$$\|x(\lambda)\|^2 = \tilde{d}_1^T F^2(\lambda)\tilde{d}_1 = \sum_{i=1}^{s} \left(\frac{\sigma_i^2}{\sigma_i^2 + \lambda}\right)^2 \tilde{d}_{1,i}^2, \qquad \tilde{d}_1 = \left\{\frac{\tilde{c}_{1,i}}{\sigma_i}\right\}_{i=1}^{s}. \tag{43}$$

It follows that

$$\frac{d\left(\|x(\lambda)\|^2\right)}{d\lambda} = 2\frac{d\left(\|x(\lambda)\|\right)}{d\lambda}\|x(\lambda)\| = 2\tilde{d}_1^T F(\lambda)\frac{dF(\lambda)}{d\lambda}\tilde{d}_1, \tag{44}$$

and thus from (37),

$$\frac{d\left(\|x(\lambda)\|\right)}{d\lambda} = -\frac{\sum_{i=1}^{s}\left(\frac{\sigma_i^2}{\sigma_i^2+\lambda}\right)^2\left(\frac{1}{\sigma_i^2+\lambda}\right)\tilde{d}_{1,i}^2}{\left(\sum_{i=1}^{s}\left(\frac{\sigma_i^2}{\sigma_i^2+\lambda}\right)^2\tilde{d}_{1,i}^2\right)^{\frac{1}{2}}}, \tag{45}$$

and hence $\|x(\lambda)\|$ is a monotonically decreasing function of λ for $\lambda > 0$. It follows, however, from the properties of $I_s - F(\lambda)$ and $F(\lambda)$ that $\|x(\lambda)\|$ has local minima and maxima for $\lambda < 0$, and that its derivative is unbounded as $\lambda \to -\sigma_i^2$, $i = 1, \ldots, s$.

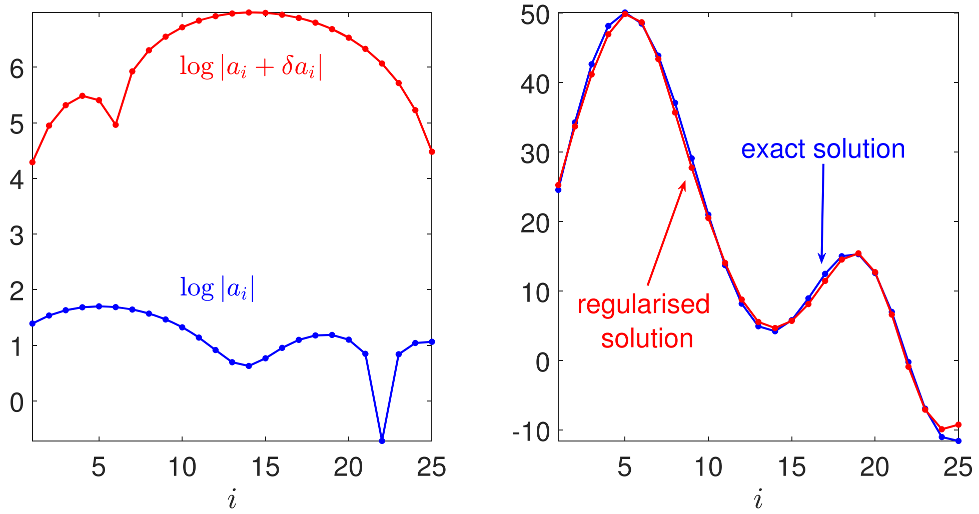

Figure 18. Left, the coefficients $\log_{10}|a_i|$ and $\log_{10}|a_i + \delta a_i|$ of the exact and noisy solutions, respectively, and right, the coefficients of the exact solution, and the coefficients of the regularised solution with $\lambda = \lambda_{\text{opt}} = 0.47014$, for Example 6.

The difference between the L-curves in Figures 16 and 17 follows from the properties of $F(\lambda)$ and $I_s - F(\lambda)$:

- If $\lambda > 0$, $F(\lambda)$ and $I_s - F(\lambda)$ are positive definite and thus $\|Ax(\lambda) - b\|$ increases monotonically, and $\|x(\lambda)\|$ decreases monotonically, as λ increases from zero.
- If $\lambda < 0$, $\|Ax(\lambda) - b\|$ and $\|x(\lambda)\|$ have local minima and maxima, and thus they are not monotonic functions as λ decreases from zero.

Figure 18 shows the coefficients a_i, the coefficients $a_i + \delta a_i$ when noise δb is added to b, and the coefficients after regularisation with $\lambda = \lambda_{\text{opt}} = 0.47014$. Figure 18 (left) shows that the regression problem is ill conditioned, and Figure 18 (right) shows that regularisation is effective in removing the effect of the perturbation δb on the coefficients a_i because the error between the exact and regularised solutions is very small. □

The examples in Sections 3, 4, 5 and 6 considered the situation in which $n \geq p$, and an example in which $n < p$ is considered in Section 7.

