This is a repository copy of *Condition estimation for regression and feature selection*.

White Rose Research Online URL for this paper:
http://eprints.whiterose.ac.uk/145647/

Version: Accepted Version

Article:
Winkler, J.R. and Mitrouli, M. (2019) Condition estimation for regression and feature selection. Journal of Computational and Applied Mathematics. ISSN 0377-0427

https://doi.org/10.1016/j.cam.2019.03.041

Article available under the terms of the CC-BY-NC-ND licence
(https://creativecommons.org/licenses/by-nc-nd/4.0/).

Reuse
This article is distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs (CC BY-NC-ND) licence. This licence only allows you to download this work and share it with others as long as you credit the authors, but you can’t change the article in any way or use it commercially. More information and the full terms of the licence here: https://creativecommons.org/licenses/

Takedown
If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.

eprints@whiterose.ac.uk
https://eprints.whiterose.ac.uk/

Condition estimation for regression and feature selection

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

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

[b] *Department of Mathematics, National and Kapodistrian University of Athens, Panepistimioupolis, GR-157 84, Athens, Greece*

j.r.winkler@sheffield.ac.uk, mmitroul@math.uoa.gr

Abstract

Regression and feature selection require the minimisation of $\|X\beta - y\|_2$ with respect to β, where $X \in \mathbb{R}^{n\times p}$, $n < p$ in feature selection and $n \geq p$ in regression. The vector β contains the coefficients of the basis functions in regression, and the weights of the features in feature selection. This paper considers the stability of β, as measured by the ratio of its relative error with respect to the relative error in y, and it is shown that the condition number $\kappa(X)$ of X is not a good measure of this stability. In particular, a large value of $\kappa(X)$ may lead to incorrect conclusions about the stability of β and it may be thought regularisation must be applied to the normal equation $X^T X\beta = X^T y$ if $\kappa(X) \gg 1$, but its application may lead to a large error in β. It is shown in this paper that (a) the presence of noise in y or the condition $\kappa(X) \gg 1$ do not imply that regularisation must be applied to the normal equation, and (b) the condition $\kappa(X) \gg 1$ does not imply that a small relative error in y yields a large relative error in β. These disadvantages of $\kappa(X)$ lead to the effective condition number $\eta(X, y)$, which provides a better measure of the stability of β due to a perturbation in y, but it is difficult to compute it reliably in some circumstances. Regularisation requires that a constraint be imposed on the solution of the normal equation, and it is shown that a constraint on $\|\beta\|_1$ can be interpreted in terms of the column sums of X, and that a constraint on $\|\beta\|_2$ can be interpreted in terms of the singular value decomposition of X. The paper contains several examples that illustrate the theoretical results.

Key words: Condition estimation; regression; Tikhonov regularisation; feature selection

1 Introduction

The determination of the equation of a curve that approximates a set of points (x_i, y_i), $i = 1, \ldots, n$, and the determination of the important features that characterise a system yield the minimisation

$$\min_{\beta} \|X\beta - y\|_2 \,, \tag{1}$$

where $X \in \mathbb{R}^{n\times p}$, $\beta \in \mathbb{R}^p$, $y \in \mathbb{R}^n$ and $\operatorname{rank} X = \min(n, p)$. In regression, the entries of X are functions of the coordinate points x_i and the p basis functions that are used for the representation of the approximating curve, y contains the data values and β contains the coefficients of the basis functions. Problems in regression require $n \geq p$, but a different situation arises in feature selection because $n < p$ and there therefore exists an infinite number of solutions β of (1) because X has a non-trivial null space [1,3,8,9,12]. The situation defined by $n < p$ also arises in supersaturated designs, in which case X is called the design matrix [2,10].

The minimisation (1) yields the normal equation

$$(X^T X)\beta = X^T y, \tag{2}$$

which has, as noted above, an infinite number of solutions if $n < p$ because $X^T X$ is a square matrix of order p and rank n. A unique solution is obtained by the addition of a constraint, which is imposed either in the 1-norm and leads to the lasso [8, §3.4.2], [11, §13.3], [12], or in the 2-norm and leads to ridge regression (also known as Tikhonov regularisation) [8, §3.4.1], [9], [11, §7.5]. The elastic net is a weighted linear combination of the lasso and ridge regression, and it therefore requires the minimisation of [13]

$$\|X\beta - y\|_2^2 + \alpha_1 \|\beta\|_1 + \alpha_2 \|\beta\|_2^2 \,, \tag{3}$$

with respect to β, where $\alpha_1, \alpha_2 \geq 0$, $\alpha_1 = 0$ yields ridge regression and $\alpha_2 = 0$ yields the lasso.

The requirement in feature selection is a sparse solution of (2), that is, a solution, many of whose components are zero, because it is desired to identify the dominant features that characterise the system. A true measure of sparsity of a vector v is $\|v\|_0^0$, which is equal to the number of non-zero entries in v, but $\|v\|_0$ does not satisfy the triangle inequality because $\|\cdot\|_l$ is not convex for $l < 1$, and thus $\|\cdot\|_0$ is not a norm. The smallest value of l that yields a norm is $l = 1$ because $\|\cdot\|_l$ is convex for $l \geq 1$. The 1-norm is therefore used for

feature selection because it is the best convex approximation of a non-convex function.

An expression for the solution of (2) for arbitrary values of n and p is obtained in Section 2, and this leads to Section 3, which considers the computation of the coefficients of a curve that approximates a set of points. It is shown that a large value of the condition number $\kappa(X)$ of X does not imply that (2) is ill-conditioned because the solution β is a function of X and y, but $\kappa(X)$ is a function of X only and it is independent of y. This leads to Section 4, which considers a refined normwise condition number of (2), called the effective condition number $\eta(X, y)$, which is a function of X and y, and therefore a better measure of the stability of (2). The discrete Picard condition is introduced and it is shown that it plays a crucial role in the characterisation of ill-conditioned equations [5,6]. If this condition is satisfied, then $\eta(X, y) \approx \kappa(X)$ and Tikhonov regularisation must be applied to (2) if $\kappa(X) \gg 1$, in which case an equation whose solution $\beta(\lambda)$ is formed, where $\lambda \geq 0$ is the regularisation parameter whose optimal value λ^* must be computed.

Tikhonov regularisation is considered in Section 5 and the error between the solutions $\beta(\lambda^*)$ and $\beta(0)$ is calculated, where $\beta(0)$ is the solution (2), that is, $\beta(0) = \beta$. It is shown that this error is small if the discrete Picard condition is satisfied, but it is large if this condition is not satisfied. It is also shown that $\beta(\lambda^*)$ is numerically stable and it is concluded that the discrete Picard condition is important for condition estimation and Tikhonov regularisation, such that if this condition is satisfied, then Tikhonov regularisation yields a numerically stable solution whose error is small.

It is shown in Section 6 that the effective condition number of the inverse problem $(X^T X)\beta_0 = X^T y_0$ and the effective condition number of the forward problem $y_0 = X\beta_0$, that is, the effective condition number of the computation of y_0 given X and β_0, satisfy an uncertainty principle because their product is equal to $\kappa(X)$. It follows that these computations cannot be simultaneously well-conditioned or simultaneously ill-conditioned if $\kappa(X) \gg 1$.

The 2-norm is used in Sections 3, 4 and 5, and this allows expressions for condition numbers and errors in terms of the singular values of X to be derived. Although this norm is frequently used, the 1-norm is preferred for feature selection because, as noted above, this norm is the best convex approximation of a non-convex function. Section 7 considers condition estimation and error analysis in the 1-norm, but the absence of a natural matrix decomposition in this norm makes it harder to derive expressions for condition numbers and errors in this norm. The paper is summarised in Section 8.