7. Simulation with a spiked covariance model

This section considers the simulation in (Kobak et al. 2020, §2) because its analysis led to the conclusion that the optimal regularisation parameter may be negative. Each vector of predictors $a_i \in \mathbb{R}^p$, $i = 1, \ldots, n$, $n = 64$, $p = 3116$, is drawn from a normal distribution with zero mean and covariance matrix $S = I_p + \rho T \in \mathbb{R}^{p\times p}$ that defines a spiked model, where $\rho = 0.1$, all the entries of $T \in \mathbb{R}^{p\times p}$ are one and the eigenvalues $\mu_i(S)$ of S are

$$\mu_1(S) = 1 + \rho p, \qquad \mu_i(S) = 1, \quad i = 2, \ldots, p. \tag{46}$$

The Cholesky decomposition LL^T of S, where $L \in \mathbb{R}^{p\times p}$ is lower triangular, allows the covariance matrix of $A := RL^T$, where the entries of $R \in \mathbb{R}^{n\times p}$ are independent random variables with zero mean and unit variance, to be equal to S,

$$\mathcal{E}\left\{A^T A\right\} = \mathcal{E}\left\{LR^T RL^T\right\} = L\mathcal{E}\left\{R^T R\right\} L^T = LL^T = S, \tag{47}$$

where $\mathcal{E}\{\cdot\}$ is the expectation operator. The vector $\tilde{b}$ is defined as

$$\tilde{b} := A\tilde{x}, \qquad \tilde{x} = \begin{bmatrix} \beta & \beta & \dots & \beta \end{bmatrix}^T \in \mathbb{R}^p, \qquad \beta = \sigma\left(\frac{\alpha}{p(1+p\rho)}\right)^{\frac{1}{2}}, \tag{48}$$

where α is a constant and $\sigma^2 = 1$. The addition of noise ϵ to $A\tilde{x}$ yields the perturbed equation,

$$b := A\tilde{x} + \epsilon, \qquad \epsilon = \{\epsilon_i\}_{i=1}^n \sim \mathcal{N}(0,\sigma^2), \tag{49}$$

and its regularised solution $x(\lambda)$ and the signal-to-noise ratio (SNR) are

$$x(\lambda) = A^\dagger(\lambda)b \qquad \text{and} \qquad \text{SNR} = \frac{\|A\tilde{x}\|}{\|\epsilon\|}, \tag{50}$$

where the form of $A^\dagger(\lambda)$ depends on the values of n and p, as shown in (6).

The conclusion in (Kobak et al. 2020) that λ_{opt} may be negative is derived from the risk $R(\lambda)$,

$$\begin{aligned} R(\lambda) &= \mathcal{E}\left\{(b - Ax(\lambda))^T (b - Ax(\lambda))\right\} \\ &= \mathcal{E}\left\{\left((A\tilde{x}+\epsilon) - Ax(\lambda)\right)^T \left((A\tilde{x}+\epsilon) - Ax(\lambda)\right)\right\} \\ &= \left(\tilde{x} - x(\lambda)\right)^T \mathcal{E}\left\{A^T A\right\}\left(\tilde{x} - x(\lambda)\right) + 2\left(\tilde{x} - x(\lambda)\right)^T A^T \mathcal{E}\{\epsilon\} + \mathcal{E}\left\{\epsilon^T\epsilon\right\} \\ &= \left(\tilde{x} - x(\lambda)\right)^T LL^T \left(\tilde{x} - x(\lambda)\right) + \|\epsilon\|^2 \\ &= \left(\tilde{x} - x(\lambda)\right)^T S\left(\tilde{x} - x(\lambda)\right) + \sigma^2, \end{aligned} \tag{51}$$

which follows from (47) and (49).

Examples 7 and 8 follow the procedure in (Kobak et al. 2020, §2.3) with SNR $= 10$. They demonstrate the theory in the previous sections and confirm that $\lambda < 0$ yields unacceptable results.

Example 7. Figure 19 shows the variation of the condition number $\log_{10}\kappa(A)$ and effective condition number $\log_{10}\eta(A,b,\lambda)$ for $p = 25, 50, 75, 100$, and $-50 \le \lambda \le 50$. The maximum value of $\kappa(A)$ is approximately $10^{1.5} = 31.6$ and thus the LS problem is well conditioned, and $\eta(A,b,\lambda) > \kappa(A)$ by several orders of magnitude if $\lambda < 0$, but $\eta(A,b,\lambda) < \kappa(A)$ if $\lambda \ge 0$. It follows that $\lambda < 0$ yields an unstable solution $x(\lambda)$, but a perturbation in $\tilde{b}$, which is defined in (48), has little effect on $x(\lambda)$ if $\lambda \ge 0$. Figures 20 and 21 show, respectively, the singular values $\log_{10}\sigma_i$, the coefficients $\log_{10}|c_i|$ and the ratios $\log_{10}|c_i|/\sigma_i$, and the singular values $\log_{10}\sigma_i$, the coefficients $\log_{10}|c_i + \delta c_i|$ and the ratios $\log_{10}|c_i + \delta c_i|/\sigma_i$, for $p = 25, 50, 75, 100$, where

$$c = U^T\tilde{b} \qquad \text{and} \qquad c + \delta c = U^T b, \tag{52}$$

and $\tilde{b}$ and b are defined in (48) and (49), respectively. The graphs in the figures are very similar and Figure 20 shows that, for each value of p, the ratio $|c_i|/\sigma_i$ does not decay to zero as i increases, and thus the discrete Picard condition (7) is not satisfied. It follows that $x(0)$ is