2 The solutions

This section considers the solution of (2) for arbitrary values of n and p, where $\operatorname{rank} X = \min(n, p)$. The SVD of X is USV^T, where U and V are orthogonal matrices,

$$
\begin{aligned}
S &= \begin{bmatrix} S_1 \; 0_{n,p-n} \end{bmatrix}, & S_1 &\in \mathbb{R}^{n\times n}, & n &< p, \\
S &= S_1, & S_1 &\in \mathbb{R}^{n\times n}, & n &= p, \\
S &= \begin{bmatrix} S_1 \\ 0_{n-p,p} \end{bmatrix}, & S_1 &\in \mathbb{R}^{p\times p}, & n &> p,
\end{aligned}
$$

the subscripts on the zero matrices indicate their order and S_1 is a square diagonal matrix whose entries are the singular values s_i of X, arranged in non-increasing order. Equation (2) has an infinite number of solutions if $n < p$,

$$
\beta = V \begin{bmatrix} S_1^{-1} \\ 0_{p-n,n} \end{bmatrix} U^T y + V \begin{bmatrix} 0_n \\ r \end{bmatrix}, \tag{4}
$$

where the terms on the right hand side are orthogonal, $r \in \mathbb{R}^{p-n}$ is an arbitrary vector and the second term on the right hand side lies in the null space of X. It follows from (4) that the minimum norm solution for $n < p$ is

$$
\beta = X^T \left(XX^T\right)^{-1} y = V \begin{bmatrix} S_1^{-1} \\ 0_{p-n,n} \end{bmatrix} U^T y,
$$

and only this solution, and not the infinite set of solutions (4), is considered in feature selection [11–13]. With this restriction, the solution of (2) is

$$
\beta = \begin{cases}
V \begin{bmatrix} S_1^{-1} \\ 0_{p-n,n} \end{bmatrix} U^T y, & n < p, \quad \text{unique minimum norm solution,} \\
V S_1^{-1} U^T y, & n = p, \quad \text{unique solution,} \\
V \begin{bmatrix} S_1^{-1} \; 0_{p,n-p} \end{bmatrix} U^T y, & n > p, \quad \text{unique solution,}
\end{cases}
$$

and these solutions can be combined into one equation,

$$
\beta = X^\dagger y = V S^\dagger U^T y, \tag{5}
$$

where the superscript † denotes pseudo-inverse. This expression for β is the solution of (2) for arbitrary values of n and p if $\operatorname{rank} X = \min(n, p)$ and only the minimum norm solution is considered for $n < p$.

3 Regression

This section considers the computation of the equation of a curve that passes through a set of points and it is shown that for a given matrix X, the addition of noise to y may cause a minor change in β or it may cause a major change in β. This result shows the problems that may arise when the condition number of X is used to draw conclusions about the stability of β due to a perturbation in y.

It follows from (5) that the condition number $\kappa_2(X)$ of X is

$$\kappa_2(X) = \max_{\delta y, y \in \mathbb{R}^n} \frac{\Delta\beta}{\Delta y} = \frac{s_1}{s_t}, \qquad t = \min(n, p), \tag{6}$$

where

$$\Delta\beta = \frac{\|\delta\beta\|_2}{\|\beta\|_2}, \qquad \Delta y = \frac{\|\delta y\|_2}{\|y\|_2}.$$

It is often stated that a large value of $\kappa_2(X)$ implies that (2) is ill-conditioned because a small relative error Δy in y leads to a large relative error $\Delta\beta$ in β, but Example 3.1 shows that this is incorrect. It is also often claimed that the presence of noise in y requires that regularisation must be applied to this equation, but this is also incorrect. Furthermore, it is shown in Section 5 that the application of regularisation to (2) when it is not required leads to a large error in β.

Example 3.1 Regression was performed on two sets of 100 data points (x_i, y_i), $i = 1, \ldots, 100$, where the points x_i lie in the interval $I = [1, \ldots, 20]$. A linear combination of 33 radial basis functions was used,

$$y_i = \sum_{k=1}^{33} a_k \exp\left(-\frac{(x_i - d_k)^2}{2\sigma_d^2}\right), \qquad i = 1, \ldots, 100,$$

where $\sigma_d = 1.35$ and the centres d_k of the basis functions are not uniformly distributed in I. The coefficient matrix X is therefore of order 100×33, y stores the function values y_i, and β stores the coefficients a_k of the radial basis functions.

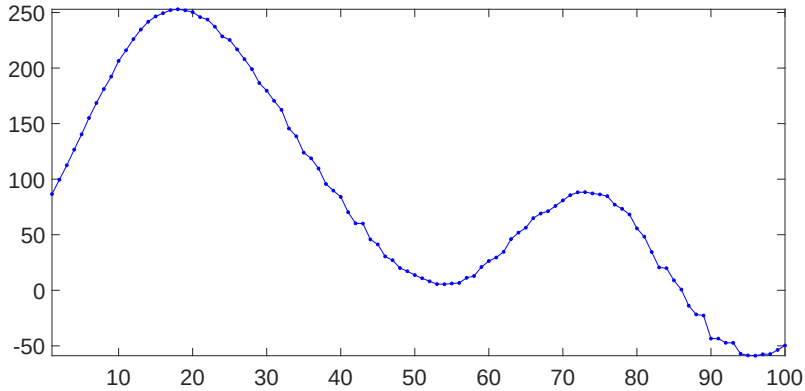

Fig. 1. A set of 100 points and their approximating curve for $y = y_1$.

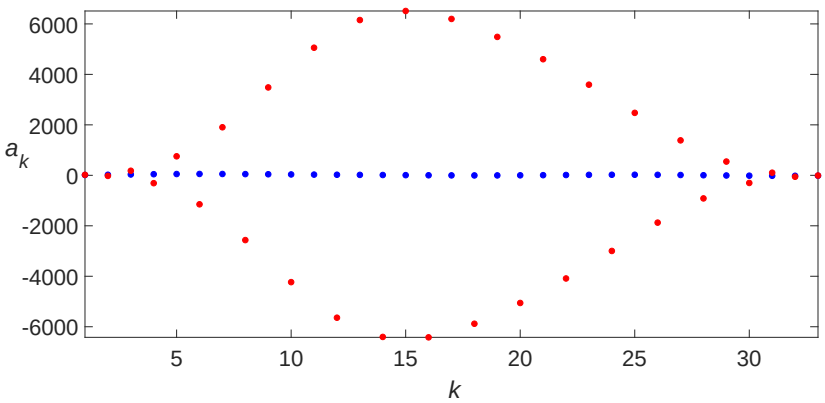

Fig. 2. The coefficients a_k for (a) $y = y_1$, ●, and (b) $y = y_1 + \delta y_1$, ●.

The first set of 100 points, which are contained in the vector $y = y_1$, and their regression curve are shown in Figure 1. Equation (2) was solved twice, once for the exact data $y = y_1$ and once for the perturbed data $y = y_1 + \delta y_1$, where the components of δy_1 are drawn from a zero mean Gaussian distribution, such that $\|y_1\|_2 / \|\delta y_1\|_2 = 2.93 \times 10^5$. The coefficients a_k for these vectors are shown in Figure 2 and it is seen that the noise has a significant effect because the error in the coefficients is large. In particular, the maximum values of $|a_k|$ for $y = y_1$ and $y = y_1 + \delta y_1$ are about 250 and 6,200, respectively.

The experiment was repeated for the second set of data points, which are contained in the vector $y = y_2$. These points and their regression curve are shown in Figure 3. Gaussian random noise of zero mean was added to each component of y, thereby forming the vector $y = y_2 + \delta y_2$, such that $\|y_2\|_2 / \|\delta y_2\|_2 = 1.21 \times 10^3$. The coefficients a_k for these vectors are shown in Figure 4 and it is seen that the noise δy_2 has little effect on the coefficients.

The condition number $\kappa_2(X)$ of X is 5.12×10^8, which would suggest that (2) is ill-conditioned and it may therefore be thought that bad results are

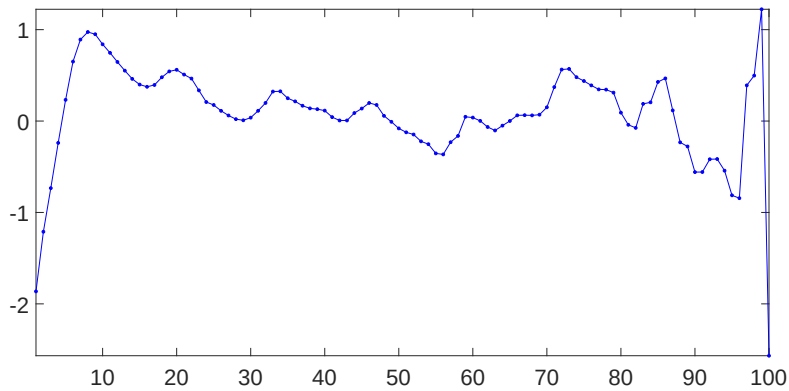

Fig. 3. A set of 100 points and their approximating curve for $y = y_2$.

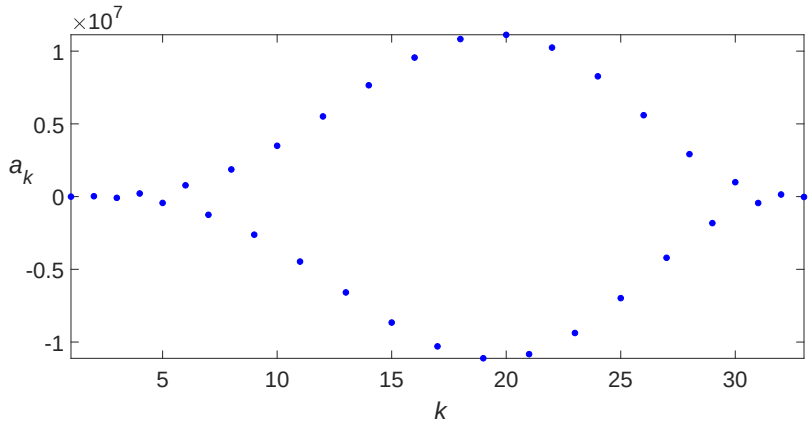

Fig. 4. The coefficients a_k for $y = y_2$ and $y = y_2 + \delta y_2$.

expected for $y = y_1 + \delta y_1$ and $y = y_2 + \delta y_2$. The computational experiments show, however, that bad results are obtained for $y = y_1 + \delta y_1$ and good results are obtained for $y = y_2 + \delta y_2$. This difference cannot be explained by the large value of $\kappa_2(X)$ because the same matrix X is used for both sets of data. $\square$

The next section considers the effective condition number of (2). This is a more accurate condition estimate because it is a function of X and y, unlike the condition number, which is a function of X only. It is shown that this accurate condition estimate allows the results in Example 3.1 to be explained.

4 The effective condition number

Example 3.1 shows that $\kappa_2(X)$ is not a good measure of the stability of β because there exist vectors y for which (2) is stable, and vectors y for which this equation is unstable. As noted above, this problem arises because $\kappa_2(X)$ is a function of X only, and β is a function of X and y. It is therefore necessary

to develop a measure of stability that is a function of X and y. This refined measure is called the effective condition number and denoted $\eta_p(X, y)$, $p = 1, 2, \infty$. [1]

Theorem 4.1 Let the relative errors $\Delta\beta$ and Δy be

$$\Delta\beta = \frac{\|\delta\beta\|_p}{\|\beta\|_p}, \qquad \Delta y = \frac{\|\delta y\|_p}{\|y\|_p}, \qquad p = 1, 2, \infty.$$

The effective condition number $\eta_p(X, y)$ of (2) is equal to the maximum value of the ratio of $\Delta\beta$ to Δy with respect to all perturbations $\delta y \in \mathbb{R}^n$,

$$\eta_p(X, y) = \max_{\delta y \in \mathbb{R}^n} \frac{\Delta\beta}{\Delta y} = \frac{\left\|X^\dagger\right\|_p \|y\|_p}{\|X^\dagger y\|_p}. \tag{7}$$

If $p = 2$, the effective condition number $\eta_2(X, y)$ is given by

$$\eta_2(X, y) = \frac{1}{s_t} \frac{\|c\|_2}{\|S^\dagger c\|_2}, \qquad c = U^T y, \tag{8}$$

where t is defined in (6).

Proof It follows from (5) that

$$\|\delta\beta\|_p \leq \left\|X^\dagger\right\|_p \|\delta y\|_p = \left\|X^\dagger\right\|_p \|y\|_p \, \Delta y,$$

and the division of both sides of this inequality by $\|\beta\|_p = \left\|X^\dagger y\right\|_p$ yields

$$\Delta\beta = \frac{\|\delta\beta\|_p}{\|\beta\|_p} \leq \left(\frac{\left\|X^\dagger\right\|_p \|y\|_p}{\|\beta\|_p} \right) \Delta y = \left(\frac{\left\|X^\dagger\right\|_p \|y\|_p}{\|X^\dagger y\|_p} \right) \Delta y,$$

from which (7) and (8) follow. □

The effective condition number is defined for $p = 1, 2, \infty$ in (7) and its form for $p = 2$ is given in (8). Subsequent analysis in this section, and Section 5, which considers Tikhonov regularisation, is restricted to $p = 2$ because extensive use

[1] The symbol p is used to denote the number of columns of X and the norm, 1, 2 or ∞, of vectors and matrices. The meaning of p is, however, clear from the context.

is made of the SVD of X. The absence of equivalent matrix decompositions for the 1- and ∞-norms makes analysis in these norms more difficult.

It is easily established that the minimum value of $\eta_p(X, y)$ is one, and that for $p = 2$, it is attained for $c = ke_t$ where k is an arbitrary constant and e_i is the ith unit basis vector. Upper bounds on $\eta_2(X, y)$ are considered in Theorem 4.2 and it is shown that the situations $n \leq p$ and $n > p$ must be considered separately.

Theorem 4.2 Equation (8) can be written as

$$\eta_2(X, y) = \frac{1}{s_t} \left(\frac{\sum_{i=1}^{n} c_i^2}{\sum_{i=1}^{t} \left(\frac{c_i}{s_i}\right)^2} \right)^{\frac{1}{2}} = \left(\frac{s_1}{s_t}\right) \left(\frac{\sum_{i=1}^{n} c_i^2}{\sum_{i=1}^{t} \left(\frac{s_1}{s_i}\right)^2 c_i^2} \right)^{\frac{1}{2}}, \tag{9}$$

and thus upper bounds on $\eta_2(X, y)$ are

$$\eta_2(X, y) \leq \begin{cases} \kappa_2(X) & \text{if} \quad n \leq p, \\ \kappa_2(X) \left(1 + \frac{\sum_{i=p+1}^{n} c_i^2}{\sum_{i=1}^{p} c_i^2}\right)^{\frac{1}{2}} & \text{if} \quad n > p. \end{cases} \tag{10}$$

Proof If $n \leq p$, then $t = n$ and thus (9) yields

$$\left(\frac{s_1}{s_t}\right) \left(\frac{\sum_{i=1}^{n} c_i^2}{\sum_{i=1}^{t} \left(\frac{s_1}{s_i}\right)^2 c_i^2} \right)^{\frac{1}{2}} = \left(\frac{s_1}{s_n}\right) \left(\frac{\sum_{i=1}^{n} c_i^2}{\sum_{i=1}^{n} \left(\frac{s_1}{s_i}\right)^2 c_i^2} \right)^{\frac{1}{2}} \leq \frac{s_1}{s_n},$$

and the result follows.

Consider now the situation $n > p$, for which $X^T(X\beta - y) = 0$ does not imply $X\beta = y$ because $X^T v = 0$ for at least one vector $v \neq 0$. If β_0 is the solution of (2), then the residual r of $X\beta_0 = y$ for $n > p$ is

$$r = X\beta_0 - y,$$

and it is easily verified that

$$\|X\beta_0\|_2^2 = \left\|XX^\dagger y\right\|_2^2 \qquad \text{and} \qquad \|r\|_2^2 + \|X\beta_0\|_2^2 = \|y\|_2^2 .$$

If the angle θ satisfies

$$\|r\|_2 = \|y\|_2 \sin\theta \qquad \text{and} \qquad \|X\beta_0\|_2 = \|y\|_2 \cos\theta,$$

then

$$\cos\theta = \frac{\|X\beta_0\|_2}{\|y\|_2} = \frac{\left\|XX^\dagger y\right\|_2}{\|y\|_2} = \left(\frac{\sum_{i=1}^{p} c_i^2}{\sum_{i=1}^{n} c_i^2}\right)^{\frac{1}{2}},$$

and thus

$$\eta_2(X,y) \leq \frac{\kappa_2(X)}{\cos\theta} \qquad \text{if} \qquad n > p,$$

from which (10) follows. □

It follows from Theorem 4.2 that if $n > p$, the minimum value of the upper bound of $\eta_2(X,y)$ occurs when $\cos\theta = 1$, that is, when $c_i = 0, i = p+1, \ldots, n$. The maximum value of this upper bound, $\eta_2(X,y) \to \infty$, occurs when $c_i = 0, i = 1, \ldots, p$.

If $n = p$, then

$$\eta_2(X,y) = \frac{1}{s_n}\frac{\|c\|_2}{\|S^{-1}c\|_2} = \frac{1}{s_n}\frac{\left\|U^T y\right\|_2}{\|S^{-1}U^T y\|_2},$$

and thus

$$\max_{y\in\mathbb{R}^n} \eta_2(X,y) = \max_{y\in\mathbb{R}^n} \frac{1}{s_n}\frac{\left\|U^T y\right\|_2}{\|S^{-1}U^T y\|_2} = \frac{s_1}{s_n} = \kappa_2(X),$$

which is attained when $U^T y$ is proportional to e_1, that is, y is aligned along the first column of U. This equation shows the difference between $\eta_2(X,y)$ and $\kappa_2(X)$ because $\eta_2(X,y)$ is equal to the maximum value of $\mu = {}^{\Delta\beta}/_{\Delta y}$ with respect to all vectors $\delta y \in \mathbb{R}^n$ but $\kappa_2(X,y)$ is equal to the maximum value of μ with respect to all vectors $\delta y, y \in \mathbb{R}^n$. It is clear that the relative errors $\Delta\beta$ and Δy for the calculation of μ are defined in the 2-norm.