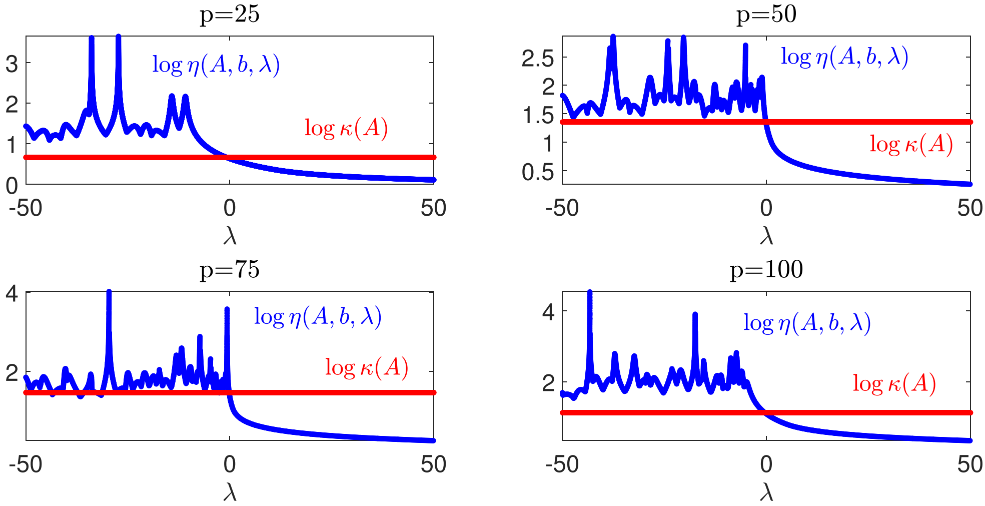

Figure 19. The variation of the effective condition number $\log_{10} \eta(A,b,\lambda)$ and condition number $\log_{10} \kappa(A)$ for $p = 25, 50, 75, 100$, for Example 7.

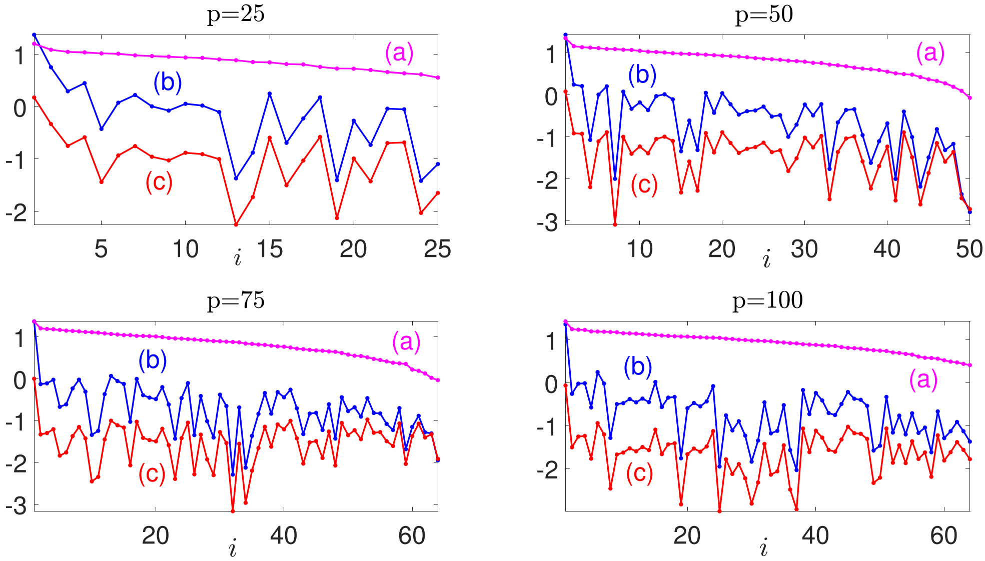

Figure 20. (a) The singular values $\log_{10} \sigma_i$, (b) coefficients $\log_{10} |c_i|$ and ratios (c) $\log_{10} |c_i|/\sigma_i$ for $p = 25, 50, 75, 100$, for Example 7.

stable, which is in accord with the low value of the effective condition number $\eta(A, b, 0)$ for each value of p, as shown in Figure 19. These graphs show that regularisation must not be applied because it requires the deletion of the small singular values of A, which would yield a large error in the regularised solution $x(\lambda)$.