It follows from (9) that $\eta_2(X,y) \approx \kappa_2(X)$ if $n \leq p$ or $c_i = 0, i = p+1, \ldots, n$, and the discrete Picard condition is satisfied [5,6].

Definition 4.1 (The discrete Picard condition) The discrete Picard condition requires that the ratio ${}^{|c_i|}/_{s_i}$ decrease to zero as $i \to t$,

$$\frac{|c_i|}{s_i} \to 0 \qquad \text{as} \qquad i \to t, \tag{11}$$

which implies that the constants $|c_i|$ tend to zero faster than the singular values tend to zero. □

If the discrete Picard condition is satisfied, then

$$\sum_{i=1}^{t} \frac{c_i^2}{s_i^2} \approx \frac{c_1^2}{s_1^2} \qquad \text{and} \qquad \sum_{i=1}^{t} c_i^2 \approx c_1^2,$$

and it follows from (9) that $\eta_2(X, y) \approx {}^{s_1}/_{s_t} = \kappa_2(X)$. More generally, it follows from $y = Uc$ that (2) is ill-conditioned if the dominant components of y lie along the columns u_i of U that are associated with small values of i. The solution β is therefore dominated by the contribution of the large singular values of X, and the small singular values are not significant. Similarly, it follows from (8) that (2) is well-conditioned if $\eta_2(X, y) \approx 1$, that is, the dominant components of y lie along the columns u_i of U that are associated with large values of i. It therefore follows that β is dominated by the small singular values of X, and the contribution of the large singular values is small.

Example 4.1 Consider the vectors y_1 and y_2 in Example 3.1. It was shown that (2) is ill-conditioned for $y = y_1$ and it is well-conditioned for $y = y_2$. These results are confirmed by the effective condition numbers of (X, y_1) and (X, y_2) because

$$\eta_2(X, y_1) = 4.62 \times 10^8 \qquad \text{and} \qquad \eta_2(X, y_2) = 7.94.$$

□

It follows from Theorem 4.1 and Example 4.1 that the effective condition number $\eta_2(X, y)$ provides a good measure of the stability of (2). More careful analysis shows, however, that if the discrete Picard condition (11) is satisfied, then $\eta_2(X, y)$ cannot be computed accurately because it is sensitive to noise δy. In particular, consider the term $\|\beta\|_2 = \left\|S^\dagger c\right\|_2$ in the denominator of (8). If noise is present, then the square of the magnitude of the perturbed solution of (2) is

$$\|\beta + \delta\beta\|_2^2 = \left\|X^\dagger(y + \delta y)\right\|_2^2 = \left\|S^\dagger(c + \delta c)\right\|_2^2 = \sum_{i=1}^{t} \left(\frac{c_i + \delta c_i}{s_i}\right)^2, \tag{12}$$

which is equal to the sum of the squares of the terms, in the presence of noise, that define the discrete Picard condition. If the magnitude of the perturbations

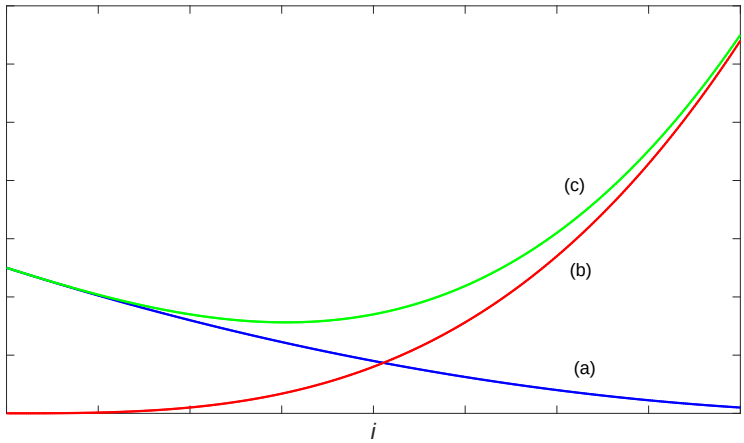

Fig. 5. The ratios (a) $|c_i|/s_i$, (b) $|\delta c_i|/s_i$ and (c) $|c_i+\delta c_i|/s_i$ if the discrete Picard condition is satisfied.

δc_i is approximately constant, $|\delta c_i| \approx \epsilon$, $i = 1, \ldots, t$, such that

$$\begin{aligned} |\delta c_i| &\ll |c_i|, \quad i = 1, \ldots, r-1, \\ |\delta c_i| &\approx |c_r|, \quad i = r, \\ |\delta c_i| &\gg |c_i|, \quad i = r+1, \ldots, t, \end{aligned} \tag{13}$$

and the discrete Picard condition is satisfied, then

$$\frac{|c_i + \delta c_i|}{s_i} \approx \begin{cases} \frac{|c_i|}{s_i} \gg \frac{|\delta c_i|}{s_i}, & i = 1, \ldots, r-1, \\ \frac{|c_r+\delta c_r|}{s_r}, & i = r, \\ \frac{|\delta c_i|}{s_i} \gg \frac{|c_i|}{s_i}, & i = r+1, \ldots, t, \end{cases} \tag{14}$$

and the ratios $|c_i|/s_i$, $|\delta c_i|/s_i$ and $|c_i+\delta c_i|/s_i$ are shown in Figure 5. It follows from (12) and (14) that if the discrete Picard condition is satisfied, then $\|\beta + \delta\beta\|_2$ is dominated by noise because of the contribution to the sum of the terms defined by $i = r+1, \ldots, t$. The ratio $|c_i|/s_i$ cannot, therefore, be computed reliably in the presence of noise, and thus the satisfaction, or otherwise, of the discrete Picard condition cannot be determined. The practical implications of this result are considered at the end of this section.

The discrete Picard condition is defined in (11) and its satisfaction guarantees that (2) is ill-conditioned. The importance of this condition is most easily seen by considering other profiles of this ratio. For example, if (13) is satisfied and the ratio $|c_i|/s_i$ is approximately constant,

$$|c_i| \approx s_i, \qquad i = 1, \ldots, t, \tag{15}$$

then $\eta_2(X, y)$ cannot be computed reliably because

$$\frac{|c_i + \delta c_i|}{s_i} \approx \frac{|\delta c_i|}{s_i}, \qquad i = r+1, \ldots, t.$$

If, however, the constants $|c_i|$ increase sufficiently rapidly with i, then

$$|c_{i+1}| \gg |c_i| \qquad \text{and} \qquad \frac{|c_{i+1}|}{s_{i+1}} \gg \frac{|c_i|}{s_i}, \qquad i = 1, \ldots, t-1, \tag{16}$$

and the magnitude of the perturbations $|\delta c_i|$ is approximately constant, $|\delta c_i| \approx \epsilon$, $i = 1, \ldots, t$, such that

$$\begin{aligned} |\delta c_i| &\geq |c_i|, \quad i = 1, \ldots, r, \\ |\delta c_i| &\ll |c_i|, \quad i = r+1, \ldots, t, \end{aligned} \tag{17}$$

then $\eta_2(X, y)$ can be computed reliably because the effect of the perturbations δc_i is small.

It is clear that the forms (11), (15) and (16) of the ratio $|c_i|/s_i$ are not satisfied exactly in practical examples, but the analysis in this section and Section 5, which considers Tikhonov regularisation, shows their importance for condition estimation and error analysis. In particular, they characterise the matrix-vector pairs (X, y) for which (2) is ill-conditioned and therefore requires regularisation, and the matrix-vector pairs (X, y) for which (2) is well-conditioned and regularisation must not be imposed.