Figure 22 shows the L-curves for $p = 25, 50, 75, 100$, and $\lambda \geq 0$, and they are significantly different from the L-curve in Figure 16. In particular, the L-curves in Figure 22 do not have a sharp corner that defines λ_{opt}, and an increase in the value of λ from $\lambda = 0$ causes little or no decrease in $\|x(\lambda)\|$, but a significant increase in $\|Ax(\lambda) - b\|$. It follows that $\lambda_{\text{opt}} =$

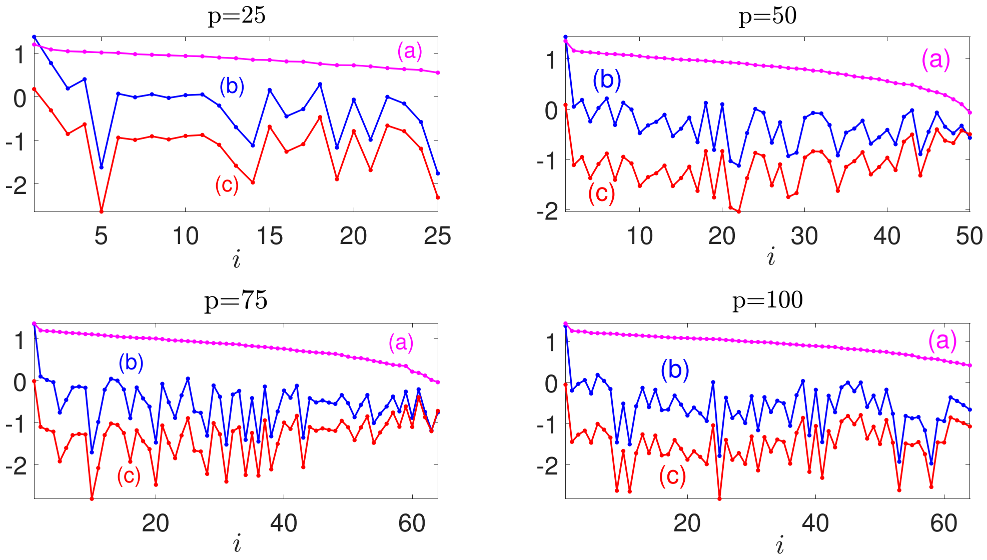

Figure 21. (a) The singular values $\log_{10}\sigma_i$, (b) coefficients $\log_{10}|c_i+\delta c_i|$ and ratios (c) $\log_{10}|c_i+\delta c_i|/\sigma_i$ for $p=25,50,75,100$, for Example 7.

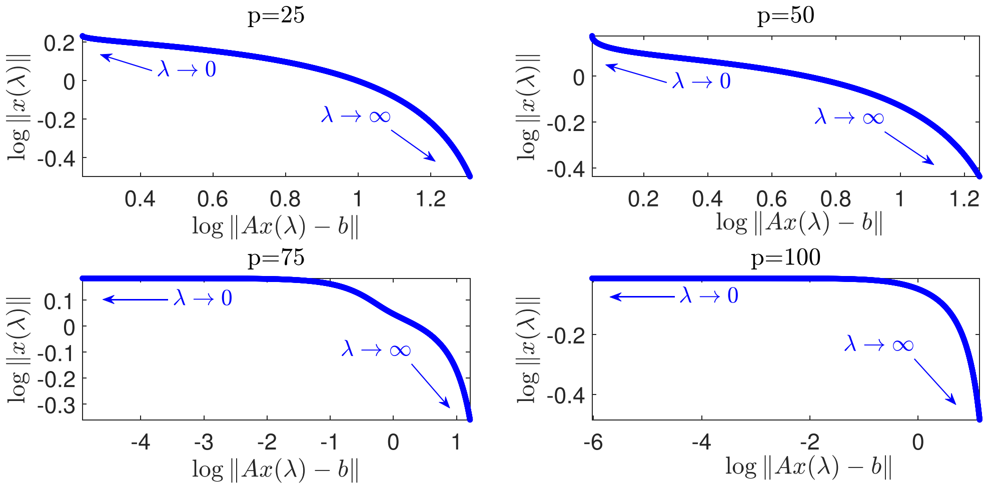

Figure 22. The L-curves for $p=25,50,75,100$, and $\lambda\geq 0$, for Example 7.

0, which agrees with the results in Figures 19 and 20, and thus regularisation must not be applied. The difference between these L-curves and the L-curve in Figure 16 arises because the solution of the regression problem in Example 6 is unstable, but the solution of the LS problem in (Kobak et al. 2020, §2.3) is stable. A parametric curve of $\log_{10}\|Ax(\lambda)-b\|$ against $\log_{10}\|x(\lambda)\|$ assumes the form of an L if the discrete Picard condition (7) is satisfied, but Figure 20 shows that this condition is not satisfied by the data in (Kobak et al. 2020, §2.3). Figure 23 shows the L-curves for $p=25,50,75,100$, and $-50\leq\lambda<0$, and it is clear that there does not exist a value of $\lambda<0$ that minimises, approximately, $\|Ax(\lambda)-b\|$ and $\|x(\lambda)\|$, and thus an optimal value of λ that is negative does not exist. This result is in accord with the L-curve in Figure 17 for Example 6.