Example 4.2 Consider the vectors y_1 and y_2 in Example 3.1. Figure 6 shows the ratios $\log_{10} |c_i|/s_i$ and $\log_{10} |c_i + \delta c_i|/s_i$ for the exact data $y = y_1$ and the perturbed data $y = y_1 + \delta y_1$, and Figure 7 shows the same quantities for the exact data $y = y_2$ and the perturbed data $y = y_2 + \delta y_2$. Figure 6 shows that the data $y = y_1$ satisfies the discrete Picard condition, but this condition is not satisfied for $i > 18$ if noise δy_1 is present. The figure suggests that $r \approx 18$, where r is defined in (13) and (14), and thus the sum (12) is dominated by noise, that is, the terms defined by $i = 19, \ldots, 33$. By contrast, Figure 7 shows that the noise δy_2 has very little effect on the ratio $|c_i|/s_i$ for the data $y = y_2$ and that the dominant term in the sum (12) is

$$\left(\frac{c_{33} + \delta c_{33}}{s_{33}}\right)^2 \approx \left(\frac{c_{33}}{s_{33}}\right)^2.$$

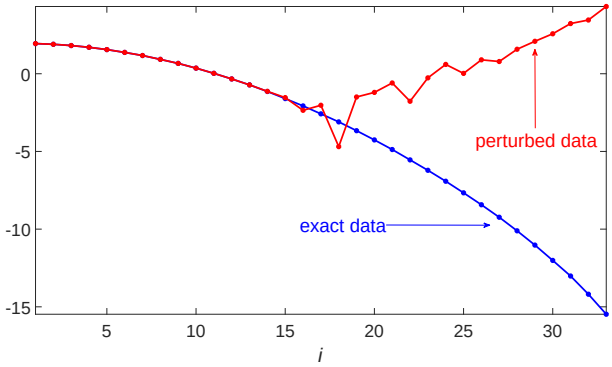

Fig. 6. The ratio $\log_{10} {}^{|c_i|}/_{s_i}$ for the exact data $y = y_1$ and the ratio $\log_{10} {}^{|c_i+\delta c_i|}/_{s_i}$ for the perturbed data $y = y_1 + \delta y_1$.

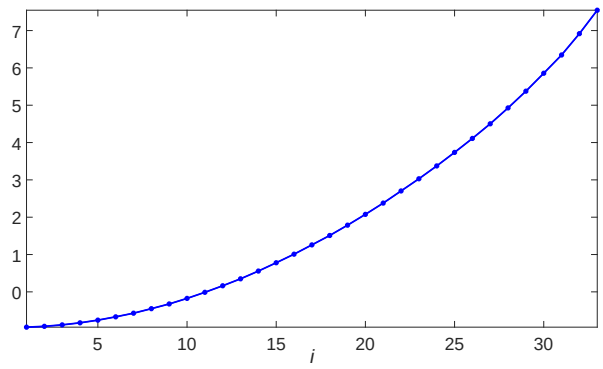

Fig. 7. The ratio $\log_{10} {}^{|c_i|}/_{s_i}$ for the exact data $y = y_2$ and the ratio $\log_{10} {}^{|c_i+\delta c_i|}/_{s_i}$ for the perturbed data $y = y_2 + \delta y_2$.

□

The analysis above shows that if (11) or (15) are satisfied, then the effective condition number $\eta_2(X, y)$ cannot be computed reliably. This is a problematic result because regularisation is required if either of these equations is satisfied, but the effective condition number cannot be used to determine this requirement. In this situation, prior information must be given in order to establish if (2) is ill-conditioned, and this determines the form of regularisation to be applied. For example, the removal of blur from an image yields an ill-conditioned equation and it is known that the theoretically exact solution (the deblurred image) satisfies the discrete Picard condition, in which case Tikhonov regularisation or truncated singular value decomposition can be used to obtain a computationally stable solution that has a small error [7, §5.6].

A different situation arises if (16) and (17) are satisfied because $\eta_2(X, y)$ is approximately independent of the singular values of X and it can be computed reliably,

$$\eta_2(X,y) = \frac{1}{s_t}\frac{\|c\|_2}{\|S^\dagger c\|_2} \approx \frac{|c_t|}{s_t\left(\frac{|c_t|}{s_t}\right)} = 1.$$

5 Tikhonov regularisation

It was shown in Section 4 that (2) is ill-conditioned if the discrete Picard condition (11) is satisfied, and it is shown in this section that, in this case, Tikhonov regularisation yields an approximate solution $\beta(\lambda^*)$ of (2), where $\lambda^* \geq 0$ is the optimal value of the regularisation parameter. It is shown in Section 5.1 that the error between this solution and the theoretically exact solution of (2), called the regularisation error, is small. An expression for the effective condition number of $\beta(\lambda^*)$ is developed in Section 5.2 and it is shown that it is much smaller than the effective condition number $\eta_2(X,y)$ of (2). The satisfaction of these two conditions implies that $\beta(\lambda^*)$ is a regularised solution of (2).

5.1 The regularisation error

The specification $\alpha_1 = 0$ and $\lambda = \alpha_2$ in (3) yields Tikhonov regularisation, which requires the solution of

$$(X^TX + \lambda I)\beta(\lambda) = X^Ty, \qquad \lambda \geq 0, \tag{18}$$

and thus

$$\beta(\lambda) = V(S^TS + \lambda I)^{-1}S^Tc, \tag{19}$$

where c is defined in (8). It is assumed $n \geq p$, that is, $t = p$, and thus this equation is written as

$$\beta(\lambda) = V\left((S^TS + \lambda I)^{-1}S^TS\right)S^\dagger c = VFS^\dagger c,$$

where

$$SS^\dagger = S(S^TS)^{-1}S^T = \begin{bmatrix} I_p & 0_{p,n-p} \\ 0_{n-p,p} & 0_{n-p,n-p} \end{bmatrix},$$

and F is a square diagonal matrix of order p,

$$F = (S^T S + \lambda I)^{-1} S^T S = \text{diag}\ \{f_i\}_{i=1}^p = \text{diag}\left\{\frac{s_i^2}{s_i^2 + \lambda}\right\}_{i=1}^p .$$

The functions f_i are called filter factors because λ determines the singular values that are removed from, and the singular values that are retained in, the solution $\beta(\lambda)$. For example, if $\lambda \approx s_r^2$ and the singular values are well-separated, then

$$\begin{aligned} f_i &\approx 1, \quad i = 1, \ldots, r-1, \\ f_i &\approx \tfrac{1}{2}, \quad i = r, \\ f_i &\approx 0, \quad i = r+1, \ldots, p, \end{aligned}$$

and thus only the first r singular values are retained in $\beta(\lambda)$. Furthermore, the condition $\lambda = 0$ yields $F = I_p$, which corresponds to the solution (5).

It is assumed y is known exactly and thus only the regularisation error, that is, the error caused by the term λI in the coefficient matrix in (18), is considered. This error is denoted $\Delta(\lambda)$ and it is considered in Theorem 5.1.

Theorem 5.1 If $p = \min(n, p)$, the regularisation error $\Delta(\lambda)$ is given by

$$\Delta(\lambda) = \frac{\|\beta(\lambda) - \beta(0)\|_2}{\|\beta(0)\|_2} = \lambda \left(\frac{\sum_{i=1}^p \left(\frac{c_i}{s_i}\right)^2 \frac{1}{(s_i^2+\lambda)^2}}{\sum_{i=1}^p \left(\frac{c_i}{s_i}\right)^2} \right)^{\frac{1}{2}} . \tag{20}$$

Proof It follows from (19) that

$$\beta(\lambda) - \beta(0) = V \left((S^T S + \lambda I)^{-1} - (S^T S)^{-1} \right) S^T c, \tag{21}$$

where $S^T S$ is a square non-singular matrix of order p, and

$$\begin{aligned} (S^T S + \lambda I)^{-1} - (S^T S)^{-1} &= \left(\left(I + \lambda (S^T S)^{-1} \right)^{-1} - I \right) (S^T S)^{-1} \\ &= \text{diag}\left\{ \left(\frac{1}{1 + \frac{\lambda}{s_i^2}} \right) - 1 \right\} \text{diag}\left\{ \frac{1}{s_i^2} \right\} \\ &= \text{diag}\left\{ \frac{-\lambda}{s_i^2 + \lambda} \right\} \text{diag}\left\{ \frac{1}{s_i^2} \right\}, \end{aligned}$$

for $i = 1, \ldots, p$. It therefore follows from (21) that

$$V^T\left(\beta(\lambda)-\beta(0)\right) = \operatorname{diag}\left\{\frac{-\lambda}{s_i(s_i^2+\lambda)}\right\}\begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_p \end{bmatrix}$$

$$= -\lambda\left[\tfrac{c_1}{s_1(s_1^2+\lambda)} \;\; \tfrac{c_2}{s_2(s_2^2+\lambda)} \;\; \cdots \;\; \tfrac{c_p}{s_p(s_p^2+\lambda)}\right]^T.$$

It follows from (19) and this equation that

$$\|\beta(0)\|_2^2 = \sum_{i=1}^{p}\left(\frac{c_i}{s_i}\right)^2 \quad \text{and} \quad \|\beta(\lambda)-\beta(0)\|_2^2 = \lambda^2\sum_{i=1}^{p}\left(\frac{c_i}{s_i}\right)^2\frac{1}{(s_i^2+\lambda)^2},$$

and the result (20) follows. □

Several methods for computing the optimal value λ^* of λ are discussed in [6, §7] and [7, §6.4], and it is assumed that λ^* satisfies

$$\begin{aligned} &\lambda^* \ll s_i^2, \quad i = 1,\ldots,r-1, \\ &\lambda^* \approx s_i^2, \quad i = r, \\ &\lambda^* \gg s_i^2, \quad i = r+1,\ldots,p. \end{aligned} \tag{22}$$

The error $\Delta(\lambda^*)$ is analysed by considering the forms of the ratio $|c_i|/s_i$, as shown in (11), (15) and (16).