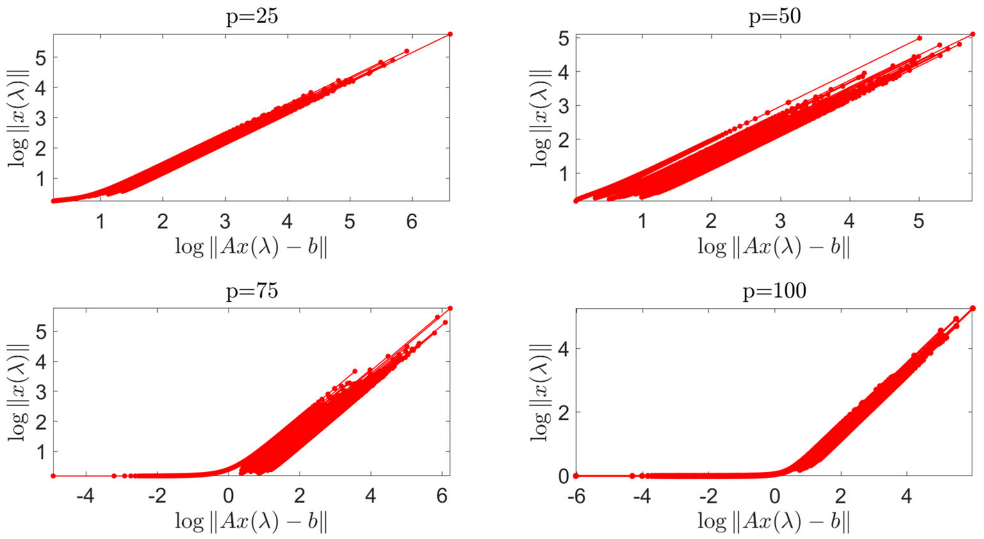

Figure 23. The L-curves for $p = 25, 50, 75, 100$, and $-50 \leq \lambda < 0$, for Example 7.

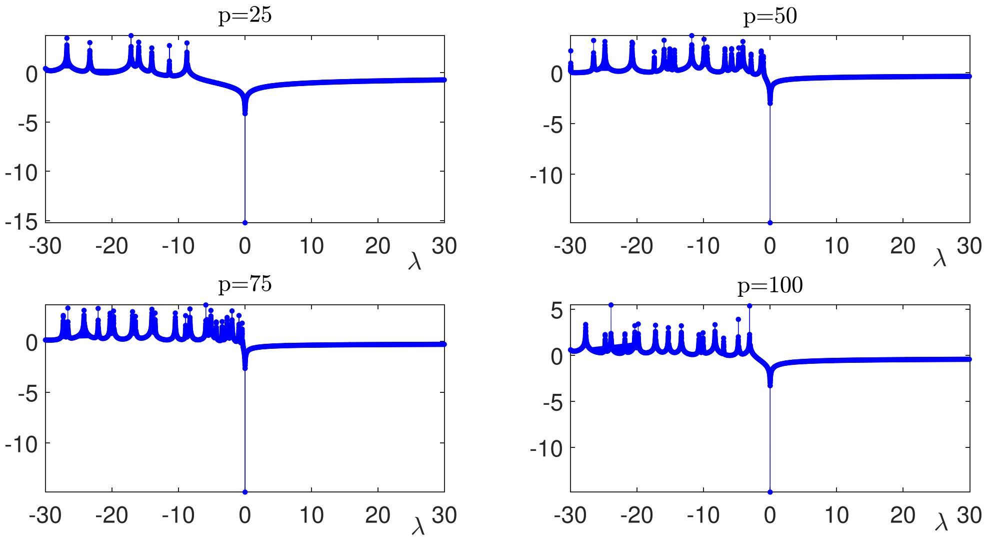

Figure 24. The variation of the relative error $\log_{10} e_{\text{rel}}(\lambda)$ with λ for $p = 25, 50, 75, 100$, and $-30 \leq \lambda \leq 30$, for Example 7.

Figure 24 shows the variation of the relative error $\log_{10} e_{\text{rel}}(\lambda)$, where $e_{\text{rel}}(\lambda)$ is defined in (32), with λ, for $-30 \leq \lambda \leq 30$. It is seen that $e_{\text{rel}}(\lambda)$ for $\lambda \geq 0$ is smaller, by up to about three orders of magnitude, than $e_{\text{rel}}(\lambda)$ for $\lambda < 0$. The differences in the properties of $e_{\text{rel}}(\lambda)$ between $\lambda < 0$ and $\lambda \geq 0$ are considered in Example 4, and they also explain the graphs in Figure 24. This analysis can be extended to determine the dependence of the risk $R(\lambda)$, which is defined in (51), with λ. This dependence is shown in Figure 25 for $-100 \leq \lambda \leq 50$ and it is seen that the properties of $R(\lambda)$ for $\lambda < 0$ are significantly different from its properties for $\lambda \geq 0$. In particular, $R(\lambda)$ has many local minima and maxima for $\lambda < 0$, but it is a smooth function of λ for $\lambda \geq 0$. This difference, which is also present in Figure 24, is due to the difference in the

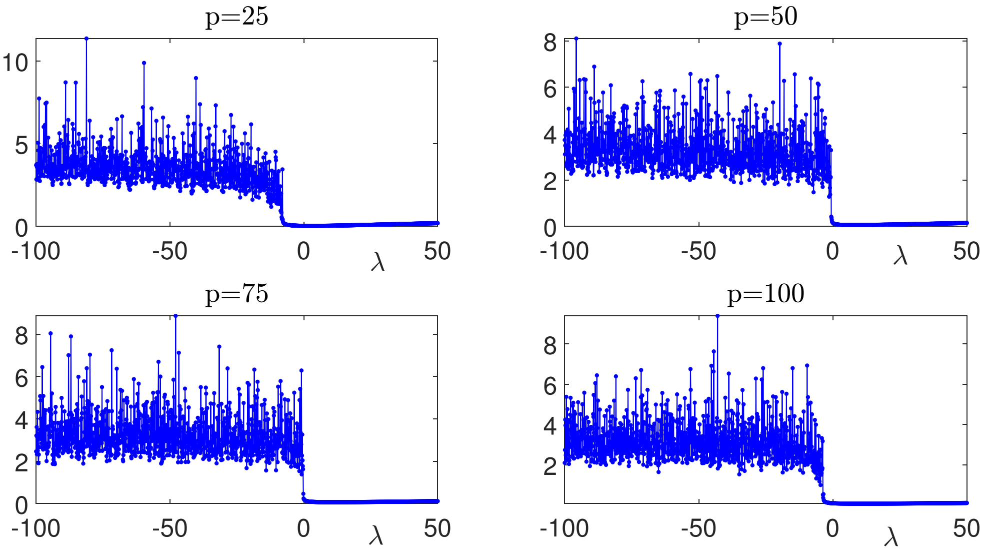

Figure 25. The variation of $\log_{10} R(\lambda)$ with λ for $p = 25, 50, 75, 100$, and $-100 \le \lambda \le 50$, for Example 7.

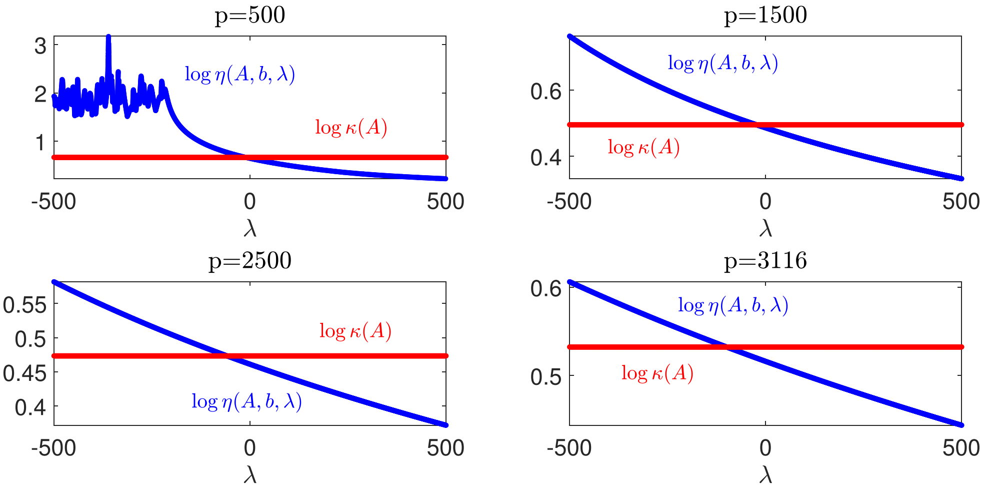

Figure 26. The variation of the condition number $\log_{10} \kappa(A)$ and effective condition number $\log_{10} \eta(A, b, \lambda)$ with λ, for $p = 500, 1500, 2500, 3116$, and $-500 \le \lambda \le 500$, for Example 8.

properties of the filters $f_i(\lambda)$ for $\lambda < 0$ and $\lambda \ge 0$, and Figures 24 and 25 show that negative values of λ yield large errors and they are therefore unacceptable. This result is in accord with the values of $\eta(A, b, \lambda)$ for $\lambda < 0$ in Figure 19. □

Example 8. It is stated in (Kobak et al. 2020, §2.3) that λ_{opt} may be negative if $p \gg n$, and this example considers this scenario. In particular, four values of the predictors that satisfy $p \gg n$, $p = 500, 1500, 2500, 3116$, and $n = 64$, were considered, and $-500 \le \lambda \le 500$. Figure 26 shows the variation of the condition number $\log_{10} \kappa(A)$ and effective condition number $\log_{10} \eta(A, b, \lambda)$ with λ. It is seen that (i) $\kappa(A) < 10$ for all λ, (ii) $\eta(A, b, \lambda) > \kappa(A)$ for $\lambda < 0$,