Case 1 The discrete Picard condition (11) is satisfied, in which case

$$\max_{i=1,\ldots,p}\left\{\frac{|c_i|}{s_i}\right\} \approx \frac{|c_1|}{s_1},$$

and thus from (20) and (22),

$$\Delta(\lambda^*) \approx \frac{\lambda^*}{s_1^2+\lambda^*} \approx \frac{\lambda^*}{s_1^2} \ll 1. \tag{23}$$

Case 2 The constants c_i satisfy (15), in which case it follows from (22) that the numerator in (20) is given by

$$
\begin{aligned}
\sum_{i=1}^{p}\left(\frac{c_i}{s_i}\right)^2\left(\frac{\lambda^*}{s_i^2+\lambda^*}\right)^2 &\approx \sum_{i=1}^{p}\left(\frac{\lambda^*}{s_i^2+\lambda^*}\right)^2 \\
&= \sum_{i=1}^{r-1}\left(\frac{\lambda^*}{s_i^2+\lambda^*}\right)^2 + \left(\frac{\lambda^*}{s_r^2+\lambda^*}\right)^2 + \\
&\quad \sum_{i=r+1}^{p}\left(\frac{\lambda^*}{s_i^2+\lambda^*}\right)^2 \\
&\approx \sum_{i=1}^{r-1}\left(\frac{\lambda^*}{s_i^2}\right)^2 + \frac{1}{4} + (p-r) \\
&\approx p-r,
\end{aligned}
$$

and since $\|\beta_0\|_2^2 \approx p$, it follows from (20) that

$$
\Delta(\lambda^*) \approx \left(\frac{p-r}{p}\right)^{\frac{1}{2}} < 1. \tag{24}
$$

Case 3 Equation (16) is satisfied and thus

$$
\Delta(\lambda^*) \approx \frac{\lambda^*\left(\frac{c_p}{s_p}\right)\frac{1}{(s_p^2+\lambda^*)}}{\left(\frac{c_p}{s_p}\right)} = \frac{\lambda^*}{s_p^2+\lambda^*} \approx 1, \tag{25}
$$

because Tikhonov regularisation filters out all the components of $\beta(\lambda^*)$. In particular, the dominant components of y lie along the columns u_i of U that are associated with large values of i, and the small singular values of X are therefore significant in $\beta(\lambda^*)$. Tikhonov regularisation filters out, however, these small singular values and thus $\beta(\lambda^*) \approx 0$.

It follows from (23), (24) and (25) that the error $\Delta(\lambda^*)$ is strongly dependent on the form of the ratio $|c_i|/s_i$. In particular:

- Case 1: The ratio $|c_i|/s_i$ decreases and thus the discrete Picard condition (11) is satisfied. The error $\Delta(\lambda^*)$ is approximately inversely proportional to the reciprocal of the square of the largest singular value.
- Case 2: The ratio $|c_i|/s_i$ is approximately equal to one and the error $\Delta(\lambda^*)$ is approximately constant and therefore independent of the singular values.
- Case 3: The ratio $|c_i|/s_i$ increases and the error $\Delta(\lambda^*)$ is approximately equal to one, which is its maximum value because $\beta(\lambda^*) \approx 0$.

These results, and the results in Section 4, show that if the discrete Picard condition is satisfied, then (2) is ill-conditioned and Tikhonov regularisation yields an approximate solution $\beta(\lambda^*)$ and its regularisation error $\Delta(\lambda^*)$ is small. The condition that $\beta(\lambda^*)$ is a regularised solution of (2) also requires that it is numerically stable, and this issue is addressed in the next section by considering the effective condition number of (18) when $\lambda = \lambda^*$.

The effective condition number of (18) is easily established from Theorem 4.1 and (19). As in Section 5.1, it is assumed that $p = \min(n, p)$, in which case $S^T S$ is a square non-singular diagonal matrix of order p. If $\Delta\beta(\lambda^*)$ is the relative error in $\beta(\lambda^*)$, then the effective condition number of $\beta(\lambda^*)$ is

$$\eta_2(X, y, \lambda^*) = \max_{\delta y \in \mathbb{R}^n} \frac{\Delta\beta(\lambda^*)}{\Delta y} = \frac{\left\|(S^T S + \lambda^* I)^{-1} S^T\right\|_2 \|c\|_2}{\|(S^T S + \lambda^* I)^{-1} S^T c\|_2}. \tag{26}$$

The terms in the numerator and denominator in this expression are considered separately and it is shown that if the discrete Picard condition is satisfied, then $\eta_2(X, y, \lambda^*) \ll \eta_2(X, y)$.

The square of the term in the denominator of (26) is

$$\left\|(S^T S + \lambda^* I)^{-1} S^T c\right\|_2^2 = \sum_{i=1}^{p} \left(\frac{s_i c_i}{s_i^2 + \lambda^*}\right)^2 = \sum_{i=1}^{p} \left(\frac{s_i^2}{s_i^2 + \lambda^*}\right)^2 \left(\frac{c_i}{s_i}\right)^2, \tag{27}$$

and since the singular values are arranged in non-increasing order, $s_i > s_j$ for $i < j$, it follows that

$$\frac{s_i^2}{s_i^2 + \lambda^*} > \frac{s_j^2}{s_j^2 + \lambda^*} \qquad \text{if} \qquad i < j.$$

Since the discrete Picard condition is satisfied and λ^* satisfies (22), it follows from (27) that

$$\left\|(S^T S + \lambda^* I)^{-1} S^T c\right\|_2 \approx \frac{|c_1|}{s_1}. \tag{28}$$

Consider now the numerator in (26), for which, from (22),

$$
\begin{aligned}
\left\|(S^T S + \lambda^* I)^{-1} S^T\right\|_2 &= \max_{i=1,\ldots,p} \left\{ \frac{s_i}{s_i^2 + \lambda^*} \right\} \\
&= \max \left\{ \max_{i=1,\ldots,r-1} \left\{ \frac{s_i}{s_i^2 + \lambda^*} \right\}, \frac{s_r}{s_r^2 + \lambda^*}, \right. \\
&\qquad \left. \max_{i=r+1,\ldots,p} \left\{ \frac{s_i}{s_i^2 + \lambda^*} \right\} \right\} \\
&\approx \max \left\{ \max_{i=1,\ldots,r-1} \left\{ \frac{1}{s_i} \right\}, \frac{1}{2s_r}, \max_{i=r+1,\ldots,p} \left\{ \frac{s_i}{\lambda^*} \right\} \right\} \\
&\approx \max \left\{ \frac{1}{s_{r-1}}, \frac{1}{2s_r}, \frac{s_{r+1}}{s_r^2} \right\} \\
&= \left(\frac{1}{s_r} \right) \max \left\{ \frac{s_r}{s_{r-1}}, \frac{1}{2}, \frac{s_{r+1}}{s_r} \right\} \\
&= \frac{\gamma}{s_r}, \qquad \frac{1}{2} \leq \gamma \leq 1,
\end{aligned}
$$

and thus from (28),

$$
\eta_2(X, y, \lambda^*) \approx \gamma \left(\frac{\|c\|_2}{|c_1|} \right) \left(\frac{s_1}{s_r} \right) \approx \gamma \frac{s_1}{s_r} \ll \frac{s_1}{s_p} \approx \eta_2(X, y),
$$

because the satisfaction of the discrete Picard condition implies that the constants $|c_i|$ decrease to zero faster than the singular values decrease to zero, and thus $\|c\|_2 \approx |c_1|$. This is the desired result because the value $\lambda^* \approx s_r^2$ implies that only the first r singular values are included in the expression for $\beta(\lambda^*)$. The combination of this result and the analysis in Section 5.1 shows that if the discrete Picard condition is satisfied, then $\beta(\lambda^*)$ is a regularised solution of (2) because it is numerically stable and the regularisation error is small.