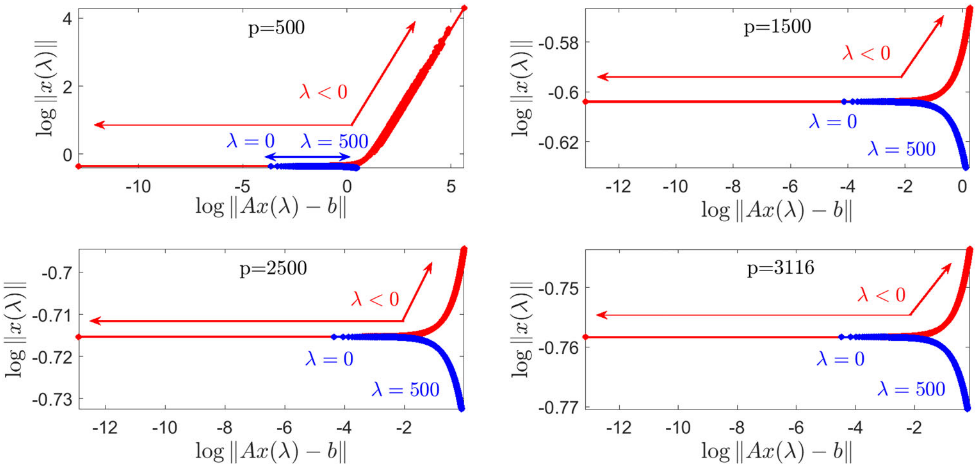

Figure 27. The L-curves for $p = 500, 1500, 2500, 3116$, and $-500 \leq \lambda \leq 500$, for Example 8.

and (iii) $\eta(A, b, \lambda) < \kappa(A)$ for $\lambda \geq 0$, from which it follows that regularisation must not be applied to $x(0)$ because it is stable. The graphs in Figures 20 and 21 for $p = 25, 50, 75, 100$, are very similar to their equivalents for $p = 500, 1500, 2500, 3116$, and they confirm that the discrete Picard condition (7) is not satisfied. Figure 27 shows the L-curves for these values of the predictors and for $-500 \leq \lambda \leq 500$, and the differences in each curve for $\lambda < 0$ and $\lambda \geq 0$ follow from the properties of the filters (11). In particular, they decrease monotonically as λ increases from $\lambda = 0$, and an increase in the value of λ is associated with an increase in the residual $\|Ax(\lambda) - b\|$ and a decrease in $\|x(\lambda)\|$, and hence the progression along the curve as λ increases from $\lambda = 0$ is unidirectional. The situation for $\lambda < 0$ is different because $\|Ax(\lambda) - b\|$ and $\|x(\lambda)\|$ may increase or decrease as λ decreases from zero, which is evident in the L-curves in Figure 23. The progression along the L-curve as $\lambda < 0$ decreases from zero is not unidirectional, and thus, as shown in Example 6, a negative value of λ cannot be considered. □

Examples 7 and 8 consider the problem in (Kobak et al. 2020, §2.3) and they show that $\lambda_{\text{opt}} = 0$, which must be compared with the result $\lambda_{\text{opt}} < 0$ in (Kobak et al. 2020, §2.5). The difference arises because the value $\lambda_{\text{opt}} < 0$ follows from the minimisation of the risk $R(\lambda)$, but the value $\lambda_{\text{opt}} = 0$ arises from the simultaneous minimisation of $\|Ax(\lambda) - b\|^2$ and $\|x(\lambda)\|^2$. In particular, $R(\lambda)$ is equal to the expected value of $\|Ax(\lambda) - b\|^2$ and thus it does not consider $\|x(\lambda)\|$, and it is minimised by $\lambda_{\text{opt}} < 0$ if $p \gg n$ and $0 < \rho \ll 1$, where ρ is defined in (46). It follows that the failure of $R(\lambda)$ to consider $\|x(\lambda)\|$ explains the difference in the values of λ_{opt} obtained by the L-curve and the minimisation of $R(\lambda)$.

8. Summary

The study of a problem in statistical learning in (Kobak et al. 2020) led to the claim that the regularisation parameter λ can be negative, and it is analysed in more detail in several problems in machine learning in (LeJeune et al. 2024, §6.2; Patil, Du, and Tibshirani 2024, pp. 24–26; Tsigler and Bartlett 2023, §8; Wu and Xu 2020, §5). This article has considered the numerical implications of the condition $\lambda < 0$ and it has been shown theoretically, from the properties of convex functions, refined condition estimation and error analysis of the regularised LS problem, and by example that it leads to unsatisfactory solutions for underdetermined and overdetermined LS problems. In particular, the solution $x(\lambda)$ for $\lambda < 0$ is unstable, even if $x(\lambda)$ is stable for $\lambda \geq 0$. Also, the residual and relative error in $x(\lambda)$ for $\lambda < 0$ are significantly larger than their values for $\lambda \geq 0$, and the L-curve shows that the optimal value of λ cannot be negative. Furthermore, the condition $\lambda \geq 0$ guarantees that the objective function in Tikhonov regularisation has a unique minimum because the matrices $A^T A + \lambda I_p$ and $AA^T + \lambda I_n$ are positive definite, but these matrices are positive definite, or indefinite, or negative definite if $\lambda < 0$. It is concluded that a negative value of λ in Tikhonov regularisation cannot be considered.