6 An uncertainty principle

It is shown in this section that the effective condition number of (2) for $y = y_0$ and $\beta = \beta_0$, and the effective condition number of $y_0 = X\beta_0$, that is, the effective condition number of the computation of y_0, given X and β_0, satisfy an uncertainty principle because these problems cannot be simultaneously ill-conditioned or simultaneously well-conditioned if $\kappa(X) \gg 1$ since the product of the effective condition numbers is equal to $\kappa(X)$. It is assumed, from (10), that

$$
n \leq p \qquad \text{or} \qquad c_i = 0,\ i = p+1, \ldots, n, \tag{29}
$$

in which case the maximum value of $\eta_2(X, y)$ is $\kappa_2(X)$, that is, $\eta_p(X, y)$, $p = 1, 2, \infty$, is finite.

It was shown in Theorem 4.1 that if $y = y_0$, then the effective condition number of the solution β_0 of the inverse problem (2) is

$$\eta_p^{\text{inv}}(X, y_0) = \max_{\delta y_0 \in \mathbb{R}^n} \frac{\Delta \beta_0}{\Delta y_0} = \frac{\left\|X^\dagger\right\|_p \|y_0\|_p}{\|X^\dagger y_0\|_p}, \qquad p = 1, 2, \infty. \tag{30}$$

Consider now the forward problem $y_0 = X\beta_0$, for which it is easily established that the effective condition number is

$$\eta_p^{\text{fwd}}(X, \beta_0) = \max_{\delta \beta_0 \in \mathbb{R}^p} \frac{\Delta y_0}{\Delta \beta_0} = \frac{\|X\|_p \|\beta_0\|_p}{\|X\beta_0\|_p}, \qquad p = 1, 2, \infty, \tag{31}$$

and thus the product of the effective condition numbers is

$$\eta_p^{\text{inv}}(X, y_0)\eta_p^{\text{fwd}}(X, \beta_0) = \|X\|_p \left\|X^\dagger\right\|_p = \kappa_p(X), \quad p = 1, 2, \infty. \tag{32}$$

It follows from this uncertainty principle that if the inverse problem is well-conditioned (ill-conditioned), then the forward problem is ill-conditioned (well-conditioned), assuming (29) is satisfied. This result cannot be obtained from $\kappa_p(X)$ because this measure of the stability of the forward and inverse problems suggests that these problems can be simultaneously ill-conditioned or simultaneously well-conditioned.

It was shown in Sections 4 and 5 that the discrete Picard condition is essential for condition estimation and regularisation in the 2-norm. Equation (32) is used in Section 7 to calculate the equivalent conditions for the effective condition number $\eta_1(X, y)$ and regularisation in the 1-norm.

7 Computations in the 1-norm

The regression of two sets of data points was considered in Example 3.1 and it was shown that a large value of the condition number $\kappa_2(X)$ does not imply that the solution β of (2) is sensitive to a perturbation in y. The equality of matrix norms implies that this result is also valid when the condition numbers $\kappa_1(X)$ and $\kappa_\infty(X)$ are used, and Example 7.1 considers this issue for the 1-norm when $n < p$, which arises in feature selection. Regularisation in the 1-norm only is obtained by specifying $\lambda = \alpha_1$ and $\alpha_2 = 0$ in (3), in which case it is required to minimise

$$\|X\beta - y\|_2^2 + \lambda \|\beta\|_1 , \tag{33}$$

with respect to β, where λ is the regularisation parameter.

The discrete Picard condition was introduced in Section 4 and its importance in condition estimation, using the 2-norm, and Tikhonov regularisation were considered in detail. It is instructive to consider the equivalent condition when regularisation in the 1-norm is used, as shown in (33).

The maximum value of the effective condition number of the inverse problem It follows from (32) that this maximum value is obtained when the effective condition number of the forward problem attains its minimum value. Equation (31) shows that this minimum value is achieved for the vector β_0 that maximises $\|X\beta_0\|_1 / \|\beta_0\|_1$, that is, β_0 is proportional to e_r where r is the index of the column of X that has maximum 1-norm,

$$\beta_0 = ke_r, \quad y_0 = kXe_r, \quad \eta_1^{\text{fwd}}(X, \beta_0) = 1, \quad \eta_1^{\text{inv}}(X, y_0) = \kappa_1(X), \tag{34}$$

and k is a non-zero constant.

The minimum value of the effective condition number of the inverse problem It follows from (30) that this minimum value is attained when y_0 is proportional to e_r where r is the index of the column of $X^\dagger$ that has maximum 1-norm,

$$y_0 = ke_r, \quad \beta_0 = kX^\dagger e_r, \quad \eta_1^{\text{fwd}}(X, \beta_0) = \kappa_1(X), \quad \eta_1^{\text{inv}}(X, y_0) = 1,$$

and k is a non-zero constant.

These results for $\eta_p^{\text{inv}}(X, y)$, $p = 1, 2$, are shown in Table 1.

p	Approx. min. value of $\eta_p^{\text{inv}}(X, y)$	Approx. max. value of $\eta_p^{\text{inv}}(X, y)$
1	Dominant components of y lie along the columns of $X^\dagger$ whose 1-norms are large.	Dominant components of y lie along the columns of X whose 1-norms are large.
2	Dominant components of y lie along the last few columns of the left singular matrix U of X.	Dominant components of y lie along the first few columns of the left singular matrix U of X.

Table 1
The forms of y for which $\eta_p^{\text{inv}}(X, y)$, $p = 1, 2$, attains, approximately, its minimum and maximum values.

Example 7.1 The matrix X is of order 90×101, where the entries in the first column are one, the entries in columns 2 to 81 inclusive are drawn from the standard normal distribution, and the entries in the last 20 columns are independent and drawn from a Bernoulli distribution with probability of success

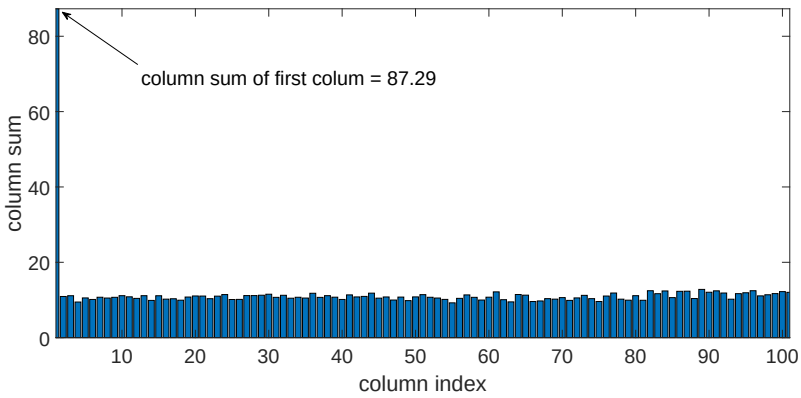

Fig. 8. The column sums of X.

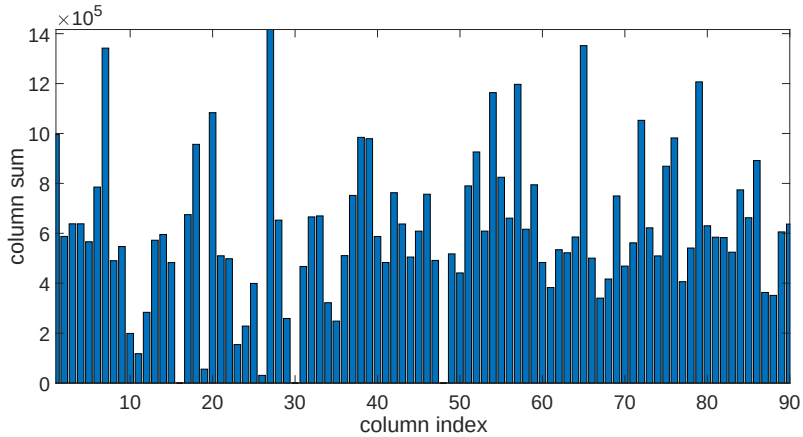

Fig. 9. The column sums of $X^\dagger$.

0.5. The entries in columns 2 to 101 inclusive are normalised, such that the mean and variance of the entries in each of these columns are zero and one, respectively. The condition number of X is $\kappa_1(X) = 1.24 \times 10^8$, and Figures 8 and 9 show the sum of the entries in each column of X and $X^\dagger$, respectively. It is seen that the 1-norm of the first column of X is 87.29, and that the 1-norm of the other columns is about 10. These norms show much greater variation for $X^\dagger$ because they range from 17.4 (column 16) to 1.42×10^6 (column 27).

A vector $y = y_1$ was chosen such that $\eta_1(X, y_1) = 4.46 \times 10^6$ and the coefficients of the minimum norm solution $\beta_1 = X^\dagger y_1$ are shown in Figure 10. It is seen that the leading coefficient of β_1 is equal to 1039 and that the magnitude of the other 100 coefficients is less than 50. This result is consistent with (34) because the column of X whose 1-norm is the maximum is the first column ($r = 1$ in (34)), and the leading coefficient of β_1 is about three orders of magnitude larger than the other coefficients.

The function `lasso` in MATLAB was used to regularise the solution by minimising (33). The mean square error was calculated using 10-fold cross-validation

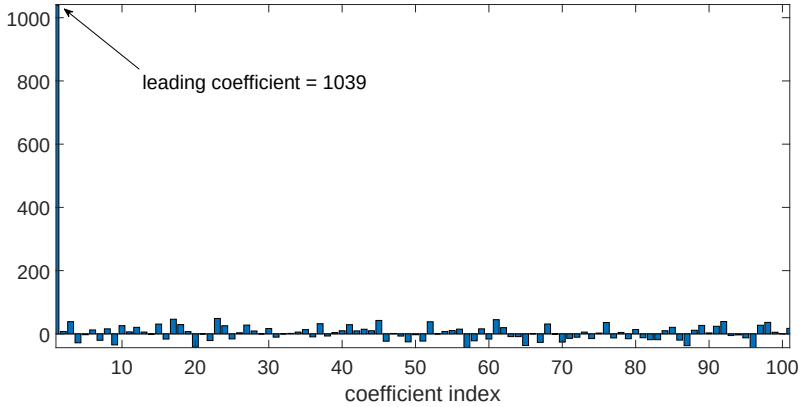

Fig. 10. The coefficients of the minimum norm solution $\beta_1 = X^\dagger y_1$.

and the value of λ^*, the regularisation parameter λ for which this error was a minimum, was $\lambda^* = 1.498$. The relative error in the regularised solution $\beta_1^{\text{reg}}(\lambda^*)$ was

$$\frac{\|X\beta_1^{\text{reg}}(\lambda^*) - y_1\|_2}{\|y_1\|_2} = 2.78 \times 10^{-2}. \tag{35}$$

The experiment was repeated with the same matrix X, but a vector $y = y_2$ was used, such that $\eta_1(X, y_2) = 22.03$. As described above for $y = y_1$, 10-fold cross-validation was used to calculate the optimal value, $\lambda^* = 1.725 \times 10^{-2}$, of the regularisation parameter λ. The relative error in the solution $\beta_2^{\text{reg}}(\lambda^*)$ was

$$\frac{\|X\beta_2^{\text{reg}}(\lambda^*) - y_2\|_2}{\|y_2\|_2} = 0.86, \tag{36}$$

which is large, and this regularised solution is therefore unsatisfactory. The maximum value of the error measures (35) and (36) is one because the coefficients of the regularised solution tend to zero as λ increases,

$$\|\beta_1^{\text{reg}}(\lambda)\|_p, \|\beta_2^{\text{reg}}(\lambda)\|_p \to 0 \quad \text{as} \quad \lambda \to \infty, \qquad p = 1, 2, \infty. \tag{37}$$

Figure 11 shows the variation of the relative error in the regularised solution with λ for $y = y_1$. The values of λ are not uniformly distributed between its minimum value ($\lambda = \lambda^* = 1.498$) and its maximum value ($\lambda = 763$). It is seen that the error is approximately equal to one for large values of λ, and this upper bound on the maximum value of the error as $\lambda \to \infty$, which follows from (35), (36) and (37), shows that the error (35) is 2.78% of its maximum value and the error (36) is 86% of its maximum value. □

The results in Example 7.1 are similar to the results in Example 3.1 because they show that regularisation can only be applied to (2) if specific conditions

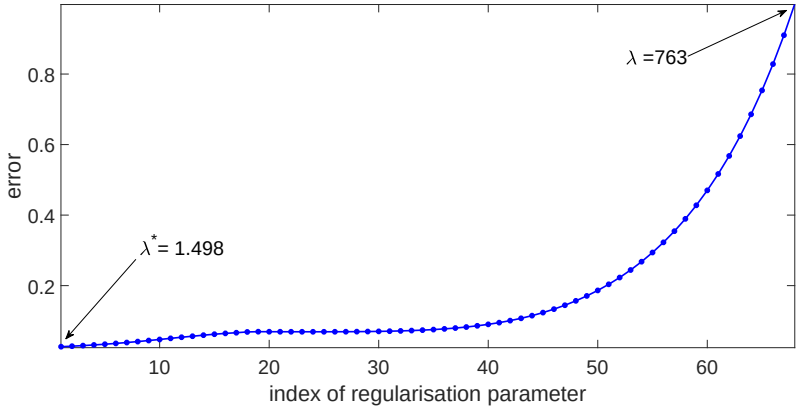

Fig. 11. The variation of the relative error in the regularised solution with the regularisation parameter λ for $y = y_1$. The optimal value λ^* of λ is marked in the figure.

between X and y are satisfied. In this circumstance, an approximate solution of (2) is a regularised solution because it is numerically stable and the regularisation error is small. If, however, these conditions are not satisfied, then the regularisation error is large and the computed solution is therefore unsatisfactory.

The smoothing function $\|\beta\|_2^2$ is the simplest regularisation function, but more general smoothing functions may need to be considered, for example, $\|L\beta\|_2^2$ where L is a discrete derivative operator. The simplest such operator is the first derivative operator,

$$L = \begin{bmatrix} 1 & -1 & & & & \\ & 1 & -1 & & & \\ & & 1 & -1 & & \\ & & & \ddots & \ddots & \\ & & & & 1 & -1 \end{bmatrix} \in \mathbb{R}^{(p-1)\times p},$$

and if $\alpha_1 = 0$ in (3), then the function to be minimised is

$$\|X\beta - y\|_2^2 + \lambda \|L\beta\|_2^2 ,$$

where λ is the regularisation parameter. This function is most conveniently considered by using the generalised singular value decomposition of the matrix pair (X, L) [4].

8 Summary

This paper has addressed condition estimation and error analysis of a regularised solution of (2), where the regularising constraint is imposed in the 1-norm or 2-norm. It was shown that the condition number of X is not a good measure of the stability of this equation, and that the effective condition number provides a better measure of this stability. This led to the establishment of the discrete Picard condition, the satisfaction of which implies that (2) is ill-conditioned. It was shown that if this condition is satisfied, then Tikhonov regularisation yields a solution that is numerically stable and has a small error.

Feature selection yields an underdetermined equation and regularisation in the 1-norm is usually applied because it is desired to compute a sparse solution of (2). It was shown that specific conditions between X and y must be satisfied in order that this sparse solution is a regularised solution. The development of expressions for condition estimation and error analysis is harder in the 1-norm than in the 2-norm because there does not exist a natural 1-norm matrix decomposition. It was shown that the column sums of X are important for analysis in the 1-norm, and this norm must be considered in more detail because of its extensive use in feature selection.

References

[1] E. Candes and T. Tao. The Dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics*, 35(6):2313–2351, 2007.

[2] J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *J. American Statistical Association*, 96(456):1348–1360, 2001.

[3] J. Friedman, T. Hastie, and R. Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22, 2010.

[4] G. H. Golub and C. F. Van Loan. *Matrix Computations*. John Hopkins University Press, Baltimore, USA, 2013.

[5] P. C. Hansen. The discrete Picard condition for discrete ill–posed problems. *BIT*, 30:658–672, 1990.

[6] P. C. Hansen. *Rank-Deficient and Discrete Ill-Posed Problems*. PhD thesis, Technical University of Denmark, Lyngby, Denmark, 1996.

[7] P. C. Hansen, J. G. Nagy, and D. P. O'Leary. *Deblurring Images: Matrices, Spectra, and Filtering*. SIAM, Philadelphia, USA, 2006.

[8] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. Springer, New York, USA, 2013.

[9] D. Kobak, J. Lommond, and B. Sanchez. Implicit ridge regularization provided by the minimum-norm least squares estimator when $n \ll p$. arXiv preprint arXiv:1805.10939, 2018.

[10] X. Lu and X. Wu. A strategy of searching active factors in supersaturated screening experiments. *J. Quality Technology*, 36(4):392–399, 2004.

[11] K. Murphy. *Machine Learning: A Probabilistic Perspective*. MIT Press, Cambridge, Massachusetts, USA, 2012.

[12] R. Tibshirani. Regression shrinkage and selection via the Lasso. *J. R. Statis. Soc. B*, 58(1):267–288, 1996.

[13] H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *J. R. Statist. Soc. B*, 67(2):301–320, 2005.