Disclosure statement

No potential conflict of interest was reported by the authors.

References

Berthet-Rayne, P., K. Leibrandt, G. Gras, P. Fraisse, A. Crosnier, and G.-Z. Yang. 2018. Inverse kinematics control methods for redundant snakelike robot teleoperation during minimally invasive surgery. *IEEE Robotics and Automation Letters* 3 (3):2501–8. doi: 10.1109/LRA.2018.2812907.

Buccini, A., O. De La Cruz Cabrera, C. Koukouvinos, M. Mitrouli, and L. Reichel. 2023. Variable selection in saturated and supersaturated designs via lp-lq minimization. *Communications in Statistics-Simulation and Computation* 52 (9):4326–47. doi: 10.1080/03610918.2021.1961151.

Cannon, A. J. 2009. Negative ridge regression parameters for improving the covariance structure of multivariate linear downscaling models. *International Journal of Climatology* 29 (5):761–9. doi: 10.1002/joc.1737.

Djennadi, S., N. Shawagfeh, and O. Abu Arqub. 2021a. A fractional Tikhonov regularization method for an inverse backward and source problems in the time-space fractional diffusion equations. *Chaos, Solitons & Fractals* 150:111127. doi: 10.1016/j.chaos.2021.111127.

Djennadi, S., N. Shawagfeh, and O. Arqub. 2021b. A numerical algorithm in reproducing kernel-based approach for solving the inverse source problem of the time-space fractional diffusion equation. *Partial Differential Equations in Applied Mathematics* 4:100164.

Djennadi, S., N. Shawagfeh, M. Inc, M. S. Osman, J. F. Gómez-Aguilar, and O. Abu Arqub. 2021. The Tikhonov regularization method for the inverse source problem of time fractional heat equation in the view of ABC-fractional technique. *Physica Scripta* 96 (9):094006. doi: 10.1088/1402-4896/ac0867.

Hansen, P. C. 1998. *Rank-deficient and discrete Ill-posed problems: Numerical aspects of linear inversion*. Philadelphia, USA: SIAM.

Hansen, P. C., J. G. Nagy, and D. P. O'Leary. 2006. *Deblurring images: Matrices, spectra, and filtering*. Philadelphia, USA: SIAM.

Hua, T. A., and R. F. Gunst. 1983. Generalized ridge regression: A note on negative ridge parameters. *Communications in Statistics- Theory and Methods* 12 (1):37–45. doi: 10.1080/03610928308828440.

Kobak, D., J. Lomond, and B. Sanchez. 2020. The optimal ridge penalty for real-world high-dimensional data can be zero or negative due to the implicit ridge regularization. *Journal of Machine Learning Research* 21:1–16.

Koukoudakis, N., C. Koukouvinos, A. Lappa, M. Mitrouli, and A. Psitou. 2025. Numerical methods in modeling with supersaturated designs. *Applied Numerical Mathematics* 208:271–83. doi: 10.1016/j.apnum.2024.02.003.

LeJeune, D., P. Patil, H. Javadi, R. G. Baraniuk, and R. J. Tibshirani. 2024. Asymptotics of the sketched pseudoinverse. *SIAM Journal on Mathematics of Data Science* 6 (1):199–225. doi: 10.1137/22M1530264.

Patil, P., J. H. Du, and R. Tibshirani. 2024. Optimal ridge regularization for out-of-distribution prediction. *Proceedings 41st International Conference on Machine Learning*, 39908–39954, Vienna, Austria.

Tsigler, A., and P. Bartlett. 2023. Benign overfitting in ridge regression. *Journal of Machine Learning Research* 24:1–76.

Ullah, I., and A. Welsh. 2024. *On the effect of noise on fitting linear regression models*, 2024, math.ST, arXiv:2408.07914v1.

Winkler, J. R. 2024. *Regularisation, overfitting and condition estimation in regression*, Submitted.

Winkler, J. R., and M. Mitrouli. 2020. Condition estimation for regression and feature selection. *Journal of Computational and Applied Mathematics* 373:112212. doi: 10.1016/j.cam.2019.03.041.

Winkler, J. R., M. Mitrouli, and C. Koukouvinos. 2022. The application of regularisation to variable selection in statistical modelling. *Journal of Computational and Applied Mathematics* 404:113884. doi: 10.1016/j.cam.2021.113884.

Wu, D., and J. Xu. 2020. On the optimal weighted ℓ_2 regularization in overparameterized linear regression. *Advances in Neural Information Processing Systems* 33:10112–23.